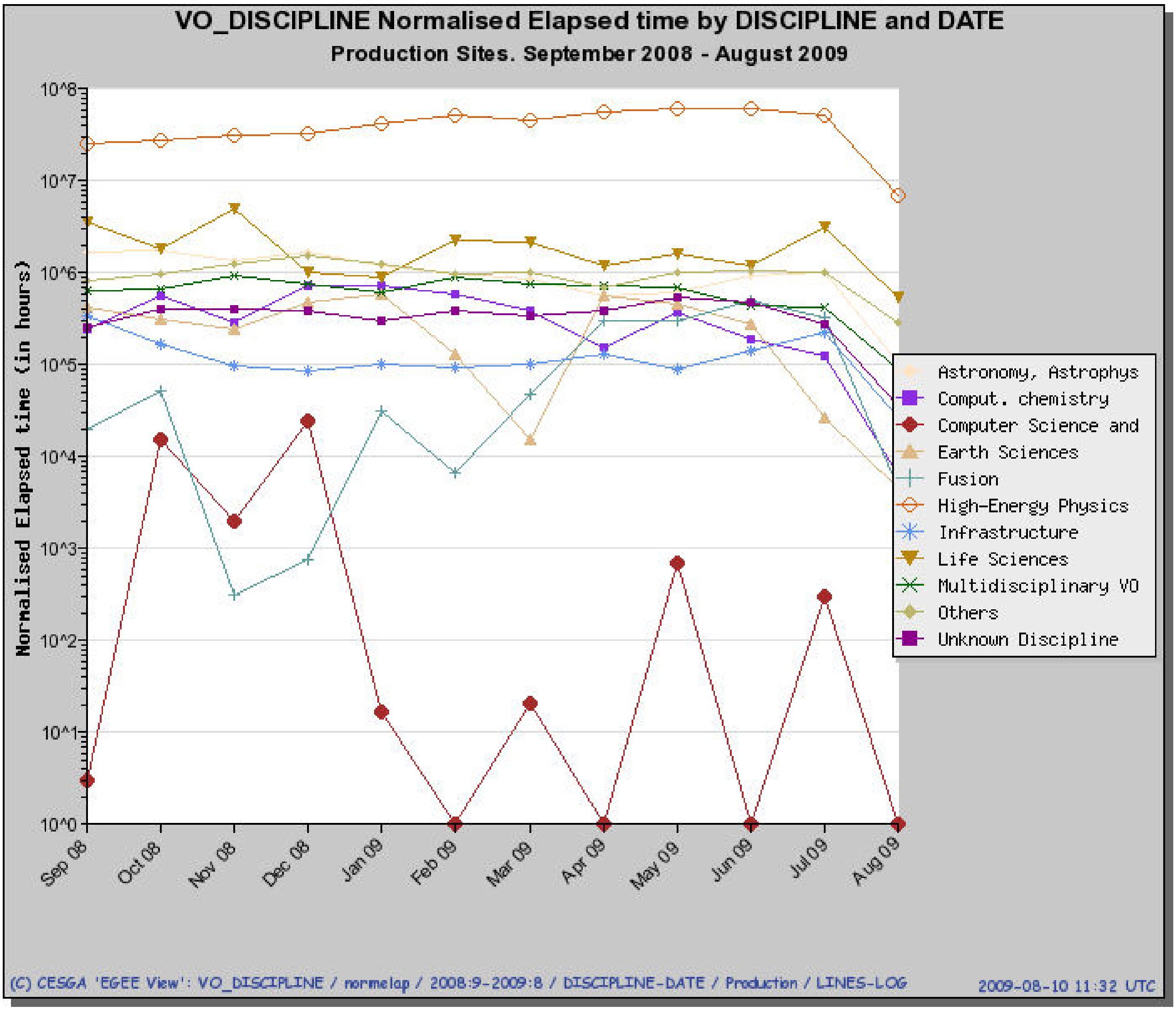
1.5.4 Activity SA4: Services for Communities of Heavy Grid Users

