Grid technologies empowering drug discovery

Andrew Chien, Ian Foster and Dean Goddette

Grid technologies enable flexible coupling and sharing of computers, instruments and storage. Grids can provide technical solutions to the volume of data and computational demands associated with drug discovery by delivering larger computing capability (flexible resource sharing), providing coordinated access to large data resources and enabling novel online exploration (coupling computing, data and instruments online). Here, we illustrate this potential by describing two applications: the use of desktop PC grid technologies for virtual screening, and distributed X-ray structure reconstruction and online visualization.

Andrew Chien
Entropia
10145 Pacific Heights
Suite 800, San Diego
CA 92121, USA;
Dept of Computer Science
and Engineering
University of California
San Diego
CA 92039, USA
Dean Goddette
Entropia
Ian Foster
Mathematics and Computer
Science Division
Argonne National Laboratory
Argonne, IL 60439, USA;
Dept of Computer Science
University of Chicago
Chicago, IL 60637, USA
tel: +1 630 252 4619
fax: +1 630 252 5986
e-mail: foster@mcs.anl.gov

▼ The rapid accumulation of digital information (e.g. from various genomes and the Protein Data Bank) and modeling knowledge (molecular and system models) for biological systems, is driving a growing use of computing in pharmaceutical research. Several decades of growth in information and computational capability were highlighted in 2001 by the sequencing of the human genome – a major scientific milestone. The tremendous human and financial stakes in the race to develop effective therapeutics and bring them to the market, motivates not only the use of computer technology to support traditional discovery techniques, but also the exploitation of novel *in silico* techniques for drug discovery. The volumes of data being generated and the amount of computing needed to process and extract meaning from those data is exploding. In fact, computational power is increasingly becoming the limiting step in drug design and discovery. This perspective is reinforced by industry pioneers such as Celera Genomics (http://www.celera.com) who used computational techniques and massively parallel supercomputers to accelerate the assembly of the human genome [1]. Celera continues to expand its computational capabilities, believing it to be a crucial capability in its drug discovery efforts.

One possible approach to meeting these needs is the 'Grid' [2,3]: a new class of infrastructure and tools that layers on today's Internet and Intranets to enable large-scale sharing of resources both within enterprises and across distributed, often loosely coordinated groups [4]. By providing scalable, secure, high-performance mechanisms for discovering and negotiating access to remote resources, Grid technologies promise to enable individual enterprises, commercial partnerships, and scientific collaborations to share resources on an unprecedented scale, and for geographically distributed groups to work together in ways that were previously impossible [5,6]. Many dozens of large collaborative 'e-science' projects worldwide are now applying these technologies in different disciplines [7–10].

Here, we briefly examine the computational challenges associated with pharmaceutical applications. We go on to discuss the role that Grid technologies can play in meeting these challenges, and, finally, present two examples of the application of these technologies to drug discovery.

Computational challenges

The amount of genomic, proteomic and molecular data available to biotechnologists is growing exponentially, doubling in as little as six months to a year. This rate of growth greatly exceeds the 18-month rate of individual-processor power doubling, as predicted by Moore's Law. As a result, researchers are increasingly motivated to use multiprocessor (parallel) systems.

This volume of data is only one of several sources driving the rapid growth in computational demands, as illustrated by the following:

- Sequence data in the NCBI Genbank database are doubling in size every 12 months. Many research activities involve sequence searching against this database, the cost of which scales with database size.

1359-6446/02/$ – see front matter PII: S1359-6446(02)02369-3

- Combinatorial chemistry has become a routine activity and is often used to create molecules en masse. However, computers are used to design and select which compounds are worth synthesizing, via, for example, virtual screening techniques that enable focused HTS by using computational analysis to select a subset of compounds appropriate for a given receptor. Here, the computational demand scales with the number of candidate molecules (e.g. tens-of-thousands to -millions or more).
- Computational demands are increasing across the board at pharmaceutical companies owing to the increasing volume both of incoming data, and data generated from technologies such as gene chips, proteomics, ADME, toxicology, virtual organs, data analysis of clinical data, and pharmacogenomics.
- Many accurate methods, such as quantum and semi-empirical methods, and free-energy calculations, are not applied on a large scale because of computational resource limitations.

