Grid and Data Mining: More Related Than You Might Think

Ian Foster
Computation Institute, Argonne National Laboratory & University of Chicago
Department of Computer Science, University of Chicago
Mathematics and Computer Science Division, Argonne National Laboratory
foster@mcs.anl.gov

Abstract

I provide a brief introduction to grid computing and outline the current state of the art in grid technology and applications. I also outline directions that I see as important for future development, paying particular attention to issues relating to data. I make the case that both grid and data mining researchers have (or should have) much in common, as both are ultimately concerned with enabling distributed communities to function effectively as they tackle complex and often data-rich problems.

1. Introduction

What do grid and data mining have in common? At one level, nothing: grid is concerned with distributed system architecture and protocols; data mining with algorithms for extracting knowledge from data. But at another level, there are intimate connections: grid is concerned with enabling large-scale collaborative problem solving, in which data is frequently a major component; meanwhile, the data to which data mining is applied is often large, distributed, and contributed by many participants, and thus data mining, writ large, is, like grid, an end-to-end, systems problem. For these reasons, I believe that the two communities have in fact much in common, and much to gain from closer collaboration.

To explore some of these connections, I provide a brief introduction to grid computing and the state of the art in grid technology and applications. I also outline directions that I see as important for future development, paying particular attention to issues relating to data.

2. Historical Notes on Grid

Part of what makes computer science so interesting is the need periodic seismic shifts in focus that occur due to exponential changes in key system parameters. While such changes do not alter fundamental principles or physical laws, they create opportunities for new applications and/or expose new technological challenges. Thus, for example, those working in parallel computing observe that every order of magnitude increase in processor count raises new issues in system architecture, and as Kleinrock observed, "Gigabit networks are really different" [31], as are petabyte datasets [28].

The ideas, technologies, infrastructures, and applications to which the label "grid" is applied arose in response to one of those periodic re-evaluations of what is possible. In the 1990s, early high-speed networks led enthusiasts to examine how those networks could be used to integrate end systems to provide new functionality and enable new high-performance applications [10, 12, 29]. It then became clear that new applications could benefit from uniform, reliable, and performant mechanisms for authentication, authorization, resource discovery, data movement, and the like. Thus we saw efforts focused on developing such mechanisms and on using them to realize increasingly ambitious application scenarios.

The term "grid" was proposed and adopted for the resulting infrastructure, by loose analogy with the power grid [19]. However, it was clear from the beginning that the scope was more than simply "computing as a utility." Study of early applications led to a recognition that grid technologies, applications, and infrastructures were concerned more generally with "flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions, and resources—what we refer to as *virtual organizations*" [24].

3. State of the Art

Grid technologies have moved beyond the research and demonstration phase and are being applied on large scales, although substantial research and development also continues. These technologies are increasingly, also, being applied to problems involving the management and analysis of large quantities of data.

First-generation grids, pioneered by the high energy physics community, have been successful in federating computing resources across many sites to provide surge capacity to participants. For example, the US-based Open Science Grid (OSG) [38] and the European Union's

Enabling Grids for eScience (EGEE) infrastructure [1] run tens of thousands of jobs per day.

Second-generation grids emphasize the delivery of data and software as services, the federation of services to meet community needs, and the construction of infrastructures designed to host services. Examples include the Earth System Grid [8], which provides access to large quantities of climate simulation data (more than 100 TB downloaded in 2006); the cancer Biomedical Informatics Grid (caBIG) [37], which encompasses data and analysis services at dozens of cancer centers across the US; and the Biomedical Informatics Research Network (BIRN) [16], which links biomedical research centers. Virtual observatories [42] are another success story, linking digital sky surveys across the globe.

Numerous other examples of large-scale Grid deployments exist in the US, Europe, and Asia. The Open Grid Forum serves as a meeting place for participants in many of these projects.

Driven by these developments, grid technologies have evolved substantially since the early 1990s. Widely adopted architectural principles provide for the use of public key infrastructure (PKI) credentials (and more recently, SAML assertions) as a basis for authentication and authorization, standardized schema for resource description, Web Services protocols for resource discovery and access, and so forth. The adoption of "industry standard" Web Services technologies as a foundation for grid protocols has been mostly positive, although standards wars have slowed availability of standard tooling. Progress on standardization of higher-level protocols has been less rapid, but some useful progress has been made: for example, the GridFTP and Storage Resource Manager (SRM) specifications.

