The Beta Grid:
A National Infrastructure for Computer Systems Research

Ian Foster
Argonne National Laboratory
and
The University of Chicago

1 Introduction

In future information infrastructures, boundaries between computing, storage, and communication will blur as these three functions become increasingly intertwined. Networks will be more than dumb "bitways" that move bits among computers, storage, and people; they will incorporate substantial embedded computing and storage. This computing and storage will, when combined with appropriate middleware services (security, resource management, instrumentation, accounting, and billing, etc.) enable dramatically enhanced functionality when compared with the best-effort delivery provided by today's Internet.

Film distributors can use such a computationally enhanced network (or *grid* [4] as it is sometimes called) for the efficient and secure distribution of digital video, using embedded resources for the caching, compression, and encryption of video streams [1]. The climate change community can use a grid to deliver climate data products ("daily mean temperature," "frost frequencies in Wisconsin," "impacts on cranberry bog yields") to scientists and policymakers; in this case the grid might not only cache datasets but also run the computations required to tailor simulation data for specific purposes. Common to these two different examples is a distributed infrastructure capable of sophisticated computational functions.

The design and application of such grids raise numerous challenging research questions. Unfortunately, no infrastructure exists to support the computer systems research that would answer those questions. For example, a researcher interested in resource management techniques for the applications above cannot easily assemble the distributed collection of computers, archives, and networks required for realistic experimental evaluation of new mechanisms. Network testbeds such as CAIRN allow experimentation with network protocols but do not incorporate significant computing or storage resources, while existing grid testbeds such as GUSTO [3] connect large amounts of compute and storage resources but do not support the dedicated, on-demand access required for experimentation.

Motivated by these concerns, we propose the creation of a *national computing systems research grid* dedicated specifically to the experimental investigation of future grid architectures. This infrastructure, which we term the Broadband Experimental Terascale Access (Beta) grid, will comprise some moderate number (20–100) of reasonably powerful compute/storage nodes distributed across the country and connected to each other and to the user community via high-speed networks. We argue that this scale is necessary—and sufficient—to allow for realistic applications-driven experimentation.

While future grids will necessarily be heterogeneous in terms of architecture and operating system, a reasonable architecture for an individual Beta Grid node is a moderate-sized PC cluster. Given this basic node configuration, Linux becomes attractive as a base operating system. Apart from cost issues, the flexibility provided by access to source code facilitates certain types of experiment, while the growing high-performance Linux community suggests that it should be possible to configure Beta Grid clusters with the software required to support high-speed, reliable delivery of data and computing to applications.

2 Grids

A distributed grid infrastructure comprises a set of high-speed computers, storage systems, and networks, plus a set of grid services (or "middleware") to coordinate the ensemble of resources.

In today's grid prototypes and testbeds [5, 3], computers and networks are typically located at network endpoints and remain under tight local control; the coordinated use of multiple resources has been demonstrated, but is not routine. In future grids, we can expect to see not only greater access to endpoint resources but also the direct embedding of substantial computing and storage resources within the network itself, under the control of network service providers. (Tomorrow's terabit routers will perhaps also contain teraop computers and terabyte caches.) These developments will enable a variety of innovative services with inherently distributed implementations. For example:

- Content distribution services, which use a combination of embedded storage (e.g., for caching or mirroring) and computation (e.g., for compression or distillation) to deliver collections of files to end user communities [1].
- Value-added services which enhance content via indexing, annotation, distillation, translation, an so forth.
- Distributed "active data repositories" [2], which generate new data from old by running owner- or user-supplied code.
- Markets for computational services, which enhance the capabilities of the ordinary desktop by providing on-demand access to substantial computing and/or storage resources (for a fee).
- Virtual communities, which use grid resources to create, store, and manage sophisticated virtual worlds.

As in today's networks, we can expect to see a variety of parallel (or layered) grid infrastructures offering services at a range of different cost, functionality, security, and performance points, from best-effort "public grids" to "virtual private grids" that offer enhanced services at increased cost.

