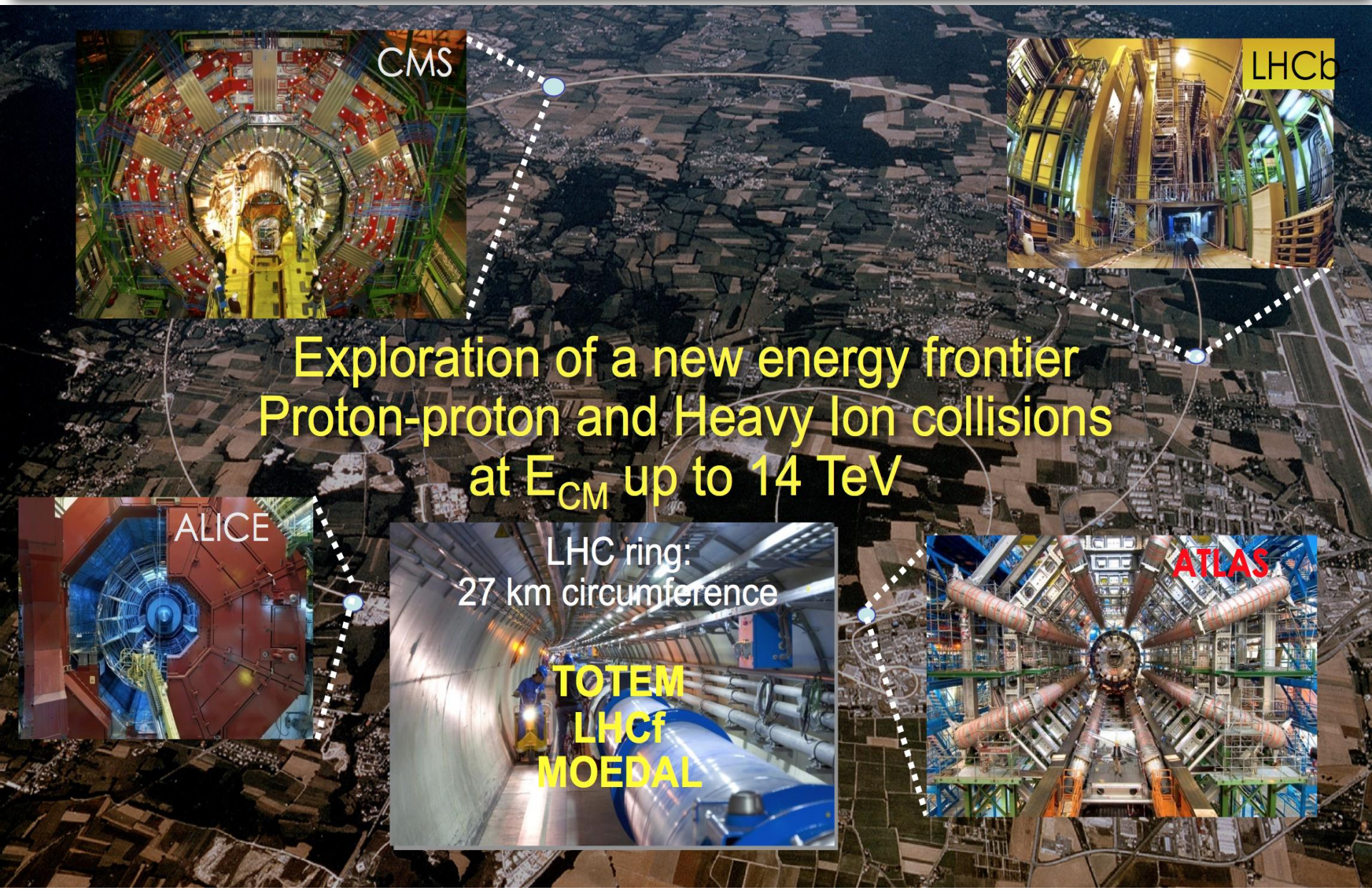
GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD
Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

Computación distribuida en el LHC y su evolución

Dr. José Flix Molina (PIC/CIEMAT)

LHC: no necesita de introducción



El desafío de datos del LHC

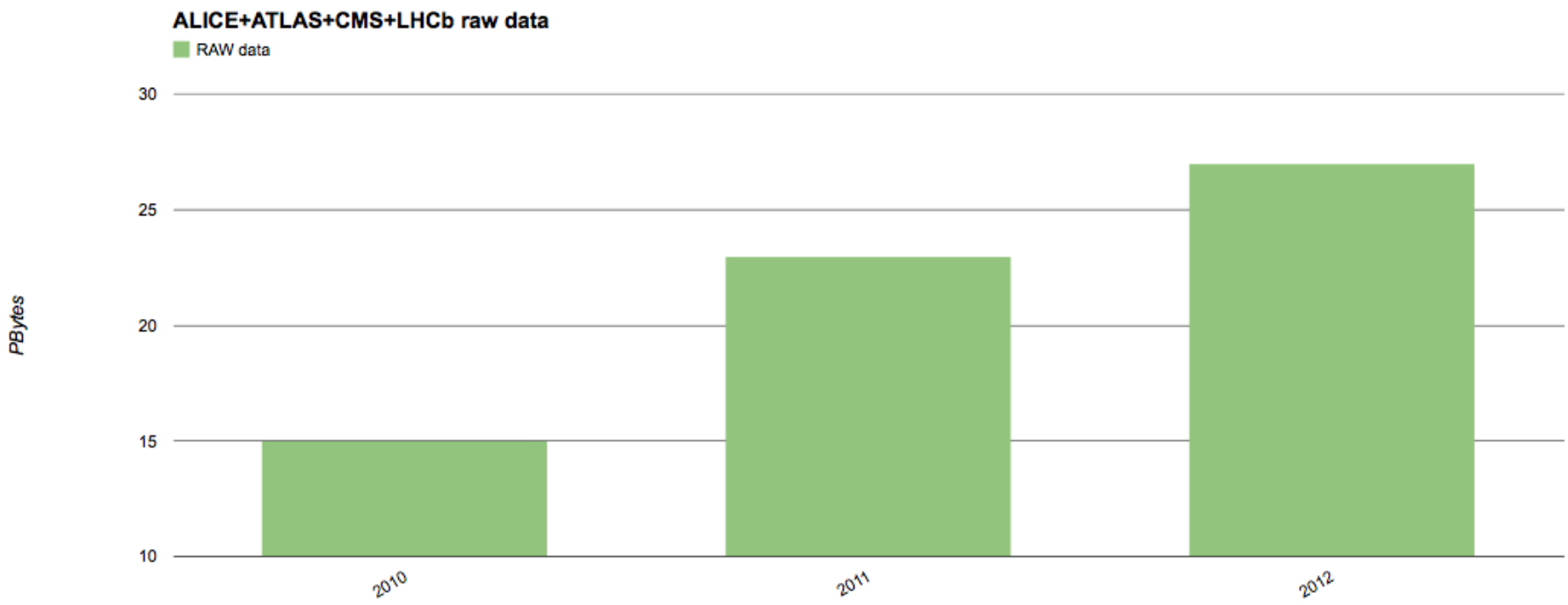
- El LHC, colisionador superconductor de hadrones @CERN, es una instalación científica **única el mundo**
- Cuatro detectores registran las colisiones *pp* o *p-ion*:
  - 20 MHz *crossing rate* → ~300 Hz *trigger* → ~ 1 GB/s → **10-15 PB/año**
  - + Datos procesados, simulados, replicas: **50 PB/año**
- 10-15 años de toma de datos → **Escala del Exabyte!**
- Se necesita una potencia de calculo sin igual → las señales buscadas ~1 suceso en cada billón (10^{12})
  - ~35 interacciones por cruce (*pile-up*) en el pico de luminosidad
  - ~1600 partículas cargadas producidas en cada colisión
- Datos analizados por <u>miles de físicos</u> repartidos por todo el mundo
- **Comparación:** Todos los datos de LEP ocupan unos pocos TBs

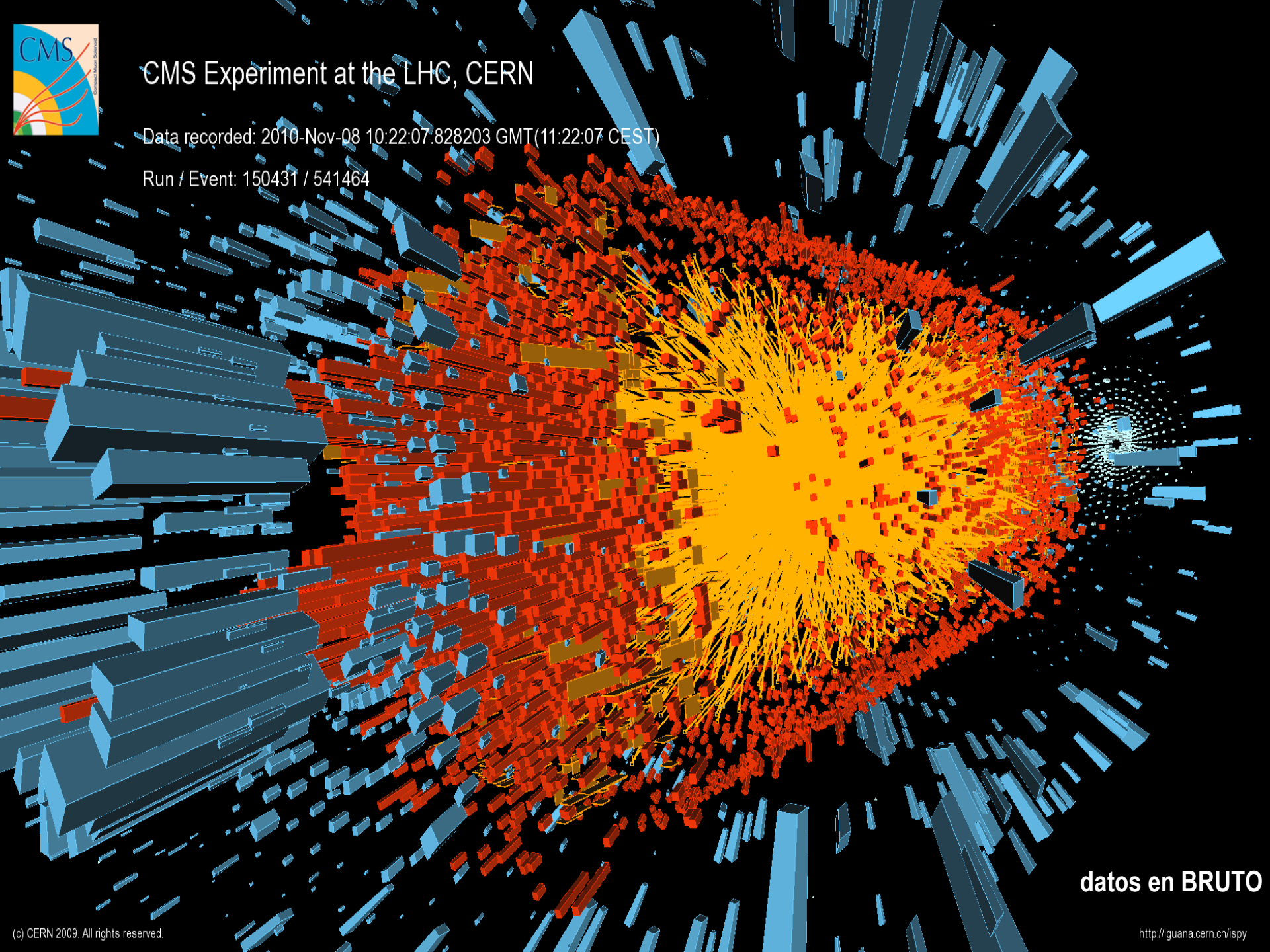
Resumen

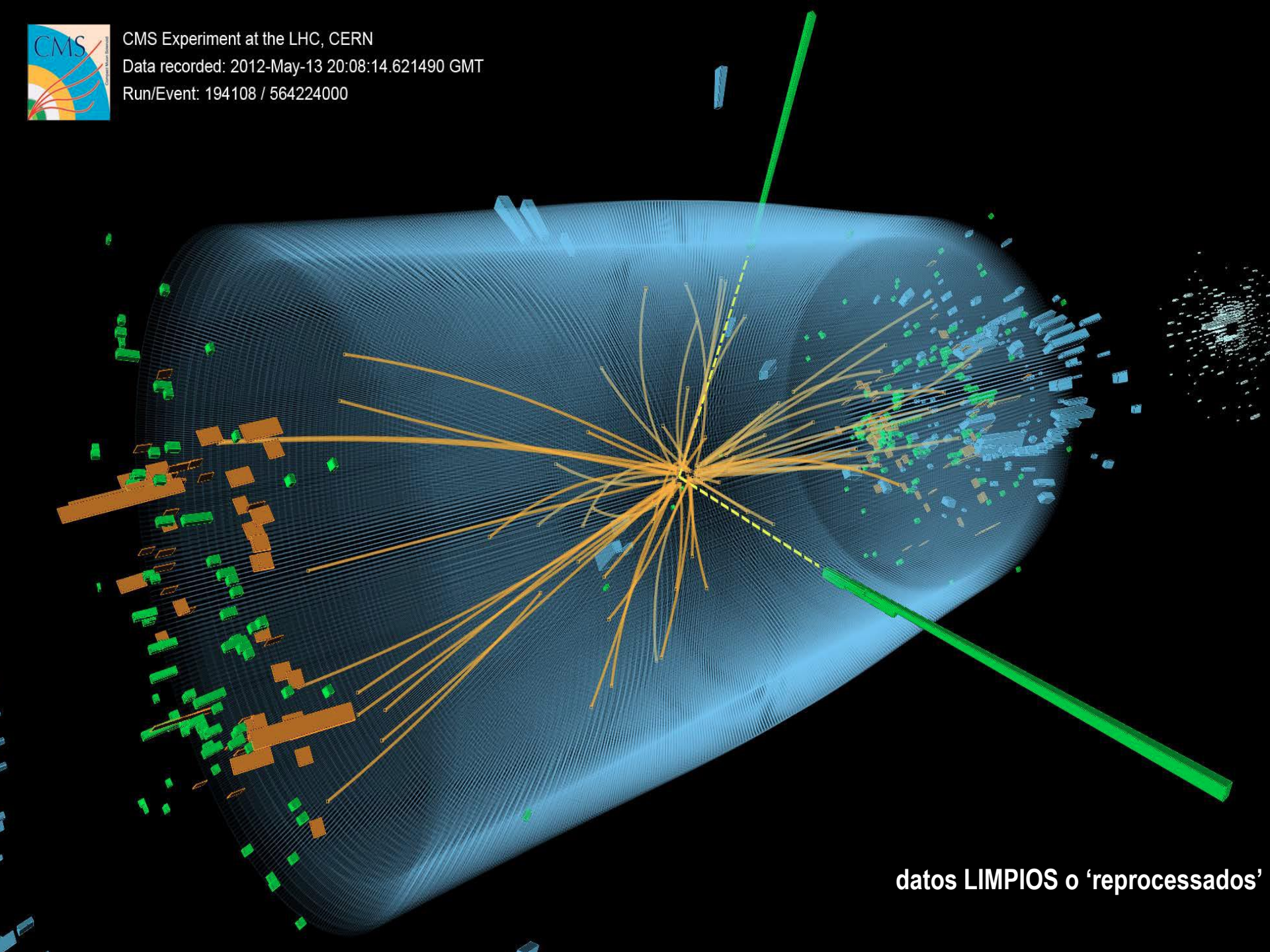
- La **importancia** del Computing en el LHC
- **Modelo WLCG** durante la gran primera toma de datos (*Run1*) y sus resultados más relevantes
- La comunidad española en WLCG
  - El **Tier-1 Español** y su contribución
- Como ha **evolucionado** el Computing en el *Run1* y los **retos** a los que se enfrenta para el *Run2* y **estrategias** para nuevas implementaciones
- Cual es el **incremento en recursos** esperado
  - Impacto para el Tier-1 Español
- **Conclusiones**

