Distributed Data Analysis in ATLAS
Top quark physics

Farida Fassi
IFIC, Valencia, Spain
PCI2010 workshop

Rabat, Morocco, October, 2011
OUTLINE

- ATLAS Computing Model: Data Format and Data Flow
- Analysis and Distributed Analysis System
- Usage of Distributed Analysis and Data Format
- Data Analysis work-Flow
  - What type of jobs you can run?
  - What type of data you can run on?
  - ATLAS release status
- Running on Grid: Practical Advise

- Top quark physics
  - Motivation
  - Search for new physics in top sector
ATLAS Computing Model

- **Tier-0**: Data Recording to tape, First Pass Processing
- **Tier-1**: CERN Analysis Facility
- **Tier-2**: Event Summary Data (ESD): ~1 MB/evt, Analysis Object Data (AOD): ~100 kB/evt, derived data (dESD, dAOD, NTUP,...) distributed over the Grid
- **Tier-3**: 10 Tier-1 centers, RAW data copy on tape, Analysis data on disk, Reprocessing
- **Tier-4**: 37 Tier-2 centers, Analysis data on disk, User Analysis
Data Formats

- **RAW:**
  - Stored in Bytestream format
  - One copy is kept at CERN (tape)
  - One distributed over Tier1’s (disk)
  - Small quantities can be copied to Tier2/group space for special study

- **ESD**
  - Produced from RAW at Tier0 (first pass reco) and Tier1 (reprocessing)
  - One ESD data copy at CERN (tape), two distributed over Tier1’s (disk)
  - Small quantities can be copied to Tier2

- **AOD**
  - Produced from ESD at Tier0 (first pass reco) and Tier1 (reprocessing)
  - At most 2 versions on disk at any given time. 2+1 copies at Tier1’s, 10+10 copies at Tier2’s

- **dESD**
  - Derived from ESD, for detector and performance studies, 10 copies across Tier2’s

- **dAOD:**
  - Derived from AOD, targeted toward physics analysis, definition driven by needs of group analyses, to be stored on Tier2 group space

- **ntuple (D3PD):**
  - Produced by running over d(ESD)/d(AOD), ntuple dumper tools must be validated (official ATLAS status), produced under group/individual control, stored in group space, or locally
Data Formats Flow

Event Source
- Monte Carlo Generation
- Simulation
- Digitization

Data
- Online ByteStream
- Conversion

Publish!
- Results

Analysis Data
- Composite particles and their daughters
- Final selections
- Visualization
- Fitting

Athena Analysis
- Selections
- Combinatorics
- Calculating Observables
- Making Ntuples

TAG
- Num of Jets, muons, etc
- Most E Jet, photon, etc
- “Bits” set by physics analysis groups

ESD
- Ex: Track w/ digits, Jets w/ cells, Trigger Info, etc...

ESD->AOD Converter
- Build “particles”
- Select good particles

AOD
- Collections of Particles (INavigable4Momentum): Photons, Electrons, Jets, Muons, etc...
What is Analysis?

- Re-reconstruction/re-calibration- often necessary.
- **Algorithmic Analysis:**
  - Data Manipulations: ESD→AOD→DPD→DPD
    - Skimming- Keep interesting events
    - Thinning- Keep interesting objects in events
    - Slimming- Keep interesting info in objects
    - Reduction- Build higher-level data which encapsulates results of algorithms

- **Interactive Analysis:**
  - Making plots/performing studies on highly reduced data.

- **Statistical Analysis:**
  - Perform fits, produce toy Monte Carlos, calculate significance.
Stages in Analysis

- Use TAG to quickly select subset of events which are interesting for analysis. (Skim)

- Starting from the AOD
  - **Stage 0**: Re-reconstruction, re-calibration, selection (AOD)
    - Redo some clustering/track fitting, calculate shower shapes, apply corrections, etc...
  
  - **Stage 1**: Selection/Overlap removal/complicated analysis (AOD/DPD)
    - Select electrons/photons → find jets on remaining clusters → btag → calculate MET
    - Perform observable calculation, combinatorics + kinematic fitting, ...

  - **Stage 2**: Interactive analysis (AOD/DPD)
    - Final selections, plots, studies.

