Mark Linesch, HP
April 23, 2007

Grid – Distributed Computing at Scale
An overview of Grid and the Open Grid Forum

Status of This Document

This document provides information to the Grid community. It does not define any standards or technical recommendations. Distribution is unlimited.

Copyright Notice

Trademark

Contents

1. Introduction

In today's global economy, organizations are being challenged to break down process, technology and people silos that inhibit the flow of information, innovation and commerce. Individuals are being challenged to work in new ways – often in collaboration with other organizations, departments, and/or disciplines. This paper explains why Grid is the IT infrastructure solution being used by leading organizations around the world to enable this knowledge-based, global economy. It explains the role of grid technologies within the broader distributed computing landscape and defines the three common categories of grid use today. It identifies the evolution of Grid technologies from application-specific solutions to dynamic, shared and service-oriented infrastructures. Finally, it describes the role that the Open Grid Forum plays in accelerating grid adoption in partnership with the grid community and the industry at large.

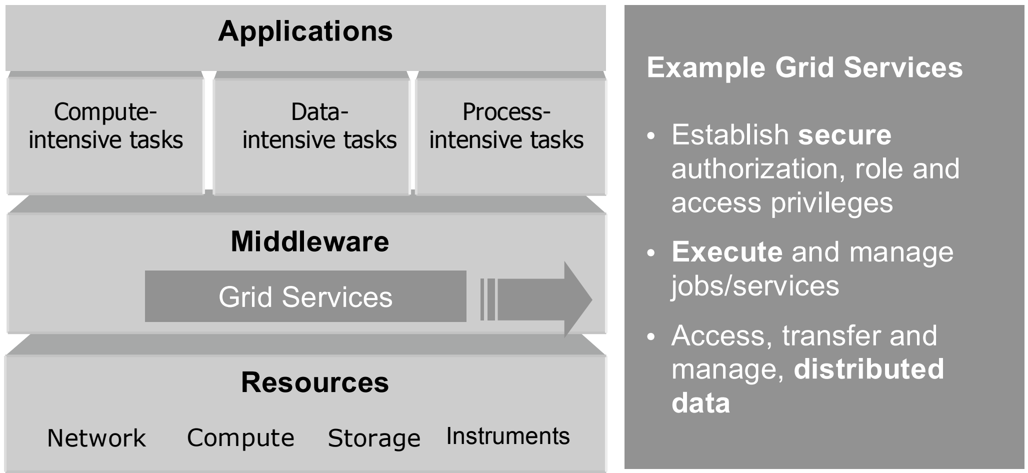
2. Grids: Distributed Computing at Scale

In the broadest sense a Grid[1] can be summed up as follows: *"Scalable distributed computing across multiple heterogeneous platforms, locations and organizations*." The term *"scalable"* refers to the need to operate and manage distributed resources as a secure, robust infrastructure – particularly as this infrastructure grows, shrinks and changes in response to user and/or organizational needs. The term "*distributed computing"* suggests that the network-connected resources (e.g. network, compute, data, and instruments) harnessed to do the work may be of different types (heterogeneous) and in different physical locations. The term *"across multiple heterogeneous platforms, locations and organizations"* refers to the reality that these distributed resources may be owned and/or managed by different entities. The notion of distributed computing as used in this definition includes a wealth of highly complex technologies, some still the focus of research. It includes a broad spectrum of deployments - from homogeneous grid-like HPC clusters at a single location to operation across a wide area networks and multiple domains of administrative control. The security, privacy, economic, and political aspects of Grids increase significantly with the introduction of Internet scale operation.

Grid is all about scaling IT –operating at scale, managing at scale, and changing at scale to support scientific collaboration by geographically dispersed researchers and dynamic business process that enable greater business agility.

As shown in the graphic, control of a grid solution is managed by middleware software, which provides a consistent, standards-based set of grid services for applications to interact securely with instrument, network, compute, information and storage resources, irrespective of their type or location.

