Virtualized High Performance Computing Infrastructure of Novosibirsk Scientific Center

Alexander Zaytsev (A.S.Zaytsev@inp.nsk.su)
Budker Institute of Nuclear Physics (Novosibirsk)
On behalf of the NSC/SCN consortium
(ICT, ICM&MG, BINP SB RAS, NSU)

ICALEPCS2011

Outline

- Organizations Involved
- HPC for Academic & Educational Centers
  - Local Trends in Russia
  - HPC Centers of Academgorodok (Novosibirsk, NSC)
  - Brief Overview of NSC/SCN (Supercomputer Network) Initiative
- HPC in HEP Related Activities of BudkerINP
  - System Integration Solutions
  - Running Production Analysis Jobs
  - Recent Results
- Lessons Learnt
- Future Plans
- Summary & Conclusion


Moscow
Novosibirsk
ICALEPCS2011
2500 km
2850 km
©2011 Google
©2011 Europa Technologies
©2011 Tele Atlas
US Dept of State Geographer

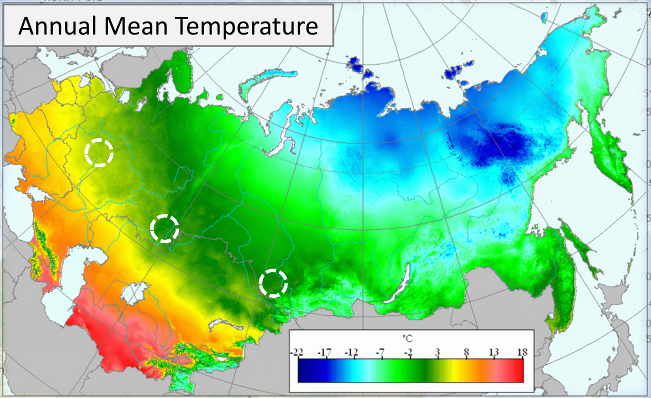
HPC in Russia (2011)

Moscow and its vicinity: more than 1.5 PFlops
- **"Lomonosov" SC (MSU): 1.37 PFlops (TOP-13)**
- **JSC (MVS-100K): 0.14 PFlops (TOP-76)**
- **Kurchatov Institute: 0.12 PFlops (TOP-85)**

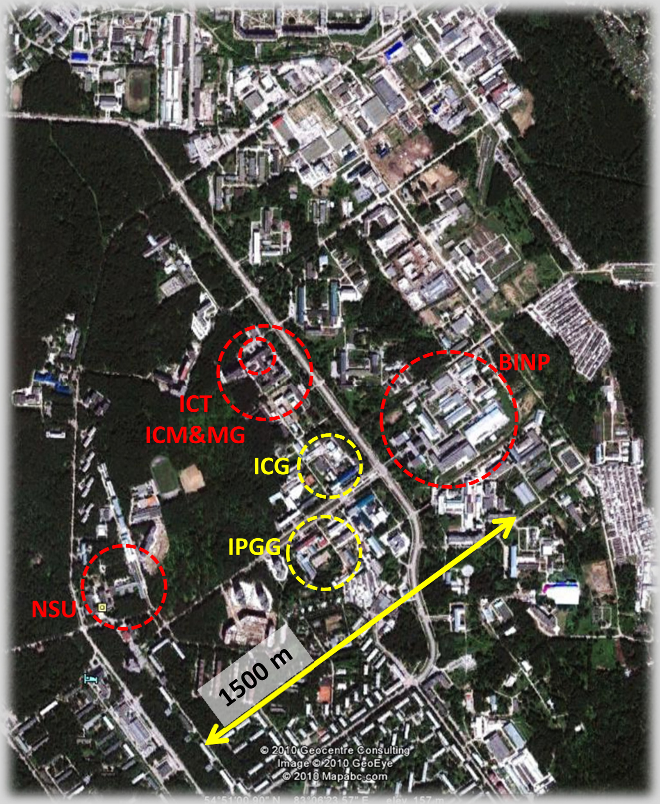
1st scientific DWDM-based network in Russia: 2x 10 Gbps (200 km: MSK-IX – JINR, Dubna)

[prospected] URAL SCN: 150 TFlops
- **SKIF-Aurora (SKIF-4): 117 TFlops**
- **SKIF-URAL: 16 TFlops**
- **USATU: 20 TFlops**

NSC/SCN: 120 TFlops >> 213 TFlops
- **NUSC: 29 TFlops >> 37 TFlops**
- **SSCC: 30 TFlops >> 115 TFlops**
- **SKIF-Cyberia: 61 TFlops**



Novosibirsk Scientific Center (NSC)



- 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.
- Most of the NSC organizations involved in HPC activities are located within the range of distances 0.5-1.5 km only (LR transceiver range)
- Multi-fiber optical links are used to interconnect the sites (SMF is used on all the bandwidth-critical segments)

NSC Organizations Involved in HPC Related Activities

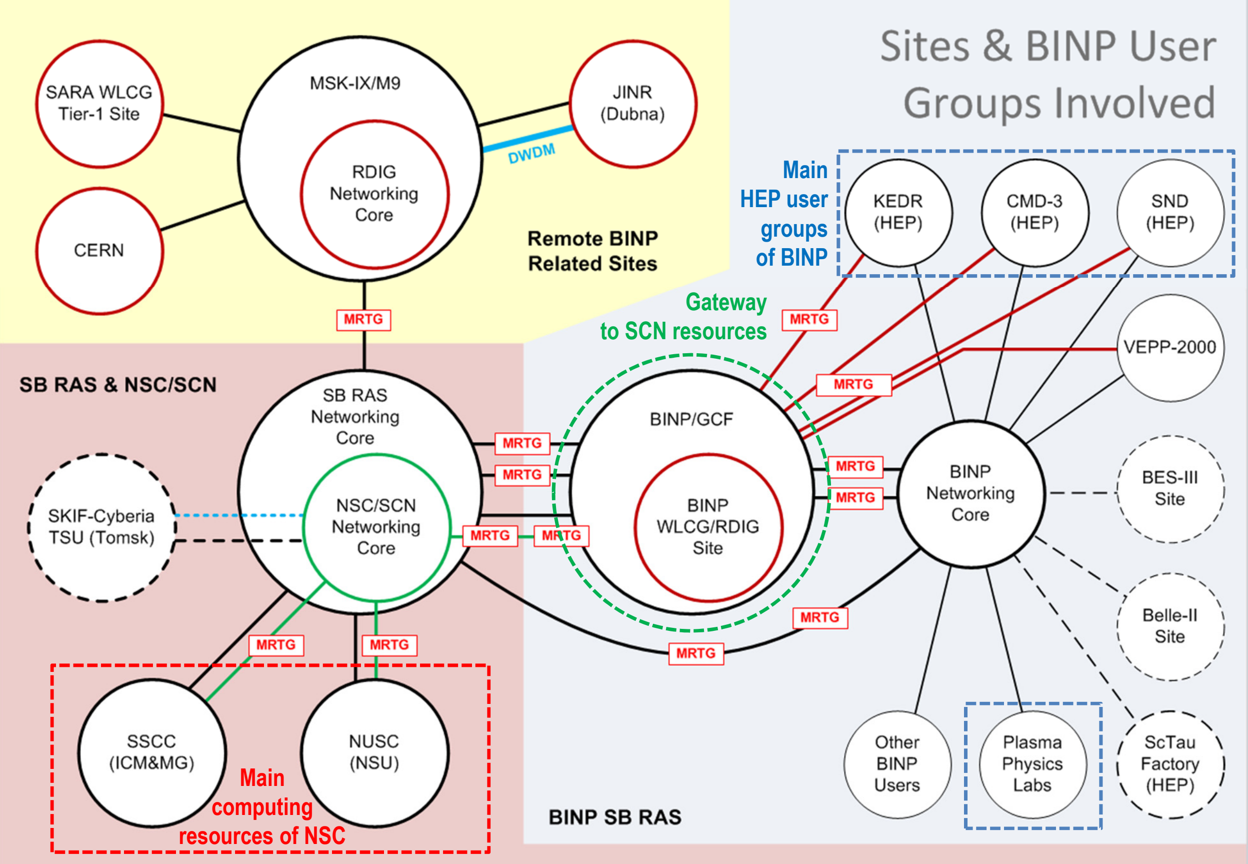
Hosting computing clusters and data storage systems:

- **Institute of Computational Technologies (ICT SB RAS)**
- Novosibirsk State University (NSU)
- Institute of Computational Mathematics and Mathematical Geophysics (ICM&MG SB RAS)
- Budker Institute of Nuclear Physics (BINP SB RAS)

Hosting user groups, exploiting the resources provided by NSC

- Institute of Cytology and Genetics (ICG SB RAS)
- Trofimuk Institute of Petroleum Geology and Geophysics (IPGG SB RAS)
- *more to join the list*

External Connectivity Schema for the Research Groups of BINP



Grid Computing Facility (GCF) at Budker Institute of Nuclear Physics (BINP)

BINP Grid Farm
CPU: 40 cores
RAM: 200 GB
HDD: 32 TB
(96 TB in 2011Q4)
Up to 80 XEN & KVM VM slots

BINP IT Facility
350 sq.m of raised floor space
Up to 250kVA (500kVA) power input (2N) & 35kW (400kW) of heat removal (N+1) capacity

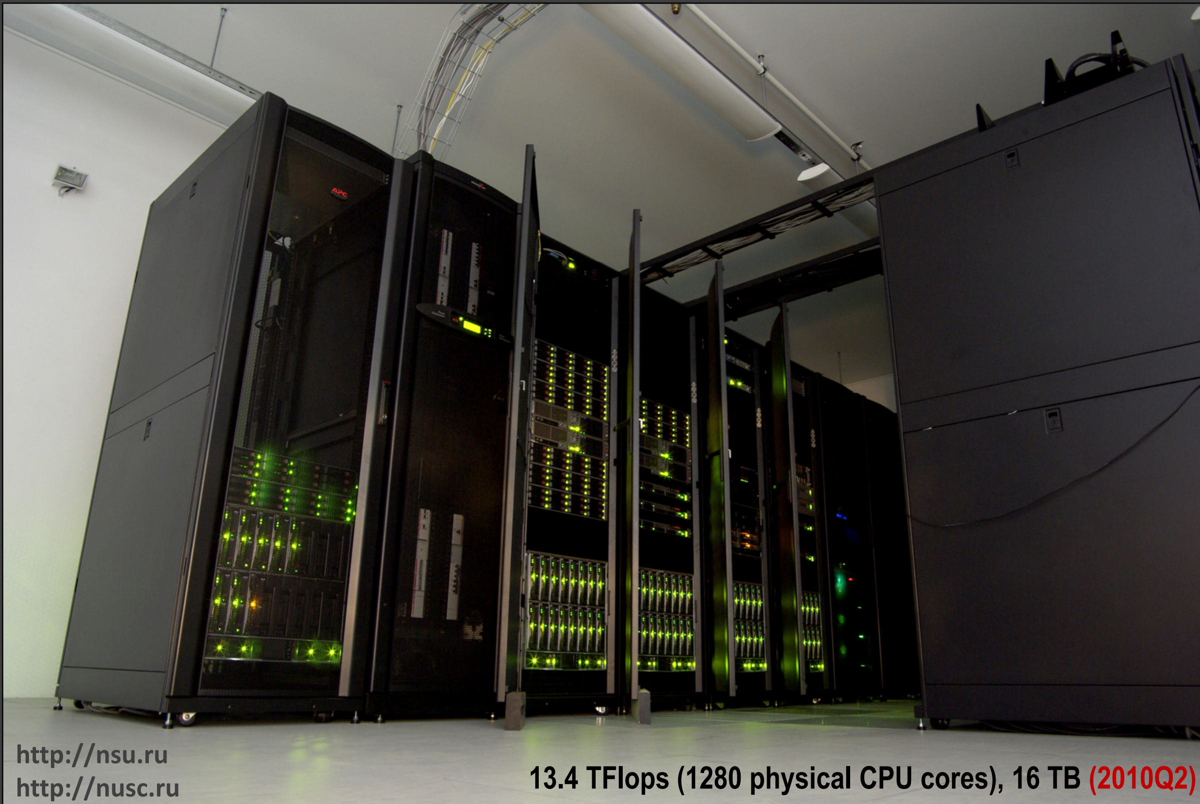
NSC/SCN Initiative

- The primary stages of the project are:
  - **STEP1:** Building a high bandwidth dedicated network infrastructure which interconnects the largest computing clusters of NSC
    - **10 GbE primary link + 2x auxiliary 1 Gbps links deployed in 2009Q2**
    - **The project was supported by RFBR grant (08-07-05031-b) and also by the internal funds of SB RAS**
  - **STEP2:** Provide the user groups across the NSC the means of accessing the interconnected clusters and performing bulk transfers of large amounts of data between the participating sites
    - **Done for BINP/GCF and NUSC since 2010Q3 – more details below**
    - **Combined effort of the participating sites**
  - **STEP3:** Create a common job handling environment by integrating the batch systems of participating clusters
    - **Done for BINP/GCF and NUSC since 2011Q1 – more details below**
    - **The same solution is proposed to be deployed on SSCC SB RAS computing resources in the near future**
    - **Many improvements are still to be made**
- The leading organization is ICT SB RAS since the very beginning and the most active participants of the project are: NUSC, BINP & SSCC

