Evolution of the Atlas data and computing model for a Tier-2 in the EGI infrastructure

Álvaro Fernández Casaní

on behalf of the IFIC Atlas computing group: AMOROS, Gabriel - VILLAPLANA, Miguel - FASSI, Farida - GONZALEZ, Santiago - KACI, Mohammed - LAMAS, Alejandro - OLIVER, Elena - SALT, Jose - SANCHEZ, Javier, SANCHEZ, Victoria
Outline

• Introduction to Atlas and IFIC Spanish Tier2
• Evolution of the Atlas data and computing model
  – Flattening the model to a Mesh
  – Availability and Connectivity
  – Job Distribution
  – Dataset replication
  – IFIC infrastructure and operations
• Example of Grid and Physics Analysis
ATLAS CENTERS

http://dashb-earth.cern.ch/dashboard/dashb-earth-atlas.kmz

Álvaro Fernández Casaní

ISGC2012, Taipei, 2th March 2012
ATLAS Computing Model: T0 Data Flow

From This morning Ueda I. talk
Computing (3900 cores & 4 PB)

- ES Tier1 at PIC Barcelona for Atlas:
  1,600 cores, 1.7 PB disk and 1.6 PB tape
- Federated ES Atlas-Tier2 (IFIC/UAM/IFAE):
  1,960 cores and 2 PB disk
- Local Tier3 farms at IFIC, UAM and IFAE:
  300 cores and 250 TB disk
  → fundamental for physics analysis

Daily average: 2780 Jobs per day

Total of 37432 millions of events processed
Spanish ATLAS Tier2

- The Spanish ATLAS Tier-2 is integrated in the WLCG Project (Worldwide LHC Computing GRID) and follows the ATLAS Computing Model

IFIC : is the coordination Center of the Spanish ATLAS Tier-2

Deployed Equipment:

<table>
<thead>
<tr>
<th>Site Name</th>
<th>HEP-SPEC06</th>
<th>Disk (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IFIC</td>
<td>6950</td>
<td>940</td>
</tr>
<tr>
<td>UAM</td>
<td>3314</td>
<td>302</td>
</tr>
<tr>
<td>IFAE</td>
<td>4993</td>
<td>470</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>15257</strong></td>
<td><strong>1712</strong></td>
</tr>
</tbody>
</table>

Human Resources:

- **14 FTE**

IFIC has reached 1 PB disk Storage!

¡The Spanish ATLAS Tier-2 has Got 2 PB Disk Storage!
IFIC Computing Infrastructure

Resources

as of February 2012

**ATLAS Tier2:**
- 48 x Xeon Quadcore E5472 3.00 GHz 16 GB RAM
- 32 x Xeon Quadcore L5520 2.77 GHz 24 GB RAM
  => 640 cores
- 20 x Sun X4500 y X4540
- 7 x Supermicro
  => 1 PB

**GRID-CSIC:**
- 48 x Xeon Quadcore E5472 3.00 GHz 16 GB RAM (Infiniband)
- 106 x Xeon Quadcore E5472 3.00 GHz 16 GB RAM
  => 1232 cores
- 5 x Sun X4540
  => 180 TB

CSIC resources (not experiment resources):
- 25% IFIC users
- 25% CSIC
- 25% European Grid
- 25% Iberian Grid

To migrate Scientific application to the GRID

Álvaro Fernández Casaní

IFIC Computing Infrastructure

ISGC2012, Taipei, 2th March 2012
Summary of year 2011

• 3,579,606 Jobs
• 6,038,754 CPU consumption hours
• 13,776,655 KSi2K (CPU time normalised)
• Supporting 22 Virtual Organizations (VOs)
Prevision 2012
For IFIC

<table>
<thead>
<tr>
<th>YEAR</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU(HS06)</td>
<td>6000</td>
<td>6950</td>
<td>6650</td>
<td>7223</td>
</tr>
<tr>
<td>DISK(TB)</td>
<td>500</td>
<td>940</td>
<td>1175</td>
<td>1325</td>
</tr>
</tbody>
</table>

- For 2012:
  - We already fulfil the CPU requirements
  - Increasing Disk (planned for April):
    - 230 TB => 4 SuperMicro x 57.6 (2 TB disks)
From Tiered to a Mesh model

Data Processing Model Revised

Auto-routing
- mesh transfer tests from every site to every site ('sonar' tests)
- the transfer system decides the site-to-site routing based on the measurements

From This morning Ueda I. talk
Connectivity

- **Network** is a key component in the evolution of the Atlas model
- **Transfers among sites measured by Atlas through FTS log**:
  - Matrix of NxN components
  - < 1 MB or > 1 GB files to split transfer and initialization components
- **IFIC** is qualified as T2 well connected to T1s and was included as multicloud site:
  - **Intercloud transfers** -> Direct transfers from/to all ATLAS T1s and process data from many T1s
Availability

