ATLAS on CSCS HPC
Bern efforts and status

Tödi, Monte Rosa, Piz Daint ...

Project Background

- Technical integration studies for running ATLAS production on CSCS HPC systems
- Phase 1: address common HPC problems
  - ATLAS Software Access (CVMFS via parrot or rsync)
  - Efficient usage of whole nodes (AthenaMP - multi-threading)
  - Interface to run ATLAS Panda jobs on HPC (ARC SSH submission)
- Phase 2: GPU code porting
  - CSCS Crays are hybrid CPU/GPU machines
  - Need GPU-enabled ATLAS workload
- Final Goal: Production project on Europe's No. 1 super computer «Piz Daint»

Project Motivation

- Overall, it might be more cost efficient to run a few HPC systems instead of a lot of clusters
- ATLAS workloads (detector simulation) can run for ~1h on single node: ideal for backfill
- Boost Switzerland's contribution to global ATLAS computing!

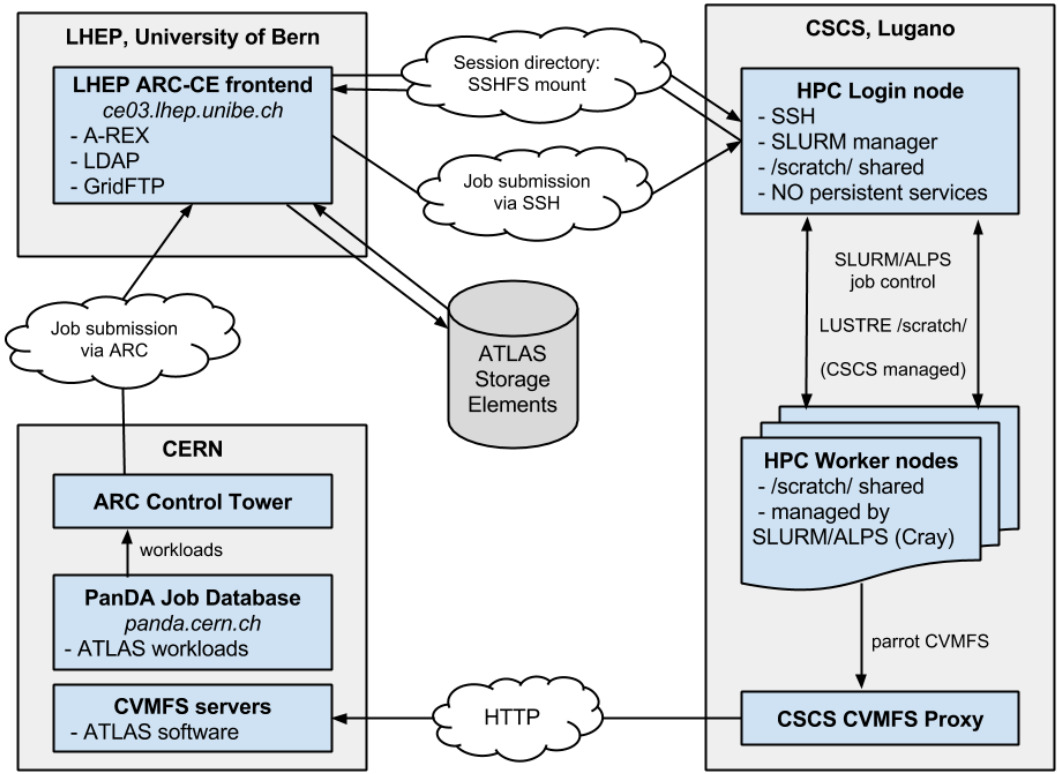
CSCS HPC resources

- Production/Flagship System: «Piz Daint»
  - CRAY XC30, 5272 nodes, Xeon CPUs (16 HT Cores) / Tesla GPUs
  - ~7 PFlops - Europe's Top 1!
  - Long-term target (with GPU code)
- Previous Flagship: «Monte Rosa»
  - CRAY XE6,1496 nodes, Dual Xeon CPUs (32 HT Cores)
  - Short-term target (scaling tests, first production runs)
- Development System: «Tödi»
  - CRAY XE7, 272 nodes, Opteron CPUs (16 Cores) / Tesla GPUs
  - Basically a scaled-down copy of Piz Daint
  - Used for testing, integration and GPU code development

Computing Grid integration?