Widely used distributed data management software includes GridFTP [4] for data movement; Data Access and Integration Services (DAIS) for access to structured data [6]; SRM [41]; and the Storage Resource Broker [7]. These components are linked with standardized authentication and authorization infrastructures [23] that allow large communities to control who can access what resources and services. One example of a data-intensive system constructed with these components is the LIGO Data Grid, which streams more than one terabyte of data per day from the LIGO gravitational wave observatory to eight sites around the world [13].

To date, open source software has been a bigger force for adoption than standards. The open source Globus Toolkit version 4 [18], which includes several of the components listed above, has seen broad adoption worldwide, with contributors from the US, Europe, and Asia. In Europe, European Union pressure for "European solutions" has led to the development of several different systems, including ExtreemOS [30], gLite [2], NorduGrid [15], OMII [3], and Unicore [40].

4. Future Directions

I mention here several directions that I think are important for the future. Whether they can be viewed as aspects of grid, data mining, or something else is open to debate, but I do believe that they are important! Some of this material is taken from another recent article [20].

I referred above to the important role that **service-oriented architecture** is already playing in certain science communities. Much work is needed on such issues as description and discovery, provenance and trust, and provisioning and scheduling in order to scale these approaches to larger scales.

One important goal must be to enable a separation of concerns and responsibilities between those who operate the physical resources that host services, those who construct services, and those who access services [21]. The US TeraGrid infrastructure [11] is pursuing this agenda via its "science gateways" program. This separation is also a major concern in industry: commercial providers of utility computing services such as Amazon's EC2 provide computing services at relatively low costs. It will be interesting to see who will ultimately become the primary suppliers of computational and storage resources.

Provenance is an important issue to address in a substantial and principled manner. Progress in science depends on one researcher's ability to build on the results of another. "Service oriented science" can make it far easier, from a mechanical perspective, for researchers to do just this, by using service invocations to perform data access, comparison, and analysis tasks that might previously have required manual literature searches, data analyses, and/or physical experiments. However, the results of these activities are only useful when published if other researchers can determine how much credence to put in the results on which they build, and in turn convince their peers that their results are credible. Ultimately we will need to automate these processes.

These observations have motivated growing interest in methods for recording the provenance of computational results. Initial work focused on databases [9, 44], but interest has broadened to encompass arbitrary computations [25, 33]. A series of workshops [34] have led to the formulation of a provenance challenge [35], in which many groups have participated. Approaches explored include the use of functional scripting languages to express application tasks [45], file system instrumentation [36], and the use of a general-purpose provenance store [33].

Another area of considerable current interest in the grid community concerns the methods used to specify and execute **large computations involving many loosely coupled activities**. Such computations arise, for example, in large-scale data analyses and parameter studies.

Considerable progress has been made, to the extent that it is now common to see computations involving tens of thousands of tasks and operating on terabytes of data running efficiently on both large supercomputers and distributed grids comprising multiple clusters. Examples of technologies used for this purpose include Condor [32], Pegasus [14], and Swift [46]. Work is required to integrate such computations into community workflows.

Research occurs within communities, and the **formation and operation of communities** can be facilitated by appropriate technology. Much progress has been made in defining relevant protocols, practices, and systems—progress to which the grid community has contributed, via for example its work on authorization architectures [17, 43]. However, many challenges remain. For example, mechanisms that work effectively for two or ten participants may not scale effectively to one thousand or one million—not necessarily because implementations cannot handle the number of entities involved, but because softer issues such as trust, shared vocabulary, and other implicit knowledge break down as communities extend beyond personal connections.

One approach to solving some scaling problems is to build infrastructures that allow clients to associate arbitrary metadata ("assertions") with data and services. Assuming that we can also determine whether such assertions can be trusted (perhaps on the basis of digital signatures, and/or yet other assertions), consumers can then make their own decisions concerning such properties as quality, provenance, and accuracy. Systems such as Wikipedia and the Flickr and Connotaea collaborative tagging systems [26] demonstrate the advantages, costs, and pitfalls of different approaches to building such community knowledge bases.

Provisioning and scheduling become important as services and data increase in popularity [5]. Large central facilities, such as those operated by national centers and commercial providers, will surely continue to be important. However, it may also be feasible to exploit distributed resources, in cases where popularity allows us to justify the high cost of migrating data [27]. To this end, we are exploring methods that *diffuse* popular data "into the grid" as popularity increases [39].