[1] See the glossary for a more detailed definition



Grids emerged in the mid-1990s, when applications used by engineers and researchers were run on High Performance Computing (HPC) clusters to side-step the high costs of supercomputers. Clusters grouped together low-cost and often underutilized resources into a scale-out, grid-like infrastructure for primarily high performance/throughput applications. These early ancestors of grids – typically implemented within a single site of an organization, running on a homogeneous suite of equipment – demonstrated the resource-sharing and scaling that grids enable for compute- and data-intensive applications.

Today grids can be found in a variety of organizations around the world in such diverse areas as collaborative scientific research, drug discovery, financial risk analysis, forecasting, design, simulation, business intelligence and transaction processing environments. Although grids have many common characteristics such as infrastructure virtualization, resource pooling and sharing, and dynamic resource provisioning, there are several broad categories of grids including: Cluster Grids, Collaboration Grids and Data Center Grids.

Cluster Grids are the most widespread grid usage model today – primarily aimed at high performance/throughput computing and the scheduling of parallel workloads across a scale-out infrastructure. They tend to be less dynamically deployed and more homogeneous in their construction.

Leading companies such as Johnson & Johnson (see sidebar) are operating multi-site cluster grids in their enterprise production environments today. Not surprisingly, the grids provide a cost effective resource for compute-intensive and data-intensive applications. J&J is an example of the growing trend by organizations to use grids to deliver to users – whether financial traders or research teams developing new drugs – a flexible and adaptive infrastructure. These and other capabilities have led Gartner Inc. to identify Grid computing as one of the top 10 technologies to watch in 2007.

Johnson & Johnson speeds drug discovery

Building on the early success of a small-scale pilot gird in 2003, J&J today has an enterprise grid pooled from resources across the world. The grid runs dozens of applications that shorten the R&D time needed to bring new drugs out of the research labs and into clinical trials and commercial launch. Grid-based applications give their scientists the tools to simulate complex chemical compounds in far less time, and at a lower cost, than with other approaches. The grid also provides the scale and reliability needed by working teams throughout J&J sites to access shared, multi-terabyte datasets and speed drug discovery.

Collaboration Grids involve multiple organizations and individuals, security domains, protocols, discovery mechanisms, and heterogeneous hardware, collaborating to share their resources to make the most effective use of it for their combined user communities. This permits the members of a distributed team (virtual organization) to exploit each others expertise in near- or real-time to achieve results that would be more difficult to achieve working separately or serially. This is the original and long-term vision of Grids.

There are many examples of collaborative grids in widespread usage around the world – often enabled through government funding to advance scientific discovery and economic development within a country or region. These "regional grid infrastructures" include: ChinaGrid, TeraGrid and the Open Science Grid in the US, the German D-Grid Initiative, Japan's National Research Grid Initiative (NAREGI), The UK e-Science Program and EGEE and EGEE-II (Enabling Grids for E-sciencE) in Europe to name a few. For instance, EGEE provides a grid used for scientific collaboration by researchers across the world on topics as varied as modeling drug effectiveness against the avian flu virus H5N1 to enabling physicists around the world to run their high energy physics experiments using the petabytes of data generated from the Large Hadron Collider (LHC) coming online during 2007. A massive grid (see sidebar), EGEE provides resources for compute- and data-intensive applications, but also provides the flexibility to meet the rapid changes in which teams – across organizational boundaries – form, operate and disband. EGEE has found Grid delivers a flexible and adaptive resource that supports new collaborative models for scientific research.

EGEE: Enabling collaborative discovery

Grid is the heart and soul of EGEE's ability to support over 100 "Virtual Organizations" of scientific endeavor. Their grid includes over 30,000 CPUs across more than 200 interconnected sites in 39 countries. A 2 Gbit/s transfer rate supports an average of 30,000 jobs per day, some using/generating multi-petabyte datasets. With this powerful architecture, EGEE can meet the evolving demands from their wide-spread communities.

