

# Development of a Tier-1 computing cluster at National Research Centre 'Kurchatov Institute'

Igor Tkachenko

on behalf of the NRC-KI Tier-1 team

National Research Centre "Kurchatov Institute"

Moscow, Russian Federation

# Current resources

## Total

- 536 Tb of disk-based storage
- Job slots:
  - 1440 HT job slots with good network
  - 2880 HT job slots with slow network
- 2Pb of tape-based storage

## ALICE

- 720 HT job slots with good network
- 156 Tb of disk-based storage
- 0 Tb of tape-based storage ☹️

# Networking

- 2.5 Gbit/sec to GEANT, 1 Gbit/sec to Gloriad and 10+ Gbit/sec as the local connectivity within RU Grid.
- LHCOPN channel to CERN being set up:
  - MOS-AMS circuit is already established and we're solving last-mile problems.
  - MOS-BUD circuit is in the process of being established: there are some problems with DWDM equipment colocation at Wigner DC, but I believe they will be solved.
- It is said that Gloriad connection from US to Amsterdam will be bumped to 10 Gbit/sec this summer, so we will have an opportunity to have more bandwidth to US. This is TBD.
- We're in the process of talking with DANTE/GEANT about getting direct connectivity for RU WLCG community directly from GEANT (resurrecting GEANT PoP in Russia); it is slow and we do believe that it will appear only next year (if we will be able to agree on political and monetary conditions with DANTE).

# Networking

*(continued)*

- Direct peering with all major RU Grid sites: JINR, IHEP (Protvino), ITEP and PNPI. And we created our own peering equipment at M9 (largest traffic exchange site in Moscow), so other sites can peer with us without any problems.
- We had also doubled our networking capacity to M9 to accommodate bandwidth requirements.
- We're currently organizing LHCONE connectivity for all our sites. US peering should come first (circuit is ready, we're in the process of setting up BGP); NORDUnet VRF connectivity is on the radar. GEANT VRF will depend on the outcome from agreement between us and DANTE.

# Storage

## **Disk based**

- EOS based installation
- 6 pool nodes
- XFS file system
- Hardware raid6

## **Tape based (in developing)**

- dCache + enstore installation
- Native xrootd support
- Up to 800Tb of tapes for first time
- 6 cache nodes
- 3 tape drives

# Jobs

- Torque with Maui as scheduler
- CVMFS based VO software area with cluster of proxy servers
- SL6 at all WNs, mostly EMI-3 as the middleware.
- 24h fair share rates
- 2Gb of RAM per core
- 6Gb of vmem per core
- 24h limit of job execution time

# Cluster management system

- Puppet-based management
  - About 5000 rows of puppet code
  - Disabled automatically manifests appliance
  - implemented 40 own classes to support our resources
- Shell-based + kickstart-based system installation
- pdsh for task automation and parallel execution on many machines

# Resource allocation policies

- CPU – easy to give, easy to take back.

As it possible allocation

- Disk + tape easy give, hard to take back.

Allocation on demand: new resources may be allocated when current resource usage  $> 80\%$

# Monitoring

- Current workflow:
  - Alexandra Berezhnaya leads the team that walks through all defined monitoring endpoints, determines and classifies problems;
  - For known problems there are described ways to handle them, so it is done in almost "don't think, act" mode;
  - For new problems and ones that can't be processed by first line, they are passed to the free administrators who handle them to the end (or, at least, are trying to);
  - Problems on our and foreign resources are the same for us: if we see the latter, we're trying to understand it and, possibly open GGUS ticket or alike. We believe that this provides early alerting and some help to other sites as well as training for our own people.
- Workflow for 24x7 staff: basically, it will be the same, but current "Operation Bible" should be updated and expanded to allow for less educated staff to handle the majority of a known problems successfully.

# IPv6 readiness in KI

- April 2014 – start of IPv6 implementation
- April-July 2014 – IPv6 test in non Grid networks
- August 2014 – IPv6 will be ready in KI.

Are Alice and all other VOs will be ready to IPv6?

# Nearest future development

- Tapes for Alice – to the end of March - April
- Good network for all nodes – to the end of Jun
- Additional storage nodes – to the end of August

# Resources perspective (for all VO)

- About 10% of sum of all Tier-1
  - About 6000 job slots
  - About 6,3Pb of disk storage
  - About 7,5Pb of tape storage

Questions, comments?