Datos en Bruto del Run1

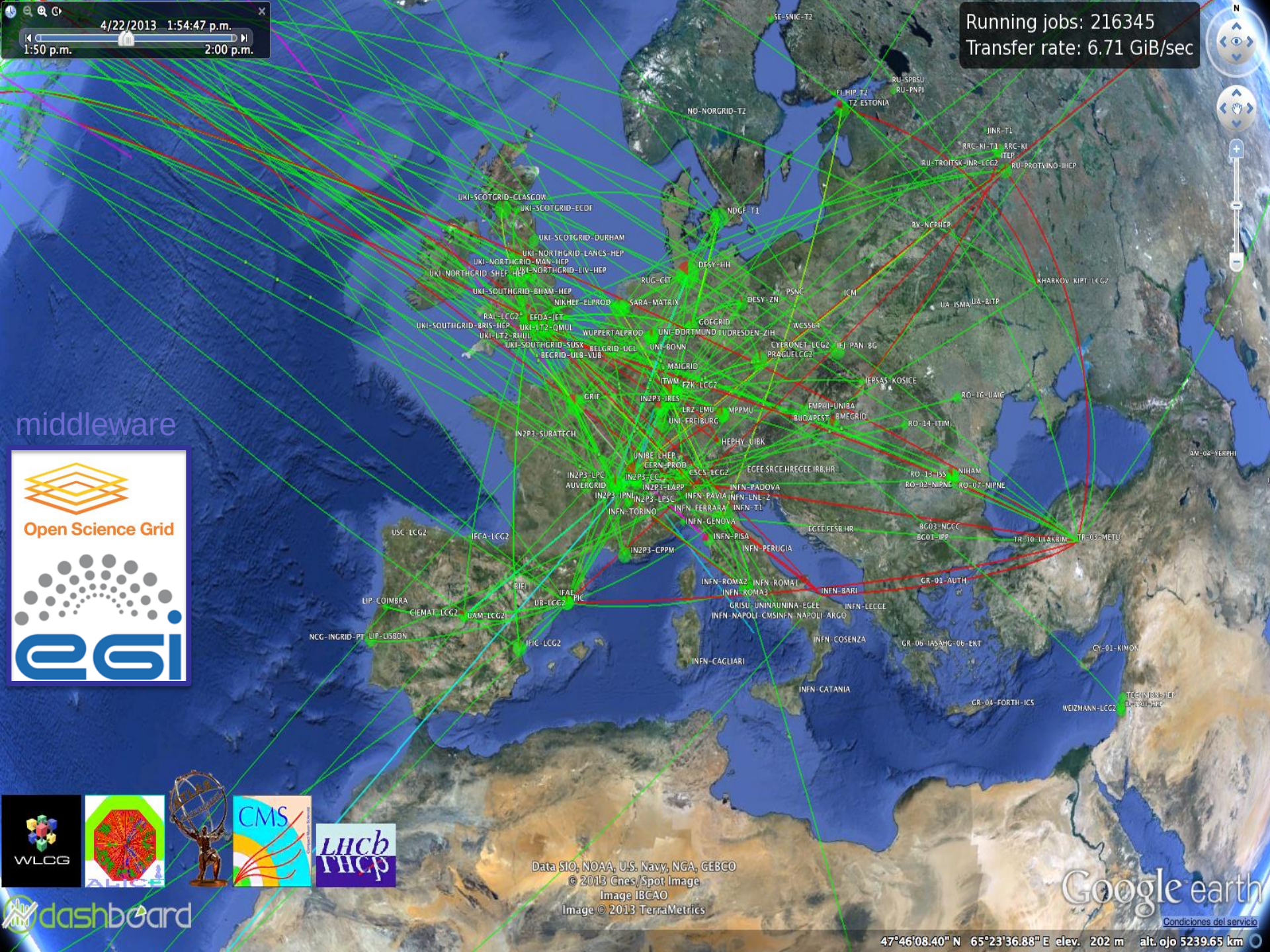
- El **incidente** de Sept. 2008 retrasó la toma de datos hasta 2010
  - Algunos meses en *commissioning* @ baja E → E_{CM} 7 TeV y luego a 8 TeV
- Durante el Run1 (2010-2013) los experimentos del LHC han registrado **billones de colisiones** *pp* y *p-ion*
  - **~70 PBs** de datos 'en bruto' registrados en el Run1

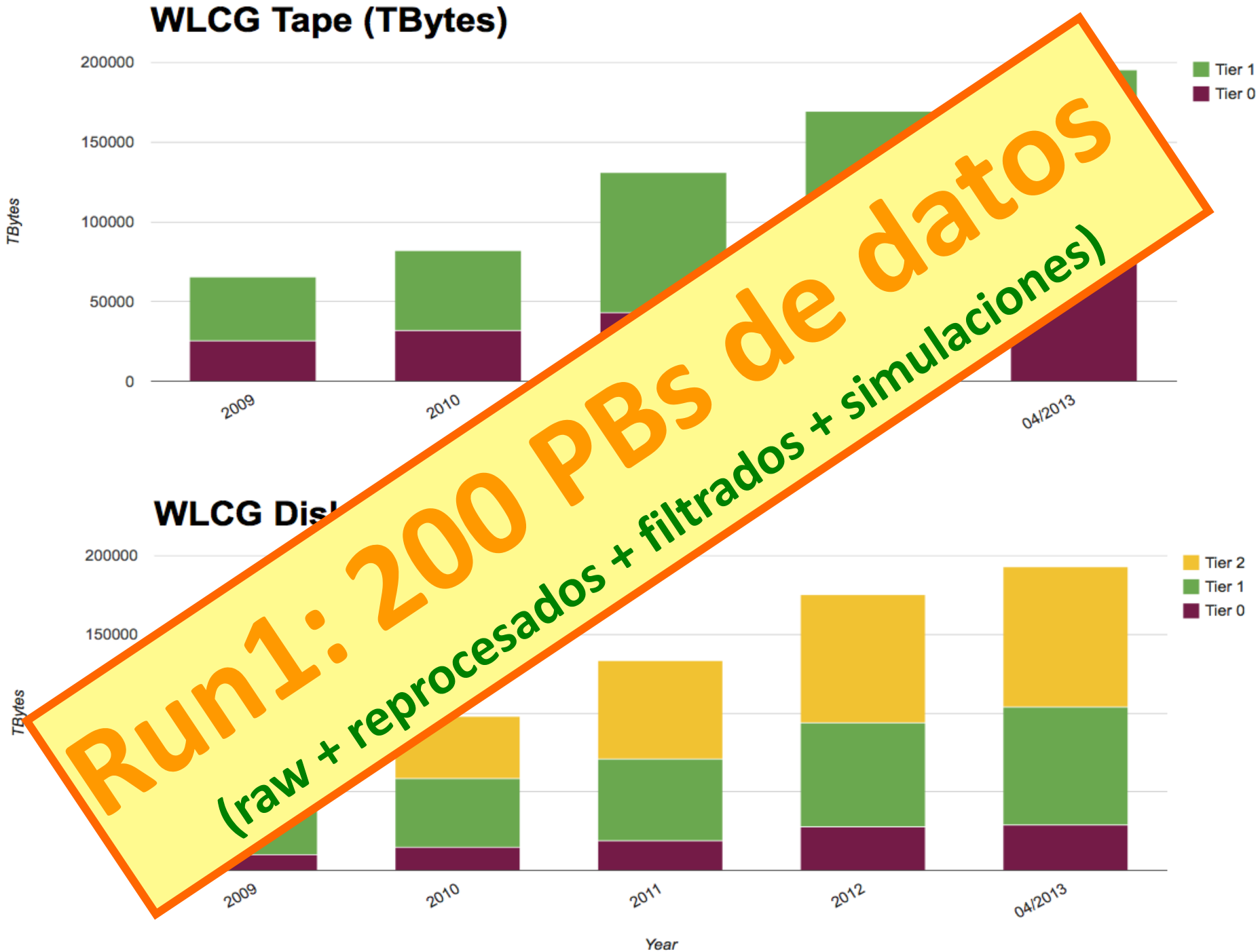



CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-08 10:22:07.828203 GMT(11:22:07 CEST)
Run / Event: 150431 / 541464
datos en BRUTO
(c) CERN 2009. All rights reserved.
http://iguana.cern.ch/ispy



datos LIMPIOS o ‘reprocessados’

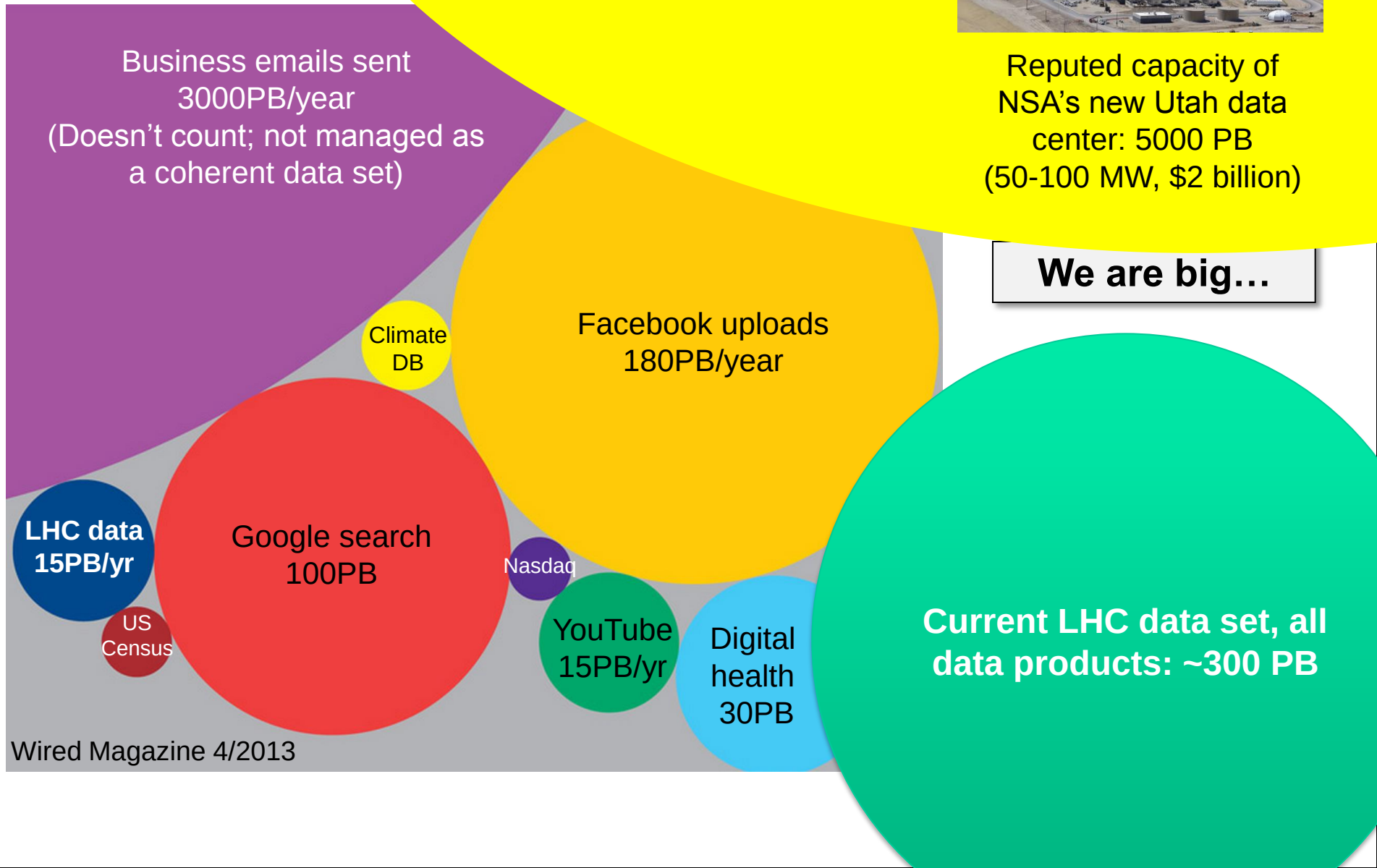

4/22/2013 1:54:47 p.m.
1:50 p.m.
2:00 p.m.
Running jobs: 216345
Transfer rate: 6.71 GiB/sec
middleware
Open Science Grid
egi
WLCG
CMS
LHCb
dashboard
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2013 Cnes/Spot Image
Image IBCAO
Image © 2013 TerraMetrics
Google earth
Condiciones del servicio
47°46'08.40" N 65°23'36.88" E elev. 202 m alt. ojo 5239.65 km


WLCG Tape (TBytes)
200000
150000
100000
50000
0
TBytes
2009
2010
04/2013
Tier 1
Tier 0
WLCG Dis
200000
150000
TBytes
2009
2010
2011
2012
04/2013
Year
Tier 2
Tier 1
Tier 0
Run1: 200 PBs de datos
(raw + reprocesados + filtrados + simulaciones)

How BIG?

Reputed capacity of NSA's new Utah data center: 5000 PB (50-100 MW, $2 billion)