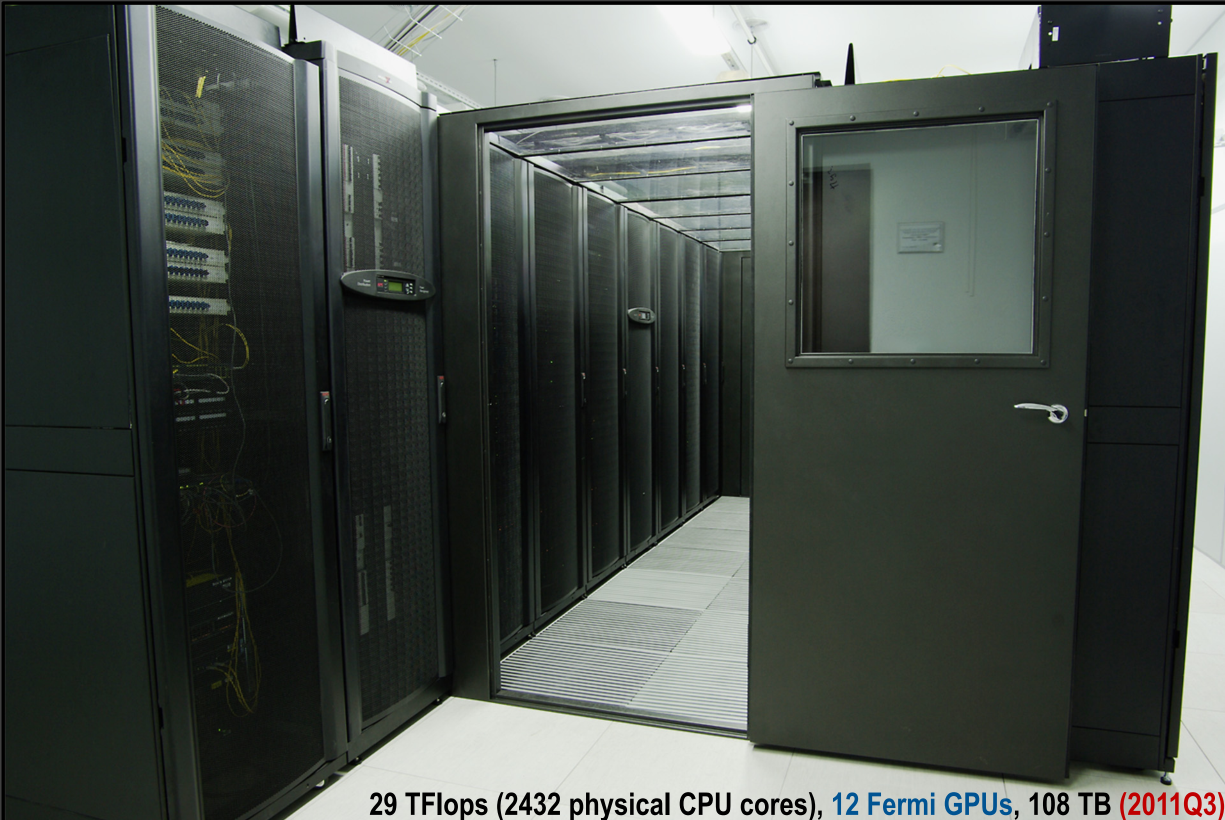
BINP/GCF Integration with NUSC (NSU)

- BINP is now supporting 3 particle detector experiments for electron-positron colliders:
  - CMD-3 and SND experiments installed at VEPP-2000 machine
  - KEDR experiment at VEPP-4M machine
    (the biggest detector experiment ever commissioned at BINP)
- Since all of these experiments were deployed before the SSCC and NUSC facilities have become available they were relying on local computing resources by design, thus they are experiencing now major difficulties with porting their software to the new environment
- The problem is proposed to be solved in three steps:
  - Virtualizing the execution environment of particular detector software by means of one of the commonly used virtualization platforms: VMware, XEN or KVM (performed by the detector experts)
  - Migrating the virtualized execution environment to the BINP Grid Farm which is acting like a gateway to the NSC/SCN resources
    (normally performed by the Grid Farm experts)
  - Replicating the VMs and exposing them to the NSC/SCN provided with the correspondent virtualization platform support
    (ideally with no detector experts involved)

Novosibirsk State University (NSU) Supercomputer Center (NUSC)

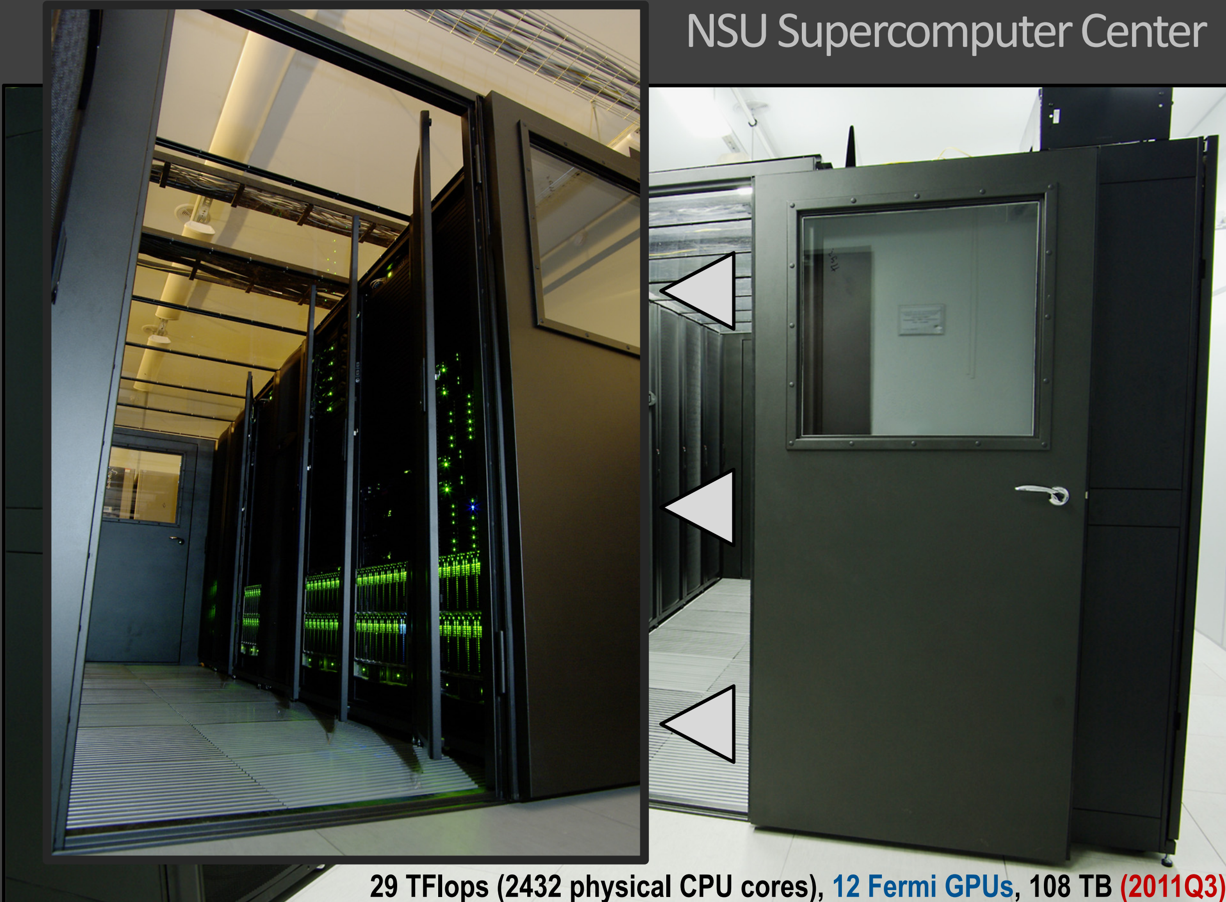


Novosibirsk State University (NSU) Supercomputer Center (NUSC)



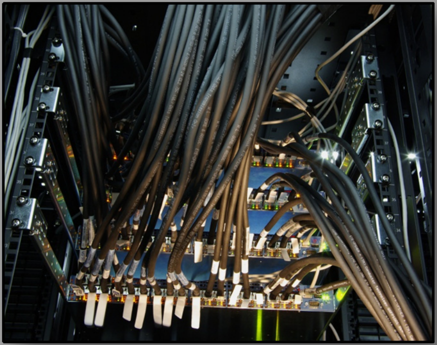
29 TFlops (2432 physical CPU cores), 12 Fermi GPUs, 108 TB (2011Q3)

