ADC& Sites Configuration, Recommendations

Andrej Filipcic

Goal

- Investigate how sites are configured, what they plan, how they follow ATLAS recommendations
- Learn about interesting ideas that could be very useful for everybody
- Figure out how to improve in the near future

Questionnaire sent to sites

- status of the cluster(s), middleware
- status of migration to CentOS 7 or any other modern OS
- status of ipv6 deployment
- status and plans for container deployment and integration (singularity, docker,kubernetes ...), virtualization
- network connectivity (LAN, WAN, LHCOPN,LHCONE)
- storage status and future, and access protocols, especially the potential use of new technologies (ceph, objectstore, ...)
- storage caching (xcache, ARC cache, ...)
- federating with other sites, shared/distributed storage, diskless sites (no storage element)

cntd.

- using other, opportunistic resources in the region (HPC, clouds, national facilities, non-pledged clusters)
- non-standard resources, eg non-x86 CPU, GPU, FPGA...
- other VOs (experiments) constraints that might impact ATLAS computing
- local computing R&D, ATLAS or non-ATLAS, even non-WLCG related
- infrastructure changes, eg center reorganization, resource migration etc...
- suggestions for improvements or changes that ATLAS should try to incorporate in the computing model, especially the ones that could lower the cost of maintenance

Responses

- Responses from ~55 sites
  - Some were better organized and provided aggregated information for the clouds
- Much more than initially expected - big thanks to all that responded
- There was a lot of information provided, far too much to present in this talk
  - We will document and organize it for reference

Migration to CentOS 7

- Many sites already migrated in 2017 and 2018
  - Some fully
  - Some sites have new hardware on CC7, the old one to be upgraded or decommissioned
- Some sites that share resources are partially locked to RH6
  - Providing separate partitions for different OSes
- Most of the remaining RH6 sites plan to migrate to CentOS 7 by June/July 2019
- Many Big T1s and T2s are providing the resources through containers transparently - OS can be chosen on demand
- Few sites would like to migrate everything, but there are difficulties with sw availability, eg Storm on CC7
- **Compute nodes are critical for ATLAS - containers**

Nodes, Batch

- >10 sites (HPCs excluded) have no middleware on the nodes
  - running Nordugrid way where it is not required
  - Accessing middleware through containers
- Most of the sites have migrated to **HTCondor or SLURM**
  - some are planning to do so this or next year - in some cases it is coupled to migration to CentOS 7
  - **Other batch systems are considered deprecated by ATLAS**
- Some sites are planning to migrate to Kubernetes/OpenStack
  - Both compute and services
  - Also directly on bare metal
- Few sites use VAC
  - Under discussion how to use them with Harvester

Computing Elements

- **The recommended CEs for ATLAS are CondorCE and ARC-CE**
- There are many sites that still use Cream-CE, ~1/2
- The migration to recommended CEs is non-trivial
  - EGI accounting
  - multi-VO support, customizations and custom gateways between CE and batch
- Main reasons to migrate:
  - Support for modern features, eg cgroups, GPUs, containers...
  - Dynamic resource allocation: score vs mcore, and shares: analysis vs production
  - CE software development and support

Storage Elements

- Most of the sites use DPM, dCache and Storm
- Some EOS and Ceph Objectstore
- Objectstores have limited SE support for now
  - Custom developments to provide gridftp and xrootd
  - No clear recipe how to integrate them transparently in ATLAS DDM
- Protocols:
  - Many sites support >5 access, transfer protocols
- **ATLAS is using gridftp, https and xrootd**

Local storage

- Many sites have nodes with local disk without shared filesystem - WLCG standard architecture
- Several large sites (~20) are shared with non-LHC activities and provide HPC-like architecture
  - GPFS, Lustre, CEPHFS
  - Typically with Storm if used for Storage Element
- Some sites would like to use more modern storage, eg CEPH Objectstore or EOS
  - Not quite clear how to transparently integrate it in ATLAS
  - Funding concerns on RAW vs useful storage if erasure coding is not used