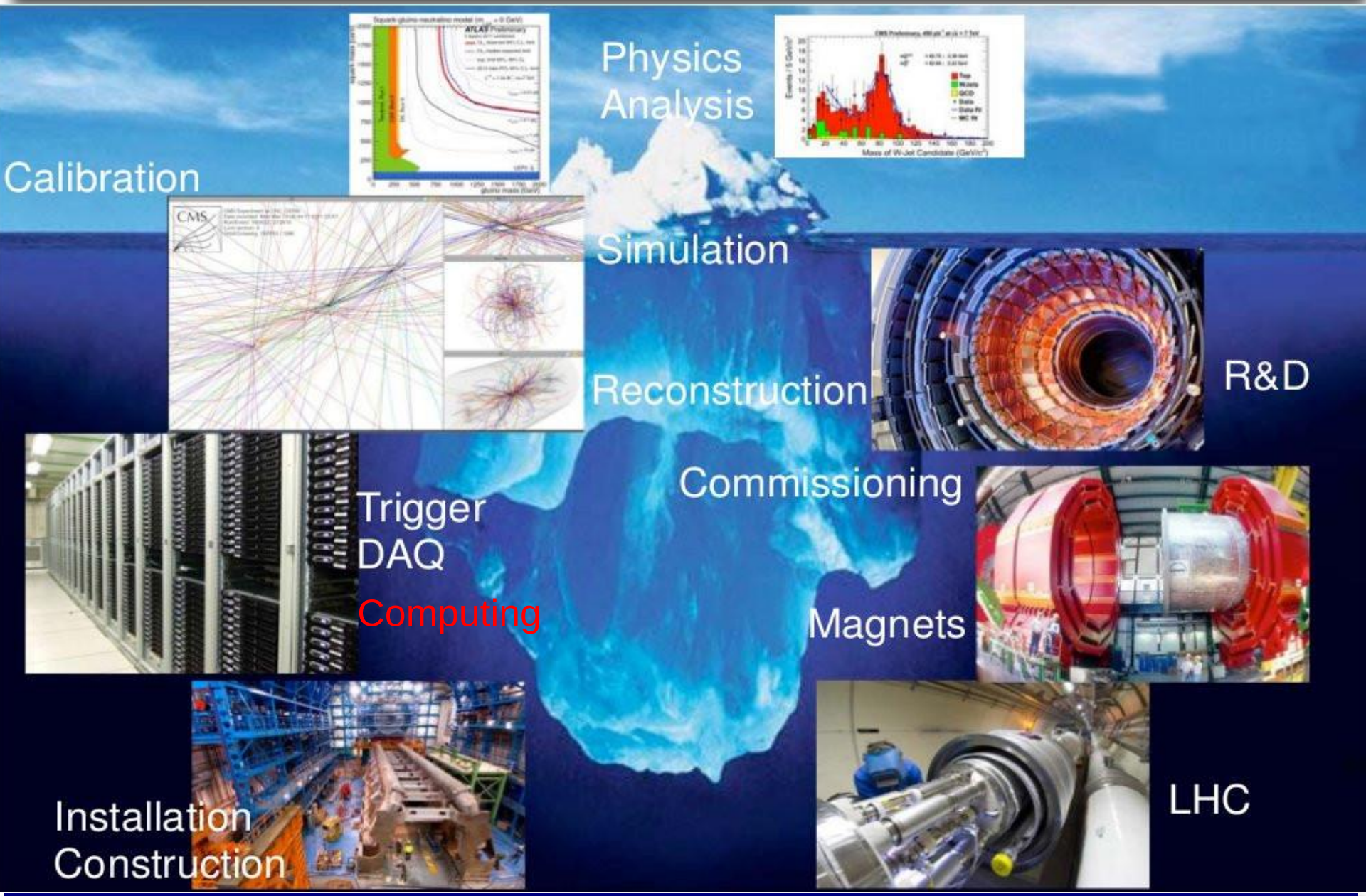
We are big…



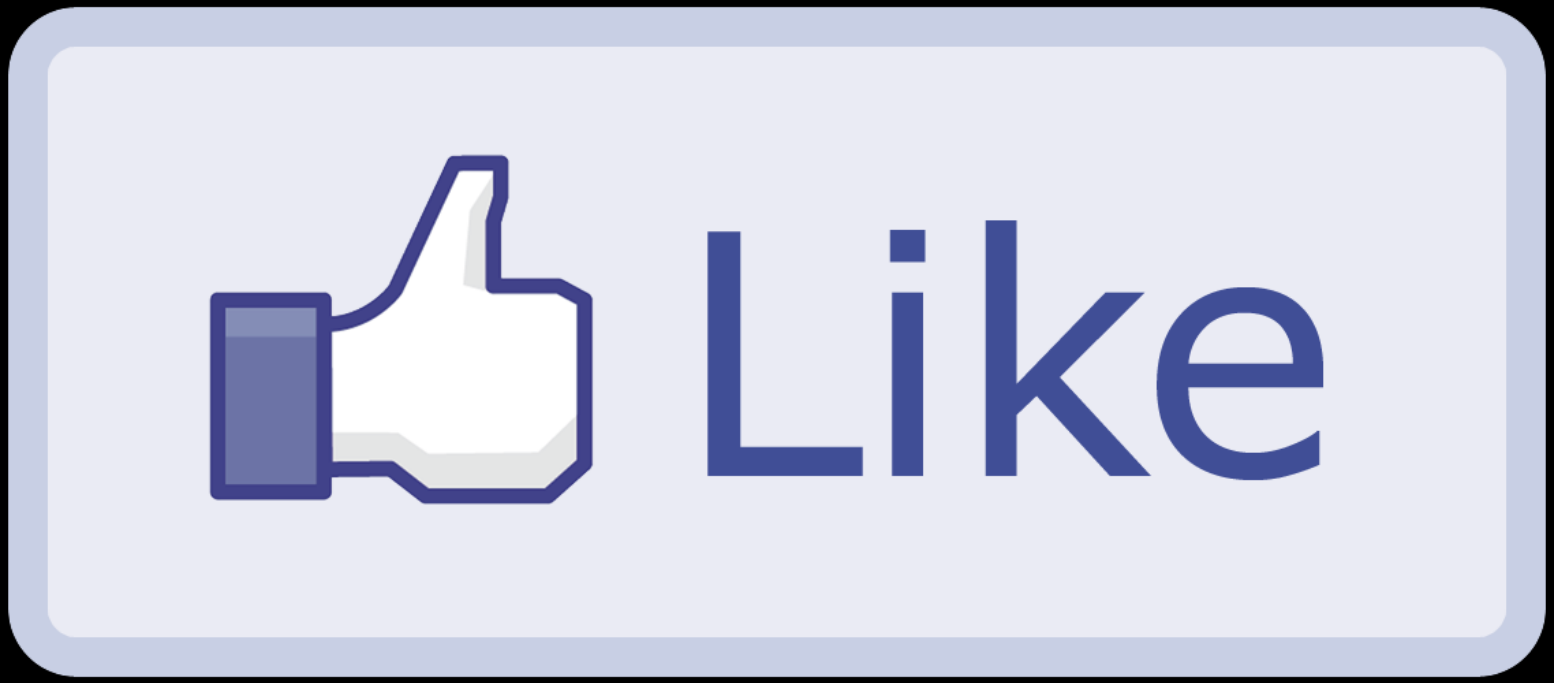
Wired Magazine 4/2013

Esfuerzo global → éxito global!

Computing: parte del esfuerzo global

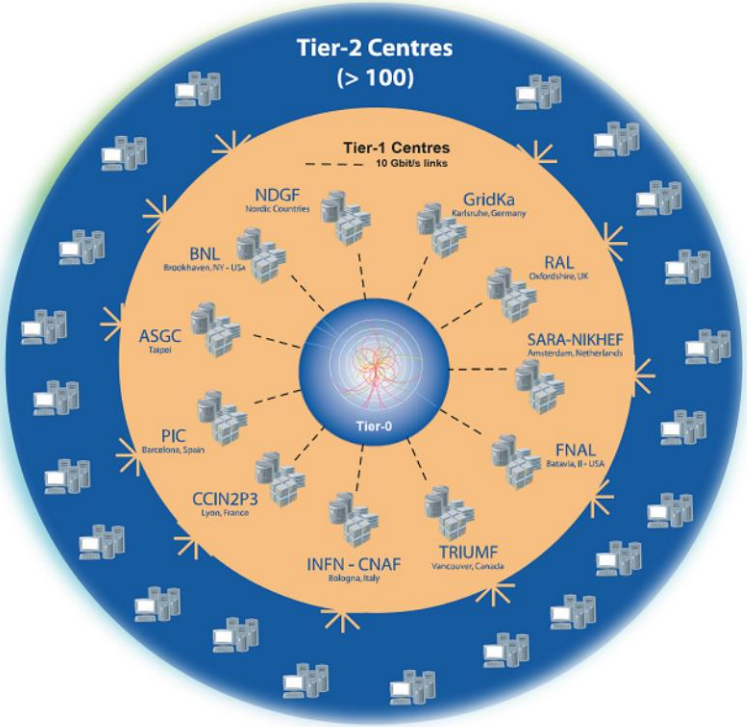


Volumen de datos sin precedentes analizados en un tiempo récord y produciendo grandes resultados científicos



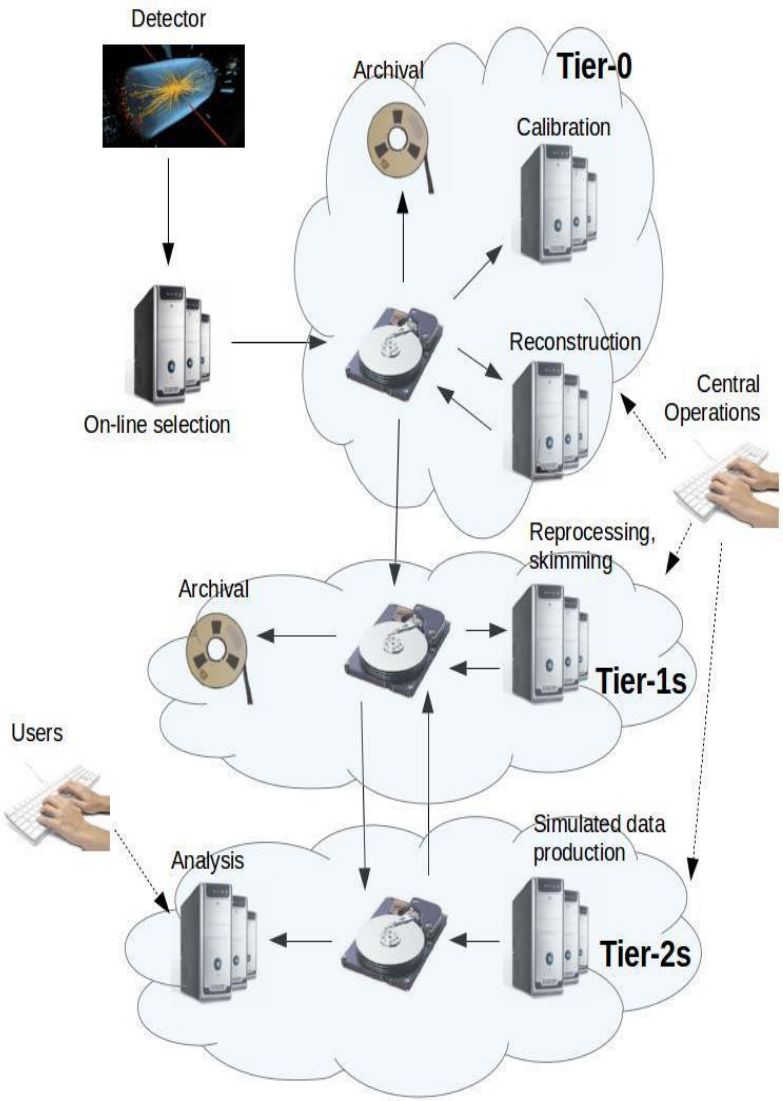
Modelo de computación WLCG

- Recursos de computación distribuidos gestionados por una **tecnología Grid** que tuvo que ser desarrollada
- Centros interconectados por **redes** privadas y/o nacionales de **gran capacidad** (Ethernet, 1-100 Gbps)
- Centros que proveen **almacenamiento masivo** (disco/cinta) y recursos de **procesamiento** (CPUs x86)
- Estructura jerárquica en **Tiers**
  - **~170 centros en 34 países**



Modelo de computación WLCG

- **Tier-0:** CERN. Primera reconstrucción, archivo de datos de larga duración
  - → **Extensión a Hungría (2x100 Gbps)**
- **11 Tier-1:** Replica de datos primarios, reconstrucción masiva y centralizada de datos, filtrado de datos, simulaciones centralizadas, análisis 'controlados'
  - → **Centros con alta calidad se servicio**
- **~150 Tier-2:** Simulación Montecarlo, análisis de datos masivos (grupos de análisis y usuarios), procesado caótico, calibración

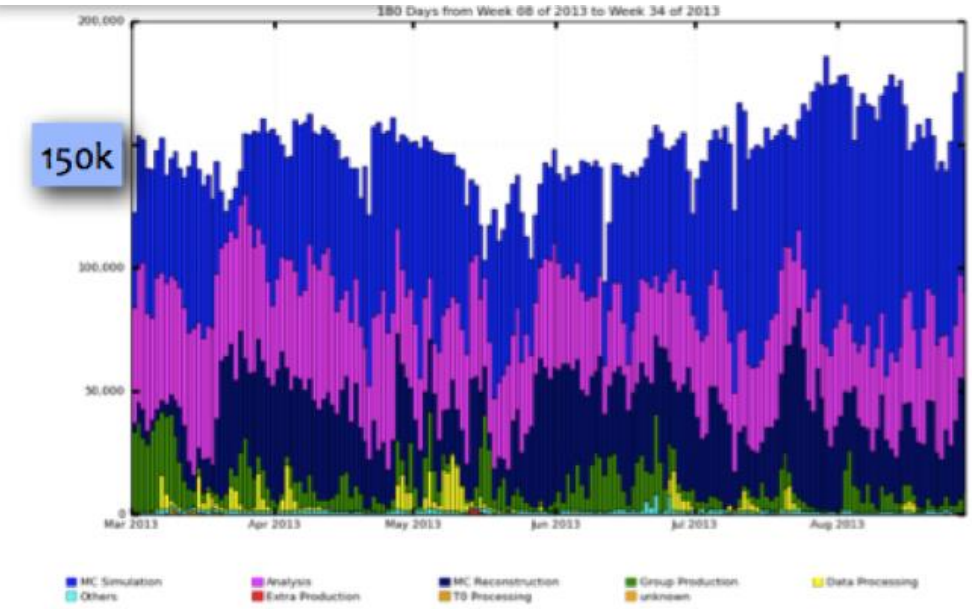


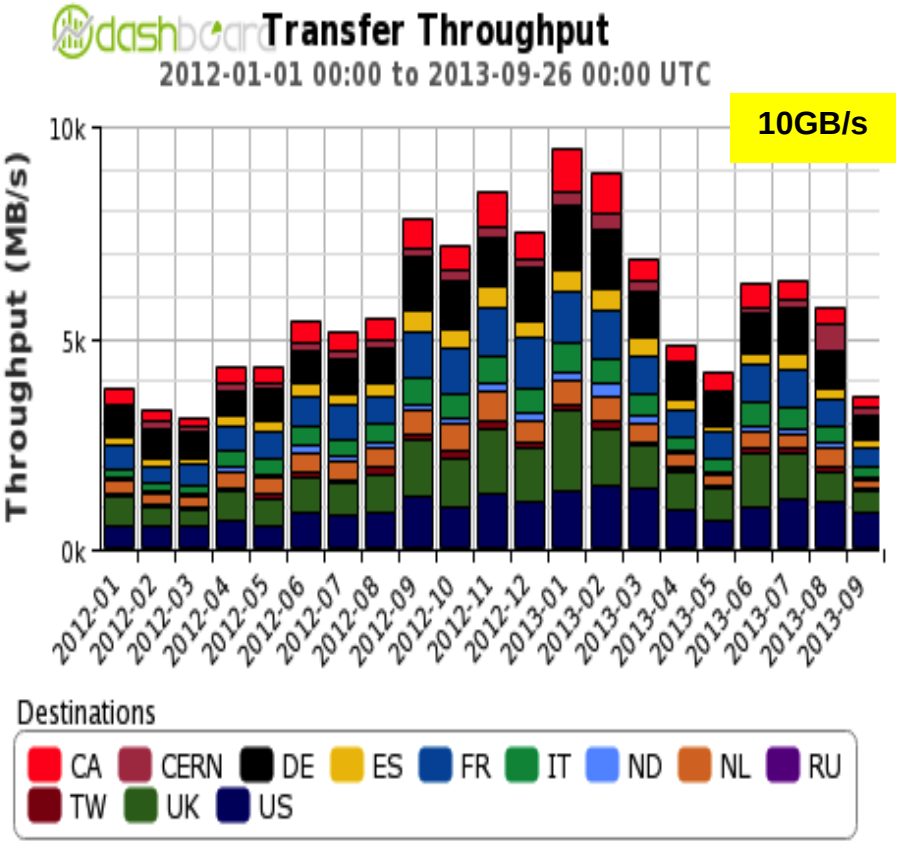
ATLAS Computing en el Run 1

150k 'slots' utilizados de forma sostenida

→ **~1.4M jobs/día** finalizados

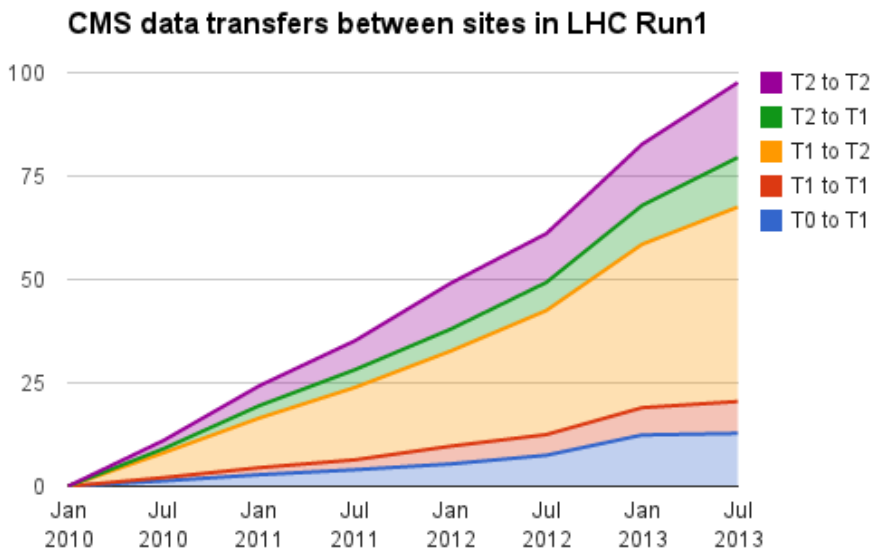
Más de **5 GB/s** de transferencias de datos a nivel mundial