NSU Supercomputer Center



29 TFlops (2432 physical CPU cores), 12 Fermi GPUs, 108 TB (2011Q3)

Novosibirsk State University Supercomputer Center

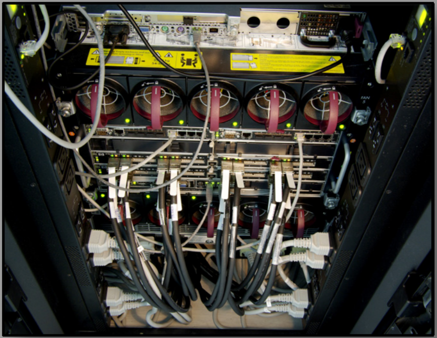


QDR Infiniband Interconnect (two-level “fat-tree“ topology)

◁

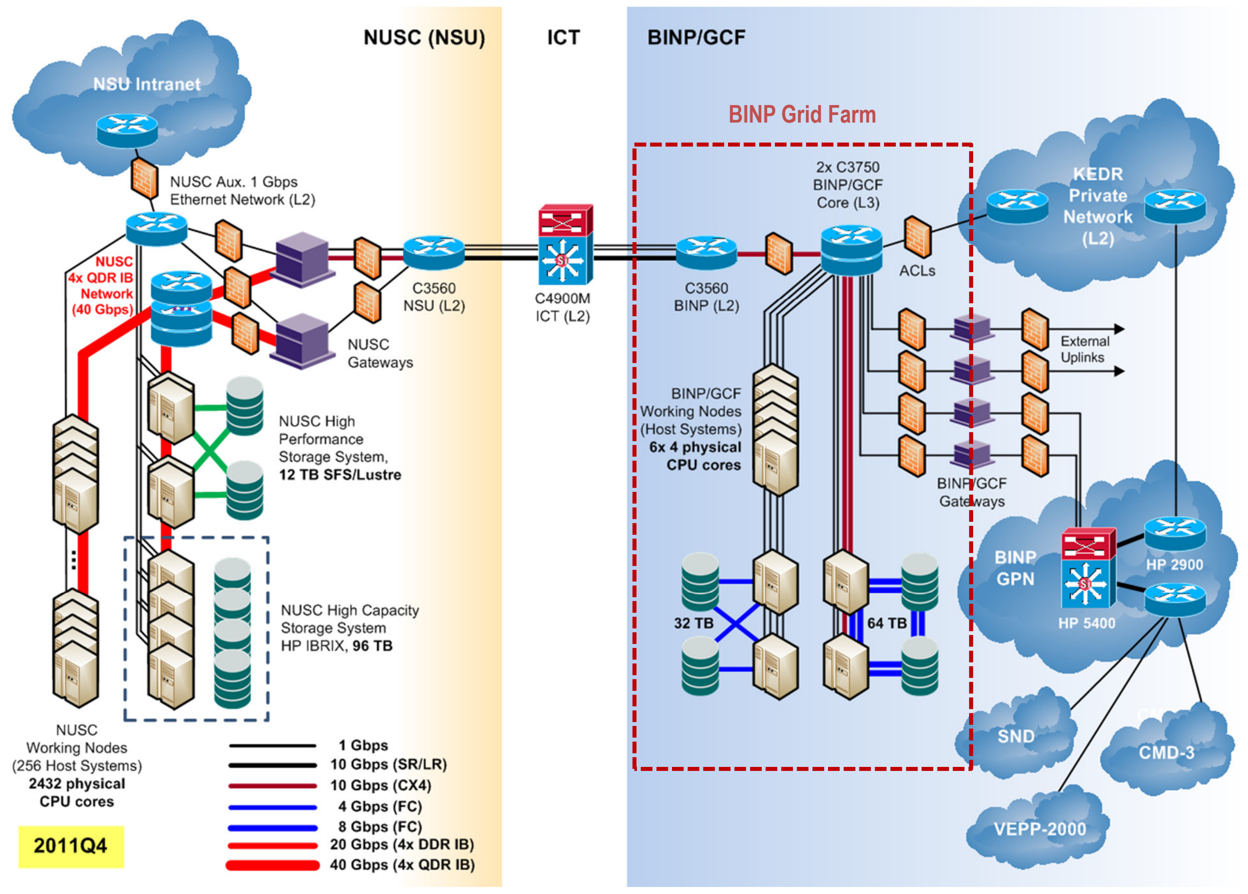
Recently added nodes, based on dual Xeon X5670 @ 2.93 GHz, 2 GB RAM/core (2011Q3)

▷



HP C7000 Blade Enclosure

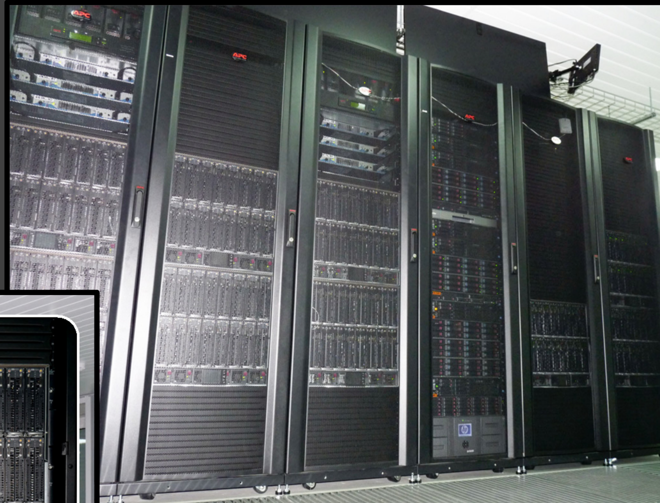
◁


NUSC (NSU)
ICT
BINP/GCF
NSU Intranet
NUSC Aux. 1 Gbps
Ethernet Network (L2)
NUSC
4x QDR IB
Network
(40 Gbps)
C3560
NSU (L2)
NUSC
Gateways
C4900M
ICT (L2)
BINP Grid Farm
2x C3750
BINP/GCF
Core (L3)
C3560
BINP (L2)
ACLs
KEDR
Private
Network
(L2)
External
Uplinks
BINP/GCF
Working Nodes
(Host Systems)
6x 4 physical
CPU cores
BINP/GCF
Gateways
NUSC High
Performance
Storage System,
12 TB SFS/Lustre
NUSC High Capacity
Storage System
HP IBRIX, 96 TB
32 TB
64 TB
BINP
GPN
HP 2900
HP 5400
SND
CMD-3
VEPP-2000
NUSC
Working Nodes
(256 Host Systems)
2432 physical
CPU cores
2011Q4
1 Gbps
10 Gbps (SR/LR)
10 Gbps (CX4)
4 Gbps (FC)
8 Gbps (FC)
20 Gbps (4x DDR IB)
40 Gbps (4x QDR IB)

Siberian Supercomputer Center (SSCC) at the Institute of Computational Mathematics & Mathematical Geophysics (ICM&MG)

SSCC was created in 2001 in order to provide computing resources for SB RAS research organizations and the external users (including the ones from industry)

