The TeraGrid: A Primer

September 2002

This document provides a general overview of the TeraGrid project, oriented primarily toward describing the TeraGrid "Grid" environment, the organization of the project, and guidelines into how additional resources will be added to TeraGrid upon completion of the initial design and deployment in 2003.

Table of Contents

1. TeraGrid: Toward a National CyberInfrastructure

The developers and users of emerging terascale application codes face many challenges. Some applications are compute intensive, requiring multiple teraflop computing systems. Others are data intensive, necessitating creation or mining of multi-terabyte data archives to extract scientific insights. Still others must be coupled to scientific instruments (e.g., radio telescopes, electron microscopes, or beam lines) for near real-time data processing and analysis. Some embody all three aspects. These Grid-enabled applications must couple geographically separated computing, storage systems, and instruments to achieve breakthroughs. For example, the LHC (Large Hadron Collider) experiments will require filtering of petabytes of data to detect 'one in a trillion' events.

Not long ago, computational scientists were forced to travel to use supercomputing facilities. Development of NSFNET, ESnet, and other national networks allowed users to interact with remote systems from their desktops. However, supercomputing is still largely centralized: users submit jobs to central supercomputers, store data on central archives, and post process data on central systems. During the early 1990s, we created 'metacomputers', tightly integrating storage, computing, and instruments. More recently, we began extending this model to integrate resources at multiple sites, moving from a client-server model to a peer-to-peer Grid approach.

Simply put, the Grid represents the next step in overcoming the tyranny of distance. Today we have an opportunity to deploy an entirely new information architecture where bandwidth, along with computing and storage, is the *enabling,* rather than gating, technology. Grid software technologies will enable on-demand construction of powerful virtual computing systems that integrate the resources needed to solve the problem at hand. Scientific collaborations will be able to easily link their data, computers, sensors, and other resources to form a single virtual laboratory. A scientist's request to run a community code could be routed to any suitable resource and output data archived on any Grid storage system; collaborative visualization of the results could occur without concern for the locations of data and users. Complex analyses of experimental data could call upon computing, data, and network resources across the country or the globe. *Realizing the Grid vision requires major enhancements to our scientific infrastructure in three key areas: high-speed networks, Grid services, and Grid-enabled terascale facilities.*

A combination of recent technological advances and strategic partnerships allows us to develop a TeraGrid infrastructure of unparalleled capability and scope. *High-speed networks* are vital to Grid applications, enabling rapid access to remote resources and allowing users to hide latency via aggressive data staging. Rapid advances in network technology mean that soon wide-area networks will be faster than internal computer networks. This change will have profound implications for all aspects of science and society. The DTF TeraGrid will represent the preeminent laboratory for exploring these network implications via the most ambitious scientific projects and by deploying the world's fastest wide area network (WAN) in support of scientific users.

Grid services are the glue that transforms a collection of distributed, independent resources into a coherent computing, storage, and collaboration fabric. Providing uniform mechanisms for user authentication and authorization, accounting, resource access, data transfer, system monitoring, and resource management, Grid services make it possible for users, applications, and tools to discover and use disparate resources in coordinated ways. The emergence of the Globus Toolkit as a *de facto* standard for Grid services makes it feasible to construct a TeraGrid that can provide coherent internal services and interoperate with other Grid systems.

2. Defining TeraGrid

TeraGrid refers to the cyber-infrastructure that is being constructed through a combination of three programs within the NSF TeraScale initiative.

- In 2000 NSF funded the **TeraScale Computing System** (TCS-1) at the Pittsburgh Supercomputer Center, resulting in a 6 TFLOPS computational resource.

- In 2001 NSF funded the **Distributed Terascale Facility** (DTF), which is in the process of creating a 15 TFLOPS computational Grid composed of major resources at Argonne National Laboratory, Caltech, the National Center for Supercomputing Applications (NCSA), and the San Diego Supercomputer Center (SDSC). The DTF will exploit homogeneity at the microprocessor level, deploying Intel Itanium architecture (Itanium2 and its successor) clusters across the four sites, in order to maximally leverage the software and integration efforts. In addition, the homogeneity will offer the user community an initial set of large-scale resources with a high degree of compatibility, reducing the effort required to move into the computational Grid environment.

- In response to the Dear Colleague letter for Principal Investigators (02-119), issued by NSF on April 25, 2002 for an **Extensible TeraScale Facility (ETF)**, the PIs have proposed an ETF that combines the TCS-1 and DTF resources into a single, 21+ TFLOPS Grid environment and supports extensibility to additional sites and heterogeneity among computational resources.

