Spanish ATLAS Tier-2 facing up to Run-2 period of LHC

CHEP2015 – Computing in High Energy and Nuclear Physics 2015, 13-17 April 2015, Okinawa, Japan

Funded by MICINN under contracts FPA2010-21919-C03 and FPA2013-47424-C3

S. González de la Hoz[1*], J. Del Peso[2], F. Fassi[1,4], Á. Fernández Casani[1], M. Kaci[1], V. Lacort Pellicer[1], A. Del Rocio Montiel[2], E. Oliver[1], A. Pacheco Pages[3], J. Sánchez[1], V. Sánchez Martínez[1], J. Salt[1], M. Villaplana[1]
on behalf of the ATLAS Collaboration

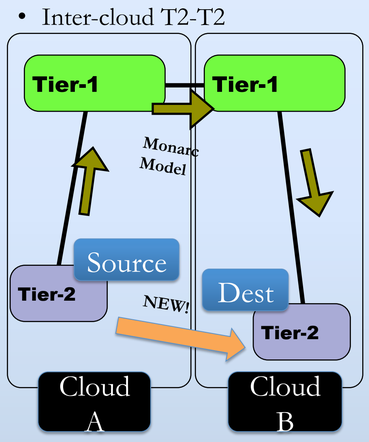
[1] Instituto de Física Corpuscular (IFIC), University of Valencia and CSIC, Valencia, Spain; [2] Departamento de Física Teórica C-15, Universidad Autónoma de Madrid, Spain;
[3] Institut de Física d'Altes Energies, Universitat Autònoma de Barcelona, Spain; [4] Mohammed V University, Rabat, Morocco
*Corresponding author

New Computing Model and ATLAS tools

- Computing system for the ATLAS experiment has performed very well during Run-1.
- Run-2, new scenario in Physics:
  - LHC will run at the design parameters
    - for the center of mass energy (13 TeV)
    - and luminosity ($L = 10^{34}$ cm^{-2} s^{-2} with 25 ns bunch spacing)
    - Event Statistics will increase a factor 5 or more.
    - Pile-up will increase slightly.
- It implies changes in Computing:
  - Improving the functioning of use of software to different platform technologies
  - MultiCore Computing ➔ to save memory
  - Cloud Computing
    - 20% of the ATLAS Computing Power
    - Virtualization
  - Breaking the data locality using remote data access (FAX)
  - Improve data management protocols (Rucio)
    - Dynamic Data Replication and Reduction

Breaking Cloud Boundaries

- **From Hierarchical tier organization** based on Monarc network topology, where sites are grouped into **clouds** for organizational reasons.
- **To a Mesh Model (without Boundaries) because:**
- We can benefit from reliable networks and from better protocols
- Breaking Restricted communications: General public network
  - Inter-cloud T1-T2
  - Inter-cloud T2-T2



Spanish ATLAS Tier-2 (ES-ATLAS-T2)

- Evolution of the Tier-2 Resources (Pledges CPU & Disk) :

T2-ES	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
CPU(HS06)	92	243	1750	5390	10308	13900	13300	18000	20600	26100	30000	36875
DISK(TB)	14	63	387	656	1107	1880	2350	2550	2800	3250	3900	6875

- Present resources provides by the ES-ATLAS-T2:

SITE	CPU (HEP-SPEC06)(pledge)	DISK (TB)(pledge)
IFIC	10478 (10300)	1400 (1400)
IFAE	6600 (5150)	875 (700)
UAM	5150 (5150)	721 (700)

The ES-ATLAS-T2 is:
- formed IFIC-Valencia (50%), IFAE-Barcelona (25%) & UAM- Madrid(25%)
- integrated in the WCLG (World Wide LHC Computing GRID) Project and follows the ATLAS Computing Model
- 5% of the total resources of all ATLAS Tier-2s
- Sites are T2D category (good connectivity among other Tier-1s & Tier-2s)

