PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2009) **367**, 2471–2481
doi:10.1098/rsta.2009.0050

Building a scientific data grid with DiGS

By Mark G. Beckett[1,*], Chris R. Allton[2], Christine T. H. Davies[3], Ilan Davis[4], Jonathan M. Flynn[5], Eilidh J. Grant[1], Russell S. Hamilton[4], Alan C. Irving[6], R. D. Kenway[1], Radoslaw H. Ostrowski[1], James T. Perry[1], Jason R. Swedlow[7] and Arthur Trew[1]

[1]*School of Physics and Astronomy, University of Edinburgh, Edinburgh EH9 3JZ, UK*
[2]*Department of Physics, Swansea University, Swansea SA2 8PP, UK*
[3]*Department of Physics and Astronomy, University of Glasgow, Glasgow G12 8QQ, UK*
[4]*Department of Biochemistry, University of Oxford, Oxford OX1 3QU, UK*
[5]*School of Physics and Astronomy, University of Southampton, Southampton SO17 1BJ, UK*
[6]*Department of Mathematical Sciences, University of Liverpool, Liverpool L69 3BX, UK*
[7]*College of Life Sciences, University of Dundee, Dundee DD1 5EH, UK*

We provide an insight into the challenge of building and supporting a scientific data infrastructure with reference to our experience working with scientists from computational particle physics and molecular biology. We illustrate how, with modern high-performance computing resources, even small scientific groups can generate huge volumes (petabytes) of valuable scientific data and explain how grid technology can be used to manage, publish, share and curate these data. We describe the DiGS software application, which we have developed to meet the needs of smaller communities and we have highlighted the key elements of its functionality.

Keywords: scientific data infrastructure; distributed data management; grid; molecular biology; computational particle physics

1. Introduction

Put simply, a *data grid* is a distributed computer infrastructure intended to manage the storage of data. The term (data grid) will be familiar to anyone who is active within the field of grid computing. This is, in part, thanks to the high-impact activities of projects such as Enabling Grids for E-SciencE (EGEE; http://public.eu-egee.org/) and TeraGrid (http://www.teragrid.org/), which

[*] Author for correspondence (george.beckett@ed.ac.uk).

One contribution of 16 to a Theme Issue ‘Crossing boundaries: computational science, e-Science and global e-Infrastructure I. Selected papers from the UK e-Science All Hands Meeting 2008’.

have substantial resources available to evolve distributed data management understanding and technology, based on the specific requirements of experiments such as the Large Hadron Collider (LHC).

In practice, data grids generally share a number of key elements. Firstly, as hinted at above, a data grid is usually intended to manage a large amount of data; hundreds of terabytes (TB) or even petabytes (PB) are typical at the time of writing. Secondly, a data grid generally holds families of the same type of data relating to experimental results, for example. Thirdly, a data grid is likely to be distributed across multiple, administrative domains and even across international boundaries.

Within this paper, we discuss our experiences of building a scientific data grid technology, called DiGS (http://www2.epcc.ed.ac.uk/∼digs/). Our motivation for doing this has been the requirements of smaller communities than those targeted by EGEE and TeraGrid, working within more modest resource constraints, but which are still able to realize significant benefits through distributed data management.

We have focused specifically on two applications: one from computational particle physics, provided by the UKQCD consortium (http://ukqcd.epcc.ed.ac.uk/), and another from molecular biology, provided by two specific research groups.

The UKQCD is a collaboration of physicists conducting research into lattice field theory, with the aim to increase the predictive power of the standard model of elementary particle interactions through numerical simulation of quantum chromodynamics (QCD). As explained in Perry *et al.* (2005), the DiGS developers have worked with the UKQCD, developing their data grid since 2001.

The Davis and Finnegan (D&F) research groups are two teams of molecular biologists, spread across two laboratories: the department of cell biology at the University of Edinburgh; and the department of biochemistry at the University of Oxford. They are interested in early development in the fruitfly *Drosophila melanogaster*, specifically investigating the process of RNA localization and anchoring. The collaboration with cell biology began in 2006, with a feasibility study (M.G. Beckett & R.S. Hamilton 2006, unpublished work) completed to confirm the use case.

