DA with GANGA and plans

Johannes Elmsheuser[1] Chun Lik Tan[2]

[1]Ludwig-Maximilians-Universität München, Germany

[2]University of Birmingham, UK

12 Dec 2006/ATLAS Distributed Analysis Meeting

Outline

Outline

Distributed Analysis in ATLAS

ATLAS offers several ways to do distributed analysis:

- PAthena/Panda on OSG
- Ganga on LCG
- Data from Production System is distributed by DDM operations team to various sites, like CERN, Lyon, Karlsruhe, BNL and is accessible by DDM/DQ2
- Analysis model implemented by PAT team foresees Athena analysis of AODs/ESDs and interactive use of Athena-aware-ROOT tuples

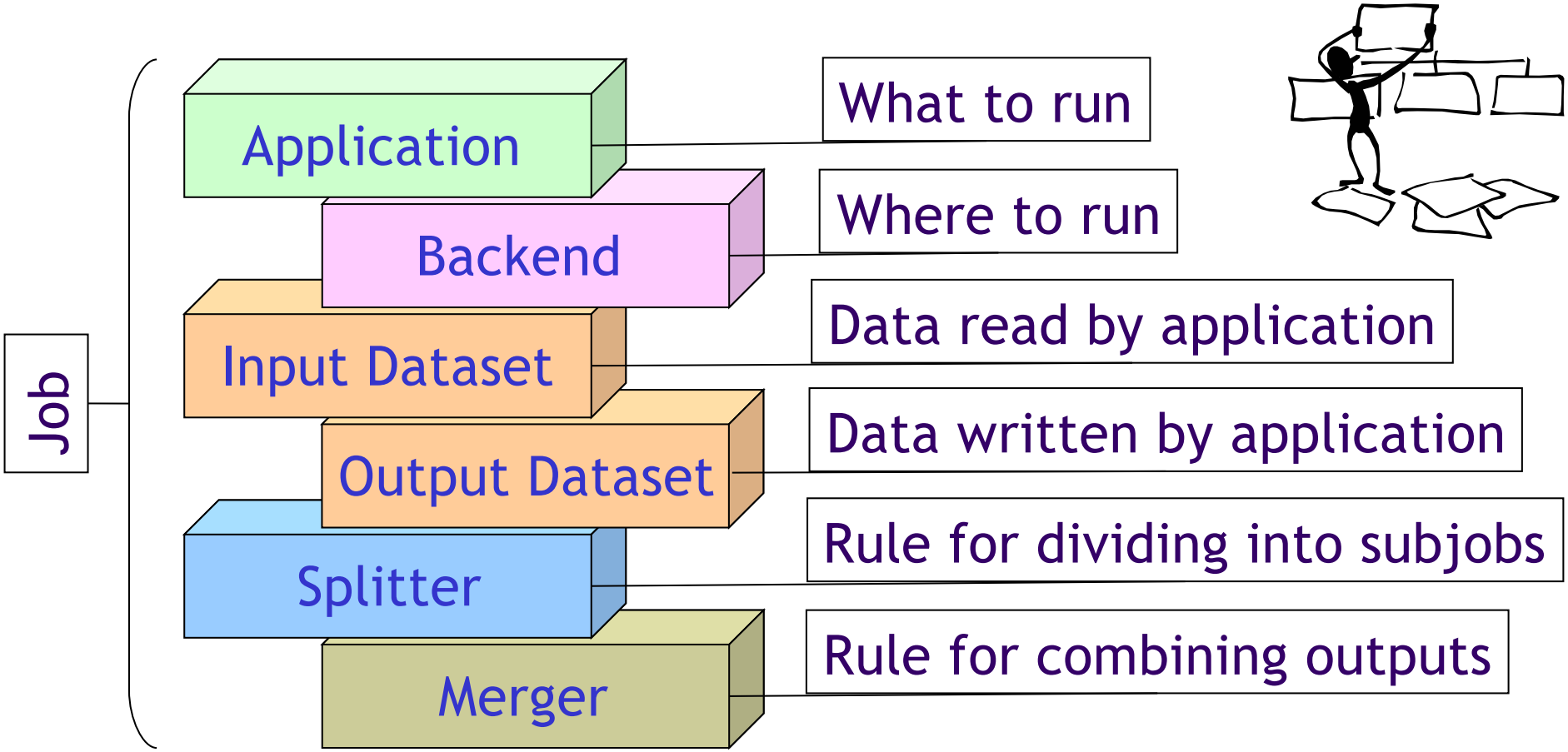
Front-end Client: GANGA

- A user-friendly job definition and management tool.
- Allows simple switching between testing on a local batch system and large-scale data processing on distributed resources (Grid)
- Developed in the context of ATLAS and LHCb :
  - have built-in support for applications based on Athena framework
  - Athena plugin aimed towards User analysis
  - AthenaMC plugin: emulate ATLAS central MC production (prodsys) at small scale, uses JobTransforms
  - Integrated with DQ2 data-management system
- Component architecture readily allows extension
- Python framework

- Development team: F.Brochu (Cambridge), U.Egede (Imperial), J.Elmsheuser (München), K.Harrison (Cambridge), H.C.Lee (ASCC), D.Liko (CERN), A.Maier (CERN), J.T.Moscicki (CERN), A.Muraru (Bucharest), V.Romanovsky (IHEP), A.Soroko (Oxford), C.L.Tan (Birmingham) and contributions past and present from many others

Ganga II

- Ganga is based on a simple, but flexible, job abstraction
- A job is constructed from a set of building blocks, not all required for every job



Ganga resources

- Project homepage:
  http://cern.ch/ganga/
- ATLAS Wiki starting page:
  https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedAnalysisUsingGanga
- Several tutorials, FAQs linked from these pages
- Please use mailing list for user questions:
  project-ganga (at) cern.ch
- Will move to Hypernews soon - discuss ATLAS specific questions.

Outline

Athena plugin in 4.2.x (I)