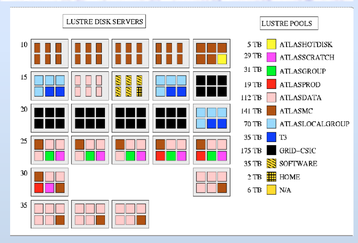
- CPU:

16 x DELL PowerEdge M620
16 cores
2 x Intel Xeon E5-2660 @ 2.20 GHz
RAM: 64 GB
HS06: 238.2

- Disk pool distribution at IFIC based on Lustre since 2009
- Current version 1.8
- 1.5 PB of data

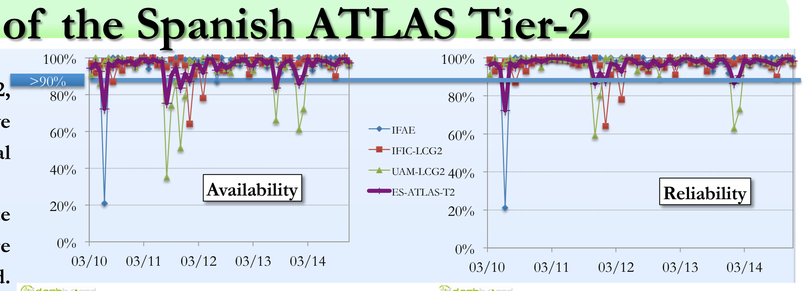
DISK:

SUN X4500: 5x(500GB) + 1x(1TB)	115 TB
SUN X4540: 13x(1TB)	495 TB
SuperMicro: 11x(2TB)	630 TB
SuperMicro: 1x(2TB)	130 TB

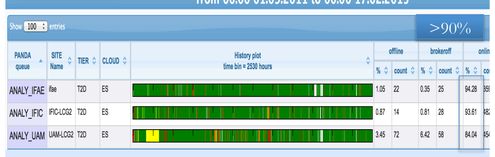


Performance of the Spanish ATLAS Tier-2

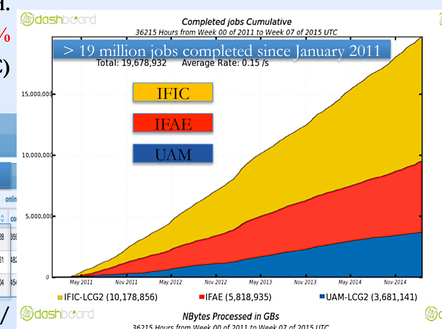
- Availability and reliability of the ES-ATLAS-T2, for the Run-1 and before starting Run-2, have always been, on average, inside the interval 90-100%.
- Sites are tested with typical analysis and Monte Carlo (MC) production jobs. Problematic sites are identified online and automatically blacklisted. Last years ES-ATLAS-T2 had more than 90% availability according to Hammer Cloud (HC) tests.

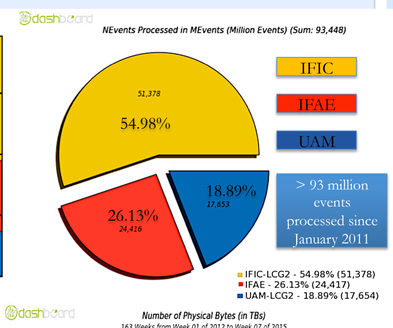


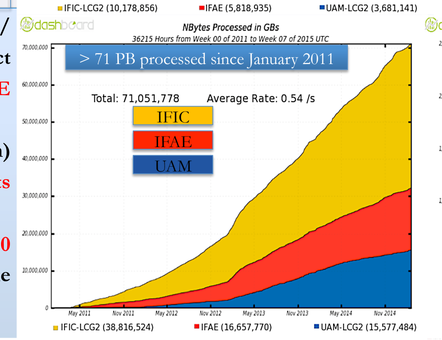
HC Tests

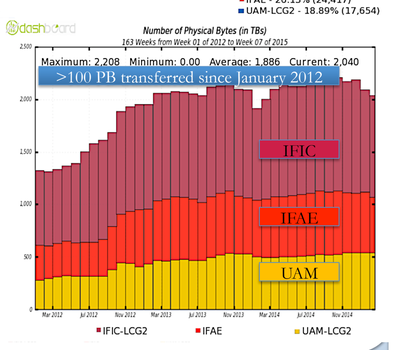


- Good connectivity implies: Direct transfers from/to Tier-0 and all ATLAS Tier-1s and Direct transfers to Tier-2 from different clouds: THREE Tier-2s tagged as T2D!
- More than 19 million jobs (analysis + production) completed!! And more than 93 million events processed!
- More than 71 PB processed and more than 100 PB transferred during the Run-1 and just before the Run-2.