Grid technologies

Pharmaceutical researchers thus need both to access more data than in the past, and to perform more computation on those data. These requirements can certainly be addressed by building ever-larger central computing facilities. However, dramatic improvements in network performance and software technologies also make it feasible to consider potentially more cost-effective alternative approaches based on the coordinated use of distributed resources, whether within an enterprise or extending outside the enterprise to business partners and service providers.

So-called 'Grid technologies' have been developed over the past several years with the goal of enabling various forms of resource sharing. One usage modality that has already been applied with considerable success is high-throughput (or distributed) computing [11–13], which involves the sharing of idle desktop workstation- or PC-cycles to solve large problems. An analysis of 500 PCs in a pharmaceutical company found that 95% were idle during the night and 85% were idle during the day (Fig. 1). Other observers have reported similarly low CPU usage numbers [12,14,15]. Thus, in any large corporation there is a supercomputer-class system available as a 'PC Grid' within the company, behind the enterprise firewall.

In other settings, sharing resources of various types in a way that is flexible, reliable and secure, with collaborators and business partners can have significant benefits. For example, a specialized device (e.g. an X-ray-source beamline, as discussed previously) can be coupled with remote databases and computers to enable 'computer-in-the-loop' data collection and analyses [16]. Secure remote access to databases can enable cross-correlations among databases maintained by different groups either within a single company, or, potentially, across different enterprises. A scientific collaboration can pool computers and storage across participating institutions to obtain instantaneous access to many more resources than available locally. A pharmaceutical company might purchase access to computers and storage from service providers to meet some short-term demand, thus obviating the need to purchase and maintain these resources locally.

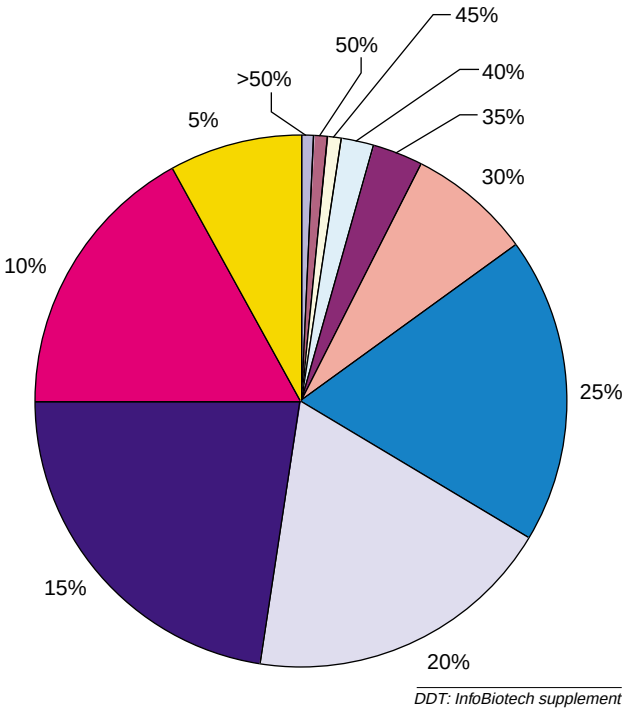

Figure 1. Central processing unit (CPU) usage of desktop PCs during the workday. The piechart represents a total of ≈500 desktop PCs in a large pharmaceutical company. Each segment represents the number of PCs with the labeled level of CPU usage (average usage over the 8-h day). Average CPU usage over all machines during the 8-h day was just over 15%, falling to less than 5% after normal working hours.