30 TFlops of combined computing performance achieved in 2011Q3 (CPU)
+85 TFlops expected in 2011Q4 (GPU)



◁ NKS-30T △

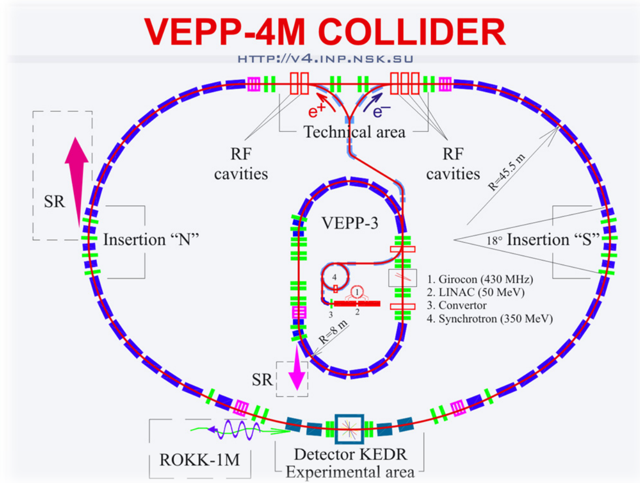
90 TB of local storage

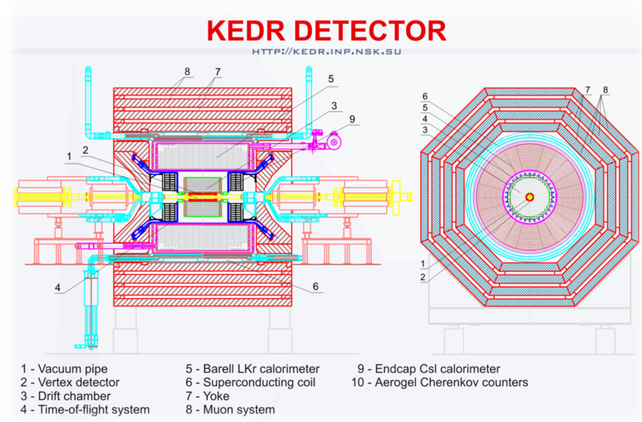
2x 70 sq.m of raised floor space

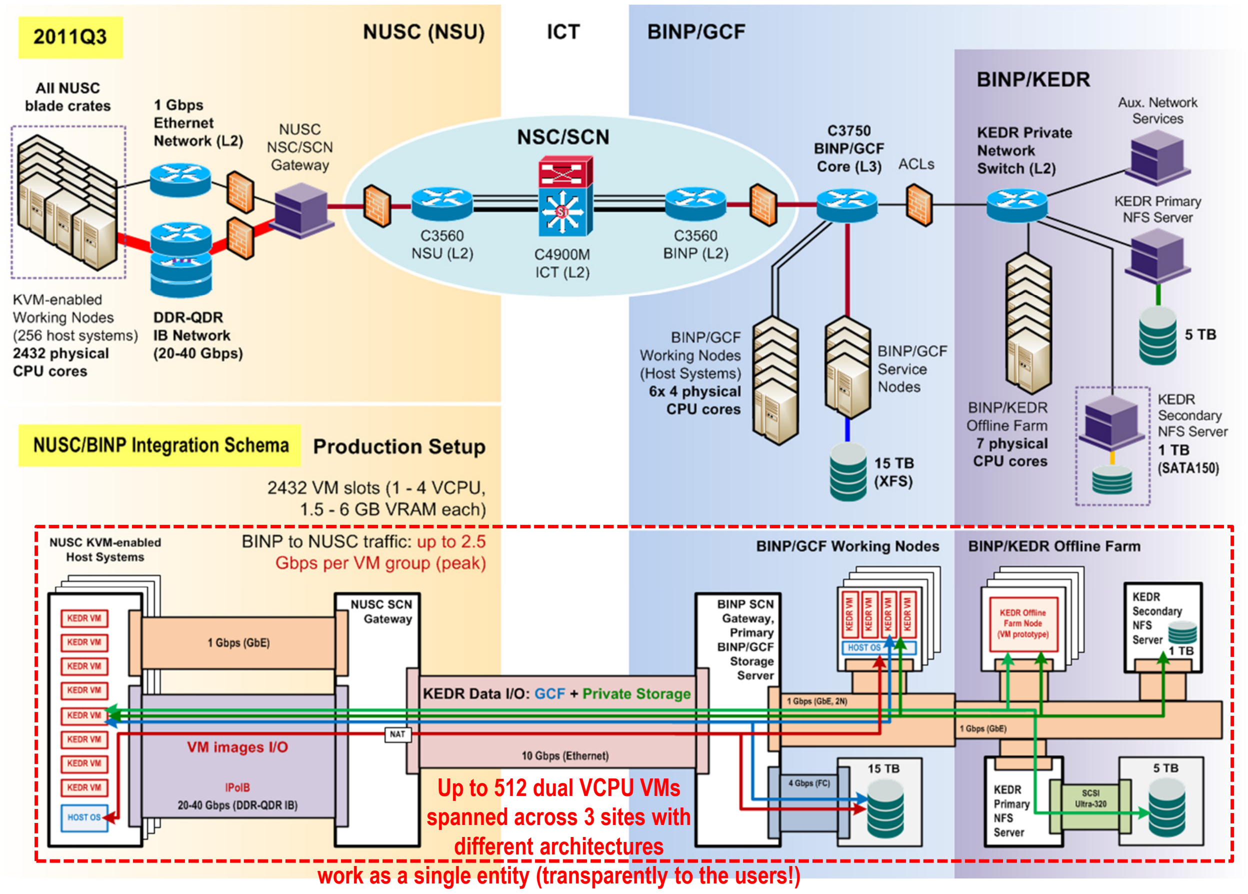
Up to 140kVA power input & 120kW of heat removal capacity (combined)

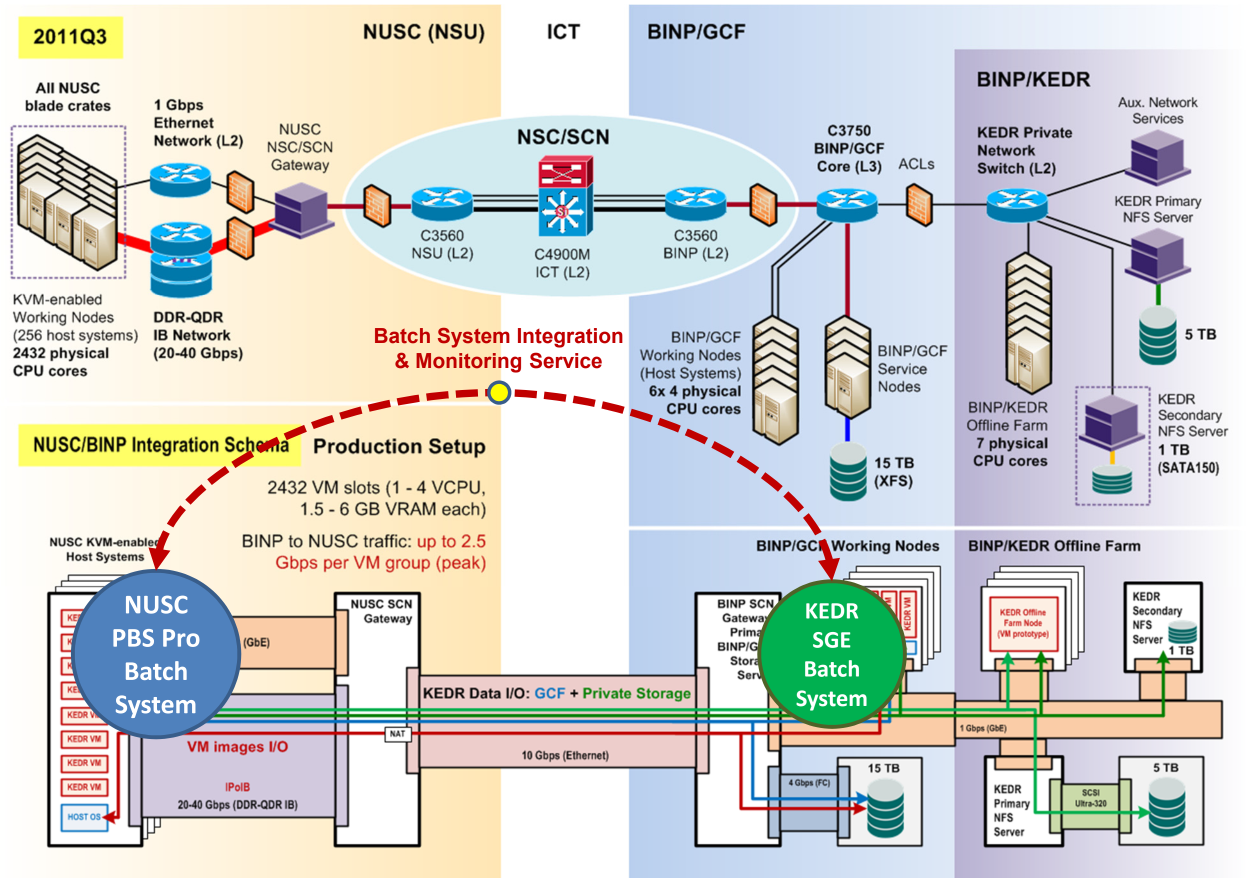
Testing the concept with KEDR experiment

