VIRTUALIZED HIGH PERFORMANCE COMPUTING INFRASTRUCTURE OF NOVOSIBIRSK SCIENTIFIC CENTER

A. Adakin, D. Chubarov, V. Nikultsev, ICT SB RAS, Novosibirsk, Russia
S. Belov, V. Kaplin, A. Sukharev, A. Zaytsev#, BINP SB RAS, Novosibirsk, Russia
N. Kuchin, S. Lomakin, ICM&MG SB RAS, Novosibirsk, Russia
V. Kalyuzhny, NSU, Novosibirsk, Russia

Abstract

Novosibirsk Scientific Center (NSC), also known worldwide as Akademgorodok, is one of the largest Russian scientific centers hosting Novosibirsk State University (NSU) and more than 35 research organizations of the Siberian Branch of Russian Academy of Sciences including Budker Institute of Nuclear Physics (BINP), Institute of Computational Technologies, and Institute of Computational Mathematics and Mathematical Geophysics (ICM&MG). Since each institute has specific requirements on the architecture of computing farms involved in its research field, currently we've got several computing facilities hosted by NSC institutes, each optimized for the particular set of tasks, of which the largest are the NSU Supercomputer Center, Siberian Supercomputer Center (ICM&MG), and a Grid Computing Facility of BINP.

A dedicated optical network with the initial bandwidth of 10 Gbps connecting these three facilities was built in order to make it possible to share the computing resources among the research communities, thus increasing the efficiency of operating the existing computing facilities and offering a common platform for building the computing infrastructure for future scientific projects. Unification of the computing infrastructure is achieved by extensive use of virtualization technology based on XEN and KVM platforms. The solution implemented was tested thoroughly within the computing environment of KEDR detector experiment which is being carried out at BINP, and foreseen to be applied to the use cases of other HEP experiments in the future.

INTRODUCTION

Over the last few years the computing infrastructure of Novosibirsk Scientific Center (NSC) located in Novosibirsk Akademgorodok [1] has improved dramatically as the new high performance computing facilities in Novosibirsk State University (NSU) [2] and various institutes of Siberian Branch of the Russian Academy of Sciences (SB RAS) [3] were established in order to be used as shared resources for scientific and educational purposes. The need for providing these facilities with the robust and reliable network infrastructure which would make it possible to share the storage and computing resources across the sites emerged instantly once the facilities have entered production. In 2008 a consortium of the following organizations:

- Institute of Computational Technologies (ICT) [4] hosting all the centralized scientific network infrastructure of SB RAS and Akademgorodok in particular,
- Novosibirsk State University (NSU) hosting NSU Supercomputer Center (NUSC) [5],
- Institute of Computational Mathematics and Mathematical Geophysics (ICM&MG) [6] hosting Siberian Supercomputer Center (SSCC) [7],
- Budker Institute of Nuclear Physics (BINP) [8] hosting a GRID computing facility (BINP/GCF) optimized for massive parallel data processing of HEP experiments (which is supposed to be used as a BINP RDIG [9] and WLCG [10] site in the near future)

was formed with the primary goal to build such a network (named later on as the NSC supercomputer network, or simply NSC/SCN) and provide it with the long term technical support. The first stage of the NSC/SCN infrastructure was deployed by ICT specialists in 2009 and it is being maintained ever since on 24x7 basis.

This contribution is focused on how the existing NSC/SCN infrastructure was used in order to build a virtualized computing environment on top of the NUSC and BINP/GCF facilities which is now exploited for running massive parallel data processing jobs related to KEDR detector experiment [11] being carried out at BINP. The prospected ways of using this environment for serving the needs of other BINP detector experiments and locally maintained GRID sites are also discussed.