Scenarios such as these are enabled, in part, by the broad deployment and rapidly decreasing cost of broadband networks. However, although high-speed networks are often necessary, they are far from sufficient. Remote resources are typically owned by others, exist within different administrative domains, run different software, and are subject to different security and access control policies. Even within a single institution, access control issues might have to be addressed. These are issues that characterize a Grid and, historically, have made the realization of such scenarios difficult. To access remote resources, an application or user must first discover that they exist, negotiate (and perhaps pay for) access, then configure their computation to use them effectively; all steps must be implemented without compromising their own security or the security of the resources on which they are computing.

A variety of technologies have been investigated over the years for these purposes. The Distributed Computing Environment (DCE) was an early attempt to provide a uniform distributed computing environment. In practice, however, deployment has been difficult. CORBA [17] is used extensively within bioinformatics for distributed computing, but does not address issues relating to either the harnessing of idle computers or inter-institutional resource sharing. Web technologies have also been investigated, but again do not address trust and performance issues that arise in practical situations. Peer-to-peer technologies have achieved significant successes within community Grids where security is not of paramount concern.

Fortunately, the past several years have seen significant technical advances that address many of these concerns. Within the research and education community, the open source Globus Toolkit™ (http://www.globus.org) has achieved wide use as a *de facto* technical solution to security, resource discovery, resource management, and other issues that arise when sharing and accessing resources across administrative boundaries. Industrial interest is also emerging within the computer industry, with 12 companies announcing support for Globus protocols and software. The complementary Condor system (http://www.cs.wisc.edu/condor) [12] is also used extensively for harnessing idle computers within a single department or institution. Other relevant research systems include the object-based Legion system (http://www.legion.virginia.edu) [18], the NetSolve (http://www.icl.cs.utk.edu/netsolve) [19] and Ninf (http://www.ninf.apgrid.org) [20] systems for remote access to numerical software, and the Nimrod system (http://www.csse.monash.edu.au) [21] for parameter studies. Data Grid technologies build on these and other technologies to address issues relating specifically to distributed management of data and computations on data.

Companies such as Entropia (http://www.entropia.com), United Devices (http://www.ud.com) and DataSynapse (http://www.datasynapse.com) provide software for harvesting idle cycles within enterprise PC Grids. One important differentiator among these different offerings is the sophistication of the 'sandboxing' technology used to protect PCs from misbehavior by the application programs. Ideally, these sandboxing technologies should control all access to the underlying system by the application, ensuring that the application cannot access the desktop user's files, applications or other resources. Some systems achieve this by supporting only Java programs, whereas others support arbitrary executables. Other relevant enterprise scheduling technologies include cluster schedulers such as the Grid Engine from Sun (http://www.sun.com/gridware), the Load Sharing Facility from Platform (http://www.platform.com), and the Portable Batch System from Veridian (http://www.pbspro.com).

A recent development that promises to unite these different technical approaches is the Open Grid Services Architecture (OGSA) recently proposed by the Global Grid Forum (http://www.gridforum.org) – a Grid community and standards organization. Most of the projects and companies listed above have made commitments to adopting OGSA standards, which integrate Globus Toolkit™ technologies with Web services mechanisms [22].

Applications to drug discovery

Two examples illustrate contemporary applications of Grid technologies to pharmaceutical problems.

Accelerating virtual screening

Computational analysis of large sets of drug-like compounds by virtual screening [23,24] enables the rational selection of those likely to be active against a chosen biological receptor. Although modern techniques in combinatorial chemistry and HTS enable large numbers of molecules to be synthesized and assayed, there are still significant limitations to the techniques owing to the time and resources required to complete the experiments in the laboratory. Virtual screening enables the identification of a focused set of compounds – selected on the basis of their properties and those of the target receptor – which leads to an enhanced hit-rate in the bioassay. Unfortunately, the size of the computational problem is large, not only because of the sheer numbers of compounds that are involved, but also because of the time required to perform a reasonable virtual screening experiment, and to explore the conformational space of the test compounds.