This resulting infrastructure is termed *TeraGrid*, and has several high-level objectives:

1. To provide an unprecedented increase in the computational capabilities available to the open research community, both in terms of capacity and functionality,
2. To deploy a distributed "system" using Grid technologies rather than a "distributed computer" with centralized control, allowing the user community to map applications across the computational, storage, visualization, and other resources as an integrated environment, and
3. To create an "enabling cyberinfrastructure" for scientific research in such a way that additional resources (at additional sites) can be readily integrated as well as providing a model that can be reused to create additional Grid systems that may or may not interoperate with TeraGrid (but are technically interoperable nonetheless).

3. Resources, Services, and Sites

This paper describes the technical and organizational approaches being employed to design and build the TeraGrid, with the objective of providing the community with insight into how TeraGrid might be extended. A key objective to using Grid technologies is to allow resources at multiple sites to be combined with little *a priori* coordination among the sites themselves. This integration creates a *virtual organization*, wherein "a number of mutually distrustful participants with varying degrees of prior relationship (perhaps none at all) want to share resources in order to perform some task." [1]

The TeraGrid is being constructed as a persistent infrastructure within which resources at the individual TeraGrid sites are autonomously managed, but with the sites supporting an agreed-upon set of fundamental Grid services on top of which user capabilities can be built. This distinction between general Grid infrastructure and the user capabilities and services has

implications for expansion of TeraGrid to include resources at new sites. As outlined below, adding resources to TeraGrid is defined on the basis of support for a core set of services and support for one or more user-level capabilities built on top of those core services. In order to sustain a high level of user support and quality of these services, it will be necessary for new sites to participate in the operational, support, and coordination efforts of the TeraGrid. More detail on this aspect of TeraGrid can be found in §5.

3.1 TeraGrid Resources

The initial TeraGrid design includes four large-scale Itanium architecture based Linux clusters at ANL, Caltech, NCSA, and SDSC. These systems collectively will provide on the order of 15 TFLOPS of computational capacity and nearly 1 Petabyte of rotating storage. In addition to this homogeneous DTF environment, extant PACI resources at SDSC and NCSA will be incorporated into the TeraGrid. With the ETF project, the PSC TCS-1 will be included as well. Figure 1 shows the resources that are to be included in TeraGrid as part of the DTF and ETF programs.