  - **Stage 3**: Statistical Analysis
ATLAS Distributed Data Analysis System

**Basic model:**
- data is pre-distributed to sites, jobs are brokered to site which has data

**Grid details mostly hidden for users:**
- automatic job splitting according dataset size

**Output registered in DQ2 and automatically collected to “home-SE”**

**New data distribution mechanism – PD2P:**
- data is distributed based on user demand

**Various ways how to submit and choice of Grid gateway**
### ATLAS Distributed Data Analysis System

**Majority of jobs (>90%) via Panda system**
(pathena or Ganga submission frontends)

**Pull-job model, pilot based job submission**
sites only see pilot-job identity

---

**Production system**

1. **ProdDB**
2. **Job**
3. **Panda Server**
4. **Submitter**
5. **User**
6. **Pilots**

- Each pilot runs on a worker node:
  1. send a request
  2. receives a job
  3. runs the job
Huge increase of the Grid-based analysis: Different users from different flavors

Septiembre 2011
Data Format Usage

DESD’s are the most accessed data format compared to AOD’s

→ Predicting data tiers usage has not worked very well

New development with data distribution model:

**Panda Dynamic Data placement at Tier2’s**→ PD2P

PanDA subscribes dataset to Tier 2, if no other copies are available (except at a Tier 1), **as soon** as any user needs the dataset

Automatic distribution of ESD & DESD stopped, AOD still continuing
MC Data Format Usage

Dataset Popularity

AOD’s are the most accessed as expected

Dataset Access (Analysis, 60 days)

Access
Data Analysis Work-flow

1. Locate the data
2. Setup the analysis code
3. Setup the analysis job
4. Submit to the Grid
5. Retrieve the results
6. Analyze the results
What type of jobs you can run on?

- **Athena jobs:**
  - Using official production transformations:
    - Event generation,
    - Simulation, pileup,
    - Digitization,
    - Reconstruction,
    - merge
  - TAG selection jobs
  - Good Run Lists tools
  - Event picking tools

- **General jobs:**
  - ROOT(CINT, C++, pyRoot)
  - ARA (AthenaRootAccess)
  - Python, user's executable, shell script
  - Merge ROOT files
What Type of Data? And How to locate data?

- User can run on all of the data types,
- however RAW/HITS are on tape, they need to request for DDM replication to disk storage

- DaTRI is an useful interface for end-user needs

- Data location

ATLAS provides several tools for data location and manipulation:

- AMI (ATLAS Metadata Interface)
- dq2 End-User Tools
- ELSSSI (Event Level Selection Service Interface)
How to locate the data? (1)

- **AMI** is a generic cataloging system
- ATLAS datasets are collections of files containing events since it is not possible to put all the data from one run into one file
- Dataset Discovery means finding the names of valid datasets to use in your analysis

How to locate the data?

- **AMI** is a generic cataloging system
- ATLAS datasets are collections of files containing events since it is not possible to put all the data from one run into one file
- Dataset Discovery means finding the names of valid datasets to use in your analysis

**Example:**

```
data10_7TeV%physics_Muons.merge.NTUP_TOP%
```

**AMI** is a generic cataloging system.

ATLAS datasets are collections of files containing events since it is not possible to put all the data from one run into one file.

Dataset Discovery means finding the names of valid datasets to use in your analysis.

**Example:**

```
data10_7TeV%physics_Muons.merge.NTUP_TOP%
```

Type here to search for datasets. Example:
How to locate the data? (2)

A Complete AMI tutorial can be found elsewhere in:

https://indico.cern.ch/getFile.py/access?contribId=21&sessionId=6&resId=1&materialId=slides&conId=115155
The data management system (DQ2) is responsible for the movement of data around the Grid.

User interaction with the system is via dq2 end-user tools: querying, retrieving, creating datasets/dataset containers, etc.

**Using DQ2:**

- DQ2 will be available from User Interface (UI) enabled machines
  - you need a valid grid certificate installed on the UI.

**On lxplus:**

- you will need to source the correct environmental settings:
  - `source /afs/cern.ch/atlas/offline/external/GRID/ddm/DQ2Clients/setup.sh` (or `zsh`)
  - Choose `sh` or `zsh`, depending on your shell type.

