ATLAS and Run-2

Andrej Filipčič
for the ADC

New Computing Model

- Less difference between Tier-1 and Tier-2
  - Long-lived data stored on stable SEs that are well connected
  - 90% availability (for analysis) – roughly 95 % of disk storage
  - Storage with <80% availability not used for data placement
  - Small Tier-2 sites (<100TB) not used for custodial ATLAS data – a size of a large dataset – fragmented SEs
- Getting rid of ATLAS tier hierarchy
  - Jobs will execute everywhere, based on input/output transfer cost – network integrated as a resource in the job brokering algorithms
  - Intermediate datasets left distributed over SEs, task chain outputs consolidated on destination SE
- Data Lifetime and extensive tape usage
  - All the datasets have a lifetime
    - RAW infinite
    - Others from 3 months to few years
  - Data untouched for a longer period of time migrated to tape
    - Tape mostly for archive storage
    - Expired datasets also removed from tape
  - ATLAS will likely need more tape than previously foreseen for Run-2
    - Analysis of Run-1 data takes longer than expected
    - Production on opportunistic resources requires more space

New services

- JEDI – dynamic job execution interface
  - Analysis and production use the same engine
  - Jobs generated in PanDA
  - Automatic lost-file recovery procedure
- ProdSys-2
  - Task definition interface
  - Integrates all production activities to the same interface
  - Task chains – automatic cleanup of intermediate datasets
  - Complex job workflows enabled
- Rucio
  - New data management system
  - Tight FTS-3 integration
  - Capable of managing much higher data and transfer volumes than DQ2

- All services in action since December 2014
- Initially many stability issues – now robust, stable, fast, flexible and ready for Run-2

Monitoring

- Many tools ATLAS dedicated, tailored to what ATLAS needs:
  - DDM Dashboard
  - Job Accounting Dashboard
  - BigPanda Monitor
  - Prodsys Monitor and user interface
  - DDM Accounting Dashboard
- Other monitoring tools not reliable enough or not useful for operations and accounting of resource usage
  - FAX dashboard, FTS dashboard, WLCG transfer dashboard
    - From time to time inconsistent with real activity.
  - REBUS OK for some information, unreliable for other (e.g. power of CPUs).
  - In most cases the issue is not with the tool itself but with the underlying infrastructure → limited usability

Production and Analysis

- Share:
  - T1 – 5% for analysis, T2 – 50% for analysis
  - Still fixed, but development of global fair-shares foreseen to manage it dynamically
- Production jobs:
  - For most, 2GB/core and 6-12 hours walltime is the target (possible with JEDI)
  - Some not so frequent jobs (eg. upgrade studies) will require extreme resources
    - More memory, 4-8GB
    - Longer execution time, 4-6 days
- Processing scale:
  - 1M jobs per day, expecting to go to 2M in few months
  - 1.2EB data processed in 2013, expecting 2 to 3 EB this year
    - 100GB/s or ~1GB/s per average site
  - Transfers between sites
    - 10-30GB/s
  - Transfers to tape
    - 1-3GB/s, including RAW distribution from CERN and tape migration from DISK

Multi vs Single core execution

- AthenaMP
  - Motivation: single process ATLAS reconstruction uses 4-5GB of RSS
  - Athena Processes share a lot of memory (COW)
  - Works with up to 32 cores
- MP Supported job types:
  - MC simulation
  - MC digitization + reconstruction
  - DATA reconstruction (including Tier-0 prompt reconstruction)
- NOT supported:
  - Merge jobs (HITS, AOD) – fast, low memory
  - Event generation (will be in the future)
  - Group production – data slimming, skimming, filtering for physics groups
  - User analysis

