# Infraestructura de e-ciencia para ATLAS en el IFIC

## E-science infrastructure for ATLAS at the IFIC

A. Fernández, S. González de la Hoz, L. March, J. Salt, R. Vives, F. Fassi, M. Kaci, A. Lamas y J Sánchez

Resumen

Bajo las siglas Large Hadron Collider (LHC) está el proyecto de Física de Altas Energías más ambicioso desde varios aspectos: científico, técnico, organizativo, etc. Se trata del mayor acelerador de partículas del mundo, con 27 Km de longitud, a varias decenas de metros bajo tierra y operando alrededor del cero absoluto. En dicho acelerador se van a realizar cuatro experimentos (ATLAS, CMS, ALICE y LHCb) que nos ayudarán a desentrañar enigmas básicos de la física tales como el origen de las masas de las partículas que conocemos o la razón del desequilibrio entre materia y antimateria. Esta previsto que el LHC empiece a operar en 2008.
El LHC supondrá un reto computacional ya que cuando el acelerador esté operativo, cada segundo se producirán 40 millones de colisiones, de las cuales sólo 100 serán de interés. El registro de cada colisión en formato digital se estima que ocupará alrededor de 1MB. Teniendo en cuenta que habrán 1010 colisiones por año, la información generada cada año será del orden de 10 PetaBytes (1 PetaByte = 1.000.000 GigaBytes). Además del reto computacional que suponen los programas de simulación y reconstrucción de sucesos, la pregunta que surge inminentemente es: ¿Dónde almacenar 10 PetaBytes anuales de manera que puedan ser procesados y analizados a posteriori por centenares de científicos distribuidos geográficamente por todo el mundo? Y la respuesta o solución es el GRID. Con ello nació en el CERN el proyecto LCG (LHC Grid Computing), donde actualmente la infraestructura del LCG está integrada en el proyecto europeo de e-Ciencia EGEE.
La distribución de los datos del LHC se hará mediante la red, la cual jugará y está jugando un papel muy importante y siguiendo una distribución jerárquica en "tiers" o capas. Una primera copia de seguridad se realizará en el CERN, único centro"tier-0", tras un procesado inicial, los datos viajarán a diversos centros "tier-1", donde serán almacenados en condiciones especiales antes de ser transferidos a centros "tier-2" y "tier-3" donde serán accedidos por los científicos.
En el caso español y para el experimento ATLAS, el TIER-2 es la federación de 3 centros (IFIC, UAM, IFAE) que proveen el almacenamiento necesario en disco en el orden de varios TeraBytes por centro, y accesibles a través del estándar SRM y protocolos como gridFtp. Los usuarios disponen de catálogos de datos distribuidos basados en WebServices, que mantienen constancia de las replicas entre los distintos tiers y que se realizan con herramientas como DQ2 desarrolladas para tal efecto. La idea es enviar jobs computacionales de los usuarios donde residen los datos de tal forma que se optimizan los recursos computacionales, pero también de red. De esta forma desde los Worker Nodes se puede acceder a los datos de forma eficiente, a través de sistemas de ficheros distribuidos como Lustre.
Por tanto, el objetivo de esta contribución es describir los servicios necesarios y la experiencia obtenida manteniendo, administrando y gestionando un centro, que es además el coordinador, que forma parte del TIER-2 Federado Español para ATLAS y de la Infraestructura TIER-3 para el Análisis de Datos en el Instituto de Física Corpuscular de Valencia, todo ello desde la perspectiva de la e-Ciencia.

Se van a realizar cuatro experimentos (ATLAS, CMS, ALICE y LHCb) que nos ayudarán a desentrañar enigmas básicos de la física

**Palabras clave:** LHC, experimento ATLAS, e-Ciencia.

Summary

The "Large Hadron Collider (LHC)", is Atlas Energies Physics' most ambitious project, in several areas: scientific, technical, organisational, etc. It is the world's largest particle accelerator (27 km long), located several dozen metres underground and operating at close to absolute zero. Four experiments will be conducted in the accelerator (ATLAS, CMS, ALICE and LHCb) that will help to unlock the basic secrets of physics, such as the origin of the known particle masses or the cause of the imbalance between matter and antimatter. The LHC is expected to begin operations in 2008.