- To create a grid proxy type:
  - `voms-proxy-init --voms atlas`
To retrieve your data from the grid to your UI machine you just use:

- `dq2-get <dataset>` or `dq2-get <container/>`

A directory with name `<dataset>` will be created and files will be downloaded into that.

For big datasets only download a few files for testing.

- Then use the grid for the bulk submission.

`dq2-get -n 1 <dataset>` This will retrieve one random file from the dataset

**Examples of listing dataset:**

List available MinBias datasets in DDM

```
dq2-ls 'data10_7TeV*physics*MinBias*'
```

Search for merged AOD’s in container datasets:

```
dq2-ls 'data10_7TeV*physics*MinBias*merge*AOD*'
```

Find location of container datasets (groups of datasets, ending with a '/'):

```
dq2-list-dataset-replicas-container data10_7TeV.00152489.physics_MinBias.merge.AOD.f241_p115/
```

List files in container dataset:

```
dq2-ls -f data10_7TeV.00152489.physics_MinBias.merge.AOD.f241_p115/
```
ELSSI – Event Level Selection Service Interface

https://atltdm10.cern.ch/tagservices/index.htm

ELSSI Website for Relational TAGS
You must have a grid certificate for access. Please choose the server nearest you:

- CERN ELSSI 04-01-02
- BNL (USA, Brookhaven, NY) ELSSI 04-01-03
- TRIUMF (Canada, Vancouver) ELSSI Portal
- RunBrowser Latest version (00-06-00)

Information Twikis:
- Tutorial: Using ELSSI, the Tag Database Portal
- Introduction to tags and tag content

Goal: Retrieve the TAG file from TAG database
- Define a query to select runs, streams, data quality, trigger chains,...
- Review the query
- Execute the query and retrieve the TAG file (a root file) to be used in Athena job

- Production Visitors statistics (awstats)
- EventLookup Visitors statistics on the old lxvm0341
- Development Visitors statistics (awstats)
More on datasets manipulation

https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DQ2Tutorial
https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DQ2ClientsHowTo

The DQ2 Tutorial

↓ What is this data management all about?
↓ Setting it up
↓ Querying
  ↓ I want to list all DDM sites
  ↓ I want to find a dataset
  ↓ I want to list the datasets in a container
  ↓ I want to list the files in a dataset
  ↓ I want to list the replica locations of a dataset
  ↓ I want to list the files in a dataset at another site
  ↓ I want to list the physical filenames in a dataset
  ↓ I want to create a Pool File Catalogue with files on a site
↓ Retrieving data
  ↓ I want to download a full dataset
  ↓ I want to download selected files from a dataset
  ↓ I want to download a dataset from a specific site
↓ Creating data
  ↓ What name should I give to my dataset?
  ↓ Where will my dataset/files be stored?
  ↓ I want to create a dataset from files on my local disk
  ↓ I want to create a dataset from files already in other datasets
  ↓ I want to create a dataset from files on CASTOR at CERN
↓ Putting it all together
↓ I still need help!
↓ Scripting

DQ2 Clients How To

↓ I need help or support! What do I do?
↓ General concepts
  ↓ Datasets and files
  ↓ Versions
  ↓ Dataset Containers
  ↓ Physics Containers
  ↓ Replicas
  ↓ Sites, mass storage systems and SRM
  ↓ When to use dq2-get or DDM subscriptions
    ↓ Few informations about DDM subscriptions
  ↓ Installing/Initializing dq2 commands
    ↓ Installing dq2 commands
    ↓ Initializing dq2 commands
      ↓ Initializing dq2 commands from CERN AFS repository
↓ Querying
  ↓ list all DDM sites
  ↓ find a dataset
  ↓ list the contents of a dataset
  ↓ list the replica locations of a dataset
  ↓ list the datasets at a site
  ↓ list the files in a dataset
  ↓ list the files in a dataset existing at a site
  ↓ list the physical filenames in a dataset
  ↓ create a Pool File Catalogue with files on a site
ATLAS Releases status: production and installation?