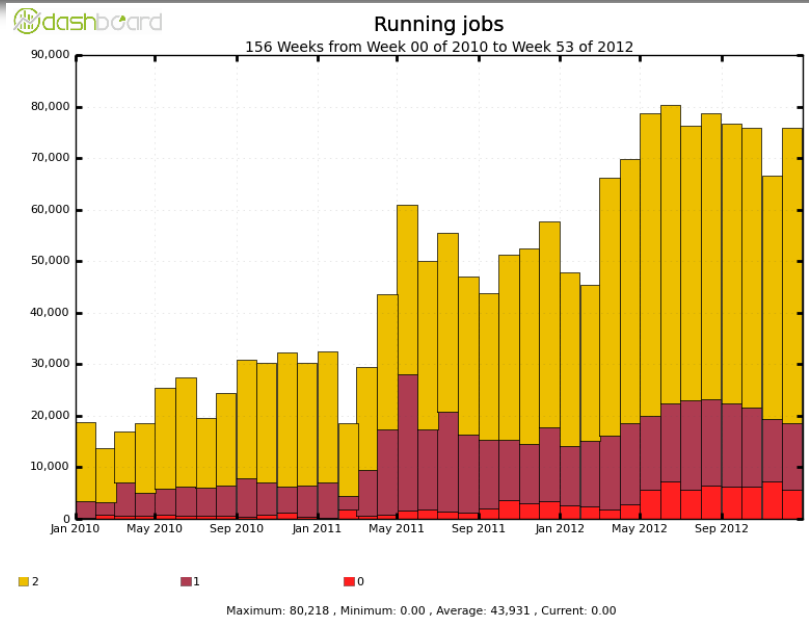


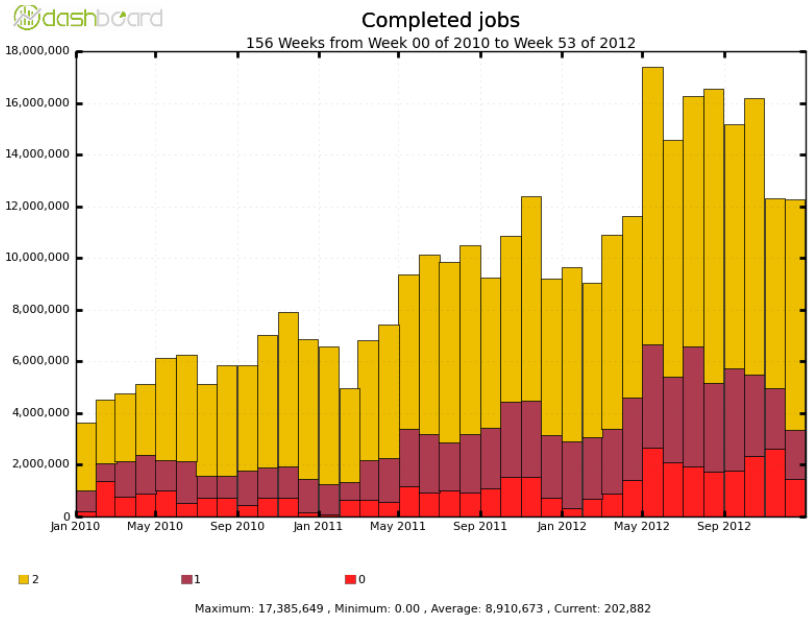


CMS Computing en el Run 1

- **~100 PB** transferidos entre centros
  - ~2/3 para análisis de datos en Tier2s (Tier1s→Tier2s)
- Saturación en uso de recursos de computación. En 2012:
  - Utilización sostenida de **~70k 'slots'**
  - **~500k jobs/día** completados







WLCG
Worldwide LHC Computing Grid

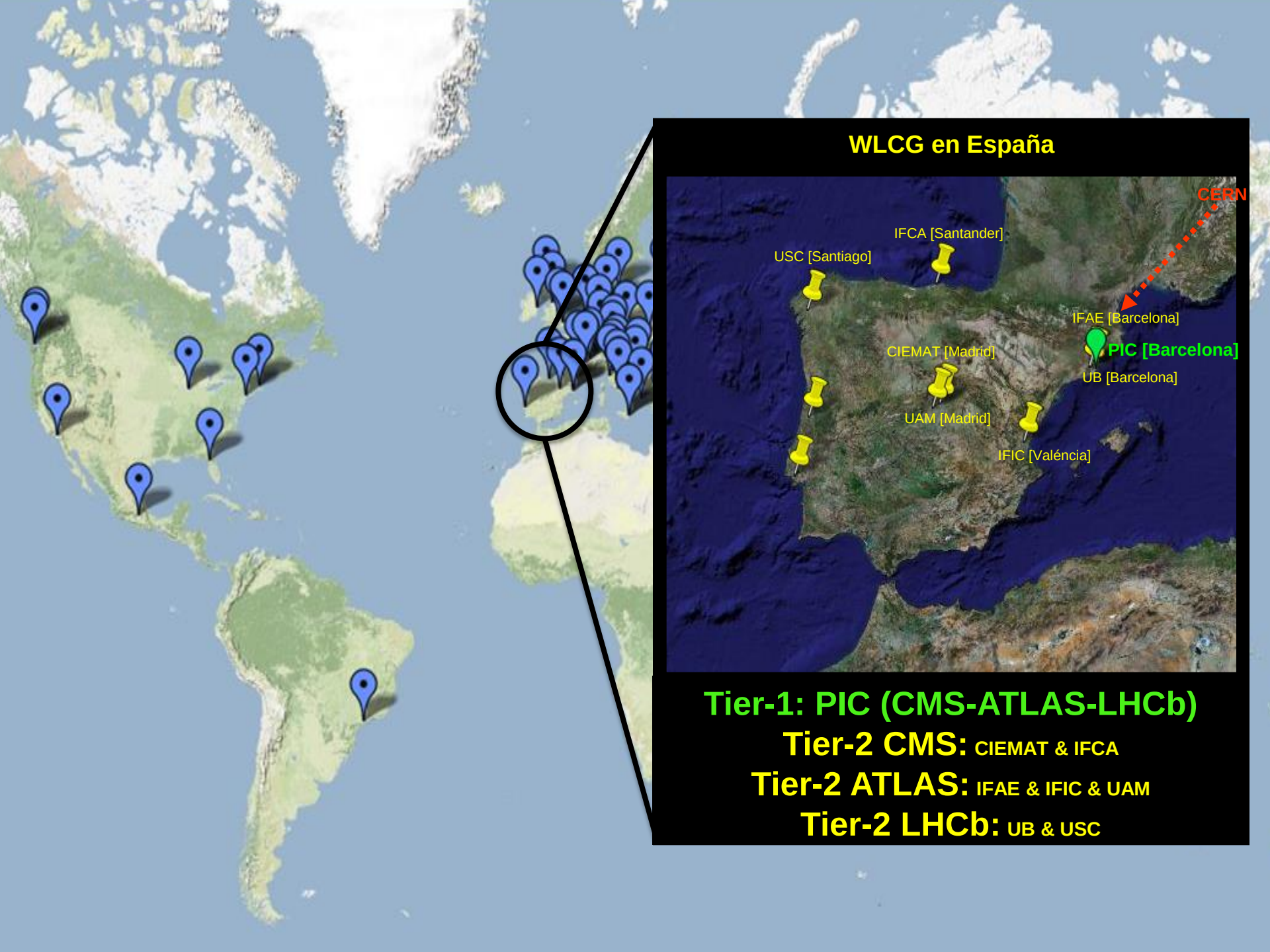
CERN

~ 10.000 físicos potenciales
Alrededor de 280.000 tare
y 300.000
~300 PB
Picos de b
Fi

WLCG
se en el Grid
ibles
cinta disponible
en transferéncias de datos
ecursos están en el CERN
dedicadas a 10 Gbps [CERN-Tier-1s]

Todos los recursos disponibles WLCG se han utilizado de forma intensiva durante el Run1 del LHC

http://wlcg.web.cern.ch/


WLCG en España
CERN
IFCA [Santander]
USC [Santiago]
IFAE [Barcelona]
PIC [Barcelona]
CIEMAT [Madrid]
UB [Barcelona]
UAM [Madrid]
IFIC [València]
Tier-1: PIC (CMS-ATLAS-LHCb)
Tier-2 CMS: CIEMAT & IFCA
Tier-2 ATLAS: IFAE & IFIC & UAM
Tier-2 LHCb: UB & USC

WLCG en España

- En el **2000** se empezó a gestionar en el CERN el WLCG
- Los grupos españoles se interesaron rápidamente en participar:
  - Se coordinaron peticiones de acciones especiales y se crean los primeros equipos y prototipos en **2001**
  - En **2005** se conforma la estructura de **1 Tier-1** y **3 Tier-2s**
    - Participación española al 5% del total para Tier-1 y Tier-2 (6,5% LHCb)
- **2005-2009**: participación en pruebas a gran escala para ver si los servicios de computación para el LHC cumplen con los requisitos establecidos por los experimentos
  - los centros españoles participan **de forma satisfactoria**, demostrando así estar listos para la toma de datos del LHC
- Excelente posicionamiento durante la 1ª toma de datos (Run1)

WLCG en España

- Participación activa en el desarrollo de nuevas herramientas
- Contribución en la gestión del software de los experimentos
- Personal en puestos relevantes
  - en WLCG
  - en los experimentos
  - en comités internacionales de computación
- Comunidad muy apreciada por sus contribuciones en Computing

Generalitat de Catalunya
Departament d'Economia i Coneixement
Secretaria d'Universitats i Recerca

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

Institut de Física d'Altes Energies IFAE

UAB
Universitat Autònoma de Barcelona

UAB CEI
CAMPUS D'EXCEL·LÈNCIA INTERNACIONAL

PIC en números

8 PBs en cinta magnética

6 PBs en disco

~4000 CPUs - procesamiento

~85% Tier1

PIC en números

Sala a suelo-alzado **de 150 m²**
~200 KVA con UPS & generadores diesel
Sistema de protección **contra-incendios** sectorizado
34 racks - **1400U** espacio para equipamiento
1000+ servidores
2 robots de cinta magnética
~40 lectoras / ~8500 cintas (40 PBs con T10KC)

+ CPD autónomo alt. eficiente de **25 m²** (**+100 KVA**)

LAN: equipado a **500+ Gbps**
WAN: 1x10 Gbps (dedicada, LHCOPN) + **2x2 Gpbs**

~25 Ingenieros/Científicos

PIC: centro de datos

Electricidad, Servidores, Software, … → **Ciencia**

Crecer de acuerdo con los requerimientos científicos

En general, es un proceso **costoso**

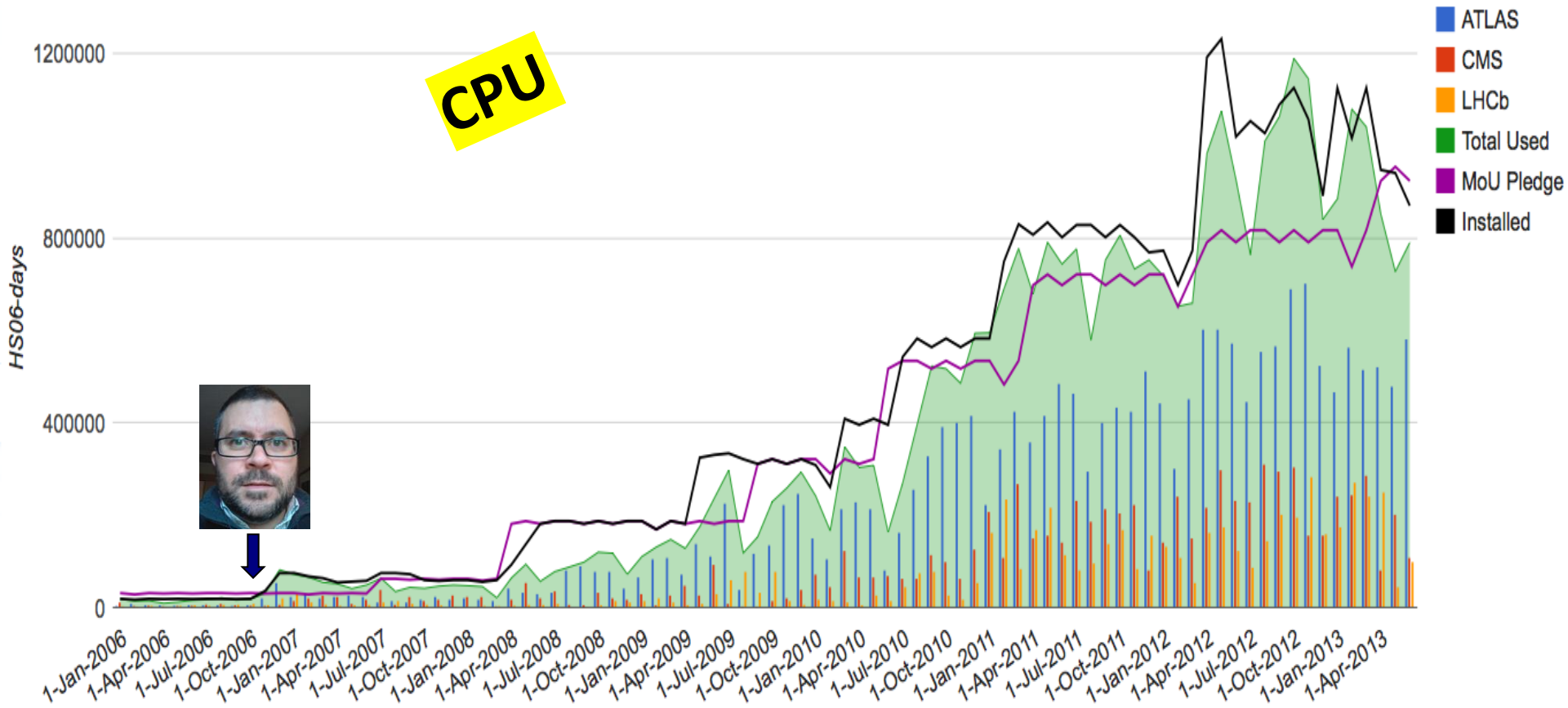
2013-2014: completa renovación del sistema de cooling

