Análisis en el Grid

Santiago González de al Hoz

1) Almacenamiento de datos de oficiales y privados.
2) Use cases.
3) Resultados del test Step09.

Almacenamiento en ATAS

- Almacenamiento en el Grid
  - Managed by the ATLAS Distributed Data Management system (DQ2)
    - Should be accessed only with the official tools; dq2-* commands, distributed analysis tools, official production system
    - Accounting by DQ2
- Almacenamiento no a la Grid para acceso local
  - No access from grid = not known to DQ2
  - Accounting by local tools
- En nuestro caso utilizamos Lustre

ATLAS spaces en la Grid

Name	Usage	T1	T2
ATLAS**DATA**DISK	official data from the detector	40%	30%
ATLAS**MC**DISK	official data from simulation	35%	25%
ATLAS**GROUP**DISK	group data	10%	20%
ATLAS**SCRATCH**DISK	temporary space for users	10%	20%
ATLAS**LOCALGROUP**DISK	locally managed space	(non-pledge)	

- Shares of the spaces are estimated for STEP09 following the ATLAS computing model
- They will be adjusted with the real experience
- 'Official data' can also be placed on 'group' spaces according to the need of groups (but treated as group data)

Control de acceso a los datos

- **Access control has been introduced**

  to the ATLAS Distributed Data Management system (DQ2) on 17 June
- Official Data Space (DataDisk, MCDisk)
  - Managed centrally
  - Group/User access is limited to read access
    - No write access, No private replications to the space
- Group/User spaces are (being) prepared
  - GroupDisk: only group data managers can put data
  - ScratchDisk: scratch space for anybody at every grid site
  - LocalGroupDisk: (possible) permanent space at the “home” institute (country)

GROUPDISK

- Only for the groups registered on Grid
  (eg. det-muon, perf-egamma, phys-beauty)
  - To store group data permanently
  - Data should be managed by the group data manager(s).
  - No direct write access – put data via official DQ2 replication
- Space name in DQ2: <SITENAME>_<GROUPNAME>
  - eg. IN2P3-LAPP_PHYS-SM, TOKYO-LCG2_PERF-JETS
  - Not in the US sites for the moment
- Quota: Accounting on daily basis (time-lag up to 1 day)
- Size: ATLAS "group space panel" will decide how much to allocate from the available resources
  - For the current tests: a small amount allocated on request
- Tools and instructions have been prepared:
  https://twiki.cern.ch/twiki/bin/view/Atlas/GroupsOnGrid

LOCALGROUPDISK

- Available at many grid sites (up to the sites)
  - To store user data permanently at "home" (where they belong to)
  - Data managed using the DQ2 tools
  - Space managed by the site or the "local" group
- Space name in DQ2: <SITENAME>_LOCALGROUPDISK
- Size: up to the site (not ATLAS requirement)
- Quota: no quota, the site (the local group) manages
- Other ways of "local" space deployment are not prevented, but they cannot be accessed with DQ2

SCRATCHDISK

- For any atlas users on Grid at every grid site
  - To store the output of user jobs run at the site
  - To upload files onto DQ2 (dq2-put)
  - Temporary storage (Renamed from USERDISK)
- Space name in DQ2: <SITENAME>_SCRATCHDISK
- Size: should scale according to the computing power at the site, but real scale is not clear yet...
- Quota: no quota yet, but possibly per user in future

- User data are to be created at first on SCRATCHDISK to protect permanent spaces from unregistered files created by failures

- Data on *_SCRATCHDISK will be deleted after ~ 1 month
  - In order to keep them on grid (DQ2);
    - **Download** them with the tool: dq2-get
    - CAUTION: Athena (POOL) files cannot be re-loaded onto DQ2 once removed
  - In order to keep the files, but not necessarily on grid (DQ2);
    - Users should request for **replication** to their permanent storage (LOCALGROUPDISK, GROUPDISK)
  - If not downloaded/replicated
    - The data will be lost

USE CASE:
Distributed analvsis data flow

- Jobs
  - Run on a grid site where input data are available on DISK
- Output data
  - are stored on the **SCRATCHDISK** space at the site **temporarily**
- The user should either
  - download the data (to their local disk) and let them be removed from DQ2
  - or request for replication (to permanent storage on grid) to keep them on DQ2

USE CASE:
Uploading data on the Grid

- Upload data
  - Created using local batch (or interactive) system
  - Upload on the **SCRATCHDISK** space at the site **temporarily**
- Replicate data
  - To the permanent space (GROUPDISK space, or LOCALGROUPDISK)

En el futuro discutir espacio no a la Grid (espacio del Tier3)
para guardar los datos

