UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA
XXXII Bienal de Física
19º Encuentro Ibérico
7 al 11 de septiembre de 2009, Ciudad Real

Uso de las tecnologías Grid en los experimentos del LHC ante la inminente toma de datos

Santiago González de la Hoz

Instituto de Física Corpuscular - Valencia

XXXII Bienal de Física

7 de Septiembre de 2009

IFIC
INSTITUT DE FÍSICA CORPUSCULAR

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
CSIC

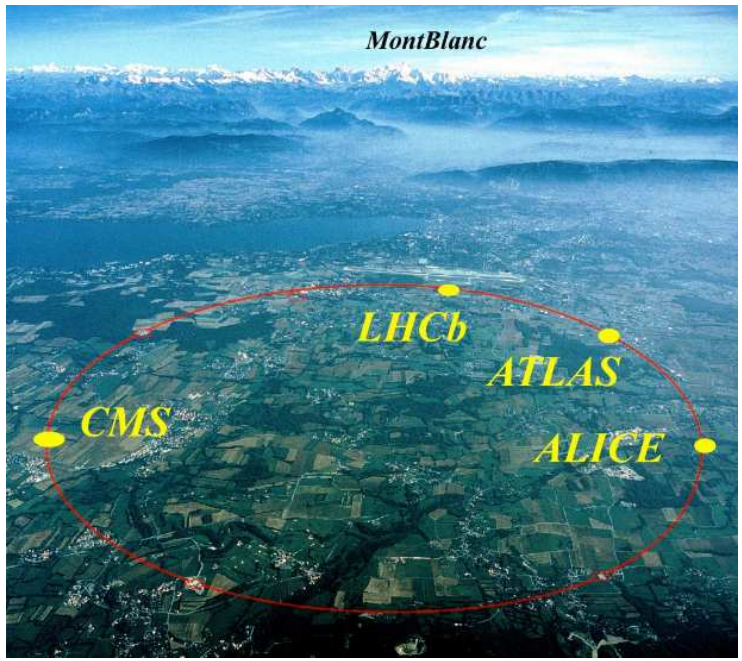
Índice

- El acelerador LHC (*Large Hadron Collider*)
- El reto del LHC desde el punto de vista de la computación
  - ¡¡¡La demanda de recursos de computación de los experimentos de física de partículas actuales no tiene precedente alguno!!!
- Las tecnologías Grid (El Grid)
- Ejemplos de aplicación de las tecnologías Grid en los experimentos de Altas Energías: ATLAS y CMS
  - Modelo de datos y de "*Computing*"
  - Uso actual
    - Recursos Españoles: Tier1, Tier2-ATLAS, Tier2-CMS
    - La Red (*Network connectivity*)
    - Producción de sucesos simulados de Monte Carlo
    - Análisis distribuido
- Conclusiones
  - ¿Estamos listos para analizar los datos cuando el LHC empiece a funcionar en el(los) próximo(s) mes(es)?

LHC

- Preguntas todavía sin respuesta
  - ¿El origen de la masa?
  - ¿De que esta hecho el 96% del universo? Materia Oscura, Energía oscura,…
  - ¿Cómo era el “universo” en los instantes iniciales justo después del Big Bang?
  - ¿Porqué hay más materia que antimateria?
  - …..

- La respuesta a estas cuestiones puede ser que estén en el LHC:
  - El instrumento científico más grande jamás construido

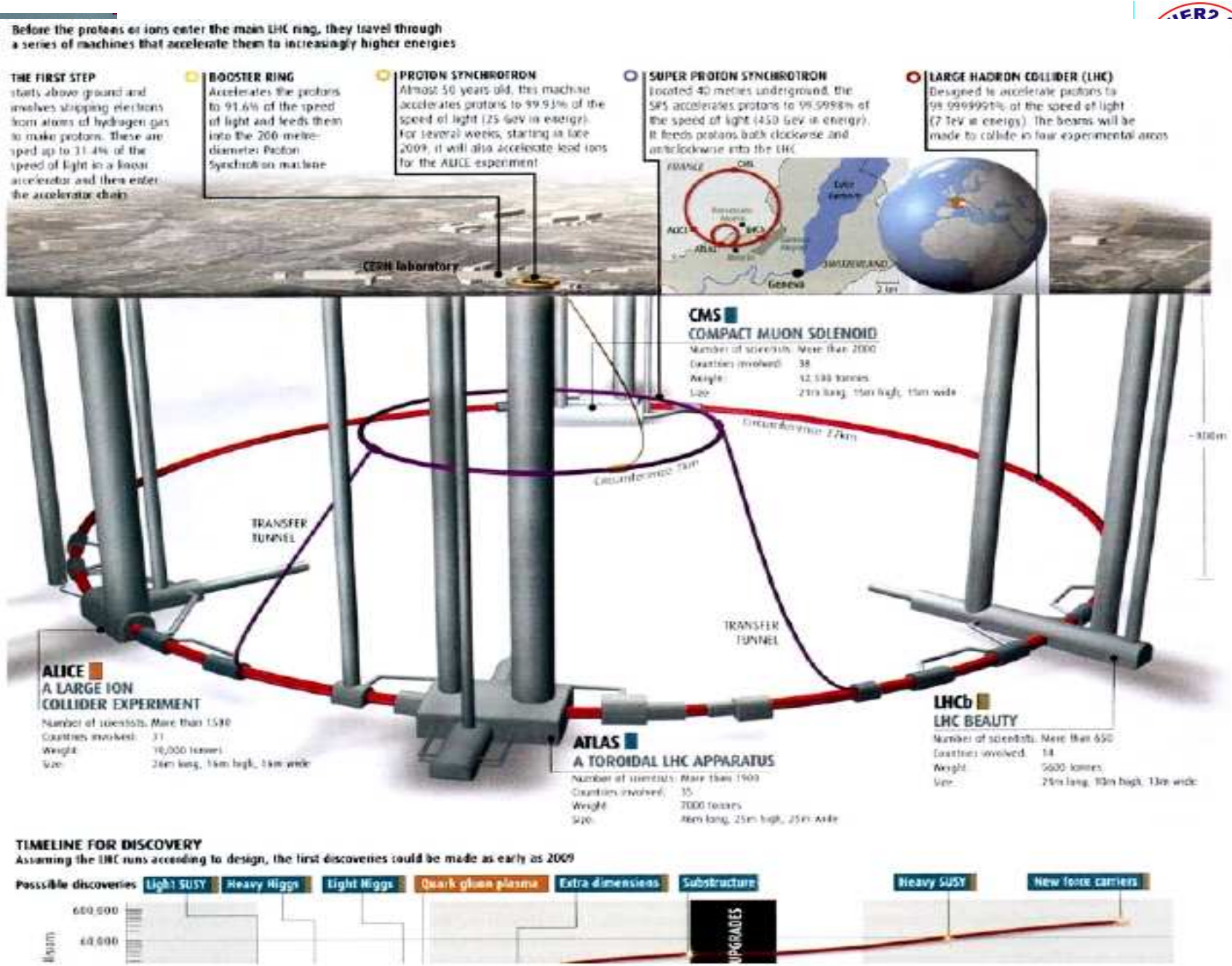


27 Km de túnel a 100 m bajo tierra

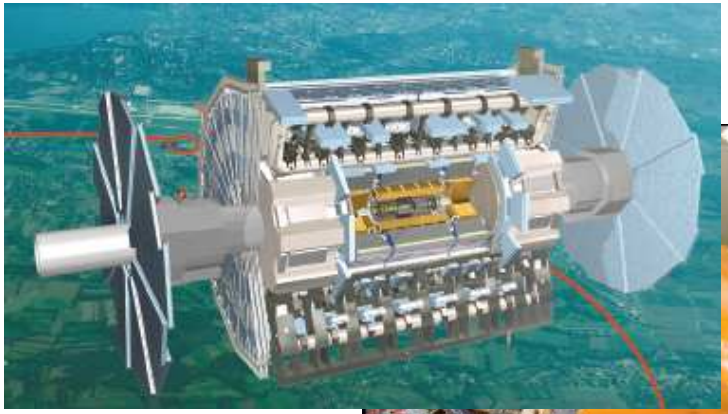
LHC: Tecnología al límite

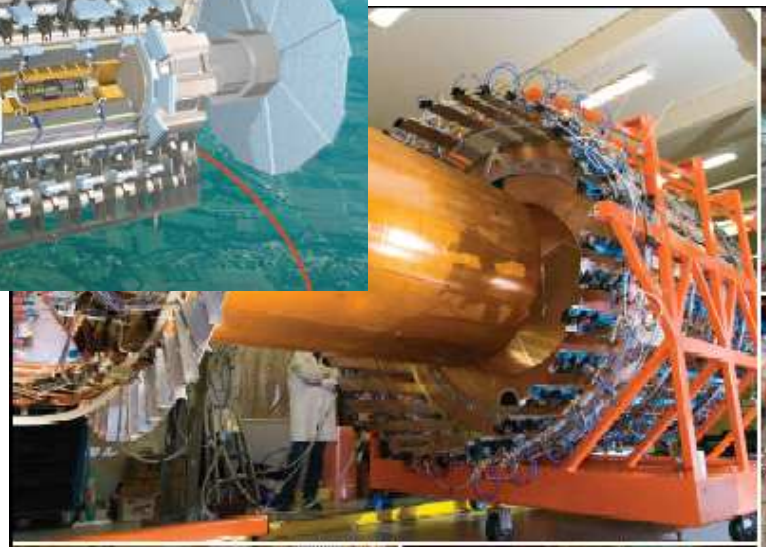


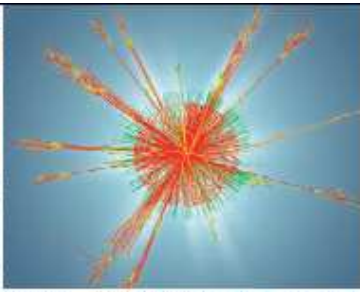

- Máquina que acelera dos haces de partículas, sentidos opuestos, hasta más del 99,9% de la velocidad de la luz.
- Los protones alcanzarán una energía de 7 TeV, con energía de la colisión 14 TeV.
- La energía de colisión de los haces de iones será de 1150 TeV (iones de plomo tienen muchos protones).
- Utiliza unos 1800 sistemas de electroimanes superconductores, funcionando a una temperatura de 1,9 K (-271ºC).
- Campo magnético generado en el LHC será de 8 Teslas


Before the protons or ions enter the main LHC ring, they travel through a series of machines that accelerate them to increasingly higher energies
THE FIRST STEP
starts above ground and involves stripping electrons from atoms of hydrogen gas to make protons. These are sped up to 31.4% of the speed of light in a linear accelerator and then enter the accelerator chain
BOOSTER RING
Accelerates the protons to 91.6% of the speed of light and feeds them into the 200 metre-diameter Proton Synchrotron machine
PROTON SYNCHROTRON
Almost 50 years old, this machine accelerates protons to 99.93% of the speed of light (25 GeV in energy). For several weeks, starting in late 2009, it will also accelerate lead ions for the ALICE experiment
SUPER PROTON SYNCHROTRON
Located 40 metres underground, the SPS accelerates protons to 99.9998% of the speed of light (450 GeV in energy). It feeds protons both clockwise and anticlockwise into the LHC
LARGE HADRON COLLIDER (LHC)
Designed to accelerate protons to 99.999991% of the speed of light (7 TeV in energy). The beams will be made to collide in four experimental areas
CERN laboratory
Geneva
CMS
COMPACT MUON SOLENOID
TRANSFER TUNNEL
TRANSFER TUNNEL
ALICE
A LARGE ION COLLIDER EXPERIMENT
ATLAS
A TOROIDAL LHC APPARATUS
LHCb
LHC BEAUTY
TIMELINE FOR DISCOVERY
Assuming the LHC runs according to design, the first discoveries could be made as early as 2009
Possible discoveries
Light SUSY
Heavy Higgs
Light Higgs
Quark-gluon plasma
Extra dimensions
Substructure
UPGRADES
Heavy SUSY
New force carriers
600 000
60 000
ilsum