- The procedure described above was applied for the KEDR detector data processing environment based on SLC3.i386 OS and a mixture of Fortran / C / C++ code being developed since 1995 which is now running under the brand new KVM-enabled SLES11.x86_64
  - VMware, XEN and KVM platforms were tested in 2009Q4-2010Q4 and the best stability and efficiency are obtained with KVM (with virtio drivers)
  - The attempt was a major success resulting in increase of computing resources available for KEDR experiment with the factor of 80 by using only 50% of full NUSC cluster capacity without making a single change in the KEDR offline reconstruction code (350 kSLOC , 100 man-years of development efforts)



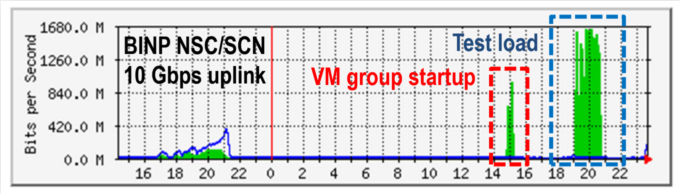


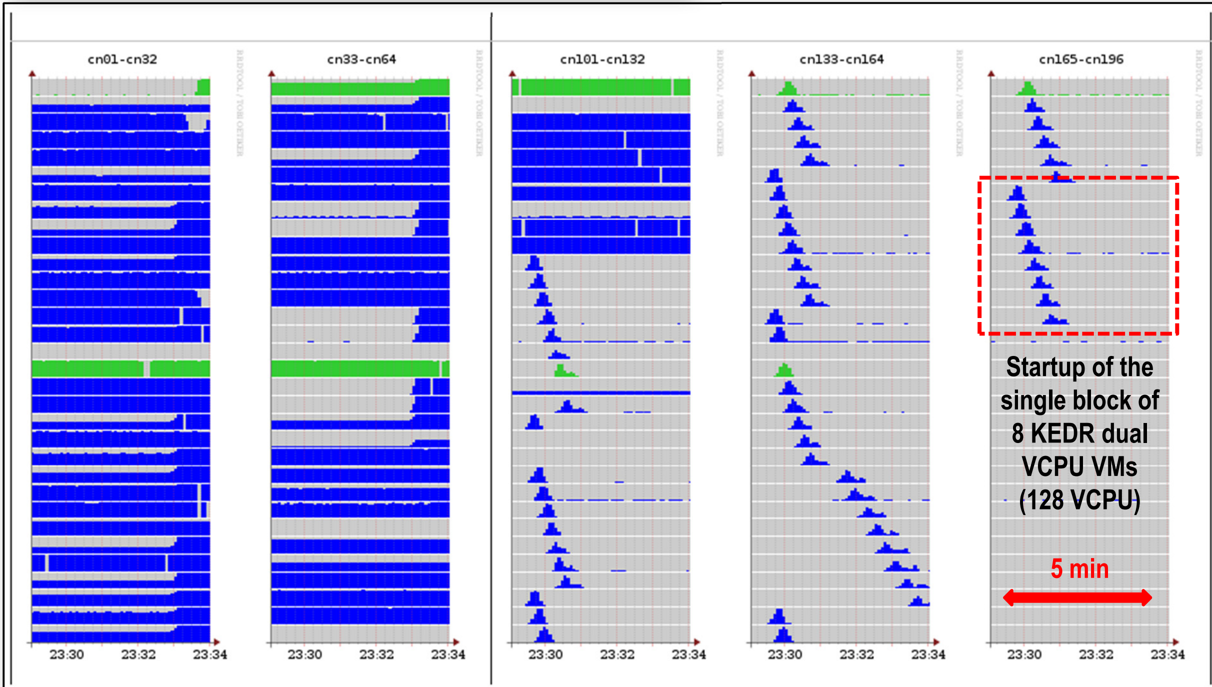

2011Q3
NUSC (NSU)
ICT
BINP/GCF
BINP/KEDR
All NUSC blade crates
1 Gbps Ethernet Network (L2)
NUSC NSC/SCN Gateway
NSC/SCN
C3750 BINP/GCF Core (L3)
ACLs
KEDR Private Network Switch (L2)
Aux. Network Services
KEDR Primary NFS Server
C3560 NSU (L2)
C4900M ICT (L2)
C3560 BINP (L2)
KVM-enabled Working Nodes (256 host systems) 2432 physical CPU cores
DDR-QDR IB Network (20-40 Gbps)
BINP/GCF Working Nodes (Host Systems) 6x 4 physical CPU cores
BINP/GCF Service Nodes
15 TB (XFS)
BINP/KEDR Offline Farm 7 physical CPU cores
5 TB
KEDR Secondary NFS Server 1 TB (SATA150)
NUSC/BINP Integration Schema
Production Setup
2432 VM slots (1 - 4 VCPU, 1.5 - 6 GB VRAM each)
NUSC KVM-enabled Host Systems
BINP to NUSC traffic: up to 2.5 Gbps per VM group (peak)
BINP/GCF Working Nodes
BINP/KEDR Offline Farm
KEDR VM
HOST OS
1 Gbps (GbE)
NUSC SCN Gateway
BINP SCN Gateway, Primary BINP/GCF Storage Server
KEDR Offline Farm Node (VM prototype)
KEDR Secondary NFS Server
1 TB
KEDR Data I/O: GCF + Private Storage
1 Gbps (GbE, 2N)
NAT
VM images I/O
10 Gbps (Ethernet)
IPoIB
20-40 Gbps (DDR-QDR IB)
4 Gbps (FC)
15 TB
KEDR Primary NFS Server
SCSI Ultra-320
5 TB
Up to 512 dual VCPU VMs spanned across 3 sites with different architectures work as a single entity (transparently to the users!)



