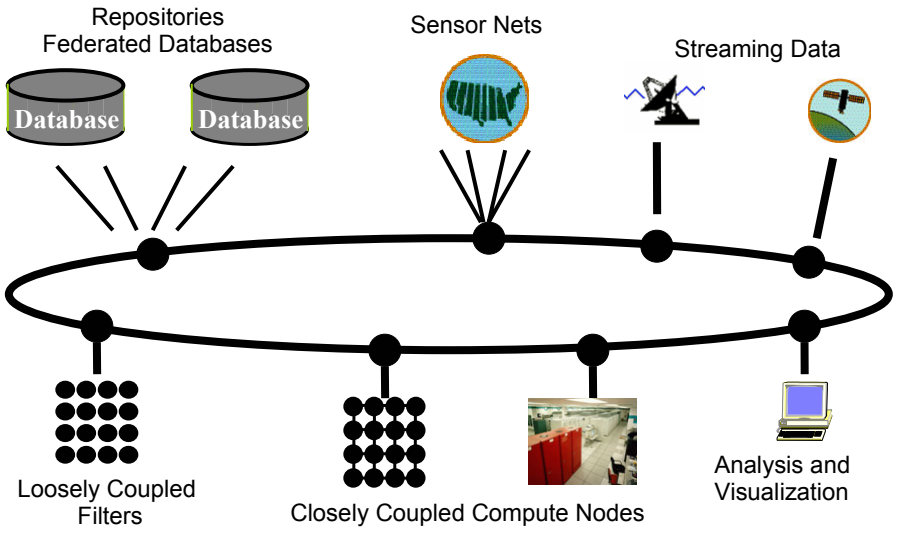
Integration of Computing and Information on Grids

Geoffrey Fox
Indiana University
Computer Science, Informatics and Physics
Community Grid Computing Laboratory,
501 N Morton Suite 224, Bloomington IN 47404
gcf@indiana.edu

We have discussed in earlier articles many aspects of Grids and here we look in more depth at some of the different capabilities they must support. There are several definitions of Grids and here we will adopt two views from our recently published book *Grid Computing: Making the Global Infrastructure a Reality* edited by Fran Berman, Tony Hey and myself. (http://www.grid2002.org/)

- Grids support e-Science representing increasing global collaborations of people and of shared resources that will be needed to solve the new problems of Science and Engineering.
- Grids provide infrastructure that will provide us with the ability to dynamically link together managed resources as an ensemble to support the execution of large-scale, resource-intensive, and distributed applications.

These descriptions emphasize the high-level user requirement and the system capability respectively. We are integrating resources, managing them and using them to support distributed collaborative engineering and science.



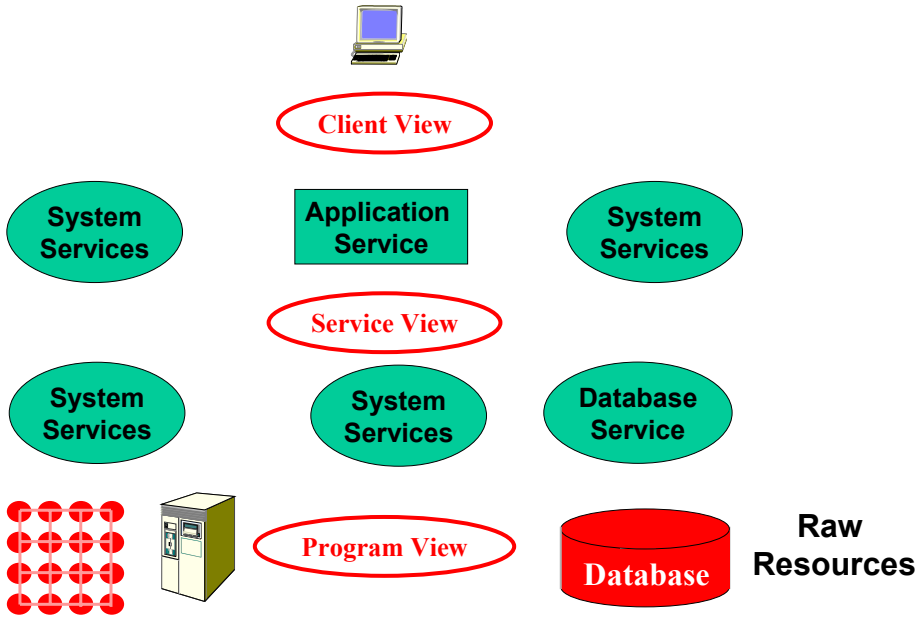
The caricature shown above shows the rich variety of resources one might wish to use in a typical e-Science scenario. We show raw data from a variety of small and large sensors and satellites; curated processed data typically stored in some variety of database; computing resources both as massively parallel machines and as a pool of independent workstations processing largely independent tasks and finally the user interface with its visualization and analysis capabilities. Grids must cope with such multiple heterogeneous resources and integrate them together. In this article, we will discuss how this requires the meeting of three worlds – that of the computer, that of the database and that of sensor. It