http://atlas-computing.web.cern.ch/atlas-computing/projects/releases/status/

Latest Production Releases

<table>
<thead>
<tr>
<th>Release</th>
<th>Patch Release</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.3.1</td>
<td>AtlasProduction 17.3.1.1</td>
<td>Upgrade studies on the grid</td>
</tr>
<tr>
<td>17.0.4</td>
<td>AtlasProduction 17.0.4.2</td>
<td>CentOS --- new migration and ROOT bug fix</td>
</tr>
<tr>
<td>16.6.7</td>
<td>AtlasProduction 16.6.7.17</td>
<td>Updates for generation, simulation and analysis-related packages</td>
</tr>
<tr>
<td>16.6.6</td>
<td>AtlasProduction 16.6.6.5</td>
<td>TAG reprocessing</td>
</tr>
<tr>
<td>16.6.5</td>
<td>AtlasProduction 16.6.5.5</td>
<td>Fast physics monitoring</td>
</tr>
<tr>
<td>16.6.4</td>
<td>AtlasProduction 16.6.4.3</td>
<td>Physics validation processing MC10b</td>
</tr>
<tr>
<td>16.6.3</td>
<td>AtlasProduction 16.6.3.7</td>
<td>Fast reprocessing of 2011 data MC10a</td>
</tr>
<tr>
<td>16.6.2</td>
<td>AtlasProduction 16.6.2.6</td>
<td>Physics validation processing</td>
</tr>
<tr>
<td>16.4.2</td>
<td>AtlasProduction 16.4.2.2</td>
<td>Upgrade studies on the grid</td>
</tr>
<tr>
<td>16.4.2</td>
<td>AtlasProduction 16.4.2.2</td>
<td>Physics validation processing</td>
</tr>
</tbody>
</table>

Release availability at sites, cache None

Caches are AtlasProduction unless otherwise indicated

Releases found (and number of sites): rel_6(2) rel_5(2) rel_4(2) rel_3(2) rel_2(2) rel_1(2) rel_0(4) Conditions(193) 5.28.00(4) 5.26.00c(7) 2.5.4(76) 17.3.1(255) 17.0.4(237) 17.0.3(258) 17.0.2(255) 17.0.1(262) 17.0.0(267) 16.7.0(106) 16.6.7(270) 16.6.6(269) 16.6.5(269) 16.6.4(269) 16.6.3(269) 16.6.2(269) 16.6.1(269) 16.6.0(124) 16.5.0(115) 16.4.2(87) 16.3.0(98) 16.2.2(271) 16.2.1(102) 16.2.0(93) 16.1.3(67) 16.1.2(67) 16.1.0(59) 16.0.3(69) 16.0.2(272) 16.0.1(87) 16.0.0(89) 16.6.6(105) 15.8.0(86) 15.6.9(274) 15.6.8(111) 15.6.7(85) 15.6.6(84) 15.6.5(86) 15.6.4(82) 15.6.3(70) 15.6.2(70) 15.6.1(274) 15.5.1(77) 15.5.0(76) 15.4.1(76) 15.4.0(80) 15.3.0(88) 15.2.0(62) 15.1.0(67) 15.0.0(82) 14.5.2(86) 14.5.1(64) 14.5.0(70) 14.4.0(85) 14.2.25(93) 14.2.24(65) 14.2.23(67) 14.2.22(65) 14.2.21(71) 14.2.20(69) 14.2.1(58) 14.2.13(65) 14.0.1(65) 14.0.10(49) 14.0.0(46) 13.0.0(69) 13.0.30(71) 13.0.20(58) 13.0.10(58) 12.0.95(14) 12.0.8(51) 12.0.7(52) 12.0.6(56) 12.0.5(49) 12.0.4(46) 12.0.3(19) 12.0.3(48) 12.0.2(36) 12.0.1(36) 12.0.0(36) 11.5.0(32) 11.0.4(2) 1.10.4(6) 0.1.35(271) 0.1.34(226) 0.1.33(122) 0.1.32(123) 0.1.31(98) 0.1.29(1) 0.1.28(3) 0.1.26(2)