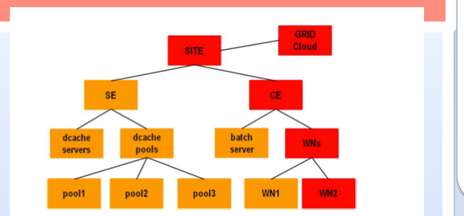


Challenges for Run-2

- Run-2 Data taking will begin in spring-summer 2015 and ATLAS Computing Model has been updated and checked in different Data Challenge tests (DC14).
- The adaptation to these changes related to our Tier-2:
- Daily work to maintain and operate the provided computing infrastructure.
- To Provide an Expert System to keep expertise and to apply automatic decisions.
- Multicore Program creating a Multicore queue in each ES-ATLAS-T2 site.
- Federated Data Storage System (FAX) to improve the use of Computing resources of our Tier-2 joining Fax.
- Improvement of Distributed Analysis.
- Event Index Project: Participation in an important ATLAS R&D activity.

Expert System

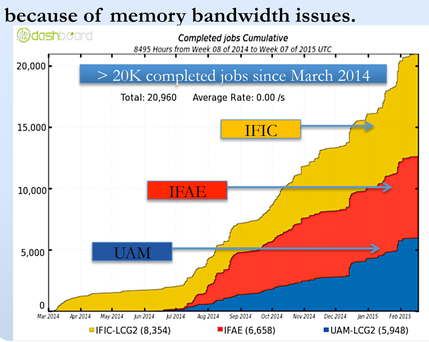
- System will go through log files to obtain information about the problem
- Include a monitoring system of the infrastructure able to present and process the information in a very simple/intuitive way
- If automatic decision can not be made/too risky ➔notify the operators



Example of the tree structure for a finite state machine of the monitoring system of the Expert System

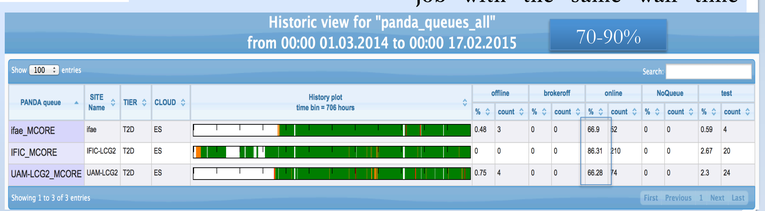
Progress in Multicore

- Athena MP: Multicore implementation of the ATLAS Software framework.
- Allows efficient memory pages sharing between multiple threads of execution whilst retaining event-based parallelism.
- Through Linux Kernel's Copy-On-Write mechanism (memory sharing technique).
- This is now validated for production and produces a significant reduction on the overall application memory footprint with negligible CPU overhead.
- However, with 32 cores or more, event-parallelism may stop scaling because of memory bandwidth issues.