IPv6 Deployment

- Most of the Storage Elements are accessible through IPv6
- Nodes and grid services - less sites (~⅔)
  - This was not a WLCG requirement
- Some sites have difficulties with deployment:
  - Issues with peering to some sites - to be addressed in the near future
- Some sites cannot do it in the near future
  - Not a high priority within the hosting university or institute
  - Lack of networking equipment
  - Needs to be further investigated

Network connectivity

- Most of the sites are connected with 10Gb/s or 20Gb/s WAN links or even faster, many considering upgrading to 100Gb/s in the next 2 years
  - Some sites reported frequent network saturation and expressed a wish to further optimize the transfers, data and job placement
  - However: many smaller sites did not report and they likely have 1Gb/s or less
  - Some sites will decommission LHCONE while upgrading to higher GPN throughput
- Several sites have fast network (10 or 25Gb/s ethernet) or fast interconnects (Infiniband, OmniPath) between the nodes and to the storage
  - Typically on HPC-like shared facilities
  - Using MPI might be feasible on more sites than we expected

Non-standard resources

- Surprisingly, a large number of sites have deployed or could offer access to GPUs
  - 18 sites from few to few 100 - mostly used by local users
  - Typically dedicated nodes or co-located clusters
- Mostly no non-x86 chips, FPGAs...
- Some sites have expertise in GPUs and could contribute
- Some sites could purchase more GPUs if requested by ATLAS
- **ATLAS encourages sites to provide GPUs**
  - We need them to develop the software and algorithms for new architectures
  - But not to a level that would affect providing the CPU pledges for now

Containers

- Many sites have already deployed singularity, some docker:
  - Running pilots inside containers, eg modern OS with centos6 image/fs
  - Provisioning of batch nodes with docker
  - Some have it standby
- ATLAS plans to:
  - Run pilot2 on host OS
  - Execute the data staging and payload execution inside middleware and custom payload containers
- **Deploy singularity everywhere: (2.6.*)**
  - Possibility to run it directly from cvmfs if OS supports unprivileged namespaces

Storage federations and diskless sites

- Most of the sites do not plan to federate the storage (yet?)
  - concerns on WAN saturation
  - concerns on high IOPs to affect storage servers if direct I/O is used
- Some sites are diskless and connected to remote (close) Storage Element
  - Either direct access to SE from the nodes
  - Through ARC-cache and XCache - positive experience, but requires shared filesystem for ARC-cache or dedicated storage servers for XCache
- Many sites participate in DOMA activities

Opportunistic access

- Many sites provide opportunistic access to co-located HPCs (see HPC part of Jamboree)
- More sites could provide access to (smaller) university HPCs or cloud infrastructure
  - Might be easier to access it with Harvester and dedicated plugins for resources with complex access
- The sites on shared facilities typically have a share for ATLAS, not much room for opportunistic usage
- **Enable preemptive queues with EventService**
  - If there are free slots on the resources

Comments & Suggestions

- Many sites said bulk of the problems are caused by hardware, power failures, network outages - not much can be done on ADC side to help with that
- Various comments:
  - Improve documentation
  - More automation to identify site and central service issues
  - Distinguish pilots running on sub-clusters (opportunistic, cloud bursting...)
  - Payload distribution optimizations, throttling high I/O or network
  - Providing containers centrally for site services, storage
  - Control the network and join the effort with other VOs, non-LHC
  - Support for AAI
- To be followed up by ADC

Conclusions

- The responses from the sites were overwhelming, a lot of interesting information - a more detailed follow up is needed to process it all
- ATLAS Sites are much more diverse than we had expected
  - Especially in architecture, many sites are HPC like
- Most of the sites are following the ATLAS and WLCG recommendations
  - But it's hard to achieve 100%.
  - ATLAS might not be able to support everything in the future (eg using Centos 6 sites if production releases require it)
- New compute, storage and network technologies are coming fast. In addition to experimenting and testing, the recommendations and best practices need to follow soon.