The Promise of Data Grids in the LHC Era

Paul Avery

Department of Physics, University of Florida, Gainesville, FL 32611-8440, USA

Abstract

The LHC era brings unprecedented challenges in information technology: (1) providing rapid access to massive data stores of 100 PB or more and (2) providing transparent access to heterogeneous computing resources throughout the world across an ensemble of networks of varying capability and reliability.

I discuss here how these challenges can be met by a hierarchical computational Grid of data analysis centers linked by Gbps networks. Such a *Data Grid* will allow physicists to play key roles in all stages of the data analysis, from development of the reconstruction programs and software infrastructure, to the extraction of first physics results. This capability will be extended to physicists at their home institutions through the use of an integrated distributed Data Grid system architecture designed for efficient petabyte-scale data access and analysis.

Finally, I describe briefly a new US initiative called **GriPhyN** (**Gri**d **Phy**sics **N**etwork)[1] which aims at developing the first production-scale Data Grids for US-ATLAS, US-CMS, LIGO (Laser Interferometer Gravitational-Wave Observatory) and SDSS (Sloan Digital Sky Survey). The network-monitoring and control software, and the modeling and optimization systems and methods, will be widely applicable functional components that should also drive the design and implementation of future distributed systems in HEP and many other fields of science, engineering and industry.

Keywords: grid, computational, lhc, middleware, network, distributed, computing, tier2, beowulf, dataset, petabyte, petascale, virtual, griphyn

1 Introduction

Realizing the scientific wealth of the LHC experiments presents new problems in data access, processing and distribution, and collaboration across national and international networks, on a scale unprecedented in the history of science. Numerous challenges in information technology must be met: (1) providing rapid access to event samples and subsets drawn from massive data stores that will reach hundreds of petabytes after 2010; (2) enabling transparent access to computing resources throughout the world; (3) extracting small or subtle new physics signals from large and potentially overwhelming backgrounds; and (4) enabling access to the data, and to the rest of the physics community across an ensemble of networks of varying capability and reliability, using heterogeneous computing resources. Solutions meeting these challenges must moreover be flexible enough to track changes in computing technology over the decades-long lifetimes of the experiments.

The computing configuration known as a *Computational Grid*[2, 3] has recently been proposed as a foundational technology for addressing these challenges. A Grid is a set of geographically separated computing resources (typically connected by high-speed networks) that can be applied to a single computational problem. Although widely separated computers have been used to solve specific problems (factorization of large numbers, SETI calculations, etc.), the key area

that must be addressed is how such resources can be mobilized *transparently*, i.e., so that the underlying details of remote resource discovery and acquisition, scheduling, distributed computing, and data transport can be hidden from the user.

Prototype Grid projects in several areas of science and engineering are planned or underway[4] in the United States and Europe. However, they do not go far enough in addressing the data intensive nature of LHC computing, discussed in more detail in the following section.

2 Data Grids and the LHC

The crux of the LHC problem is petabyte-scale data. Although proposed computing and data handling systems and network throughputs are large by present standards, they will not support on-demand access, transport or reconstruction of more than a minute fraction of the data, even in compact pre-processed forms.

Within this sobering framework, studies[5] have shown that a distributed hierarchy of facilities linked by high speed networks offers the most promising distribution of computing resources, representing a balance between (1) proximity of the data to central computing and data handling resources, especially for large-scale production-oriented jobs; (2) proximity of frequently accessed data to the users; (3) making efficient use of limited network bandwidth; especially transoceanic; (4) making appropriate use of regional and local computing and data-handling resources; and (5) involving scientists and students from each world region in the analysis and the physics.

2.1 A candidate LHC Grid hierarchy

These considerations plus similar broad trends in other scientific fields and industry have led the US-based groups in ATLAS and CMS to propose a particular form of hierarchical Grid as an appropriate architecture to meet the demands of LHC data analysis. The proposed hierarchy consists of five levels of resources:

- Tier 0: The central facility at CERN
- Tier 1: A national regional center, e.g. Fermilab (CMS) or Brookhaven (ATLAS)
- Tier 2: A center in one region of the country
- Tier 3: The computing resources of a single institutional group
- Tier 4: An individual workstation.

This configuration, which has broad applicability to all the nations participating in the LHC, is termed a *Data Grid*, reflecting the design strategy that each Tier is defined chiefly by the storage and I/O throughput capabilities of each of its elements. The Data Grid is more than just a clever redefinition of existing resources because it makes it possible for the first time to coherently mobilize these resources, thus extending the range and number of analyses that can be performed. Stated succinctly, the whole is greater than the sum of its parts.

The Tier 2 centers represent a novel and interesting computing resource, occupying a role that can be thought of as the "geometric mean" of a national computing center and an institutional center. This scaling argument is reflected in several ways: (1) the planned number of US Tier 2 sites (5 per experiment), (2) the size of each site (10–20% of the Tier 1 center) and (3) the Tier 2 mission, which would be somewhat less production oriented than that of the higher level Tiers and thus more agile in responding to changing analysis priorities and needs. Since Tier 2 centers have less need for 24×7 operation, they can be constructed from cheaper commodity components and supported by a local team with physicist help, further reducing costs.

In this model, a Tier 2 center would consist of a medium-scale (128 – 256 node) Linux computer farm, with attached commodity disk, connected to a high-speed LAN switch (sometimes

called a “Beowulf” cluster), a data server with attached RAID array for serving physics data objects and possibly a small tape robot with a capacity several times that of the disk cache. It is envisioned that a multi-gigabit backbone will connect the Tier 2 sites to each other and to the Tier 1 centers and CERN, thus enabling the rapid data movements necessary to balance the computing and I/O loads across the Grid. Universities and other institutes would connect to the backbone through existing or upgraded connections (Internet 2 in the US), completing the hierarchy.