El experimento ATLAS del LHC



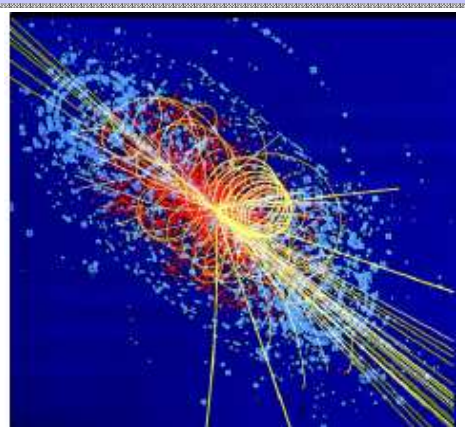


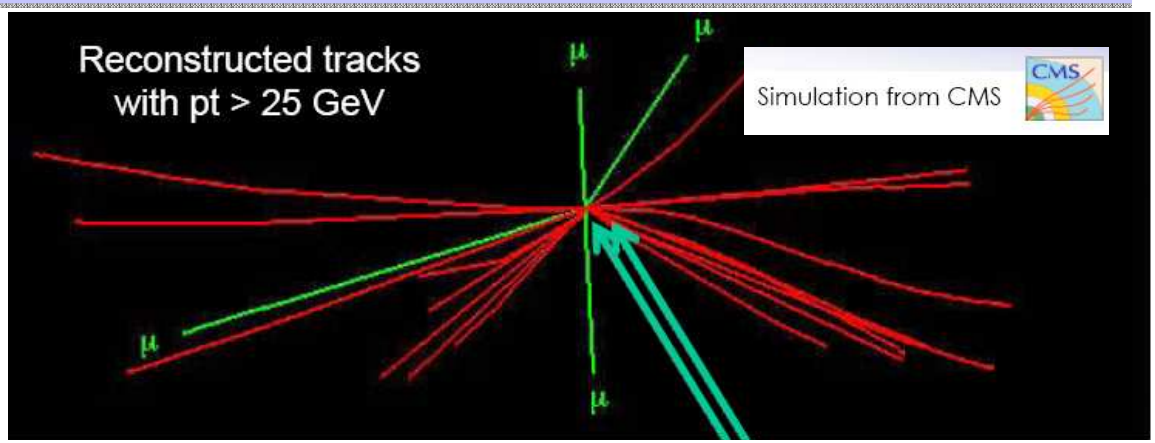


25 m de alto
46 m de largo

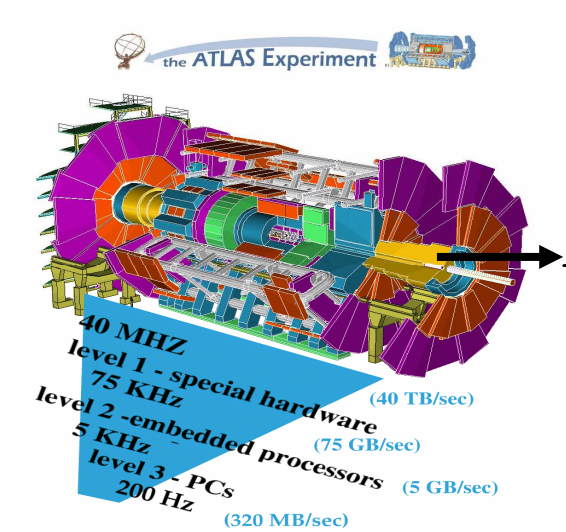
Detector de mayor volumen jamás construido para la física de partículas

El reto de la computación

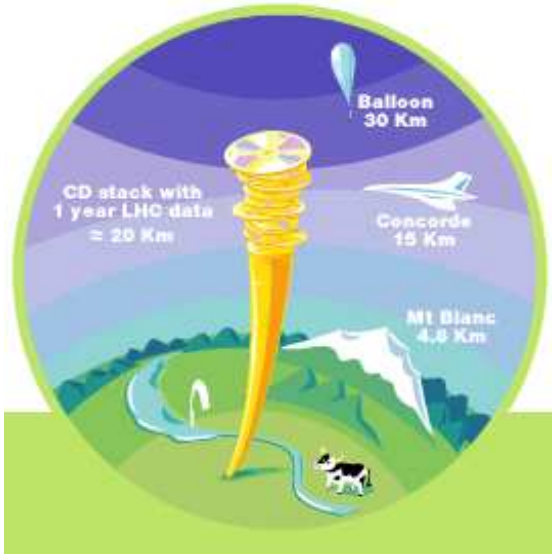




- Reconstrucción de los sucesos que ocurran en los detectores. 40000 PCs de hoy en día para hacer el reprocesado.

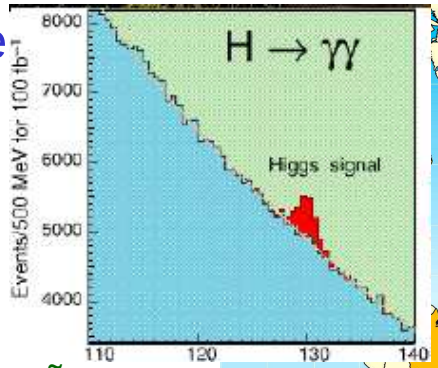


- Adquisición de datos
  - 320 MB/sec
  - 200 Hz ~10^9 sucesos/año
  - Tamaño medio del suceso (*raw data*): 1.6 MB/suceso
- Almacenamiento
  - 5-8 PB/año (frecuencia de toma de datos 320MB/sec)



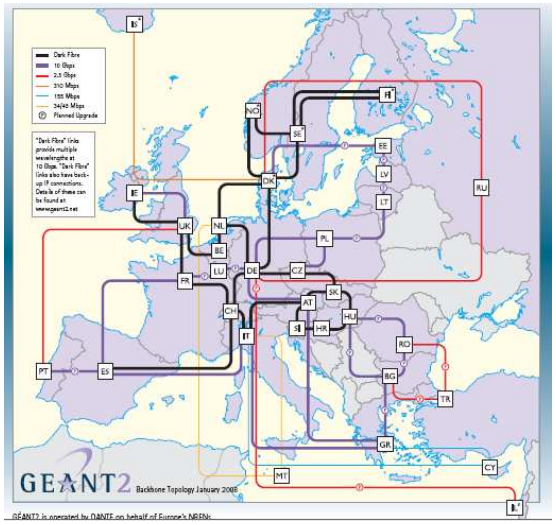
El reto de la computación

- Señal/ruido: 10^{-9} (Reprocesado de datos)
- Volumen de datos
  - Alta frecuencia * gran número de canales de cada detector * 4 experimentos
    - 15 petabytes de datos nuevos cada año
- Potencia de computación
  - Complejidad de cada suceso * núm. de sucesos * miles de usuarios
    - 40k PC's rápidos de hoy en día
    - 45 PB de almacenamiento en disco
- Distribución de los recursos a nivel mundial
  - Recursos localizados en regiones y países.
  - Análisis distribuido
  - Red estable y rápida



Europe: 267 centers/institutes, 4603 users
Rest of the world: 208 centers, 1632 users

Tecnologías Grid



Las tecnologías Grid

- Una propuesta en analogía con la red eléctrica ("electrical grid"):
  - Recursos de computación distribuidos, fiables y accesibles desde cualquier punto.
  - Interacción sencilla, sin que el usuario perciba la complejidad de la infraestructura.
  - Organiza un acceso eficiente a los datos.
  - Ejecuta el trabajo e informa al usuario.

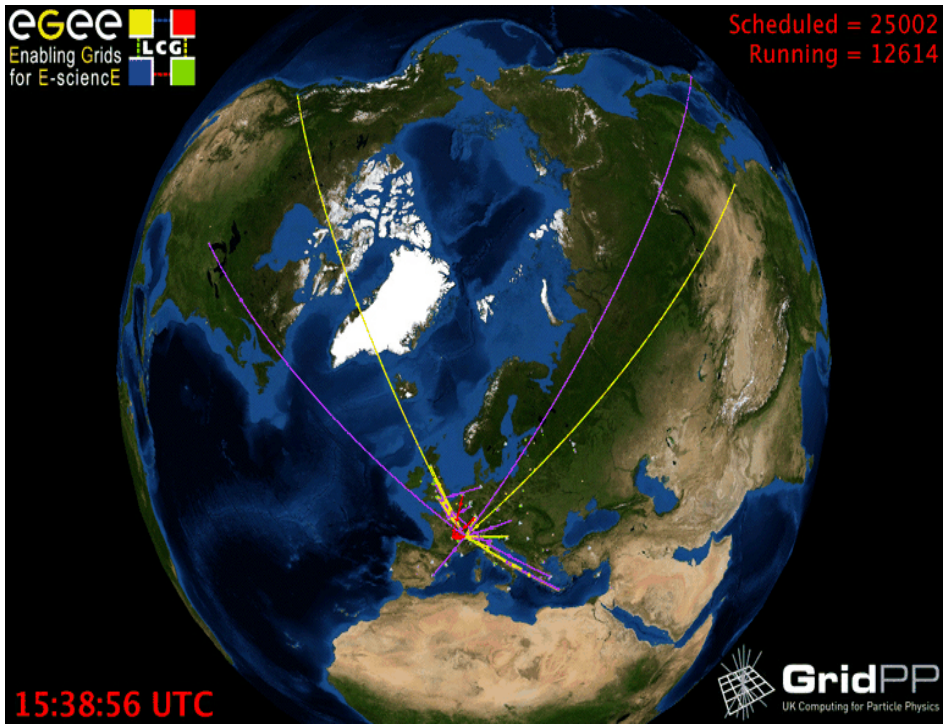
- El GRID no es:
  - Un *cluster* de PCs
  - Una mejora de Internet
  - Un proyecto, sino un conjunto de tecnologías

- El GRID intenta resolver problemas actuales de la Sociedad de la Información:
  - Acceso rápido a bases de datos/almacenamiento
  - Proporcionar su procesado y análisis utilizando potencia de cálculo distribuida y potentes facilidades de visualización
  - Mediante la utilización de la red

- El CERN y los grupos españoles participan activamente en dos proyectos
  - EGEE: Enabling Grid for E-Sciencie in Europe
  - WLCG: Worldwide LHC Computing Grid

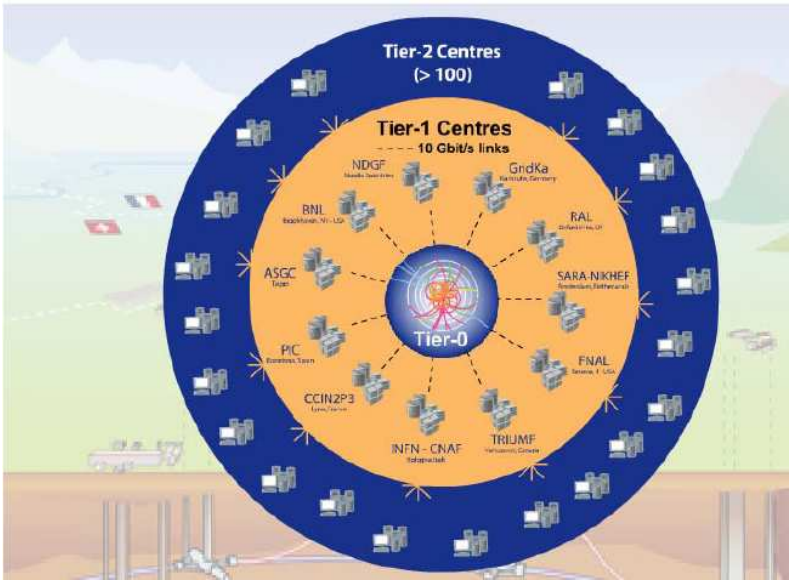
Las tecnologías Grid

- EGEE: Las tecnologías Grid puestas al servicio de cualquier campo de la ciencia. Entre ellos las Altas Energías