Data Center Grids span 1 or more enterprise data centers and are in many ways as complete technically as collaboration Grids. They involve the complete dynamic life cycle of service deployment, provisioning, management and decommissioning as part of their normal operation. At first glance, they may appear to be missing the aspect of multiple administrative domains, but that is typically an illusion. While the funding may come from a single source, and the administration carried out by a single organization, there is typically just as much tension among the various user entities as in a Collaboration Grid.

Examples of Data Center Grids can be found in companies such as Amazon, eBay, and Google that support Internet-scale data centers and in organizations that provide IT as a "utility service" (either public or private). Operating at scale, managing at scale, and changing at scale is the life-blood of these organizations and many have invested significant time, effort and intellectual property in their data center architectures and technologies ahead of the availability of standard, "off-the-shelf" capabilities being available in the market. However, even these leading edge firms acknowledge the importance of a comprehensive portfolio of production-proven grid-enabled products that are now becoming available today from open source providers and commercial vendors. As these capabilities become more available, enterprise organizations are increasingly utilizing grid and related technologies to re-architect their enterprise data centers –

eBay: Managing Commerce at Scale

"We have a grid-like computing architecture in which network-based scaling delivers the availability, adaptability and cost-effectiveness demanded by our business model. Standards-based interoperability and components will be an essential part of our ongoing ability to run applications on our grid platform."

Paul Strong, Distinguished Research Engineer, eBay breaking down existing application and information silos and moving toward shared and service oriented infrastructures. Enabling a more flexible and dynamic relationship between the changing needs of business and the supporting IT infrastructure is a journey that many IT organizations have embarked on.

3. Grids and other Distributed Computing Concepts

Grids are at the heart of the IT architecture journey - from silo'd, statically bound applications and resources managed by manual processes to a new world of shared, dynamically provisioned resources that reliably delivers application services to users. Grid is a core concept in modern distributed computing architectures and is aligned with other important distributed computing technologies such as virtualization, service-orientation and data center automation. In a sense, Grids are enabled by virtualization, automation and service-orientation technologies and also integrate these technologies into a unifying solution – particularly across functional and organizational boundaries.

A basic characteristic of Grids is the capability to virtualize applications, information and other IT resources such as networks, servers, storage, and desktop computers. Virtualization is the logical representation of a resource that is separated – abstracted – from its physical implementation. Virtualization frees applications and information from being statically bound to dedicated physical IT infrastructure such as servers or storage. Resources can be pooled, shared and aggregated, whether they are in the same building or across the world. Virtualization is typically thought of in terms of IT infrastructure resources such as computers, storage or networks. Grid takes the virtualization concept to another level and virtualizes the application and information resources that rely on the IT infrastructure. It is important to note that many Grids today are built mostly from non-virtualized IT resources (e.g. computers, storage and networks that are NOT virtualized) and yet the virtualization of the application and the information still allows the Grid to create an environment where resources can be pooled shared and re-purposed easily. Grids also complement one of the hottest trends in virtualization - server virtualization, by pooling virtualized and non-virtualized server resources *across* a wide range of operating systems and platforms that are controlled and managed as a common resource. This is in contrast to server virtualization, which is the partitioning of resources *within* a single physical system.

Grids also support and exploit an increasingly popular architectural style for building and managing applications referred to as service-oriented architecture or SOA. SOA is a way of architecting software to support repeatable tasks (i.e. services). Services are modular building blocks isolated from the specific internal implementation details of other services and the underlying IT resource infrastructure. They are defined by the specific interfaces they publish and also hide the details of their implementation. Each service has a name, a purpose, and policies for things such as security and service levels. Services can be composed of and/or utilize other services to complete a specific task or job. The purpose of a service may be as simple as retrieving information or as complex as executing a business process. Service oriented architecture is a natural style both for use in implementing grid management middleware and as a pattern for applications intended to be hosted in grids. Services are largely written as relatively small containable functions and then aggregated or composed to build applications. Grids provide an ideal unifying infrastructure on which to run such loosely-coupled, composed, service-oriented applications given their capability of managing heterogeneous IT resources scaled across organizational and geographic boundaries. Grid middleware software components also exploit service-oriented concepts in their design and operation. Based on broadly adopted Internet and web services standards, these “Grid Services” enable the discovery of appropriate resources upon which to run applications, help to describe, execute and manage jobs, access and move data, and in general enable a robust and secure environment for the execution of a variety of scientific, engineering, and business applications.