The LHC will be a computing challenge in that, once the accelerator is operational, there will be 40 million collisions a second, of which only 100 will be of interest. The digital recording of each collision is estimated to require 1MB. Taking into account that there will be 1010 collisions a year, the information generated annually will be approximately 10 PetaBytes (1 PetaByte = 1,000,000 GigaBytes). In addition to the computing challenge posed by the simulation and event reconstruction programs, the question that immediately comes to mind is: Where do you store 10 PetaBytes a year so that they can be processed and analysed subsequently by hundreds of scientists located around the world? The answer or solution is the GRID. With it, the LGC (LHC Grid Computing) project was born at the CERN, where the LGC infrastructure is currently integrated into the European EGEE e-Science project.

The LHC data will be distributed on the network, which will play and is playing a very important role, following a hierarchical tier distribution. An initial back-up will be made at the CERN (the only "tier-0" centre). Following preliminary processing, the data will be sent to several "tier-1" centres, where they will be stored under special conditions before being transferred to "tier-2" and "tier-3" centres where they will be accessed by the scientists.

Para el experimento ATLAS, el TIER-2 es la federación de 3 centros: IFIC, UAM e IFAE

In the case of Spain, for the ATLAS experiment, TIER-2 is the federation of 3 centres (IFIC, UAM, IFAE), which provide the storage disk space necessary (several TeraBytes per centre), accessible through the SRM standard and protocols such as gridFtp. Users have a distributed data catalogue based on WebServices, which log the replications between the different tiers, carried out with tools such as DQ2, developed specifically for that purpose. The idea is to send computing jobs of the users where the data reside, optimising computer resources, as well as the network. In this way the Worker Nodes can be used to access data efficiently, through distributed file systems such as Lustre.

The aim of this contribution is to describe the necessary services and experience obtained by maintaining, administering and managing a centre, which is also the co-ordinator, that is part of the Federated Spanish TIER-2 for ATLAS and the TIER-3 Infrastructure for Data Analysis at the Corpuscular Physics Institute of Valencia, all from the point of view of e-Science.

**Keywords:** LHC, ATLAS experiment, e-Science.

## 1. Introducción, contexto y objetivos

El Large Hadron Collider (LHC) es el proyecto de Física de Altas Energías más ambicioso desde varios aspectos: científico, técnico, organizativo, etc. Se trata del mayor acelerador de partículas del mundo, con 27 Km de longitud, a varias decenas de metros bajo tierra y operando alrededor del cero absoluto. En dicho acelerador se van a realizar cuatro experimentos (ATLAS, CMS, ALICE y LHCb) que nos ayudarán a desentrañar enigmas básicos de la física tales como el origen de las masas de las partículas que conocemos o la razón del desequilibrio entre materia y antimateria.
Esta previsto que el LHC empiece a operar en 2008.

ATLAS es un detector de propósito general para el estudio de colisiones protón-protón de altas energías

ATLAS es un detector de propósito general para el estudio de colisiones protón-protón de altas energías. ATLAS, al igual que los otros experimentos del LHC, supondrá un reto computacional ya que cada segundo se producirán 40 millones de colisiones, de las cuales sólo 100 serán de interés. El registro de cada colisión en formato digital se estima que ocupará alrededor de 1MB. Teniendo en cuenta que habrán 1010 colisiones por año, la información generada cada año será del orden de 10 PetaBytes (1 PetaByte = 1.000.000 GigaBytes).
Además del reto computacional que suponen los programas de simulación y reconstrucción de sucesos, la pregunta que surge inminentemente es: ¿dónde almacenar 10 PetaBytes anuales de manera que puedan ser procesados y analizados a posteriori por centenares de científicos distribuidos geográficamente por todo el mundo? Y la respuesta o solución es el GRID. Con ello nació en el CERN el proyecto LCG (LHC Grid Computing), donde actualmente la infraestructura del LCG está integrada en el proyecto europeo de e-Ciencia EGEE.