is as always hard to draw clearly lines between different ways of doing e-Science and different ways of looking at data. Data can be streamed from sensors, stored in large files, produced as a visualization or science output of a program, or stored in databases; further we have the conventional pipeline Sensor—Data—Information—Knowledge describing increasing refinement as bits are passed through some analysis process. We here focus on two extreme models

- "Run-a-job" Grids where one executes some linked programs that typically fetch and store data in files and run on multiple remote computers that may or may be high performance (massively parallel)
- "Information" Grids where one accesses a set of databases holding either meta-data or the raw information itself.

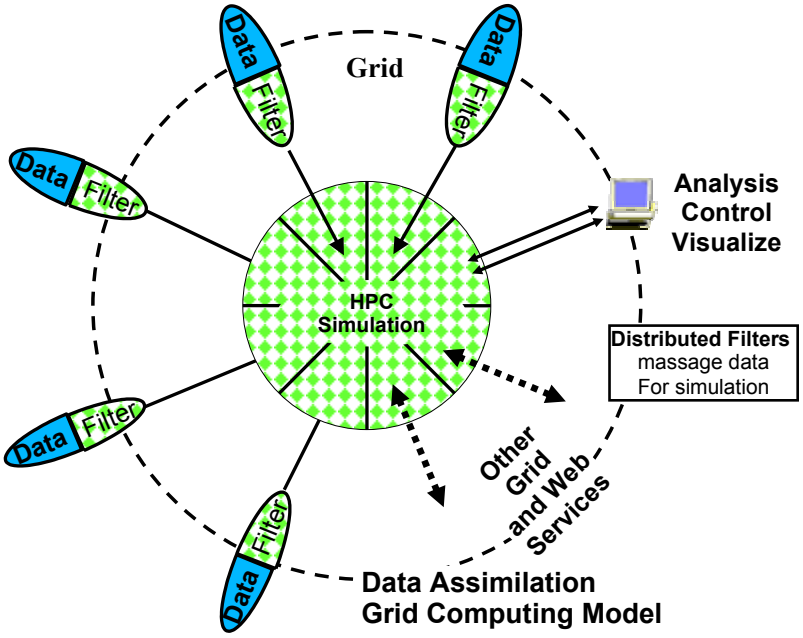
"Run-a-job" Grids at their simplest, support remote "shells" that allow the user command line interfaces to executing jobs on remote computers. Sophisticated systems like Globus and Condor allow the jobs and their associated files to be on geographically dispersed machines; further they support the scheduling of many simultaneous jobs. The challenge of this type of Grids is clearly seen in the initial stages of the analysis of data from particle physics experiments. The international science collaborations formed for experiments at the CERN LHC accelerator will process the raw data from their detectors on a mammoth world wide network of computers which will need to support tens of thousands of simultaneous jobs running reliably 24 hours a day reading the tens of petabytes of data produced each year. Clearly the networking, reliability and management issues are of staggering complexity.

Information Grids are typified by applications like virtual observatories and bioinformatics where the typical service is accessing a database. In each case, the database holds either the meta-data or highly processed data itself from the many experiments in astronomy and biology. Bioinformatics has developed this model over many years with several sites across the world responsible for ensuring the data is properly entered and maintained in the database. This data curation activity seems likely to grow in importance in many fields and become an essential feature and driver of Grids.