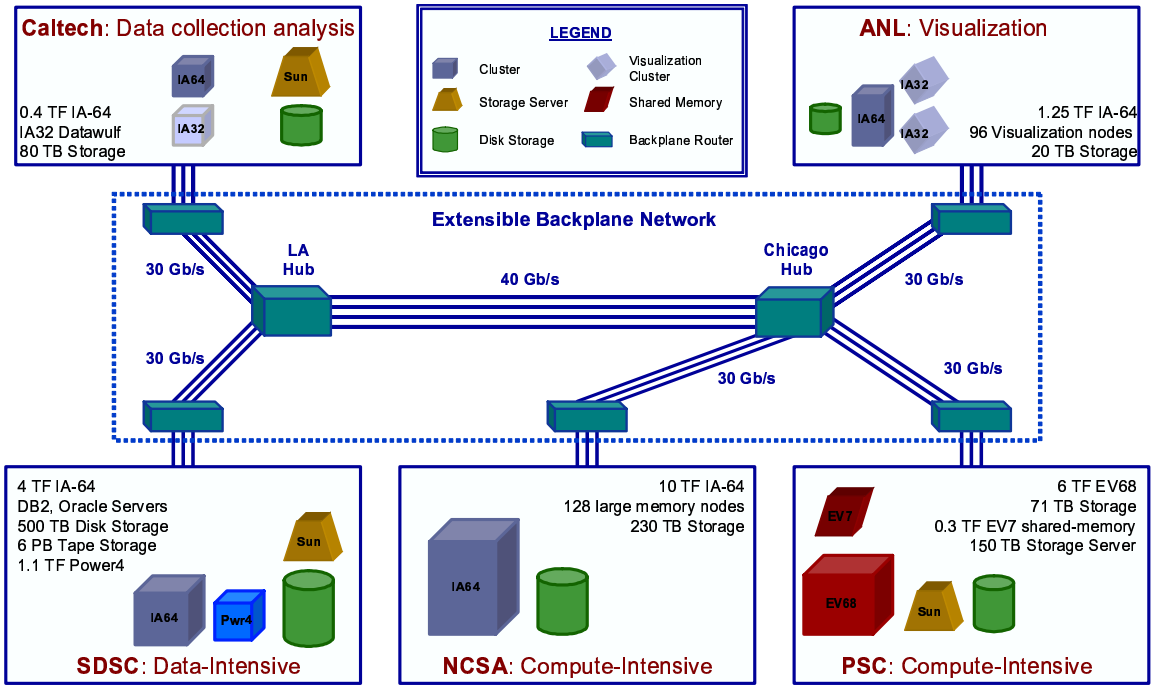

Figure 1: TeraGrid Resources

The TeraGrid architecture and approach is aimed at creating an infrastructure in which the whole is more than the sum of its parts. Simple extension of a working design through replication would not encourage users to utilize novel resources beyond their usual bailiwicks. To that end there has been a conscious effort toward specialization by the TeraGrid sites. This specialization has had significant implications in resource acquisition by the individual centers, with the compute-intensive PSC and NCSA centers investing more in large clusters than the data-intensive, visualization, and data-analysis resources at SDSC, ANL, and Caltech, respectively.

3.2 Architecture and Approach: Grid Services

One approach to building the TeraGrid could be to simply federate the PACI environments, allowing users to select from among resources that are provided by the NCSA Alliance, SDSC

NPACI, and PSC. This approach would provide a larger pool of resources within which users could continue to compute in what is basically a client-server environment. However, PACI allocations are specific to individual resources and thus any given user has access to only a subset of the PACI resources. The PACI program has demonstrated that this is a highly scalable model, and one that provides tremendous benefit to the user community by offering a wider selection of architectures with uniform user support structure and processes for applying for resources.

Another approach to building the TeraGrid would be the "virtual machine room" approach whereby policies and specific implementations would be identical across all sites, with centralized and/or committee-based control, management, and security. One could imagine the four DTF sites attempting to deploy the homogeneous Itanium architecture clusters with identical software configurations and policies. However, even within DTF the clusters are not homogeneous in their individual configurations (some have more memory, others more disk, others have graphics hardware…) and thus require different software configurations despite using the same microprocessors. Agreements could also be made across DTF sites to coordinate operational details such as user IDs and home directories. This approach would be difficult initially, but conceivably possible. It would, however impose unrealistic barriers to entry for additional resources at other sites, or indeed even at existing TeraGrid sites. Further, as difficult as such an approach would be initially it would be nearly impossible to sustain such lockstep coordination over time, as the only way to sustain such an approach over time would effectively amount to atrophy.

The approach that the TeraGrid project has adopted is to use existing Grid software technologies to support the integration of resources into what appears to be a "virtual system" but is in fact composed of resources independently controlled by individual sites. The virtual system will involve a set of agreed-upon "service specifications" that describe the capabilities and behavior of a TeraGrid resource without mandating a particular implementation. In this sense, TeraGrid involves two layers of architecture: the basic software components (Grid Services) and a set of application services (TeraGrid Application Services) implemented using these components. Each of these layers is described below.

3.2.1 Components for Basic, Core, and Advanced Grid Services

The TeraGrid architecture will build on the deployment of three foundational layers of Grid services, each of which builds upon services provided by the layer beneath. These layers are shown in Table 1. Within each layer the TeraGrid project has selected a set of software components to provide the necessary services.

Table 1: Grid Services Layers

Service Layer	Functionality	TeraGrid Implementation
Advanced Grid Services	super schedulers, resource discovery services, repositories, etc.	SRB, MPICH-G2, distributed accounting, etc.
Core Grid Services (Collective layer)	TeraGrid information service, advanced data movement, job scheduling, monitoring	GASS, MDS, Condor-G, NWS
Basic Grid Services (Resource layer)	Authentication and access Resource allocation/Mgnt Data access/Mgmt Resource Information Service Accounting	GSI-SSH, GRAM, Condor, GridFTP, GRIS

Advanced Grid Services, layered on top of the Basic and Core grid services, include enhancements required (for example, those listed in Table 1) for TeraGrid as well as additional capabilities beyond the initial scope of TeraGrid as funded within the DTF and ETF projects. However, many TeraGrid participants are involved in research and development of advanced Grid services and thus it is expected that these will be offered at a future date on TeraGrid.

Core Grid Services (collective layer) build on Basic Grid Services, and involve a combination of components and configuration strategies with a focus on the coordination of multiple resources (ubiquitous infrastructure services). For example initially TeraGrid is using the Globus Monitoring and Discovery Service (MDS) software component to discover and access system configuration and status information. The application of this service to TeraGrid will involve the definition of schema to represent system information within MDS. This schema will be used to store TeraGrid hardware, software, and configuration details within MDS, and is being developed in cooperation with other large-scale Grid projects such as the GriPhyN project in the US and the European DataGrid project.