Highlights: (see various release notes/wiki for more)

- Inputdata:
  - `DQ2Dataset`: Integration with DQ2/DDM 0.2.x
  - Code cleanup in input data staging, query DQ2 DB only once
- Outputdata:
  - Data is stored on closeSE of the grid site the job is running
  - `ATLASOutputDataset`: generic method to store data for local or grid jobs
  - `DQ2OutputDataset`: Register output files as user datasets in DQ2, download with `dq2_get`
- Output Merging:
  - Download remote files in `ATLASOutputDataset` with `job.outputdata.retrieve()`
  - Merge output files with `hadd` or `addAANT` with `job.outputdata.merge()`

Athena plugin news in 4.2.x (II)

- Various:
  - Add dashboard monitoring support by Benjamin Gaidioz: job and job progress monitoring
  - Shared inbox support, offers possibility to effectively use new glite RB bulk submission
  - Introduction of TAG analysis based on ROOT files
  - Rewrite of GangaAtlas/scripts/athena, much more options for shell command line submission
  - Enable Athena 12.0.x
  - Add (non)-compile option to `Athena.prepare()` method
- Documentation:

  `https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedAnalysisUsingGanga`

  `https://twiki.cern.ch/twiki/bin/view/Atlas/GangaTutorial42Athena12031`

AthenaMC plugin news in 4.2.x

Emulate ATLAS central MC production system at small scale - uses prodsys jobTransforms

- prototype data management with built-in dependencies to Curl, restrict range of Grid sites
- supports ATLAS releases up to 12.0.1, change in prodsys transformations prevents event generation in 12.0.2 and beyond, have a fix for that available in CVS
- Can process input data from either legacy DDM and DQ2, output data is registered in DQ2

ATLAS Job definition

- Job definition at shell command line for local desktop:

```
athena AnalysisSkeleton_topOptions.py
```

- Athena Job definition at shell command line for the GRID:

```
ganga athena \
  --inDS csc11.005320.PythiaH170wwll.recon.AOD.v11004107 \
  --outputdata AnalysisSkeleton.aan.root \
  --split 3 \
  --lcg \
  --ce ce-fzk.gridka.de:2119/jobmanager-pbspro-atlasS \
  AnalysisSkeleton_topOptions.py
```