NSC SUPERCOMPUTER NETWORK DESIGN AND IMPLEMENTATION

The supercomputer network as it is implemented now, has a star topology and based on 10 Gigabit Ethernet technology. As it is shown in Fig. 1, the central switch of the network is located in ICT and connected to each of the remote participating sites by means of two pairs of SMF G.652 fibers, two of which are equipped with the pair of long range (LR) 10 GbE optical transceivers and the remaining ones are used for two independent 1 Gbps wavelength-division multiplexing (WDM) technology based auxiliary control and monitoring links. The only exception is the link between the ICT and SSCC facilities which is less than 200 meters long and currently deployed over the MMF fiber equipped with the short range (SR) 10 GbE transceivers.

NSC/SCN is by design a well isolated private network which is not supposed to be directly exposed to the general purpose networks of the organizations involved.

#A.S.Zaytsev@inp.nsk.su

Each of the sites connected to the NSC/SCN infrastructure is equipped with the edge switch, used for the centralized access management and control, though the client side connections of the 10 Gbps uplinks are implemented individually on each site, reflecting the architectural differences between them. All the network links are continuously monitored by means of MRTG [12] instances deployed on sites.

The following approaches and protocols are now exploited for exposing computing and storage resources of the interconnected facilities to each other across the NSC supercomputer network:

- OSI Layer 3 (static) routing between the private IPv4 subnets,
- IEEE 802.1Q VLANs spanned across all the NSC/SCN switches,
- Higher level protocols for storage interconnect and also InfiniBand and RDMA interconnect across the sites over the Ethernet links (experimental).

RTT value observed between the sites in the network is less than 0.2 ms. The maximum data transfer rate between the sites for unidirectional TCP bulk transfer over the NSC/SCN network measured with Iperf [13] is equal to 9.4 Gbps. No redundancy implemented yet for the 10 Gbps links and the routing core of the network, but these features are planned to be added during the future upgrades of the infrastructure, along with increasing the maximum bandwidth of each link up to 20 Gbps, while preserving the low value of RTT for all of them. The prospects for extending the NSC/SCN network beyond Akademgorodok are also being considered.

NSC/SCN COMMON VIRTUALIZED COMPUTING ENVIRONMENT

Generic Design Overview

Since the early stages of deployment the NSC supercomputer network is interconnecting three facilities which are quite different from the point of view of their primary field of application and amount of resources available:

- NUSC at NSU: high performance computing (HPC) oriented facility (HP BladeSystem c7000 based solution running SLES 11 x86_64 [14] under control of PBS Pro [15] batch system) provided with 29 TFlops of combined computing power, 108 TB of total storage system capacity, and KVM [16] based virtualization solution,
- SSCC at ICM&MG: HPC oriented facility (30 TFlops of combined computing power plus 90 TB of total storage system capacity),
- GCF at BINP: parallel data processing and storage oriented facility (0.2 TFlops of computing power running SL5 x86_64 [17] plus 32 TB of centralized storage system capacity) provided with both XEN [18] and KVM based virtualization solutions.