At the **Basic Grid Services** level (**resource layer**) there is also a combination of component selection and TeraGrid specific implementation. The resource layer is defined by a focus on sharing single resources (negotiating access, controlling use). An example of this is authentication. TeraGrid has chosen an X.509 certificate-based authentication scheme that utilizes the Grid Security Infrastructure (GSI) protocol. The TeraGrid project evaluated both a centralized approach (all users must obtain a TeraGrid authentication certificate from a centrally operated TeraGrid CA) and an approach that will allow the acceptance of certificates from approved Certificate Authorities. For scalability reasons, the TeraGrid project has chosen not to set up a TeraGrid specific CA but rather to document TeraGrid certificate policy requirements and to accept certificates from Certificate Authorities that meet those requirements. Another example is resource access. TeraGrid uses the Grid Resource Allocation and Management (GRAM) protocol as the basis for secure remote access to computational resources. Work is ongoing to define the operational procedures required to support the use of GRAM at individual sites: for example, whether advance reservation is required.

In summary, then, the approach that TeraGrid is taking is to define interfaces (protocols, schema, operational procedures) that a site must support to participate in TeraGrid, without making any statements about the ways in which these interfaces are implemented. A detailed specification of these interfaces at the three Grid service layers is being developed within TeraGrid and will be made available to potential TeraGrid partners in the near future.

The implementation of these interfaces at a particular site is helped by the availability of a standard software release. To this end, TeraGrid is utilizing the NSF Middleware Initiative's (NMI) software release as the Grid software base, working closely with NMI to ensure that any TeraGrid improvements are incorporated into NMI releases. Prospective TeraGrid sites are encouraged to use this release as well. The NMI release may be found at http://nsf-middleware.org/.

3.2.2 TeraGrid Application Services

With all TeraGrid resources operating the appropriately configured basic and core Grid services, it will be possible to submit computational jobs to run on any TeraGrid computational resource. This capability has already been demonstrated within a large number of Grid projects; however most implementations assume a simple standard runtime, or "hosting" environment. (E.g., they might require that every resource run Redhat Linux version 7.2.) The TeraGrid project, with input from a large number of application teams, is developing a set of specifications for application services that are hosted on top of the basic and core Grid services. The initial set of services being defined is shown in Table 2.

Table 2: Example TeraGrid Application Services

Service	Objective
Basic Batch Runtime	Supports running static-linked binaries
Advanced Batch Runtime	Supports running dynamic-linked binaries
Scripted Batch Runtime	Supports scripting (including compile)
On-Demand / Interactive Runtime	Supports interactive applications
Large-Data	Supports very large data sets, data pre-staging, etc.
File-Based Archive	Supports GridFTP interface to data services

As an example of this services approach, the following is an initial set of specifications for the TeraGrid Basic Batch Runtime environment, illustrating the level of detail within the specification framework (but not necessarily this final or complete specification).

Table 3: Example TeraGrid Application Service Specification

Specification	**Basic Batch Runtime**
Objective	*To support execution of compiled binaries*
Assumptions / Policy	• *1 user per node, dedicated during runtime* • *nodes accessible for login (minimal environment)* • *static linked binaries* • *attributes and values published in MDS (using XML schema currently being developed)*
System Software	Query for values (e.g. OS release, kernel level)
Node Hardware	Query for values (e.g. IA64 nodes, n CPU/node)
Software and Libraries	MPICH-G (min version)
I/O	• Local space: $TG_Local • Home space: $TG_Home • Global (shared) space: $TG_Global • Node communication (protocol(s), naming) • Note in/out streaming (to other notes within cluster, to other clusters, outside clusters) • GridFTP
System Commands & Utilities	• Path $TG_UTIL • Min set (e.g. tar, gzip, gsi-scp, gsi-ssh, etc.)

It should be noted that TeraGrid Application Service specifications are still under development and discussion, using a prototype TeraGrid environment ("TeraGrid Lite") as a testbed for evaluation and definition by TeraGrid application and systems participants. Furthermore, while initial TeraGrid sites are expected to support a common set of TeraGrid Application Service specifications, new resources added to the TeraGrid may elect to focus only on a subset.

Finally, new service specifications are expected to be added over time by both current TeraGrid sites and by new sites. For example, ANL is developing a Batch Rendering Service whereby data will be streamed or pre-staged to the ANL cluster for image rendering.

3.3 Defining TeraGrid Service "Compatibility"

By defining the TeraGrid in terms of services and technical specifications, we enable compatibility verification in an objective way that can be automated. As each TeraGrid service definition is finalized, a corresponding software verification program will be written to both verify the

compatibility of the implementation and test functionality and performance. As new services are added to TeraGrid, new verification modules will be made available.

Note that it is not anticipated that all TeraGrid resources will support the full complement of TeraGrid Application services (Table 2). For example, a data service may not support computation services. However, it is expected that all resources will support all basic Grid services (Table 1) and that most resources will also support all core Grid services (Table 1).