5. Summary

Both grid and data mining researchers seek to enhance the abilities of individuals and communities to solve complex problems. In pursuing this goal, we need to take a system-level [22] view, in which we study and seek opportunities for optimization in every aspect of the problem solving process, not only by the individual researcher or within an individual laboratory, but also within and across communities.

For example, we may determine that (as I have argued here) service oriented architectures can be used to distribute and thus accelerate the processes of publishing, discovering, and accessing relevant data and software; that the encoding of provenance information can facilitate the reuse of data; and that software support for building communities can promote the collaborative development of knowledge. Many other opportunities to facilitate distributed problem solving (whether data-intensive or otherwise) can easily be identified.

In examining these issues, I have focused on the concerns of scientists and science. Scientists are certainly not alone in grappling with these issues. However, science is perhaps unique in the scope and scale of its problems and the subtlety of the questions that the methods discussed here can be used to answer. We may expect that methods developed for science can find application elsewhere, even as scientists look increasingly to computer science and information technology for tools that maximize the time that they spend thinking.

Looking forward, there is much to be gained from grid and data mining researchers making common cause, due to their common interest in these and other topics.

6. Acknowledgments

This work was supported in part by the Mathematical, Information, and Computational Sciences Division of the Office of Advanced Scientific Computing Research, U.S. Department of Energy (DE-AC02-06CH11357).

7. References

1. *Enabling Grids for eScience (EGEE)*, http://public.eu-egee.org, 2007.
2. *GLite Grid Middleware*, http://glite.web.cern.ch, 2005.
3. *U.K. Open Middleware Infrastructure Institute (OMII)*, www.omii.ac.uk, 2005.
4. Allcock, B., Bresnahan, J., Kettimuthu, R., Link, M., Dumitrescu, C., Raicu, I. and Foster, I. The Globus Striped GridFTP Framework and Server *SC'2005*, 2005.
5. Appleby, K., et al., Oceano - SLA Based Management of a Computing Utility. in *7th IFIP/IEEE International Symposium on Integrated Network Management*, (2001).
6. Atkinson, M., et al. Data Access, Integration, and Management. in *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, 2004.
7. Baru, C., Moore, R., Rajasekar, A. and Wan, M., The SDSC Storage Resource Broker. in *8th Annual IBM Centers for Advanced Studies Conference*, (Toronto, Canada, 1998).
8. Bernholdt, D., et al. The Earth System Grid: Supporting the Next Generation of Climate Modeling Research. *Proceedings of the IEEE*, *93* (3). 485-495. 2005.
9. Buneman, P., Khanna, S. and Tan, W.-C., Why and Where: A Characterization of Data Provenance. in *International Conference on Database Theory*, (2001).