- ATLAS needs to automate the workflow to run bulk Production Jobs
- HPC facilities (e.g. CSCS) disallow installing and exposing services on login nodes
  - We were explicitly asked not to run any Grid services on the HPC systems.
- Typically SSH access for job submission
- … we can deal with this!

Computing Grid integration!



ARC/Grid interface at LHEP

- Takes jobs and submits them to the target system by SSH
- Does all data handling (e.g. communication with storage elements)

Target HPC System

- Just gets the jobs via SSH + sbatch submit
- Only needs outbound access to CVMFS

Conclusion: What works ...

- Software access (CVMFS) through *parrot*
- Manually run ATLAS jobs
  - Sherpa event generation, single-core
  - Geant4 detector simulation, multi-core (AthenaMP)
- Interface to run jobs from the ATLAS Computing Grid on the machines
  - Running stable for small-scale tests
  - Currently used in Munich for SuperMUC integration

Conclusion: What works ...

Bern UBELIX T3	2584	215+1377
Geneva ATLAS T3	656	0+0
Gordias at hepia	224	0+48 (queue inactive)
LHEP HPC TEST	4352	32+1776
Lugano PHOENIX T2	3266	159+2479

Jobs at ce03.lhep.unibe.ch - Google Chrome

giis.lhep.unibe.ch/jobstat.php?host=ce03.lhep.unibe.ch&port=2135&status=Running&jobdn=all

Jobs at ce03.lhep.unibe.ch

Jobname	Eigner	Status	CPU (min)	Queue	CPUs
1 mc12_8TeV.189658.gg2VVPythia8_AU2CT10_ggH125p5_50SMW_gg_VV_2e2nue_m2l4_2pt3.simul.e2872_s1831_tid01461702._004009.job	mihostet	INLRMS:R	238	day	1
2 mc12_8TeV.189701.gg2VVPythia8_AU2CT10_ggH125p5_VV_2mu2nu_m2l4_2pt3.simul.e2872_s1831_tid01461712._001301.job	mihostet	INLRMS:R	38	day	1

