Facilidad de Análisis:
ATLAS TIER3 en Valencia

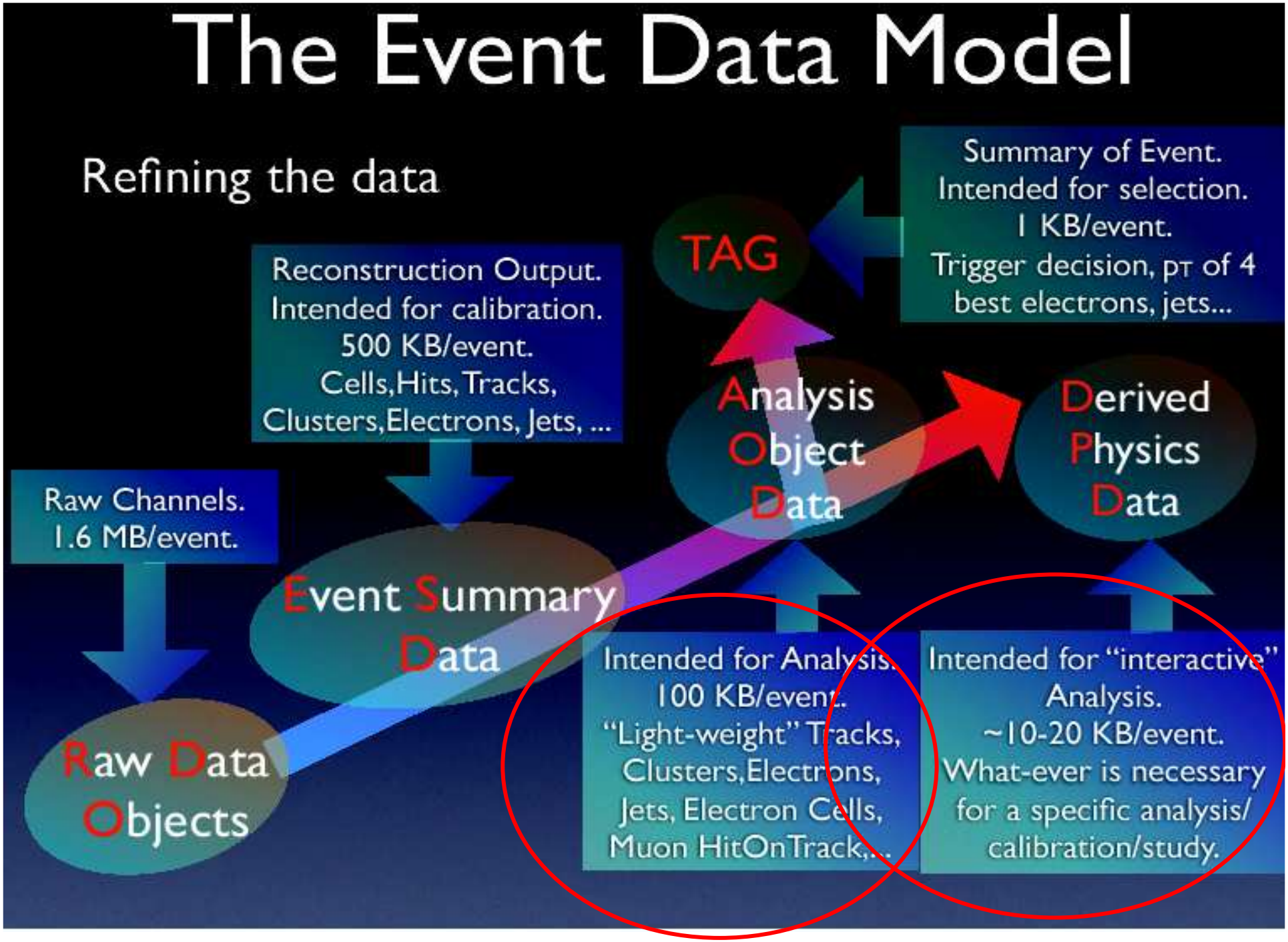
S. González de la Hoz

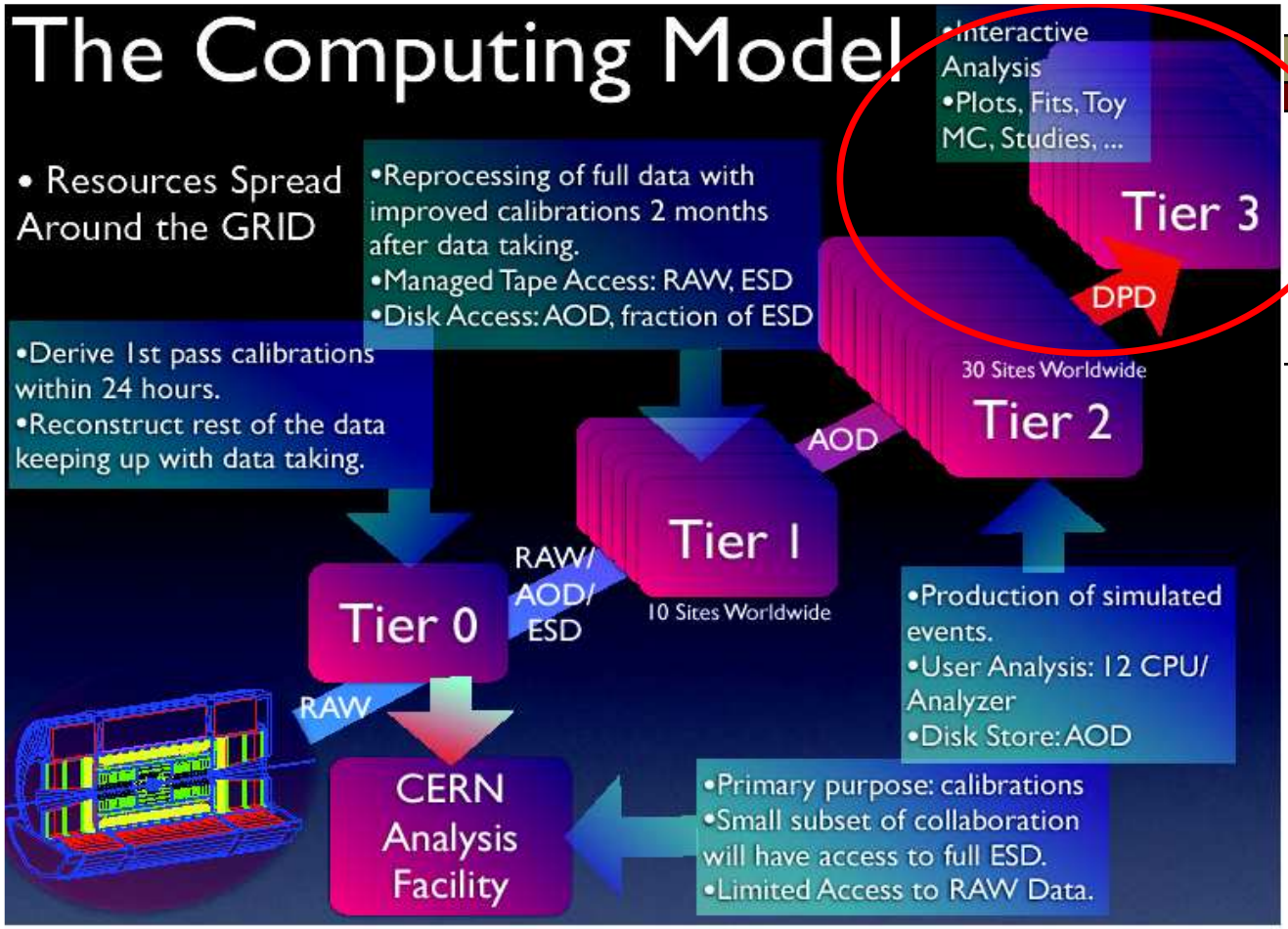
IFIC – Instituto de Física Corpuscular de Valencia