The criteria for Tier 2 site selection should be flexible enough to adapt the Data Grid hierarchy to particular characteristics of the host nation(s), but as a general rule they should be geographically dispersed, able to connect reasonably easily to high-speed networking, and be located in areas that can take advantage of skilled R&D personnel. In the US, for example, intellectual resources are concentrated in widely dispersed universities which are (generally) well-connected to high-speed networks. This situation coupled with federal funding priorities argues for Tier 2 siting at universities in different US regions. The US groups are considering other flexibility options, for instance allowing Tier 2 centers to specialize in different subdetectors and/or physics analyses, while preserving location independence of the users.

2.2 Physical implications of the Data Grid

The organization of computing resources in a Data Grid hierarchy provides important advantages, some of them serendipitous. The primary advantage is *better resource utilization* due to the fact that computing resources are part of a larger unified system. A unified system is simply better than a fragmented one at averaging over spikes in usage and discovering, scheduling and utilizing resources, which in turn extends the effectiveness of these resources and permits them to be used more efficiently. This is an important consideration in an era where computing resources are expected to be small relative to the demands placed on them.

The distributed nature of the Data Grid hierarchy is also *balanced* with respect to data locality and network utilization. Massive datasets would be kept close to the central computing and data handling facilities of CERN or the national Tier 1 centers while frequently used physics data would be cached at the Tier 2 centers or at the institutes, depending on its size and frequency of access. This adaptive distribution of data relies on the dependable operation of a high-speed network fabric that is most efficient (i.e., avoids bottlenecks) when data is distributed throughout the fabric.

2.3 Human implications of the Data Grid

There are also important human considerations that favor configuring resources as a Data Grid rather than concentrating them in a national center or at CERN. First, a laboratory cannot reasonably manage or help many hundreds or even thousands of users without imposing strong rules and in general limiting user flexibility. The ability to leverage resources, maintain control and set priorities is far easier at the Tier 2 centers which are small enough to respond flexibly and relatively quickly to new situations. Second, the Tiers are clearly differentiated in functionality, a fact that allows complementary funding and targeted initiatives such as the one described in the next Section. In the US, for example, DOE prefers to fund the Tier 1 laboratories while NSF prefers to fund universities. Third, the Data Grid, through the use of Tier 2 centers and Tier 3 institutes, increases the effective involvement of scientists and students, regardless of location.

Finally, the Data Grid provides the means to implement a global (or even regional) collaboration strategy for prioritizing and routing computing requests, ensuring that resources are utilized efficiently while delivering acceptable turnaround times to the entire ensemble of tasks. Developing a successful Grid-based global strategy involves determining, under dynamically changing and

even incompletely measured conditions, an "optimal" solution to several problems: (1) Meeting the demands and prioritizing the requests of hundreds of users from local and remote communities who need transparent access to local and remote data in disk caches and tape stores; (2) Structuring and organizing the data, and providing the tools for locating, moving and scheduling data transport between tape and disk and across networks, so that it can be accessed and/or delivered rapidly and transparently; and (3) Ensuring that the overall system is dimensioned correctly to meet the aggregate need.

3 The GriPhyN Project

Many of the ideas discussed in this paper are embodied in a large-scale effort called **GriPhyN** (**Gri**d **Phy**sics **N**etwork)[1] which aims at developing the first production-scale Grids for US-ATLAS, US-CMS, LIGO (Laser Interferometer Gravitational-Wave Observatory) and SDSS (Sloan Digital Sky Survey). **GriPhyN** is a joint project by a collaboration of computer scientists and senior physicists and computing personnel from the four experiments. They propose to design, develop, prototype, deploy and field-test a new generation *Petascale Virtual Data Grid* (PVDG) that will meet the data-intensive computational needs of the diverse community of thousands of scientists spread across the globe. The key concept they add is that of *virtual data*, a term which recognizes that all except irreproducible raw experimental data need 'exist' physically only as the specification for how they may be derived or computed. The grid may instantiate zero, one, or many copies of derived or computed data depending on probable demand and the relative costs of computation, storage and transport. These considerations are not trivial: in high-energy physics today, over 90% of data access is to derived or computed data.

The success of the **GriPhyN** project will require multifaceted advances in Information Technology, including new scalable information models, network-distributed resource management constructs, and interactive workflow management models based on resource discovery coupled to system state tracking and forward prediction[3]. While the new generation Grid-based systems to be developed will be applicable to large-scale data-intensive problems in many fields of science, engineering, and eventually industry and commerce, the initial field of development and testing will be that of the experiments in this proposal. Information systems of this size and complexity providing transparent access to massive sets of raw and processed data will be needed in the coming decades as a central element of our information-centric society.

References

1 `www.phys.ufl.edu/~avery/mre/`.

2 I. Foster, C. Kesselman (eds.), "The Grid: Blueprint for a New Computing Architecture", Morgan Kaufmann, San Francisco, California, 1999; `www.globus.org`.

3 See talks by M. Livny and W. Johnston in these proceedings.

4 A not exhaustive list of grid projects includes (1) NASA's Information Power Grid (discussed by W. Johnston at this meeting); (2) GUSTO (`www-fp.globus.org/overview/testbeds.html`) undertaken by Globus collaborators, (3) the $\beta-$Grid project (`dsl.cs.uchicago.edu/beta/`), a research grid being developed by the University of Chicago, (4) the MILAN research project (`www.cs.nyu.edu/milan/`) and (5) NetSolve (`www.cs.utk.edu/netsolve/`). In addition, there are some EU projects.

5 MONARC 98/1 and 99/3. See `www.cern.ch/MONARC/docs/monarc_docs.html`.