10. Catlett, C. In Search of Gigabit Applications. *IEEE Communications Magazine* (April). 42-51. 1992.
11. Catlett, C. and others. TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications. in *High Performance Computing and Grids in Action*, 2007.
12. Catlett, C. and Smarr, L. Metacomputing. *Communications of the ACM*, *35* (6). 44-52. 1992.
13. Chervenak, A., Schuler, R., Kesselman, C., Koranda, S. and Moe, B., Wide Area Data Replication for Scientific Collaborations. in *6th IEEE/ACM Int'l Workshop on Grid Computing* (2005).
14. Deelman, E., et al. Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems. *Scientific Programming*, *13* (3). 219-237. 2005.
15. Eerola, P., et al. The NorduGrid production Grid infrastructure, status and plans *4th International Workshop on Grid Computing* 2003.
16. Ellisman, M. and Peltier, S. Medical Data Federation: The Biomedical Informatics Research Network. in *The Grid: Blueprint for a New Computing Infrastructure (2nd Edition)*, Morgan Kaufmann, 2004.
17. EU DataGrid *VOMS Architecture v1.1*, http://grid-auth.infn.it/docs/VOMS-v1_1.pdf, 2003.
18. Foster, I. Globus Toolkit Version 4: Software for Service-Oriented Systems. *Journal of Computational Science and Technology*, *21* (4). 523-530. 2006.
19. Foster, I. The Grid: Computing without Bounds. *Scientific American*, *288* (4). 78-85. 2003.
20. Foster, I. Man-Machine Symbiosis, 50 Years On. in Grandinetti, L. ed. *Advances in High Performance Computing*, 2007.
21. Foster, I. Service-Oriented Science. *Science*, *308*. 814-817. 2005.
22. Foster, I. and Kesselman, C. Scaling System-level Science: Scientific Exploration and IT Implications. *IEEE Computer* (November). 32-39. 2006.
23. Foster, I., Kesselman, C., Pearlman, L., Tuecke, S. and Welch, V., The Community Authorization Service: Status and Future. in *Computing in High Energy Physics (CHEP)*, (2003).
24. Foster, I., Kesselman, C. and Tuecke, S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. *International Journal of Supercomputer Applications*, *15* (3). 200-222. 2001.
25. Foster, I., Voeckler, J., Wilde, M. and Zhao, Y., The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration. in *Conference on Innovative Data Systems Research*, (2003).
26. Golder, S. and Huberman, B.A. The Structure of Collaborative Tagging Systems. *Journal of Information Science*, *32* (2). 198-208. 2006.
27. Gray, J. *The Economics of Distributed Computing*. Microsoft Research MSR-TR-2003-24, http://research.microsoft.com/research/pubs/ 2003.
28. Gray, J. and Szalay, A. Science In An Exponential World. *Nature*, *440* (23). 2006.
29. Grimshaw, A., Weissman, J., West, E. and Lyot, E. Metasystems: An Approach Combining Parallel Processing and Heterogeneous Distributed Computing Systems. *Journal of Parallel and Distributed Computing*, *21* (3). 257-270. 1994.
30. Johnson, I., Lakhani, A., Matthews, B., Yang, E. and Morin, C. XtreemOS: Towards a Grid Operating System with Virtual Organisation Support *UK eScience All Hands Meeting*, 2007.
31. Kleinrock, L. The Latency/Bandwidth Tradeoff in Gigabit Networks; Gigabit Networks are Really Different! *IEEE Communications Magazine*, *30* (4). 36-40. 1992.
32. Litzkow, M. and Livny, M. Experience with the Condor Distributed Batch System. in *IEEE Workshop on Experimental Distributed Systems*, 1990.
33. Miles, S., Groth, P., Branco, M. and Moreau, L. The requirements of using provenance in e-science experiments. *Journal of Grid Computing*, *5* (1). 1-25. 2005.
34. Moreau, L. and Foster, I. (eds.). *International Provenance and Annotation Workshop (IPAW'06)*. Springer LNCS, 2006.
35. Moreau, L., et al. The First Provenance Challenge. *Concurrency and Computation: Practice and Experience*. 2007.
36. Muniswamy-Reddy, K., Holland, D., Braun, U. and Seltzer, M., Provenance-Aware Storage Systems. in *2006 USENIX Annual Technical Conference*, (Boston, MA, 2006).
37. Oster, S., et al. caGrid 1.0: A Grid Enterprise Architecture for Cancer Research *AMIA Annual Symposium*, 2007.
38. Pordes, R., et al., The Open Science Grid. in *Scientific Discovery through Advanced Computing (SciDAC) Conference*, (2007).
39. Ranganathan, K., Iamnitchi, A. and Foster, I. Improving Data Availability through Dynamic Model-Driven Replication in Large Peer-to-Peer Communities *Global and Peer-to-Peer Computing on Large Scale Distributed Systems Workshop*, 2002.
40. Romberg, M., The UNICORE Architecture: Seamless Access to Distributed Resources. in *8th IEEE International Symposium on High Performance Distributed Computing*, (1999), IEEE Computer Society Press.
41. Shoshani, A., Sim, A. and Gu, J. Storage Resource Managers: Essential Components for the Grid. Nabrzyski, J., Schopf, J. and Weglarz, J. eds. *Resource Management for Grid Computing*, 2003.
42. Szalay, A. and Gray, J. Scientific Data Federation: The World Wide Telescope. in *The Grid: Blueprint for a New Computing Infrastructure (2nd Edition)*, Morgan Kaufmann, 2004.
43. Welch, V. *Globus Toolkit Version 4 Grid Security Infrastructure: A Standards Perspective*, http://www.globus.org/toolkit/docs/4.0/security/GT4-GSI-Overview.pdf, 2004.
44. Woodruff, A. and Stonebraker, M., Supporting Fine-Grained Data Lineage in a Database Visualization Environment. in *13th International Conference on Data Engineering*, (1997), 91-102.
45. Zhao, Y., Dobson, J., Foster, I., Moreau, L. and Wilde, M. A Notation and System for Expressing and Executing Cleanly Typed Workflows on Messy Scientific Data. *SIGMOD Record 34* (3). 37-43 2005.
46. Zhao, Y., et al., Swift: Fast, Reliable, Loosely Coupled Parallel Computation. in *IEEE International Workshop on Scientific Workflows*, (Salt Lake City, Utah, 2007).