DiGS has been created (and continues to be developed) for these two communities. It is a *grid application* that combines the disparate storage resources of a community to give a unified view of a (distributed) repository. In short, DiGS supports the management, sharing, publication and curation of collections of data.

In §2, we explain the complementary requirements of the two target applications. We explain how, while apparently being quite different in nature, these two applications share a common need for a distributed data management system.

We then describe the typical make-up of a data grid infrastructure, with reference to our own experience from developing DiGS. We highlight the fundamental components that are essential to an effective data management solution, and describe how these components fit together to form a data grid.

In §5, we consider how DiGS delivers real benefits to the target scientific communities, helping them to get more from their data and to mature collaborations that would otherwise be problematic.

2. Requirements

The two applications considered in this paper are very different in their nature, though they have a number of crucial similarities in their requirements for distributed data management, as described below.

(*a*) *Types of scientific data*

For the UKQCD, the data of interest are generated (or consumed) by lattice gauge field theory computations. These are stochastic in nature and individual files are only meaningful if collected together to form an *ensemble*. The format of the constituent files is completely within the control of the community, being generated by the community's software. These files have a size, at the time of writing, of approximately 1 gigabyte (GB) and ensembles generally contain thousands of files, implying that an ensemble typically represents in excess of 1 TB of data. File size is only constrained by available computing power, and sizes in excess of 10 GB are predicted to be commonplace by the end of the decade.

By contrast, the D&F groups are concerned with image data acquired from specialized microscopes, in proprietary formats, as dictated by the particular manufacturer. As for the UKQCD, images are often collected together into *experiments*—for example, capturing a time sequence of images for a live specimen and documenting analysis techniques that are applied to them. The size of a typical image is in the range 10 megabytes (MB) to 1 GB, with potentially hundreds of images in an experiment. A storage requirement of the order of 1 TB is not uncommon for an experiment.

For both applications, the usual mode of access to the data is *write once and read many times*. This reflects the fact that files contain the results of experiments, which need to be preserved without alteration for reproducibility. This mode of access is significant, influencing the architecture of the underlying data grid, as explained in §3. For the D&F groups, there is a requirement to support occasional file modifications, for log book information that accompanies image data and encapsulates a history of any analysis that has been performed.

(*b*) *Make-up of community*

The D&F groups, as noted above, are split between two laboratories in Edinburgh and Oxford, and the researchers in these groups frequently need to share data. The groups also undertake projects with collaborators from other centres, both in the UK and internationally. These collaborations can involve four or five centres, each contributing different expertise while requiring access to the same data.

The UKQCD consortium includes significant contributions from more than 10 UK-based institutions. As a mature consortium, it is commonplace for the UKQCD to share their resources and work in dynamic collaborations for specific projects.

(*c*) *User interface*

If one spends a little time observing the two groups, one begins to appreciate the gulf that exists between their preferred computer environments. A Unix workstation is the tool of choice for a physicist, who appreciates the flexibility of the command line, the ability to create batch scripts or more substantial programs, and even to implement algorithms in an architecture-specific machine code.

By contrast, a molecular biologist is likely to be more comfortable using high-level (graphical) applications that distance them from the underlying system; for example, working at a laboratory bench on a Mac laptop.

(*d*) *Collated requirements*

For the UKQCD, the main motivation for setting up a data grid is to combine storage capacity available at collaborating sites to provide a resource capable of hosting the volumes of data generated by specialized facilities, such as the QCDOC supercomputer (Boyle *et al.* 2004). At the time of writing, the group has a requirement for 100 TB of storage, though this will reach of the order of PB within the next 5 years.

Implicit in the above requirement is a need for secure access to data from distributed locations, and a need to support significant peaks and troughs in data generation and consumption rates, as dictated by individual projects.

The primary data management requirement of the UKQCD is high availability of data, achieved by replication of datasets to (at least) two geographically distinct sites. Data replication is mainly intended to improve data availability—for example, to mitigate for downtime on a storage resource—though replication also leads to better system resilience, reducing the potential for loss of data due to corruption.