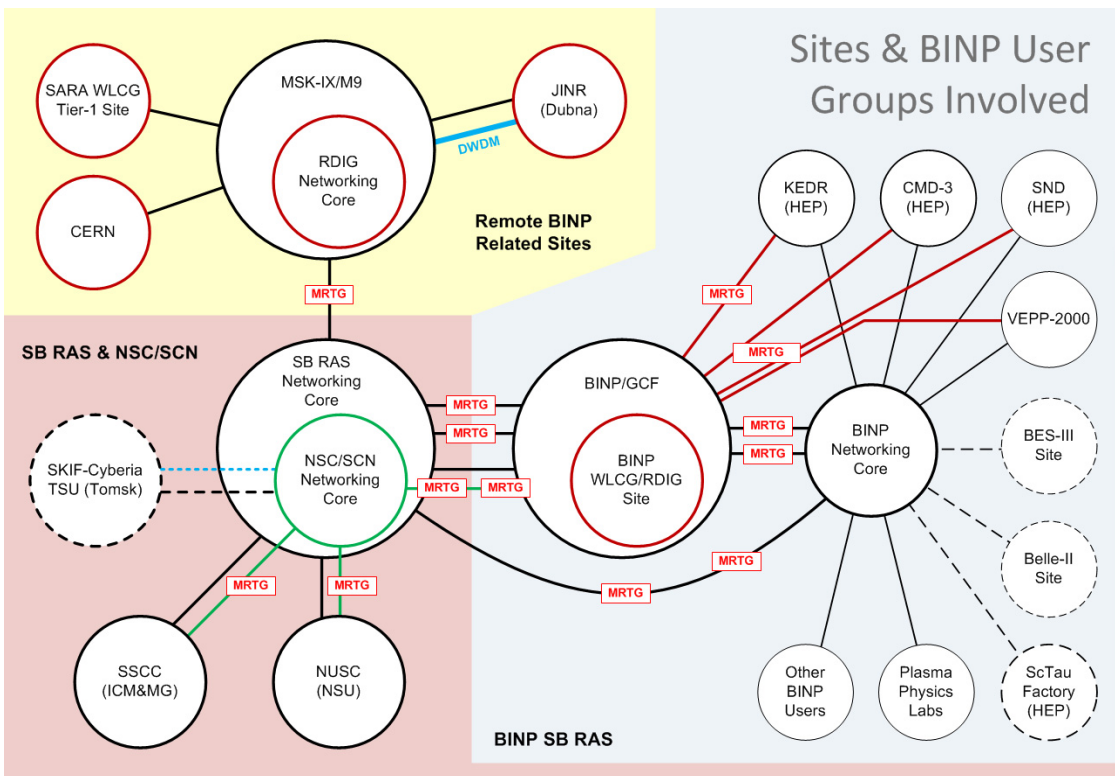

Figure 1: Networking layout of the NSC supercomputer network. Links to the user groups which are still to be established are represented by the dashed lines/circles, and the primary 10 GbE links – by the green lines. MRTG network statistics gathering points are also shown.

Considering the obvious imbalance of computing resources among the listed sites an initiative has emerged to use the NSC/SCN capabilities for sharing the storage resources of BINP/GCF with NUSC and SSCC facilities, and at the same time allow BINP user groups to access the computing resources of these facilities, thus creating a common virtualized computing environment on top of them. The computing environment of the largest user group supported by BINP/GCF (described in the next section) was selected for prototyping, early debugging and implementation of such an environment.

Computing Environment of KEDR Experiment

KEDR [11, 19] is a large scale particle detector experiment being carried out at VEPP-4M electron-positron collider [20] at BINP. The offline software of the experiment was being developing since late 90's. After several migrations the standard computing environment was frozen on Scientific Linux CERN 3 i386 [21] and no further migrations are expected in the future. The amount of software developed for KEDR experiment is about 350 kSLOC as estimated by the SLOCCount [22] tool with the default settings. The code is written mostly in Fortran (44%) and C/C++ (53%). The combined development effort invested into it is estimated to be more than 100 man-years. An overall size of experimental data recorded by KEDR detector since 2000 is 3.6 TB which are stored in a private format. Sun Grid Engine (SGE) [23] is utilized as a standard batch system of the experiment.

All the specific features of the computing environment mentioned here are making it extremely difficult to run KEDR event simulation and reconstruction jobs within the modern HPC environment of the NUSC and SSCC facilities, thus making it an ideal candidate for testing the common virtualized environment infrastructure deployed on top of the BINP/GCF and NUSC resources.

Implementation and Validation Procedures

The following candidates for a solution of the problem stated above were carefully evaluated at NUSC facility while trying to find an optimal configuration of the virtualized environment capable of providing both high efficiency of using the host system CPU power and the long term virtual machine stability at the same time:

- VMware server [24] based solution requiring minimal changes in the native OS of NUSC cluster – *ruled out due to the low performance and poor long term stability*,
- XEN based solution identical to the one deployed on BINP/GCF resources – *ruled out as it required running a modified version of Linux kernel which was not officially supported by the hardware vendor of the NUSC cluster,*
- KVM based solution which have shown the best performance and long term stability while running SLC3 based VMs among all the evaluated candidates and therefore *picked up for the final validation and running the KEDR detector production jobs*.