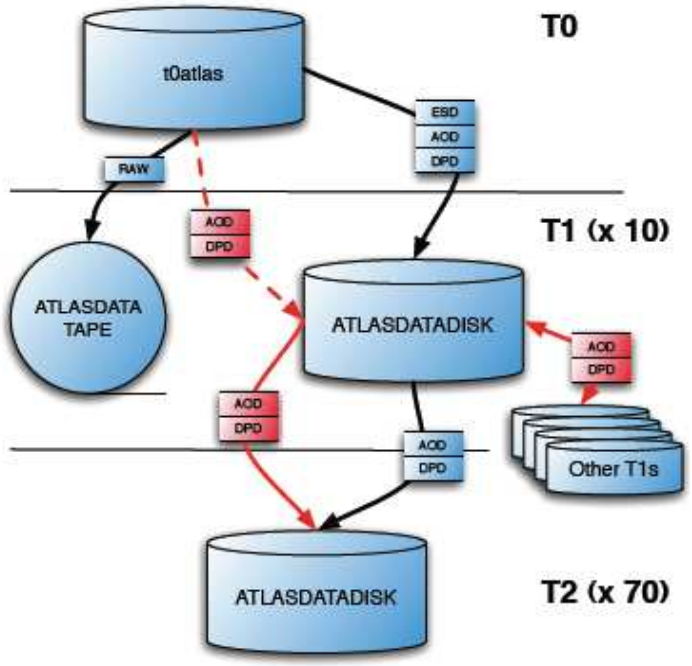
STEP09

- **Primer test de uso real, es decir trabajos de simulación montecarlo y análisis corriendo a la vez!!!! Y además los trabajos privados de los usuarios.**
- **Fechas**
  - Mayo 25 - 31 Setup
  - **Junio 2 - 14 Corriendo**
- **Actividad de los Tier-2s**
  - Los Tier-2s participaron produciendo los datos simulados de Montecarlo y corriendo trabajos de análisis
    - 50% simulación, 50% análisis (25% ganga, 25% pilot jobs).
  - Almacenamiento de los datos necesarios para el análisis
- **Distribución de datos**
  - El volumen de datos distribuidos a los Tier-2s fue de 112TB durante las 2 semanas que duró el test (**50% IFIC, 25% IFAE, 25% UAM)**
- **Tier-1**
  - Participaron almacenado los Raw Data y los datos producidos por la simulación de MC en los Tier-2s. Reprocesando los Raw Data y transfiriendo los datos procesados a los Tier-2s.

Distribución de los datos

- **ATLAS**
  - **DDM (*Distributed Data Management*)**
- **Gestión de datos siguiendo la topología (Tier-0→Tier-1s →Tier-2s)**
- **Basado en herramientas de transferencia de datos Grid**
  - **Registro de los datos en el catálogo**
  - **Distribución de datos siguiendo la política adecuada**
  - **Priorización, obtener datos del sitio más cercano, etc..**

- Data taking and first reconstruction passes
  - RAW and ESD from CERN → distributed to T1 sites (1, 2 copies respectively, RAW to tape)
  - AOD and DPD from CERN distributed to all T1 sites (10 copies)
  - AOD and DPD from CERN distributed to T2 from their parent T1 (1 to 2.7 copies per cloud)
- Reprocessing at Tier-1s →
  - AOD and DPD distributed from all T1s to all other T1s
  - AOD and DPD from all T1s distributed to all T2s from their parent T1
    - In STEP this involved an extra T0→T1 step, but this is a minor perturbation - - - →

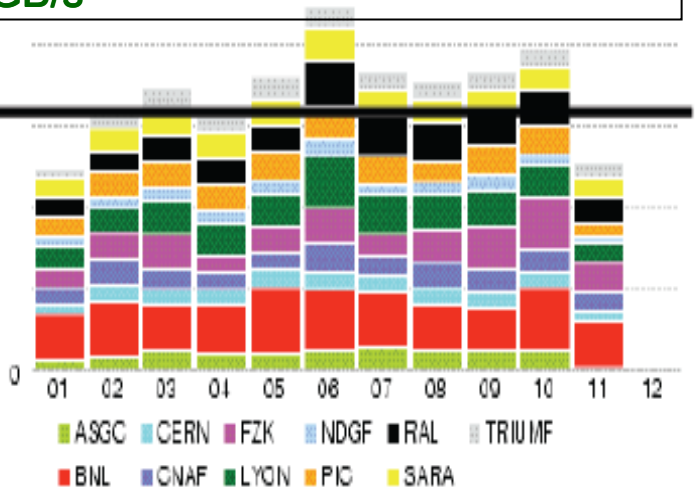


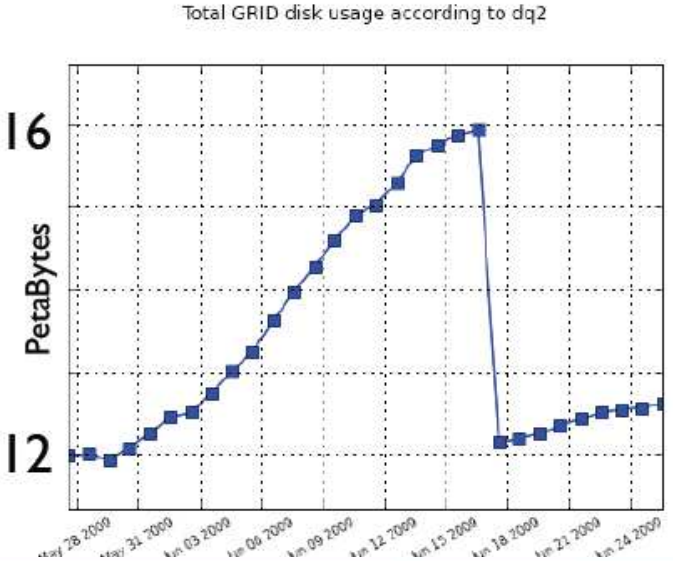
Resultados distribución de datos

- **Resultados para ATLAS**
  - **4PB de datos distribuidos**
    - **Ficheros grandes!!! Raw data**
  - **Correcto funcionamiento de los Tier-1s**
  - **54/67 Tier-2s funcionado correctamente**
  - **Problemas: Inestabilidad de los SE, problemas con el espacio, cuellos de botellas en la red, transferencia de datos**