II Reunión ATLAS-Valencia

Modelo Jerárquico de Tiers

- **Event Filter Farm at CERN (EF)**
  - Situada al lado del Experimento, transforma los datos en un *stream* hacia el Tier 0
- **Tier 0 Center at CERN**
  - Raw data ➔ Se almacenan en el CERN y en los Tier 1
  - Rápida producción de Event Summary Data (ESD) y Analysis Object Data (AOD)
  - Envío de ESD, AOD a los Tier 1 ➔ Mass storage at CERN (Tier 1)
- **Tier 1 Centers distributed worldwide (10 centers, PIC-Barcelona)**
  - Re-reconstrucción de los raw data, producción de nuevos ESD, AOD (entre ~2 meses después de la llegada y un año)
  - Acceso a los ESD y AOD
- **Tier 2 Centers distributed worldwide (~30 centers, Spanish Tier2 Federation: IFIC-Valencia, IFAE- Barcelona y UAM-Madrid)**
  - Almacenamiento de algunos dataset (ESD, AOD) en función de la demanda de los grupos de análisis y de los detectores.
  - Simulación de Montecarlo, produciendo ESD, AOD ➔ Tier 1
- **CERN Analysis Facility (CAF)**
  - Fácil acceso a los datos ESD y RAW/calibración
  - Calibración, optimización del detector, algunos análisis - vital en las primeras etapas. **Focalizada a la calibración.**
- **Tier 3 Centers distributed worldwide (Responsabilidad de cada centro). No definido en el modelo (no hay un MOU como en el caso del Tier1 y Tier2)**
  - **Recursos locales "diferentes" a los utilizados en un Tier1 y Tier2 para:**
    - **Análisis de Física: Con ESD y AOD para desarrollo de código de análisis.**
    - **Análisis Interactivo: No requiere una conexión con los ESD o AOD, solo con los datos provenientes de DPD (Ntuplas).**


The Event Data Model
Refining the data
Summary of Event.
Intended for selection.
1 KB/event.
Trigger decision, pT of 4
best electrons, jets...
TAG
Reconstruction Output.
Intended for calibration.
500 KB/event.
Cells,Hits, Tracks,
Clusters,Electrons, Jets, ...
Analysis
Object
Data
Derived
Physics
Data
Raw Channels.
1.6 MB/event.
Event Summary
Data
Raw Data
Objects
Intended for Analysis.
100 KB/event.
"Light-weight" Tracks,
Clusters,Electrons,
Jets, Electron Cells,
Muon HitOnTrack,...
Intended for "interactive"
Analysis.
~10-20 KB/event.
What-ever is necessary
for a specific analysis/
calibration/study.



http://twiki.cern.ch/twiki/pub/Atlas/AODFormatTaskForce/AODFormatTF_Report.pdf

- AOD - Input para el análisis con Athena
- **DPD** – Derive Physics Data. Output del análisis con Athena. Input para el análisis interactivo utilizando ROOT (interés por parte de la comunidad de físicos).
  - Típicamente una "ntupla"
  - Prácticamente son copias de los datos de los AOD
- AOD/DPD Merger: Se usará la misma tecnología
  - Idea es que AOD se puedan analizar con ROOT
  - Y que DPD se puedan analizar también con Athena

Table 2-2 The assumed event data sizes for various formats, the corresponding processing times and related operational parameters.

Item	Unit	Value
Raw Data Size	MB	1.6
ESD Size	MB	0.5
AOD Size	kB	100
TAG Size	kB	1
Simulated Data Size	MB	2.0
Simulated ESD Size	MB	0.5
Time for Reconstruction (1 ev)	kSI2k-sec	15
Time for Simulation (1 ev)	kSI2k-sec	100
Time for Analysis (1 ev)	kSI2k-sec	0.5
Event rate after EF	Hz	200
Operation time	seconds/day	50000
Operation time	days/year	200
Operation time (2007)	days/year	50
Event statistics	events/day	10^7
Event statistics (from 2008 onwards)	events/year	$2 \cdot 10^9$

- AOD Event/year:
  - **200TB**
- DPD Event/year
  - 10KB/event → **20TB**

- El Tier2 español en 2008 dispondrá de 387 TB
  - ¿Suficiente para nuestro análisis?
- **Para el modelo de análisis:**
  - **¿Tenemos que almacenar todos los AODs?**
  - **¿Cuantas versiones de ellos?**
  - **¿Y cuántos ESD?**
  - **¿Y los datos de Monte Carlo?**
  - **¿Y los DPD?**

Spanish ATLAS T-2 assuming a contribution of a 5% to the whole effort

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

Resumen del ATLAS Tier3 workshop (Enero 2008)

https://twiki.cern.ch/twiki/bin/view/Atlas/Tier3TaskForce

- Difícil de definir un Tier3. Diferentes tipos
  - Si tiene un Tier2 cerca o no, personal, etc..
  - Su tamaño (núm. Físicos, recursos, etc..)
- Parece que tiene que ser una mezcla entre infraestructura Grid y local/interactiva
- El Tier2 almacenaría los AOD mientras que el Tier3 los DPD y estos se analizan con ROOT
- Estimación:
  - 25 cores (~12000 euros) y 25 TB (~22000 euros) por análisis (no por usuario). **En total 34000 euros por análisis**.
    - 30-10 KB/event (10^9/year) → 1 mes se podrían analizar todos
- Conectado al Grid: SE basado en SRM, CE y UI
  - Si hay CE se puede hacer instalación del software de ATLAS automáticamente. Sino a mano
  - En cualquier caso se necesita un sistema de ficheros compartidos

Resumen del ATLAS Tier3 workshop (Enero 2008)

- Hardware y SO: Software de ATLAS funciona bien para SL(C)4
- Instalación: Depende del tamaño del centro. Si es pequeño a "mano" si es grande se necesita una herramienta de mantenimiento e instalación (Ej. Quattor, etc..)
- Interactividad: PROOF y un sistema de ficheros compartidos (escalabilidad buena si hay 5-10 análisis en un instituto)
- Almacenamiento (escalable y *reliable*):

Storage System	Local Protocol	Load Balancing	Externally Secure	POSIX Access	Single Namespace	Installation Load	Maint Load	Quotas	Cost
NFS	bad	N	N	Y	N	low	high	Y	$0
Lustre	Y	Y	w/SRM	Y	Y	medium	medium	Y	$0
GPFS	Y	Y	w/SRM	Y	Y	high	medium	Y	$$$
xrootd	Y	Y	w/SRM	mkdir/rmdir do nothing	Y	medium	low	partitions	$0
DPM	Y	Y	Y	special commands	Y	medium-high	low-medium	partitions	$0
dCache	Y	Y	Y	metadata	Y	high	low-medium	partitions	$0

¿Recursos en los Centros o en el CERN? Discutido en diversas reuniones T1-T2

- En el CERN siempre se pueden poner recursos. En principio sólo es cuestión de poner dinero, ellos gestionan esos recursos
  - Entonces, ¿por qué este modelo de *computing* jerárquico?
    - ¿Por qué no están todos los Tier-1s y todos los Tier-2s en el CERN?
- Si algún día se decide crear y mantener una infraestructura en el centro, cuanto antes se empiece a ganar experiencia mucho mejor
  - Yo gestiono estos recursos, en el CERN no.
- ¿Qué pasa con los Físicos de nuestro centro en el CERN?
  - ¿De cuánta gente estamos hablando?
    - Que utilicen los recursos que les da el CERN
    - Que se conecten a los recursos de su centro
- ¿Qué pasa con los Físicos que se quedan en su Centro?
  - Que utilicen los recursos de su centro
  - Que se conecten a los recursos del CERN