La distribución de los datos se realizará mediante la red, la cual jugará y está jugando un papel muy importante en la accesibilidad de los datos

El modelo de computación para el LHC en general y de ATLAS en particular es de base jerárquica de TIERs (centros de cálculo-almacenamiento de diferentes dimensiones y funcionalidades). La distribución de los datos se realizará mediante la red, la cual jugará y esta jugando un papel muy importante en la accesibilidad de los datos.
Una primera copia de seguridad se realizará en el CERN, único centro"tier-0", tras un procesado inicial, los datos viajarán a diversos centros "tier-1", donde serán almacenados en condiciones especiales antes de ser transferidos a centros "tier-2" y "tier-3" donde serán accedidos por los científicos. Las funcionalidades del TIER2 establecidas en la colaboración ATLAS son las siguientes:

- Servicios de almacenamiento en disco permanente y temporal para ficheros de datos y bases de datos.
- Suministrar capacidades de análisis para grupos de trabajo de Física. Posibilitar la operación de una instalación de un Sistema de Análisis de Datos para 'Usuarios Finales' que dé servicio a unas 20 líneas de análisis.
- Suministrar datos de simulación de acuerdo con los requisitos de los experimentos.
- Dar acceso a los Servicios de Red para el intercambio de datos con los TIER-1.

En el caso español y para el experimento ATLAS, el TIER-2 es la federación de 3 centros (IFIC, UAM, IFAE) que proveen el almacenamiento necesario en disco en el orden de varios TeraBytes por centro, y accesibles a través del estándar SRM y protocolos como gridFtp.

## 2. Recursos y servicios del TIER-2/IFIC

El conjunto de recursos para computación del IFIC proporciona actualmente una potencia de cálculo de 132KSi2k y los recursos de almacenamiento un total de 34 TB en disco y de un máximo de 134 TB con robot de cintas (actualmente se dispone 1.7 TB utilizados). Estos recursos están distribuidos en diferentes nodos de distintas características. En la red, los enlaces físicos subyacentes pertenecen a la red académica contando con enlaces ópticos dedicados entre los nodos principales (T0-T1), en el caso español Cern-Pic con enlace de 10Gbps provisto por RedIRIS y GN2. El IFIC dispone de un enlace de 1Gbps con el punto de acceso de RedIRIS, siendo después la conectividad de 10 Gbps entre los centros del Tier-2.

El servicio FTS permite establecer canales de comunicación unidirecccional punto a punto entre los distintos centros de recursos de la infraestructura

### 2.1. Sistema de Transferencia de Ficheros y Almacenamiento:

El servicio FTS (File Transfer Service) permite establecer canales de comunicación unidirecccional punto a punto entre los distintos centros de recursos de la infraestructura. Al estar organizado el proyecto LCG en una arquitectura jerárquica o de tiers, estos canales FTS implementan esta organización en la comunicación de datos. Desde el Tier-0 (CERN) se establecen 2 canales (entrada/salida) por cada Tier-1 asociado, y de igual manera desde los Tier-1 los Tier-2. El IFIC establece los canales con su Tier-1 asociado (PIC), y formando parte de la región T-2 española.
Se han realizado transferencias de datos a nivel global de más de 10 millones de ficheros en los últimos 6 meses, constituyendo alrededor de 9 PetaBytes de datos.
El FTS tiene una interfaz basada en WebServices y está implementado en JAVA, y se encarga de realizar la negociación de los canales y contactar con los endpoints de cada sitio, de forma que es transparente para el usuario. Se encarga de establecer la comunicación con los SRM (Storage Resource Managers) de cada centro de recursos, origen y destino de cada comunicación, realizando el transporte usualmente con GridFTP sobre TCP.
En el IFIC se dispone de varios elementos de almacenamiento (SE o Storage elements) que proporcionan los endpoints SRM. "IFICTAPE" es un SE que representa un sistema jerárquico de almacenamiento con 1.7TB de disco como front-end, y con una librería de cintas como backend proporcionando una capacidad de almacenamiento potencial de varios TB. "IFICDISK" es un SE basado únicamente en disco, proporcionando actualmente 34 TB con un servidor de discos SUNx4500, accesible mediante LUSTRE, un sistema de ficheros paralelo.