- AthenaMC Job definition at shell command line for the GRID:

```
ganga --config-path=GangaAtlas/Atlas.ini athena --release=11.0.41 \
  --prodname=tutorial --process=single_e_Et40 --random=1102362401 \
  --runnum=000002 --jobopt=DC3.007004.singlepart_e_Et40.py --evgen \
  --transpath=JobTransforms-11-00-41-07.tar.gz --maxevt=30
  --transform=csc.evgen.singleparticle.trf \
  --ce=ce04.pic.es:2119/jobmanager-lcgpbs-atlas
```

- See tutorials for Job definition examples at IPython shell or GUI

Outline

Ganga plans for the upcoming releases (I)

Next major release 4.3 is scheduled for January/February.

Athena plugin

- Migration to `dq2_client` library and more user DQ2 interaction
- Dashboard monitoring in production
- TNT plugin from Mike Kenyon and Caitriana Nicholson, Analysis with TAG DB and job splitting
- Panda backend
- GangaRobot, run automated simple analysis job on a regular basis
- Interest from NorduGrid
- Investigate further Job monitoring possibilities

Ganga plans for the upcoming releases (II)

AthenaMC plugin

- Support Local back-end on with distribution kits
- Removed dependencies to Curl
- Support local input and output files
- Fix problem with evgen transformations, extends support to ATLAS 12.0.2 and beyond
- Future developments: support lxplus and LSF, Castor on lxplus

Plans for the GANGA Core

- Enhanced support for job slicing
- Implementation of GANGA Remote Workspace, i.e. overcome 10MB LCG inbox restriction
- Implementation of timestamps of actions in a GANGA job.
- Introduction of new Internal Services component in GANGA
- Improve the flexibility of GANGA job state machine. Allow certain job states to be manually changed by the user: - provide possibility to resubmit subjobs - avoid resource leak
- Introduce type information for Ganga object schema items to remove certain type ambiguities when assigning values e.g. evt_max = '1' and evt_max = 1
- Introduce lazy reading of repository at startup. GANGA startup may be long if repository has many jobs. This is particularly important for GANGA in scripting mode i.e. when repository loading is not required.
- Improved GANGA config file handling e.g. enable config templating
- Introduce intermediate config levels for developers to alleviate the debug log statement 'screen deluge' problem.

Plans for the GANGA GUI

- Execute GANGA plugin export methods that potentially take some time in threads.
- Add help documentation
- Implement GANGA config file modification from within the GUI
- Move to Qt4 (eventually)
- Introduce filters in the job monitoring panel to allow monitoring of a subset of the jobs.
- AMI dataset browser integration
- Automated GUI dialog integration. Define a simple framework to allow developers to add customised windows to the GANGA GUI. Initial work will be geared towards integrating the GANGA Plotter and the LHCb bookkeeping database browser.
- Investigate using the new job.peek() functionality in the GUI.

Plans for GANGA documentation

- Agreed on standard format for CLI help() documentation
- Investigate the use of Doxygen for the autogeneration of API documentation
- User manual structure finalised. - Detailed generic GANGA user manual - Experiment-specific manual will be self-contained and serve as startup guides.
- More frequent documentation updates

Miscellaneous News

- Ganga Developers day was held in London on 4th-5th December (http://indico.cern.ch/conferenceDisplay.py?confId=7528)
- Results from the user feedback:
  - Likes: eases Grid job submission, job-splitting, DQ2 integration
  - Wants: job resubmission and diagnosis, ease DQ2 interaction
- Distributed analysis session/tutorials at the LCG-IndiaGrid Workshop
- An ATLAS Distributed Analysis tutorial will be held in Edinburgh 1st-2nd February (http://www.nesc.ac.uk/esi/events/737/)

Outline

Conclusions

- Presented the status and plans for Ganga
- Increasing number of users and encouraging feedback
- Improvements and new features for the new version Ganga 4.3