Requisitos mínimos para un Tier3

- The ATLAS software environment, as well as the ATLAS and grid middleware tools, allow us to build a work model for collaborators who are located at sites with low network bandwidth to Europe or North America.
- The minimal requirement is on local installations, which should be configured with a Tier-3 functionality:
  - A Computing Element known to the Grid, in order to benefit from the automatic distribution of ATLAS software releases
  - A SRM-based Storage Element, in order to be able to transfer data automatically from the Grid to the local storage, and vice versa
- The local cluster should have the installation of:
  - A Grid User Interface suite, to allow job submission to the Grid
  - ATLAS DDM client tools, to permit access to the DDM data catalogues and data transfer utilities
  - The Ganga/pAthena client, to allow the submission of analysis jobs to all ATLAS computing resources

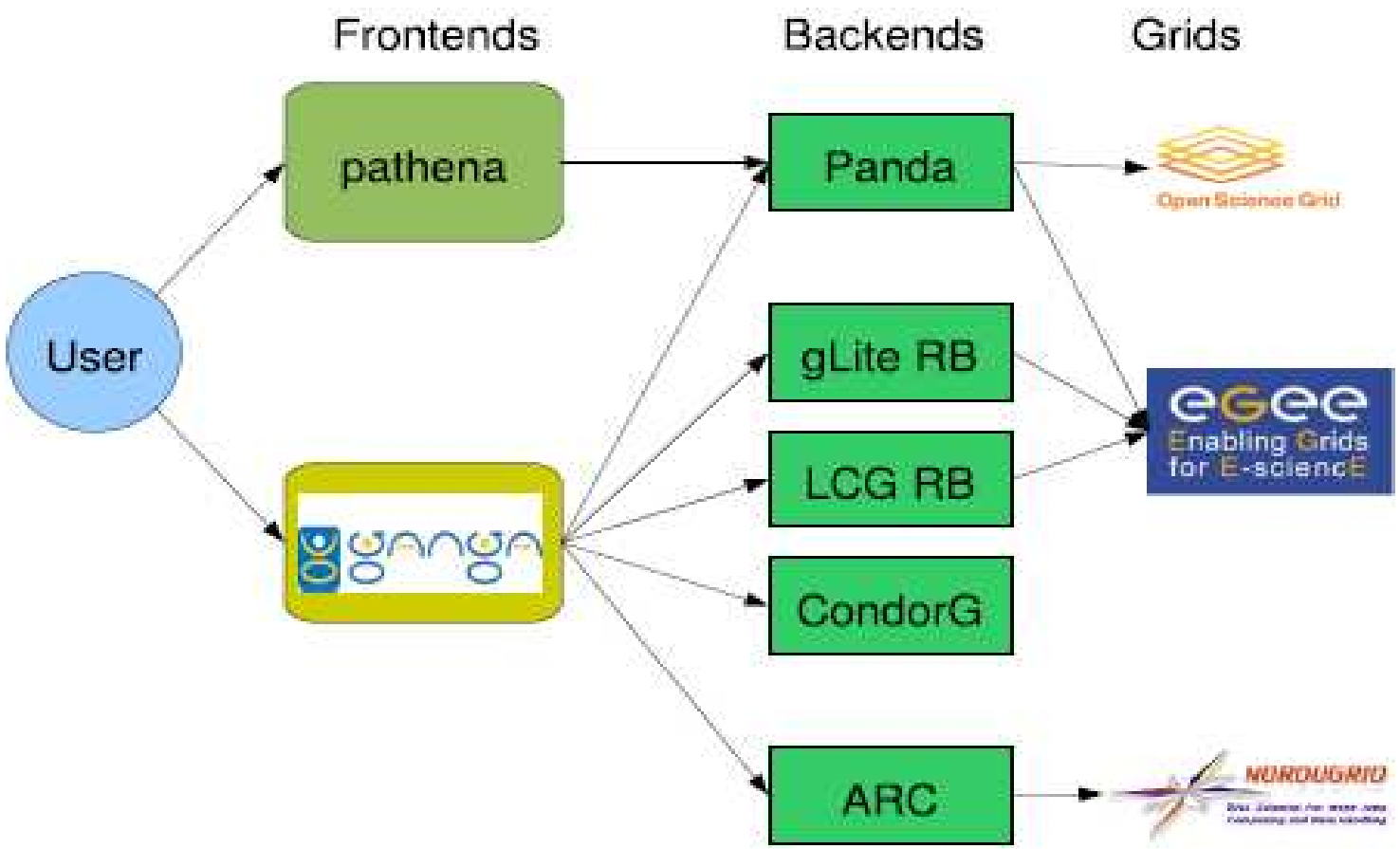
Nos lo da nuestro Tier2

Tenemos del Tier2, tendría que haber uno para el Tier3

Instalado en el IFIC

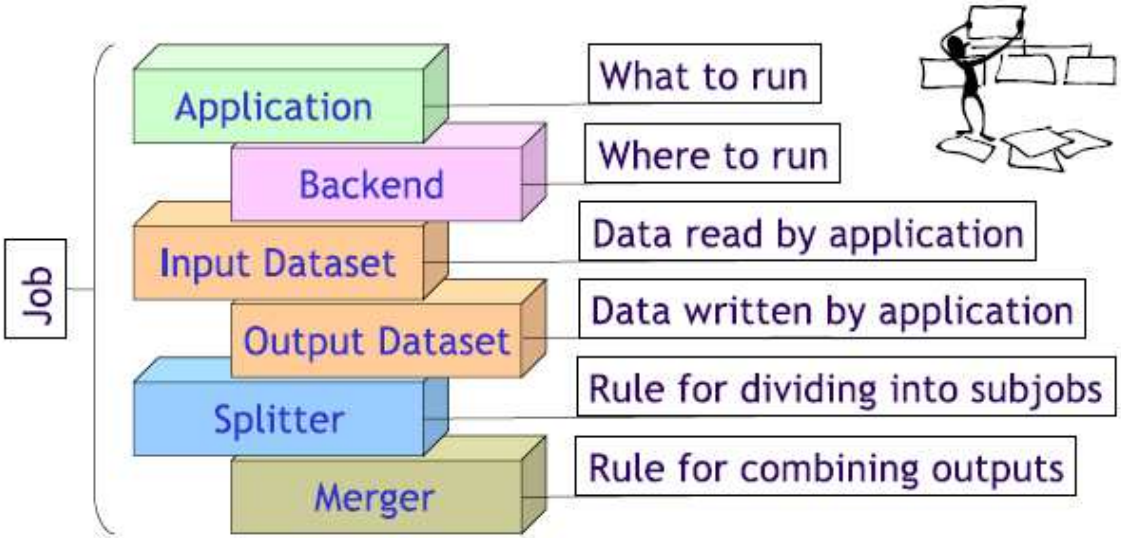
Dario Barberis: The ATLAS Computing Model

Situación actual del Análisis distribuido (utilizando el GRID)



Ganga

https://twiki.cern.ch/twiki/bin/view/Atlas/DistributedAnalysisUsingGanga



Job definition within GANGA IPython shell:

```
j = Job()
j.application=Athena()
j.application.prepare(athena_compile=False)
j.application.option_file='$HOME/athena/12.0.5/InstallArea/jobOptio
j.splitter=AthenaSplitterJob()
j.splitter.numsubjobs = 3
j.merger=AthenaOutputMerger()
j.inputdata=DQ2Dataset()
j.inputdata.dataset='csc11.005145.PythiaZmumu.recon.AOD.v11004103'
j.inputdata.match_ce=True
j.outputdata=DQ2OutputDataset()
j.outputdata.outputdata=['AnalysisSkeleton.aan.root']
j.backend=LCG()
j.submit()
```

Instalado en el IFIC

- Cualquier PC puede ser un **User Interface (UI)** en AFS:
  - source /afs/ific.uv.es/sw/LCG-share/sl4/etc/profile.d/grid_env.sh
  - source /afs/ific.uv.es/sw/LCG-share/sl4/etc/profile.d/grid_env.csh

  - source /afs/ific.uv.es/sw/LCG-share/sl3/etc/profile.d/grid_env.sh
  - source /afs/ific.uv.es/sw/LCG-share/sl3/etc/profile.d/grid_env.csh