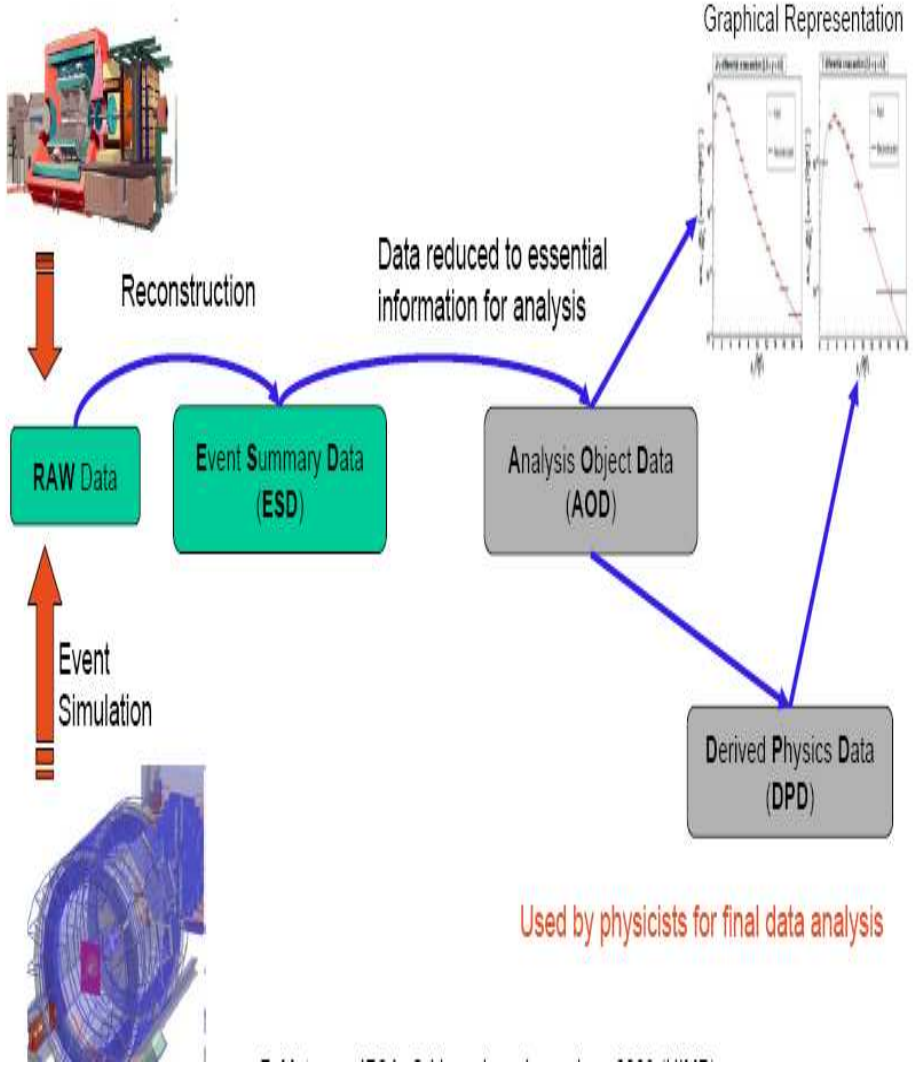


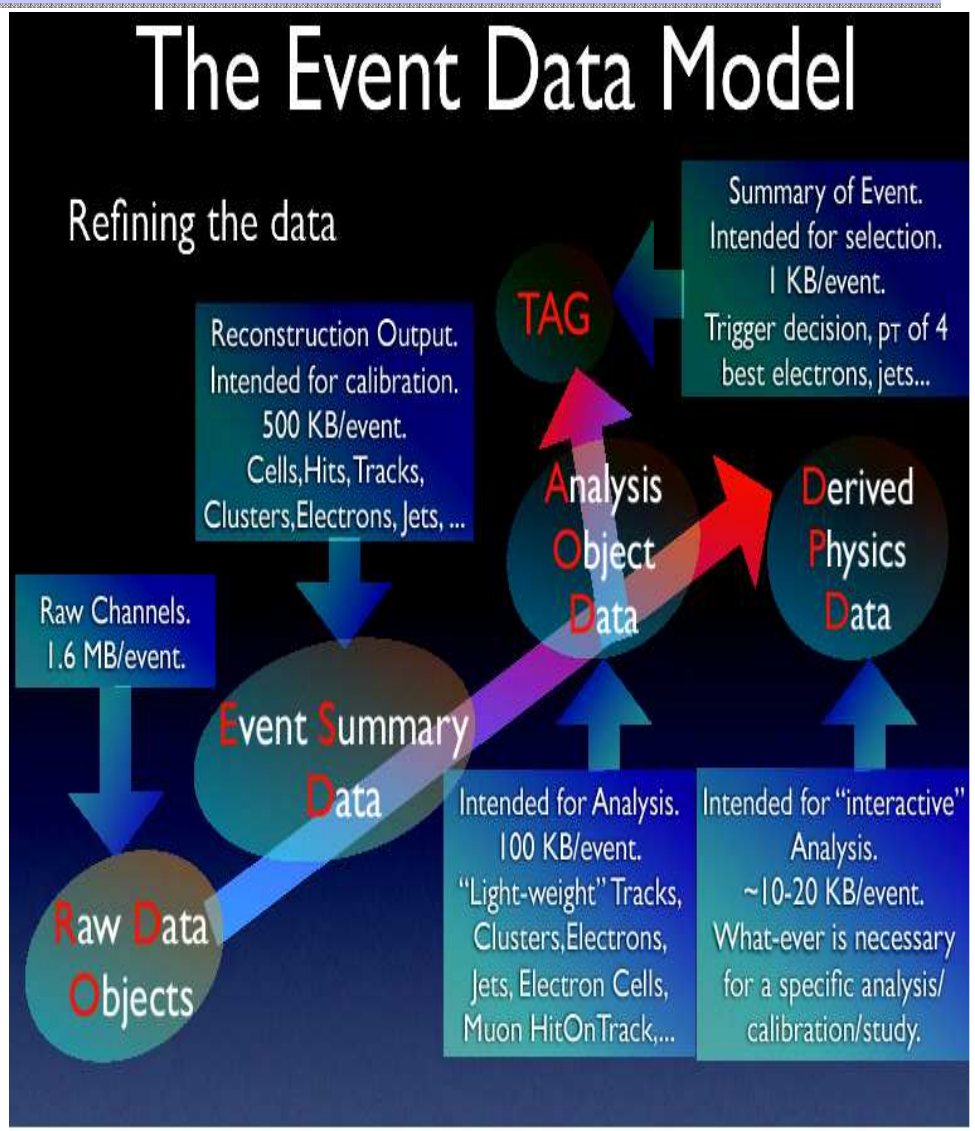
- La mayor infraestructura utilizando las tecnologías Grid y de carácter multidisciplinar en el mundo.
- 240 instituciones: alrededor de 300 centros en 50 países.
- 10000 usuarios accediendo a unas 80000 CPU.
- Más de 100000 trabajos ejecutándose al mismo tiempo.

Las tecnologías Grid

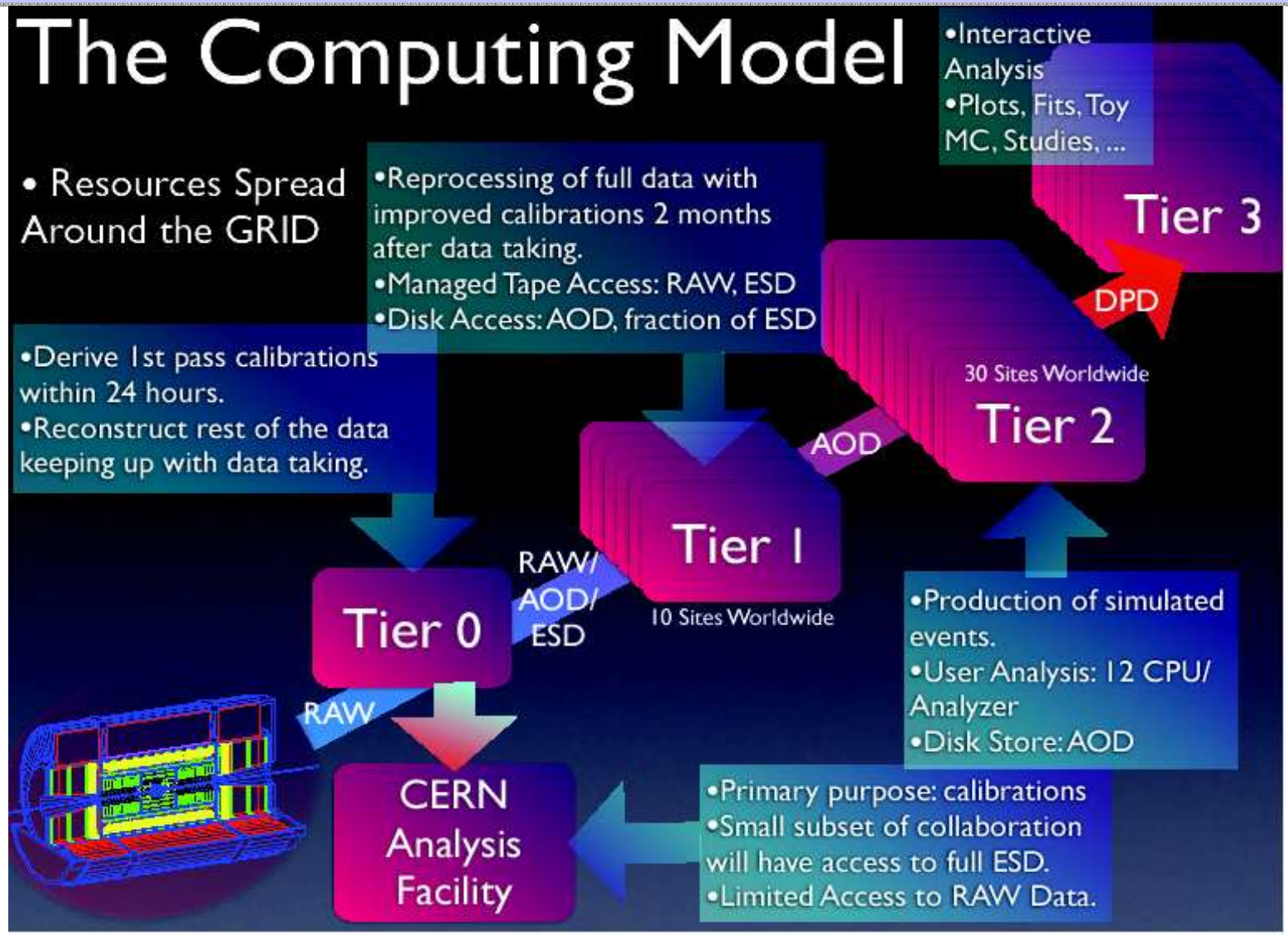
- WLCG
  - Estaba claro que todos los recursos no iban a estar en un centro, ya que este sería demasiado grande y complejo de gestionar.
  - La alternativa consistió en pequeños centros (pero aún así grandes) distribuidos alrededor del mundo y conectados entre si a través de las tecnologías Grid.
  - EL grupo Monarc del CERN propuso una estructura con diferentes tipos de centros: **Tier**
    - Un Tier-0 en el CERN
    - Unos pocos centros grandes (5-10/experimento) en diferentes países: Tier-1
    - Tier-2 (20-25/experimento)
    - Tier3 (pequeñas granjas locales cercanas a los grupos de física)
  - Enviar los trabajos al centro/sitio adecuados
  - Optimizar los recursos alrededor del mundo
  - Optimizar el acceso d los datos
  - Autentificación y autorización
  - Monitorización de la ejecución de los trabajos: Estado y resultado
  - 

Modelo de datos

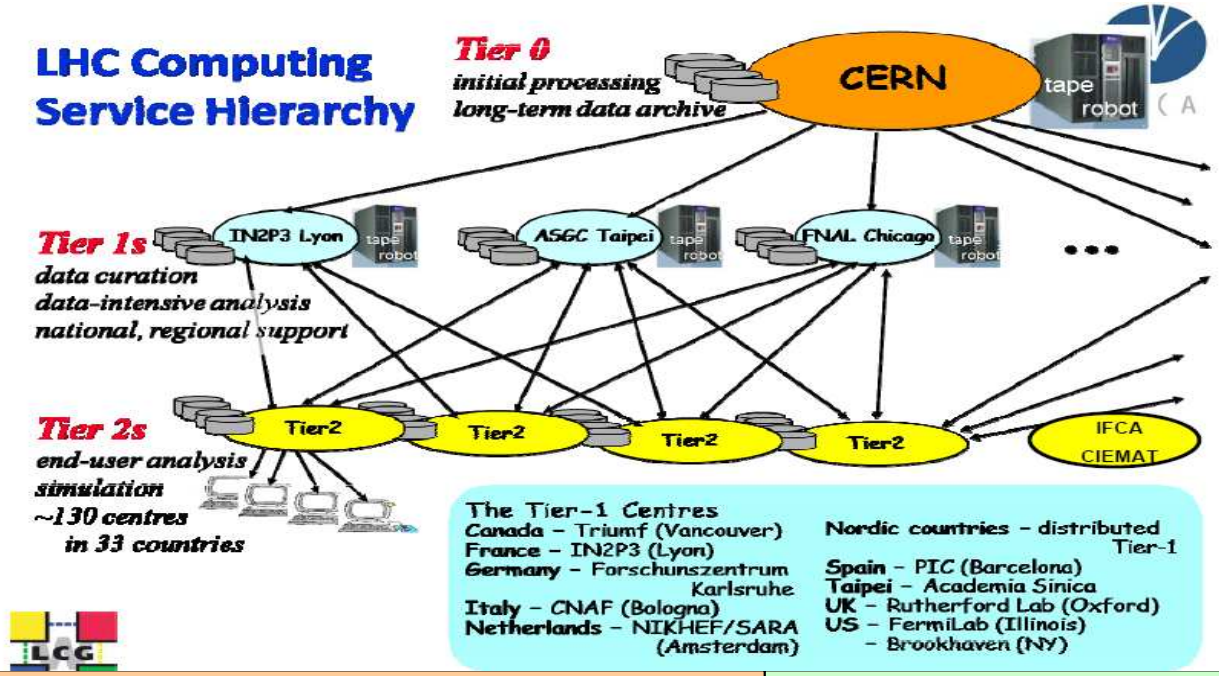




Modelo de Computación para ATLAS



Modelo de Computación para LHC



- Tier-2s:
  - Corre el análisis
  - Producción de la simulación por Montecarlo
  - Importa *datasets* desde cualquier Tier-1 y exporta los datos de Montecarlo