One approach to this problem is to use specialized high performance computing (HPC) systems [25]. However, high-end HPC systems are expensive, owing to the considerable investment required for specialized packaging and interconnection technologies. These systems are therefore commonly reserved for problems that require high interprocessor performance. At the other end of the HPC scale are the increasingly popular Linux and Beowulf clusters [26], which are built from commodity components and can provide significant amounts of computational power at a lower cost per node than mid-range compute servers. However, the cost of building, maintaining and operating these systems can be significant, owing to the range of open-source software and systems involved. The use of both high-end and Beowulf systems also requires expertise in algorithm parallelization, which is often difficult to find.

High-throughput (distributed) Grid computing can bring significantly more computational power to problems that do not require high interprocessor performance (a typical Beowulf cluster has 64 nodes, while an enterprise could have hundreds or even thousands of PCs). This increased

computational power enables not only more virtual screening in less time, but also the implementation of more complex and accurate methods to reduce the number of false-positives. This strategy fits well with a current trend in virtual screening: separation of the docking step (identifying the most probable binding mode of this compound to a specific receptor) from the scoring step (determining how tightly this docked molecule binds relative to other molecules), which can involve more complex algorithms than those needed for docking.

Figure 2 shows the docking throughput of a distributed computing system installed in a pharmaceutical company, as a function of the number of PCs used. Performance increases linearly with the number of machines, and compares favorably in performance with both high-end multiprocessors [an SGI system (http://www.sgi.com)] and Beowulf clusters.

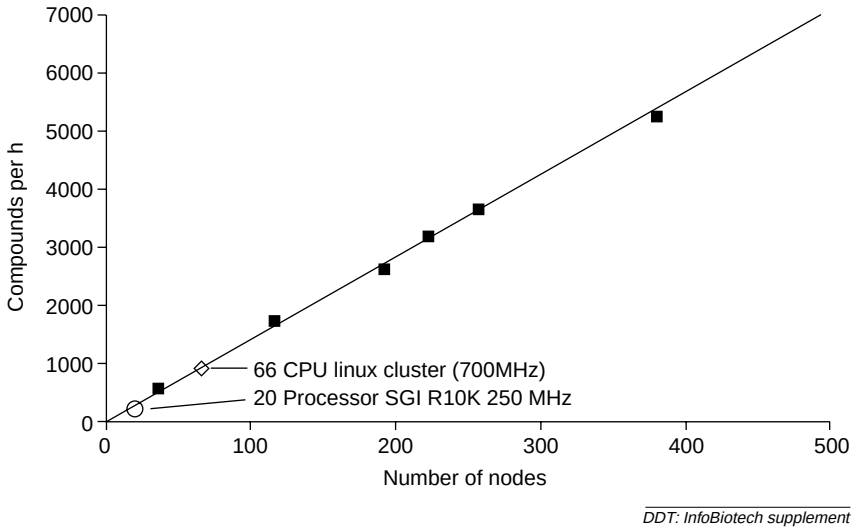

Figure 2. Docking throughput using DOCK [29] (expressed as compounds per h) increases linearly with the number of nodes. A standard job of 50,000 pharmaceutical compounds was run on a Grid using either a 20 central processing unit (CPU) Origin 2000 or a 66 node Linux cluster. The total number of nodes was varied from 1 to 500, and the result shows that desktop Grid performance increases linearly with the number of nodes. This means that virtually unlimited computing power can be gathered from desktop Grids for applications such as docking.

X-ray source data analysis

The second example is the enhancement of advanced scientific instrumentation by linking with remote computers. The facility in question is the Advanced Photon Source (APS), a high-brilliance X-ray source located at Argonne National Laboratory (http://www.anl.gov/) that is used for X-ray crystallography, micro-tomography, and other applications. The application couples an APS beamline with remote computers to enable the transfer, filtering and reconstruction of beamline data at a rate similar to that at which they are acquired.

In most current synchrotron-based X-ray crystallography systems, a user performs the processing and reconstruction calculations after the experimental period is finished. The inability to examine reconstructions directly after data has been acquired makes it difficult to make crucial decisions during the experimental period, resulting in a poor use of beam time. A common consequence is that most experimenters accumulate large numbers of unprocessed datasets. The online reconstruction system therefore presents a significant advantage over existing systems by providing users with rendered reconstruction results for analysis immediately after the data are acquired.