- Utilizar el cliente de **dq2(DDM)** en AFS:
  - source /afs/ific.uv.es/project/atlas/software/ddm/current/dq2.csh
  - source /afs/ific.uv.es/project/atlas/software/ddm/dq2_user_client/ setup.sh.IFIC
  - dq2-list-dataset-site IFIC

- Y el cliente de **Ganga**:
  - source /afs/ific.uv.es/project/atlas/software/ganga/install/etc/setup-atlas.sh
  - Source /afs/ific.uv.es/project/atlas/software/ganga/install/etc/setup-atlas.csh

- Por supuesto instalado en los UI (ui02 y ui03.ific.uv.es) del Tier2

The Express Stream of ATLAS Data

Algunos grupos han expresado su interés (ej Tilecal Valencia)

- Need to define what calibration and express streams the group would like to have replicated in Spain. Need to develop common strategy with this other Spanish groups and the Tier1
  - As TileCal experts the groups needs calibration streams.
    - TileCal community is in the process of defining the composition and specifics of calibration streams
      - Preliminarily: laser triggers or the charge injection triggers generated in the empty bunches. Use muon streams
  - At the same time, due to physics interests and for the sake of competitiveness, the group also needs access to express streams (see next slide)

Preliminary list of signatures for the Express Stream

List still under discussion and

Streams interesting for Physics/Detector performance preferences of the group

Decay or signature	Motivation
$Z \rightarrow l^+l^-$	calibration and data quality
minimum bias	data quality
lepton pair with high mass	alert on rare events
$B \rightarrow \mu^+\mu^-$	alert on rare events
≥ 3 high p_{T} leptons	alert on rare events
☑ lepton + jets + ETmiss	calibration
$W \rightarrow l\nu$	calibration and data quality
☑ large missing E_{T}	alert on rare events
lepton with large p_{T}	alert on rare events
☑ large ΣE_{T}	alert on rare events
large M_{eff}	alert on rare events
high multiplicity of trigger objects	alert on rare events

¿Qué son Express Stream?

http://indico.cern.ch/getFile.py/access?contribId=2&resId=0&materialId=0&confId=a06527

- Motivations for the Express Stream
  - Calibration
    - "Physics calibration data streams for rapid processing
  - Check of general data quality
    - Only sample that can be used for as rapid and complete check of the data quality
  - Rapid alert on interesting physics events
    - Monitoring about rare avents while the data are being taken.
- Pero Tienen una vida de 48 horas, después existe el Stream

- **¿El Tier-2 o Tier-3 tiene que tener datos de este tipo?**
  - **¿Se van a hacer análisis por parte de nuestros usuarios sobre este tipo de datos?**
  - **¿ Y los Stream of data?**

Prototipo Tier3 en el IFIC

Acordado en Septiembre 2007

<table>
<tr><td rowspan="2">Desktop
Or
Laptop
(I)</td><td>Atlas Collaboration Tier2 resources (Spanish T2)
(II)</td></tr>
<tr><td>Extra Tier2: ATLAS Tier3 resources (Institute)
(II)</td></tr>
<tr><td colspan="2">PC farm to perform interactive analysis (Institute)
(III)</td></tr>
</table>

Fase 1 (ordenador personal de cada usuario): HECHA!

- Desktop (Funcionalidad similar a la lxplus del CERN):

a) Software de ATLAS visible desde AFS-IFIC (Athena+Root+Atlantis+ …)

- **Permite correr trabajos de forma interactiva antes de enviarlos al Grid** (grandes producciones)

b) User Interface (UI) (Glite; middleware Grid)

- **Permite buscar datos del Grid y copiarlos en el propio ordenador**
- **Permite el envío de trabajos al Grid**

En lo que hemos trabajado hasta ahora, faltaría actualización

Fase 1 (ordenador personal de cada usuario):

- Otra posibilidad discutida en el ATLAS Tier3 task force:
(https://twiki.cern.ch/twiki/bin/view/Atlas/AtlasComputing?topic=Tier3TaskForce)

a) Instalar algunos User Interface y al menos un CE dedicados al Tier3:

- **Permite tener el software de ATLAS instalado automáticamente como en el Tier2**
- **El usuario se tiene que conectar a dicho UI**
- **Se instalarían las *releases* de producción, pero podrían instalarse de desarrollo bajo petición**
- **Ven Lustre (forma de lectura), AFS y permiten enviar trabajos al Grid**
- Existen 2 (ui02 y ui03.ific.uv.es) y se instalarán otros 2 en breve

b) El cliente de Ganga se tendría que seguir instalando en AFS

Fase 2 (Acoplamiento con el Tier2):

Aprovechar la experiencia de haber montado un Tier2

- Nominal:

a) Recursos para toda la colaboración, cumpliendo las especificaciones en TB (SE) y CPU (WN)

- Extra (Tier3):

a) **WNs y SEs de uso preferente por usuarios del IFIC**

b) **Recursos extras que aseguran disponibilidad (para ejecutar trabajos)**

c) Producciones privadas de AOD y DPD

d) Análisis de los AOD utilizando el GRID (correr sobre millones de sucesos)

!!!!!

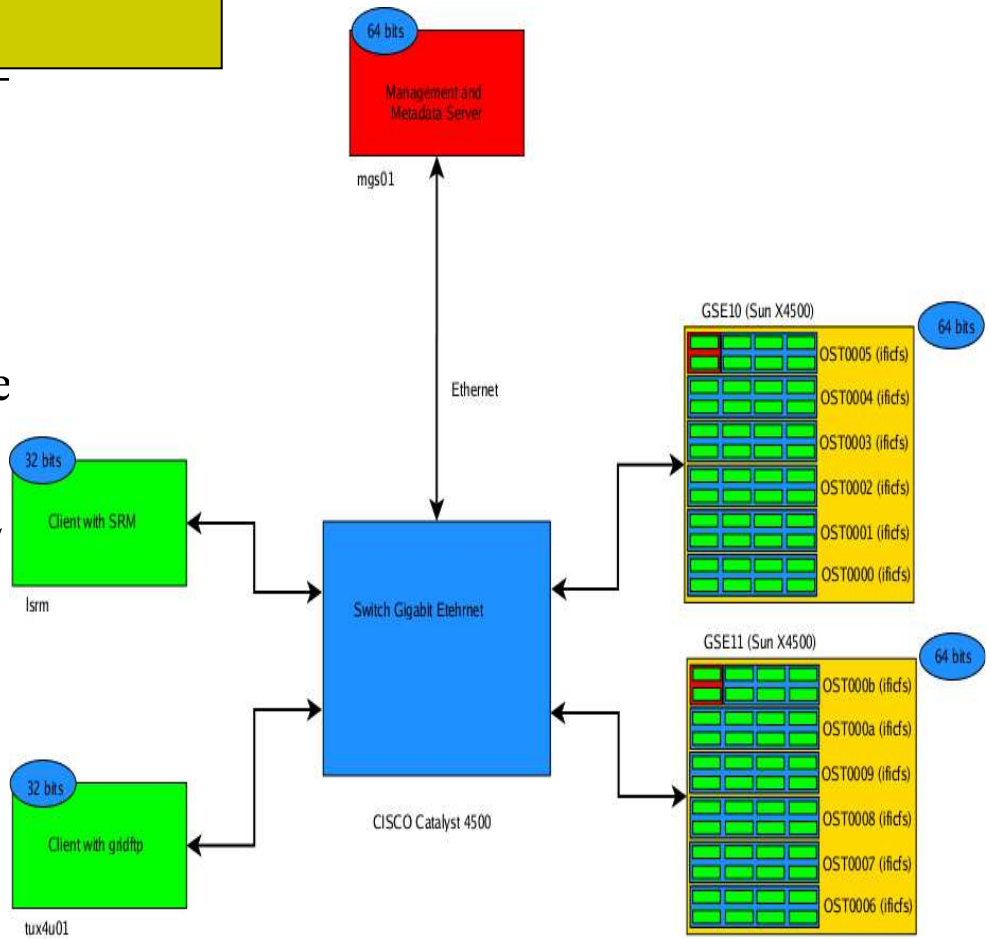
Fase 3 (Requisitos especiales):