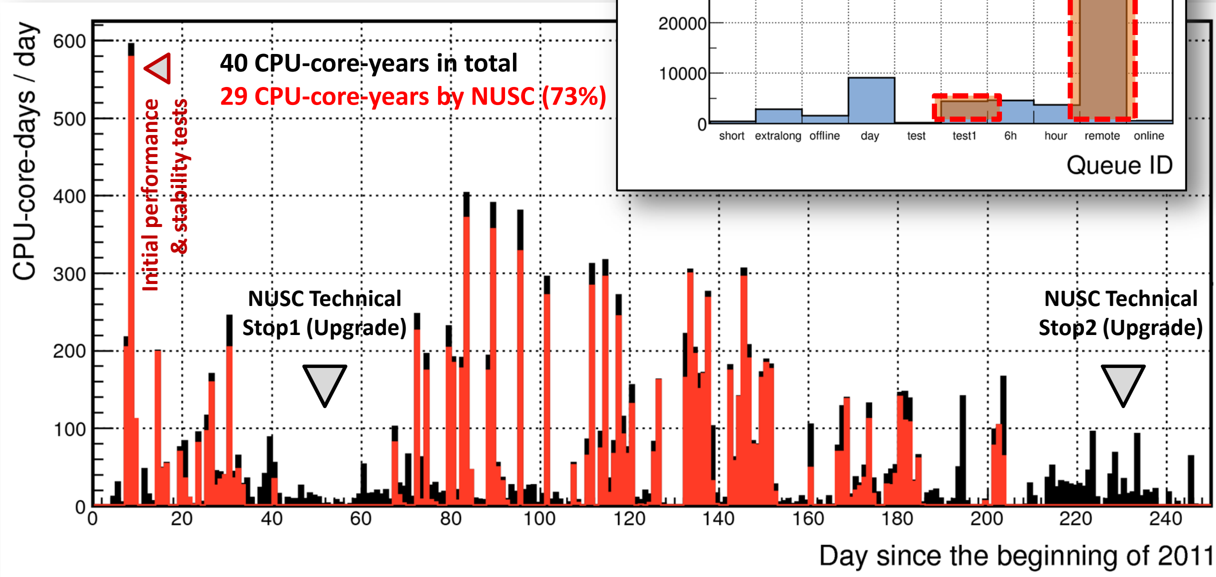
2011Q3
NUSC (NSU)
ICT
BINP/GCF
BINP/KEDR
All NUSC blade crates
1 Gbps Ethernet Network (L2)
NUSC NSC/SCN Gateway
NSC/SCN
C3560 NSU (L2)
C4900M ICT (L2)
C3560 BINP (L2)
C3750 BINP/GCF Core (L3)
ACLs
KEDR Private Network Switch (L2)
Aux. Network Services
KEDR Primary NFS Server
KVM-enabled Working Nodes (256 host systems) 2432 physical CPU cores
DDR-QDR IB Network (20-40 Gbps)
Batch System Integration & Monitoring Service
BINP/GCF Working Nodes (Host Systems) 6x 4 physical CPU cores
BINP/GCF Service Nodes
BINP/KEDR Offline Farm 7 physical CPU cores
5 TB
15 TB (XFS)
KEDR Secondary NFS Server 1 TB (SATA150)
NUSC/BINP Integration Schema
Production Setup
2432 VM slots (1 - 4 VCPU, 1.5 - 6 GB VRAM each)
BINP to NUSC traffic: up to 2.5 Gbps per VM group (peak)
NUSC KVM-enabled Host Systems
NUSC PBS Pro Batch System
KEDR SGE Batch System
BINP/GCF Working Nodes
BINP/KEDR Offline Farm
NUSC SCN Gateway
(GbE)
KEDR Data I/O: GCF + Private Storage
NAT
VM images I/O
IPoIB
20-40 Gbps (DDR-QDR IB)
10 Gbps (Ethernet)
KEDR VM
HOST OS
KEDR Offline Farm Node (VM prototype)
KEDR Secondary NFS Server
1 TB
1 Gbps (GbE)
4 Gbps (FC)
15 TB
5 TB
KEDR Primary NFS Server
SCSI Ultra-320

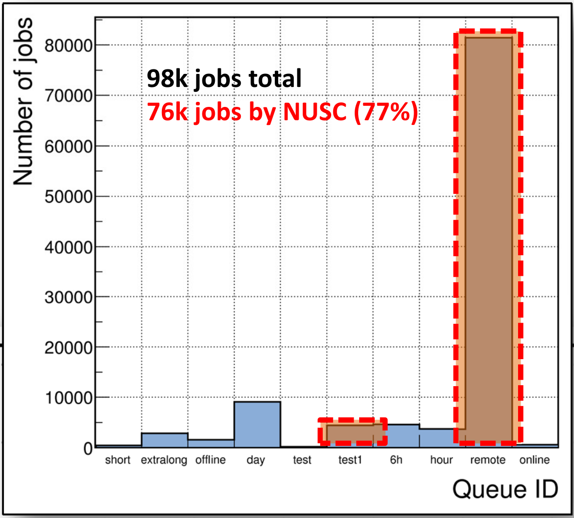
Sample test run with
8 blocks X 8 hosts X 8 dual
VCPU KEDR VMs = 1kVCPU
(NUSC, KVM, HT-enabled)



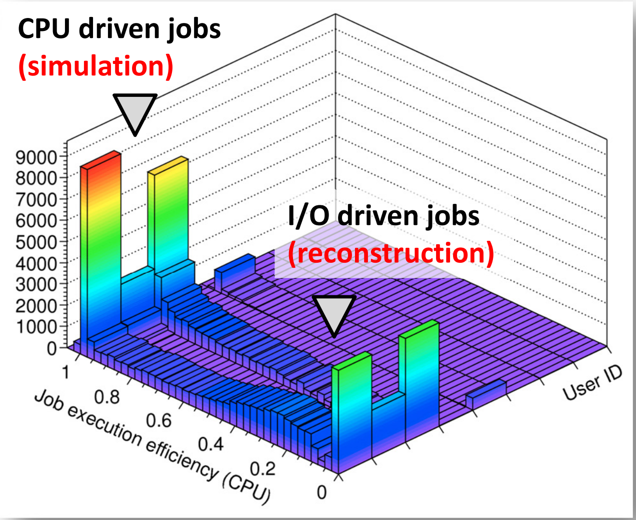


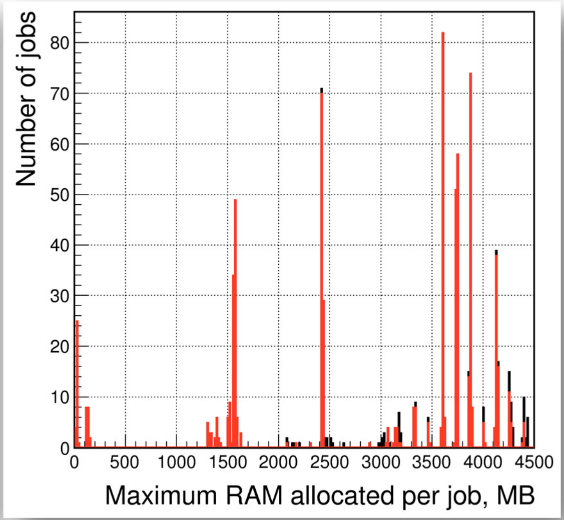
Running KEDR production analysis & simulation jobs in 2011 (BINP/GCF + NUSC)