Requerido cuando funcione el LHC:
1-2 GB/s

3GB/s

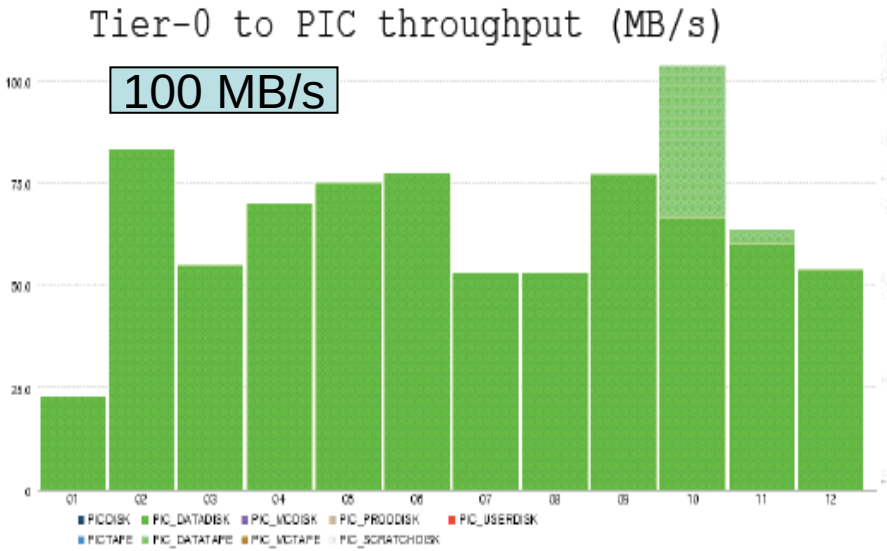


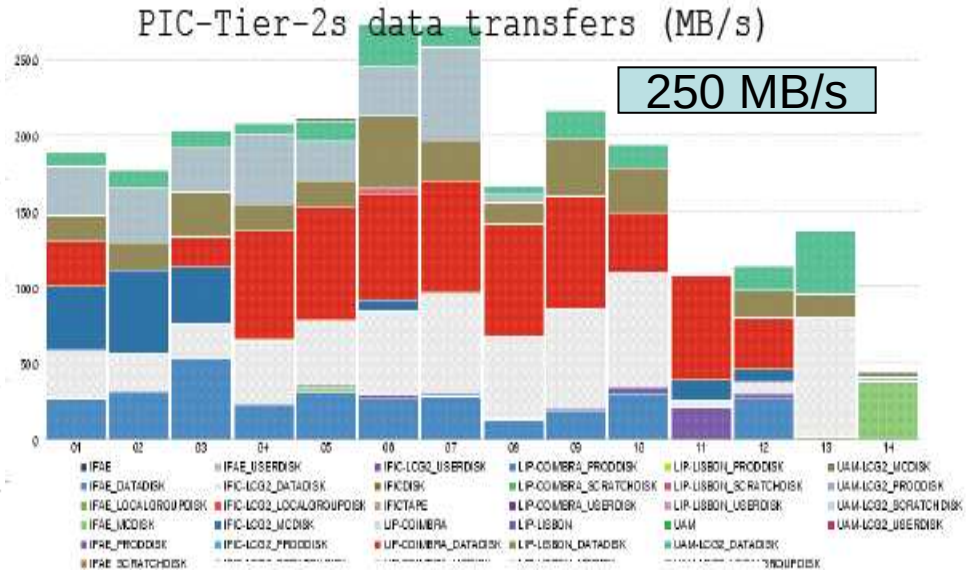


Cloud	Efficiency	Transfers Throughput
ASGC	99%	397 MB/s
BNL	84%	1128 MB/s
CERN	100%	334 MB/s
CNAF	98%	561 MB/s
FZK	85%	556 MB/s
LYON	96%	620 MB/s
NDGF	84%	137 MB/s
PIC	93%	429 MB/s
RAL	99%	838 MB/s
SARA	53%	262 MB/s
TRIUMF	100%	297 MB/s