- Análisis interactivo de DPD que se pueden analizar con ROOT
- Instalar granja **PROOF** (Root en paralelo):
  a) Granja fuera del Grid
  b) De unos pocos trabajadores (~20)
     - Menor tiempo de ejecución aprovechando paralelismo
  c) Buena conexión al sistema de almacenamiento.
     - Acceso rápido a los datos
     - Misma tecnología que el que se utiliza en el Tier2 (Lustre)

StoRM + Lustre

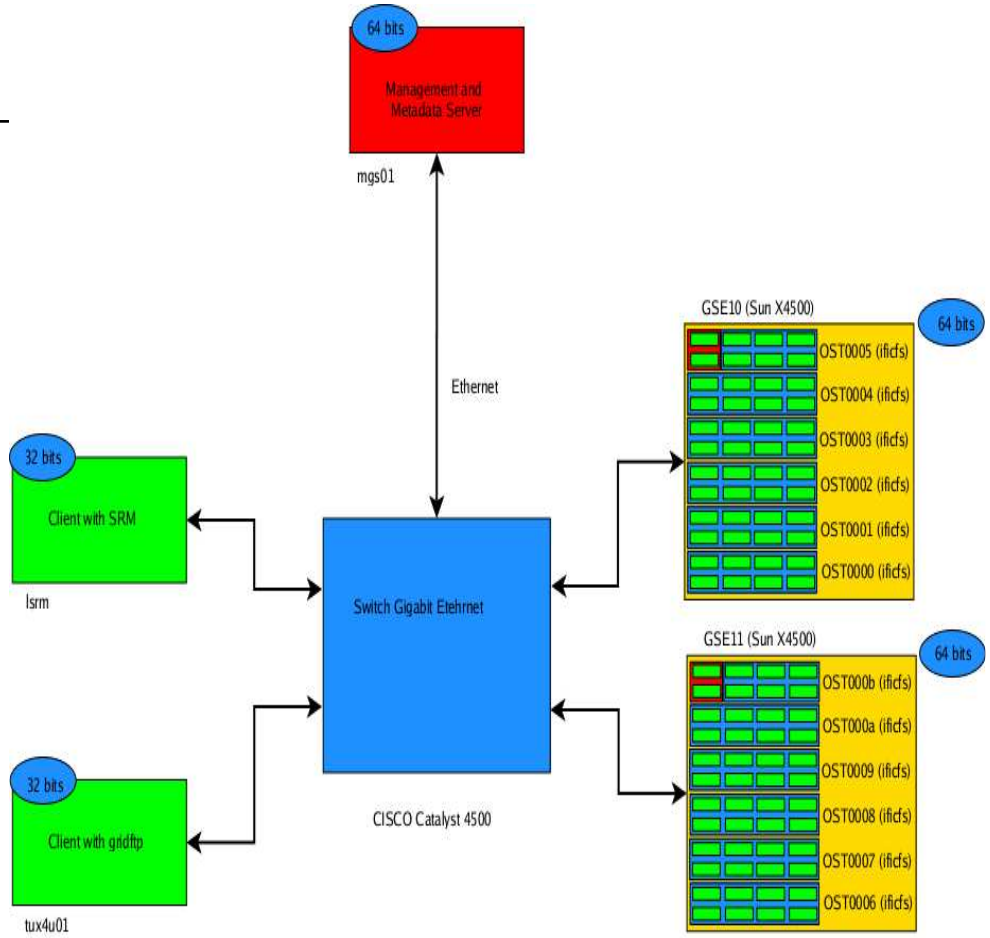
Como está en el TIER2

- StoRM
  - Posix SRM v2
  - Under testing. Being used in preproduction farm.
  - Temporally using UnixfsSRM (dCache SRM v1) for production in Tier2
- Lustre in production in our Tier2
  - High performance file system
  - Standard file system, easy to use
  - Higher IO capacity due to the cluster file system
  - Used in supercomputer centers
  - Free version available
  - Direct access from WN
  - www.lustre.org



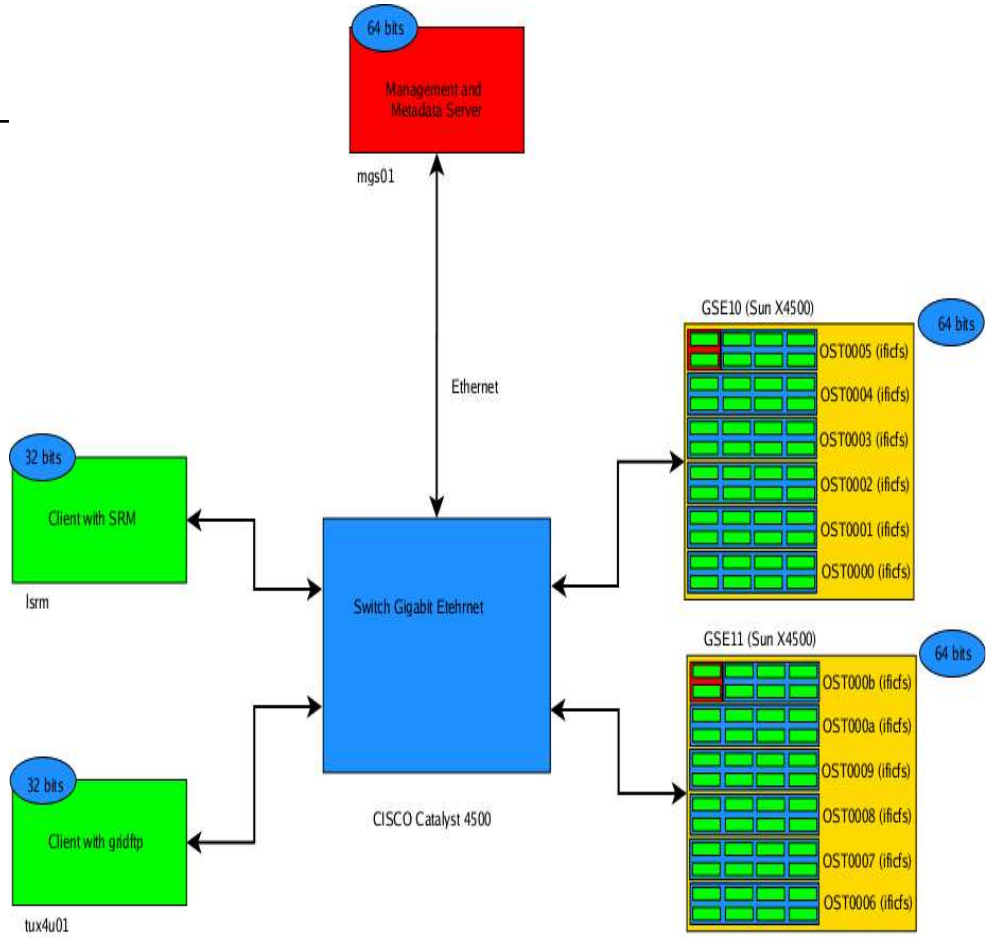
StoRM + Lustre

- Hardware
  - Disk servers:
    - 2 SUN X4500 (two more in place to be installed in the near future, used for testing)
    - 34 TB net capacity
  - Connectivity:
    - Switch Gigabit CISCO Catalyst 4500
  - Grid Access:
    - 1 SRM server (P4 2.7 GHz, GbE)
    - 1 GridFTP server (P4 2.7 GHz, GbE)
  - Lustre Server:
    - 1 MDS (Pentium D 3.2 GHZ, R1 disk)