Los servicios de accounting son importantes tanto para monitorización como para mantener un histórico del uso de los recursos

### 2.2. Job Accounting:

Los servicios de accounting son importantes tanto para monitorización como para mantener un histórico del uso de los recursos. En la actualidad se mantiene el accounting de trabajos computacionales por VO (Organización Virtual o grupo de usuarios con infraestructura e intereses comunes), por grupos de trabajo (dentro de una misma VO) y que son identificables con atributos específicos de los proxies con los que se envían los trabajos (groups/roles de los proxys VOMS), y con granularidad más fina, por usuarios específicos.
El acceso a estos datos es privado y sólo pueden acceder las personas autorizadas para ello (usuario, o Vo-manager como gestor de la VO). Técnicamente el accounting se implementa con APEL, que se basa en unos sensores que se ejecutan en los Computing Elements (elementos computacionales donde se ejecutan los trabajos) y que recogen los logs del sistema gestor de recursos local (LRMS) y los envían a un servidor central que mantiene las operaciones del grid, denominado Grid Operations Centre. Luego esta información puede sea accedida a través de una interfaz web, y de forma que el acceso a estos datos es privado y sólo pueden acceder las personas autorizadas para ello (usuario, o Vo-manager como gestor de la VO).

Existe un sistema nuevo denominado DGAS que no está implantado en producción, y que se diferencia en que no realiza un postproceso como APEL, sino que realiza el accounting por trabajo en tiempo real.
Se está desarrollando y comenzando a implantar el accounting de datos almacenados, aunque tampoco está implantado en producción en la actualidad ya que requiere de nuevos esquemas de información que serán incorporados progresivamente.

## 3. Transferencia y gestión de datos

El IFIC, como uno de los centros que forman el Tier-2 español para el experimento ATLAS, está participando en los diferentes ejercicios de transferencia y distribución de datos de dicho experimento; Service Challenge 4 (SC4), Fuctional tests y recientemente en la toma de datos de rayos cósmicos (M4.), como se puede observar en la figura 1 (a).

FIG. 1 (A). TRANSFERENCIAS REALIZADAS DENTRO DE LOS EJERCICIOS DESARROLLADOS EN SEPTIEMBRE DE 2007 (M4, FUNCTIONAL TESTS Y PRODUCCIÓN MONTE CARLO)

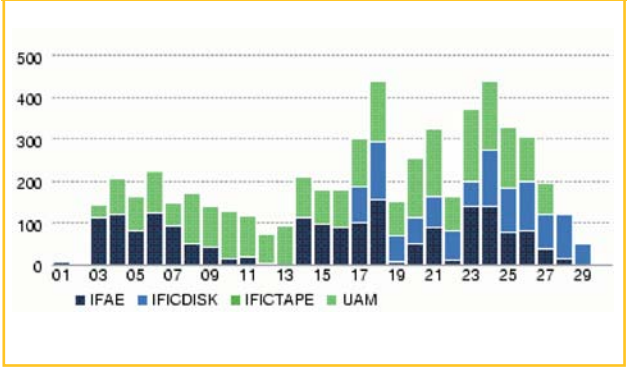
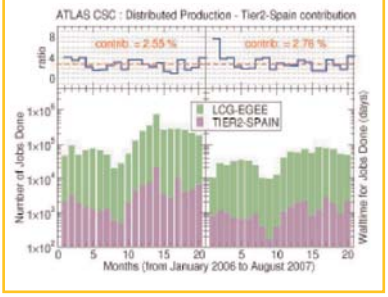



FIG. 1 (B). CONTRIBUCIÓN DE LA INFRAESTRUCTURA DE E-CIENCIA DEL IFIC (TIER2) A LA PRODUCCIÓN DE DATOS SIMULADOS DEL EXPERIMENTO ATLAS




El IFIC está participando en los diferentes ejercicios de transferencia y distribución de datos del experimento

Estos datos se almacenan en el IFIC utilizando LUSTRE (sistema de ficheros poxis IO) en disco, teniendo una capacidad actual de 34 TB. El almacenamiento de estos datos es uno de los objetivos primordiales del Tier2 ya que serán utilizados por los físicos para su posterior análisis. Por lo tanto permitir y asegurar su acceso a sus usuarios es primordial en este tipo de infraestructura de e-Ciencia. Se ha creado una página web, donde se da información de todas las muestras y datos almacenados en los centros del Tier-2, para información de sus usuarios, correspondientes a todos estos ejercicios: http://ific.uv.es/atlas-t2-es. Estos datos están almacenados en los SE del Tier-2 y registrados en sus respectivos catálogos: catálogo Grid general de LCG/EGEE (LFC) y catálogo del experimento ATLAS (DDM/DQ2).