Hammercloud: for ATLAS
- Sites are tested with 'typical' analysis jobs
- Problematic sites are identified online and automatically blacklisted
- Protect user jobs from problematic sites

EGI has own different availability tools:
- Based on ops VO
- Can have different results:
  Last month **IFIC had > 90% availability** according to HC (ALPHA SITE) while just 64% for EGI (due to a config issue)

http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteviewhistorywithstatistics?columnid=564&view=shifter\n
IFIC LAST MONTH

T1 LFC catalog
Movement to T0

http://dashb-atlas-ssb.cern.ch/dashboard/request.py/siteviewhistorywithstatistics?columnid=564&view=shifter%20\n
IFIC LAST MONTH
Job Distribution at Tier-2

- Tier 2 usage in data processing
  - More job slots used at Tier2s than at Tier1
    - Large part of MC production and analysis done at Tier2s
- More Analysis jobs and “group production’ at Tier-2s
  - Less weight at Tier-1s
- Production shares to limit “group production” jobs at Tier-1, and run at Tier-2s
  - Production shares to limit “group production” jobs at Tier-1, and run at Tier-2s
- Analysis share reduced at Tier-1s
Replication of datasets in Es-Cloud last month

- Es-Cloud Tier 2 sites getting more datasets
- Bit more that Tier-1 Pic

IFIC is alpha T2D site (good connectivity and availability) -> candidate for more datasets replication

IFIC Storage resources

- **SUN thumpers:**
  - X4500 + X4540
  - Lustre v1.8

- **New Supermicro servers:**
  - SAS disks, capacity 2 TB per disk and 10 Gbe connectivity

- Srmv2 (STORM)

- 3 x gridftp servers

---

**SUN X4500:** 5x(500GB) + 1x(1TB) = 112 TB
**SUN X4540:** 13x(1TB) = 442 TB
**SuperMicro:** 7x(2TB) = 403 TB

**TOT:** 954 TB

**Pledge 2011:** 940 TB
### Storage tokens

**FEBRUARY 2012**

<table>
<thead>
<tr>
<th>TOKEN</th>
<th>ASSIGNED</th>
<th>USED</th>
<th>FREE</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALIB</td>
<td>10.0</td>
<td>6.7</td>
<td>3.3</td>
<td>66.0</td>
</tr>
<tr>
<td>DATA</td>
<td>759.8</td>
<td>502.9</td>
<td>257.0</td>
<td>66.0</td>
</tr>
<tr>
<td>PROD</td>
<td>20.7</td>
<td>3.6</td>
<td>17.1</td>
<td>17.0</td>
</tr>
<tr>
<td>GROUP</td>
<td>113.2</td>
<td>35.8</td>
<td>77.3</td>
<td>31.0</td>
</tr>
<tr>
<td>SCRATCH</td>
<td>51.2</td>
<td>36.3</td>
<td>14.9</td>
<td>70.0</td>
</tr>
<tr>
<td>HOT</td>
<td>4.9</td>
<td>3.5</td>
<td>1.4</td>
<td>71.0</td>
</tr>
<tr>
<td>*LOCAL</td>
<td>76.8</td>
<td>60.1</td>
<td>16.7</td>
<td>78.0</td>
</tr>
<tr>
<td><strong>sum</strong></td>
<td><strong>959.8</strong></td>
<td><strong>588.7</strong></td>
<td><strong>371.0</strong></td>
<td><strong>61.0</strong></td>
</tr>
</tbody>
</table>

LOCAL is for T3 local users, out of T2 Pledges. Numbers reflect actual usable space for Filesystem.
IFIC Lustre Filesystems and Pools

- With the Lustre release 1.8.1, we have added pool capabilities to the installation.
- Allows us to partition the HW inside a given Filesystem
  - Better data management
  - Assign determined OSTs to a application/group of users
  - Can separate heterogeneous disks in the future

4 Filesystems with various pools:
/lustre/if ic.uv.es Read Only on WNs and UI. RW on GridFTP + SRM

/lustre/if ic.uv.es/sw. Software: ReadWrite on WNs, UI (ATLAS USES CVMFS)

/lustre/if ic.uv.es/grid/atlas/t3 Space for T3 users: ReadWrite on WNs and UI

xxx.if ic.uv.es@tcp:/homefs on /rhome type lustre. Shared Home for users and mpi applications: ReadWrite on WNs and UI