StoRM + Lustre

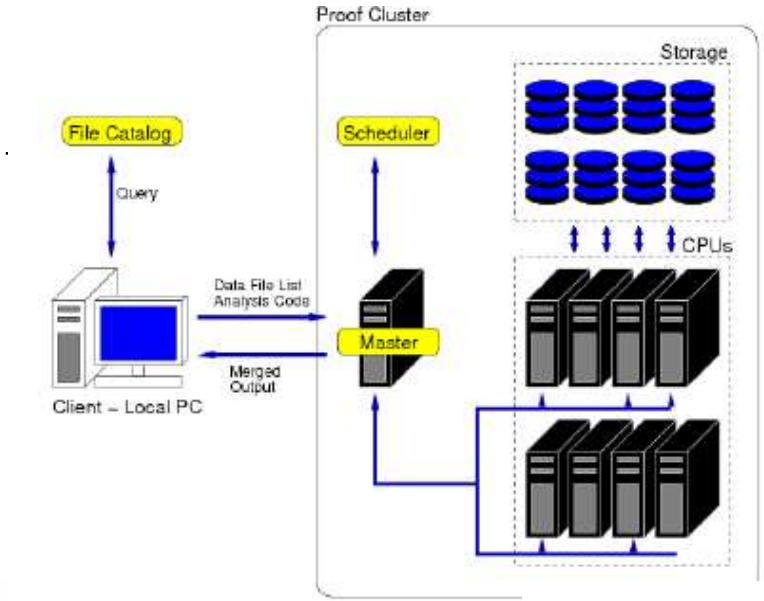
- Plans
  - Put StoRM in production (SRM v2)
  - Add more gridftp servers as demand increases
  - Move the Lustre server to a High Availability hardware
  - Add more disk to cope with ATLAS requirements and use
  - Performance tuning

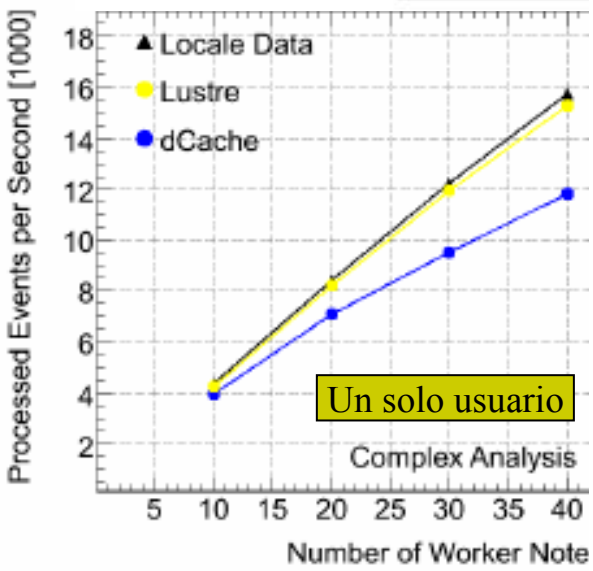


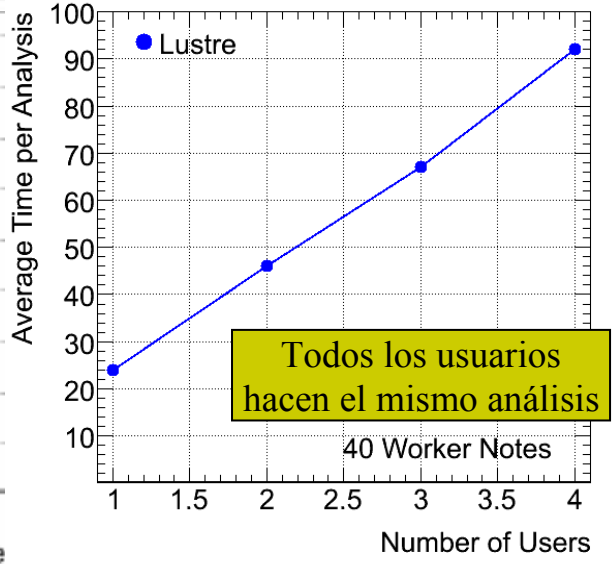
Granja fuera del GRID

- Hemos comprado dos máquinas para instalar PROOF y hacer pruebas con Lustre:
  - PE1950 III 2 Quad-Core Xeon E5430 2.66GHz/2x6MB 1333FSb
  - 16 GB de memoria (2GB por core; 8x2GB)
  - 2 discos de 146 GB (15000 rpm)
- Bajo test de rendimiento de disco y CPU
  - Disco RAID0 (*Data Stripping*): Datos distribuidos entre los discos de manera equitativa. Aumenta el rendimiento tanto de lectura como de escritura.
    - Hardware y Software

