Spanish ATLAS Tier-1 &Tier-2 perspective on computing over the next years

CHEP2018 – Computing in High Energy and Nuclear Physics 2018, 9-13 July 2018, Sofia, Bulgaria

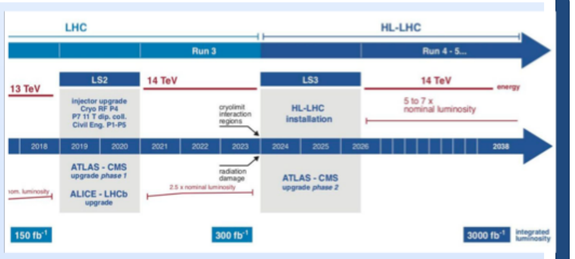
Funded by MICINN under contracts FPA2016-75141-C2-1-2-R and FPA2016-80994-C2-2-R

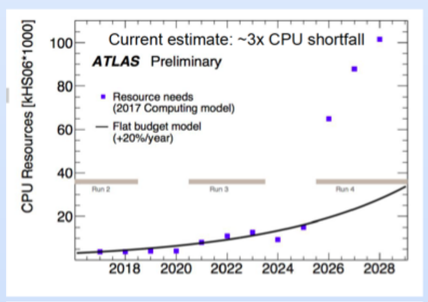
S. González de la Hoz[1]*, C. Acosta-Silva[3,4], J. Aparisi Pozo[1], M. Delfino[3,4], J. del Peso[2], Á. Fernández Casani[1], J. Flix[4,5], E. Fullana Torregrosa[1], C. García Montoro[1], J. Lozano Bahilo[1], A. del Rocio Montiel[2], A. Pacheco Pages[3,4], J. Sánchez Martínez[1], J. Salt[1], A. Vedaee[3,4] on behalf of the ATLAS Collaboration

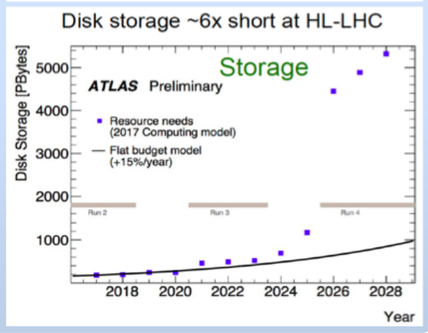
[1] Instituto de Física Corpuscular (IFIC), University of Valencia and CSIC, Valencia, Spain; [2] Departamento de Física Teórica y CIAFF, Universidad Autónoma de Madrid, Spain;
[3] Institut de Física d'Altes Energies (IFAE), Universitat Autònoma de Barcelona, Spain; [4] Port d'Informació Científica (PIC), Barcelona, Spain;
[5] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (Ciemat), Madrid, Spain
*Corresponding author

ATLAS Computing Challenges at HL-LHC Period

- HL-LHC computing demands far outstrips the budget
- ATLAS physics already is computed-limited
  - Drives our computing approach today
- On this foundation, developing new approaches that
  - Leverage powerful networks, the most crucial factor to LHC computing success
  - Minimize/optimize storage needs
  - Make the most of use opportunistic resources (HPCs)
- Working assumptions:
  - Flat facilities budget, 20%/yr capacity growth, constant euros
  - Development effort is sustained
- HL-LHC CPU estimations showed a ~3x shortfall with respect to the flat budget model
  - This is still without leveraging co-processor/GPUs
- ~6x shortfall by today's estimate in Storage on Disk
  - Storage shortfall is our biggest problem
  - A way to dramatically reduce our storage footprint is grow the use of tape
    - This is hard: tape is slow and complicates workflow orchestration





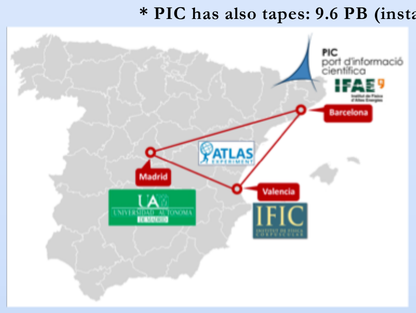


Spanish ATLAS Tier-1 & Tier-2 Federation

➢Present resources provided by the ES-ATLAS-T2:

SITE	CPU (HEP-SPEC06)(installed)	DISK (TB)(installed)
PIC-Tier-1	40024	2400
IFIC-Valencia	26751	2146
IFAE-Barcelona	10420	980
UAM-Madrid	10358	1220

* PIC has also tapes: 9.6 PB (installed)



The ES-ATLAS-T2 and Tier-1:

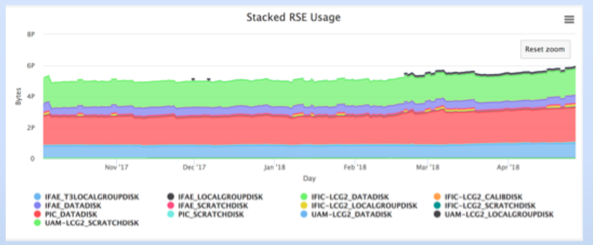
- Tier-1: PIC-Barcelona
- Tier-2: IFIC-Valencia (50%), IFAE-Barcelona (25%) & UAM- Madrid(25%)
- integrated in the WCLG (World Wide LHC Computing GRID) Project and follows the ATLAS Computing Model
- 4% of the total resources of all ATLAS Tier-2s and 5% of all Tier-1s
- Sites are T2D category (good connectivity among other Tier-1s & Tier-2s)

➢Site Availability and Performance, based on WLCG information

SITE	Availability (2017)	Reliability (2017)
PIC	98.9%	99.2%
IFAE	98,62%	98,92%
UAM	98,75%	99,41%
IFIC	98,33%	98,58%

* All sites with Availability and Reliability greater than 98%

Disk Usage in the last 6 months (6 PB storage in Spanish sites) :



- Disk pool distribution at IFIC based on Lustre since 2009: Current version 2.1.6; 2.1 PB of data
- Disk pool distribution at IFAE/PIC and UAM based on dCache since 2006: Current version 3.2 ; 4 PB of data

The Event Index Project

➢The ATLAS Event Index is a catalogue of ALL events that includes event identification, trigger and references to locate events in data files. It uses an industry Big data noSQL DB, which stores all real and simulated data, and /or Oracle storage database for fast datasets discovery.

➢ It currently contains 63 Billion entries for 2016 and 17 Billion entries for 2017 real data and MC in Hadoop.

➢The IFIC-Tier2 group is the coordinator of the Data production task where Tier0 index merged physics AODs, and Grid jobs collect info from EVNT and AOD datasets as soon as they are produced and marked "ALL EVENTS AVAILABLE" in AMI. DAOD can be indexed on demand.

➢ IFIC has developed the Data Collection tools and currently has put into production the new Object Store based approach as temporary storage with very good results.

➢The new Data Collection Supervisor controls the system and keeps the system running with consistent data.

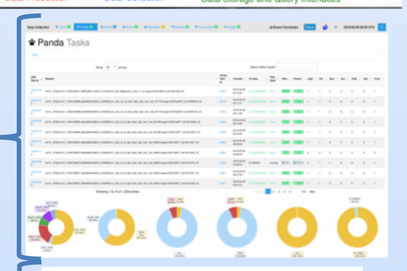
➢ IFIC developers recently added a new feature that allows to get trigger counts and overlapping in "human time" scale, processing millions of events in less than 2 minutes. Figure shows a heatmap of the trigger overlaps of a dataset.

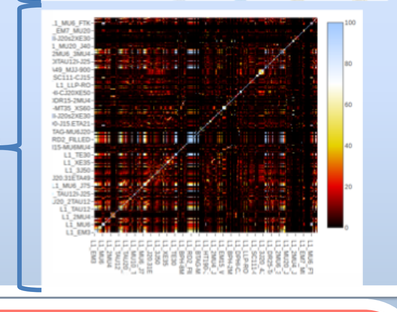
➢ IFIC is participating in the evolution of the Event Index to support the future challenges of Run3 and beyond.