Nuevas **refrigeradoras** +potentes +eficientes

introducir Free-cooling indirecto

Doblar recursos en los próximos **3 años**

Contribución PIC a WLCG: CPU

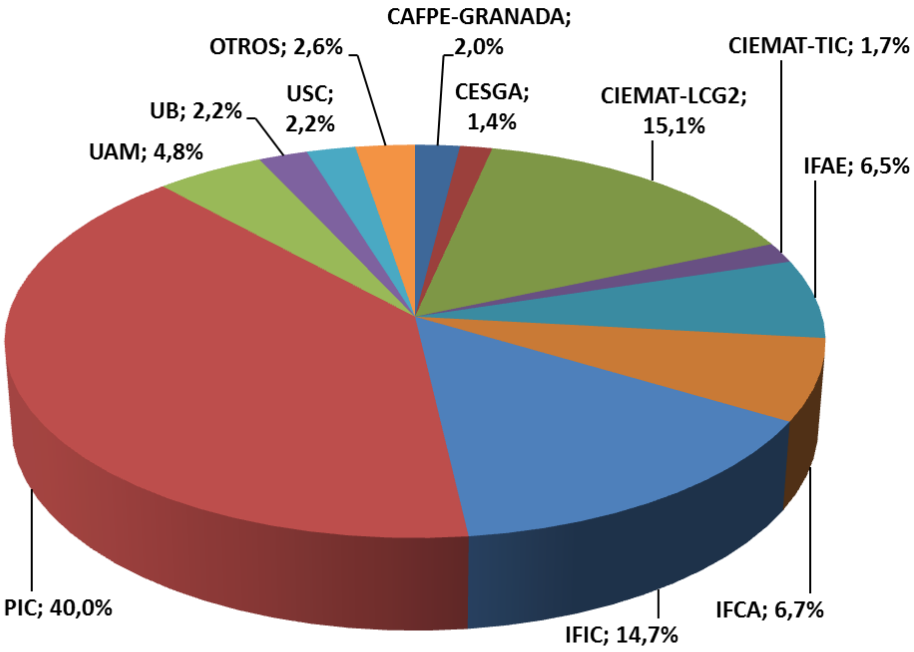


http://accounting.egi.eu

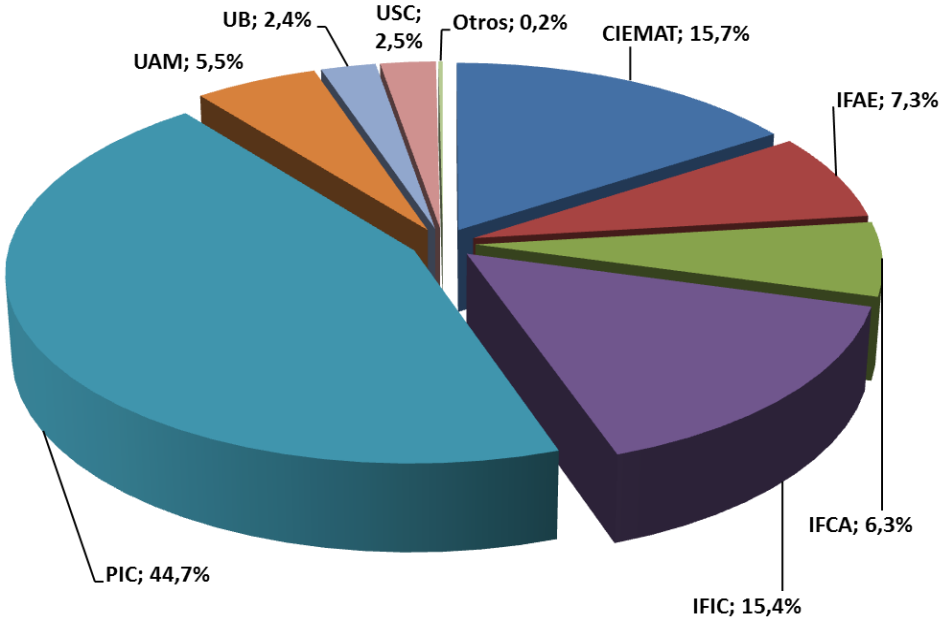
CPU normalizado – Grid España

- Total de CPU normalizado utilizado por trabajos GRID por centro en España **[2010-2013]**
  - ATLAS, CMS y LHCb son el 88% del total

Todas las VOs

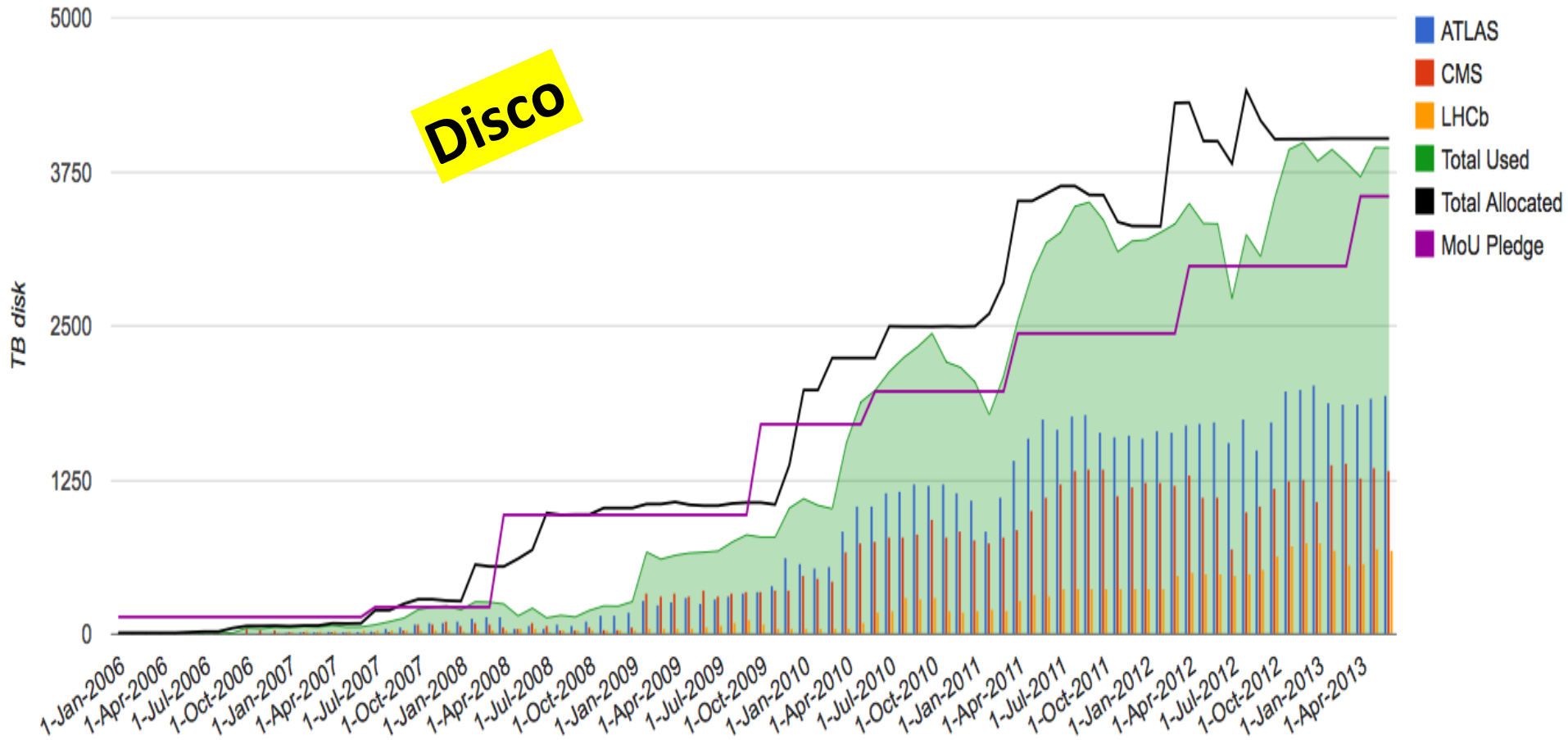


VOs: ATLAS, CMS y LHCb



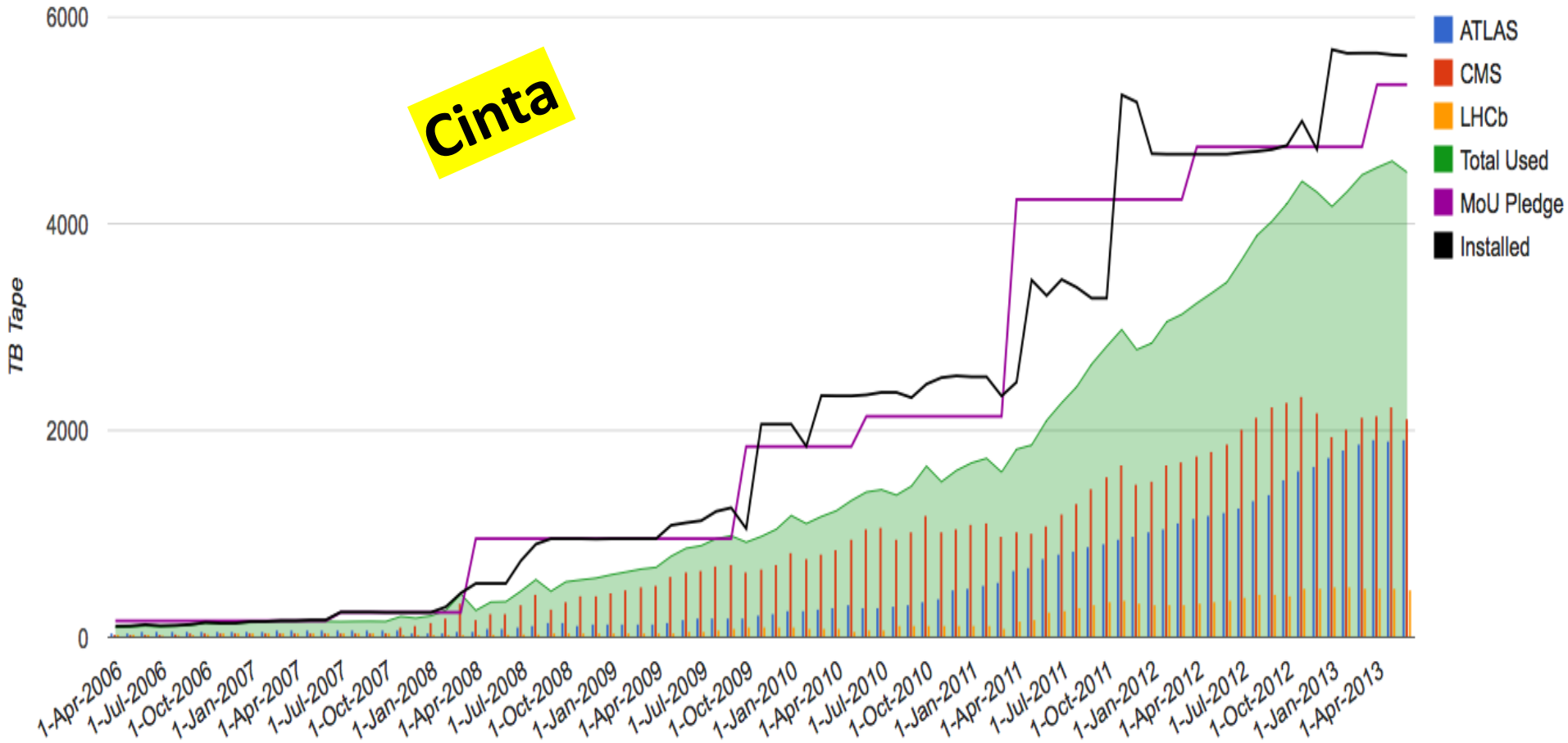
http://accounting.egi.eu

Contribución PIC a WLCG: Disco



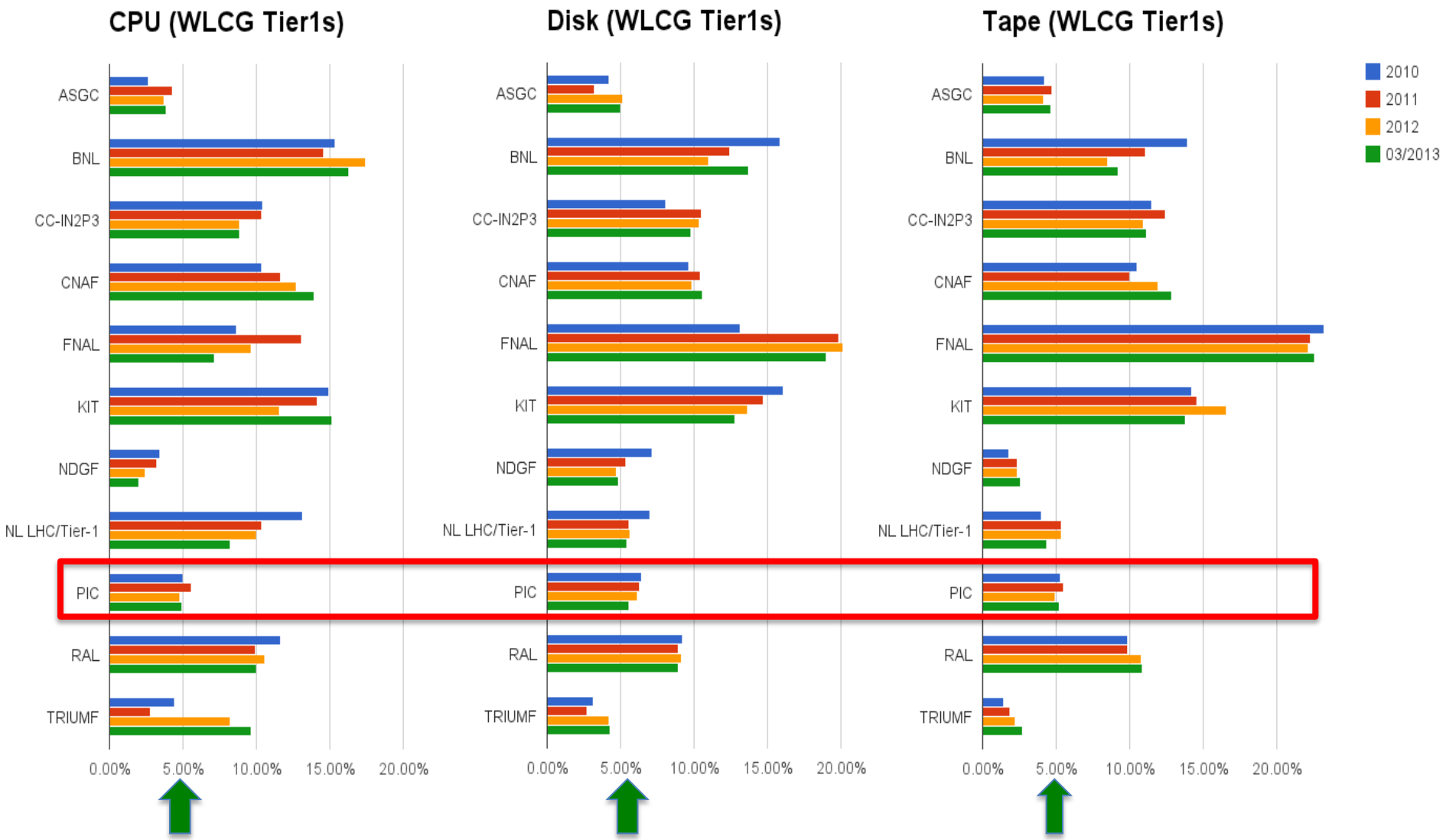
http://accounting.egi.eu

Contribución PIC a WLCG: Cinta



http://accounting.egi.eu

Contribución PIC a WLCG

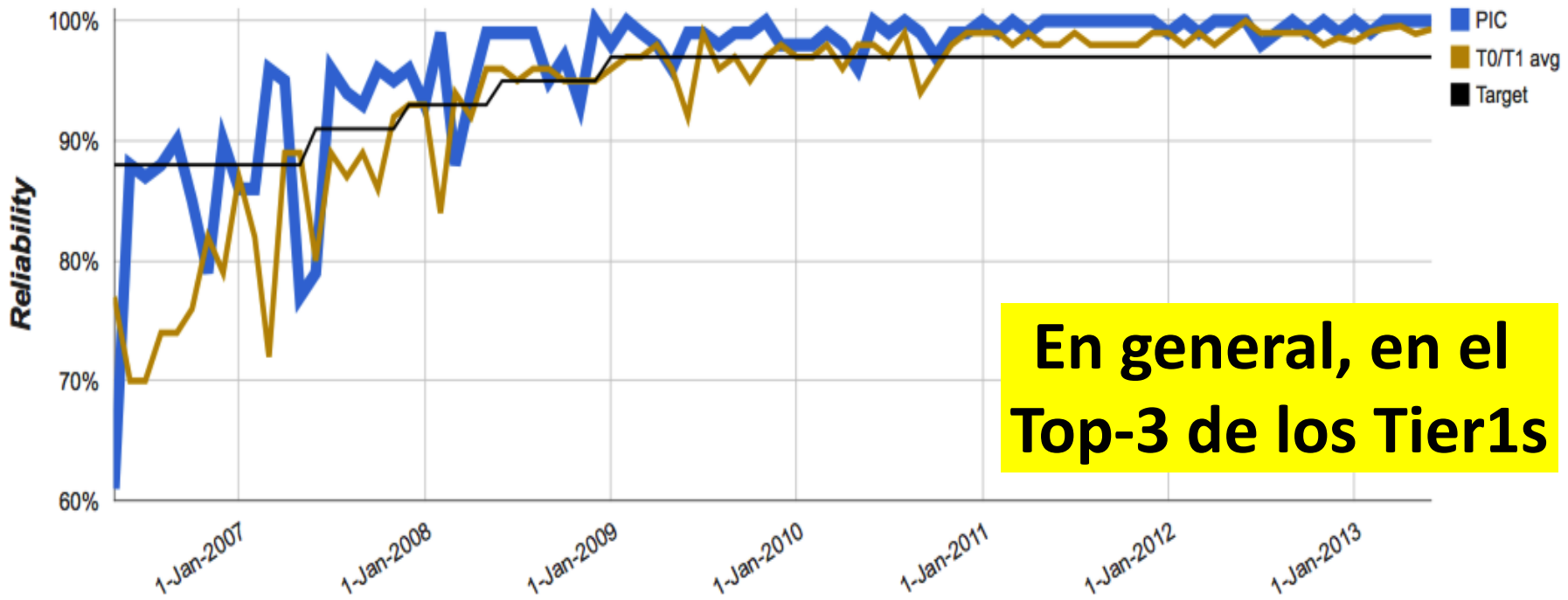


✔ Uso estable de los recursos del PIC durante todo Run1, al nivel esperado de contribución

http://accounting.egi.eu

La fiabilidad del PIC Tier1

- Los servicios Tier1 tienen que ser **extremadamente fiables**:
  - 4h = tiempo máx. de interrupción del servicio no deseado
  - 6h = tiempo máx. de degradación en importación de datos Tier0
- Los servicios críticos en el PIC funcionan en modo 365x24x7 **[MoD]**