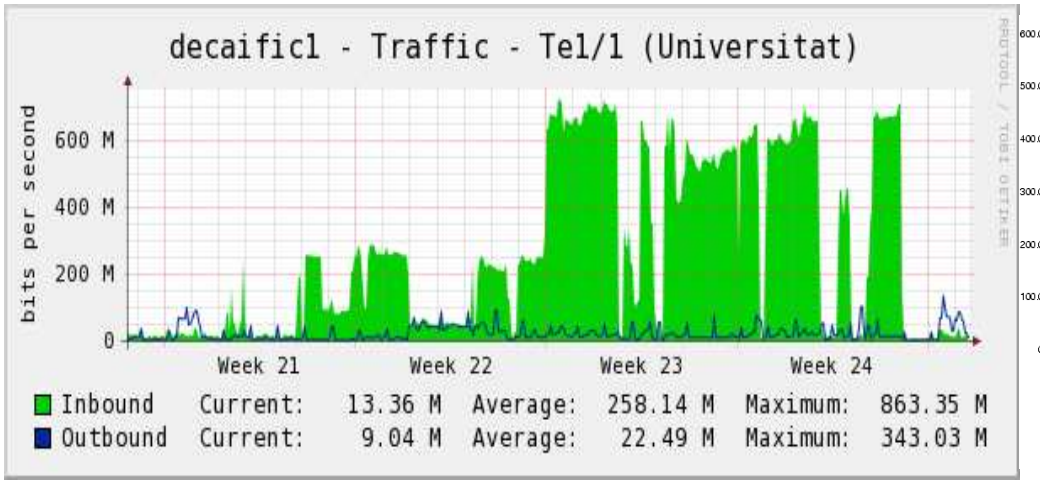
Peaks of 5.5GB/s

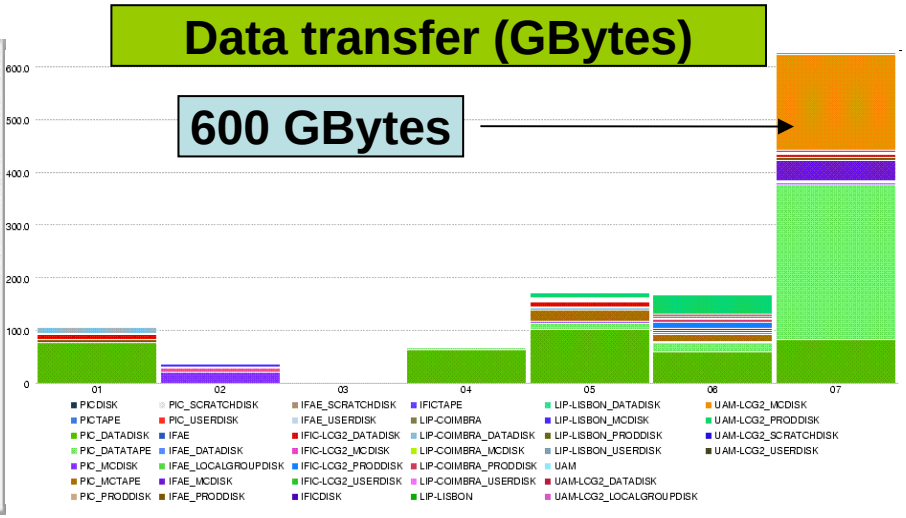
Resultados distribución de datos en los centros españoles





- **Picos de 600Mbps observados en el IFIC**

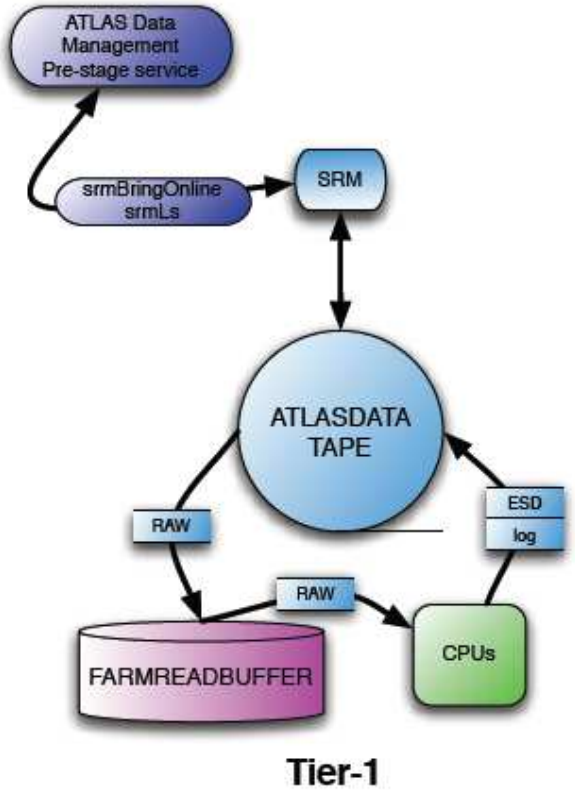




Reprocesado de datos

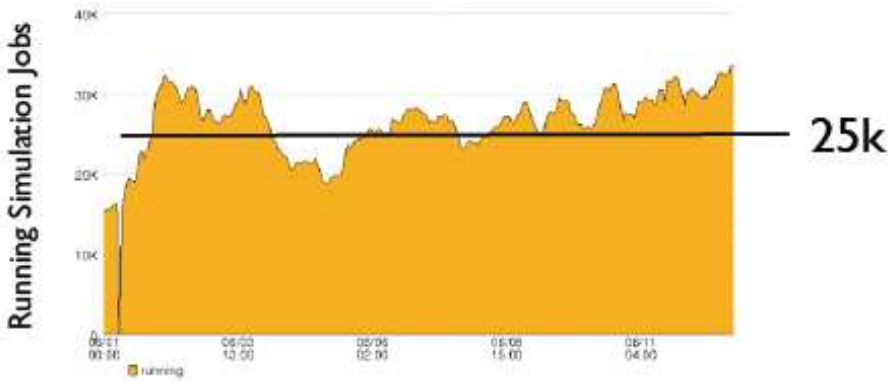
- **Validar el reprocesado de datos utilizando las cintas de los Tier-1s.**
  - ***Pre-stagein* desde la cinta**
  - **Producir ESD de los datos RAW**
  - **Escribir los nuevos datos ESD en las cintas**
- **Objetivo reprocesar más rápido de 200 HZ el cual es la frecuencia nominal de la toma de datos**
  - **2 Tier-1s reprocesaron a 400 Hz (x2)**
  - **5 Tier-1s reprocesaron a 1000 Hz (5)**