- Tier-0:
  - Lee los datos directamente DAQ
  - Primer procesado de los datos
  - Almacena todos los datos
  - Los distribuye a los Tier-1s
  - Corre la calibración

- Tier-1s:
  - Almacenamiento de los datos procesados
  - Re-procesado de los datos
  - Corre filtros sobre los datos para su distribución
  - Almacena los datos de simulación
  - Distribuye los datos a los Tier-2s

Recursos Españoles

- Tier-1 está en el PIC (*Port d'Informarció científica*) Barcelona
- Cada Tier1 se agrupa con un conjunto de Tier-2s formando lo que se conoce como "*cloud*".
- El Tier-1 español ofrece:
  - Almacenamiento y recursos para el procesado de datos de tres experimentos del LHC: ATLAS, CMS y LHCb.
  - Los experimentos del LHC almacenarán una copia de los datos en el CERN y distribuirán otra copia a los diferentes Tier-1s.
  - ~10% de los "*raw data*" estarán almacenados en el PIC.
  - *Optical Private Network* (OPN) entre el Tier-0 y los Tier-1s
  - Más de 9 Petabytes de datos in/out en 2008 en el PIC.

1 KSI2K = 1 K SPECint 2000 (CPU)

SPECint is a computer benchmark specification for CPU's integer processing power. It is maintained by the Standard Performance Evaluation Corporation (SPEC). SPECint is the integer performance testing component of the SPEC test suite. The first SPEC test suite, CPU92, was announced in 1992. It was followed by CPU95, CPU2000, and CPU2006. The latest standard of SPECint is CINT2006 (aka SPECint2006).

		2007	2008	2009	2010	2011	2012	2013
CPU (kSI2K) required	ATLAS	172	865	1226	1960	2687	3417	4872
	CMS	289	477	1058	2516	3292	4099	6201
	LHCb	37	167	307	633	962	1215	1263
	TOTAL	498	1509	2591	5109	6941	8731	12336
Disk (Tbytes) required	ATLAS	114	512	902	1595	2168	2743	4176
	CMS	79	358	630	1113	1513	1915	2915
	LHCb	21	97	170	301	409	518	788
	TOTAL	214	967	1702	3009	4090	5176	7880
Tape (Tbytes) required	ATLAS	68	385	681	1182	1767	2439	2819
	CMS	140	487	974	1677	2519	3358	5186
	LHCb	18	81	189	543	963	1456	2981
	TOTAL	226	953	1844	3402	5249	7253	10986

Tier-2 español para CMS

CIEMAT
- ~1000 KSI2k
- 500 slots
- storage: 220 TB

IFCA
- 750-1500 KSI2k
- 300 slots dedicated to CMS
- additional opportunistic usage of local resources (+400 and more on peaks)
- storage: 160 TB

Total

- CPU:
  - ~1750-2500 KSI2k
- Storage
  - 380 TB

ramping up ~linearly to cope with the expected increase of needs

Tier-2 español para ATLAS

Año	2006	2007	2008	2009	2010	2011	2012
CPU(KSI2k)	925	2336.11	17494.51	26972.76	51544.64	69128.42	86712.2
Disk (TB)	289	1259.04	7744.37	13112.04	22132.3	31091.45	40050.92

Todos los Tier-2s de ATLAS Tier-2

Year	2006	2007	2008	2009	2010	2011	2012
CPU(KSI2k)	46	117	875	1349	2577	3456	4336
Disk (TB)	14	63	387	656	1107	1555	2003

ATLAS Tier-2 español contribución 5%

Worker Nodes

Tape Storage Robot

PC Farm (former one)

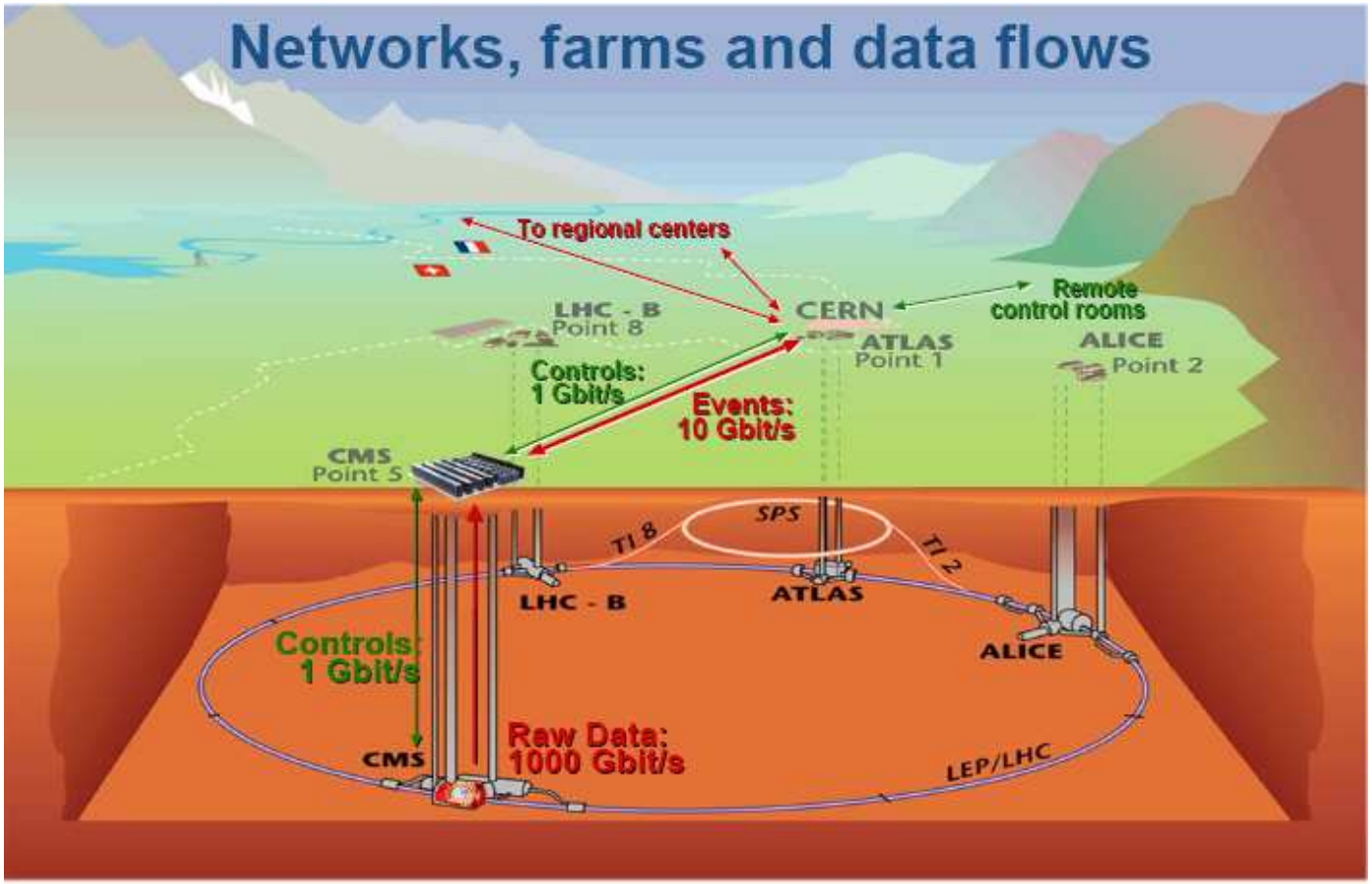
Human Resources:

13'5 FTE

New Worker Nodes
+
Disk servers

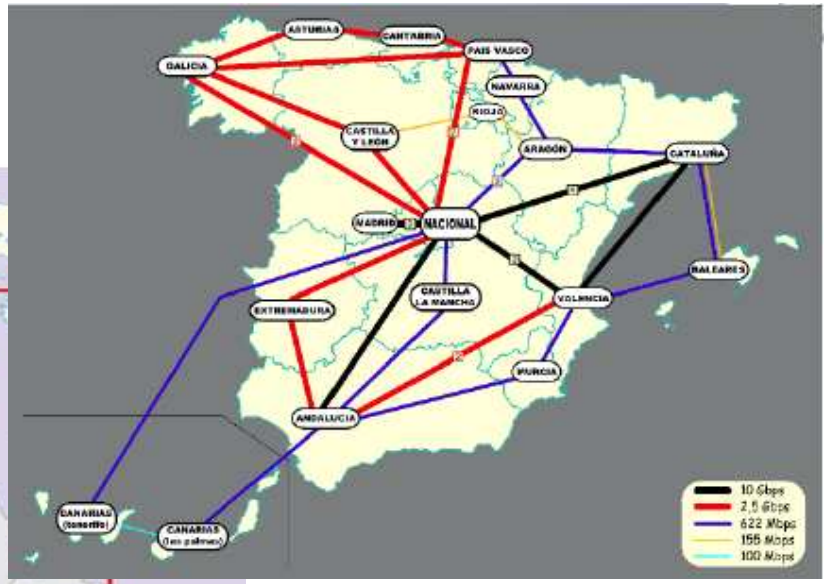
La Red

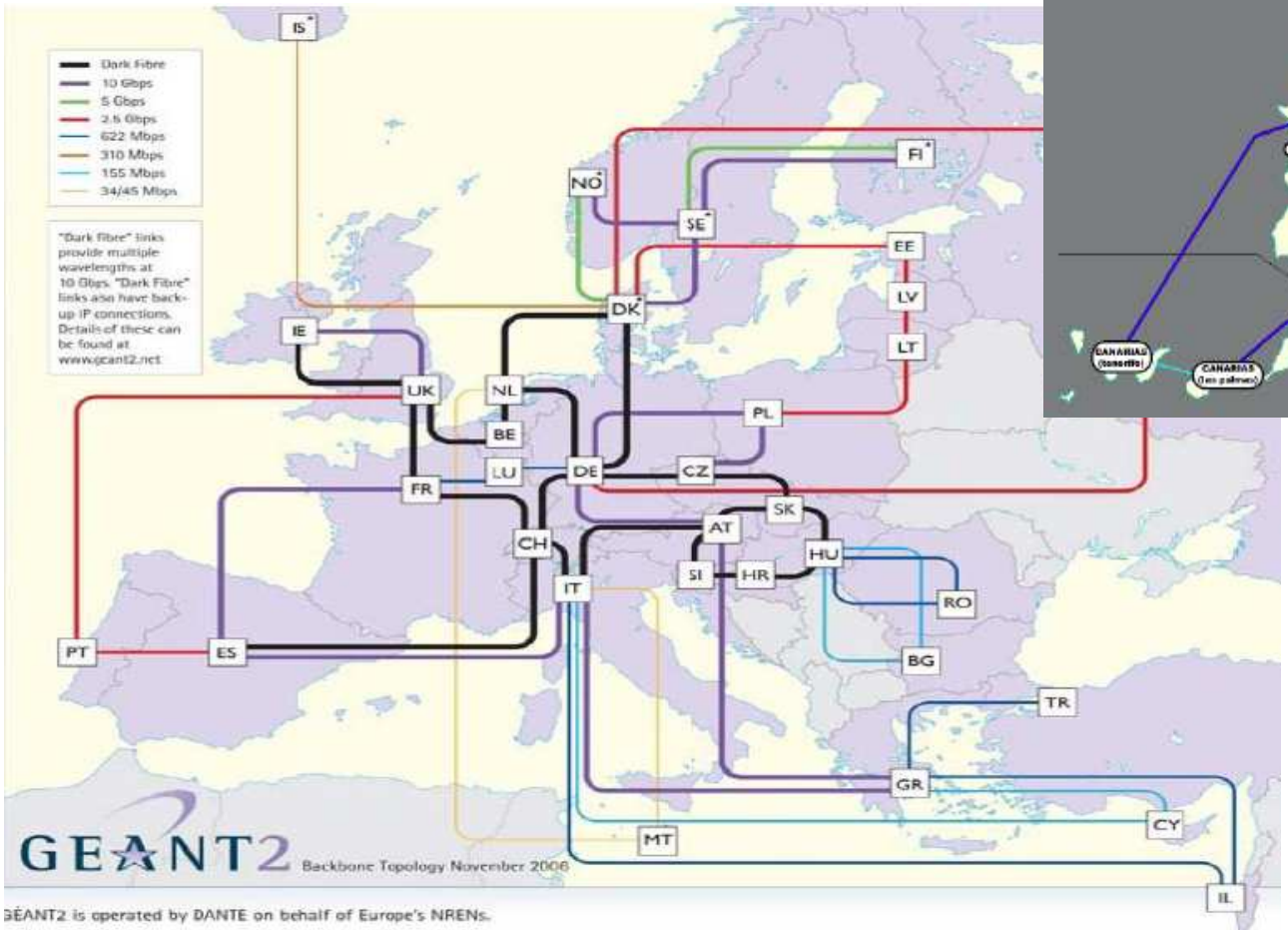
- Alta demanda de la red (LAN/WAN) por el flujo de datos