Cloud | Site | Release | Conditions | Caches |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>ANALAY ALBERTA-WG1</td>
<td>17.3.1</td>
<td>17.3.1.1 17.3.1.2 None</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>ANALAY ALBERTA-WG1</td>
<td>17.3.0</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>ANALAY ALBERTA-WG1</td>
<td>17.0.3</td>
<td>AtlasPhysics-17.0.3.1.1 AtlasPhysics-17.0.3.3.1 AtlasPhysics-17.0.3.5.1 17.0.3.1 17.0.3.2 17.0.3.3 17.0.3.5 MuonCalibration-17.0.3.1.1 MuonCalibration-17.0.3.3.1 None</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>ANALAY ALBERTA-WG1</td>
<td>17.0.2</td>
<td>AtlasPhysics-17.0.2.10.1 AtlasPhysics-17.0.2.3.1 AtlasPhysics-17.0.2.9.1 17.0.2.1 17.0.2.10 17.0.2.2 17.0.2.3 None</td>
<td></td>
</tr>
</tbody>
</table>
Setup the Analysis Job-Panda (1)

- Using PanDA (Production and Distributed Analysis System for ATLAS)

**Client Tools:**

**Panda Analysis Dashboard**

The current pathena version is 0.3.62. The current ganga version is 5.6.10. It is recommended you use the current versions.

**Documentation on user analysis with Panda:**
- Distributed analysis on Panda - overview page
- Client tools for Panda analysis jobs
- pathena: how to submit athena analysis jobs
- prun: how to submit ROOT and general jobs
- pbook: bookkeeping for Panda analysis jobs
- psequencer: how to perform sequential jobs/operations

**Frequently asked questions:**
- Full FAQ
- How is job priority calculated?
- Online/Offline status of sites

- PanDA client consists of tools to submit or manage analysis jobs on PanDA
  - **pathena**
    - How to submit Athena jobs
  - **prun**
    - How to submit general jobs
    - (ROOT, python, sh, exe, ...)
  - **pbook**
    - Bookkeeping (retry, kill) of analysis jobs
  - **psequencer**
    - How to perform sequential jobs/operations
    - (e.g. submit job + download output)
Setup the Analysis Job-Panda (2)

- Submit the Athena Job to the Grid with pathena
- A consistent user interface to Athena

How to submit Athena jobs to Panda

- When you run Athena with:
  
  ```
  $ athena jobOptions.py
  → all you need to do is run pathena: 
  $ pathena jobOptions.py
  --inDS inputDatasetName
  --outDS outputDatasetName
  ```

Very useful examples
Setup the Analysis Job-Ganga

Tutorials how to work with Ganga

The tutorial pages describing how to work with GANGA with many examples:

Quick Start:
• A Quick Start Guide how to execute different ATLAS specific workflows with Ganga

Full Ganga Tutorial:
• A full tutorial covering all the major aspects of using Ganga with Athena on the grid. Kept up to date with both the latest Ganga and Athena releases

Non-CERN Tutorials:
• Ganga Tutorial with special setup at the DESY-HH NAF (Sep 2010)

AthenaMC specific:
• Detailed notes about AthenaMC

Old or Special Versions:
• Tutorial Ganga 4.4.x on NorduGrid (October 2007)

Frequently Asked Question for Distributed Analysis using GANGA

+ Getting Your Outputs
  • Where are my output datasets?

+ Common Job Failures
  • "My job failed - how do I find out why?"
  • "Most of my jobs completed but one failed - how do I resubmit?"
  • "My job seems to be stuck in completing state. What can I do?"
  • "My job ran but failed with a 'file not found' style error on a long dataset filename with rfto: at the beginning"
  • "My job just stays in submitted state and then fails"
  • "My job seems to have failed because it couldn't import the Pf module"
  • "There is an error when staging out the data: problems in output stage-out. Could not read output file..."
  • "There is an error 'NameError: name 'include' is not defined' in my stdout"

+ Datasets and DDM Policy
  • "How do I find out where a dataset is available?"
  • "Where has my data gone and where can I send it?"
  • "How can I write my small output/log file directly to my desktop?"
  • "How do I write data to a LOCALGROUPDISK?"
  • "How do I request a dataset replication to another site?"
  • "Failed" jobs producing output files: how to avoid duplicates?