Looking beyond the UK, the consortium is a member of the International Lattice Data Grid (ILDG; Coddington *et al.* 2007), a community of scientists from countries as far spread as Japan, Australia and the USA, who have ambitious targets to share their data and to accelerate progress in the field of lattice QCD. The UKQCD has adopted ILDG-wide standards for the format and annotation of data; support for (and promotion of) these standards is another requirement for the consortium.

For the D&F groups, the main motivation for a data grid is to give researchers, at the two laboratories, access to a common view of shared data and an ability to analyse the data at the particular site where the appropriate tooling and expertise are available. This use case encapsulates the aims of an in-progress pilot study. Based on its success, the intention is to widen the application to include other centres.

Given the user interface preferences of the biologists, a key requirement is to integrate the data grid functionalities into the existing (i.e. familiar) client tools. To this end, the study has focused on the work of the Open Microscopy Environment (OME; http://www.openmicroscopy.org/) that has developed standards for formatting and describing microscope image data, along with tools to support these standards.

3. Architecture of a data grid

Based on the requirements above, we have developed DiGS, a Globus Toolkit-based application (http://www.globus.org/toolkit) that links together low-cost, commodity Linux-based storage to provide a scientific data repository for

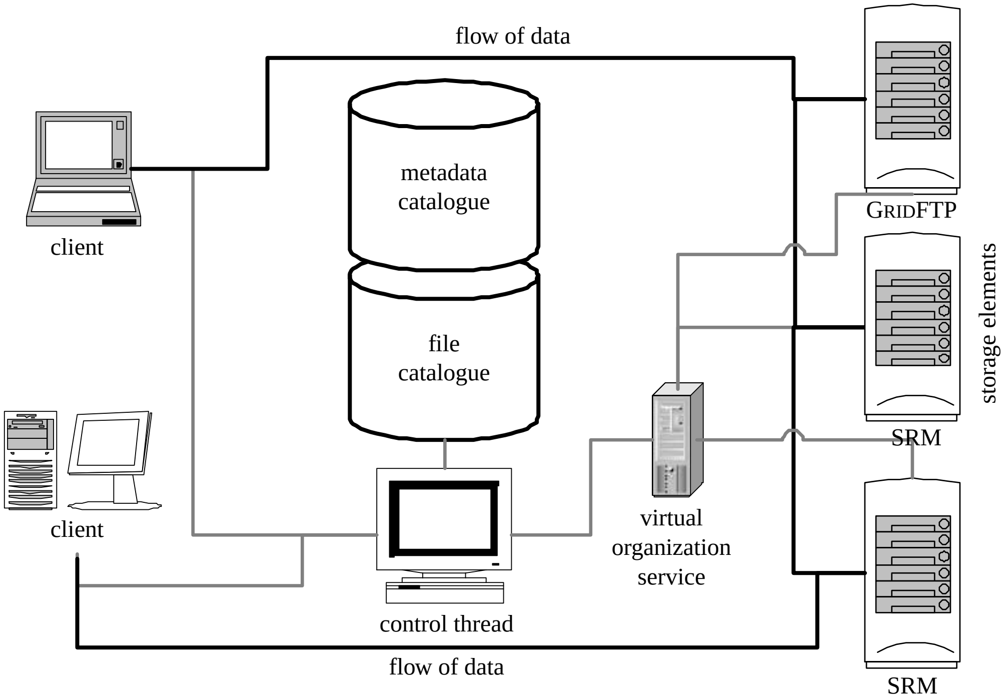

Figure 1. The make-up of a DiGS data grid infrastructure. (SRM is Storage Resource Manager.)

collaboration. A typical DiGS installation is shown in figure 1, with key components: a file catalogue (FC) that tracks the locations of data; a metadata catalogue (MDC) that holds scientific annotations of data; storage elements (SEs) that hold the actual data; a virtual organization (VO) service that maintains the VO of users of the system; clients that allow these users to manage their data; and a control thread (CT) that maintains the health of the system. We describe each of the components in more detail below.

(*a*) *File catalogue*

At the heart of a data grid is a service called the FC, which maintains a mapping between a unique and persistent identifier for a dataset (the logical filename (LFN)) and one or more pointers to actual copies of the dataset.