La Red

- Géant2/Rediris





Uso actual (real): Test Step09

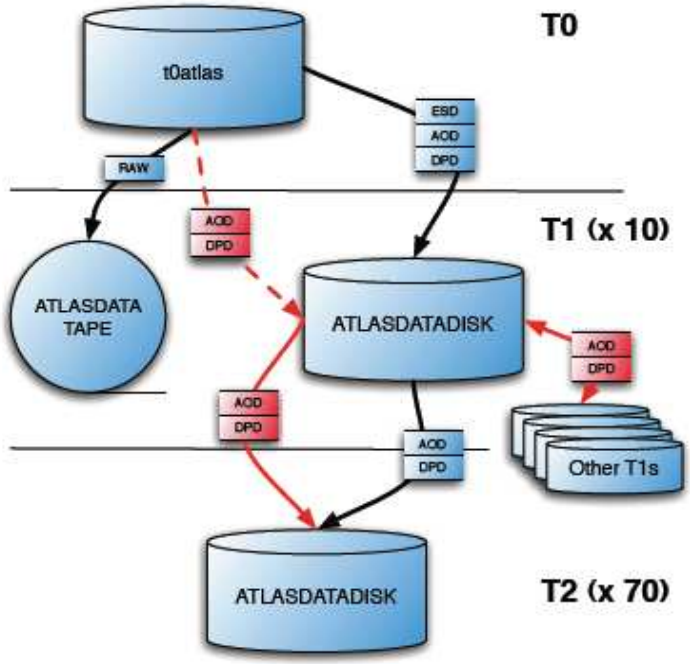
- **Primer test de uso real, es decir trabajos de simulación montecarlo y análisis corriendo a la vez!!!!**
- **Fechas**
  - Mayo 25 - 31 Setup
  - **Junio 2 - 14 Corriendo**
- **Actividad de los Tier-2s**
  - Los Tier-2s participaron produciendo los datos simulados de Montecarlo y corriendo trabajos de análisis
    - 50% simulación, 50% análisis.
  - Almacenamiento de los datos necesarios para el análisis
- **Distribución de datos**
  - El volumen de datos distribuidos a los Tier-2s fue de 11TB 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.

https://twiki.cern.ch/twiki/bin/view/LCG/WLCGStep09

Distribución de datos

- CMS
  - PhEDEx (*Physics Experiment Data Export*)
- 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.

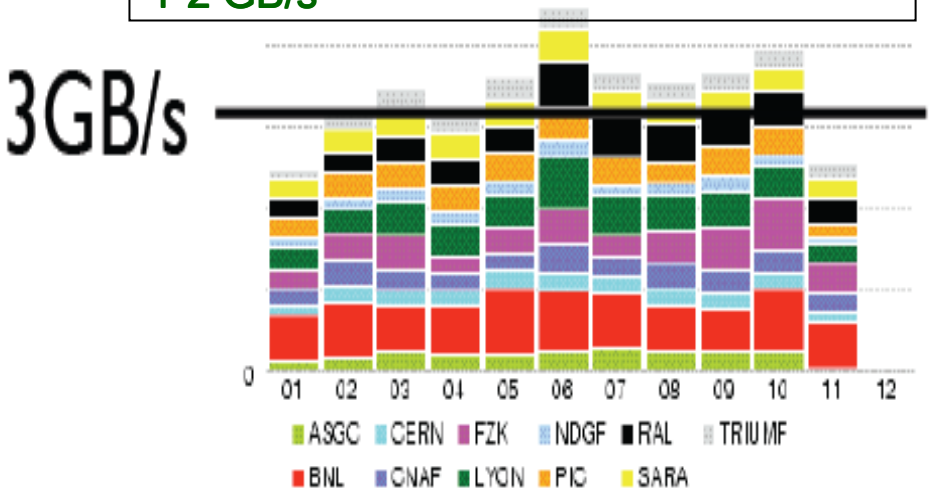
- Data taking and first reconstruction passes
  - RAW and ESD from CERN → distributed to T1 sites (1, 2 copies respectively, RAW to tape)
  - AOD and DPD from CERN distributed to all T1 sites (10 copies)
  - AOD and DPD from CERN distributed to T2 from their parent T1 (1 to 2.7 copies per cloud)
- Reprocessing 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 - - - →

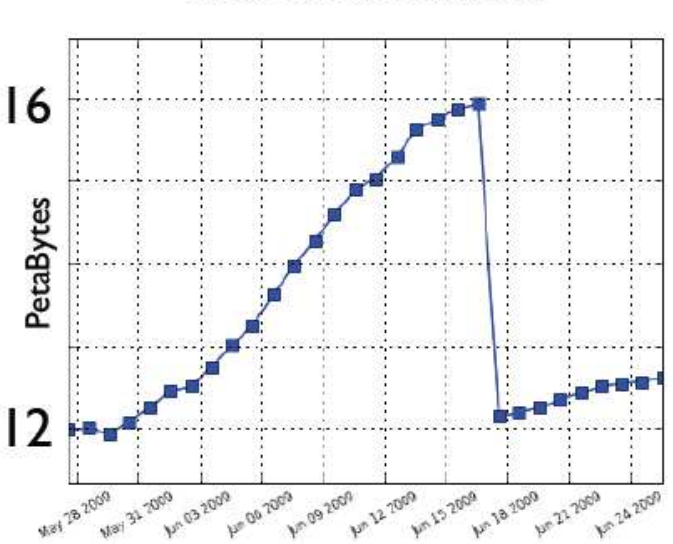


Resultados distribución de datos

- **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

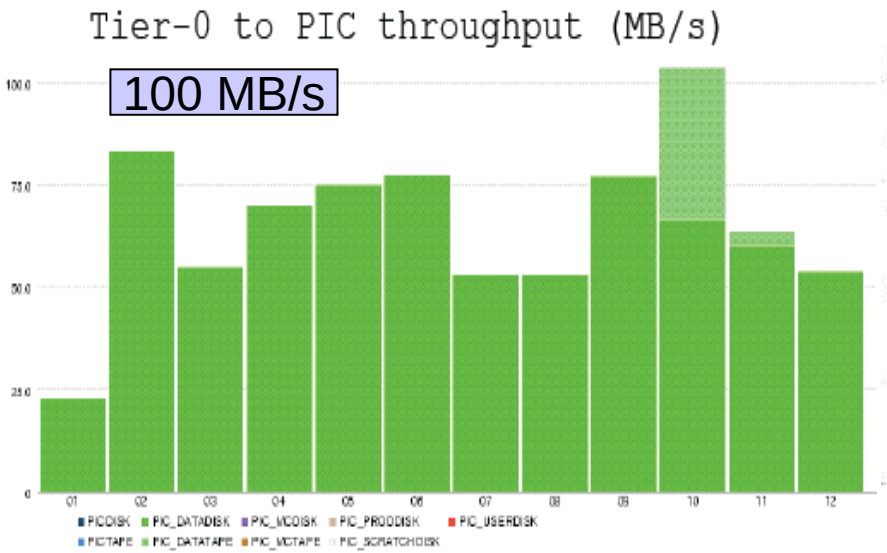


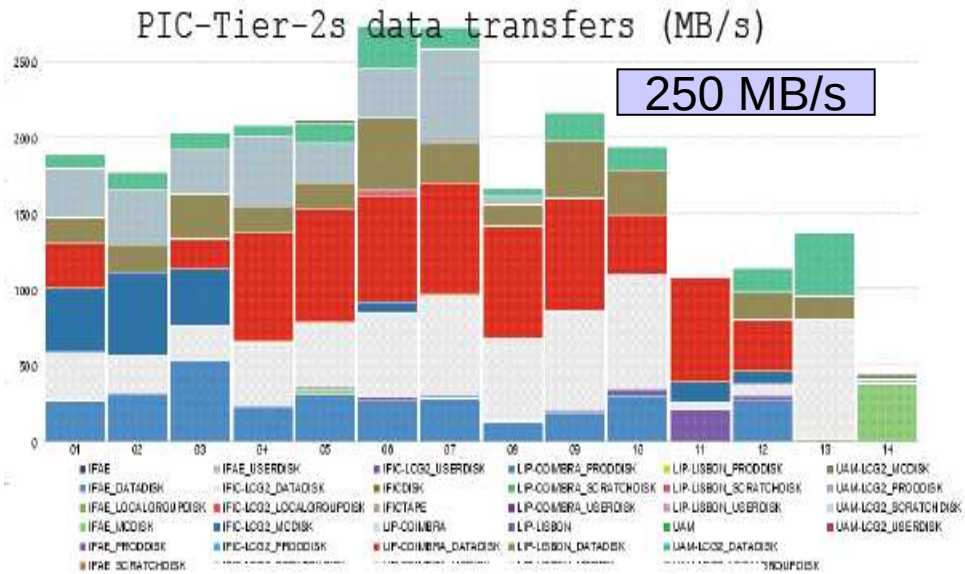


		Transfers
Cloud	Efficiency	Throughput
ASGC	99%	397 MB/s
BNL	84%	1128 MB/s
CERN	100%	334 MB/s
CNAF	98%	561 MB/s
FZK	85%	556 MB/s
LYON	96%	620 MB/s
NDGF	84%	137 MB/s
PIC	93%	429 MB/s
RAL	99%	838 MB/s
SARA	53%	262 MB/s
TRIUMF	100%	297 MB/s