1.5.4.1 Activity Description

The support of the services needed by the present heavy users of the grid, is a key Service Activity in this Proposal, in response to the 1.2.1.2 sub-call. These services are build over the basic middleware components that in year 2008 (and beforehand) are provided by EGEE and other interoperable projects, and include both middleware services (e.g. FTS) and the frameworks that relay on the elementary components and services for performing functions tailored for specific communities (e.g. the Dashboard for the monitoring of the grid for specific VO's). The EGI support to these services includes their deployment, operation, and evolution for adapting to the needs both of the community that originally used the service, and of a larger set of users. The deployment and operation is part of this Activity, while the evolution is treated in JRA2

The European Grid Infrastructure is presently actively used by some scientific disciplines that heavily relay on this infrastructure for carrying on their research activity. The graph below, showing the monthly usage in the year from September 2008 to August 2009 by the different Scientific disciplines, gives clear indication of this massive usage.



The graph is taken from the EGEE accounting portal, the unit is Hours.1kSI2k elapsed time; in very first approximation a 10^6 in the vertical scale corresponds to about 2000 typical processors continuously working for the full month.

The Communities identified as Heavy Users Communities (HUCs) are

- High Energy Physics (HEP)
- Life Sciences (LS)
- Astronomy and Astrophysics (AA)
- Computational Chemistry and Material Sciences (CCMT)
- Earth Sciences (ES)

Besides the massive usage of the Grid, the HEP and LS have played a pilot role in EGEE, giving decisive contributions for bringing the grid at production quality level, via feedback on the deployed services efficiency and functionalities, stress tests of the infrastructure and selected components, etc. It is expected now that all the HUCs will be able to play a similar role for the services of their interest included in this Activity, and at least in some case for specific aspects of the EGI grid (e.g. further scalability in case of HEP)

A key element for the satisfaction of the needs of these users, and for ensuring the continuity and enlargement of the role of "pilot users", vital for the grid, is providing them with the higher level services they need. In general these services are already used by the community, with some support by EGEE and the other interoperable projects.

The first objective for SA4 is that Heavy Users Communities experiment no disruption of their activity with the transition of the e-infrastructure to the EGI support. In fact EGI aims at increasing the satisfaction of these main users, also in view of expanding the Grid usage within the disciplines they belong to, and toward new disciplines. SA4 will grant to the communities the continuity and good integration of services in the general grid, and will shape the provision of these services so that they are efficiently used by all the presently interested user communities and extend their usage to new communities.
SA4 will work so that most services will become more standard and easy to configure, deploy and operate. Some services will retain interest only for a small fraction of the user communities, and in perspective will be supported primarily by these communities with reduced contribution from EGI, while some other services will become more general in their use and will in the future become integral part of the EGI general infrastructure.

1.5.4.2 Assumptions on the general services available outside SA4

The services provided in SA4 are the services needed by the HUCs and not included in the general EGI infrastructure, nor supported by any other trusted provider.

The following Table.SA4.1 provides a list of the general services available outside SA4; when the service is not included in the general EGI infrastructure the provider is indicated

Compute Element from gLite and ARC		WLCG
LCG CE?	?	WLCG – until CREAM CE everywhere in full production
SE from GLite and ARC		What is the SE from gLite & ARC?
dCache	dCache.org	WLCG
DPM	CERN	WLCG
StoRM	INFN	WLCG
Castor	CERN	WLCG
SRM		WLCG
GridFTP		WLCG
Info.services from gLite and ARC		WLCG
Accounting from gLite and ARC		WLCG
Authorization Services from gLite and ARC (e.g. VOMS, MyProxy, SCAS LCMAPS, LCAS, gLExec)		WLCG
Work Load Management services (e.g. WMS)		WLCG

Each HUC provides the indication of with detailed services in the list are of interest for it, specifying the use, the importance etc. The VOs supported by WLCG (ALICE, ATLAS, CMS and LHCb) require the above services in full production. CASTOR is deployed at the Tier0 and 3 Tier1s (ASGC,CNAF, RAL). dCache is installed at the other Tier1 sites and many Tier2s, although the majority of such sites deploy DPM. SToRM is also deployed at CNAF and a small number of Tier2s.