T1	Base Target	Result	Comment
ASGC	10 000	4 782	Many batch system and basic setup problems
BNL + SLAC	50 000	99 276	Also ran high priority validation and other tasks
CNAF	10 000	29 997 ★	
FZK	20 000	17 954	Big tape system problems pre-STEP; no CMS
LYON	30 000	29 187	Very late start due to tape system upgrade, then good
NDGF	10 000	28 571 ★	
PIC	10 000	47 262 ★	
RAL	20 000	77 017 ★	
SARA	30 000	28 729	Tape system performance very patchy
TRIUMF	10 000	32 481 ★	Also ran high priority validation and other tasks

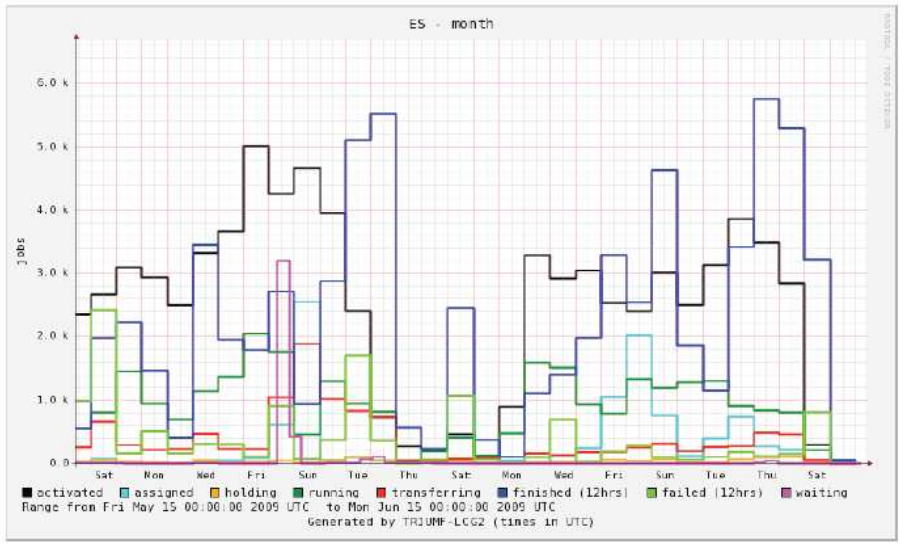


Producción de MonteCarlo

La simulación durante el step09 para ATLAS produjo 12 millones de sucesos llenando cualquier recurso libre

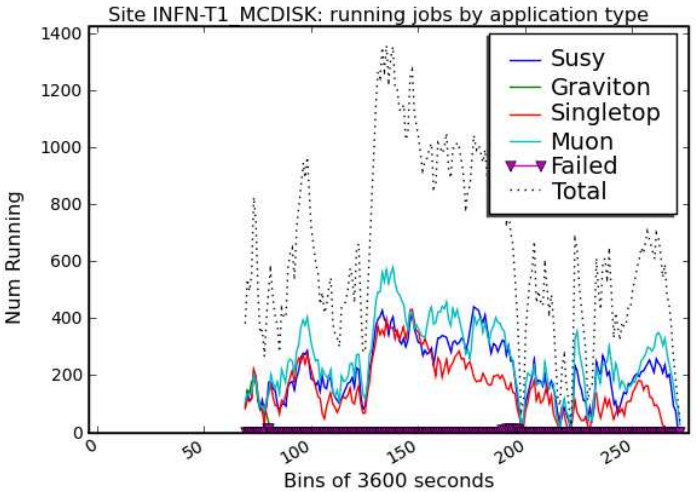


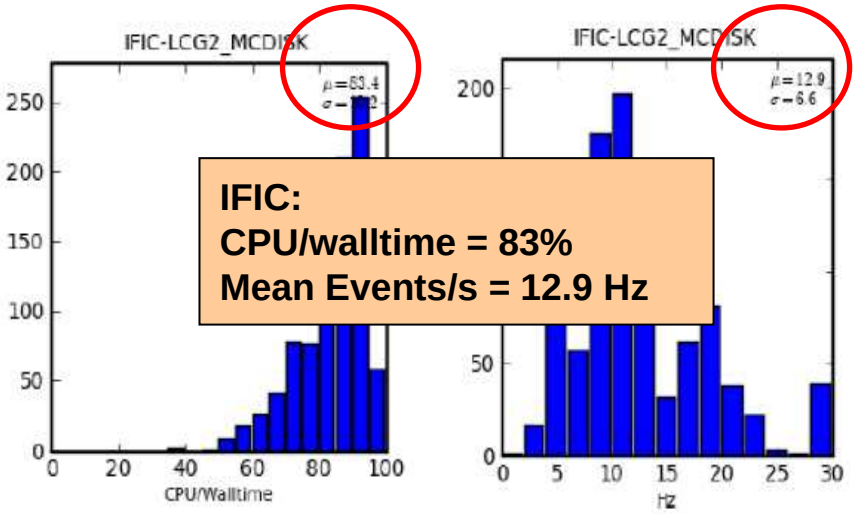
- **Durante los step09:**
  - **La producción de MC en el Tier-2 español ha estado entre 1K y 1.5 K trabajos por día alcanzando una eficiencia mayor del 90%**



Trabajos de Análisis para ATLAS durante los STEP09

- Corrieron 4 análisis reales sobre AOD
  - 50% producción de MC
  - 50% análisis
- 1M de trabajos enviados, con una eficiencia del 83.4%
- 26.3 B sucesos procesados
- Mean Events/s = 7.7Hz
- Mean CPU/Walltime = 39%