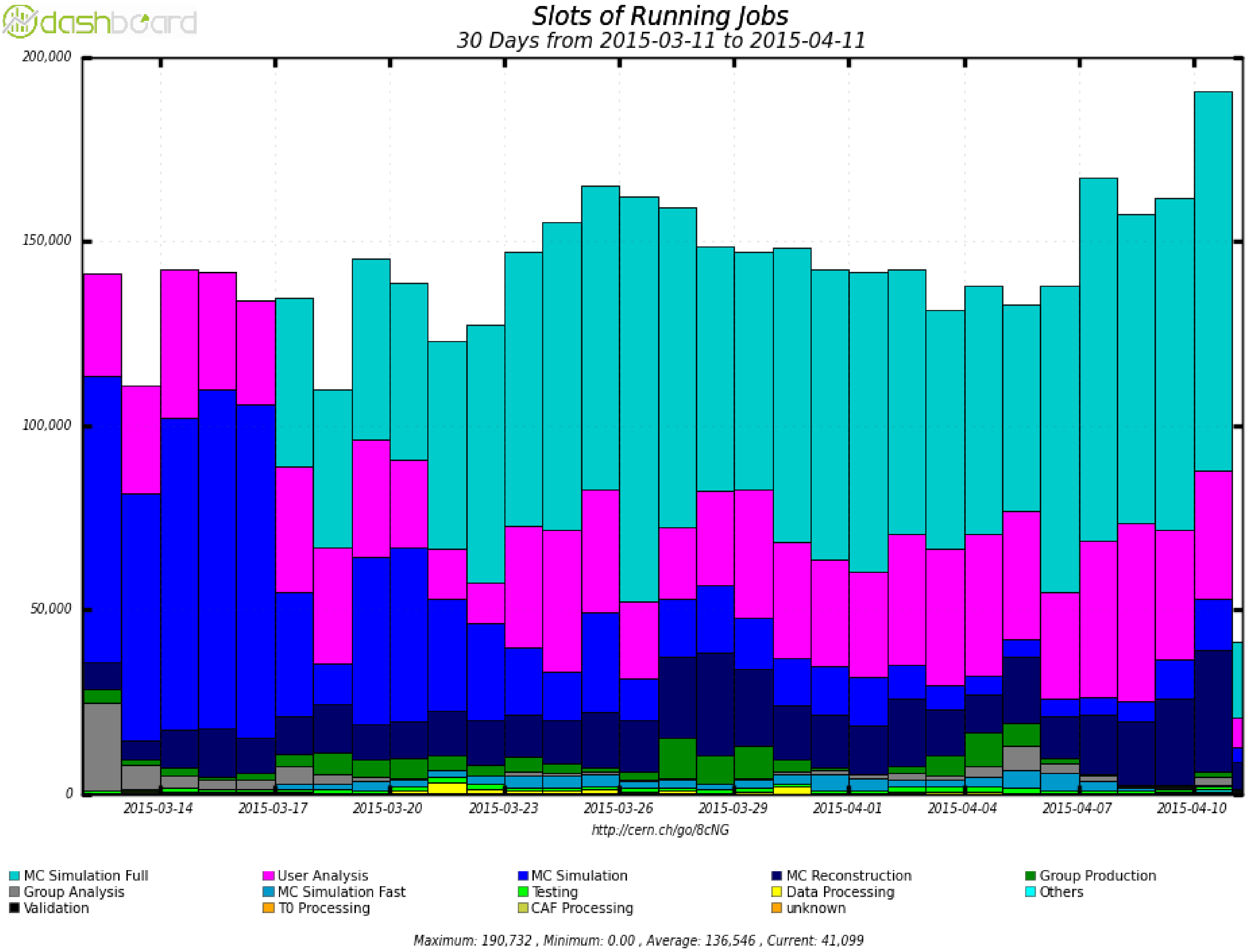
Production job resource usage

- Single core jobs will all use <2GB of RSS
  - Except in rare cases
- Multicore jobs:
  - MC simulation – 3GB of RSS for 8-core job (2GB on score)
  - MC digitization+reconstruction – 14-16GB of RSS for 8-core job (5GB on score)
    - There might be special cases which use more
- Wall vs CPU time:
  - Job initialization and completion (merging) spend ~15min in single-process mode
  - Target – 6h walltime (48h cputime) job – 96% efficiency