These concepts appear both promising and exciting. However, their practical realization and application remain extremely challenging tasks, because of the novelty of the concepts and the range of technologies involved. The following are just seven examples of areas in which research advances appear to be required:

- Security technologies able to protect data and code in a shared infrastructure that performs computation and stores data.
- Distributed management and monitoring technologies able to maintain system integrity, protect against misuse, and detect problems in a highly distributed environment.
- Algorithms, languages, compilers, and libraries for parallel computing across multiple systems.
- Adaptive computation techniques able to provide reliable service in an inherently unreliable and stable environment.
- Scalable resource management strategies for arbitrating between competing demands on network, storage, and computing resources.
- Scalable data distribution mechanisms, ranging from distributed file systems and digital libraries to "content channels" [1].
- Advanced network protocols: for example, multicast protocols that use distillation to deal with clients with different capabilities.

3 A Research Infrastructure

The complexity and novelty of the grid computing concepts outlined in the preceding section mean that large-scale experimentation will be required before we can hope to understand the true nature of the problem. This experimentation in turn requires access to prototype grid systems with sufficient functionality and complexity to allow for realistic experimentation.

Unfortunately, no such prototype systems exist. Current testbeds either are dedicated to specific research purposes (e.g., network protocols, active networks [7], agent technologies, content distribution [1]) or are not easily accessible to grid researchers (e.g., the NCSA Alliance's National Technology Grid [5], the Globus Ubiquitous Supercomputing Testbed Organization [3]).

We believe that progress in grid concepts and technologies requires the creation of a dedicated infrastructure designed to support research by the computer systems community on problems such as those listed above. This Broadband Experimental Terascale Access (Beta) grid will allow technologies that appear promising in the laboratory to be deployed and evaluated at a realistic scale.

3.1 Beta Grid Physical Architecture

The need for realistic experiments places demands on the Beta Grid's physical architecture in terms of size, geographical distribution, and network connectivity. Specifically, we argue that a minimum useful size is 20 sites (or *nodes* as we will call them here), with national distribution and high-speed (OC12–OC48: 0.6–2.4 Gb/s) network connectivity;

a total of at least 1 teraop/s of compute capability; and tens of terabytes of attached storage. One hundred or more nodes would be desirable to allow for more effective exploration of scalability issues. These numbers are justified as follows.

The rapid increase in network speeds and the deployment of advanced networks such as Abilene means that to be interesting, the Beta Grid needs to be designed from the outset to support OC48 (2.4 Gb/s) networks. This requirement has significant implications for node architecture.

In principle, we want Beta Grid nodes to be embedded in the network. In practice, this means that they should be located at sites where high-speed access to external networks is possible. Good geographical distribution is important so as to best approximate the structure of future national-scale grids, reduce load on backbone networks, and ensure reasonable proximity to users.

Individual nodes must be sufficiently powerful to permit interesting experiments. With an OC48 network, it would be desirable for a node to be able to source or sink data at 300 MB/s either from memory or from disk, be able to perform at least minimal processing on that data (e.g., 50 instructions/byte=15 BIPS), be able to support hundreds of concurrent streams, and support a local cache that can store say two hours of data (2 TB).

We believe that a credible Beta Grid system needs to comprise at least 20 nodes. This scale permits good national coverage and provides aggregate compute power that is competitive with moderate-scale supercomputers (300 BIPs) and storage capacity that can hold an interesting fraction of a petabyte archive (40 TB). Scaling the system to 100 nodes makes for a much more interesting system: 1.5 TIPs and 200 TB.

3.2 Beta Grid Software Architecture

The usage model that we envision for the Beta Grid also places demands on the software and services that need to be provided.

An infrastructure of the size proposed for the Beta Grid must necessarily be shared by a large user community in order to justify the investment. Hence, the Beta Grid architecture must incorporate security, resource management, accounting, billing, and other infrastructure to support shared use by at least hundreds of users.

An infrastructure distributed across 20–100 sites and subject to modification by a community of hundreds will necessarily be inconsistent, buggy, and subject to unexpected failure. Automated software update, configuration, fault detection, and intrusion detection mechanisms will be required. These goals in turn motivate a requirement for ready access to information about the structure and state of Beta Grid resources. Extensive instrumentation and a data collection infrastructure are required to allow for measurement of system behavior.