In contrast to a conventional file system, the location of data within a data grid is not fixed, and is likely to change over time. This can happen, for example, when a storage resource is added or withdrawn, or when data are replicated to a new site. The primary function of the FC is to keep track of any such changes; the uniqueness and persistence of the LFN is essential for this function.

A data grid may hold more than one instance of a dataset. Reasons for having multiple copies (*replicas*) include: improved availability of data, in the event that a resource is unavailable; data resilience, to mitigate against the risk of data corruption; and convenience, with data located close to sites where they are required.

There are a number of FC implements available, such as the gLite FC (http://glite.web.cern.ch/) and storage resource broker (SRB; http://www.sdsc.edu/srb/). DiGS uses the Globus Toolkit Replica Location Service (RLS). It is a wrapper around a relational database (MySQL (http://www.mysql.org/), by default), providing a high-level interface that exposes functions typically required for FC operation. RLS allows the definition of additional fields for dataset records and, using this function, DiGS includes file-based metadata such as file length, checksum and author (submitter). This information is useful for other components of the system.

Completing the mapping between a dataset's LFN and its locations is critical to locating the actual data. For this reason, the catalogue should be a highly available service. To this end, DiGS replicates the catalogue contents to a second server that automatically takes over (in read-only mode) if the primary server fails.

(b) *Metadata catalogue*

The MDC provides scientifically meaningful annotations of data that can be searched by a user. Whereas, the FC stores information useful to data grid services, the MDC holds information interesting to the user.

The form of the scientific annotation is inherently application-specific, although typically has several characteristics. Firstly, an annotation may refer to a single dataset or to a collection of related datasets (for example, an experiment). A key use of the MDC is to identify interesting data without accessing the actual storage resources. Secondly, an annotation is small in size in comparison with the dataset(s) it describes. The intention is that the MDC can be hosted as a central service—for example, in a database—in order that searches are both fast and effective.

Within a community, it is advantageous to have agreement on the form and content of scientific annotations (or metadata). This helps to reduce the risk of ambiguous specification or misidentification of data, and assists with data curation.

A technology such as extensible markup language (XML; http://www.w3.org/XML/) can greatly facilitate the formalization of a scheme for annotations. It promotes the use of hierarchical markup, includes a mechanism for enforcing schema definitions and provides devices for extending schemata while maintaining backward compatibility.

For the two applications considered herein, community-wide standards for metadata are being developed. For the UKQCD, this is a remit of the ILDG (Maynard & Pleiter 2005). For the D&F groups, standards are provided by OME (Swedlow *et al.* 2003). In both cases, XML is the adopted markup and has proved an effective vehicle for the types of query that users want to make. Users may include free text comments within the annotation. However, this practice is discouraged, as such comments introduce the possibility of ambiguity or misinterpretation, adversely affecting the curation process.

Unfortunately, while XML is ideally suited to capture and curate scientific metadata, the syntax of the markup is likely to be unfamiliar to a user. For this reason, it is crucial that they are insulated from raw XML by appropriate client tooling. This is a focus for the development of DiGS, as is discussed later.

The separation of the MDC and FC is by no means a universal approach. For example, the SRB provides a single catalogue (the MCAT) for both file location information and scientific annotations. From our perspective, there are several advantages to separating the function of the MDC and FC. Firstly, the structure of the FC and MDC is likely to be different. The FC is a registry, typically accessed to retrieve information about a given LFN. This specificity of use implies that steps can be taken to optimize the registry for the most common types of look-up. The MDC, by contrast, is a data resource that can have arbitrary structure, as dictated by the application, and be queried in numerous manners. Secondly, information contained in the MDC and FC may be subject to different levels of access control. For example, in the UKQCD the MDC is publicly available, whereas the FC contains more sensitive information available only to registered users. Hosting the MDC and FC separately simplifies the application of access control policies.

(*c*) *Storage elements*

Fundamental to a data grid is the SE, where files are actually stored, and from (to) which files can be downloaded (uploaded). A data grid usually has numerous SEs that (as indicated by the requirements above) are distributed across the sites contributing to the grid.