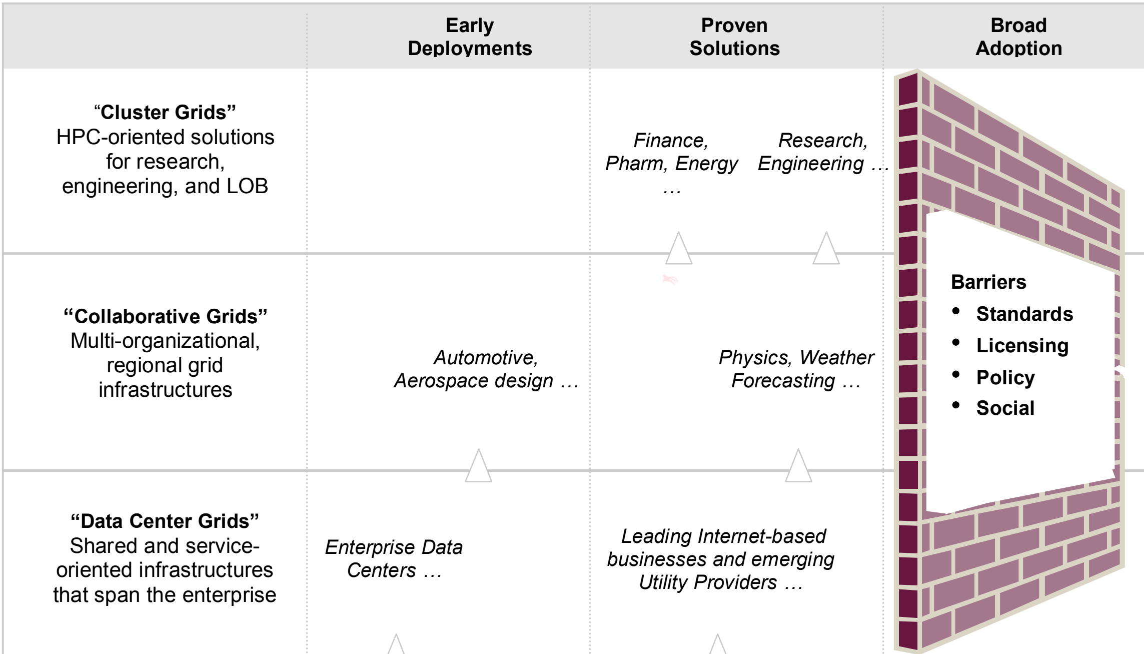
Grids can also provide a unifying framework for policy-based automation. Automation is critical to managing the complexity of distributed systems - insuring required service levels are maintained based on pre-defined policies in a cost-efficient manner. For instance, Grid Services provide a control structure for the scheduling and provisioning of resources, the graceful failover of resources and the dynamic scaling of resources to meet workload demand based on policies set by the organization and/or IT professional. This ability to dynamically discover, assemble, operate and release the resources needed to accomplish a given application task securely is an important characteristic of grid technologies and a requirement for the cost-effective operation and management of distributed systems at scale.

