Tier-1 in Kurchatov Institute: status before Run-2 and HPC integration

Alexandra Berezhnaya, Yury Lazin, Ilya Lyalin, Alexander Rogovskiy, Eygene Ryabinkin, Igor Tkachenko, Vasily Velikhov
National Research Centre "Kurchatov Institute"

Tier-1 activity

Since 2011 Kurchatov Institute and JINR were doing joint government-funded project for creating another Tier-1 centre for Large Hadron Collider.

In September 2012 WLCG Overview Board accepted proposition for Russian Tier-1. It was decided that Kurchatov Institute will host ALICE, ATLAS and LHCb, while JINR will work for CMS.

Planned resources for RRC-KI-T1 at the beginning of Run-2 were set to:

- 71000 HS06 of computing power (worker nodes)
- 6.3 PB of disk-based storage
- 7.4 PB of tape-based storage

During 2014 we agreed with c-RRB that resource split between ALICE, ATLAS and LHCb will be 40:40:20.

Under the hood

We currently run Torque and Maui for our computing field and at our current scale they work pretty well. Though we spent quite a long time (since our Tier-2 had overgrown something like 400 job slots) patching their sources and tuning LRMS components to cope with large number of jobs.

There was a project to move to Slurm (since we're very much familiar with it: all our HPC resources are using this batch system), but this is a work in progress and it isn't yet finished.

For disk-based storage we use dCache for ATLAS/LHCb and CERN EOS for ALICE. We have fully separated instances for each VO. Each dCache installation runs its own database cluster (PostgreSQL) with two nodes in a master/slave mode (2 pairs use shared storage and one runs with replication).

For tape-based instance we have dCache as a front-end software which handles protocols (GridFTP, xrootd, dCap) and access to disk pools that act as fast caches. On the backend we have Fermilab's Enstore which interfaces our tape library, IBM TS3500.

Some neat tricks

For stagein/stageout we use NFS which is shared across CREAM CE and worker nodes, but is mounted to the different places, say /var/cream on CE and /var/wn on WNs. This allows us to use Torque's `$usecp` directive and frees from hostbased SSH, while allowing to avoid keeping home directories for grid accounts at NFS too.

We use small file aggregation capabilities provided by Enstore, so tiny files are packed together into larger ones before they really go to the physical tapes. This greatly improves tape bandwidth.

For ALICE tape pools (which use xrootd), we had faced the problem that xrootd leaves zero-sized files when client drops connection without doing `close()`. There can be a lot of such files and putting them onto tape is not a good idea. And here dCache comes to the rescue: we have read-write pools that accept incoming files and from there large files are moved to the read-only pools that are connected to tape (new files are flushed to tape from them). And these pools are read-only, so ALICE can recall files from tape onto these pools and use them for reconstruction, analysis or whatever.

It costs almost nothing to make stateless services to be clustered via CARP (Common Address Redundancy Protocol): there is Linux implementation, it mostly works and its configuration is rather simple. For example, one can cluster BDII nodes easily: they are really stateless.

Squids can be clustered too and if all their clients sit in the same L2 domain (or you can make Squids to be present in multiple L2 domains), you can use FreeBSD's master/master CARP implementation. This makes Squids not only highly redundant, but also adds performance boost; especially with HTCP which allows cluster members to sniff each other's request results. CVMFS and Frontier clients are more than happy with this approach.

One can split networks into external ones (which are typically limited by the capacity of external channels, so ≈100 Gbit/sec) and internal network which provide high-bandwidth fabric (≈10 Tbit/sec) for moving data between SEs and WNs. That's easier to handle than an all-in-one network that should fulfill both requests. And it doesn't cost millions nowadays.

ATLAS

- We were validated back in 2012 without tapes: reprocessing of 2011's 2.76 TeV data yield 98% efficiency with 54 TB of data and 37 000 tasks
- Since then we rolled out tape system for ATLAS and added small file aggregation service that was used to verify our ability to store small log files on tape. ATLAS is happy with it and keen to use such a tool.
- Crunching ATLAS jobs daily for 3 years.
- With full-capacity hardware pool (that is now on-site and being installed) we will be fully validated once again in a coming month.
- Russian Tier-2 sites will form RU cloud: this will be a gradual process, but it should finish successfully in a coming trimester.