Implementing test and verification software will be a useful tool in the definition of TeraGrid services as well as in verifying, in essentially real-time, what compatible resources are available on TeraGrid. Beyond functionality, it is also expected that these test modules might be augmented to report on performance. In time this could be used to periodically adjust allocation charging rates on individual resources (see §3.4) or, eventually, to support charging schemes directly tied to current performance.

3.4 Defining TeraGrid Participation: Allocations and Accounting

At one level, TeraGrid compatibility is defined in a purely technical sense based on the implementation of services (as described in § 3.2). As with current NSF computational resources, the TeraGrid will support users based on allocations. At present, the PACI program supports individual allocations on a per-resource basis, and these allocations are exchangeable only through specific action by users with PACI staff manually affecting the move of allocated time among resources within the PACI program. This manual exchange must occur before the allocation is used, and there is no automated process.

TeraGrid will implement a distributed accounting approach whereby TeraGrid allocations can be used on any TeraGrid resource. Initially there will be a peer-review allocation committee for TeraGrid users who do not require the particular capabilities of any one site, and thus would request time on the TeraGrid in general, with a single allocation that can be used at any TeraGrid site. Use will be debited from a single TeraGrid account, initially through a central usage database.

In order to support this, the TeraGrid sites are working together and within the Global Grid Forum (www.gridforum.org) to develop a standardized usage record format, associated information exchange protocols, and a cooperatively managed database system for tracking usage of individual projects on TeraGrid resources. Structurally, this will involve several components:

a) a standard format for exchange of usage information
b) a set of information exchange protocols that will allow for
    i) querying a resource for usage by a particular project or set of projects
    ii) support for authorization via querying an allocations information service for "balance" information associated with a particular project (how much allocation remains in that project's "account")
    iii) reporting usage of a given project (e.g. a resource reporting to the central usage database after a job has completed)
c) an allocations information service that can be used to track, query, and report on usage and allocation balances

Note that many of these components exist within the PACI program and elsewhere: many projects use locally developed "standard" usage record formats for example, but each local system uses its own record format, and thus there is no common standard. The use of a common record format also assumes a standardized unit of allocation that can be mapped at an exchange rate to a particular resource. Given that the NSF centers already use standardized units for

reporting, and these involve agreements about relative value of resources, the TeraGrid accounting environment will implement existing processes (and not attempting to create more advanced concepts such as Grid free markets or economies, initially).

Participation in the TeraGrid infrastructure through making resources available that involve usage allocations and tracking (in some cases, such as an information service, usage may not need to be allocated or tracked) will require that the resource supports the TeraGrid accounting information exchange protocols and specifications.

Note that scalability and extensibility is a critical factor in this approach. By defining the necessary information exchange protocols and record formats, local implementation is left to the local site. More importantly, sites need not change their local accounting systems in order to be compatible with TeraGrid. Rather, sites will implement software to (a) translate between local usage record formats and TeraGrid usage record formats, and (b) interact with TeraGrid accounting via defined protocols for a finite set of exchanges (e.g., query, report, authorize).

"Exchange rates" are anticipated to be implemented using the standardized unit scheme that has been in place for many years in the PACI program and its predecessor, the NSF Supercomputer Centers program.

3.5 Operations and User Support

The support of users between and across platforms is another aspect of the TeraGrid that must be handled correctly in a collaborative manner between TeraGrid sites. For this reason, an essential component of the TeraGrid effort is the creation of a TeraGrid Operations Center that operates in a distributed manner to provide a single point of contact for receiving and responding to user issues. The distributed Operations Center is defining interfaces and tools for monitoring the status of TeraGrid resources at the sites and a method to facilitate trouble ticket information sharing between the sites to provide a coordinated response to issues that may arise within and across the sites.

User services are also being coordinated across TeraGrid sites, addressing issues ranging from training, documentation and consulting to day-to-day support processes and capabilities.

4. Connecting to the TeraGrid Backplane

It is anticipated that large resources that are integrated into TeraGrid will require extremely high-bandwidth data transfer commensurate with the resource, generally in excess of several Gigabits/second initially. *It is not the case, however, that participation is defined in terms of high-bandwidth connections.*

The TeraGrid project team has designed a "backplane" interconnection network using design principles appropriate for "machine room" networks, oriented toward supporting bursts, or peak rather than average demand. As such, the design and operational model for the TeraGrid backplane is quite different from a general purpose Internet backbone, as will be outlined below.