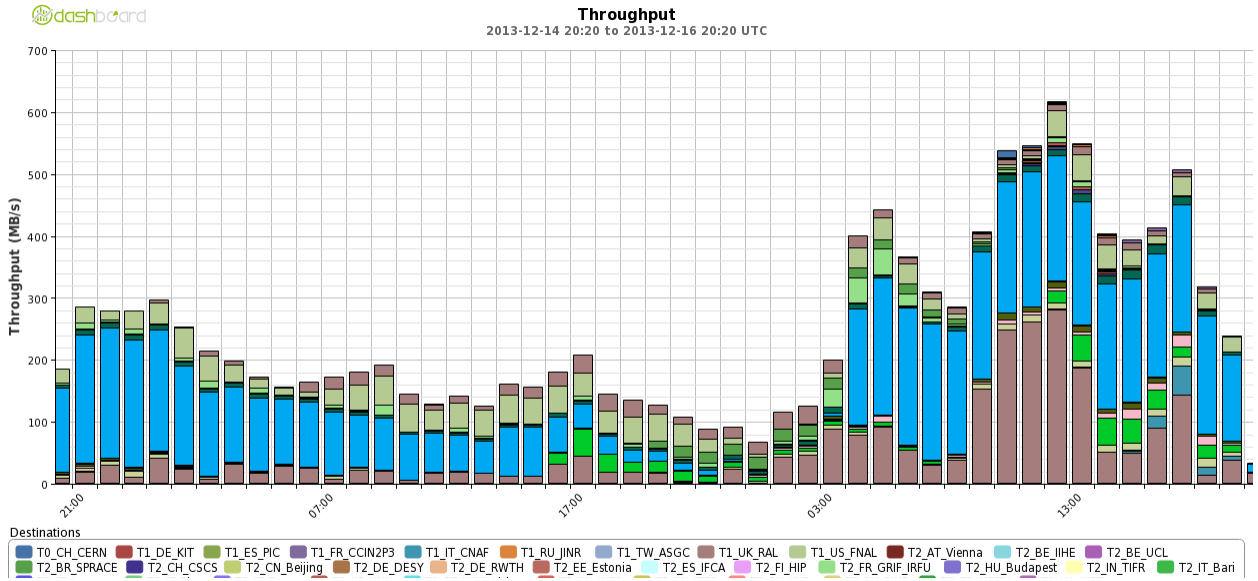
En general, en el Top-3 de los Tier1s

BACK TO THE PAST

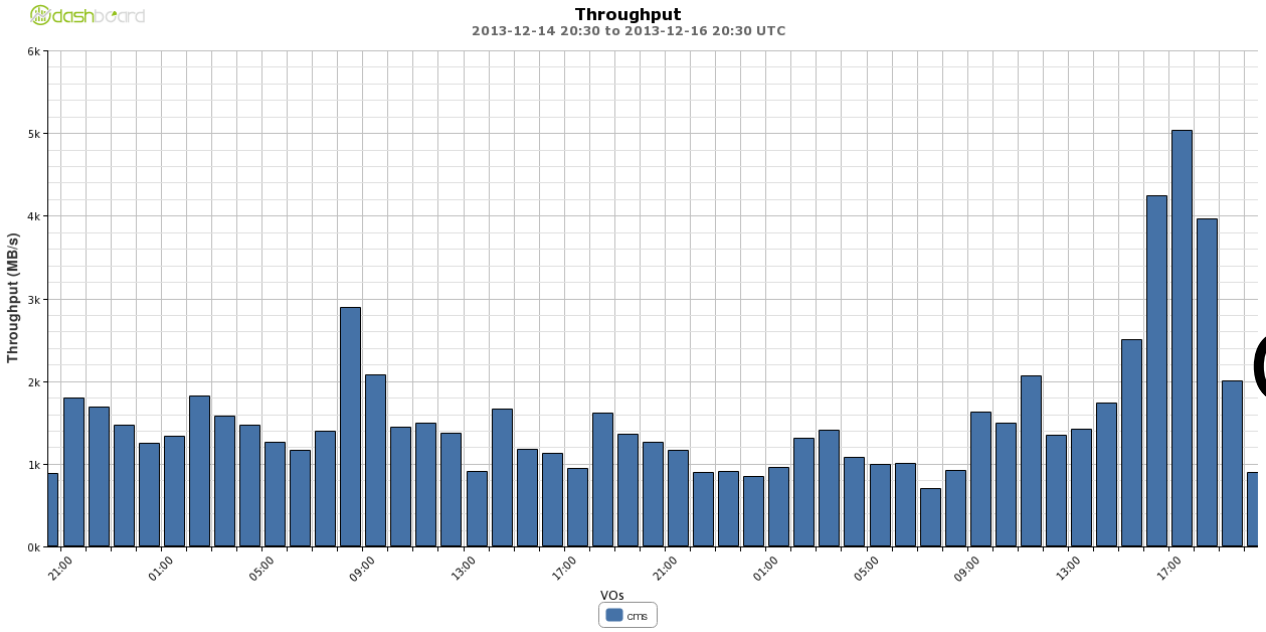
Evolución del Computing en Run1

- Gran aumento de **capacidad/estabilidad de la red**
- Optimización en la distribución de datos
  - Relajación en la replicación jerárquica de los datos
  - Se permite una **replicación de datos abierta entre centros**
- Optimización en el acceso a datos
  - Pre-localizar datos en los centros y enviar trabajos a los datos
  - Permitir el **acceso a datos remotos**: en caso de fallo de acceso local, *overflow* de los sitios ocupados, centros sin disco, etc…
  - Despliegue de federaciones ***xrootd***, protocolo que permite el acceso remoto a datos

CMS: jobs leyendo datos remotos



xrootD



GridFTP

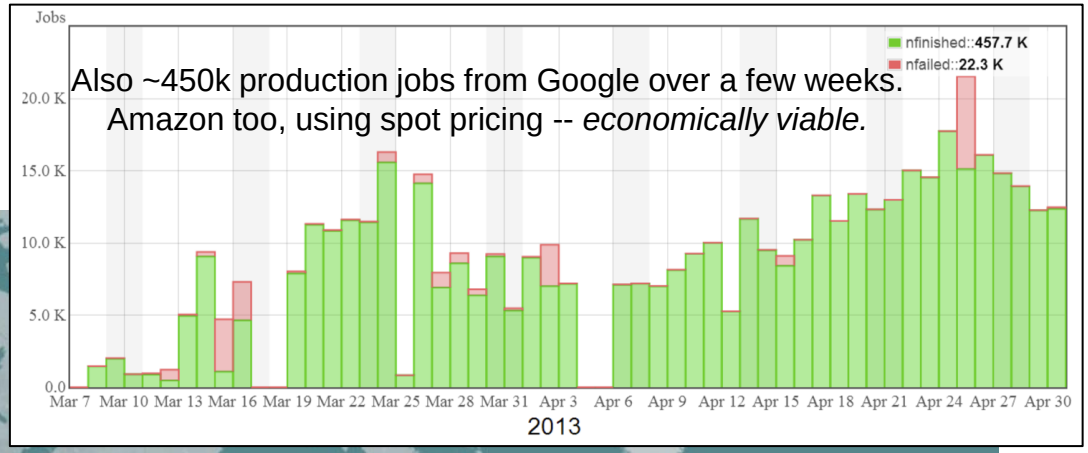
Evolución del Computing en Run1

- Se optimiza la **distribución de software** de experimentos
  - Se liquida el sistema de instalaciones locales
  - Utilización del CERNVM *File System* (**CVMFS**)
  - Caché HTTP optimizado para la entrega de software
  - Muy eficiente y escalable

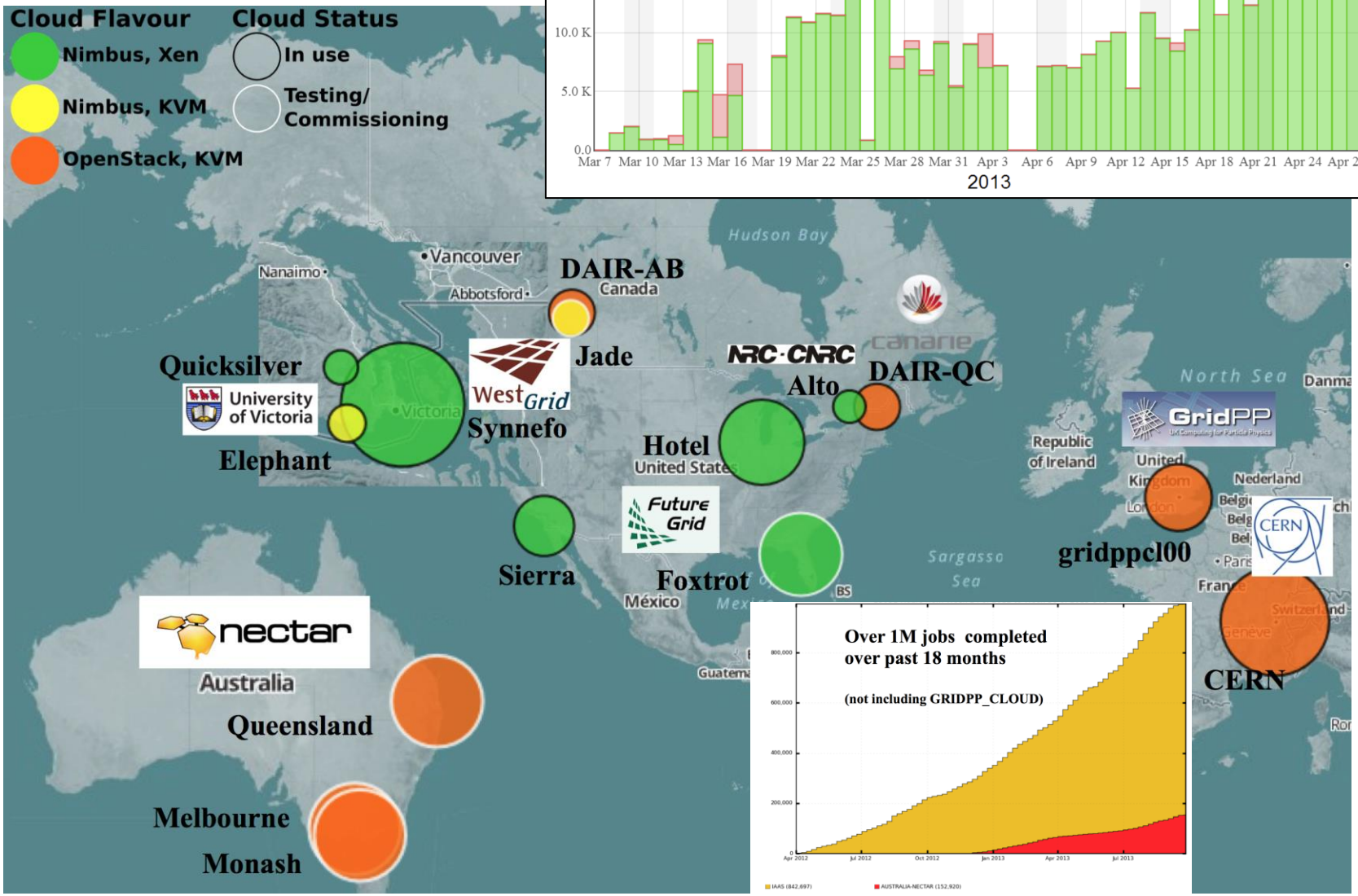
- Mejoras en la **gestión de trabajos**
  - Los experimentos pasan a usar un modelo de ***pilot jobs***, recibiendo tareas desde una cola centralizada
  - Uso más eficiente de los recursos

- Exploración de técnicas **Cloud Computing** y **uso oportunista** de recursos del HLT o *clusters* HPC

“Grid of Clouds” used by ATLAS

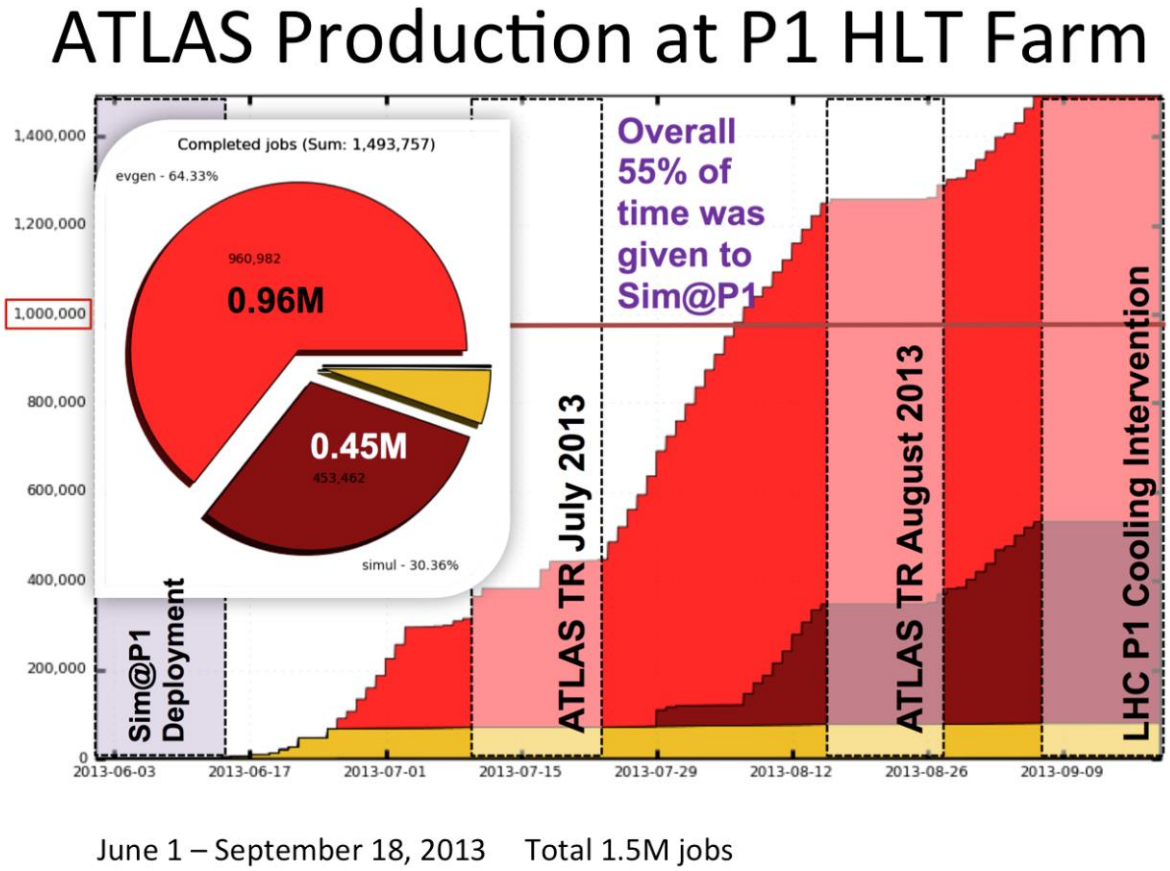


Also ~450k production jobs from Google over a few weeks.
Amazon too, using spot pricing -- *economically viable.*



