Building a High Performance Computing Infrastructure for Novosibirsk Scientific Center

A. Adakin\textsuperscript{1}, S. Belov\textsuperscript{2}, D. Chubarov\textsuperscript{1}, V. Kalyuzhny\textsuperscript{3}, V. Kaplin\textsuperscript{2}, N. Kuchin\textsuperscript{1}, S. Lomakin\textsuperscript{4}, V. Nikultsev\textsuperscript{1}, A. Sukharev\textsuperscript{2}, A. Zaytsev\textsuperscript{2,5}

\textsuperscript{1} Institute of Computational Technologies (ICT), Novosibirsk, Russia
\textsuperscript{2} Budker Institute of Nuclear Physics (BINP), Novosibirsk, Russia
\textsuperscript{3} Novosibirsk State University (NSU), Novosibirsk, Russia
\textsuperscript{4} Institute of Computational Mathematics and Mathematical Geophysics (ICM&MG), Novosibirsk, Russia
\textsuperscript{5} E-mail: A.S.Zaytsev@inp.nsk.su

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 (ICT), and Institute of Computational Mathematics and Mathematical Geophysics (ICM&MG). Since each institute has specific requirements on the architecture of the 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. Recently 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 of participating institutes, thus providing a common platform for building the computing infrastructure for various scientific projects. Unification of the computing infrastructure is achieved by extensive use of virtualization technologies 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 upcoming future.

1. 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).

The consortium 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 by them ever since on 24x7 basis.

This contribution is focused on the use of the existing NSC/SCN infrastructure for building a virtualized computing environment on top of NUSC and BINP/GCF facilities which is now exploited for running massive parallel data processing jobs related to KEDR detector experiment [11] at BINP. The prospected ways of using this environment for serving the needs of other local HEP experiments and BINP RDIG/WLCG site are also discussed.

2. NSC Supercomputer Network Design and Implementation

The supercomputer network as it is currently implemented has a star topology and based on 10 Gigabit Ethernet. As it is shown in Fig. 1, 2 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 WDM technology [12] based auxiliary control and monitoring links. The only exception is the link between 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 implemented as a well isolated private network which is not supposed to be directly exposed to the general purpose networks of host organizations by design. 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 [13] 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 and (experimental) InfiniBand and RDMA interconnect across the sites over the Ethernet links.

Since all the optical links involved are less than 2 km long over the fiber, the average ICMP ping 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 [14] 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 foreseen to be added during the future upgrades of the infrastructure, along with increase of 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 discussed.

3. Common Virtualized Computing Environment for BINP/GCF and NUSC Facilities

3.1. Generic Design Overview

Since the early stages of deployment the NSC supercomputing network is interconnecting three computing centers which are quite different from the point of view of their primary field of application and amount of resources available. These computing centers are:
**Figure 1.** Networking layout overview of the organizations participating in the NSC/SCN project. Links/user groups which are still to be established are shown by the dashed lines/circles, all the 10 Gbps links of the NSC supercomputer network – by the green lines, and virtual private network (VPN) links within the existing channels – by the red ones.

**Figure 2.** Detailed networking layout of BINP/GCF computing facility connected to the NSC/SCN infrastructure via dedicated optical 10 Gbps link (according to the state which is to be achieved by the end of 2011). The key components of BINP general purpose network (GPN) infrastructure are shown as well.
**Figure 3.** Layout of NUSC and BINP/GCF computing clusters interconnected by the NSC supercomputer network (status expected by the end of 2011Q1).

**Figure 4.** 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 event reconstruction and simulation jobs (present status).
• NUSC at NSU: high performance computing (HPC) oriented facility (HP BladeSystem c7000 based solution running SLES 11 x86_64 [15] under control of PBS Pro batch system [16], 13 TFlops of combined computing power plus 16 TB of total storage system capacity, including 4.3 TB on HP EVA based storage system),
• SSCC at ICM&MG: HPC oriented facility (17 TFlops of combined computing power),
• GCF at BINP: parallel data processing and storage oriented facility (0.2 TFlops of computing power running SL5 x86_64 [17] plus 25 TB of centralized storage system capacity) provided with XEN [18] and KVM [19] based virtualization solutions serving multiple independent BINP user groups (as it was illustrated above in Fig. 1).

Considering the obvious misbalance of computing and storage 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. Basic layout of the BINP/GCF data processing farm and NUSC computing cluster interconnected by the NSC/SCN 10 Gbps network is shown in Fig. 3.