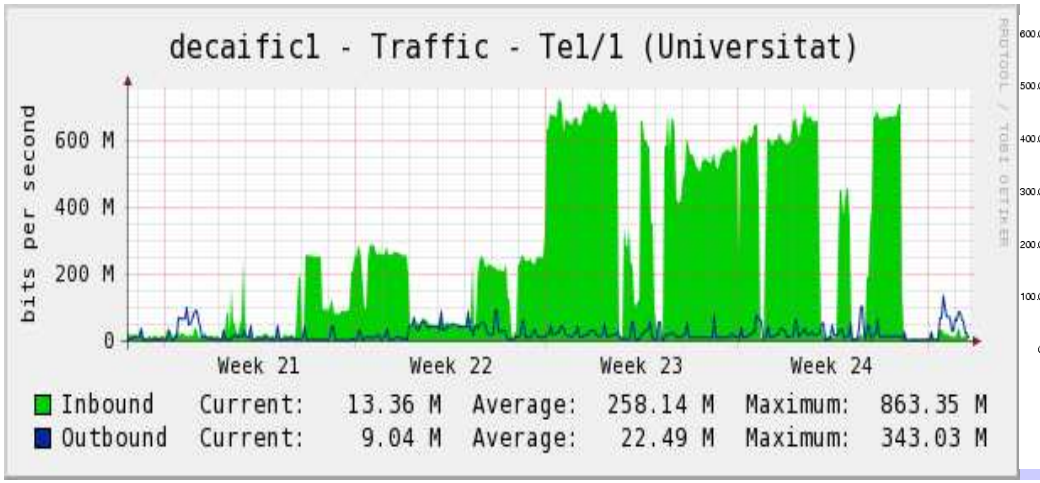
Peaks of 5.5GB/s

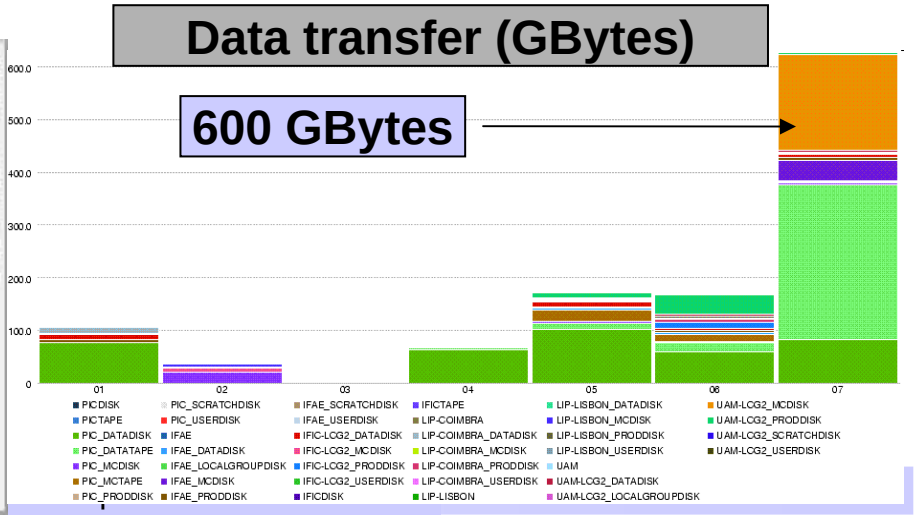
Resultados distribución de datos en los centros españoles





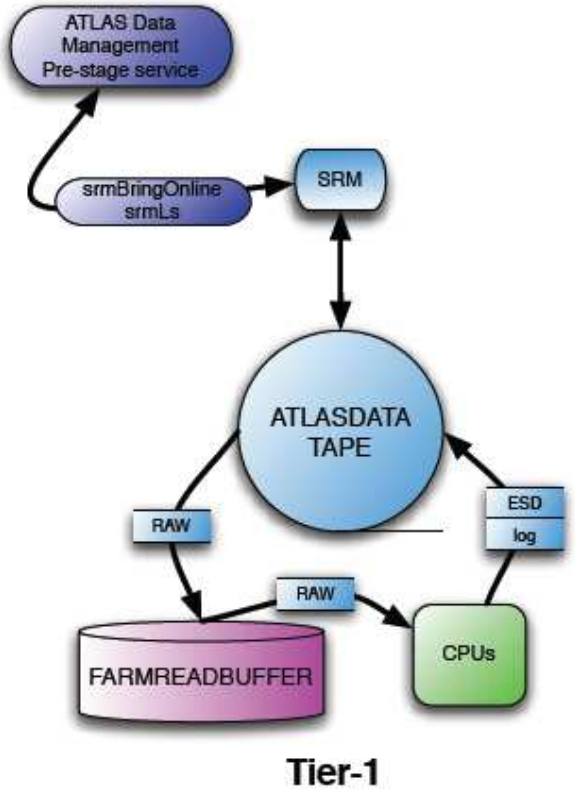
- **Picos de 600Mbps observados en el IFIC**





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)**

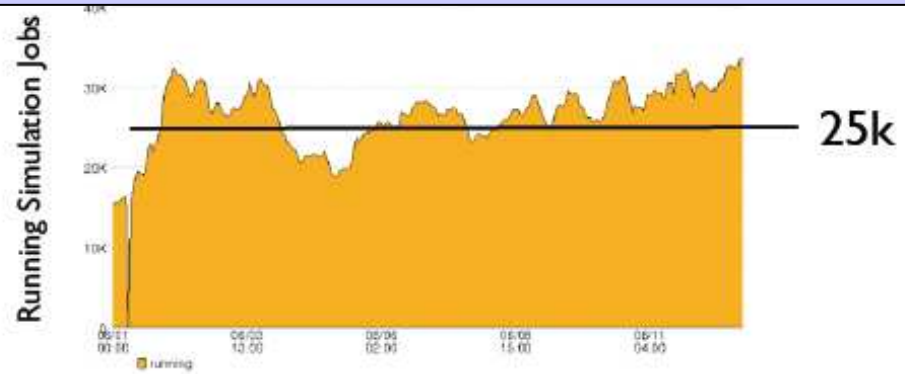


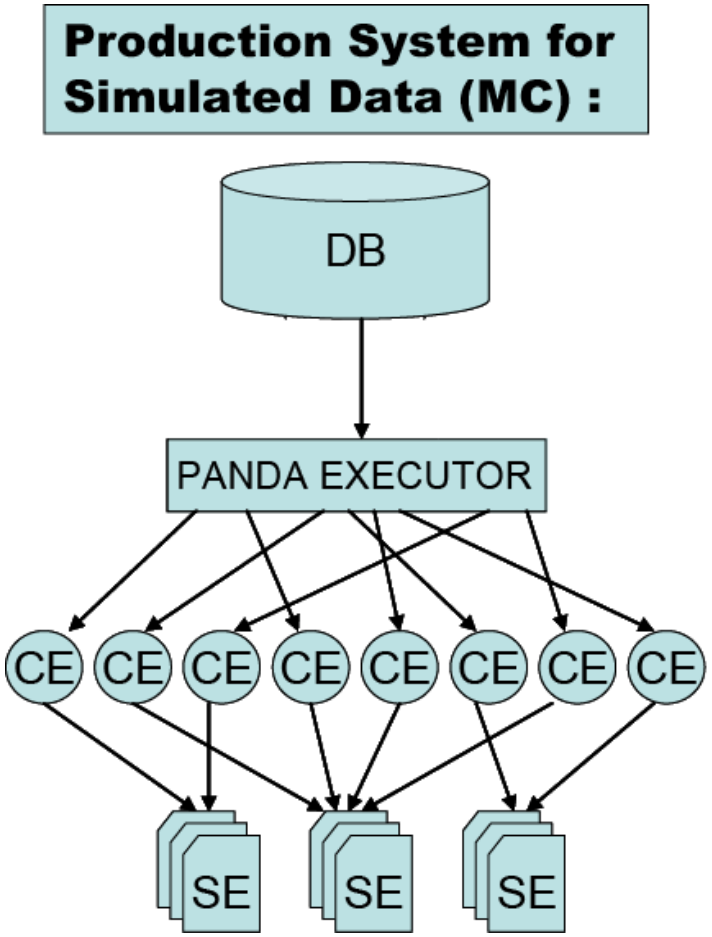
T1	Base Target	Result	Comment
ASGC	10 000	4 782	Many batch system and basic setup problems
BNL + SLAC	50 000	99 276	Also ran high priority validation and other tasks
CNAF	10 000	29 997 ★	
FZK	20 000	17 954	Big tape system problems pre-STEP; no CMS
LYON	30 000	29 187	Very late start due to tape system 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

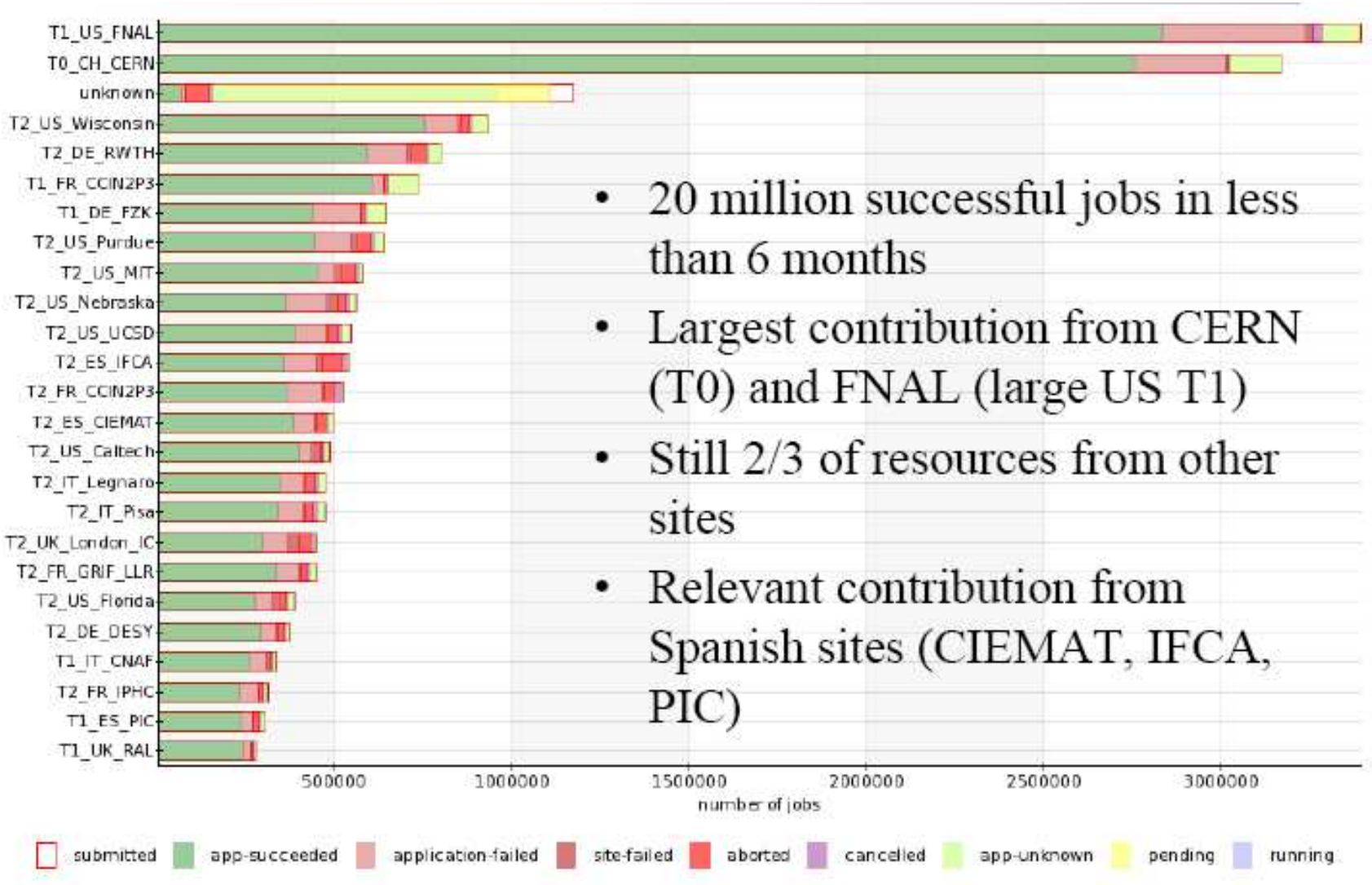
- Cada experimento tiene un sistema automático de producción de trabajos de simulación.
  - En continuo desarrollo y validación desde hace muchos años (2002)
  - Diferentes tests llevados a cabo "*challenges*" para ver como funciona el sistema cuando enviamos millones de trabajos de simulación.
    - Data Challenge, Service Challenge Computing, STEP09, etc..