Unlike the FC and MDC, there is likely to be less capacity to control the form and configuration of SEs, since they are likely to be administered by site-specific support teams, and be of varying capacity, capability and reliability. The SE provision of a data grid is likely to be heterogeneous in nature, including protocols such as Hypertext Transfer Protocol (HTTP (S)), File Transfer Protocol (FTP) and GridFTP. This heterogeneity has always presented a problem for file sharing in distributed collaborations, though the situation has improved significantly in recent years thanks to the introduction of the Storage Resource Manager (SRM) protocol (Donno *et al.* 2008). Trends in EGEE and TeraGrid suggest that the grid community is converging towards SRM as a simple means to communicate with site-specific fabrics. In the same way that SRM is gaining acceptance as the standard for interactions with SEs, GridFTP (part of the Globus Toolkit) is becoming a widely accepted service for file transfer operations.

At the time of writing, DiGS combines the roles of SE management and file transfer service into a single entity, GridFTP. Thus far, this has satisfied the needs of the two applications, though it imposes several limitations. Firstly, GridFTP only supports online (disk-based) storage. Use of mass storage, such as tape, is not possible. Secondly, to overcome some of the limitations of GridFTP as an SE manager, some DiGS components need to be installed onto the SE. Given that an SE may be within the control of a third-party entity, it may not be possible to install these components on every storage resource.

To overcome these deficiencies, the DiGS team is planning to separate the roles of an SE manager and the file transfer service, in the next version of the software.

(*d*) *Security and virtual organization management*

For a data grid that spreads across more than one administrative domain, access control becomes problematic, as it depends on the (possibly incompatible) policies of the different providers. Moreover, a simple authorization implementation based on direct user registration with each resource is generally not practical.

Within the academic community, at least, grid infrastructures rely on X.509 digital certificates (Welch *et al.* 2004) to allow a user to authenticate to remote resources, and to be attributed with appropriate privileges to complete their work.

The grid security infrastructure (GSI)—part of Globus—encapsulates the functionality to handle X.509 certificate-based authentication and authorization, providing: a distributed security infrastructure spanning organizational boundaries; support for secure communications between grid resources; and *single sign-on*, for completion of multiple tasks without reauthenticating.

The management of a grid community—a VO—is handled by a VO management service, such as VOMS (Alfieri *et al.* 2005), which aggregates membership information in a form that is accessible to data grid resources.

In terms of authentication and authorization, the interface between the grid and the local infrastructure is still relatively immature. For the Globus Toolkit, the reference approach (*grid-mapfile* (Alfieri *et al.* 2005)) attributes credentials to the holder of a certificate by mapping them to a local account on the target host.

DiGS uses GSI to authenticate a request to the data grid and then maps the presenter to a low-privilege, generic account on the local resource, to complete the request. In this way, the security of grid data is delegated to the local resource—and in the case of file access, is handled by file permissions—rather than being controlled in the grid layer. A more effective approach would involve making authorization decisions at the grid layer—for example, within the GridFTP service code. However, at the time of writing, this functionality is not supported by Globus.

(*e*) *Client tooling*

As we have noted above, the user base for DiGS encapsulates a range of levels of computer confidence. The experience a user has with the data grid significantly affects both their effectiveness with and enthusiasm for the system. That a grid infrastructure is a complicated entity is inescapable, though this complexity should not be apparent to the user. Furthermore, the system needs to fit seamlessly into the user's existing process of work.

With this in mind, we have developed very different client tools for the two groups with which we are working. For the UKQCD, command-line tools support data grid operations such as list, commit and retrieve. These can easily be incorporated into shell scripts to facilitate the automation of bulk data transfers. A graphical client—the DiGS browser—has also been built, allowing a user to search the MDC, without any knowledge of XML or the underlying query language (XPath).

For the D&F group, the DiGS browser is being adapted to fit into the existing (microscope-based) image acquisition processes. Implicit to this task is the integration of the browser with the OME client tools.

(*f*) *Control thread*

The CT is a persistent and autonomous process, often hosted on a dedicated node, which continuously monitors and validates the integrity and availability of the data grid. It performs a number of tasks, as noted below.