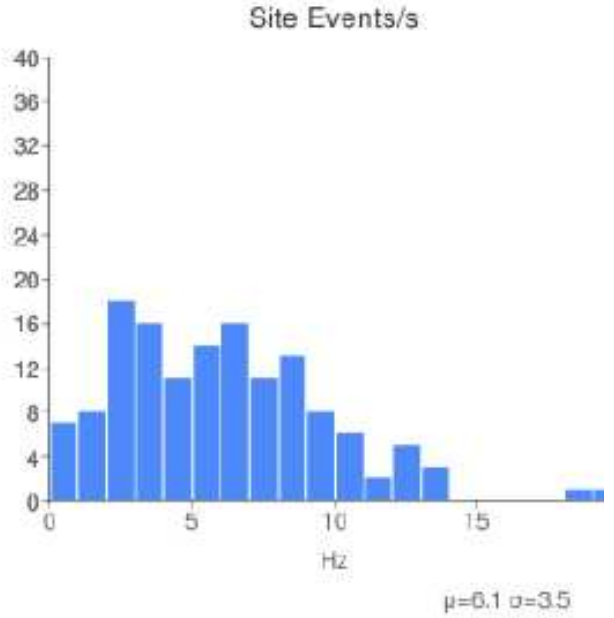




CLOUD	SUBMITTED	RUNNING	COMPLETED	FAILED	Efficiency	# Files
CA	0	0	32110	9848	0.77	87117
DE	0	0	132103	44853	0.75	557395
ES	0	0	62113	10651	0.85	236690
FR	0	0	143628	22927	0.86	561978
IT	0	0	52464	7101	0.88	311061
NG	0	0	14551	2919	0.83	20179
NL	0	0	36327	30376	0.54	154452
T0	0	0	0	1151	0.00	0
TW	0	0	19459	4910	0.80	86293
UK	0	0	143670	38190	0.79	439084
US	0	0	153609	9770	0.94	467732
TOTAL	0	0	790034	182696	0.81	2921981

Ganadores por centro: Sucesos procesados por segundo (Hz)

		#comp	Hz	
1.	ANALY_CSTCDIE	121	19.2	**
2.	ANALY_TW-FTT	14634	18.4	
3.	ANALY_DESY-ZN	8277	13.7	
4.	IFIC-LCG2_MCDISK	1330	13.6	
5.	TAIWAN-LCG2_MCDISK	175	13.2	**
6.	grid03.unige.ch	2136	12.6	**
7.	ANALY_LAPP	27711	12.6	
8.	ANALY_IHEP	7291	12.4	
9.	LIP-LISBON_MCDISK	4113	12.3	
10.	ANALY_PIC	32077	12.0	
11.	ANALY_TOKYO	21271	11.4	
12.	JINR-LCG2_MCDISK	68	11.0	**
13.	ANALY_BHAM	2640	10.7	
14.	ANALY_DESY-HH	14437	10.6	
15.	ANALY_LRZ	7514	10.5	
16.	TOKYO-LCG2_MCDISK	14674	10.4	
17.	ANALY_GRIF-LAL	16974	10.2	
18.	DESY-ZN_MCDISK	17870	10.2	

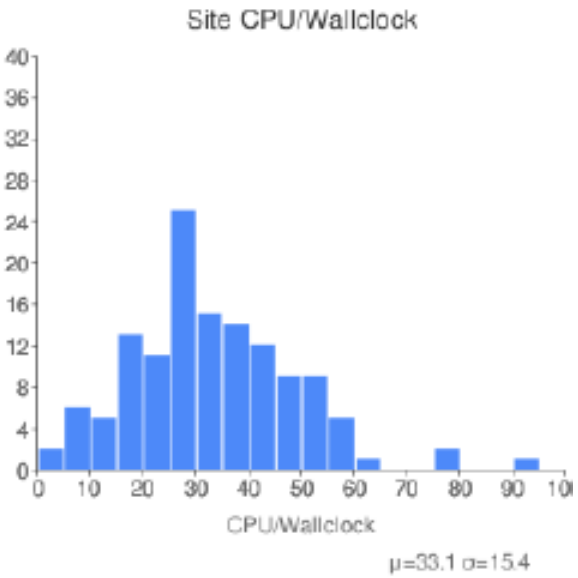


Hz = avg(#events/(stoptime-starttime))
Note: potential bias between Panda/WMS

* site did not run many jobs

Ganadores por centro: CPU/Wall time

		#comp	CPU/W	
1.	TAIWAN-LCG2_MCDISK	175	94.2	**
2.	JINR-LCG2_MCDISK	68	79.5	**
3.	IFIC-LCG2_MCDISK	1330	75.8	
4.	ANALY_SLAC	13086	64.8	
5.	ANALY_BEIJING	8653	57.7	
6.	ANALY_LAPP	27711	57.3	
7.	DESY-ZN_MCDISK	17870	57.2	
8.	ANALY_CSTCDIE	121	57.0	**
9.	ANALY_DESY-ZN	8277	55.7	
10.	ANALY_LIV	7642	54.8	
11.	UKI-SCOTGRID-DURHAM_MCDISK	22	54.8	**
12.	TR-10-ULAKBIM_MCDISK	1406	52.7	
13.	ANALY_FREIBURG	3712	51.8	
14.	ANALY_TW-FTT	14634	51.5	
15.	ANALY_CSCS	17728	51.3	
16.	ANALY_PIC	32077	51.0	
17.	TOKYO-LCG2_MCDISK	14674	50.9	
18.	ANALY_SHEF	5106	50.8	