+ Advanced Workflows
  • "$Gangashell" is trying to run a RedExCommon style analysis and it just fails. What am I doing wrong?"
  • "$Gangashell" do I use production releases?"
  • I'm getting Database errors when running Athena on the grid"
Output Datasets and DATRI

Files that are stored on the grid will be placed within Containers and Datasets.

- The output of your Grid jobs will end up in a dataset.
- Datasets and filenames must be uniquely named

User data must follow: `user.<nickname>.xxx`

<table>
<thead>
<tr>
<th>User:jobID</th>
<th>Created</th>
<th>Latest</th>
<th>Jobs</th>
<th>Pre-run</th>
<th>Running</th>
<th>Holding</th>
<th>Finished</th>
<th>Failed</th>
<th>Cancel</th>
</tr>
</thead>
<tbody>
<tr>
<td>In:mc10_7TeV.105009.J0_pythia_jetjet.merge.AOD.e574_s934_s946_r2299_r2300_tid358712_00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Out:user.farida.mc10_7TeV.105009.J0_pythia_jetjet.merge.AOD.e574_s934_s946_r2299_r2300_v1/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DaTri interface

Allows large data transfers to be scheduled to avoid disruption.

- For user datasets, much faster to download from local T2 SCRATCHSPACE than job site.
Running Jobs on the grid: Practical Advice

1) Always test your job locally before submitting to the Grid
2) Use the latest version of pathena/Ganga
3) Submit your jobs on container datasets (merged)
4) If your dataset is on tape, you need to request for data replication to a disk storage first
5) Do not specify any site name in your job submission, pathena/Ganga will choose the best site available
6) Check release and installation status

Submitting

Where and how you will store your output datasets?
- Request a data replication: your output files will stay as a dataset on Grid.
  ➔ You need to freeze your dataset before requesting for replication
- Download onto your local disk using dq2-get

Note that: by default, user datasets are created on “SCRATCHDISK” at the site where the jobs run.
This disk space is not a permanent place, users have ~ 15 days to retrieve the data
Replication is needed if users want to see their data on the grid

Retrieving your outputs from the Grid
Useful links

**Tutorials**

**DQ2:**
- [https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2ClientsHowTo](https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2ClientsHowTo)
- [https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2Tutorial](https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2Tutorial)

**DA on PanDA:**
- [https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DAonPanda](https://twiki.cern.ch/twiki/bin/viewauth/Atlas/DAonPanda)
- [https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2ClientsHowTo#AfterCreatingDataset](https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2ClientsHowTo#AfterCreatingDataset)

**Get help**
- Subscribe to [atlas-dist-analysis-help forum](https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasDAST):
Top quark physics
Motivation (1)

- **Top quark:**
  - Discovered 1995 at Fermilab
  - By far the heaviest quark
    - it decays before it can hadronize
    - investigation of a bare quark
  - Large coupling to the Higgs boson
    - top plays an important role in the SM

- **A Window to new physics**
  - Many models couple preferentially to top
  - New particles may decay to top
  - Top quark plays an important role in many BSM models

- **Main background in many new physics scenarios (e.g. SUSY)**

- **Very useful to calibrate detector**
  - Jet energy scale, b-tagging efficiency
Motivation (2)

Since the Top discovery, based on a few events, top quark physics has evolved into precision physics with a large number of measurements:

- **Production**: $t\bar{t}$ cross section (CS), differential CS, $t\bar{t}$ asymmetries, single top, etc.
- **Decay**: $W$ helicity, spin correlations, etc.
- **Properties**: Mass, widths, lifetime, charge, etc.

Much more to come at LHC, which will produce loads of $t\bar{t}$ pairs.

Top quark production rate

<table>
<thead>
<tr>
<th></th>
<th>Tevatron</th>
<th>LHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top pairs</td>
<td>1-2 per h</td>
<td>30-40 per h</td>
</tr>
<tr>
<td>Single Top</td>
<td>1/2-1 per h</td>
<td>10-20 per h</td>
</tr>
</tbody>
</table>

2011 data taking period

LHC is a top factory
Top pairs Decay Channels

- Top quark pair production via strong interaction (dominant process)
- Electroweak production of single top quarks (factor 2-3 smaller than pair prod.)