PROOF con Lustre en Munich

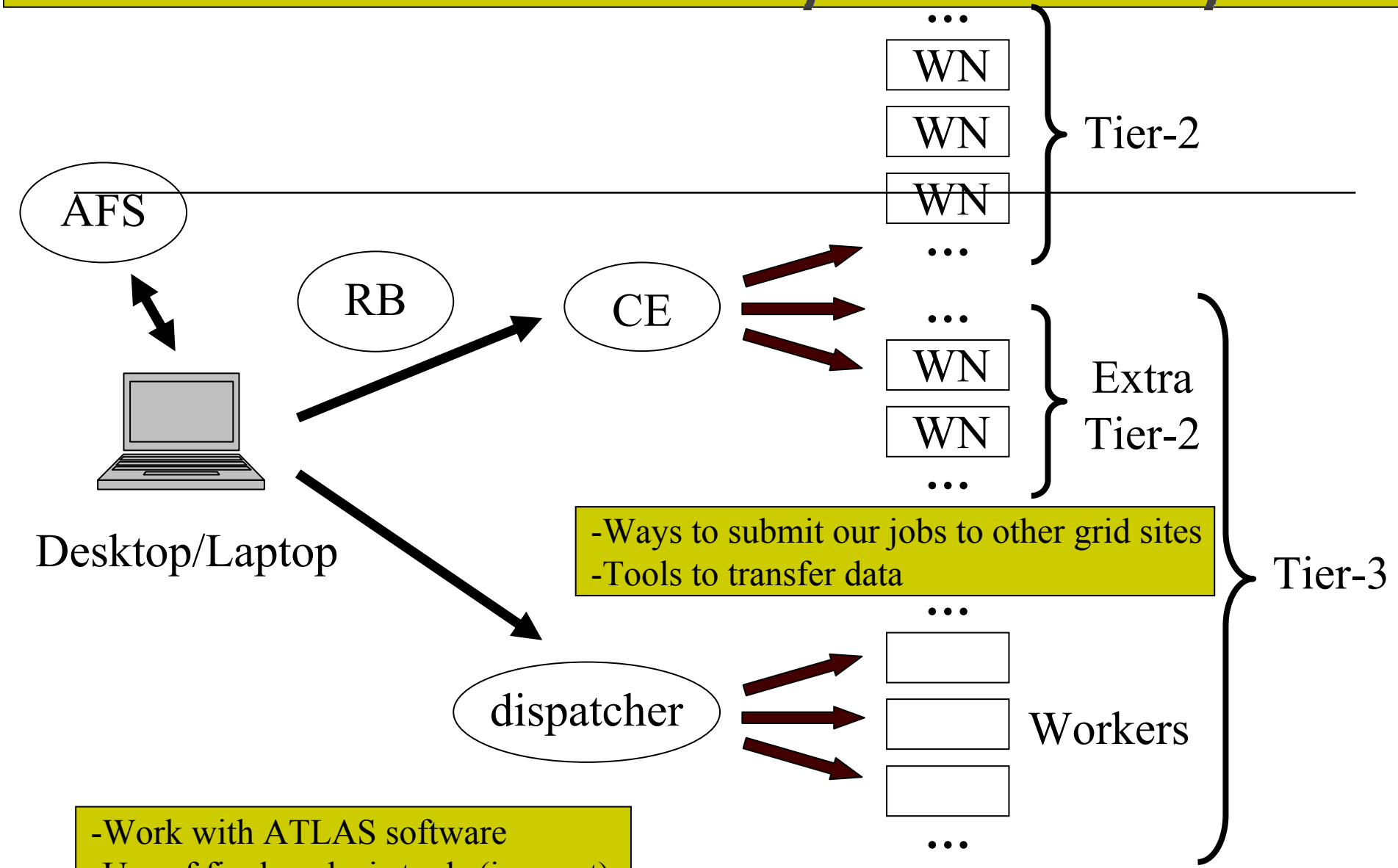






- Producimos DPD desde las AOD y se analizan con PROOF
- PROOF administra el paralelismo en nuestro *cluster* local
- ~~10 nodos utilizados y cada uno~~:
  - 2 dual core processor
  - 2.7 GHz y 8 GB RAM
- Dataset ($Z \rightarrow e^+e^-$) con 1.6 millones de sucesos
- Cada suceso tenia un tamaño de 4 KB, en total unos 6GB
- Los datos se guardaron localmente en cada nodo, utilizando los sistemas de ficheros Lustre y d-Cache
- Número de sucesos procesados en función del número de trabajadores utilizados.
- Lustre muestra un resultado equivalente al almacenamiento local
- Se observa que el rendimiento aumenta de forma lineal hasta los 40 procesadores
- Como se esperaba el tiempo para hacer un análisis es proporcional al número de usuarios

IFIC Valencia Analysis Facility



Conclusiones

- Los Tier3s:
  - Se utilizarán mayoritariamente para análisis tanto interactivo como en *batch* de los datos DPD.
  - El formato de estos DPD se está todavía discutiendo dentro del grupo "ATLAS Analysis Model".
  - Recursos locales, diferentes a los Tier1 y Tier2.
  - El tamaño de estos "centros" pude ser tan pequeño como un PC de despacho o tan grande como una granja Linux.
  - Apoyo por parte de los Tier1 y Tier2 en término de experiencia (instalación, configuración, etc. de las versiones del software de ATLAS y *middleware* Grid) y servicios (data storage, data service, etc.) es fundamental.
  - En el IFIC de Valencia estamos trabajando en la creación de esta infraestructura a nivel de configuración y software que se adecue principalmente a las necesidades de análisis de dichos DPDs.

Ejemplo de uso de un Tier3
(basado en la experiencia de Luis, Elena y Miguel)

- Análisis interactivo de Ntuplas.
  - No es necesario acceso a los datos desde donde estas Ntuplas se han generado
- Desarrollo de código de análisis.
  - Se necesita un copia local de una pequeña cantidad de sucesos ESD, AOD o quizás RAW
- Correr pequeñas pruebas locales antes de en enviar una gran cantidad de trabajos a los Tier1s o Tier2s utilizando el Grid.
  - También necesito una pequeña cantidad de datos copiados localmente, igual que el caso anterior.
  - Incluso igual necesito tener acceso a los TAG data
- Correr trabajos vía Grid en los Tier1s o Tier2s pero copiando los AOD (o quizás raramente los ESD) en el Tier3 para un posterior análisis.
- Analizar los anteriores AOD usando Athena (Con Ganga).
- Producción de muestras privadas de Monte Carlo de interés especial para los análisis que se lleven a cabo en el instituto.

Backup

ATLAS Spanish Tier2

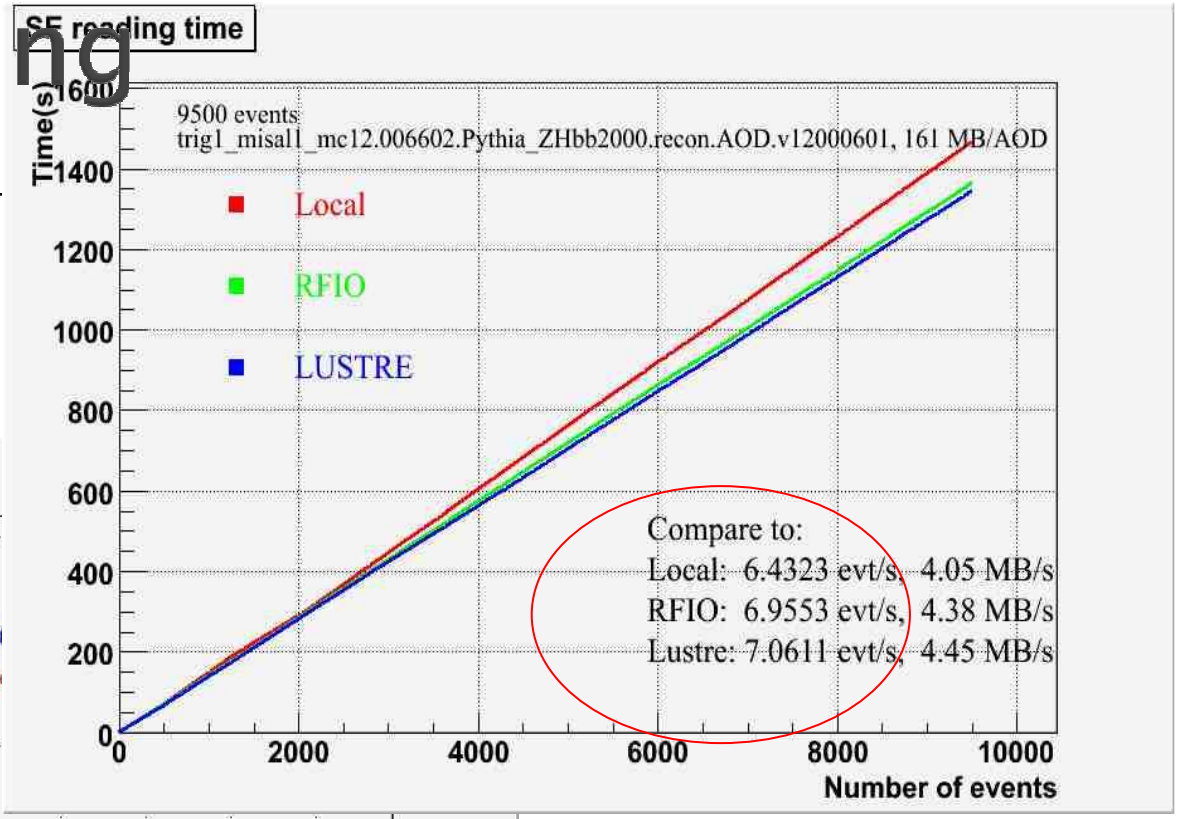
- Distributed Tier2: UAM(25%), IFAE(25%) and IFIC(50%)