Grid: ~1.4M jobs/día

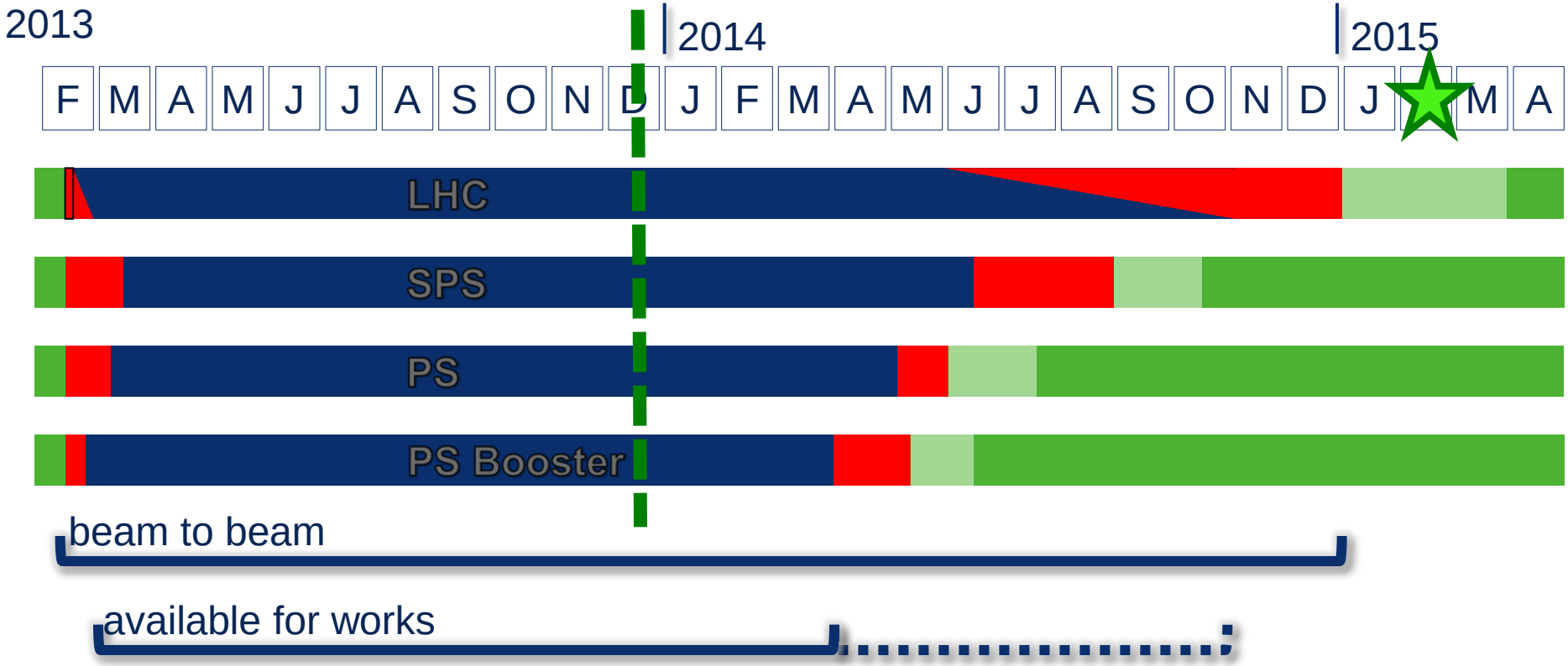
Opportunistic Resources ATLAS HLT



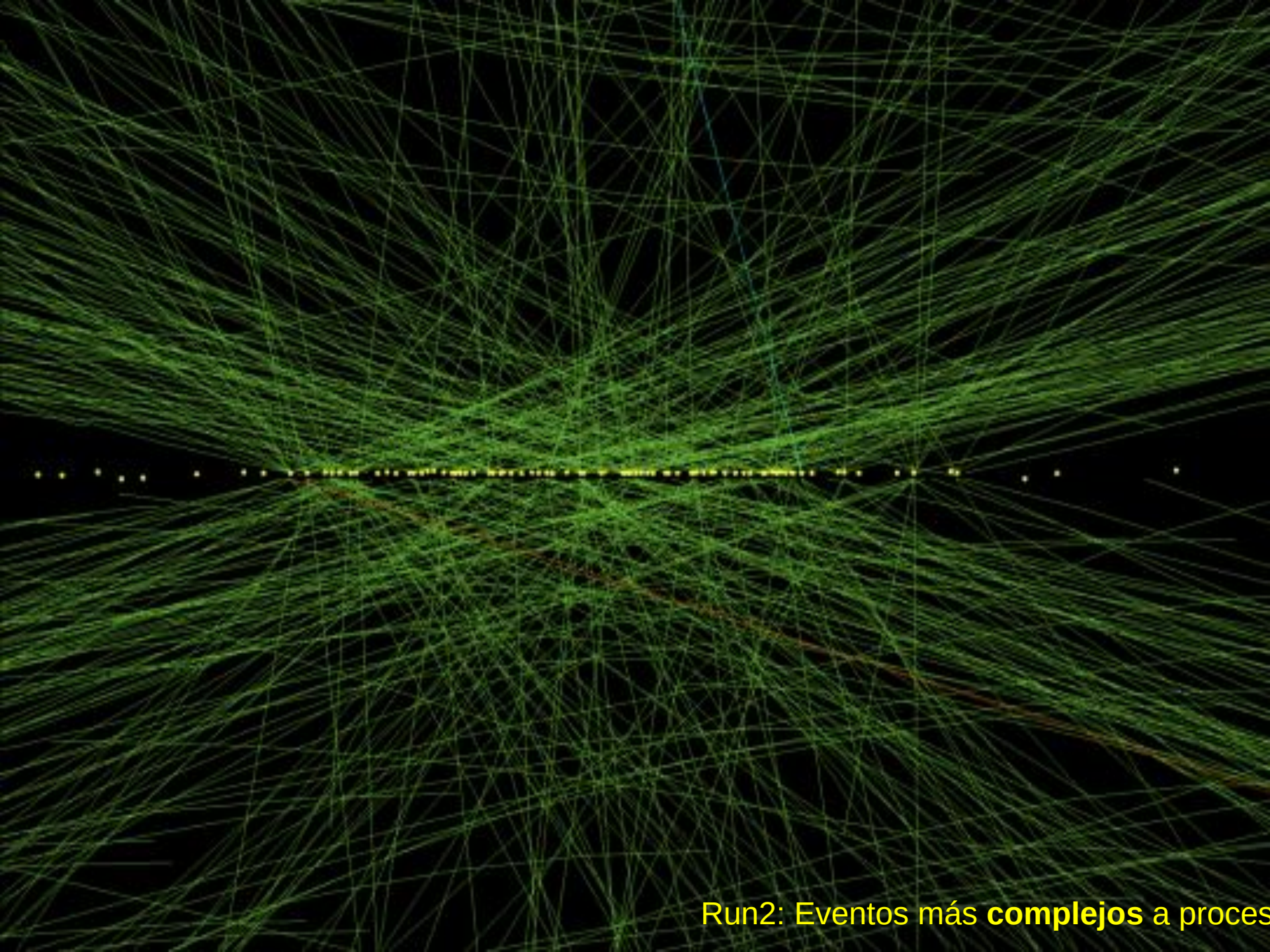
The largest ATLAS grid site when running

- LHCb usa su granja HLT desde principios de 2013
  - supone un ~20% de sus recursos
- CMS, ATLAS también han construido sus plataformas Cloud HLT basadas en **OpenStack**

LHC Long Shutdown 1 (LS1)



***Beam* previsto para Febrero de 2015 (la física en ~Mayo'15)**



Run2: Eventos más **complejos** a proces

Retos de Computing hacia el Run2

- Un **buen sistema** de Computación distribuido del LHC en el Run1, pero el Run2 a partir de 2015 plantea nuevos desafíos
  - Aumento de la **energía** y **luminosidad** del LHC
- Eventos más **complejos** a procesar (+*pile-up*)
  - Mayor tiempo de reconstrucción de eventos [CMS ~2x]
  - Más memoria RAM para analizar cada evento
- Incremento en los ***trigger rates***:
  - Mejor determinación propiedades del Higgs + Nueva física
  - Factor 2x-2.5x en producción de datos (ATLAS/CMS)
- Necesita un aumento sustancial de los recursos de computing, que es probable que **no nos podamos permitir**

Actualización del LHC Computing

- Éste período de ***shutdown*** es muy útil para:
  - Analizar todo lo aprendido del Run1
  - Estimar qué recursos necesitaremos de caras al Run2
  - Estudiar la evolución del Computing (y su coste)

- Planificar, adaptar y/o **desarrollar nuevas herramientas** de Computing para 2015 y más allá:
  - Mientras seguimos operando la infraestructura!
  - Asumiendo que, en la coyuntura actual, no se va a disponer de un incremento de financiación considerable (*o nulo*)

- En discusión en los experimentos, en colaboración con el CERN IT, WLCG, y expertos de los experimentos – WLCG Computing Model Evolution

Estrategia de Computing para el Run2

- Aumentar los recursos en WLCG tanto como sea posible
  - Ajustándose a una situación presupuestaria **restringida**
  - Aprovechando la evolución de costes en la tecnología - podría ser factible un ~25% de crecimiento anual
  - **Hacer un uso más eficiente y flexible de los recursos**

- **Reducir** las necesidades de CPU y almacenamiento
  - menos pases de reprocesamiento, menos simulación, formato de datos más compacto, reducir las replicaciones de datos,…

- **Replicación dinámica** e inteligente de datos populares y permitir **acceso remoto**
  - Replicación automática de ‘datos calientes’ y borrado de ‘datos fríos’, I/O remoto

Estrategia de Computing para el Run2

- **Romper con los límites** entre los niveles de computación
  - Ejecutar reconstrucción, simulación y análisis en centros Tier-1/Tier-2 indistintamente
- Producción centralizada de ***group analysis datasets***
  - Limitar el "análisis caótico" sólo a lo que realmente es específico de los usuarios
  - Eliminar redundancias en el procesamiento/almacenamiento, reduciendo la carga de trabajo operativo para usuarios
- Acceso a **recursos oportunistas**
  - Clusters HPC, nubes académicas o comerciales, computación voluntaria, …

Acceso a nuevos recursos en el Run2

- Uso de las **granjas HLT** para el procesado de datos
  - Durante largos períodos de tiempo sin toma de datos, o incluso en períodos *inter-fill* del acelerador

- Adoptar **arquitecturas avanzadas**
  - # de transistores en CPUs (Moore's Law), pero Velocidad de *Clock* está saturada a 2.x GHz desde hace 10 años… Multi-core, co-procesadores, concurrencia…
  - Run1 está procesado bajo Enterprise Linux en procesadores x86
  - Procesadores de múltiples núcleos, CPUs de bajo consumo, GPUs → Un entorno heterogéneo!
  - Paralelizar las aplicaciones de procesamiento será *clave*
    - Desarrolladores se enfrentan a esto, lo que es una buena noticia
      - Gèant4, ROOT, librerías matemáticas, …

Evolución de la red (Network)

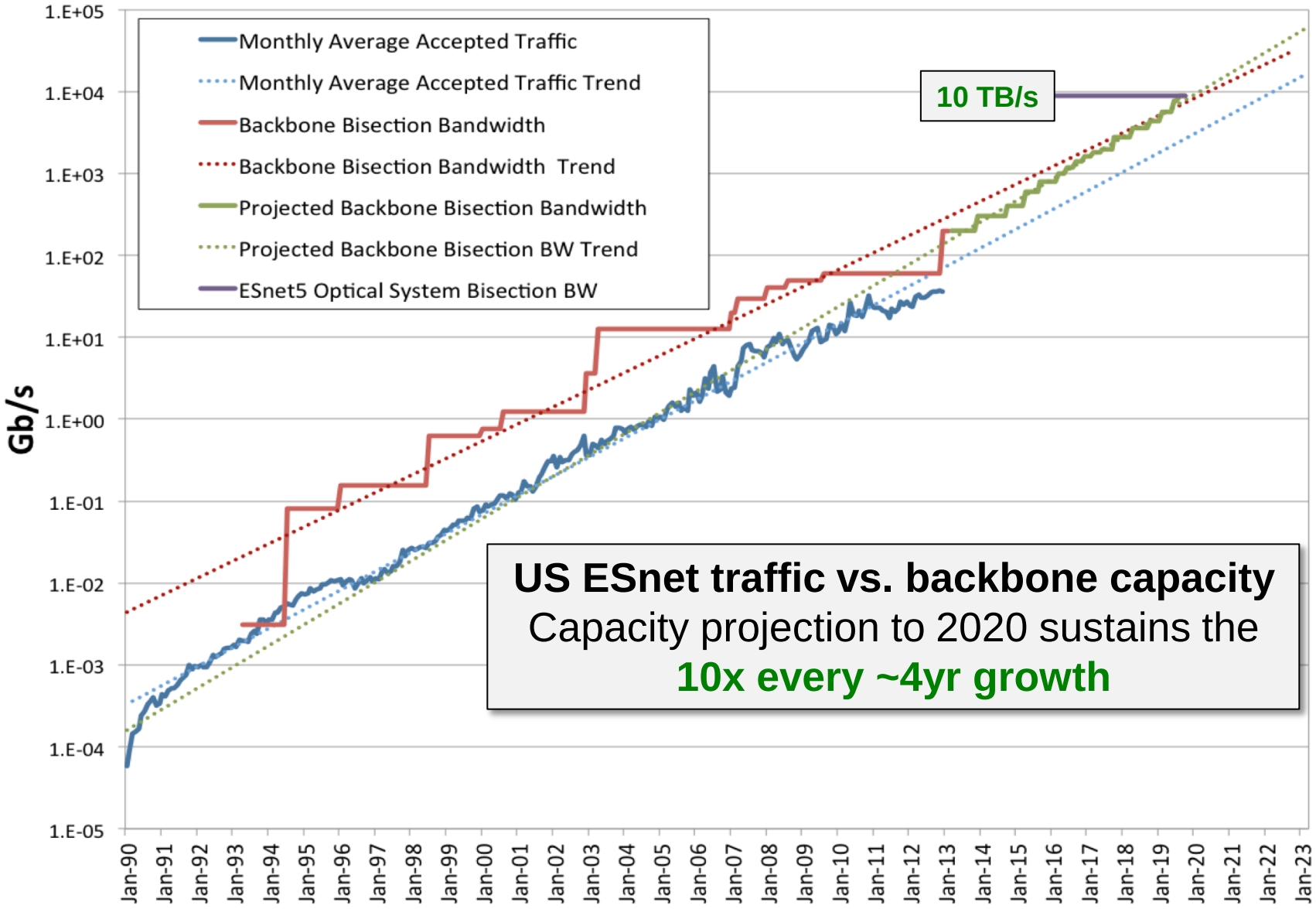
- HEP ha sido **pionera** en el uso intensivo de las redes internacionales de investigación, y continúa a la cabeza
- Redes optimizadas para **flujos masivos de datos**
  - Probando el primer enlace en producción transatlántico a 100 Gbps
- Sacar el **máximo partido** de la red → ciencia a un coste menor!
  - Importante diseñar los flujos de trabajo en torno a este hecho
- Redes de próxima generación permiten que las aplicaciones interactúen con la red y podamos usarla aún más eficientemente (**OpenFlow**)