> *“Modern IT infrastructure has grown and sprawled and interwoven without a lot of systematic pre-planning. It’s really no wonder that it’s exceedingly complex, discouragingly inflexible, and alarmingly fragile; it’s no wonder that it’s a complete nightmare to manage.*
>
> *Unless we begin to grow IT in a thoughtful and architected way, we’re just continuing to enlarge the pile – and all the problems that go along with that.*
>
> *Grid is essential to IT because it’s the architecture for gracefully adding new resources, and for dynamically adjusting resource allocations over time. Grid is all about scaling IT – not just for adding stuff to a pile, but for operating at scale, managing at scale, and changing at scale. There really is no alternative.”*
>
> Jonathan Eunice
> Founder and Principal IT Advisor
> Illuminata

4. Grid Adoption within Organizations and the Overall IT Industry

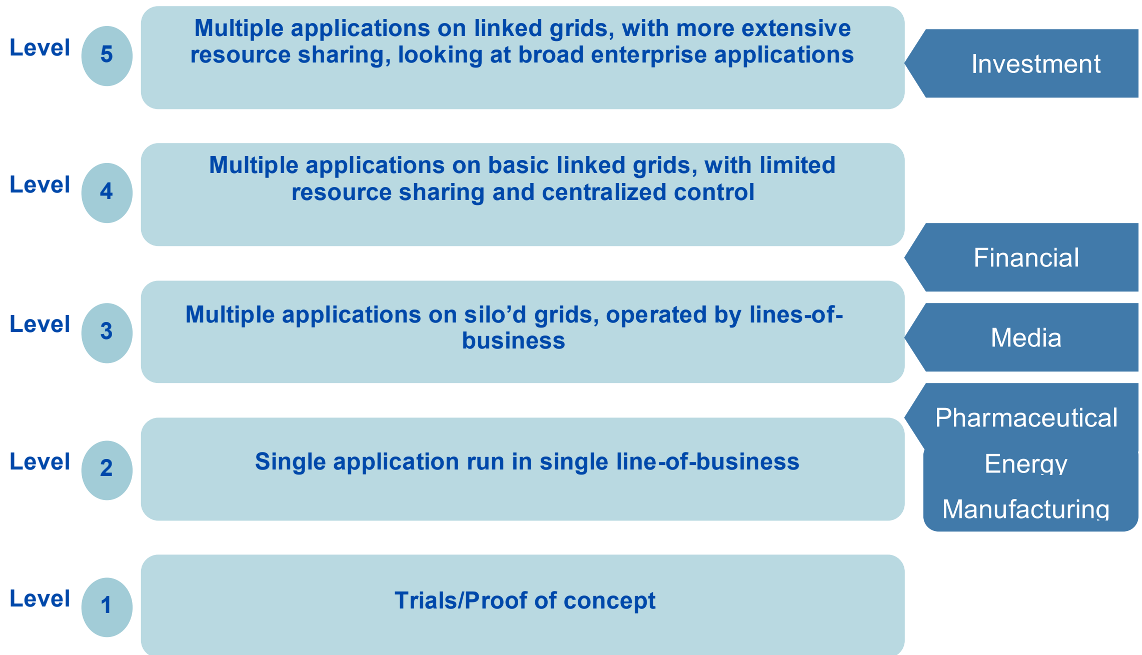
Adoption of grid solutions can be viewed from both an overall IT industry perspective and an organizational perspective.

From a broad IT industry perspective, the adoption model for grid is progressing in 3 phases as indicated by the chart below: (1) early adoption; (2) proven solutions and (3) broad adoption.



Phase 1, "early adoption" is primarily an exercise in handcrafting solutions. In the "proven-solutions" phase, grid-enabled software is available from vendors and grid success stories in specific scientific disciplines and industry sectors are available. These proven solutions provide real-world examples of the benefits and risks of grid deployment and enable other organizations to leverage the successful experiences of the early pioneers. In the "broad-adoption" phase, mainstream users begin adopting grids and packaged solutions are available from a variety of providers. Moving to broad adoption requires the "lessons learned" from early adoption, the "success patterns" from proven solutions and breaking through key standard, software licensing, security/administration policy, and social barriers associated with distributed systems. Standards are particularly important for broader, more mainstream adoption because they enable organizations to quickly and inexpensively connect grids within their organizations and/or to grids of external organizations (e.g., trusted partners, research collaborators). Breaking through these barriers and ensuring standards are in place is the work of the Open Grid Forum.