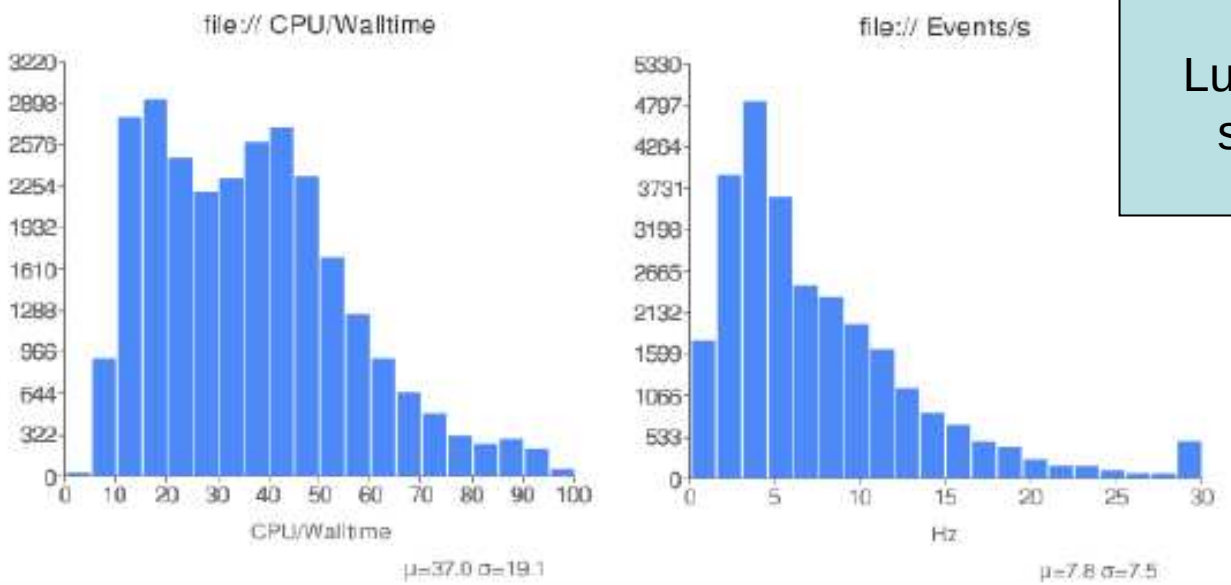


For gLite: cpu/wall = "Percent of CPU this job got..." as reported by time athena...
*For Panda: cpu/wall = 100 * cpuConsumptionTime / cpuConversion / (endTime - startTime)*