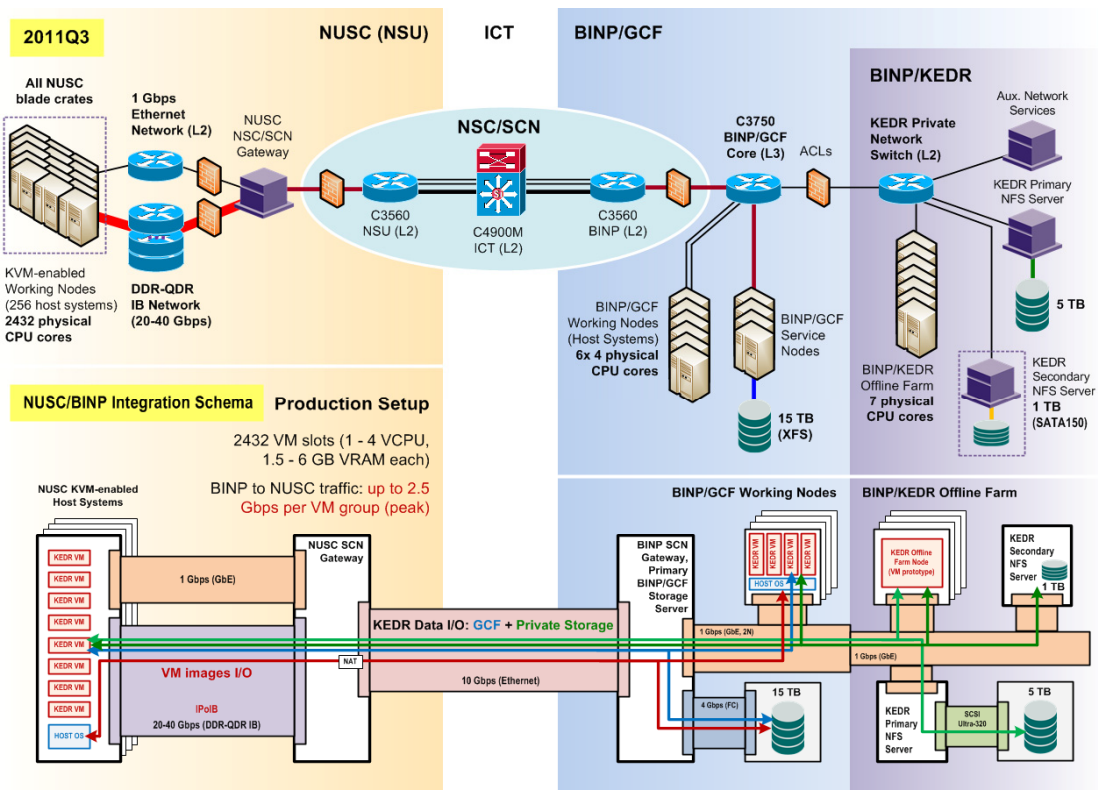

Figure 2: Networking and storage interconnect schema of the common virtualized computing environment spanning across the NUSC and BINP/GCF clusters as it is being used now for running the KEDR detector production jobs.

In addition the KVM based solution was validated during the large scale tests involving up to 512 dual VCPU virtual machines of KEDR experiment running up to 1024 experimental data processing and full detector simulation jobs in parallel controlled by the KEDR experiment private batch system.

All the stages of deployment of the KVM based virtualization environment on the computing nodes of NUSC cluster were automated and now handled via the standard PBS Pro batch system user interface. Furthermore, a generic integration mechanism has been deployed on top of the virtualization system which handles all the message exchange between the batch systems of KEDR experiment and NUSC cluster thus delivering a completely automated solution for the management of the NSC/SCN virtualized infrastructure.