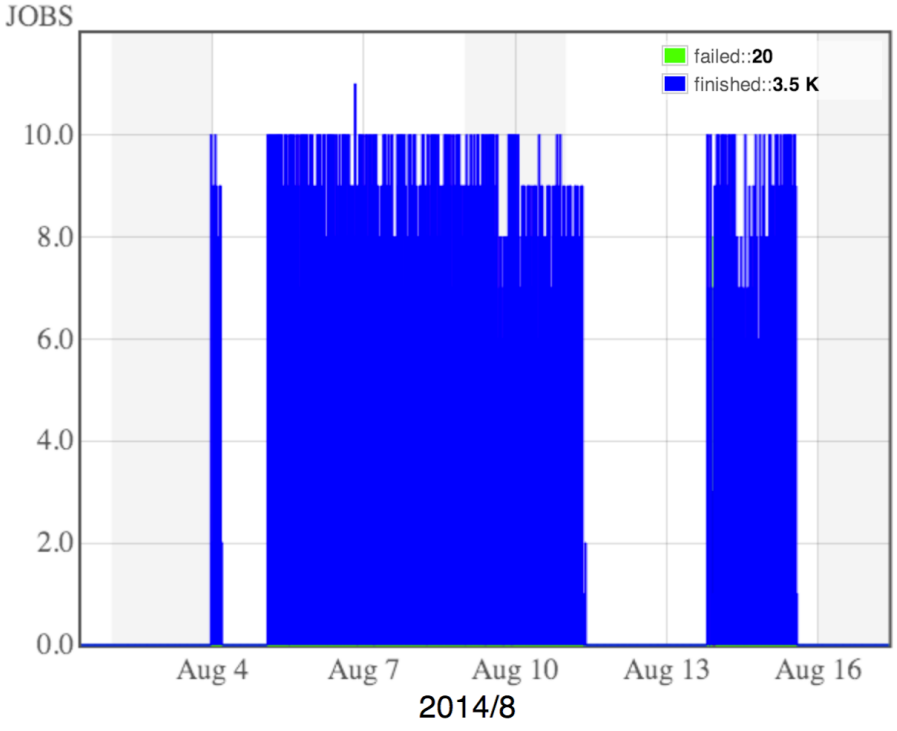
```
[2014-05-03 20:05:52] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: ACCEPTED: parsing job description
[2014-05-03 20:05:52] [Arc.JobDescriptionParserPlugin] [INFO] [10055/3] String successfully parsed as nordugrid:xrsl.
[2014-05-03 20:05:52] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: ACCEPTED: moving to PREPARING
[2014-05-03 20:05:52] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: PREPARING from ACCEPTED
[2014-05-03 20:06:02] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: SUBMIT from PREPARING
[2014-05-03 20:06:02] [Arc.JobDescriptionParserPlugin] [INFO] [10055/3] String successfully parsed as nordugrid:xrsl.
[2014-05-03 20:06:02] [Arc.JobDescriptionParserPlugin] [INFO] [10055/3] String successfully parsed as nordugrid:xrsl.
[2014-05-03 20:06:02] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: state SUBMIT: starting child: /usr/share/arc/submit-SLURM-job
[2014-05-03 20:06:03] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: state SUBMIT: child exited with code 0
[2014-05-03 20:06:03] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: INLRMS from SUBMIT
[2014-05-03 20:06:10] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: ACCEPTED: parsing job description
[2014-05-03 20:06:10] [Arc.JobDescriptionParserPlugin] [INFO] [10055/3] String successfully parsed as nordugrid:xrsl.
[2014-05-03 20:06:10] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: ACCEPTED: moving to PREPARING
[2014-05-03 20:06:10] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: PREPARING from ACCEPTED
[2014-05-03 20:06:32] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: SUBMIT from PREPARING
[2014-05-03 20:06:32] [Arc.JobDescriptionParserPlugin] [INFO] [10055/3] String successfully parsed as nordugrid:xrsl.
[2014-05-03 20:06:32] [Arc.JobDescriptionParserPlugin] [INFO] [10055/3] String successfully parsed as nordugrid:xrsl.
[2014-05-03 20:06:32] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: state SUBMIT: starting child: /usr/share/arc/submit-SLURM-job
[2014-05-03 20:06:34] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: state SUBMIT: child exited with code 0
[2014-05-03 20:06:34] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: INLRMS from SUBMIT
[2014-05-03 21:14:58] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: Job finished
[2014-05-03 21:14:58] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: FINISHING from INLRMS
[2014-05-03 21:15:28] [Arc] [INFO] [10055/3] MAwKDm2Ut0jnYxaMun2n7QwpABFKDmABFKDmknLKDmABFKDmfqWen: State: FINISHED from FINISHING
[2014-05-03 21:27:44] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: Job finished
[2014-05-03 21:27:44] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: FINISHING from INLRMS
[2014-05-03 21:28:14] [Arc] [INFO] [10055/3] WQ2MDmKVt0jnYxaMun2n7QwpABFKDmABFKDm1uLKDmABFKDmfYpCEm: State: FINISHED from FINISHING
```

```
Trf:PerfMonSvc.__write_out_pmon_data 2014-05-03 21:12:37,816 INFO Writing out colle
Py:PerfMonSvc        INFO  --> [ntuple.pmon.dat] =>   12.000 kb
Trf:PerfMonSvc.__write_out_pmon_data 2014-05-03 21:12:37,907 INFO  --> [ntuple.pmon
Py:PerfMonSvc        INFO  --> [ntuple.pmon.stream] =>   70.109 kb
Trf:PerfMonSvc.__write_out_pmon_data 2014-05-03 21:12:37,915 INFO  --> [ntuple.pmon
Py:PerfMonSvc        INFO  --> [ntuple.pmon.pmonsd.txt] =>   26.903 kb
Trf:PerfMonSvc.__write_out_pmon_data 2014-05-03 21:12:37,932 INFO  --> [ntuple.pmon
Py:PerfMonSvc        INFO Writing out collected data... [ntuple.pmon.gz] =>   16.98
Trf:PerfMonSvc.__write_out_pmon_data 2014-05-03 21:12:37,948 INFO Writing out colle
ApplicationMgr       INFO Application Manager Finalized successfully
Py:Athena            INFO replacing PoolFileCatalog.xml by MP version
Trf:Athena.replacePFC 2014-05-03 21:12:37,956 INFO replacing PoolFileCatalog.xml by
ApplicationMgr       INFO Application Manager Terminated successfully
Py:Athena            INFO leaving with code 0: "successful run"
Trf:Athena.exit 2014-05-03 21:12:37,966 INFO leaving with code 0: "successful run"
CoralApplication Info Delete the COOL CoralApplication...
CoralApplication Info Delete the COOL database service
RalDatabaseSvc Info Delete the RalDatabaseSvc...
RalDatabaseSvc Info Purge the connection pool
RalDatabaseSvc Info Reset the ICS pointer
RalDatabaseSvc Info Delete the RalDatabaseSvc... DONE
CoralApplication Info Delete the CORAL connection service
CORAL/Services/ConnectionService Info Deleting the ConnectionPool
CoralApplication Info Delete the COOL CoralApplication... DONE
```