KEDR Production Jobs Performance Achieved in 2011 (BINP/GCF + NUSC)





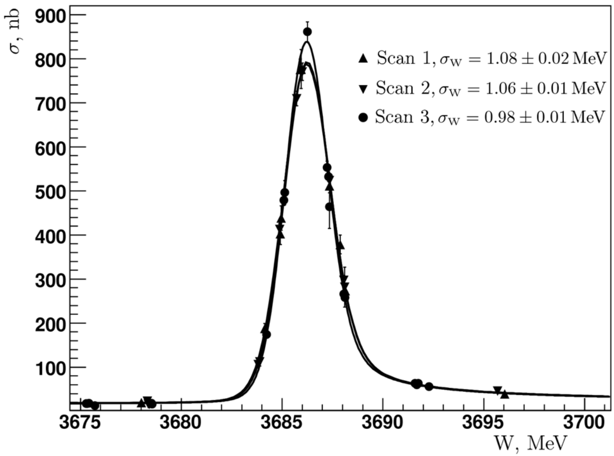
Further I/O optimization is required for the jobs dealing with the RAW experimental data (e.g. off-loading I/O intensive operations to BINP/GCF side).

KEDR analysis code is proven to increase performance by **15%** with 4 GB/CPU core and shows **+60%** overall performance gain with HT mode enabled.

Recent Results by KEDR Experiment (2011)

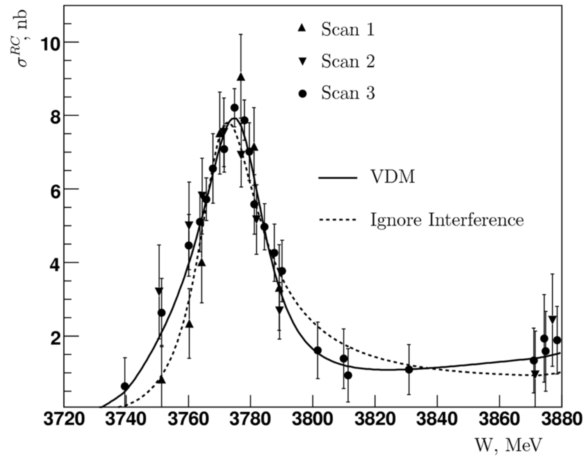
obtained by using NUSC resources via NSC/SCN

Measurement of main parameters of the ψ(2S) resonance: http://arxiv.org/abs/1109.4215



The multihadron cross section as a function of the c.m. energy for three scans in the ψ(2S) region

Measurement of ψ(3770) parameters: http://arxiv.org/abs/1109.4205



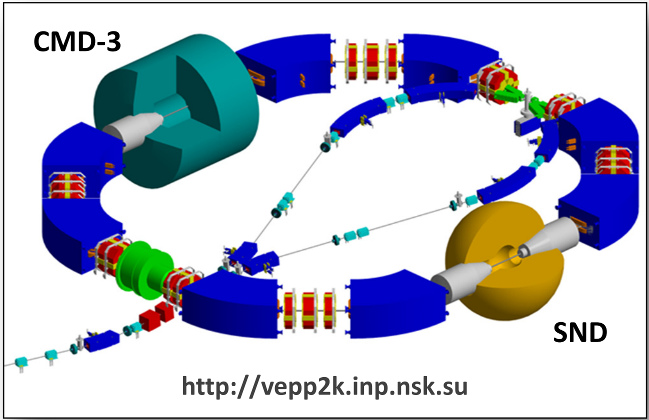
Cross section of ee → hadrons vs. c.m. energy in the vicinity of ψ(3770)

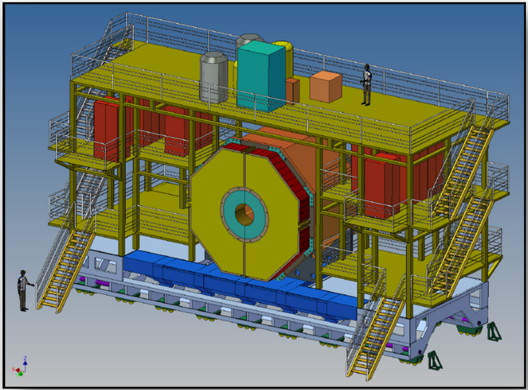
NUSC & BINP/GCF Integration: What We've Learnt So Far

- The virtualization solution works fine for the real life example of particle detector experiment:
  - Long term VM stability obtained: >1 month at NUSC, >1 year at BINP Grid Farm (lower limits)
  - Most of the underlying implementation details are hidden for the users
  - No changes were needed for detector offline reconstruction / simulation software and/or its execution environment
  - KEDR & NUSC batch system integration mechanism is in production since 2011Q1
- The solution obtained for KEDR detector applies to all particle experiments currently running at BINP (and maybe others as well)
- Main benefits:
  - Ability to freeze software and its execution environment and (exactly!) reproduce it when needed
  - Ability to use free capacity of supercomputer sites in order to run much more simple (from HPC point of view) single threaded detector software
  - Ability to eliminate the compatibility issues triggered by continuous OS and detector computing farm hardware upgrades (including migration to the extreme nodes)

Inviting More User Groups to Join the Activity

- VEPP-2000 collider and its detectors deployed at BINP:
  - Running reconstruction and simulation jobs for SND and CMD-3 detectors
  - Running ANSYS Multiphysics jobs for VEPP-2000 itself
- Super c-Tau Factory and its detector yet to be constructed at BINP:
  - Running simulation jobs
  - Prototyping software execution environment for the future ScTau TDAQ and offline data processing facilities





NSC/SCN: Future Plans

- Continue the development activities on **STEP3** of NSC/SCN project (batch system integration across participating sites)
  - **High performance storage system integration of NUSC, SSCC (HP IBRIX) and BINP/GCF (PVFS2/OrangeFS/Ceph) sites**
  - Further network performance tuning activities
  - Get more user groups involved in the NSC/SCN initiative
- Building 0.5-1.0 Gbps VPNs to RDIG networks in Moscow region (currently limited by 100 Mbps on the MSK-IX side):
  - Getting direct access to the Geant Network (GN3) resources
  - **Making it possible to deploy BINP RDIG/WLCG site & SB RAS National Nanotechnology Network (NNN) site components over the NSC/SCN resources**
- Finding resources for prospected future extensions of NSC/SCN network (10 GbE or multiple 10 GbE links over DWDM) to the following destinations:
  - **Tomsk (SKIF-Cyberia at TSU): 61 TFlops since 2011Q3**
  - Chelyabinsk (SKIF-Aurora at SUSU): 117 TFlops since 2011Q3

Summary & Conclusion

- The situation with the supercomputer infrastructure of Russia is dramatically improved over the last decade, especially in the European part of the country
- Though the issues with the lack of national broadband (DWDM-based) scientific networks are still where especially for the geo-graphically remote sites (such as Novosibirsk and its vicinity)
- NSC/SCN project represents an attempt to unify the computing resources of Novosibirsk (and hopefully, the entire Siberia in the future) by means of building the dedicated regional broadband networks between the major supercomputer sites and imple-menting a common computing and storage environment on top of the existing machines where needed
- NSC/SCN project has achieved a major success within the Novosibirsk Scientific Center up to the moment, and we are now looking for the ways how to extend its reach beyond NSC

Questions
& Discussion