* site did not run many jobs

Ganadores de acceso más rápido a los datos

file:// protocol at CNAF, IFIC, LIP

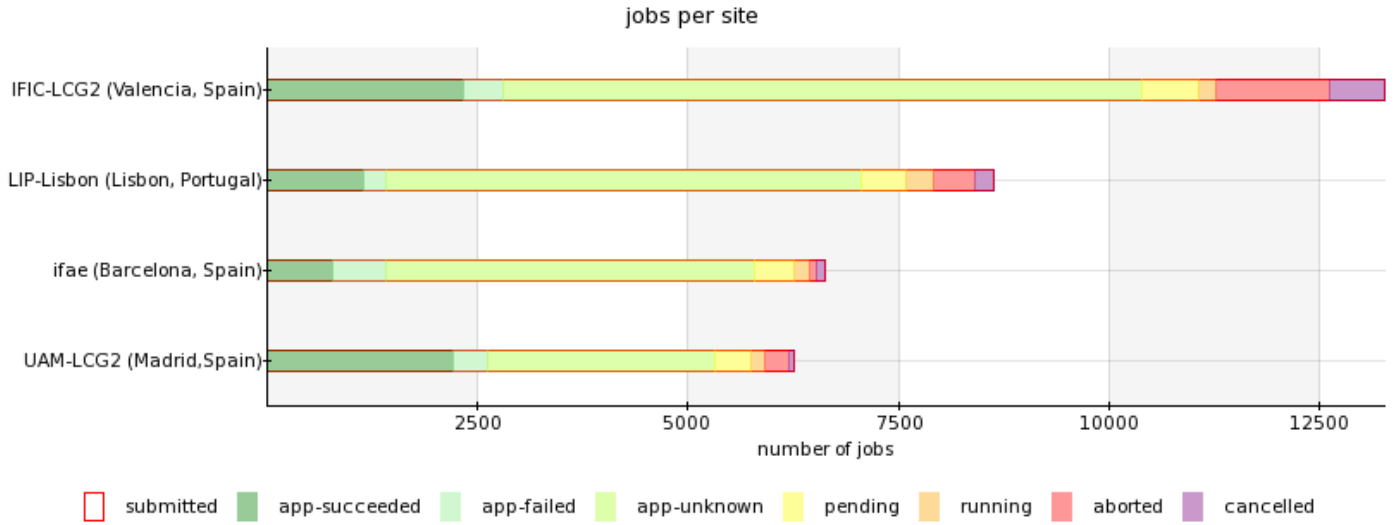


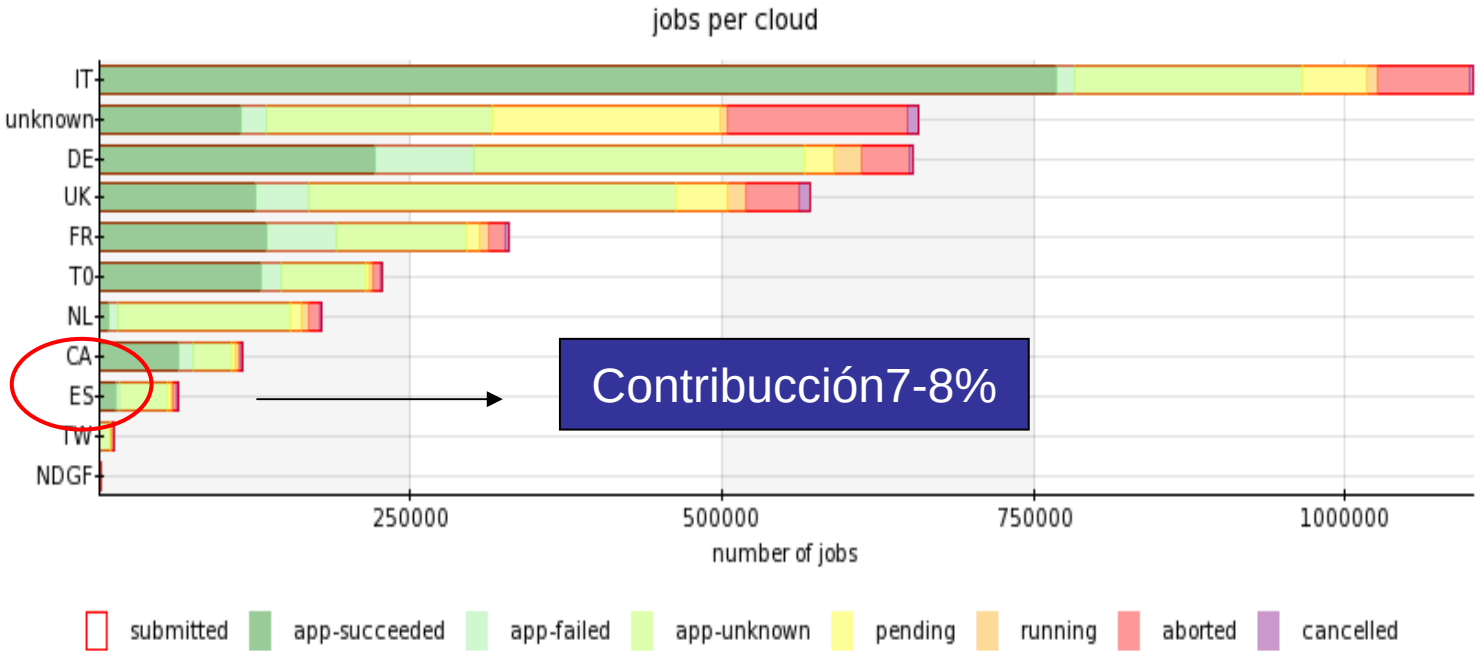
Lustre file system

Better than dcap/rfio and FileStager:
- only a few very slow jobs
- still the storage is the bottleneck

Trabajos análisis en 2009

En la "*cloud*" del PIC
Enero-Julio 09





Contribucción7-8%

Por "clouds"
Enero-Julio 09

Conclusiones Step09

- Sistema de computación para el LHC está en operación
  - Tests a nivel de transferencia de datos, simulación y análisis distribuido.
  - Tests que llevan el sistema al límite (stress-tested)

- **¡¡¡Listos para cuando empiece la toma real de datos en el LHC!!!**

- **Nuevo test a principio de Septiembre**

Recursos de computación 2010

ATLAS	Old 2009	Current 2010	Ratio	CMS 2010
Inputs	*2006*	*Revised*		
CERN				
CPU (kHS06)	53.6	67	1.25	101.8
Disk (PB)	4.0	4.0	1.00	4.2
Tape (PB)	9.7	9.0	0.93	10.4
Tier-1				
CPU (kHS06)	113.6	217	1.91	110.8
Disk (PB)	20.9	21.9	1.05	13
Tape (PB)	15.8	14.2	0.90	20.6
Tier-2				
CPU (kHS06)	108	240	2.22	200.7
Disk (PB)	13.3	20.9	1.57	10.7

Main differences 2009-2010 (for same live time, 6 Msec) from:

- updated event size (x1.5 for ESD, AOD)
- DPD added
- updated simulation CPU/evt: x3(η coverage) x2(hadronic shower model) x3(high-p_T samples)
- number of fully-simulated events: 900 M now (was ~200 M)
- number of AOD/DPD copies at Tiers-1 from 10 to 2 in 2010 to reduce resource requirements

If 2009 requests updated for event size and sim CPU, increase is mainly in Tier-2 disk (35%)

Main differences ATLAS-CMS:

- CMS: more CPU at CERN due to larger impact of pile-up on reconstruction
- ATLAS: more CPU at Tier-1s mainly because of CPU/event for G4 simulation (x8 larger due to η coverage and hadronic shower model)
- ATLAS: more disk at Tier-1s and less tape (don't want to rely on tape …)
- ATLAS: more disk at Tier-2s: 10 copies of AOD/DPD (one per cloud); CMS has only 2 copies

Our motivation: make data access easy and fast for everyone in the Collaboration