Conclusion: What works ...

2014/8

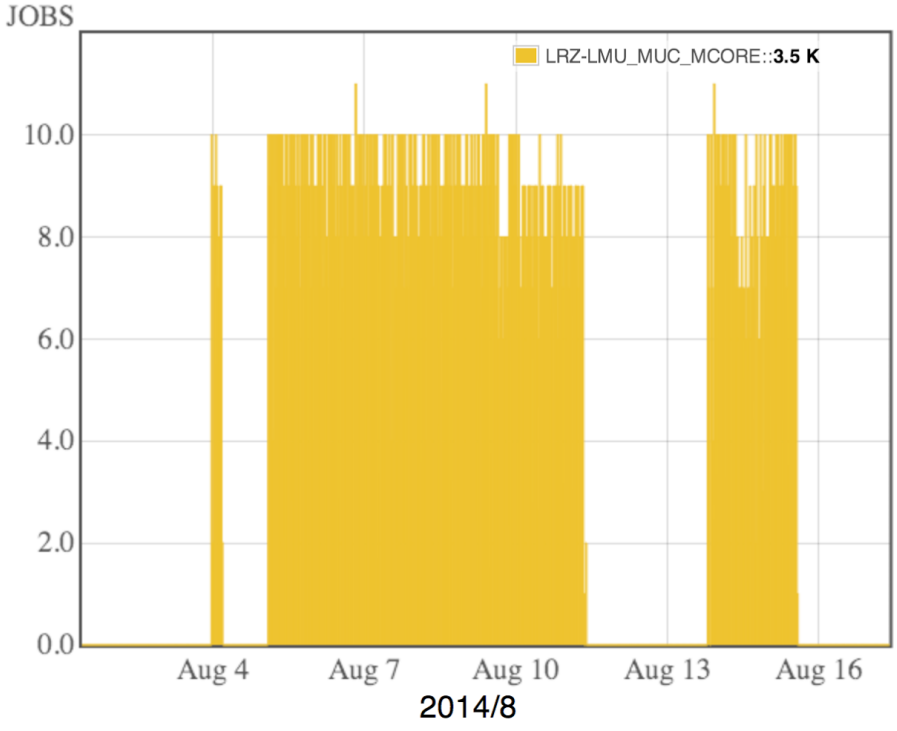
The jobs/hour for all sites progress



2014/8

2014/8

The jobs/hour/site progress



2014/8

Future Plans

- Large-scale tests on Monte Rosa
  - Request submitted to CSCS in June
  - ATLAS Detector Simulation (Geant4), MT CPU code
  - Start with a few parallel jobs, increase up to O(100)
- Optimization and GPU code porting
  - Most promising target: Geant4 = detector simulation
  - ~50% of all ATLAS offline computing time
  - Project at CERN (Geant core team) and elsewhere
  - Even small optimizations / GPU offloads can help!

Thanks for your attention!

Questions?