Multicore vs singlecore allocation

- 80% of CPU resources are expected to execute multi-core jobs
  - MC simulation, MC reconstruction, DATA (re)processing
  - Only analysis and group production expected to spend significant time in single-core mode
  - There will be periods of high requirements for single-core resources – e.g. group production campaigns, massive analysis before conferences – enforcing the share of mcore vs score is not recommended to sites
- Most of ATLAS resources have configured for multicore – <site>_MCORE queues
  - Some smaller Tier-2 sites still missing
- However, the experience from past 2 months:
  - ATLAS was not able to use more than 90k out of 180k maximum cpu slots in multicore mode
  - When no single-core load available – 50k slots left unused or unavailable to ATLAS
- Difficulties:
  - Expensive draining of slots for multicore execution
  - Static partitioning of mcore vs score in batch systems
  - Fair-sharing score vs mcore
  - Interference with score of other experiments reducing the mcore slots
- Not enough experience – further site optimizations needed
  - Planning a discussion with representatives of WLCG sites after CHEP
  - WLCG can help a lot

Last month production & analysis



W]
Ok

Job resource allocation

- JEDI probes the job consumption
  - Scout jobs report resource usage and the rest of the task runs with reliable limits
  - Walltime, cputime – RAW for now – normalized to HSPEC06 in the future
  - Memory – VMEM of Athena process for now – RSS, PSS, VMEM of the full job and it's steps in the near future using SMAPS
- Data transfer cost – used by JEDI to assign a job to “the least expensive” site
- VMEM → RSS
  - VMEM allocation makes little sense - RSS (PSS) is the quantity telling how much physical memory is actually used
  - RSS + SWAP << VMEM (not always, but true for ATLAS jobs) – jobs can allocate plenty of VMEM even if there is little SWAP on the node
- Batch systems and cgroups:
  - Modern batch systems (HTCondor, SLURM, …) support RSS allocation through cgroups
    - In addition, cgroups limit CPU access – job cannot use unallocated cores
  - Old batch systems can only kill on per-process user-space estimate of RSS or VMEM
    - Not reliable for multicore jobs
- ATLAS will introduce a site RSS/VMEM flag in AGIS to differentiate between the new and old batch limits:
  - Automatically handled by APF and aCT submission systems
  - Jobs will monitor their RSS usage and will kill the payload when the requested RSS is exceeded to ensure the stability of WNs on sites without RSS memory limits

Grid and opportunistic resources

- During Run-1, ATLAS has used ~50% more resources than pledged.
- Opportunistic resource usage will be even more enhanced during Run-2
- Grid resources on WLCG sites – many sites have more CPUs than pledged
  - Can be used for production when not used by local ATLAS physics groups or by other users
  - Some sites provide access to general purpose academic clusters
- Clouds – many sites provide (experimental) academic or commercial (through grants) allocations to ATLAS
- HPC – supercomputers – ~10 big machines (>100k cores each), more in the future, are explored by ATLAS
  - MC simulation or event generation – no outbound connectivity required
- Volunteer Computing (ATLAS@Home) – home PCs are not that fast or efficient (30% cpu efficiency on average), but there are many
  - ATLAS volunteer base grew in 1 year to 20k volunteers and 6k concurrent jobs
  - No wide campaign yet
- ALL the opportunistic resources are transparently included in the ATLAS production system

- Diverse job execution platforms require:
  - Flexibility of ATLAS software
  - Flexibility of grid middleware – many common assumptions on site/node setup are not valid any more

Tools for opportunistic resources

- More than 50% of the ATLAS CPU time is used for MC event generation and simulation – this is the workload targeted for most of the opportunistic usage
  - Low I/O, High CPU, No DB connectivity required
- Yoda – AthenaMPI
  - Athena MP spawning several nodes and communicating through MPI for event processing distribution
  - Supercomputers tuned for big parallel jobs, on some of them, single-node jobs make no sense
- HPC pilot
  - Runs on edge service (login node) and manages transfers and PanDA communication
  - Submits batch job to HPC
- ARC-CE
  - Computing Element that can separate input/output transfers and job execution
  - Job can execute on a node without external network connectivity
  - Works also on a remote server connecting to HPC with SSH (SSHFS) transparently
- arcControlTower (aCT)
  - “pilot factory” which grabs the payload from PanDA and sends it to ARC-CE service on HPC
  - also used for Nordugrid WLCG resources since 2008.
- ATLAS is putting a lot of effort to leverage the opportunistic resources and will continue to do so while balancing the effort with the effectiveness of the various resources