From an organizational perspective, adoption often progresses from a simple trial to a first successful implementation running a single specialized application. However, successful grids deployments are infectious and often lead to larger, more complex grid infrastructures that span multiple locations and extend to trusted partners and collaborators – ultimately influencing an organization's overall enterprise architecture. The illustration below from the 451 Group, a leading grid and distributed computing analyst organization captures this progression from specialized solution to mainstream IT architecture – highlighting adoption by leading industries to date.



Source: 451 Group, January 2007
Grid Computing - The State of the Market

Today, organizations are adopting grid solutions and achieving significant benefits even without having all the standards in place for interoperability. This is particularly true for enterprises running the same version of one of the popular grid middleware software products. Interoperability becomes more critical as organizations connect grids to other grids within their organizations or with other organizations that utilize different grid middleware software. For those grids to come together and interoperate, they need to speak the same language.

5. The Open Grid Forum's Role in Accelerating Grid Adoption

As grid solutions become widely adopted, the need for interoperability and standards increases. The Open Grid Forum (OGF) is a Standards Development Organization (SDO) dedicated to developing open standards for grid interoperability. It serves as a global forum where the grid community gathers to identify common requirements, develop best practices and share use cases. OGF specifications are adopted and productized by software providers, referenced by SDOs working on complementary standards, and utilized by end users deploying grids.

The Open Grid Forum has a goal that commercial and academic organizations will build operational grids using OGF-defined, standards-based components by 2010. This work is well underway, however much more effort is needed to develop and mature specifications. In January 2007, OGF published its roadmap document entitled, "Technical Strategy for the Open Grid Forum 2006-2010"[2]. The document provides the distributed computing community with a sense of the maturity of existing grid

[2] The most current version of the document can be accessed at http://forge.ogf.org/sf/go/doc13748

specifications and gaps. It identifies six high priority capabilities that need to be developed including:

- **Grid Security*:*** to securely transfer data, authenticate users, and authorize access to resources

- **Application Provisioning***:* to discover, describe, provision and manage the lifetime and lifecycle of software

- **Job Submission:** to submit jobs, query the status of running jobs, and cancel jobs that are executing on a distributed system

- **File Movement:** to move data and manage data including the ability to “cancel,” “suspend,” and “resume” when appropriate

- **Data Provisioning:** to handle files, databases, caching, transport, metadata, and federation at both the data and storage levels

- **Grid Application Programming Interfaces (APIs):** to provide programming interfaces and abstractions that provide stability across middleware technologies and the underlying protocols as they evolve

OGF recognizes that it takes cooperation and collaboration across the entire distributed computing community to effectively build open standards. For instance, many OGF standards are based on the foundational protocols, information, and web services standards developed by other SDOs, including W3C, IETF, SNIA, DMTF, and OASIS. OGF proactively engages in liaison activities with these SDO’s and they, in turn, look to OGF as uniquely chartered to define interoperable grid architectures, specifications and community practices.

In addition, the vendor and open source community are the consumers of specifications and are key partners in their development. End users have a particularly critical role to play in encouraging their vendors to deliver software based on industry standards and to provide software licensing models to support shared and service-oriented IT infrastructures.

OGF also engages extensively with other grid-related organizations throughout the world to ensure that we align globally but also communicate and collaborate locally. OGF affiliates include: Grid Consortium Japan, Grid Forum-Netherlands, Grid Forum-Korea, The Israeli Association of Grid Technologies (IGT), and Grid Forum-Singapore.

6. Membership in the Open Grid Forum

As a community-initiated organization, OGF involves more than 300 organizations from 50 countries. OGF has extensive engagement with the largest national/regional grid initiatives in 25 countries, such as TeragGrid™ and Open Science Grid (U.S.), EGEE (Europe), NAREGI (Japan), APAC (Australia), and UK eScience (U.K.). Leading hardware, software, and solutions vendors such as EMC, Fujitsu-Siemens, Hewlett-Packard, IBM, Intel, Microsoft, Oracle, Platform Computing, and Network Appliance are also actively engaged. Finally end users of grids such as Boeing, Micron, Shell Exploration, and eBay actively participate in the work of the organization.