- Top decay $\sim 100\% \rightarrow Wb$
- decay channels: (3 leptons, 2 quark families, 3 colour, 2 branches)
  - dileptonic: $3 \times 3 = 9$ $\rightarrow 11\%$
  - semileptonic: $3 \times 2 \times 3 \times 2 = 36$ $\rightarrow 44\%$
  - all-hadronic: $2 \times 2 \times 3 \times 3 = 36$ $\rightarrow 44\%$

  $\Sigma = 81$

**Experimental Signatures**

Analysis strategy depends on W decay modes:

- Dilepton channel (without tau)
  - 5%: 2 charged leptons, 2 jets, 2 $\nu$
- lepton+jets channel (without tau)
  - 30%: 1 charged leptons, 4 jets, 1 $\nu$
- all-jets channel
  - 44%: 6 jets (At least 2 jets are b jets)
Looking for narrow resonances  Model independent in top sector

Reconstruction of the $tt$ system with help of kinematic fit

- Considered modes: $e+\text{jets}$, $\mu+\text{jets}$

Reconstruction of $m_{tt}$ done in 3 steps

Reconstruction of leptonic $W$ (with $E_T$ as $p_T(\mu)$)

- 2 real solutions → keep both
- imaginary solutions → modify MET

jet-parton association by $\chi^2$ minimisation

- The method accounts for the $W$ and the top reconstructed
  masses, $P_T(t\bar{t})$ and $H_T$ fraction
- kinematic fit to improve resolution

Results
Good agreement in $m_{tt}$ with SM
No significant signal observed

Very promising Analysis with
$> 3.8 \text{ fb}^{-1}$ of data so far
Searches for new physics at Charge asymmetry $tt$ production (1)

- Charge asymmetry is sensitive to the left-right couplings in top pair production.
- Different left and right handed couplings of a beyond-SM resonance lead to different rapidity distributions of top and anti-top.
- $Z'$ or axigluons are candidates for this effect.

At Tevatron a forward-backward asymmetry is measured using $tt$ rapidity difference $\Delta y$.

- Large asymmetry value found with a deviation from SM of $\sim 2\sigma$.
- For events with $M_{tt}>450$ GeV the deviation from SM is $\sim 4\sigma$ [hep-ex/1101.0034]
Searches for new physics at Charge asymmetry $tt$ production (2)

- At LHC:
  - $gg$ fusion symmetric
  - Asymmetry only from small $qq$ fraction
    - Due to smaller $qq$ fraction in initial state predicted asymmetries at LHC are much smaller
- At LHC an asymmetry can be visible in $|\eta_t| - |\eta_{\text{antitop}}|$.

**Measurement of charge asymmetry consist of following parts:**
- Selection of top pair events
- Reconstruction of top quark pair four momenta to estimate $|\eta_t| - |\eta_{\text{antitop}}|$.

- $A_C = (N^+ - N^-)/N^+ + N^-$
- $N^+/N^-$ are the number of events with positive/negative values of $|\eta_t| - |\eta_{\text{antitop}}|$.
  - Predicted in SM $A_C = 0.0130(11)$
  - Would indicate BSM if there is deviation
    - E.g. Axigluon/Z$'$
  - Measured at CMS:
    - $A_C = 0.060 \pm 0.134(\text{stat}) \pm 0.026(\text{sys})$, CMS-PAS-TOP-10-010
SUMMARY

I

✓ Review of the Data Analysis machinery in ATLAS is presented.
✓ Main focus is given to the Distributed Data Analysis WorkFlow

II

✓ LHC era is the golden period for Top Physics
✓ Searches for new Physics is very well motivated
✓ Technical knowledge is available
✓ Learning experience allows Searches adventure feasible
Backup
ATLAS Computing changes

Dynamic data placement:
- Pre-defined, static data distribution according to fixed shares of certain data types found to be inefficient
  - lots of data is centrally distributed to many sites but hardly used
  - lots of data gets used but not covered by central distribution
- DatRI: User/group dataset replication triggered by users
- System in DQ2, large volumes must be confirmed by local managers
- users copy only data they really need (in general …)
- PD2P: Panda dynamic data placement replication based on analysis request
  - triggers replication of popular datasets to different site
  - most T2 data to be replicated this way