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



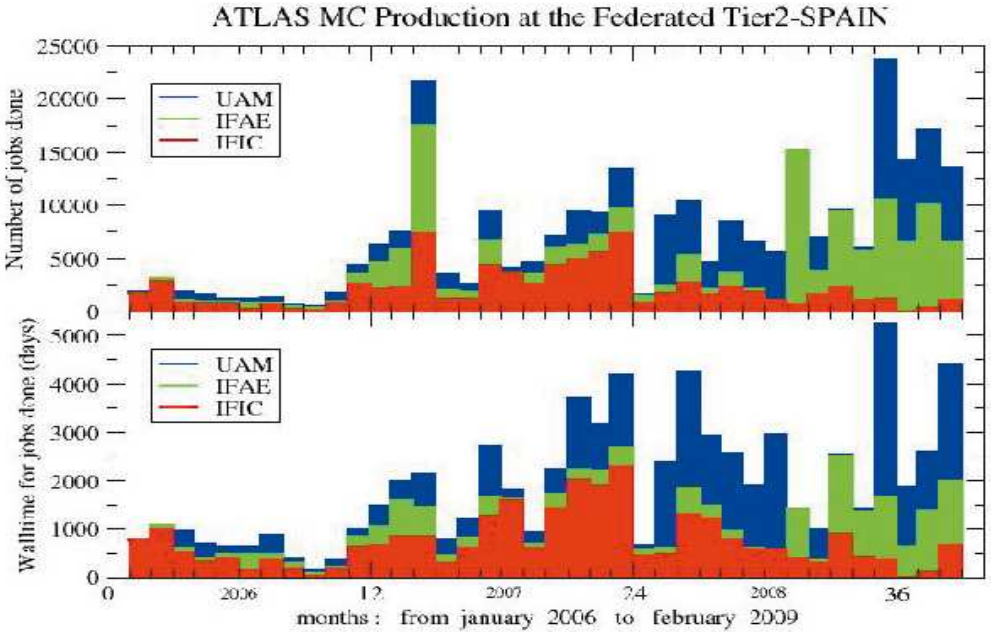


Número de trabajos Grid en 2009 (CMS)



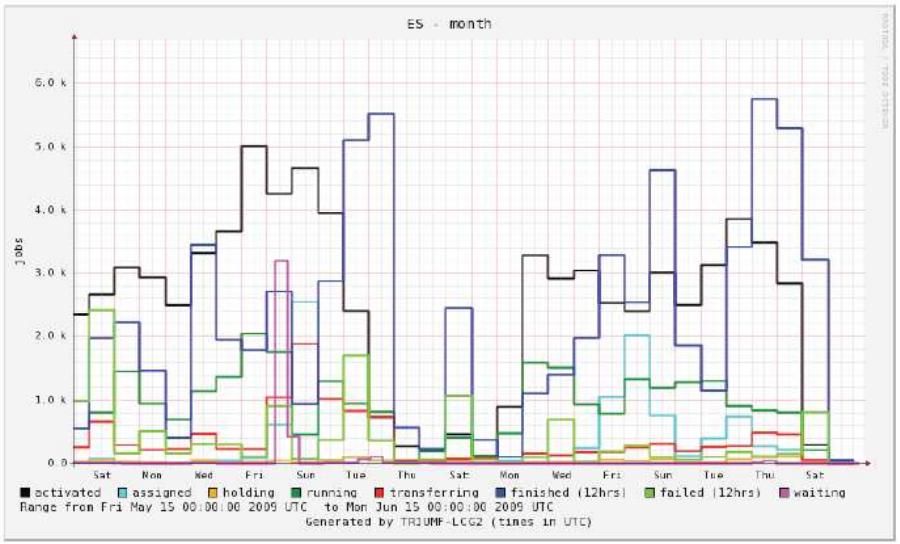
- 20 million successful jobs in less than 6 months
- Largest contribution from CERN (T0) and FNAL (large US T1)
- Still 2/3 of resources from other sites
- Relevant contribution from Spanish sites (CIEMAT, IFCA, PIC)

Producción de MC en el Tier-2 español de ATLAS



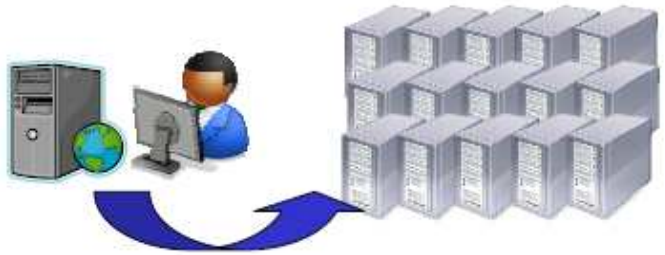
- **Desde Enero 2006:**
  - **La eficiencia del Tier-2 español es bastante estable y está alrededor del 95% en los últimos meses.**
  - **La contribución española la producción total es del 2.5%.**

- **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%**

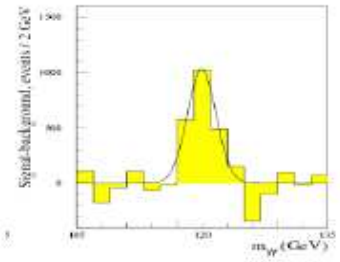


Análisis en el LHC

- Test, mejora algoritmo, etc..: Análisis interactivo sobre pequeñas muestras de sucesos corriendo en el ordenador o granja local.

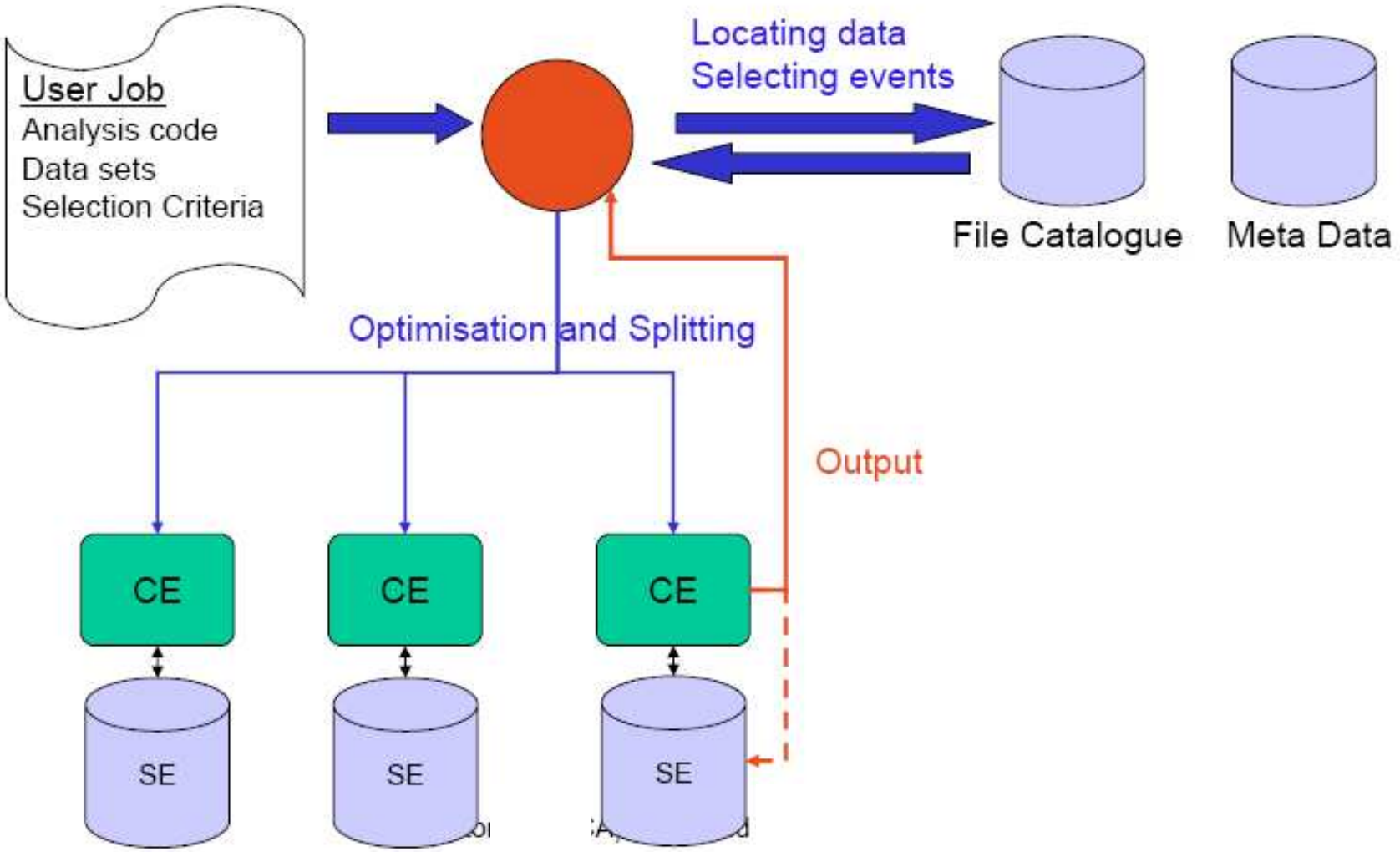


- Producción de AOD a partir de ESD, correr sobre grandes muestras de sucesos, datos distribuidos a lo largo del mundo, etc..Análisis distribuido



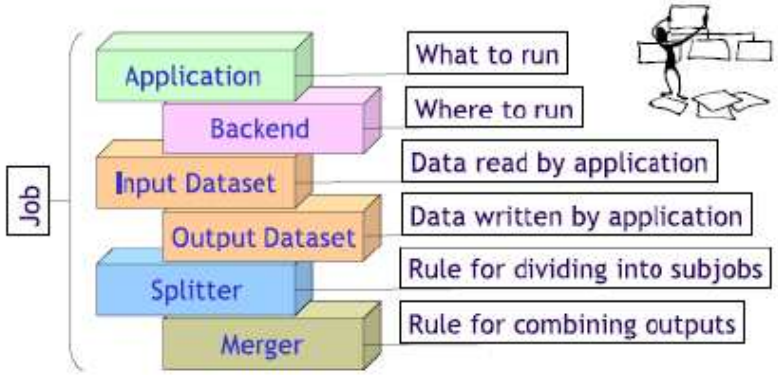
- La salida enviada al usuario para ser analizada de forma interactiva y producir el "*plot*" final