4.1 Purpose and Use Modalities

One of the key design criteria for the TeraGrid interconnection network was to provide sufficient capacity such that all TeraGrid resources appeared to be part of a single physical facility. In short, this meant that bandwidth was optimized for peak requirements, as is typically done in the design

of a computer room network, rather than for the average load of the aggregation of many users at many sites. As with a machine room network, the TeraGrid interconnection network is designed to scale to a much smaller number of sites and resources than a general purpose Internet backbone network such as Abilene or ESnet. For this reason we refer to the interconnection network as a "Backplane."

The technical architecture and management plans for the TeraGrid backplane assume on the order of a dozen (or small number of dozens of) sites. Further, the TeraGrid backplane is intended to be dedicated to the data transfer needs of TeraGrid resources, as contrasted with general interconnectivity of the sites or general Internet access. It does not replace the general purpose Internet connectivity of TeraGrid sites, or between TeraGrid sites and end users, but rather is a dedicated, separate resource.

4.2 Technical Description of TeraGrid Backplane

The initial DTF network design was based on a full mesh of 10 Gb/s wavelengths, and this resulted in the requirement for four 10 Gb/s wavelengths between Chicago and Los Angeles. Each of the four DTF sites arranged for dark fiber and/or wavelength services to Chicago or Los Angeles, the closest locations where commercial 10 Gb/s wavelength services were available. (DTF sites are located in smaller cities or suburbs, whereas 10 Gb/s wavelength services are typically only available between major metropolitan areas).

With ETF, the backplane was re-designed by creating two general purpose hubs in carrier-neutral fiber collocation facilities in Chicago and Los Angeles, interconnected with the four 10 Gb/s wavelengths. ETF sites are connected to the closest hub with three 10 Gb/s wavelengths. With the introduction of heterogeneity in cluster architecture, particularly related to the local cluster interconnection topologies and external I/O strategy, the TeraGrid project elected to isolate the local resources from the TeraGrid backplane using a dedicated border router. The resulting TeraGrid backplane is thus a backbone topology with hubs and spokes, extensible by adding new sites to existing hubs or by extending the backbone with additional hubs.

The "TeraGrid Backplane" as discussed here includes the hub routers, all border routers, and the wavelengths interconnecting these routers. Simply put, the border routers are considered to be part of the backplane rather than a site resource, and thus are selected, configured, and managed cooperatively by the TeraGrid networking team[i].

4.3 Connection Guidelines: Expanding a Shared Backplane

It is important to note several implementation guidelines that result from the TeraGrid backplane design and management plans:

1. Additional sites will connect, using one or more 10 Gb/s wavelengths, to the closest TeraGrid hub (currently in Chicago or Los Angeles). Connecting sites will cover the costs associated with connecting to the hubs, including necessary interfaces in the hub routers, telecommunications access charges, etc.

2. Border routers, which are "backplane" components rather than "site" resources, will be dedicated to interfacing TeraGrid resources at each site. Border routers were selected

[i] Typically a border router is considered to be part of the site, connected to an external network. However, the TeraGrid architecture considers the border routers to be part of the core of the network. Typically called "Site Border Routers," it is actually more appropriate in the TeraGrid backplane architecture to term these routers "backplane border routers."

within the TeraGrid through a cooperative evaluation effort, examining many potential solutions with the objective to arrive at the best solution for the TeraGrid. Selection of border routers to be added to the TeraGrid backplane will be subject to the same "global optimization" strategy.

3. Border routers will be tightly coupled to the local TeraGrid resource using dedicated LAN equipment, with no intervening shared equipment or firewall. It is anticipated that border routers will connect resources at a single site, and that no border router will serve as both a border router and a hub router. Primarily this is intended to keep the backplane hierarchy from expanding beyond the current 2 layers (hub, borders), however this does not preclude the possibility for embedding resources at hub sites and connecting these to the backplane, should this become desirable.

4. It is anticipated that additional TeraGrid hubs may be required in order to facilitate extensions to identified collections of resources within a region. In order to assure maximum extensibility and flexibility, new hubs will be sited at locations with rich fiber connectivity from a critical mass of suppliers, and in carrier-neutral collocation facilities. The TeraGrid networking team will be responsible for the design and specification of these hub facilities, including selection of hub routers.

4.4 Bandwidth: Requirements and Strategies

As can be seen in Figure 1, current TeraGrid sites are providing a wide range of capacity and functionality. For example, the ANL site will provide a set of specialized visualization services using a relatively small (1.25 TF) cluster and Caltech will provide a set of advanced data collection analysis capabilities, with a 0.4 TF cluster. Like the larger clusters at NCSA and SDSC, the ANL and Caltech clusters have 1 Gb/s (Gigabit Ethernet, GbE) interfaces on all nodes and thus can exchange data with other TeraGrid nodes via dozens of Gb/s streams. PSC's LAN similarly will use multiple GbE channels to interconnect the Quadrics switch with the TeraGrid backplane at tens of Gb/s.