1.5.4.3 Task Description

1.5.4.3.1 TSA4.1 SA4 Management

The SA4 management comprises the full time Activity Manager, with the responsibility of supervising the services and coordinating their provision with the relevant communities.
He/she is assisted by representatives of the relevant communities, with the expertise necessary for providing feedback on the working of the services and input on the modification that may be needed, in configuration, operation and deployment of the services for the specific communities. The effort needed from the representatives of the communities depend from the amount and complexity of the services requested by the specific community

The SA4 manager will be a member of the MCB

1.5.4.3.2 TSA4.2 Hosting of Community specific Services

Provision and operation by a small number of NGIs of Core Grid Services (O-N-8) explicitly needed to support this user community, but of potential benefit to other communities.
These centres will be experts and provide an SLA around the hosting of services such as FTS, LFC, Hydra, AMGA and VO specific services.

*This list is an **example** of services and the writers from the specific communities should introduce here the services of their interest,if any; for each service please provide*

- *The description and the motivations*
- *The evaluation of the effort*
- *The NGI(s) proposed for hosting and their share of the effort: CERN ≥ 50%*
- *Indication of SLA/SLD*
- *Interested HUC(s): HEP(WLCG)*

The four main experiments at CERN's Large Hadron Collider (LHC) – namely ALICE, ATLAS, CMS and LHCb – rely on a worldwide virtual computing facility that is implemented using grid technology. The architecture of this system is based upon a tier model that consists of CERN, eleven major regional or national data-intensive "Tier1" centres as well as over one hundred "Tier2" sites that are located close to end users. In order to facilitate data movement between these sites – which has run at 1PB/day over prolonged periods – and to enable the required data to be located, file transfer and catalogue services are required. These are implemented on top of *glite* components: namely the File Transfer Service (FTS) and the LHC (aka "local") File Catalog (LFC). These services run at the Tier0 and Tier1 sites and involve both application servers as well as backend database systems. Whilst the LFC has been adopted by numerous other HEP and non-HEP communities, the use of the FTS is currently predominantly WLCG only. However, as more and more communities face limitations from power and cooling and move to distributed systems, the need for reliable wide-area file transfer services over high capacity network links can be expected to grow. It may also be needed in *cloud* computing environments to transfer data both into and out of "the cloud".

Both of these middleware components were carefully designed with service deployment and reliability in mind: they permit resilient and scalable deployment models (load balanced front-ends, database clusters as back-ends etc, "rolling upgrades" of both middleware and underlying O/S components and even migration of hardware!) WLCG Tier1 sites are nonetheless recommended to have 1 FTE to manage the database services behind these and other required services (detector conditions and / or storage services), whereas the Tier0 runs a more complex setup with – in the case of LHCb – data replication for the LFC to read-only replicas at the Tier0 and all Tier1 sites. Thus, to deliver these services a total of 1 FTE is required per Tier1 with a minimum of 2 FTEs at the Tier0. The sites involved are CERN and the 7 WLCG Tier1 sites in Europe (IN2P3 in France, FZK in Germany, NIKHEF/SARA in the Netherlands, NDGF in the Nordic countries, PIC in Spain, CNAF in Italy and RAL in the UK – the remaining Tier1 sites in North America and Asia would not expect to be funded through this task, although a similar amount of effort, adjusted to the number of VOs supported by each site, is required).

An additional component of service delivery that is today covered by the CERN team for both FTS and LFC is consultancy and / or development of operations tools and procedures for various "house-keeping" operations, such as change in DB back-end configuration, bulk operations on namespace, bulk changes to FTS channel configurations and so forth. Additional work is also required at the level of database replication, again to handle configuration changes and / or to resynchronise sites after prolonged outages or other problems. Such operations occur on average once per month and require ~20% of an FTE, included in the overall estimate below.

Effort: up to 9 FTEs, 50% of which would be co-funded by the above mentioned sites and / or parent NGI.

1.5.4.3.3 TSA4.3 Hosting of VO specific Services

Provision and operation by a small number of NGIs of Core Grid Services (O-N-8) explicitly needed to support this VO, but of potential benefit to other users.
Justified if the VO users are a relevant fraction of the Grid users and/or they use a relevant fraction of the grid resources, or if the service is foreseen to become of more general interest during this Project.