Una de las actividades principales de la infraestructura de e-Ciencia en el IFIC (TIER2) es contribuir a la generación y producción de datos simulados

## 4. Producción masiva de datos simulados

Una de las actividades principales de la infraestructura de e-Ciencia en el IFIC (TIER2) es contribuir a la generación y producción de datos simulados del experimento ATLAS. A partir del año 2002 se iniciaron discusiones sobre las necesidades de cálculo para el experimento, llegándose a la conclusión de que el tratamiento de los datos del LHC supone, por el volumen y complejidad de los mismos, un reto sin precedentes en experimentos anteriores de física de partículas. La respuesta ha consistido en la elaboración de un modelo basado en las tecnologías GRID. Los proyectos GRID (ej. EGEE/LCG) para el LHC tienen una estructura jerarquizada, definiéndose los centros llamados TIER0, TIER1 y TIER2. Entre los objetivos del TIER2 tenemos la producción masiva de datos simulados y el análisis de datos por cada

usuario del experimento. De este último se hablará con más detalle en un apartado posterior. Los trabajos de producción son tanto enviados como recibidos en la infraestructura del IFIC. Desde enero hasta agosto de 2006 más de 60.000 trabajos fueron enviados desde el IFIC a otros laboratorios de la colaboración utilizando las tecnologías GRIDs. En ese periodo la colaboración ATLAS envío un total de 400.000 trabajos con una media diaria de 4.000. A su vez, nuestra infraestructura de e-Ciencia lleva recibiendo trabajos de simulación de distintos institutos de la colaboración ATLAS, habiendo procesado unos 85.000 trabajos de un total de 3.300.000 y habiendo gastando unos 22.700 días de tiempo de reloj de CPU sobre un total de 820.000 días como se puede observar en la figura 1-b. La contribución del IFIC a la producción de datos del experimento ATLAS ha sido de un 2.7%.

## 5. Servicios ofrecidos

### 5.1. Aplicación de análisis distribuido

Entre los objetivos del TIER2 está la producción masiva de datos simulados y el análisis de datos por cada usuario del experimento

Ganga es una herramienta implementada en python fácil de usar desde el punto de vista del usuario para la definición de trabajos y su envío a infraestructuras basadas en recursos GRIDs. La colaboración ATLAS ha tomado como estrategia el desarrollo de este tipo de herramientas para cumplir con las necesidades específicas de sus usuarios en la definición, construcción y ejecución de sus aplicaciones basadas en los programas de análisis y simulación de Monte Carlo del experimento. Desde este punto de vista Ganga permite enviar trabajos de forma local, lo cual es interesante para poder validar aplicaciones, o a diferentes centros/institutos para un procesado a mayor escala utilizando los recursos GRID del proyecto LCG/EGEE.

Un trabajo en Ganga se define siguiendo el esquema en bloques mostrado en la figura 2-a. El usuario debe especificar el tipo de software que quiere ejecutar (aplicación) y dónde quiere procesarlo (backend, de forma local o no). Además los trabajos pueden requerir de datos de entrada (input dataset) y/o especificar dónde quiere guardar los datos de salida (output dataset). Otra característica de Ganga es la posibilidad de dividir un mismo trabajo en varios subtrabajos para ser procesados en paralelo combinando el resultado final en un solo fichero de salida. Ganga se basa en un sistema plugin, lo cual hace que sea una herramienta fácilmente extensible y adaptable a las necesidades del usuario. La prueba de ello es que Ganga se utiliza en otro tipo de aplicaciones fuera de la Física de Altas Energías como puede ser la clasificación de imágenes en bases de datos.

Desde Enero de 2007, 968 personas utilizan la herramienta Ganga para enviar sus trabajos a centros con recursos GRIDs

Desde Enero de 2007 un total de 968 personas utilizan esta herramienta para enviar sus trabajos a centros con recursos GRIDs, de ellas 597 pertenecen al experimento ATLAS y 10 al IFIC. Aproximadamente 50.000 trabajos fueron enviados a dichos centros desde septiembre de 2007. En la figura 2-(b) se puede observar que aproximadamente unos 500 trabajos corrieron en la infraestructura de e-ciencia del IFIC.