We believe that realistic evaluation of grid technologies requires that real application developers use the infrastructure to solve real problems. The scale just proposed should be sufficient to generate interest among application scientists; however, the Beta Grid architecture must also be designed to enable the co-existence of (experimental) applications and experimental services.

A grid infrastructure designed to enable experimentation with a wide variety of technologies—from distributed computing models to file systems and network protocols—must provide a core set of services that allow applications to manipulate various Beta Grid nodes in a uniform fashion, while also allowing for selective enhancement of functionality.

4 Building the Beta Grid

We believe that the Beta Grid outlined above can be created quickly, if adequate resources are available, via the coordinated development of a set of *Beta Grid service specifications*, identifying the basic services that are required to be in place at every Beta Grid node, and a *Beta Grid reference implementation* of those services. The service specifications ensure that the Beta Grid can evolve in an open fashion, while the reference implementation allows rapid deployment.

A variety of Beta Grid hardware and software architectures can be and should be supported: heterogeneity is inevitable and should be embraced, not resisted; standardization needs to occur at the level of APIs, not implementations. However, performance and price concerns suggest a reference implementation based on PC clusters loaded with Linux and some other software technologies listed below. We sketch here what form such a reference implementation might take, noting areas in which development is required.

A 32-processor rack-mounted cluster loaded with as much disk as can be chassis-mounted (currently around 2 TB) seems a reasonable hardware configuration. In order to enable this cluster to source or sink OC48 flows, Gb/s internal networking is required. This basic configuration translates to 640–3200 processors, 40–200 TB of disk, and (assuming 256 MB memory per processor) 160–800 GB of memory for an 20–100 node Beta Grid system.

On the software side, the core of the reference implementation would be the software distribution being developed by the high-performance Linux community, comprising Linux with modifications (e.g., to networking code) required for high-performance execution. This basic system would need to be extended with a local (node) resource management system allowing space sharing or time sharing of processors within a Beta Grid node (e.g., PBS or the Maui Scheduler, integrated perhaps with the Distributed Soft Realtime Scheduler) and high-speed (parallel) I/O capabilities (e.g., DPSS [6] and/or a parallel file system, with MPI-IO support).

We would use services provided by the Globus toolkit [3] to provide the basic mechanisms required for coordinated use of resources within multiple nodes. For example:

- The Grid Security Infrastructure and associated Globus services can be used for authentication and authorization, providing secure single sign-on/run anywhere capabilities via public key technologies.
- The Globus Architecture for Reservation and Allocation provides a basis for reservation, allocation, monitoring, and control of compute, network, and storage resources.
- The Global Access to Secondary Storage provides basic support for uniform data access; integration with the Storage Resource Broker (SRB) may provide a basis for distributed storage management.
- Finally, the Metacomputing Directory Service provides a mechanism for publishing and discovery properties of Beta Grid resources and applications.

The services just listed provide a reasonable basis on which to start Beta Grid development. However, additional work is required before we can produce a truly useful infrastructure. New software is required for user management, that is, the registration of new users and the establishment of required trust relationships at different nodes. Monitoring, instrumentation, logging, and intrusion detection represent another general area in which work is required. More work is needed on system configuration management, with the goal (ideally) of allowing for the automated propagation of system changes to Grid nodes. Accounting and billing services will also be required.

5 Related Projects

We review other infrastructure projects and proposals with similarities to what we propose here.

The Cooperative Advanced Interagency Research Network (www.cairn.net) provides a programmable routing infrastructure to support experimentation with network protocols. However, it does not incorporate significant computational or storage capabilities at its endpoints. Nor do the Active Node Transfer System (ANTS) [7] and the D'Agents (agent.cs.dartmouth.edu/network/ testbed, which support active network and agent research, respectively.

The Globus Ubiquitous Supercomputing Testbed Organization (GUSTO: www.globus.org) links thousands of processors at some 80 sites around the world, with the Globus toolkit providing uniform security, scheduling, data access, and other services. GUSTO is the largest existing grid infrastructure and has been used to support a range of interesting research projects. However, because GUSTO resources are not dedicated to the computer systems research community (indeed, they are frequently already overcommitted for other purposes), the use of resources at multiple sites can require considerable coordination.