The writers from the specific communities should introduce here the services of their interest, for each service please provide

- *The description, and the justification*
- *The evaluation of the effort*
- *The NGI(s) proposed for hosting and their share of the effort: CERN ≥ 50%*
- *Indication of SLA/SLD*
- *Interested HUC(s): HEP (WLCG)*

Building on the powerful generic infrastructure of the underlying grids that they use, the LHC experiments have developed important complementary services particularly in the areas of data and workload management, as well as in support for analysis services. Such services, which extend the capabilities of the infrastructure by exploiting knowledge of the experiment's computing model, data placement policies and/or information in metadata repositories, allow these massive international communities to maximise the benefit of the grids that they use. For example, PhEDEx, the CMS data movement system, is able to source files belonging to a larger dataset (a concept that does not exist at the underlying FTS layer) from alternative sites, leading to additional robustness and performance. As much as 50% of the data – possibly more – may be retrieved from such a source: functionality that cannot – by design – be provided at the FTS layer.
Whilst these experiment-specific solutions typically address today individual VOs, each such VO consists of thousands of users worldwide and corresponds to significant usage of grid infrastructure. Furthermore, experience has shown that some such solutions not only become adopted by other HEP VOs but later spread into additional communities and should be considered an important source of innovation (the driving force being the raw requirement but the actual realisation through the significant competence within these VOs must also be considered as a major source of "unfunded" effort that can benefit other communities worldwide). The "classic example" in this vein is AliEn, which is a lightweight Grid framework that is built around Open Source components using the Web Services model. It has been initially developed by the ALICE collaboration as a production environment for the simulation, reconstruction, and analysis of Physics data in a distributed way. The architecure of AliEn provided a blueprint and a starting point for developing the gLite architecture. Other examples include the use of generic pilot jobs, originally adopted by both ALICE and LHCb but now becoming the preferred mode for all 4 LHC experiments.

This activity is of particular importance now as we enter the exploitation phase of the world's largest scientific machine – the Large Hadron Collider at CERN – and will allow us to capitalize on the investment made by the European Commission through its funding of three phases of the Enabling Grids for E-sciencE project. This has resulted in large scale production use of world-class Grid-based solutions by many key communities and has established Europe's leadership in this area. In the short to medium term it is expected that this will lead to significant advances in our basic understanding of the Universe around us, whereas in the longer term major spin-offs, both related to the advances in science as well as in Information Technology, can be expected.

The detailed service requirements by experiment are listed below.

ALICE

AliEn is the Grid middleware used by the ALICE experiment. It provides the components needed for Workload and Data management, including MC production, raw data reconstruction and user analysis.
AliEn uses WLCG infrastructure for services and support.

AliEn is also used by other communities within HEP, such as CBM and PANDA, and also in other HUCs, including life science. Support for AliEn is therefore not only essential for the ALICE experiment but also will encourage technology transfer – one of the goals of EGI.

ALICE specific services needed in the EGI infrastructure can be divided in three areas:

- Workload Management:
  - interfaces to the CREAM-CE and the LCG-CE
  - Implementation of job priorities and quotas
- Data Management:
  - Integration with all supported transfer and storage systems, including xrootd, CASTOR, dCache, FTS
  - Improve file distribution to the sites
  - User storage quotas
- Site & production support:
  - MC production
  - Production support
  - VOBox support and maintenance
  - Integration of the AliEn services into WLCG services

ALICE needs two FTEs to accomplish all these items.

ATLAS

ATLAS depends on the Grid infrastructure (a federation of EGEE, OSG and NDGF sites) for its off-line computing activities. The Grid is used for the simulation, for the data reconstruction and all the analysis activities. It is also used for part of the calibration work.
ATLAS has built its system on the foundation of these Grid infrastructures. As an example in the data management system (DDM) the data transfer is delegated to FTS, the storage functionality to SRM, the site local cataloging to LFC. The DDM monitoring is an integral part of the Dashboard system.

As a part of the preparation for data taking, ATLAS has optimised its system with particular emphasis on operation and support issues. Wherever possible common systems have been contributed to and adopted, for example, the Dashboard as the backbone for all of the ATLAS monitoring and Ganga as the interface to physics analysis job submission. These systems are described in TSA4.4.