Figure 2 shows the predominant architecture that TeraGrid sites are using to connect TeraGrid resources to the TeraGrid backplane. Each node in the cluster has a single Gb/s LAN interface (GbE) that is connected to a GbE LAN switch. In each LAN switch shown here, 16 GbE connections enter from the cluster and a single 10 GbE exits to a central "cluster aggregation switch," where three such channels are passed through to the TeraGrid backplane (the border router). The cluster aggregation and border routers are separate for two main reasons. First, this separation allows for local cluster configuration changes, outages, or experiments to be done without affecting the operation of the TeraGrid backplane. Conversely, it allows for experiments and other operations that might disrupt the TeraGrid backplane without affecting the operation of the local TeraGrid cluster. Second, the requirements for switching and routing network traffic over LANs versus WANs are quite different. Enterprise IP routers (such as are being used for the border routers and hub routers) are designed to handle the necessary buffering and associated requirements for long-delay, high bandwidth wide area networks. LAN switches, on the other hand, are optimized for short-delay, low-latency connectivity as would be expected in a LAN environment.

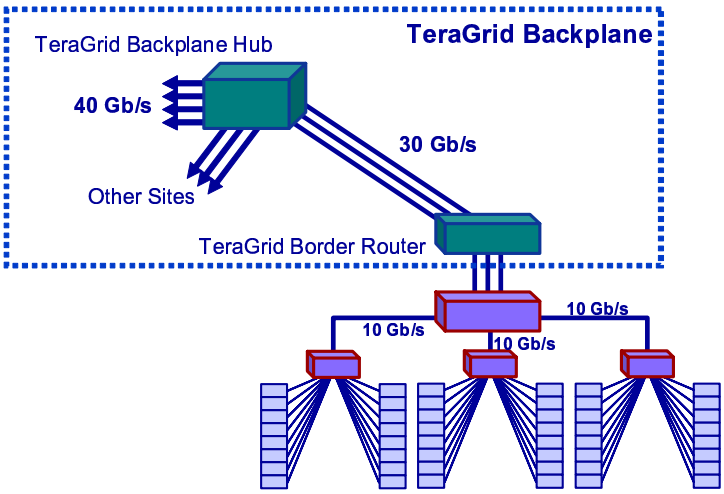

Figure 2: *Example* LAN connectivity to TeraGrid Backplane. Each rack shown here contains 8 nodes, each node with 1 GbE. Aggregate cluster bandwidth capacity shown is 48 Gb/s. In practice, TeraGrid sites are typically using LAN switches that support 2-300 GbE and 2-3 10 GbE interfaces per switch.

4.5 TeraGrid Backplane Management

A TeraGrid network team has been formed with a minimum of one technical representative from each site. This team cooperatively manages all TeraGrid hub routers and border routers, which together constitute the TeraGrid backplane. In practice this means that ANL (lead networking partner) is responsible for management of hub routers and each site is responsible for management of the local border router. The entire team is notified of all changes made to any TeraGrid hub or border router. Each new TeraGrid site will participate in this team.

The TeraGrid backplane will be managed with a 99% uptime ("two nines") availability goal. This allows for brief planned outages for maintenance, upgrades, reconfiguration and experiments of various types ranging from applications experiments with dedicated network resources to middleware experiments investigating network control to low-level networking experiments. In some cases, hub and border routers will be included in disruptive experiments (e.g. middleware making changes to router configurations) under controlled conditions. Over time we anticipate growing interactions between middleware and network management, and this will require extensive cooperation among sites and tolerance for experimentation.

4.6 TeraGrid Backplane Acceptable Use Policies

Traffic across the TeraGrid backplane will be limited to TeraGrid resources. Routing policies are expected to exclude for example, user desktop machines or transit traffic.

The TeraGrid backplane will connect to Abilene at the backplane hubs in Chicago and Los Angeles for the exchange of a limited set of pre-approved routes to optimize user access to TeraGrid resources. As noted earlier, however, TeraGrid sites are required to keep separate Internet connectivity, as the TeraGrid backplane is dedicated to TeraGrid interconnection and cannot be used for general Internet access.

The TeraGrid backplane may also be connected to other networks (ESnet, NREN) as the needs and requirements develop.