In general it's much cheaper to transport data than to store it

T. Wenaus @ CHEP'13

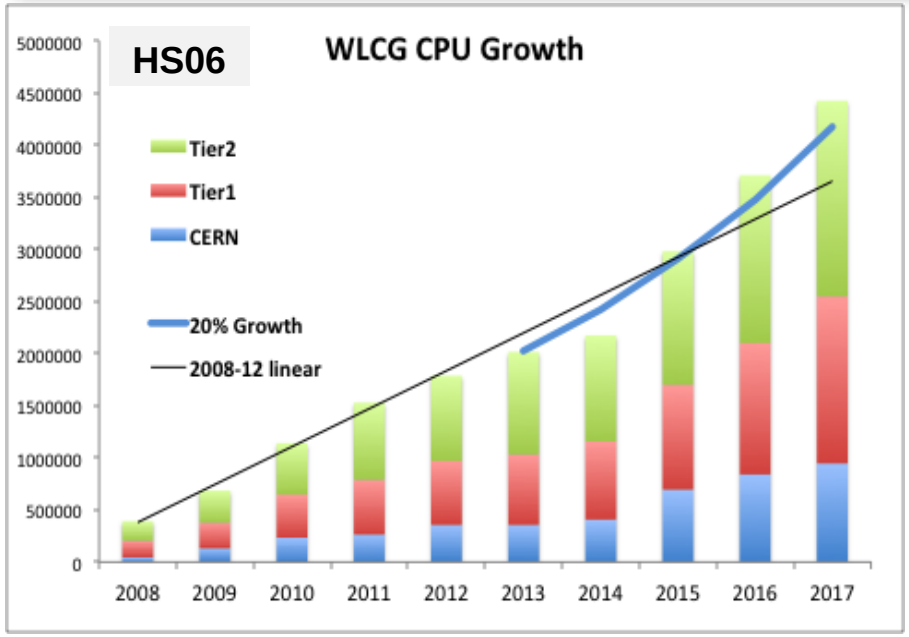
Spotify

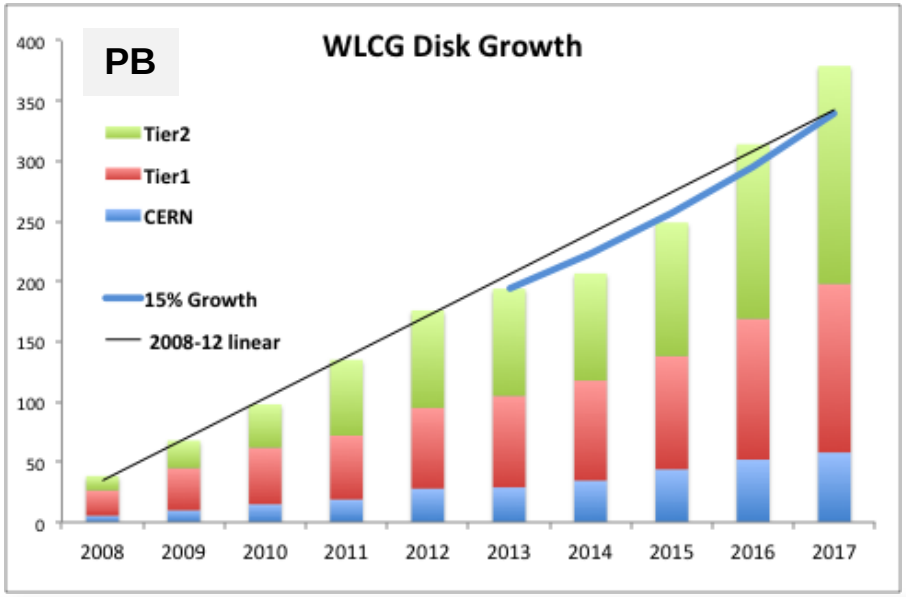
Evolución de la red (Network)



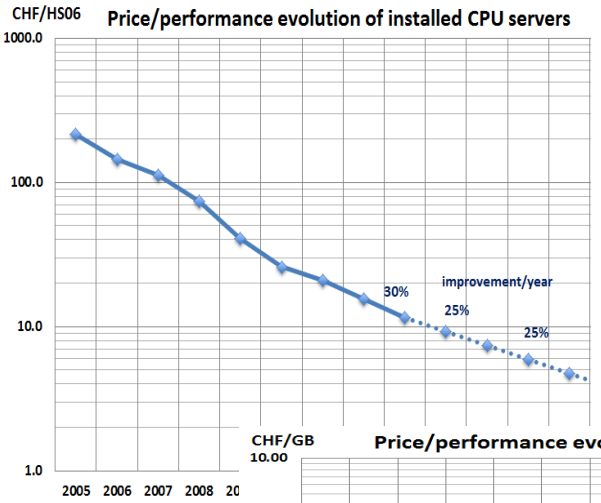
US ESnet traffic vs. backbone capacity
Capacity projection to 2020 sustains the
10x every ~4yr growth

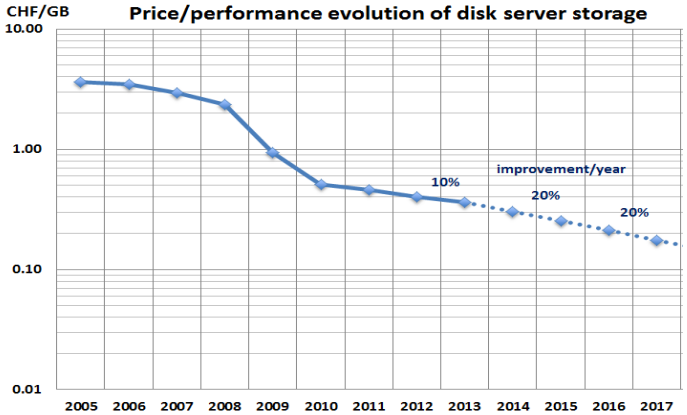
Incremento recursos de Computing



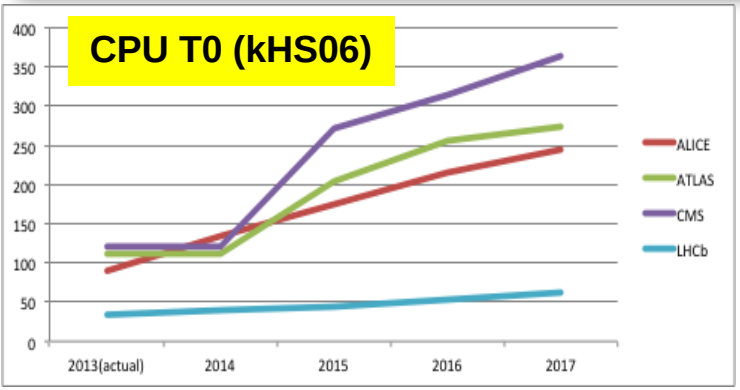


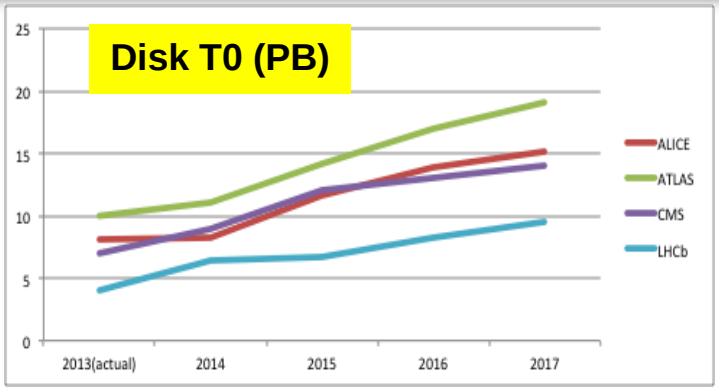
Gracias a la evolución tecnológica y el impacto de la actualización del Computing (draft), *con una financiación como la actual se prevé que se podrían proporcionar los recursos futuros*

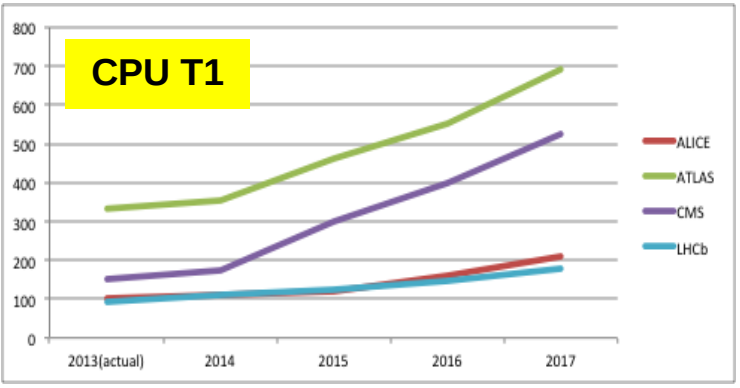


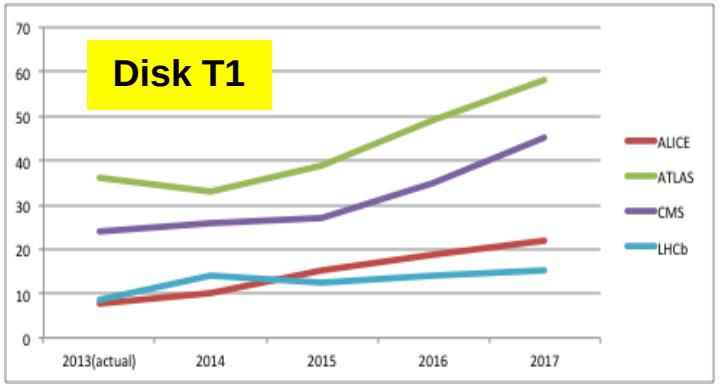


Incremento recursos de Computing





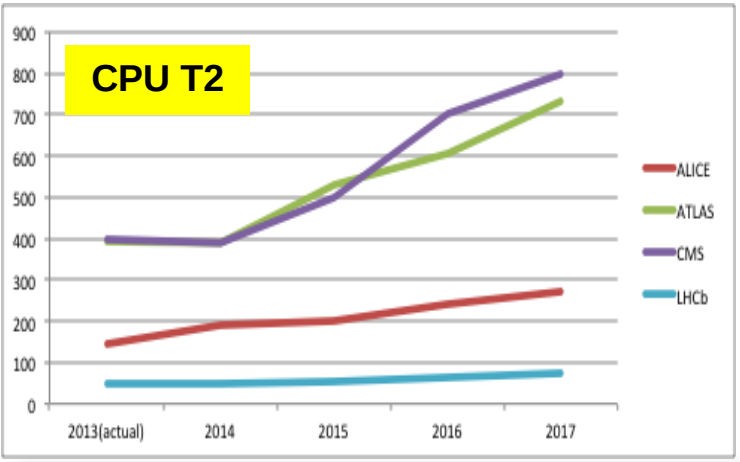


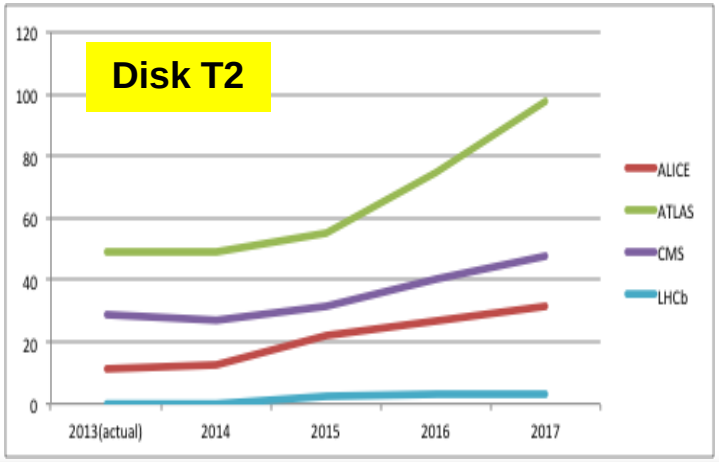


Gran aumento a partir de 2015

Gran aumento de la CPU para CMS en el T0

Gran petición de disco por parte de ATLAS





Recursos futuros en el PIC

Pledges 2014-2017

		2013*	2014**	2015**	2016***	2017***	
CPU (HS06)	ATLAS	16,269	18,105	22,950	28,152	35,241	
	CMS	8,415	8,925	15,300	20,400	26,775	
	LHCb	5,610	7,150	7,995	9,620	11,505	
	Total	30,294	34,180	46,245	58,172	73,521	x2.5
Disk (TB)	ATLAS	1,683	1,683	1,887	2,499	2,958	
	CMS	1,326	1,326	1,326	1,785	2,295	
	LHCb	439	910	813	923	1,001	
	Total	3,448	3,919	4,026	5,207	6,254	x2
Tape (TB)	ATLAS	2,040	2,244	3,315	4,284	5,508	
	CMS	2,601	2,805	3,749	5,100	6,885	
	LHCb	551	722	1,352	1,983	2,716	
	Total	5,192	5,771	8,416	11,367	15,109	x3

** 2013: Confirmed and approved by the Scrutiny Group at the C-RRB of 30th October 2012* [18]
*** 2014/2015: Confirmed and approved by the Scrutiny Group at the C-RRB of 29th October 2013* [19]
**** 2015/2017: Tentative estimates from "Computing Model Update" document Ref. CERN-RRB-2013-089* [20]

Conclusiones

- El sistema de computación distribuido del LHC ha funcionado **extremadamente bien** durante el Run1
  - Sabemos cómo entregar, adaptando el sistema si es necesario
- Las **redes** son excelentes, los **modelos de computación** son flexibles y suficientemente adaptables para explotar los recursos de computación disponibles
- El sistema de computación necesita afrontar nuevos **retos**
  - Gran **aumento** de los recursos necesarios a partir de 2015
  - Contexto de **presupuestos limitados**
- Utilizar los recursos de la forma más **eficiente** posible
  - Por ejemplo, el disco… uno de los recursos más caros
  - Para ello, se prevé un importante programa de desarrollo

Conclusiones

- Acceso a **recursos oportunistas** y **Cloud Computing**
- Exploración de **nuevas arquitecturas** de computación y procesamiento
- Evolucionar hacia un **acceso de datos** más dinámico y **computación paralela** distribuida
  - Gracias a la implantación de redes de alta capacidad y estabilidad
  - Gracias a la mejor explotación de los procesadores multi-core
- Las extrapolaciones sin estas consideraciones conducen a costes inaceptables. **Objetivo:** evolucionar hacia un sistema más dinámico, eficiente y flexible, **minimizando costes**…

… o …

Gracias!

Dr. José Flix Molina (PIC/CIEMAT)

jflix@pic.es

@JosepFlixMolina @pic_es

PIC WLCG Tier-1 Project Manager
WLCG Operations Co-coordinator
CMS Resource Management Coordination