The final layout of networking and storage interconnect schema developed for the KVM based virtualization environment deployed on the NUSC resources is shown in Fig. 2. Note that all the virtual machine images and input/output data samples are exported to the nodes of the NUSC cluster directly from BINP/GCF and KEDR experiment storage systems through the NSC/SCN infrastructure. It is foreseen that the similar solution is proposed to be deployed on SSCC side in 2011Q4 which would result in an increase of the amount of computing resources accessible via the NSC/SCN up to approximately 150 TFlops of computing power and 300 TB of shared storage capacity by the end of 2011 (taking into account the upgrades which are yet to be completed).

CONCLUSION

The supercomputer network of the Novosibirsk Scientific Center based on 10 Gigabit Ethernet technology which was built by the consortium of institutes located in Novosibirsk Akademgorodok, and currently provides a robust and high bandwidth interconnect for the largest local computer centres devoted to scientific and educational purposes. Although nowadays the NSC/SCN infrastructure is geographically localized within a circle of 1.5 km in diameter, it may be extended to the regional level in the future in order to reach the next nearest Siberian supercomputing sites.

The NSC supercomputer network once constructed made it possible to build various computing environments spanned across the resources of multiple participating computing sites, and in particular to implement a virtualization technology based environment for running typical HEP-specific massive parallel data processing jobs serving the needs of detector experiments being carried out at BINP. The solution implemented was tested thoroughly on a large scale within the computing environment of KEDR experiment in 2011Q1 and have been used for running production jobs ever since, delivering up to 75% of computing resources required by KEDR experiment over the last 9 months of continuous operation, resulting in a significant speed-up of the process of physical analysis, e.g. [25, 26].

Recently in 2011Q3 the other two local detector experiments (CMD-3 and SND detector at VEPP-2000 collider [27]) being carried out at BINP have successfully adopted the virtualization solution previously built for KEDR detector in order to satisfy their own needs for HPC resources. The solution obtained is foreseen to be used as a template solution for making a fraction of NUSC computing resources available for deployment of the gLite [28] worker nodes of BINP RDIG/WLCG resource site, Russian National Nanotechnology Network (NNN) resource site of SB RAS maintained by ICT, and also for prototyping the future TDAQ and offline data processing farms for the detector experiment at Super c-Tau Factory electron-positron collider proposed to be constructed at BINP over the upcoming 10 years. An experience obtained while building the virtualization based solution for running production jobs of BINP detector experiments which is compatible with modern high density HPC solutions, such as those deployed at NUSC and SSCC facilities might be of interest for other HEP experiments and WLCG sites across the globe.

ACKNOWLEDGEMENTS

This work is supported by the Ministry of Education and Science of the Russian Federation [29] and grants from the Russian Foundation for Basic Research [30].

REFERENCES

[1] http://en.wikipedia.org/wiki/Akademgorodok
[2] http://www.nsu.ru
[3] http://www.nsc.ru/en/
[4] http://www.ict.nsc.ru
[5] http://www.nusc.ru
[6] http://www.sscc.ru
[7] http://www2.sscc.ru
[8] http://www.inp.nsk.su
[9] http://www.egee-rdig.ru
[10] http://cern.ch/lcg/
[11] http://kedr.inp.nsk.su
[12] http://oss.oetiker.ch/mrtg/
[13] http://sourceforge.net/projects/iperf/
[14] http://www.novell.com/products/server/
[15] http://www.pbsgridworks.com
[16] http://www.linux-kvm.org
[17] http://www.scientificlinux.org
[18] http://www.xen.org
[19] NIM A478 (2002)420-425
[20] http://v4.inp.nsk.su
[21] http://linuxsoft.cern.ch
[22] http://www.dwheeler.com/sloccount/
[23] http://wikipedia.org/wiki/Sun_Grid_Engine/
[24] http://www.vmware.com/products/server/
[25] http://arxiv.org/abs/1109.4205
[26] http://arxiv.org/abs/1109.4215
[27] http://vepp2k.inp.nsk.su
[28] http://glite.cern.ch
[29] http://mon.gov.ru
[30] http://www.rfbr.ru