3.2. Computing Environment of KEDR Detector Experiment
KEDR [11, 20] is a large scale particle detector experiment being carried out at VEPP-4M electron-positron collider [21] at BINP. The offline software of the experiment was being developing since late 90th. After several migrations the standard computing environment was frozen on Scientific Linux CERN 3 i386 [22] and no further migrations are planned for the future.

The amount of software developed for KEDR experiment is 350 kSLOC as estimated by the SLOCCCount [23] 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 about 100 man-years. An overall size of experimental runs data recorded since 2000 is 3.6 TB which are stored in a private format. Sun Grid Engine (SGE) [24] 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 facility, thus making it an ideal candidate for testing the common virtualized environment infrastructure deployed on top of the BINP/GCF and NUSC resources.

3.3. Implementation and Validation of the Solution Obtained
The following candidates for a solution of the problem stated above were carefully evaluated 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 stability at the same time:
• VMware server [25] 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 was picked up for the final validation and running the production jobs of the KEDR experiment.

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

The final state of networking and storage interconnect schema developed for the KVM based virtualization environment deployed on the NUSC resources is shown in Fig. 4. Note that all the virtual machine images and input/output data are exported to the nodes of the NUSC cluster directly from
BINP/GCF and KEDR experiment storage systems through the NSC/SCN infrastructure. All the stages of deployment of the KVM based virtualization environment on the host systems of NUSC cluster are now automated and handled via standard PBS Pro batch system user interface. Currently the integration mechanism between the KEDR experiment and NUSC cluster batch systems is being prototyped that would give us a completely automated solution for the management of the virtualized infrastructure of the BINP/GCF and NUSC facilities.

4. Conclusion
The supercomputer network of the Novosibirsk Scientific Center based on 10 Gigabit Ethernet technology which was built by the consortium of institutions 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-oriented massive parallel data processing jobs serving the needs of accelerator and particle physics experiments being carried out at BINP.

The solution implemented was tested thoroughly on a large scale within the computing environment of KEDR experiment implying up to 1024 data processing threads running simultaneously. It has been considered ready for production starting from 2011Q1. It is also foreseen to be used as a template solution for making a fraction of NUSC computing resources available for deployment of the gLite [26] Worker Nodes of the BINP RDIG/WLCG Site. An experience obtained while building the virtualization based solution for running computing tasks of the KEDR detector experiments which is compatible with the modern high density HPC solutions, such as the one exploited at NUSC or SSCC might be of interest for other HEP experiments and WLCG sites across the globe.

5. References
[1] Novosibirsk Akademgorodok: http://en.wikipedia.org/wiki/Akademgorodok
[2] Novosibirsk State University (NSU): http://www.nsu.ru
[3] Siberian Branch of the Russian Academy of Sciences (SB RAS): http://www.nsc.ru/en/
[4] Institute of Computational Technologies (ICT): http://www.ict.nsc.ru
[5] Novosibirsk State University Supercomputer Center (NUSC): http://www.nusc.ru
[6] Institute of Computational Mathematics and Mathematical Geophysics (ICM&MG): http://www.sscc.ru
[7] Siberian Supercomputer Center (SSCC): http://www2.sscc.ru
[8] Budker Institute of Nuclear Physics (BINP): http://www.inp.nsk.su
[9] Russian Data Intensive Grid Consortium (RDIG): http://www.egee-rdig.ru
[10] Worldwide LHC Computing Grid: http://cern.ch/lcg/
[11] KEDR detector experiment at BINP: http://kedr.inp.nsk.su
[12] WDM technology: http://en.wikipedia.org/wiki/Wavelength-division_multiplexing
[13] Multi Router Traffic Grapher (MRTG): http://oss.oetiker.ch/mrtg/
[14] Iperf project homepage: http://sourceforge.net/projects/iperf/
[15] SUSE Linux Enterprise Server (SLES): http://www.novell.com/products/server/
[16] PBS Pro batch system: http://www.pbsgridworks.com
[17] Scientific Linux (SL): http://www.scientificlinux.org
[18] XEN virtualization platform: http://www.xen.org
[19] KVM virtualization platform: http://www.linux-kvm.org
[20] V.V. Anashin at al., “Status of the KEDR detector”. NIM A478(2002)420-425.
[21] VEPP-4M collider at BINP: http://v4.inp.nsk.su
[22] Scientific Linux CERN (SLC): http://linuxsoft.cern.ch
[23] SLOCCount tool homepage: http://www.dwheeler.com/sloccount/
[24] Sun Grid Engine: http://wikipedia.org/wiki/Sun_Grid_Engine/
[25] VMware server virtualization solution: http://www.vmware.com/products/server/
[26] gLite Grid Middleware: http://glite.cern.ch