FIGURA 2 (A). DEFINICIÓN DE TRABAJOS EN GANGA

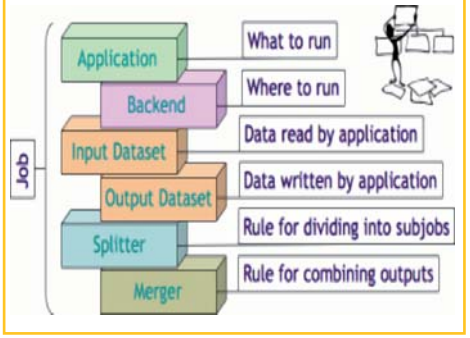
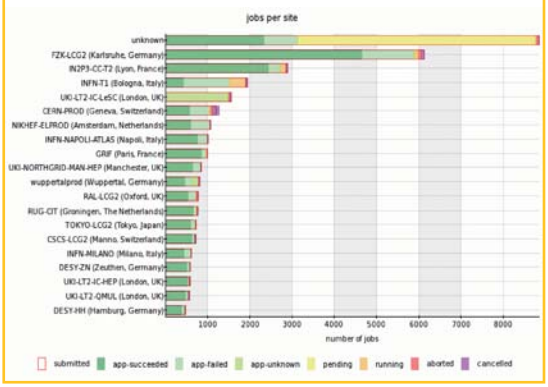



FIGURA 2 (B). DISTRIBUCIÓN DE TRABAJOS A DIFERENTES CENTROS CON RECURSOS GRIDS



### 5.2. Aplicación de Prioridades de Trabajos (Job Priorities)

Una forma de establecer prioridades en la ejecución de trabajos a nivel local de cada sitio es la implementación de políticas de fair-share de los recursos a través de los LRMS. Técnicamente es posible identificar cada trabajo con el proxy del usuario que lo ha enviado, y los proxies basados en VOMS permiten no sólo identificar al usuario, sino tambien establecer una serie de atributos como el grupo o grupos a los que pertenecen dentro de la VO (groups/roles).

Podemos utilizar esta facilidad para implementar políticas de prioridades no sólo a nivel de Vos, sino políticas intra-VO. De esta forma se establecen en cada sitio políticas, en el caso del IFIC tenemos asignado un fair-share 70% de tiempo de cómputo para ATLAS, y el resto para las otras Vos soportadas. De este porcentaje, se establecen las siguientes politicas intra VO definidas por grupos:

atlas:atlas 40% of atlas
atlb:/atlas/Role=production 60% of atlas
atlc:/atlas/Role=software no FS (but more priority, only 1 job at a time)
atld:/atlas/Role=lcgadmin no FS (but more priority, only 1 job at a time)

Esto significa que los usuarios normales obtienen potencialmente el 40% del tiempo de ATLAS, y que aquellos más especializados como los usuarios de "producción" obtienen el 60%. La aplicación de estas prioridades no es pre-emptiva con los trabajos en ejecución (por ejemplo si está todo el clúster ocupado con trabajos normales de atlas y llegan de producción no se eliminarán los que ya se están ejecutando), sino que son asintóticas a alcanzar.

Usuarios con los roles de instaladores de software (Role=software) o administradores (Role=lcgadmin) no envían muchos trabajos, y no tienen un tiempo de fair share asignado, pero tienen una alta prioridad de ejecutarse prontamente una vez han llegado al sistema.

Técnicamente es posible identificar cada trabajo con el proxy del usuario que lo ha enviado

## 6. Conclusiones

El IFIC dispone de una infraestructura de e-Ciencia que atiende las necesidades que se derivan de su papel como TIER-2 dentro del contexto del Modelo de Computación de ATLAS. Las diferentes actividades utilizan un considerable número de recursos e involucran diferentes aspectos del LCG. La Red es y será una pieza fundamental en el buen desarrollo y éxito de la infraestructura.

La Red es y será una pieza fundamental en el buen desarrollo y éxito de la infraestructura

**Santiago González**
(Santiago.González@ific.uv.es)
**Farida Fassi**
(Farida.Fassi@ific.uv.es)
**Álvaro Fernández**
(Alvaro.Fernandez@ific.uv.es)
**Mohamed Kaci**
(Mohamed.Kaci@ific.uv.es)
**Luis March**
(Luis.March@ific.uv.es)
**Jose Salt**
(Jose.Salt@ific.uv.es)
**Javier Sánchez**
(Javier.Sanchez@ific.uv.es)
**Roger Vives**
(Roger.Vives@ific.uv.es)
**Alejandro Lamas**
(Alejandro.Lamas@ific.uv.es)

Instituto de Física Corpuscular (CSIC-Universitat de València)