The DDM is the core of the system and it is still in evolution due to changes in the underlining technology and on the experience collected. The final DDM validation will clearly arrive with the full-scale heavy “chaotic” access from final users (as opposed to massive simulation and reconstruction activities). A set of readinesses challenges (the last one being STEP09 in June 2009) have confirmed the robustness of the system so far but this still requires continuous interaction with the sites and middleware providers (asking for configuration and middleware changes respectively).

In the last year a lot of work has been invested (by CERN IT and by ATLAS) to improve the entire software process of the ATLAS services with the goal to use the same procedures in place for baseline Grid services. This operational task needs to be finalised and a full integration of ATLAS and site procedures (primarly CERN) is still going on.

The counterpart of this activity on the user community is the distributed ATLAS support system, based on the integration of existing tools to deal with users and site issues This system (now in production in its first version for the last 12 month) has the potential to be beneficial for the entire HEP and other large user base communities.

Future evolution:
More streamlined operations and procedures will allow ATLAS to sustain the operation of its system in the long run. Most of the experience is quite VO independent and will be shared within the project.
Impact:
As demonstrated by the example of Dashboard, Ganga, User Analysis Support, ATLAS feeds back experience and development (coming from the experiment itself) into the Grid community (notably EGEE in the past 6 years). The DDM is actually a laboratory to understand multipetabyte data management (on a scale of 100+ sites and 1000+ users) which will be a major contribution to Grid technology and to the EGI in particular.
Effort:
The total estimated effort to support these services worldwide is 3 FTEs, 50% co-funded through CERN, INFN and WLCG.

CMS

CMS relies on a globally distributed computing model for its data processing and analysis, performed by more than 2000 scientists world-wide. The computing model is based on a hierarchy of tiered regional computing centres operating services for data management, transfers and workload management. The CMS Remote Analysis Builder (CRAB) tool allows the access to the distributed data, storage and processing resources in a transparent way. The PhEDEx Data Service provides access to information from the central PhEDEx database, as well as certificate-authenticated managerial operations such as requesting the transfer or deletion of data. The Data Service is integrated with the 'SiteDB' service for fine-grained access control, providing a safe and secure environment for operations. Below is a more detailed description of the services and the manpower required.

CRAB is a CMS specific user friendly interface to the WLCG resources and CMS Data Management tools. It supports access to OSG, WLCG and ARC resources. It uses the core CMS Workload Management code base and the WLCG provided API's to WMS, and provides a client-server solution to job submission. Users interact with a thin client which delegates work to a "centrally" operated server. The servers are well known to the CMS analysis community and run and are supported centrally and at remote centres. Long term maintenance of CRAB, for instance to handle updates to WMS API's and the CMS wrapper (BossLite) is necessary for both CRAB and the wider Workload Management system. Support, and significant training, of administrators of the servers is needed for the long term viability of the tool.

Impact:
CRAB is the recommended and supported CMS analysis tool. It takes care of all transfer, catalogue and job submission operations. Migration, and associated testing, to more recent externals needs

significant attention. Support of centrally run servers, and support for the CRAB server to operations staff at T2 centres is required.

Effort: 1 FTE

SiteDB is a key component in the CMS computing. It is used for resource planning, providing "human readable" names in monitoring and job matching services and for providing authorisation to Data & Workload management systems (for instance, every PhEDEx request, or every job from CRAB makes a call to SiteDB). Maintenance effort is needed to keep it inline with developments in VOMS, Apache/mod_ssl and the surrounding CMS code base. SiteDB also acts as a top-level portal into monitoring software (SLS, PhEDEc, Dashboard, SAM etc.).

Impact:
Providing seamless integration with existing secured services eases the lives of every active member of CMS. Careful and close monitoring and tuning of the application is necessary for all services relying on SiteDB.

Effort: 0.5 FTE

PhEDEx is the CMS data placement tool. It manages FTS transfers to place data at sites as dictated by CMS policy. As the authoritative source of data location information in CMS it is also used, via a sophisticated RESTful data service, to identify which sites are compatible with analysis, skimming and re-reconstruction jobs.