Análisis distribuido

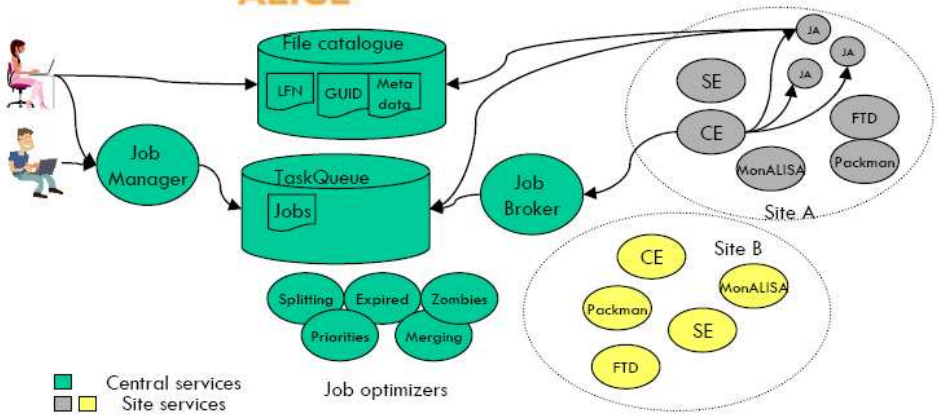


Análisis distribuido

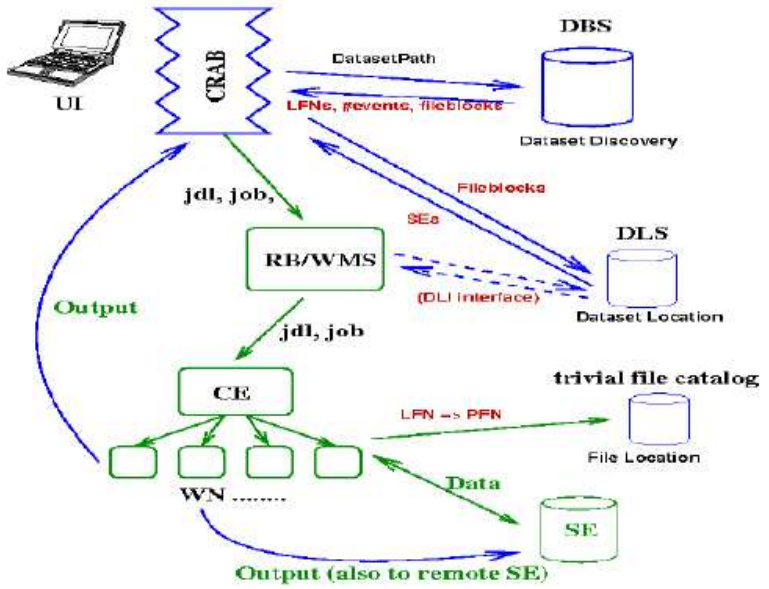
ATLAS LHCb GANGA



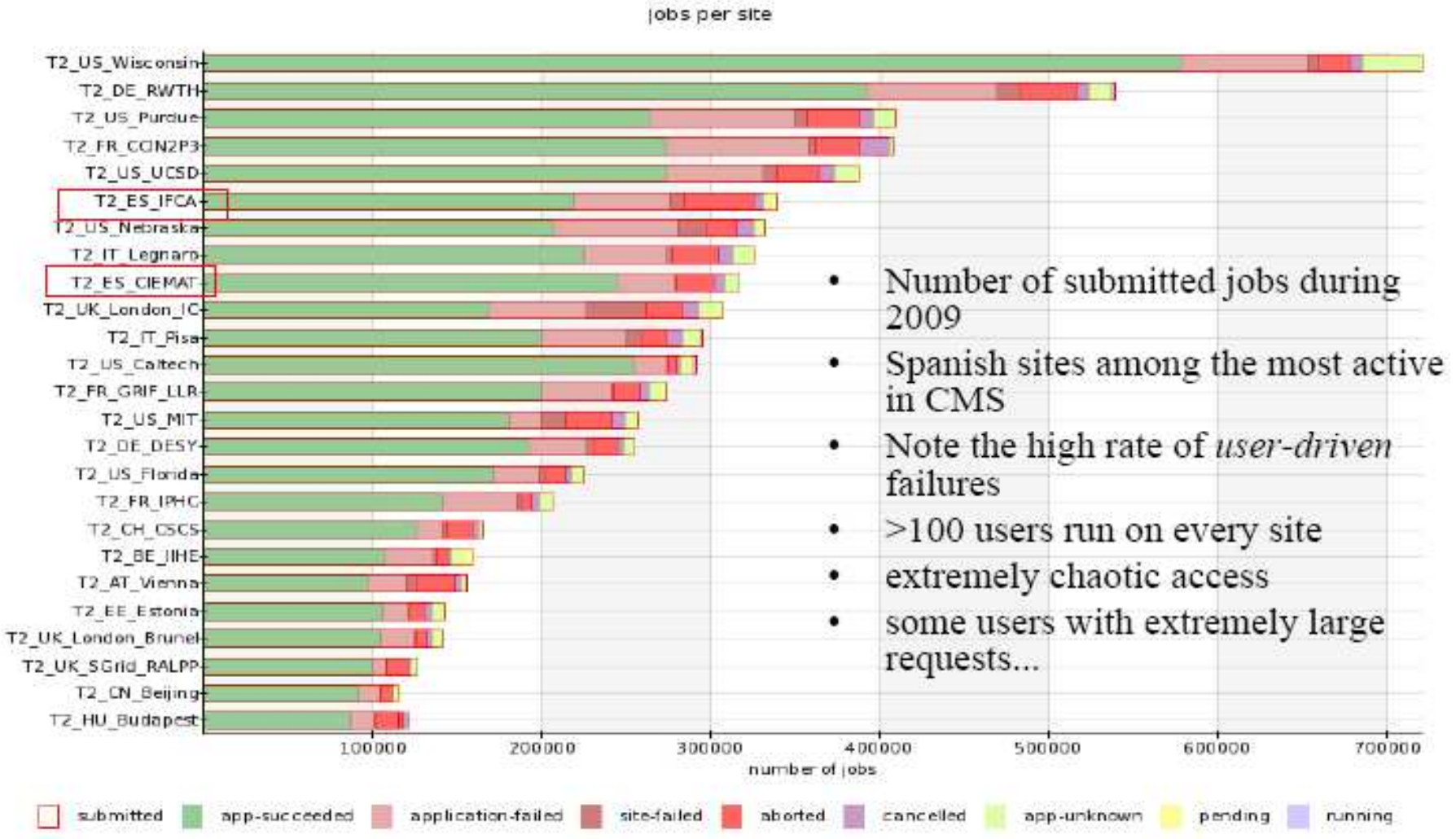
ALICE AliEn²



CMS CRAB



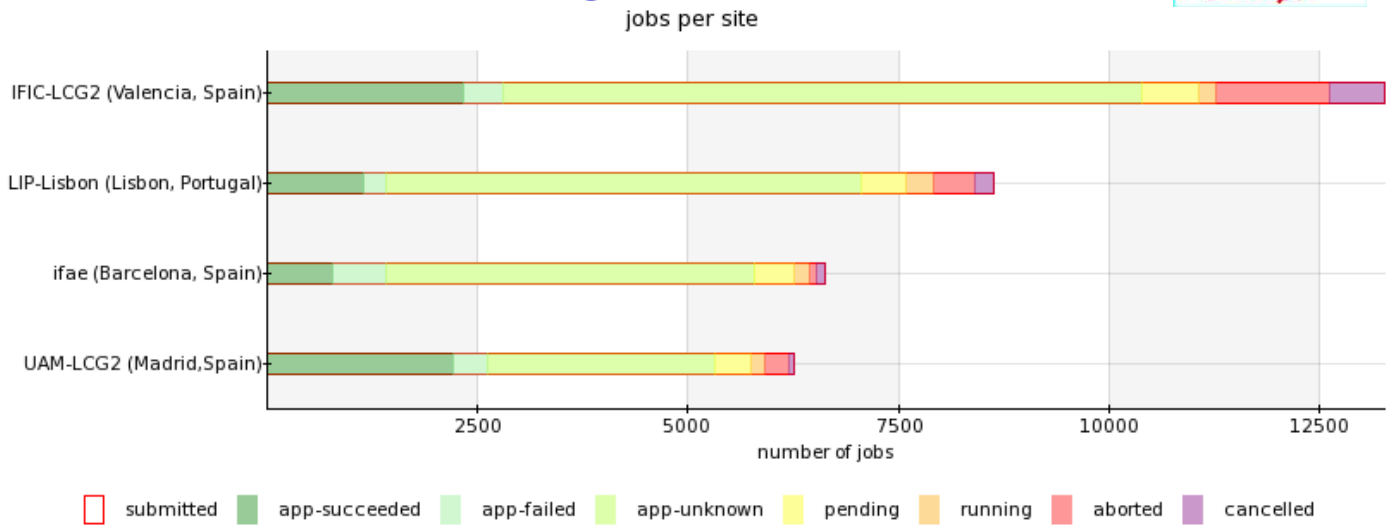
Trabajos de Análisis en 2009 para CMS

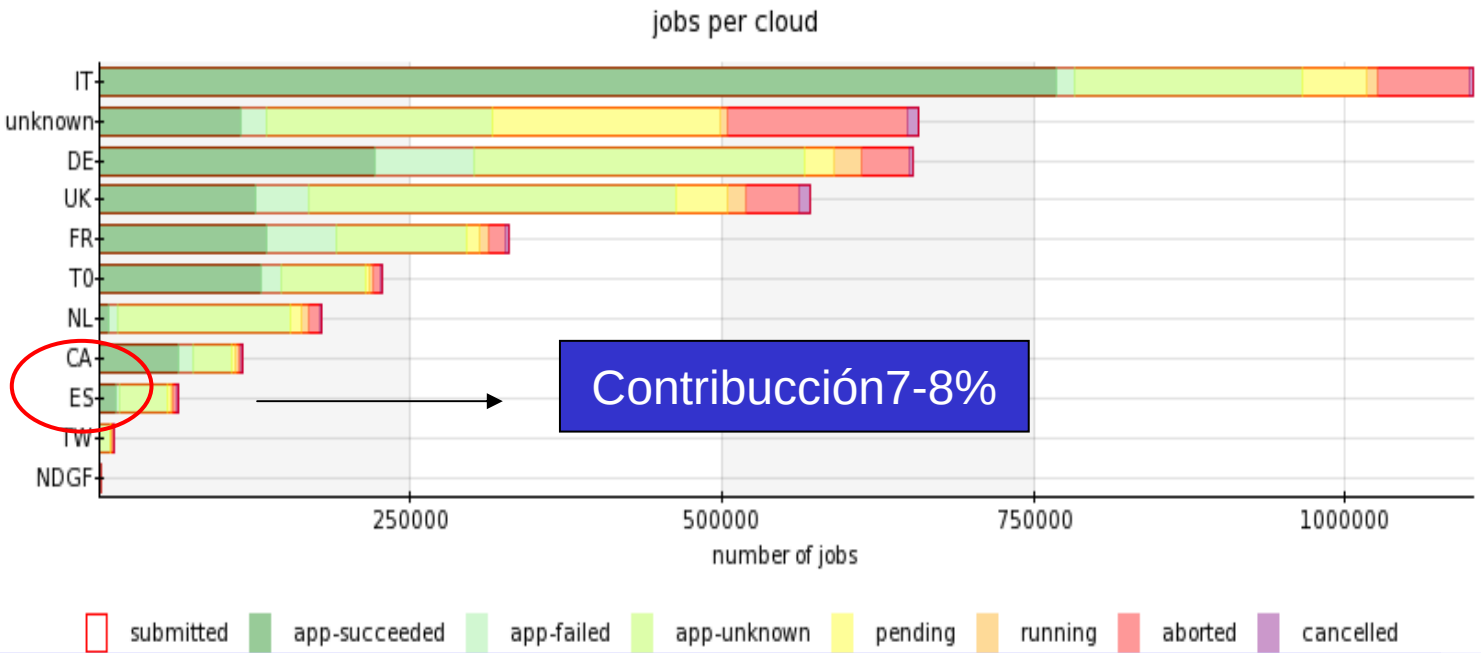


- Number of submitted jobs during 2009
- Spanish sites among the most active in CMS
- Note the high rate of *user-driven* failures
- >100 users run on every site
- extremely chaotic access
- some users with extremely large requests...

Trabajos de Análisis en 2009 para ATLAS

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



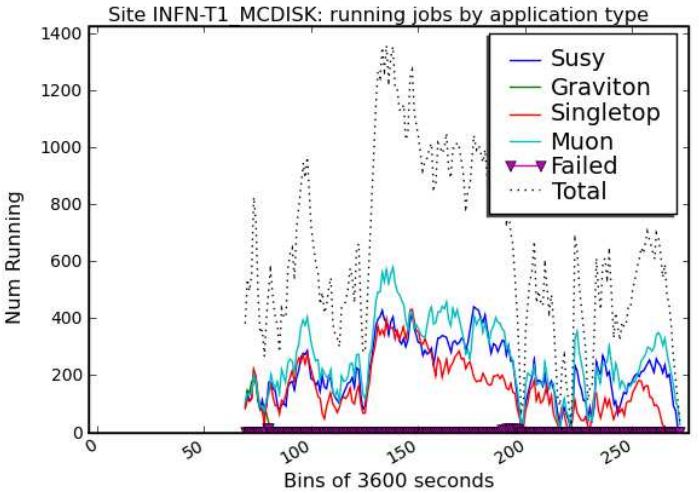


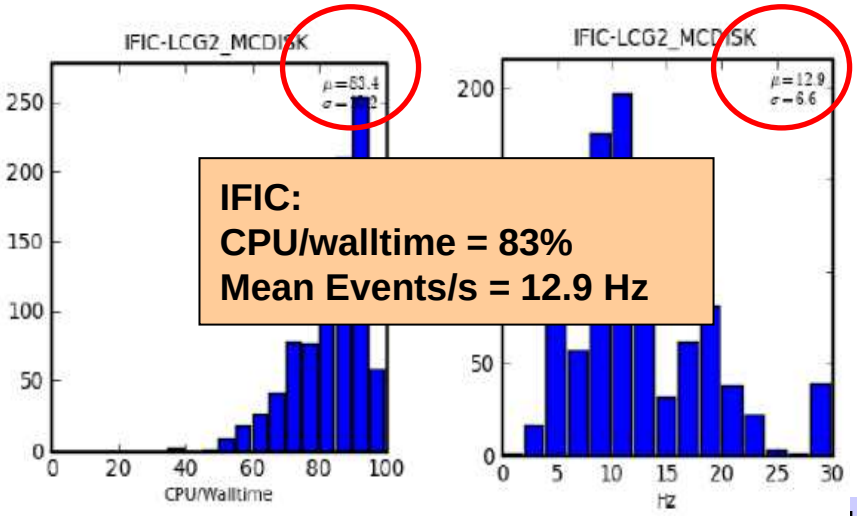
Contribucción7-8%

Por "clouds"
Enero-Julio 09

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

Recursos de Computación para el 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

Conclusiones

- Se ha visto que la infraestructura a nivel de computación requerida por los experimentos de Altas Energías actuales es de un tamaño sin precedente alguno.
  - Tanto a nivel de tamaño como de complejidad
- Las solución adoptada por el CERN ante tal reto ha sido un sistema distribuido basado en las tecnologías Grid.
  - Sistema jerárquico de Tiers (Tier-0, Tier-1s, Tier-2s, Tier-3s)
- Los experimentos utilizan las herramienta actuales provistas por las tecnologías Grid
  - Aunque también hay desarrollos independientes por parte de cada experimento (adaptar tales herramientas a sus necesidades)
- 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!!!**