The NCSA Alliance's National Technology Grid [5] and NASA's Information Power Grid projects are constructing substantial distributed computing infrastructures to support computational science and engineering applications. However, as with GUSTO, these infrastructures are not primarily designed for the computer systems research community.

A Computing and Communications Environment for Software Systems (ACCESS: www.cs.washington.edu/homes/tom/access) is a proposed experimental environment that would involve 50–100 small (10–20 node) clusters to support systems research in such areas as networking protocols, Internet measurement, and Web caching protocols. Administration is simplified by a uniform architecture with dedicated control PCs and remote administration tools. ACCESS and Beta Grid have similarities but emphasize different classes of problem, with ACCESS focusing on low-level systems issues and the Beta Grid seeking to enable realistic application-driven evaluation of grid technologies. To date, ACCESS remains just a proposal.

Finally, the very interesting Internet 2 Distributed Storage Infrastructure project (DSI: dsi.internet2.edu is concerned with providing efficient access to Web, video, and other Internet-based educational content [1]. An infrastructure comprising multiple storage clusters (IBM RS6000 web servers with 72 GB of disk and 900 GB of tape) has been deployed and is being used to support experiments with mirroring, caching, and other strategies for the transparent delivery of content. Like the Beta Grid, DSI is concerned with exploring applications for embedded network resources; unlike the Beta Grid, its focus is on a single application, namely content distribution.

6 Next Steps

The creation of a usable Beta Grid requires that three tasks be pursued concurrently. All three need to be pursued cooperatively by the community if we are to succeed.

The first task is the creation of the required software base. The considerable interest that we have encountered in the Beta Grid concept suggests that by leveraging the work already going on in the Linux and Grid communities, it should be possible to deploy a basic reference implementation quickly. The more advanced services that do not yet exist can then be incorporated as they are developed under other funding. A key is the effective coordination of the many relevant efforts that are already under way; the NSF PACIs, DOE NGI program, NASA IPG, and Grid Forum (`www.gridforum.org`) are important in this regard.

The second task is the creation of a Beta Grid prototype that couples existing clusters to which access is provided on a volunteer basis, in order to support initial experimentation and demonstrate feasibility. Experiences such as I-WAY and GUSTO show that such community-based testbeds can be effective as long as the scope of the planned experiments is kept appropriately small.

The third task is the funding, creation, and deployment of the Beta Grid proper. We believe that the substantial infrastructure and long-term access required for research progress will require dedicated platforms and hence investment by funding agencies and interested commercial organizations. This investment ($4–10M) is substantial, although it is not large when we consider the broad spectrum of important research projects that will be enabled.

Acknowledgments

I gratefully acknowledge useful discussions with Peter Beckman, Andrew Chien, Remy Evard, Carl Kesselman, Miron Livny, Rick Stevens, Brian Toonen, and Steven Tuecke. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38; by the Defense Advanced Research Projects Agency under contract N66001-96-C-8523; by the National Science Foundation; and by the NASA Information Power Grid program.

References

[1] M. Beck and T. Moore. The Internet2 distributed storage infrastructure project: An architecture for internet content channels. *Computer Networking and ISDN Systems*, 30(22-23):2141–2148, 1998.

[2] Renato Ferreira, Tahsin Kurc, Michael Beynon, Chialin Chang, Alan Sussman, and Joel Saltz. Object-relational queries into multidimensional databases with the active data repository. *International Journal of Supercomputer Applications*, 1999.

[3] I. Foster and C. Kesselman. Globus: A toolkit-based grid architecture. In *[4]*, pages 259–278.

[4] I. Foster and C. Kesselman, editors. *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan Kaufmann Publishers, 1999.

[5] R. Stevens, P. Woodward, T. DeFanti, and C. Catlett. From the I-WAY to the National Technology Grid. *Communications of the ACM*, 40(11):50–61, 1997.

[6] B. Tierney, W. Johnston, L. Chen, H. Herzog, G. Hoo, G. Jin, and J. Lee. Distributed parallel data storage systems: A scalable approach to high speed image servers. In *Proc. ACM Multimedia 94*. ACM Press, 1994.

[7] David J. Wetherall, John Guttag, and David L. Tennenhouse. ANTS: A toolkit for building and dynamically deploying network protocols. In *IEEE OPENARCH'98*, 1998.