Impact:
PhEDEx is used by every CMS site. Maintaining contact with CERN/IT, FTS and CASTOR developers, keeping abreast of SRM developments etc. is vital for successful service operation. Also, guiding/pushing site testing of pre-releases is required. Engaging effort to assist with deployment of new releases, especially in the regime of significant change to the underlying grid tools, enables CMS to distribute, and hence analyse, data more effectively.

Effort: 1 FTE

The CMS Data/Workload Management services use HTTP(S) based services to query, retrieve, or update information from other components. The ease of deploying HTTP based services, with their already well defined failure modes and API (REST), is extremely attractive, and will become more common, and more important, over time.

Impact:
Maintenance effort is needed to ensure these services are operated in secure ways. The services need to be extended and the scale of access grows with increased activity during data analysis. Novel operational techniques may be required to provide the level of service expected by the CMS collaboration.

Effort: 1 FTE

LHCb

LHCb Grid access is based on the DIRAC integrated Workload Management and Data Management systems (WMS and DMS). DIRAC fully depends on middleware and services currently provided by EGEE and the WLCG. The LHCb Computing Operations also use the EGEE infrastructure services. Grid Analysis using DIRAC is provided through the ganga framework.

The DIRAC framework for WMS and DMS implicitly or explicitly makes use of the following components:

- Storage: Storage Element technologies (Castor, dCache, StoRM and DPM), interfaced through SRM (addendum to SRM 2.2 for WLCG).
- Data Management middleware and services: LFC as file catalog, gfal and lcg_utils as high-level access to SRM (using the python binding), FTS for file transfers. LHCb has its dedicated LFC services at each Tier1 (replication from the CERN main repository using WLCG 3D).
- Workload Management middleware and services: DIRAC uses the pilot job paradigm for WMS. Currently the pilot job submission goes through gLite-WMS instances to LCG-CE. When CREAM becomes available at all Tier1s for LHCb and DIRAC has been instrumented, pilot job submission to CREAM CEs will be direct submission, no longer using the gLite-WMS.

Site performance is monitored and reported through the SAM framework and the Dashboard services and tools.

The total estimated effort to support these services worldwide is 10.5FTEs, 50% co-funded through CERN, INFN and WLCG. The approximate breakdown by VO is ALICE: 2, LHCb: 2, ATLAS: 3, CMS: 3.5

The gLite VO box is a pre-requisite for these services.

1.5.4.3.4 TSA4.4 Support of Frameworks

The frameworks integrate different components and services for performing functions tailored for specific communities or VOs
An example are the VO Dashboards: VO Dashboards have been found to be very useful by large VOs to provide a VO view of the infrastructure for their community. Other examples may be GANGA, PHEDEX, DDM, WISDOM

This task in case of Dashboard includes

- *hosting of the service*
- *support of the Dashboard framework which provides necessary components for constructing of the monitoring applications*
- *integration of the VO-specific tests driven by a particular user community*
- *support in instrumentation of the VO-specific workload and data management systems for reporting of the monitoring data in a generic way*
- *constructing high-level cross-VO view (at the scope of the single site or at the global scope) via integration of multiple VO-specific monitoring systems*

This will also draw on the generic service monitoring infrastructure and tests maintained by the NGIs.

The content is analogue in the case of the other Framework and should be described by the specific writers

The writers from the specific communities should introduce here the frameworks of their interest, if any; for each framework please provide

- *The description, and the justification*
- *The evaluation of the effort*
- *The NGI(s) proposed for contributing and their share of the effort*
- *Indication of SLA/SLD if applicable*
- *Interested HUC(s): HEP(WLCG) and others*

In order to perform production and analysis tasks across a highly distributed system crossing multiple management domains powerful and flexible monitoring systems are clearly needed. To respond to the LHC experiments' requirements in this area, the experiment **Dashboard** monitoring system was originally developed in the context of the EGEE NA4/HEP activity. This framework, not only supports multiple grids / middleware stacks, including glite, OSG and ARC (NDGF), but is also sufficiently generic as to address the needs of multiple other communities including but not limited to HUCs. Furthermore, it covers the full range of the experiments' computing activities: job monitoring, data transfer (see FTS and VO services above) as well as site commissioning. It also addresses the needs of different categories of users, including:

- Computing teams of the LHC VOs;
- VO and WLCG management;
- Site administrators and VO support at the sites;
- Users running their computational tasks on the grid infrastructure.