In Information Grids today, one naturally starts with a "service" model where information is extracted by sending messages to a "database service". This is rather different from the classic Unix shell "run-a-job" model. In the figure above we depict the typical multi-tier "service-architecture" view of a Grid and overlay the "program" "service" and "client" levels. Information Grids are naturally discussed at the "service" level which has a richer functionality but typically lower bandwidth than the "program view" where the "run-a-job" Grids operate. We have in earlier articles discussed the OGSA (Open Grid Service architecture) approach which will unify Grids giving all capabilities a service view. Today most information Grids adopt a pure Web service approach and most "run-a-job" Grids use Globus; these will be unified as OGSA gets adopted with Globus Toolkit 3 (GT3) implementing these ideas with preliminary implementations already available. We discussed Grid Computing Environments in the March/April Issue. Here one has already seen many successful service models built on top of Globus and so we can expect the more systematic approach of OGSA to be successful. A critical development for Information Grids is the OGSA-DAI (DAI is Data Access and Integration) software produced by the UK e-Science program. This has designed a Grid service interface for essentially all important databases (including both relational and XML) and as OGSA-DAI matures we can expect it to be the standard building block of Information Grids. As well as providing a uniform interface to databases, OGSA-DAI will support distributed query across multiple databases and the integration of filters that provide customized views of a data source as depicted below.

These filtered views would be described initially as a Web Service specified by WSDL (Web Service Definition Language). This WSDL Specification is easily upgradeable to be Grid Service compatible by adding the extra OGSI (Open Grid Services Infrastructure) features. So we see a reasonably clear model for building new Information Grids with Web Service based filters linked to OGSA-DAI wrapped databases.



In fact most applications need to combine the features of Information and "run-a-job" Grids and we finish by briefly discussing "complexity" grids so-called as they support the emerging fields of Geocomplexity and Biocomplexity. The figure shows one natural way to look at the problem as generalized data assimilation where the Grid manages distributed data sources which could be databases (biology), sensor nets (environment) or streams from multiple satellites (earth science). This data can be preprocessed in a distributed fashion as it is pruned and transformed for use in large scale simulations. Note that the anticipated "data deluge" is expected to produce so much data that substantial filtering could be needed to project the data onto those components of greatest value to guide the simulation. Actually one could consider particle physics to have the general structure of the figure. The data corresponds to the raw events from the accelerator; the filters correspond to the initial processing to produce data summary tapes; the data assimilation "simulation" corresponds to the physics analysis phase. In the latter case one typically needs a large parallel machine for the PDE dominated fields like climate, ocean, and weather simulations. In the particle physics case, largely uncoupled machines are necessary to support Monte Carlo simulations and data analysis.

Complexity Grids represent one hybrid of "run-a-job" and information grids emphasizing the integration of experimental data with simulations and analysis tools. The typical computing environment in any organization forms another such hybrid and we can correspondingly expect a growing interest in Enterprise Grids with as a special case Campus Grids supporting university communities.

We have sketched two types of Grids – Information and "run-a-job" and some aspects of their integration. This is necessarily vague as we are only now building the core interfaces and infrastructure. However enough is understood today that one can start to build these Grids with some confidence that ones work will not require drastic changes as the field evolves.

References

- **Globus:** http://www.globus.org and *Implementing Production Grids* by Bill Johnston (Chapter 5 of http://www.grid2002.org)
- **Condor:** http://www.cs.wisc.edu/condor/ and *Condor and the Grid* by Douglas Thain, Todd Tannenbaum, and Miron Livny (Chapter 11 of http://www.grid2002.org)
- **Grid Service Interface to Information Grids:** http://www.ogsadai.org and *Databases and the Grid*: Paul Watson (Chapter 14 of http://www.grid2002.org)
- **OGSA and GT3**: http://www.globus.org/ogsa/ and *The Physiology of the Grid* Ian Foster, Carl Kesselman Jeffrey M. Nick and Steven Tuecke (Chapter 8 of http://www.grid2002.org)
- **OGSI Open Grid services Infrastructure:** http://www.gridforum.org/ogsi-wg/
- **e-Science and Data Deluge:** http://www.escience-grid.org.uk/ and *The Data Deluge: An e-Science Perspective* by Tony Hey and Anne Trefethen (chapter 36 of http://www.grid2002.org)
- **Virtual Observatories:** http://www.us-vo.org, http://www.astrogrid.org/ and *Grids and the Virtual Observatory* by Roy Williams (chapter 38 of http://www.grid2002.org)
- **Particle Physics Grids:** http://eu-datagrid.web.cern.ch/eu-datagrid/**,** http://www.griphyn.org and *Data Intensive Grids for High Energy Physics* by Julian Bunn and Harvey Newman (chapter 39 of http://www.grid2002.org)
- **Bioinformatics Grids:** http://www.mygrid.info/, http://www.discovery-on-the.net/ and *The New Biology and the Grid*: Phil Bourne Kim Baldridge (chapter 40 of http://www.grid2002.org)