The system described here acquires computing resources dynamically, using Globus Toolkit™ technologies, and then

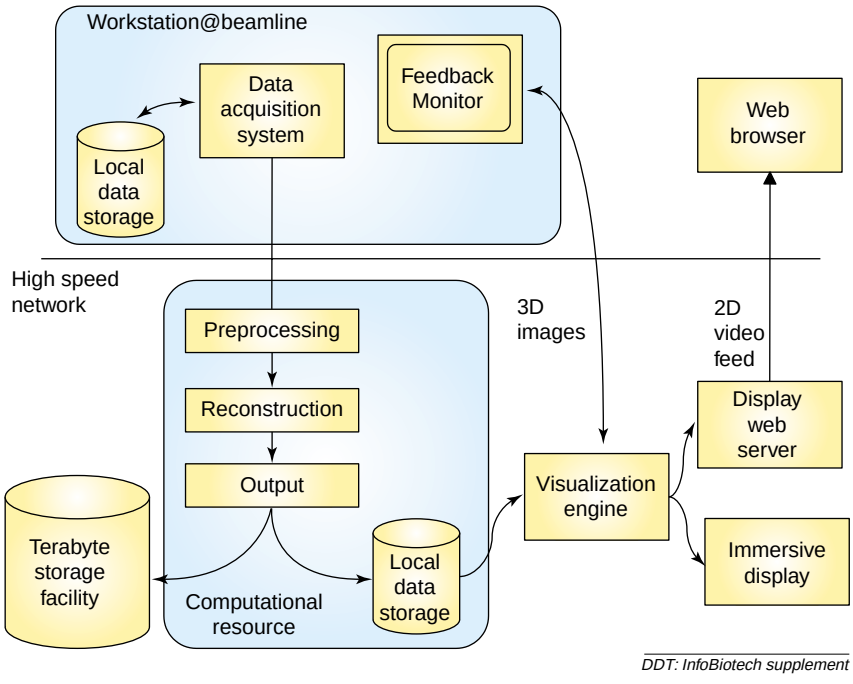

Figure 3. Schematic of an online data reconstruction system used at Argonne National Laboratory's Advanced Photon Source (http://www.anl.gov).

starts processes that comprise the preprocessing and reconstruction pipeline on those resources [27]. This pipeline then transfers the data to the computing resources [either a parallel computer (Fig. 3) or workstations]; approximately 80 processors are typically used. The data are then preprocessed and reconstructed, generating output data that is written to archival storage as well as passed to a rendering engine that generates quasi-real-time output for visualization by scientists. Examination of these data might then lead to changes in the experimental setup. The same system has also been used for reconstruction of data from electron microscopes [28].

We believe that the same motivations, computational structures, and technologies can find application in many other areas of biotechnology, owing to the increased ease with which data can be generated, and the increased difficulty with which data is reconstructed. Coupling the two procedures can increase both the efficiency of data collection, and the quality and accessibility of the processed data.

Concluding remarks

Although Grid technologies are only beginning to be used by leading drug discovery companies, their promise is clear, and the momentum behind them is building. Because desktop PC Grids can provide tremendous computing power at low cost, they are already being used in many enterprises. With increasing commercial use of Globus Toolkit™ technologies and the Open Grid Services Architecture, the novel capabilities that server Grids can provide are also gaining commercial acceptance. We expect that, within a few years, Grid technologies will dramatically increase productivity and resource efficiency in a wide range of drug discovery companies.