Future Evolution:

The future evolution of the project is driven by the requirements of the LHC community which is preparing for LHC data taking at the end of 2009.

The main strategy is to concentrate effort on common applications which are shared by multiple LHC VOs but can also be used outside the LHC and HEP scope. Examples of such applications are: generic job monitoring application and user task monitoring, FTS monitoring, site status board, VO-specific site availability based on the results of tests submitted via Site Availability Monitor (SAM).

Impact:

Reliable monitoring is a necessary condition for establishing and maintaining production quality of the distributed infrastructure. Monitoring of the computing activities of the main communities using this infrastructure in addition provides the best estimation of its reliability and performance.

The importance of flexible monitoring tools focusing on the applications has been demonstrated to be essential not only for "power-users" but also for single users.

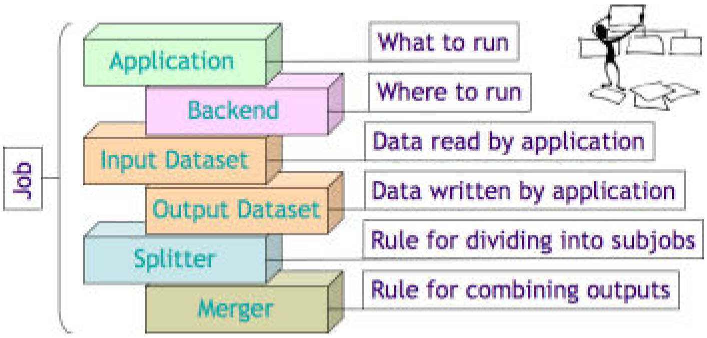
For the power users (such as managers of key activities like large simulation campaigns in HEP or drug searches in BioMed) a very important feature is to be able to monitor the resource behaviour to detect the origin of failures and optimise their system. They also benefit from the possibility to "measure" efficiency and evaluate the quality of service provided by the infrastructure. Single users are typically scientists using the Grid for analysis data, verifying hypotheses on data sets they could not have available on other computing platforms. In this case the monitoring / dashboard is a guide to understand the progress of their activity, identify and solve problems connected to their application.

This is essential to allow efficient user support by "empowering the users" in such a way that only non-trivial issues are escalated to support teams (for example, jobs on hold due to scheduled site maintenance can be identified as such and the user can decide to wait or to resubmit).

Effort: 4 FTEs, 50% co-funded through CERN / WLCG.

Ganga is an easy-to-use frontend for job definition and management, implemented in Python. It has been developed to meet the needs of the ATLAS and LHCb for a Grid user interface, and includes built-in support for configuring and running applications based on the Gaudi / Athena framework common to the two experiments. Ganga allows trivial switching between testing on a local batch system and large-scale processing on Grid resources.

The main paper reference is *Ganga: a tool for computational-task management and easy access to Grid resources; arXiv:0902.2685v1* , to be published in Computer Physics Communications (doi:10.1016/j.cpc.2009.06.016).



A job in Ganga is constructed from a set of building blocks. All jobs must specify the software to be run (application) and the processing system (backend) to be used. Many jobs will specify an input dataset to be read and/or an output dataset to be produced. Optionally, a job may also define functions (splitters and mergers) for dividing a job into subjobs that can be processed in parallel, and for combining the resultant outputs. Ganga provides a framework for handling different types of application, backend, dataset, splitter and merger, implemented as plugin classes. Each of these has its own schema, which places in evidence the configurable properties.

As it is based on a plugin system, Ganga is readily extended and customised to meet the needs of different user communities. Activities outside of ATLAS and LHCb where Ganga is successfully used include Geant4 regression tests and image classification for web-based searches.

The number of Ganga users has steadily increased and today there are several hundred grid users using the tool in their daily work, some 25% of whom are not from HEP VOs. Whilst these other VOs and the successful gridification of numerous associated applications in a wide range of fields including Fusion, Material Sciences, Accelerator Studies and Biomedical applications, the effort requested here would focus on production service deployment to the WLCG VOs ATLAS and LHCb in the critical early years of the LHC's operation.

Effort: 2 FTEs, 50% co-funded through CERN / WLCG.