Transfer and Access Protocols

- The protocol ZOO still present, ATLAS is trying to reduce the number of used protocols
  - Removing unused data movers from the pilot code
  - Recommending the sites to rely on the protocols below
- gridftp – the only common to all WLCG SEs
  - FTS-3
- SRM – still used
  - download/upload to/from nodes on most of the sites
  - Tape access
  - 3rd party transfers initiation
  - Bulk deletions
- https/WebDAV
  - download/upload to/from nodes used at some sites
  - Not yet production ready for direct access – remote I/O
  - But modern, commercial storage endpoints are primarily http based
  - Maturity of the service is still too low for production
- xrootd
  - The recommended way for direct access – LAN or WAN remote I/O
  - ATLAS software tuned and data optimized (TTreeCache…)
- Other protocols are deprecated by ATLAS , although still supported

WAN data access

- Job overflow model prototype
  - PanDA manages the amount of analysis jobs that can read from remote storage
  - The job input locations are fixed to source sites based on cost-matrix calculations
- Careful planning of amount of overflow
  - 100GB/s of average total input traffic
  - overflow of the order of 10% of all jobs – comparable to the transfer traffic between the sites
  - WAN data access must not affect the ATLAS production and transfers
- Evaluated mostly within US – applying it worldwide will face extra difficulties:
  - Insufficient international bandwidth
  - Commercial links
  - Storage at a site can be under stress due to increased number of connections
- Will be gradually tested and explored during Run-2

Network connectivity

- Networks proved to be very reliable
  - less and less frequent issues
  - although problems on international links can take a long time to solve
- Some singularities are still a problem
  - connectivity between academic and commercial networks affecting some site pairs
- Monitoring is an open issue – many sources of network monitoring
  - FTS3 information (much better than in FTS2)
  - experiment tests (DDM sonar, cost-matrix)
  - perfSONAR
- Tests measure different things and in principle they are all needed
- They need to reach maturity to use them for workload and data scheduling
- Issues with stability of perfSONAR-based network monitoring infrastructure – latest becoming more stable

Databases

- Databases are rock solid – very good collaboration between ATLAS and CERN IT DBAs
- More ATLAS applications on non-Oracle databases, as
  - MySQLOnDemand (e.g. Hammercloud, aCT)
  - Hadoop (e.g Event Index).
- Event index replacing the TAG database as catalog of ATLAS events and metadata
  - Being commissioned and filled now.
- CVMFS and Frontier for condition data will not change for Run-2
  - Deployed on Tier-0 + some Tier-1s
  - Decommissioning of BNL Frontier had no visible impact.

Shifts

- Some changes with respect to Run-1:
  - ADCoS, DAST remain the same
  - Comp@P1 is gone – no more 24/7 monitoring of Tier-0 processing and export
  - Computing Run Coordinator (CRC) is brand new – replacing AMOD
- CRC remains the frontend for computing operations with WLCG and sites
  - CRC requires less expertise on ATLAS computing than former AMOD
  - CRC contacts the ADC experts in case of problems and is not required to fix them
  - (ATLAS) WLCG site experts have enough knowledge and are invited to participate
  - Note that CRC counts as a Class-2 shift
  - More on: https://twiki.cern.ch/twiki/bin/view/AtlasComputing/CRC

Issue Reporting Tools

- Tools:
  - GGUS for issues with sites
  - ELOG for internal ATLAS communication
  - Mailing lists for discussions
- Meetings
  - sites are warmly invited to attend ADC weekly.
  - ADC morning meeting at 9:00 to define the planof the day – people can join if they wish to discuss something. Between
- WLCG "daily" meeting, WLCG ops coord and GDB:
  - a lot of reporting, some overlap and little room for technical discussion.
- We had a very successful jamboree with sites in December for Run-2 preparation – One of those events every year?

Conclusions

- ATLAS Computing Infrastructure is ready for Run-2
- The core services were redesigned and provide an extensible interface for future extensions
- The overall performance has still some hiccups due to commissioning of new workflows:
  - Multicore job scheduling
  - Complex job resource requirements
  - Changes in data placement and extensive tape usage
- Many new features are foreseen which will be gradually implemented during the Run-2 with no disruption to ATLAS production and analysis
  - Overflow jobs
  - Global fair-shares
  - Global tasks