5. Organization and Participation Issues

5.1 Construction and Architecture Processes

The TeraGrid project is a construction project intended to build a specific Grid for a particular group of users, within the constraints that resources must be controlled and tracked as with other NSF-funded high performance computing facilities. The management structures and processes required to design, deploy, and operate the TeraGrid are not intended to scale nor extend beyond a relatively small number of TeraGrid sites. Rather, it is expected that many sites, and other operational Grids, will be able to interoperate with the TeraGrid based on standard protocols and service level agreements. For example, the various sites participating in the International Virtual Data Grid Laboratory (www.ivdgl.org) are expected to link with TeraGrid in support of specific data-intensive science experiments.

The TeraGrid systems at ANL, Caltech, NCSA and SDSC are designed to be homogeneous with respect to node configurations, software environments and applications support. This homogeneity was driven by the desire to develop a comprehensive cyberinfrastructure that was usable and supportable as the TeraGrid. The responsibility for designing and implementing the TeraGrid systems at these four sites rests with working groups in the areas of Clusters, Data, Networking, Grids, Applications, Visualization, User Services, Operations and Performance Evaluation.

Incorporating heterogeneous systems in the construction of the ETF, first at PSC with the TCS-1 and later at SDSC with Power4-based systems, are the mechanism to develop and support the heterogeneous ETF software environment by the addition of staff from PSC to these working groups.

As part of the construction process for the TeraGrid, a set of deliverables and milestones have been developed that are being addressed and met by the working groups and the five ETF sites. This implementation of the heterogeneous TeraGrid will enable the incorporation of additional sites that are able to provide resources of interest to the user community, compatible grid services as specified in the TeraGrid documentation and verified by the compatibility test suites that are being developed by the working groups.

5.2 Construction Processes and Management Model (included for information)

The TeraGrid project requires an extremely high level of cooperation and collaboration between the sites to be able to develop, deploy and support the computational environments and users of the overall system. Many of the issues (technical and policy) span multiple technical areas and multiple sites, such as those related to accounting and account management. As such, it is critical that the best possible technical planning and implementation is in place and this must exist within the existing framework of the sites to function effectively and draw on the experiences and expertise at each site.

For this reason, the leadership and management of the TeraGrid project implementation and construction involves an overall project manager working with a team of Site Leads. Site Leads are responsible for ensuring that the overall implementation and policies at their respective sites are consistent and supportive of the TeraGrid project goals and the Technical Working Groups which focus on technical issues in critical areas of the TeraGrid, and that appropriate staffing is assigned at those sites to construct and operate TeraGrid.

5.2.1 Site Leads

The Site Lead is a senior person from each site who is responsible for the overall implementation of the TeraGrid at their respective site. Each Site Lead will:

- oversee the activities (development and deployment) on site.
- serve on the leadership team of the TeraGrid Operations Center (TOC).
- identify the contents of new versions of the DTF software and hardware configuration.
- manage integrated software prototype deployment, with a period of friendly user testing prior to production deployment.

It is the responsibility of each Site Lead to ensure that the implementation and policies at the site are consistent with the TeraGrid goals and policies.

5.2.2 Technical Working Groups

The Technical Working Groups are composed of staff and researchers from each site who are responsible for the technical aspects of the TeraGrid in specific areas. There are currently nine domains for which working groups have been formed:

- Applications
- Clusters
- Data
- Grids
- Network
- Operations
- Performance Evaluation
- User Service
- Visualization

The Working Groups provide the technical leadership of the TeraGrid for their respective areas, producing requirements and specifications, documentation and implementations. The Working Groups have varying levels of responsibility for architecture, implementation, and operations with the overall emphasis for the groups shifting to operational and user support during the production phase of the TeraGrid.

As the TeraGrid evolves with the addition of new sites and resources, they should be included in the operational aspects of the technical Working Groups. The group of Site Leads will expand to include operational leads from the additional sites, who will be responsible for the compliance of their sites with the implementation requirements and user environments for scientific users of the TeraGrid.

5.3 TeraGrid Site Operational Support Teams

Collectively, participation in the TeraGrid Site Leads and Working Group activities will result in the formation of site TeraGrid teams that consist of a minimum of the following functions:

- Site Lead
- Networking
- Resource administration (e.g. cluster administrator)
- Grid software
- User Services (including accounting and allocation support)

The TeraGrid experience to date indicates that in the initial stages of implementation each of these functions requires a substantial effort on the part of a local expert, with some requiring 25-50% FTE per function, partly due to the need for local implementation of TeraGrid protocols and

specifications (e.g. accounting record exchange protocols). In steady state operations it is anticipated that the effort required will be reduced to 10-25% FTE for each function.

6. References

[1] The Anatomy of the Grid: Enabling Scalable Virtual Organizations. I. Foster, C. Kesselman, S. Tuecke. *International J. Supercomputer Applications*, 15(3), 2001.