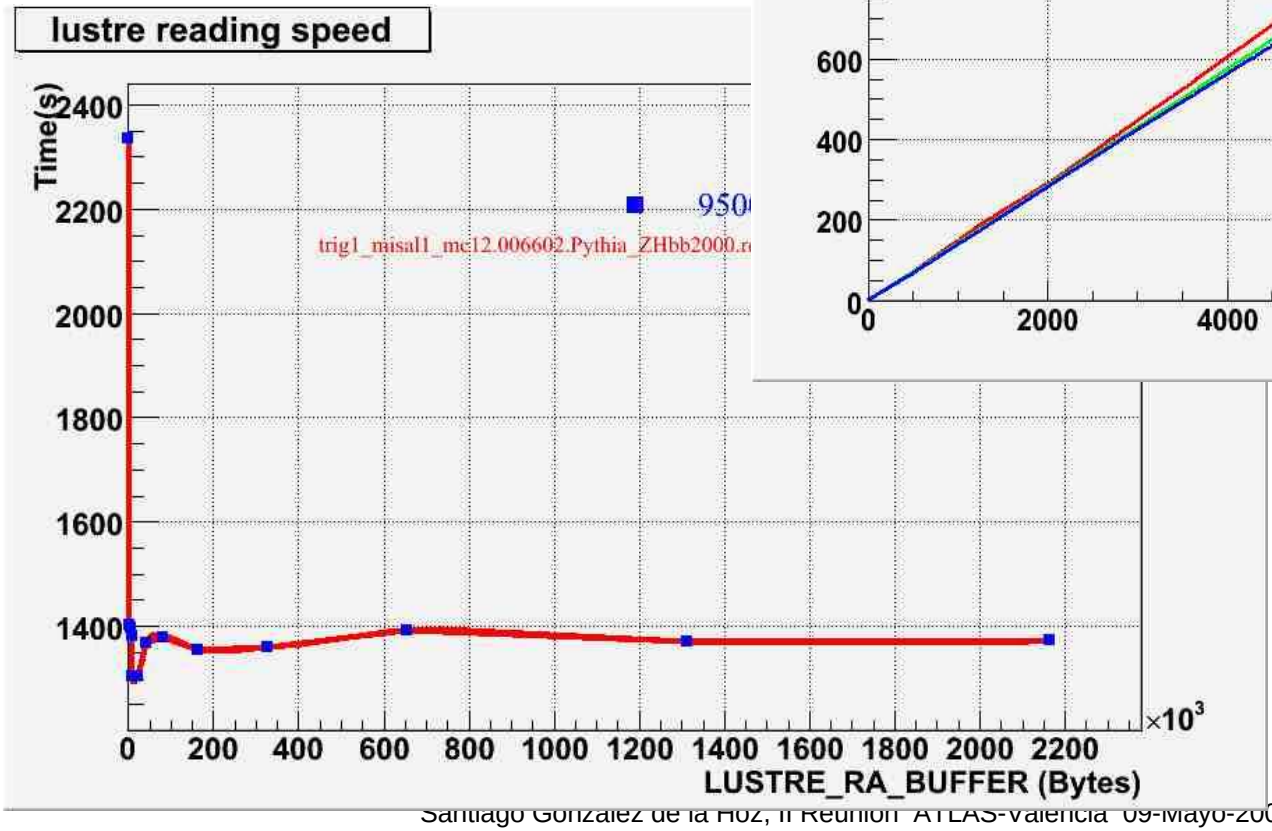
	SE
IFIC	Lustre+StoRM
IFAE	dCache/disk+SRM posix
UAM	dCache

- Inside our Tier2 two SE options are used. In case that Lustre won't work as expected we will switch to dCache

Lustre Tuning

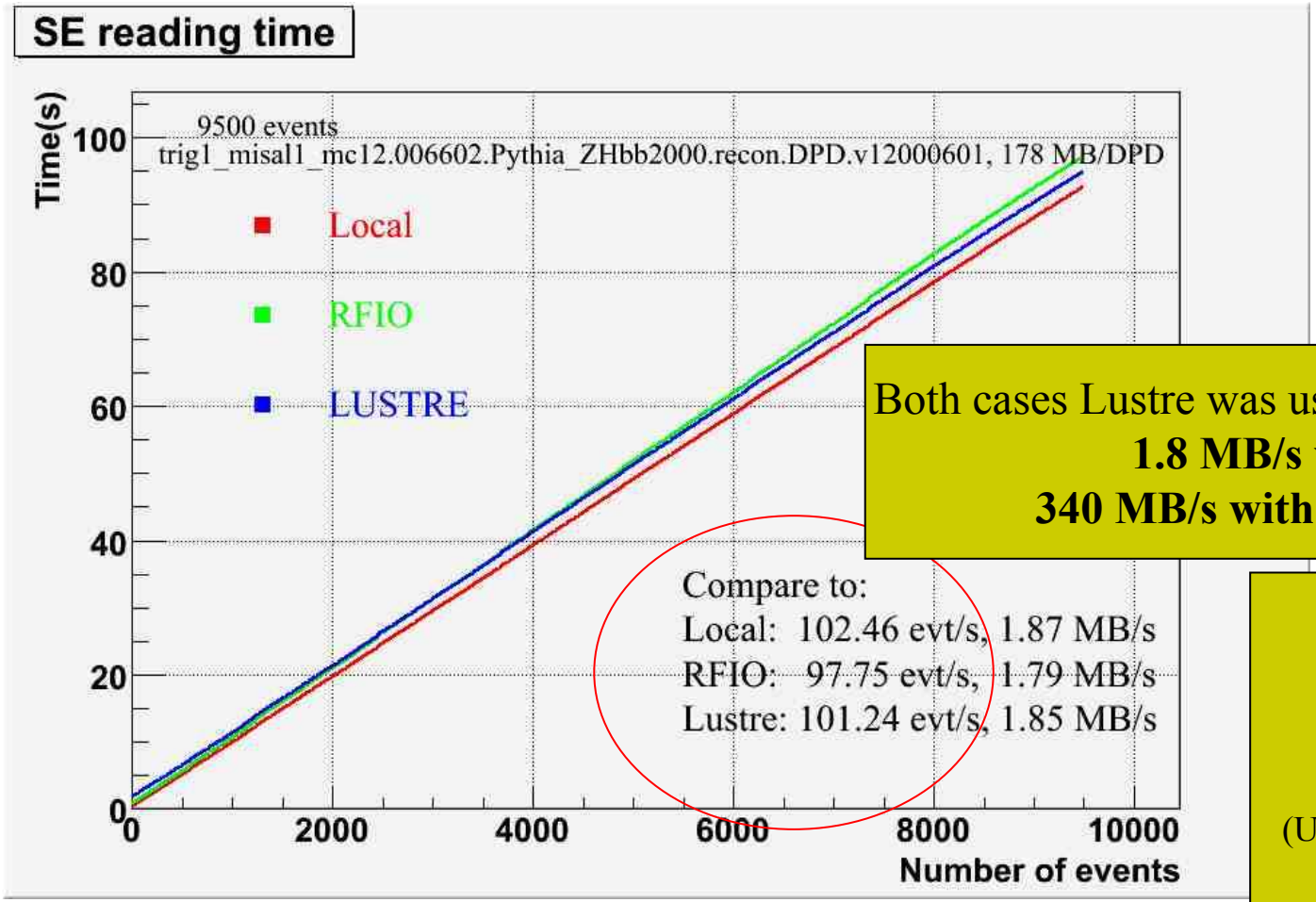
Tests:
RFIO without Castor





Athena analysis 12.0.6
AOD's
4 MB/s CPU limited and
Athena

The same test with DPD



Both cases Lustre was used and data in the cache
1.8 MB/s with Root
340 MB/s with a simple "cat"

2 x Intel Xeon 3.06GHz
4 GBytes RAM
1 NIC Gigabit Ethernet
HD: ST3200822AS
(Using a 3Ware card: 8006-2LP)

Summary

- From her/his desktop/laptop individual physicist can get access to:
  - IFIC Tier2-Tier3 resources.
  - ATLAS software (Athena, Atlantis, etc..), DDM/dq2 and Ganga tools.
- IFIC Tier3 resources will be split in two parts:
  - Some resources coupled to IFIC Tier2 (ATLAS Spanish T2) in a Grid environment
    - AOD analysis on millions of events geographically distributed.
  - A PC farm to perform interactive analysis outside Grid
    - To check and validate major analysis task before submitting them to large computer farms.
    - A PROOF farm will be installed to do interactive analysis.

¿Qué son Express Stream?

http://indico.cern.ch/getFile.py/access?contribId=2&resId=0&materialId=0&confId=a06527

- The Express Stream, like any other stream of data, will be defined by the trigger selection criteria.
- The computing model assumes that the first-pass event reconstruction will be completed in 48 hours since the data was taken (in the Tier-0 operations).
  - The bulk of the processing will begin after 24 hours.
- The data of the Express Stream, and of the calibration estreams, will be reconstruct in less than 8 hours.
- It will be possible to achieve feedback from the Express Stream in a few hours for the following reasons:
  - The data volume will be small (15% of the total)
  - Existing calibration constants can be used
  - A regular and rapid offline processing of the Express Stream will be made a part of the operation,
  - Useful feedback can be obtained already online, by monitoring the trigger rates that contribute to the Express Stream