Different T2 / T3 ATLAS pools, and separated from other Vos
Better Management and Performance
CVMFS at IFIC (+SQUID)

- Installed in all our WNs and UIs since September 2011
  - Easy installation (only 2 configuration files)
  - 20 GB/repository
  - There is not a dedicated partition
- Using the same SQUID as frontier (sq5.ific.uv.es)
  - The squid server is pointing to the public replicas (CERN, BNL, RAL)
- The performance until now is so good.
- We are monitoring SQUID via CACTI (SNMP)

- Reduced job setup time
- Better performance for analysis jobs
- Reduced load on Lustre MDS
- No more need to had a central disk hosting all ATLAS software versions
Storm SRM server

- Access like a local file system, so it can create and control all the data available in disk with a SRM interface.
- Coordinate data transfers, real data streams are transferred with a gridFTP server in another physical machine.
- Enforce authorization policies defined by the site and the VO.
- Developed Authorization plugin to respect local file system with the corresponding user mappings and ACL's
IFIC Network

- Data Network based on gigabyte ethernet.
- 10GBe network

Cisco 4500 – core centre infrastructure.
Cisco 6500 – scientific computing infrastructure

Data servers:
- Sun with 1GB connection. Channel bonding tests were made aggregation 2 channels
- SuperMicro with 10 Gbe WNs and GridFTP servers with 1GB

Recent Upgrade FTP servers to satisfy connectivity requirements for alpha sites
Example of Grid & Physics Analysis

- Distributed Computing and Data Management Tools based on GRID technologies have been used by the IFIC physicists to obtain their results
- As an example, the Boosted Top candidate presented by M. Villaplana
Tier2 and Tier3 examples from Spain

- At IFIC the Tier3 resources are being split into two parts:
  - Resources coupled to IFIC Tier2
    - Grid environment
    - Use by IFIC-ATLAS users
    - Resources are idle, used by the ATLAS community
  - A computer farm to perform interactive analysis (proof)
    - outside the grid framework

- Reference:
  - ATL-SOFT-PROC-2011-018
Tools for Distributed Analysis

• For ATLAS users, **GRID tools** have been developed:
  
  – For **Data management**
    • Don Quijote 2 (DQ2)
      – Data info: name, files, sites, number,…
      – Download and register files on GRID,…
    • Data Transfer Request (DaTri)
      – Extension of Dq2 including quota management
      – Users make request a set of data (datasets) to create replicas in other sites (under restrictions)
    • ATLAS Metadata Interface (AMI)
      – Data info: events number, availability
      – For simulation: generation parameter, …
  
  – For **Grid jobs**
    • PanDa Client
      – Tools from PanDa team for sending jobs in an easy way for user
    • Ganga (Gaudi/Athena and Grid alliance)
      – A job management tool for local, batch system and the grid
Daily user activity in Distributed Analysis

- An example of Distributed Analysis in heavy exotic particles
  - Input files
  - Work flow:

1. Test the analysis locally - input files downloaded by DQ2
2. Submit 1º job to GRID - creating an output file with only the info interesting.
   - input: real/simulation data
   - time job around 20 hours
3. Submit 2º job to GRID (normally at Tier-3)
   - analyzing: particles reconstruction, selections..
   - input: output of 1º job
   - time job around 2 hours
Daily user activity in Distributed Analysis

- 1) A python script is created where requirements are defined
  - Application address,
  - Input, Output
  - A replica request to IFIC
  - Splitting
- 2) Script executed with Ganga/Panda
  - Grid job is sent
- 3) Job finished successfully, output files are copied in the IFIC Tier3
  - Easy access for the user

Just in two weeks, 6 users for this analysis sent:
- 35728 jobs,
- 64 sites,
- 1032 jobs ran in T2-ES (2.89%),
- Input: 815 datasets
- Output: 1270 datasets
Analysis Jobs in IFIC

Peak on February 2012

http://panda.cern.ch/server/pandamon/query?dash=analysis

due to T1 LFC consolidation this week
User experience

• Users are mostly satisfied with infrastructure and tools

• Functionality of dataset replication is transparent to users, but eventually they learn that it is in action:
  – “a job was finally sent to a destination that did not originally had the requested dataset”
  – “I saw later my original dataset has been copied”
  – “I just realized because I used dq2 previously to check where the dataset was”

• Replication is effective for official datasets, not for user generated ones:
  – “We see that is was not replicating our homemade data”
Summary

• Evolution of Atlas data model was presented, and how this affects the IFIC Tier-2 inside the EGI infrastructure.
  – Improvement of the Tier-2 resources usage
  – Changes in operations and infrastructure

• It was presented a real example on how users are working and their experiences:
  – Dynamic data replication useful and transparent