Active membership in OGF provides the benefits of:

- **Insight** into technical directions, best practices, and the evolution and adoption of standards
- **Influence** on the priorities of the organization and the opportunity to be in leadership roles
- **Recognition** as a leader in the development of next generation of grid and distributed computing, or for individual contributions

> *The benefits of OGF membership to SAS*
>
> *"Membership in OGF provides SAS a unique global forum in which we can exchange ideas with other leading organizations across the world and create common solutions for our customers to improve their business performance and reduce costs. The OGF standards and best practices are an important part of our plans to support our customers' business intelligence, data integration and analytic requirements across grid and distributed computing infrastructures.*
>
> Cheryl Doninger, R&D Director, SAS Institute Inc.

We encourage everyone interested in grids to participate in our community as we work to ensure the pervasive adoption of grids.

For more information on OGF, please visit one of the following:

General Information	www.ogf.org
Current List of Members	http://www.ogf.org/Members/members_members.php
Membership Program	http://www.ogf.org/About/abt_getinvolved.php
How to Get Involved	http://www.ogf.org/About/abt_getinvolved.php

7. Glossary

Capability	A set of one or more *services* that together provide a function that is useful in a *Grid* context. [6]
Enterprise	The entire [business] organization[3]
Grid	A system that is concerned with the integration, *virtualization*, and *management* of *services* and *resources* in a distributed, heterogeneous environment that supports collections of users and resources (*virtual organizations*) across traditional administrative and organizational domains (*real organizations*). [6]
Grid Service	The formal definition of this term is deprecated. In general use, a Grid service is a *Web service* that is designed to operate in a *Grid* environment, and meets the requirements of the Grid(s) in which it participates. [6]
High Performance Computing	The use of supercomputers and computer clusters, that is, computing systems comprised of multiple (usually mass-produced) processors linked together in a single system with commercially available interconnects.[4]
Line of Business (LOB)	Divisions of a company responsible for the production and creation of the organization's products and/or services. IT, HR and Accounting are not lines of business.[5]
Open Grid Services Architecture (OGSA)	Developed by OGF, OGSA is a grid-specific implementation of Web services that work with XML, SOAP and WSDL (among others) across multiple types of transport protocols (e.g., HTTP, SMTP).
Provisioning	The activity of specifying, *reserving*, *allocating* and *deploying* the set of *resources* required to accomplish a task. [6]
Resource	A resource is an *entity* that is useful in a Grid environment. The term usually encompasses entities that are pooled (e.g. hosts, software licenses, IP addresses) or that provide a given capacity (e.g. disks, networks, memory, databases). However, entities such as processes, print jobs, database query results and *virtual organizations* may also be represented and handled as resources. [6]
Service Level Agreement (SLA)	A contract between a provider and a user that specifies the level of service that is expected during the term of the contract.They might specify availability requirements, response times for routine and *ad hoc* queries, and response time for problem resolution (network down, machine failure, etc.). [6]
Services-Oriented Architecture (SOA)	This term is increasingly used to refer to an architectural style of building reliable distributed systems that deliver functionality as *services*, with the additional emphasis on loose coupling between interacting services. [6]
Web Services (WS)	A software system designed to support interoperable machine- or application-oriented interaction over a network. [6].

[3] "The Computer Desktop Encyclopedia" by Alan Freedman, Second Edition. 1999
[4] See http://en.wikipedia.org/wiki/High_performance_computing
[5] See http://it.csumb.edu/departments/data/glossary.html
[6] See http://www.ggf.org/documents/GFD.81.pdf

8. Intellectual Property Statement

9. Disclaimer

10. Full Copyright Notice