➢This evolution foresees the extension to what is called an Event Whiteboard, including new capabilities like labelling individual events for building Virtual Datasets.

➢As part of this evolution IFIC is testing new backend storage technologies like Kudu, to support the new model and requirements.

General Overview diagram of Event Index





HPCs in Spain: Use of Extra Resources

➢ Explore ATLAS/HEP production and analysis workflows on HPC infrastructure. Acting as managers of external (HPC) resource providers.

➢ Software stack delivery

- ➢Singularity/docker/shifter containers
- ➢CernVM-FS

➢TEST done by IFAE & IFIC in Spanish HPCs: MareNostrum4 and Lusitania2

- ➢ARC Computing Element installed at IFIC & IFAE/PIC.
- ➢Dedicated PANDA queues set-up in AGIS for both HPCs.
- ➢Running simulation jobs using the ATLAS production system in an automatic way.

➢Combining MareNostrum4 and Lusitania2 together we got 300kh of CPU since September 2017.

➢ Number of ATLAS simulation events: more than 4 M in Lusitania2, 1000 events/job, about 2 h 20 m per job on average. More than 2.5 M in MareNostrum4.

Lusitania2

MareNostrum4

Federation Storage

▪The HL-LHC expects an increase of more than twenty times for the data permanent storage w.r.t. present.

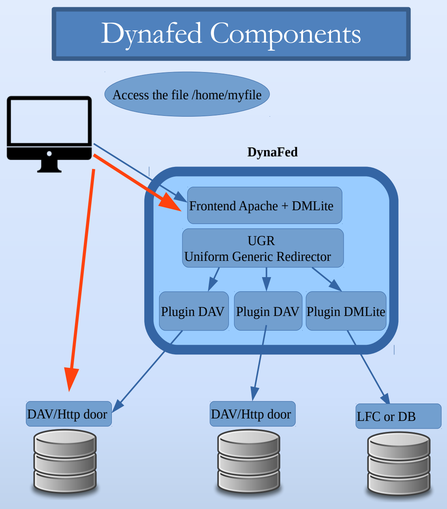
▪Spanish sites are going to participate in the CERN Data Lakes project to federate storage and reduce costs (see ATLAS core computing box).

▪ Spanish institutes are exploring solutions to federate their storage. First we have considered the validation of the tool DynaFed which meets the requirements we expect:

- ▪Aggregates storage and metadata
- ▪Exposing standard protocols that support redirections and WAN data access
- ▪A unique endpoint is presented as federated access against the site endpoints, which may be a variety of protocols (webdav, dpm, S3).

▪Spanish ATLAS sites are planning to contribute to the Eulake project, which is based on the EOS system to build Data Lakes.

Dynafed Components



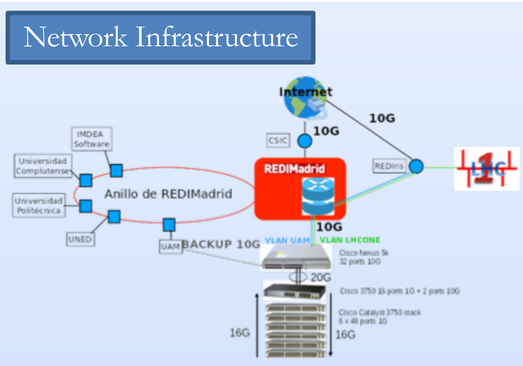
Network: LHCONE in Spain

▪LHCONE is the collaborative network implemented for the LHC research community as a controlled access routed IP internet.