- — It checks if new datasets are awaiting insertion into the data grid.
- — It checks if sufficient replicas of each dataset are present on the data grid.
- — It monitors the health (availability and capacity) of SEs, and checks whether their content is consistent with the file metadata held in the FC.
- — It manages file transfers, retrying or revising (by choosing alternative source and/or destination) any that fail. This implies that, typically, a user does not encounter (or deal with) failed data transfers.

The CT is a data grid component that is specific to DiGS; it is not a typical component of a data grid, though we note that the Globus Toolkit includes a data replication service that has some overlap in terms of functionality.

There are several philosophies that govern the operation of the CT, which we believe are important considerations for grid applications more generally. Firstly, the CT maintains a target state for the data grid, which it converges towards through a sequence of small steps (such as file transfers and replications). The target state is defined as a set of policies, describing how many copies of each file should be maintained, how much free space should be preserved and so on. If the target state were to not change, and if all of the grid resources persisted and functioned correctly, then the CT would—after some time—reach the target state. However, in practice, this rarely happens because the target state is always changing, due to user intervention (for example, adding new datasets) and changes in the underlying fabric (hardware failures, new resources being added and so on). As changes occur, the CT simply adjusts its course towards the new target state.

Secondly, the CT is designed to assume that no operation (i.e. a step towards the target state) is guaranteed to succeed. Wherever possible, the CT changes the grid in atomic steps, which either succeed and achieve the desired result, or fail and result in no change—for example, creating a (single level) new directory on a remote resource can be regarded as an atomic operation. Unfortunately, not all steps towards the target state can be reduced to atomic operations. Consider, for example, a file transfer, which could fail in progress, leaving a partial file stub on the destination resource. Whatever precautions one takes, one cannot eliminate the risk of this type of failure. Thus, for such cases, checks are included in the operation of the CT to rectify such problems after the event.

4. Scientific impact

For the two communities considered, the overriding objective of setting up a data grid has been to enhance their ability to progress their science. Both groups have achieved this objective, as we explain below.

The UKQCD has had a DiGS system since 2002. In the 6 years to date, they have amassed a significant portfolio of lattice QCD primary data, both for the consortium and—through the ILDG—for the wider community. At the time of writing, the data grid contains over 70 000 datasets (equating to approx. 30 TB), which is readily available to the 80 or so registered users of the system.

As a pertinent example of the importance of the data grid to the UKQCD, we consider a recent study of flavour-singlet mesons and glueballs using an improved staggered fermion formulation (Gregory *et al.* 2008). For this study, a family of primary datasets was generated on the QCDOC super-computer (Boyle *et al.* 2004) in Edinburgh, using specially optimized code from the Columbia Physics System suite. These data were analysed—using QCDOC and computer cluster resources in Liverpool and Glasgow—with the Chroma software suite (Edwards & Joo 2005) to measure relevant statistical correlations and hence determine the mass of subatomic particles (such as the η and η' mesons and glueballs). The primary data, which amount to 7.2 TB, took 3 years to generate. They were uploaded to the data grid as they were produced, from where they were immediately available as required by the teams in Liverpool and Glasgow. The data grid now acts as an archive for these valuable data, which have since been published to the ILDG through the DiGS system.

Two other examples of high-impact UKQCD projects, for which the DiGS system has been of significant and acknowledged benefit, are Gockeler *et al.* (2006) and Allton *et al.* (2008).

The D&F groups have, at the time of writing, been trialling a DiGS system for approximately 6 months and, although only being available to a subset of users in a pilot phase, key benefits of the system have already been realized. For example, a previous study of bioinformatics searches, for particular structures that direct RNA localization, was undertaken by the Oxford and Edinburgh laboratories (Hamilton *et al.* 2009). In the absence of a data grid, images were shared in an ad hoc manner, with datasets being passed between laboratories without tracking or any form of version control. By contrast, the next such study will use the data grid, so the whole team has access to the most up-to-date image data and it is metadata, reducing the capacity for mistakes.

Looking to the future, the real benefit of the DiGS system to cell biologists goes beyond the distribution of microscope image files, which could (to some extent) be handled with sFTP, for example. The strength of DiGS is the inclusion of scientific metadata. Without the metadata, the images would likely be annotated in an ad hoc manner—for example, in free text. Such ad hoc annotation is totally inadequate on the scale of a typical collaboration, which hosts thousands of microscopic images.