References

1 Golden, F. and Lemonick, M.D. (2000) The race is over. *Time Magazine* July

2 Foster, I. (2002) The grid: a new infrastructure for 21st century science. *Physics Today 55*, 42–47

3 Foster, I. and Kesselman, C., eds (1999) *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann

4 Foster, I. *et al.* (2001) The anatomy of the grid: enabling scalable virtual organizations. *International Journal of High Performance Computing Applications 15*, 200–222 (http://globus.org/research/papers/anatomy.pdf)

5 National Research Council (1993) *National Collaboratories: Applying Information Technology for Scientific Research*, National Academy Press

6 Teasley, S. and Wolinsky, S. (2001) Scientific collaborations at a distance. *Science 292*, 2254–2255

7 EU DataGrid Project (2001) The DataGrid Architecture. DataGrid-12-D12.4-333671-3-0 (http://www.eu-datagrid.org)

8 Johnston, W.E. *et al.* (1999) Grids as production computing environments: the engineering aspects of NASA's information power grid. In *Proc. 8th IEEE Symposium on High Performance Distributed Computing*, IEEE Press

9 Stevens, R. *et al.* (1997) From the I-WAY to the National Technology Grid. *Commun. ACM 40*, 50–61

10 Szalay, A. and Gray, J. (2001) The world-wide telescope. *Science 293*, 2037–2040

11 Anderson, D.P. and Kubiatowicz, J. (2002) The worldwide computer. *Sci. Am.* 3, 40–47

12 Livny, M. (1999) High-throughput resource management. In *The Grid: Blueprint for a New Computing Infrastructure*, (Foster, I. and Kesselman, C., eds), pp. 311–337, Morgan Kaufmann

13 Sullivan, W. *et al.* (1997) A new major SETI project based on project SERENDIP data and 100,000 personal computers. In *Astronomical and Biochemical Origins and the Search for the Life in the Universe*, Editrice Compositori

14 Mutka, M. and Livny, M. (1991) The available capacity of a privately owned workstation environment. *Performance Evaluation 12*, 269–284

15 Ryu, K.D. and Hollingsworth, J. (2000) Exploiting fine grained idle periods in networks of workstations. *IEEE Transactions on Parallel and Distributed Systems 11*, 683–698

16 Johnston, W. (1999) Realtime widely distributed instrumentation systems. In *The Grid: Blueprint for a New Computing Infrastructure*, (Foster, I. and Kesselman, C., eds) pp. 75–103, Morgan Kaufmann

17 Object Management Group (1998) *Common Object Request Broker: Architecture and Specification* (Revision 2.2), Document 96.03.04 (http://www.omg.org)

18 Grimshaw, A.S. and Wulf, W.A. (1997) The legion vision of a worldwide virtual computer. *Commun. ACM* 40, 39–45

19 Casanova, H. and Dongarra, J. (1997) NetSolve: a network server for solving computational science problems. *Int. J. Supercomputer Appl. High Performance Comput. 11*, 212–223

20 Nakada, H. *et al.* (1999) Design and implementations of Ninf: towards a global computing infrastructure. *Future Generation Computing Systems 15*, 649–658

21 Abramson, D. *et al.* (1995) Nimrod: a tool for performing parameterised simulations using distributed workstations. In *Proc. 4th IEEE Symp. on High Performance Distributed Computing*, IEEE Press

22 Graham, S. *et al.* (2001) *Building Web Services with Java: Making Sense of XML, SOAP, WSDL, and UDDI*, Sams

23 Bajorath, J. (2002) Virtual screening in drug discovery: methods, expectations and reality. *Curr. Drug Discov.*

24 Walters, W.P. *et al.* (1998) Virtual screening: an overview. *Drug Discov. Today 3*, 160–178

25 Kaufmann, W.J. and Smarr, L.L. (1993) *Supercomputing and the Transformation of Science*, W. H. Freeman

26 Sterling, T. *et al.* (1999) *How to Build a Beowulf*, MIT Press

27 Wang, Y. *et al.* (2001) A High-throughput X-ray microtomography system at the advanced photon source. *Rev. Sci. Instruments 72*, 2062–2068

28 Smallen, S. *et al.* (2000) Combining workstations and supercomputers to support grid applications: the parallel tomography experience. In *Heterogeneous Computing Workshop* pp. 241–252, IEEE Press

29 Meng, E.C. *et al.* (1992) Automated docking with grid-based energy evaluation. *J. Comp. Chem. 13*, 505–524