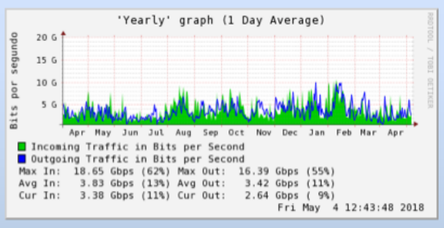
- ▪ UAM joined February 2017
  - ▪ Migration to LHCONE through REDIMadrid.
  - ▪10 Gbps, with a backup line.
  - ▪IPv6: dual-stack all Tier-2 services
- ▪ PIC-IFAE joined 2010
  - ▪ Connected through RedIRIS to LHCOPN/LHCONE at 20 Gbps, with a backup line.
  - ▪IPv6: dual-stack all Tier-1 & Tier-2 services.
- ▪ IFIC
  - ▪To be included in LHCONE infrastructure.
  - ▪IPv6: dual-stack all Tier-2 services: Gridftp, SRM, FAX, WEBDAV, perfsonar.

▪HL-LHC will require to increase the network bandwidth by a factor 10.

Network Infrastructure



Daily LHCONE traffic in RedIRIS in the last year (for PIC, IFAE & UAM)



Data Analysis (DA)

▪Activity carried out in close collaboration with the ATLAS Physicists, keeping substantial CPU resources for data analysis.

▪HEP analysis pipelines on HPC facilities using GPUs.

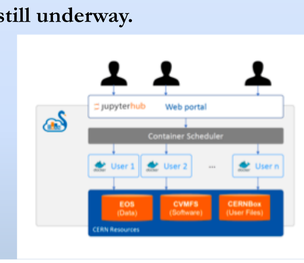
- ▪ Optimizing analysis workflows for novel compute/memory/storage capabilities.
- ▪ Utilizing industry tools like Apache Spark: enabled to read ROOT files natively.
- ▪ More and more frequently Physics analysis will be performed using Machine Learning methods, that would imply the use of GPUs and much more memory than provided by ATLAS standard worker nodes.

▪ SWAN – service for web-based analysis:

- ▪ Users can do analysis using just their browsers. Cloud-based analysis model.
- ▪ Jupyter Notebooks on demand. Use personalized docker container instance and scale to larger resources like Spark clusters.
- ▪It is a first step towards a scalable & interactive distributed data analysis environment.

▪IFIC's effort to implement SWAN was mainly focused on adapting this service to the IFIC environment.

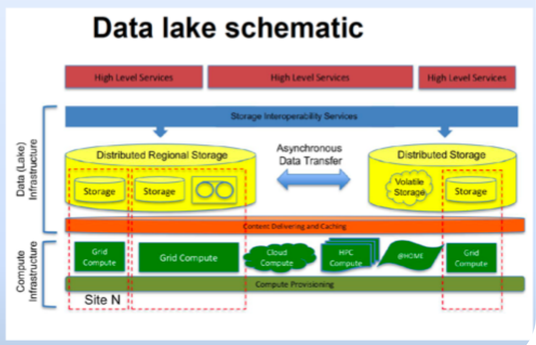
▪ Currently the IFIC implementation is in a "proof-of-concept" stage so that the service works properly but the full platform is still underway.



ATLAS Core Computing

▪ PIC and IFAE are involved in the Working Groups of the WLCG project to develop tools and services for the HEP Software Foundation.

- ▪ Data Lake WG project: integrated consolidation of distributed storage (and compute) facilities, leveraging high-bandwidth networks.
- ▪Computing Resource Information System: distinct description of physical Computing/Storage resources provided by Sites and their usage by Experiment-specific Computing Model.



Conclusions

- Spanish ATLAS Tier-1 &Tier-2 infrastructure started in 2005. It is offering 5% and 4% of all ATLAS Tier-1 & Tier-2 resources respectively.
- ATLAS will need x20 more resources at HL-LHC with respect to today. Flat budget and +20%year from technology evolution fills part of this gap but there is still a factor x5 missing. Storage looks like the main challenge to address.
- The ATLAS Spanish LHC computing community is participating to establish the main guidelines for the next years (HL-LHC). Moving towards common implementations: Data Management, Data Lakes, HPC, Resources Federations, Grid resources, Machine/Deep learning facilities, and increasing the network bandwidth.