5. Conclusions

In this paper, we have demonstrated how small scientific groups, working with modest resources, can reap benefits from the use of a data grid—a technology more usually associated with large-scale, high-impact experiments akin to the LHC. We have described the typical make-up of a data grid, building on our experience of developing DiGS. We have explained the crucial role of catalogue services in maintaining both file management information and scientific annotations, and have highlighted the use of VO management tools to mitigate for the complexities of securing data in a distributed environment. We have also described how the CT, a non-standard data grid component unique to DiGS, can help to maintain the integrity of the infrastructure. Finally, we have summarized some recent scientific successes that have been achieved with the help of DiGS to support the management, publication, sharing and curation of the data.

Development of DiGS is supported by the GridPP programme, an STFC-funded collaboration between particle physicists and computational scientists from the UK and CERN, who are building a grid for particle physics. *A Data Grid for Cell Biology* is an activity supported by the STFC PIPSS Programme.

References

Alfieri, R., Cecchini, R., Ciaschini, V., dell'Agnello, L., Frohner, Á., Lőrentey, K. & Spataro, F. 2005 From gridmap-file to VOMS: managing authorization in a grid environment. *Future Gen. Comput. Syst.* **21**, 549–558. (doi:10.1016/j.future.2003.10.002)

Allton, C. *et al.* 2008 Physical results from 2+1 flavor domain wall QCD and SU(2) chiral perturbation theory. *Phys. Rev. D.* **78**, 114 509. (doi:10.1103/PhysRevD.78.114509)

Boyle, P. A. *et al.* 2004 QCDOC: a 10 teraflops computer for tightly-coupled calculations. In *Proc. 2004 ACM/IEEE Conf. on Supercomputing.*

Coddington, P. *et al.* 2007 Towards an interoperable International Lattice Data Grid. In *Proc. XXV Int. Symp. on Lattice Field Theory.*

Donno, F. *et al.* 2008 Storage resource manager version 2.2: design, implementation, and testing experience. *J. Phys. Conf. Ser.* **119**, 062 028. (doi:10.1088/1742-6596/119/6/062028)

Edwards, R. G. & Joo, B. 2005 The Chroma software system for lattice QCD. *Nucl. Phys. Proc. Suppl.* **140**, 832. (doi:10.1016/j.nuclphysbps.2004.11.254)

Gockeler, M., Horsley, R., Irving, A. C., Pleiter, D., Rakow, P. E., Schierholz, G., Stüben, H. & UKQCD and QCDSF collaborations 2006 A determination of the lambda parameter from full lattice QCD. *Phys. Rev. D* **73**, 014 513.

Gregory, E. B., Irving, A. C., McNeile, C. & Richards, C. M. 2008 A high statistics study of flavour-singlet mesons with staggered fermions. (http://arxiv.org/abs/0810.0136)

Hamilton, R. S., Hartswood, E., Jones, C., Vendra, G., Van De Bor, V., Finnegan, D. & Davis, I. 2009 A bioinformatics search pipeline, RNA2DSearch, identifies RNA localization elements in *Drosophila* retrotransposons. *RNA* **15**, 200–207. (doi:10.1261/rna.1264109)

Maynard, C. M. & Pleiter, D. 2005 QCDml: first milestones for building an International Lattice Data Grid. *Nucl. Phys. B* (*Proc. Suppl.*) **140**, 213–221. (doi:10.1016/j.nuclphysbps.2004.11.116)

Perry, J. T. *et al.* 2005 QCDgrid: a grid storage resource for quantum chromodynamics. *J. Grid Comput.* **3**, 113–130. (doi:10.1007/s10723-005-9005-5)

Swedlow, J. R., Goldberg, I., Brauner, E. & Sorger, P. K. 2003 Informatics and quantitative analysis in biological imaging. *Science* **300**, 100–102. (doi:10.1126/science.1082602)

Welch, V., Foster, I., Kesselmann, C., Mulmo, O., Pearlman, L., Gawor, J., Meder, S. & Siebenlist, F. 2004 X.509 proxy certificates for dynamic delegation. In *Proc. 3rd Annual PKI R&1D Workshop.*