Panda Resources:
IFAE_MCORE
IFIC_MCORE
UAM-LCG2_MCORE

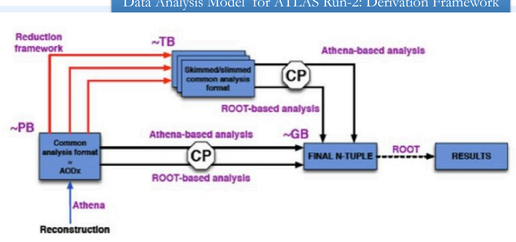
- Multicore@ES-ATLAS-T2 :
1. Each site has created a Multicore queue with dedicated nodes. Since March 2014 efficiencies between 70-90% and more than 20k completed jobs.
2. No mix between Multicore and Monocore is possible in a CPU server nowadays:
   - Simplification of brokering & allocation of cores in nodes running Monocore jobs.
   - Usual approach: assign 8 cores/ job with the same wall time



Distributed Analysis (DA) in Run-2

- DA in ATLAS has worked very well during Run-1 but several improvements are needed to speed-up the workflow and to take profit from the powerful computer architectures.
- Running same program using all the cores available in the same server (MULTICORE parallelism).
- New developments for computing and data access: JEDI, RUCIO.
- DERIVATION FRAMEWORK:
  - Performs bulk data processing to produce targeted samples.
  - To produce automatically common data collection for all the physics groups.
- ES-ATLAS-T2 through the IFAE has contributed to the coordination of DA for two years and now is the Grid Distributed Production (GDP) coordinator. In the future DA will merge with GDP.

Data Analysis Model for ATLAS Run-2: Derivation Framework

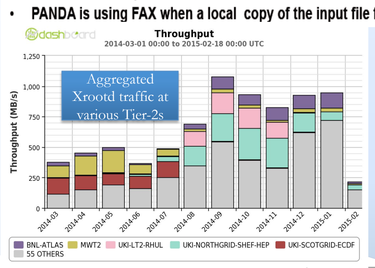


- Computing resource constraints for Run-2 motivates a limit of the nb. of outputs 10.
- Target output size for derivation: 1-10TB.
- Data reduction can be achieved either by skimming or slimming.
- Book-keeping tools will read the sum of weights of events removed.

Towards a Federated Data Storage System

- In Run-1 the processing of data has been done with strong requirement in the CPU-Data locality with non positive effects, like: Multiplication of data replicas in different sites and inability to run ATLAS jobs on disk-less infrastructures.
- Solution: a strategy to use data federation using XRootD (so-called Federated ATLAS XrootD - FAX).
- A way to unify direct access to a diversity of storage services used by ATLAS, bringing together Tier-1, Tier-2 and Tier-3 storage resources into a common namespace accessible from anywhere.

ES-ATLAS-T2:
- IFAE: already serving its disk space within FAX
- IFIC & UAM: will proceed to be activated in the federation
- PANDA is using FAX when a local copy of the input file fails



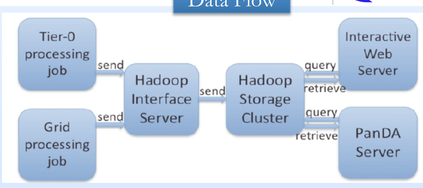
The Event Index project

- A complete catalog of ATLAS events: all events, real and simulated data and all processing stages.
- Contents:
  - event identifiers
  - online trigger pattern and hit counts
  - references (pointers) to the events at each processing stage in all permanent files on storage

1) Information for the Event Index is collected from all production jobs running at CERN and on the Grid and transferred to CERN using a messaging system (ES-ATLAS-T2 is participating actively in this task)

2) This info is reformatted, inserted into the Hadoop storage system and internally catalogued and indexed

Data Flow



3) Query services (CLI and GUID) implemented also as web services query the EventIndex and retrieve the information

Fore further information see Dario and Javier's talks in this conference

Conclusions and Perspectives

- Run-2 Data taking will begin in April 2015 and ATLAS Computing Model has been updated.
- Essential purpose of Tier-2s is to provide computing infrastructure carrying increasing both CPU & Disk
- Aspects related to our Tier-2:
  - To provide an Expert System to keep expertise and to apply automatic decisions
  - Each ES-ATLAS-T2 site has created a Multicore queue (shared with Monocore queue)
  - Federated Data Storage system to improve the use of Computing resources of our Tier-2
  - Improvement of Distributed Analysis is needed and the Derivation framework is emerging as the key to success on Run-2
  - Participation in an important ATLAS R&D activity (Event Index project) in collaboration with other computing groups