ALICE

- EOS is moving the data and we're mostly happy with it. It sometimes exhibit hot spots, but we're trying to identify and address them. At some point we will possibly do RAIN, but not now.
- Tape validation brought an interesting problem with dCache and zero-sized files, but we were able to pinpoint and fix it.
- ALICE is distributed by nature, so not really a new cloud. However, Predrag has some ideas about federations and we will be trying to taste it and possibly implement.
- We have an iteresting EOS testbed use-case that extends Geneva-Wigner case: sites from Russia will try to run distributed EOS instance and see what it will bring to the table and if such construction will be usable and reliable.

LHCb

LHCb mostly "just works". With the current expansion of storage we will be able to get the higher share of the data and excercise our LHCb worker nodes more thoroughly.

Networking

- Together with JINR we run LHCOPN ring Moscow – Amsterdam – Budapest – Moscow and capacity is 10G on each segment. Supplemented with CERN channels from Geneva and Budapest we have diamond-like topology that allows both Tier-1 run at 10G most of the time without interfering with each other.
- We have 10G local connectivity to most of big Russian Tier-2 sites (PNPI most likely will be upgraded to 10G during this year), so local cloud-like structure will rest on the sufficiently fat channels.
- We have access to LHCONE, we are planning to bring up new VRF for LHCONE. Unfortunately, we have problems with GEANT network, so our VRF will not be fully meshed.

Run-2 readiness

- All hardware we promised in the rollout plan is delivered to us and being installed and validated.
- Tape system is ready, new disk-based storage is rolled out, worker nodes are on the go too.
- MoU is still in the ministry that will at some point approve and sign it, but nobody really knows, when. We have some preliminary arrangements with VOs who care about MoU about QoS guarantees from the higher management of Kurchatov Institute.

You say "HPC"?

In Kurchatov Institute we also run supercomputing facilities and this is done inside the same division that runs Grid resources. In fact, we got our first supercomputer after 3 years of operation as a Grid-only shop.

For the long time these efforts were disjointed, though we were trying to apply experience gained in one field to another and vice versa. CVMFS on our HPC nodes is one of such examples, efforts to use Slurm for Grid is another one.

Interestingly enough, our supercomputers were always "not the real ones", like BlueGene or K: they were based on the commodity hardware, with local disks and outbound network access. Of course, we are using fast network fabric that for the past 7 years was InfiniBand: we started from DDR and are using all later incarnations. And our supercomputing clusters that are based on the "heavy" cores do look like normal Grid nodes, but with added InfiniBand.

And this year, together with expansion of our Tier-1 we are also getting new 500 TFLOPS cluster for HPC. And we have architectured both solutions to work together: worker nodes from HPC and Grid have just the same specifications (apart from InfiniBand cards).

To allow HPC worker nodes to reach Tier-1 storage we're employing IP over InfiniBand from the HPC side and InfiniBand-to-10GigE gateways made by Mellanox that are plugged into our 15 Tbit/sec internal switching fabric for Tier-1.

There are two HPC node images: for doing normal processing and for working in Grid environment. We won't be provisioning these nodes instantly (fully on-demand), but HPC resources will be available for Grid use on a rather short notice (like one or two days). They will be unpledged, they will be plugged directly into the worker node field, so no external Grid users will ever see other queues, resources or alike: this is completely transparent reshuffling.

Not only worker nodes have the same specifications, we run many similar core-level services (Squid, CVMFS, NAT, firewalling) on both Grid and HPC. So nodes in both domains will be able to use the same set of services familiar both to the workload they process and to the resource administrators.

This is just one strategy of using HPC resources on the Grid, there are others. For example, BigData laboratory in Kurchatov Institute together with ATLAS, ALICE and many other institutions are adopting workload and tools to be used in HPC environment. We also take part in this activity as the resource provider and system architects who are skilled both in Grid and HPC.

Diversity is good, though, so we're trying to provide our resources to our users via different ways in attempt to find the best ones for specific cases.

Conclusion

conclusion (kən-kloō'zhən) *n.*

1. The close or last part; the end or finish: *the conclusion of the festivities.*

Nope, it is not the end, it is just the beginning ;)

Learn Russian and then find out more at http://computing.kiae.ru/