Building
Tomorrow's Enterprise

White Paper

Infosys®

Sharpen your business edge with Grid Computing

Pankaj Misra

The rapid spread of IT systems in the aftermath of the dotcom era has led to server sprawl. Applications run in silos and most servers are underutilized. Companies are quickly realizing that this means sub-optimal RoI. They are also seeing the benefits of collaboration and sharing not just at the data level but also at the infrastructure layer – something that is made feasible through Grid Computing.

Jun 2006

As a business leader, would you accept 30% of your employees sitting idle, or being busy only for a couple of hours a day? If not, then why accept low utilization levels for assets like servers?

Business users, designers and architects use conventional wisdom and plan for "peak" workload. They are willing to spend a little more money than run the risk of the system crumbling under peak demand. Applications and hosting infrastructure are therefore designed to handle peak workloads.

While this assumption is largely valid, it is also true that "peak workload" for all departments within a company may not occur at the same instant. When we consider that applications are usually hosted on dedicated infrastructure, it becomes clear that while a few servers are busy serving peak loads, other servers in the same company (whether in the same or different locations) are running idle. This is clearly sub-optimal utilization of resources. Viewed differently, the investment thus locked in could easily have been used in other ways to build and strengthen competitive advantage.

Distributed isolated systems tend to run underutilized in almost all companies. Unix servers have a minimum average underutilization of 30%, while for desktops, the figure is a staggering 95%. Many companies are actively upgrading desktops to the latest chips in the market. However, do you really need a 3.14 Ghz chip to browse, create a Word document or a presentation? Imagine the load of CPU cycles wasted!

It is in this context that this paper describes Grid Computing as a credible way to virtualize computing, storage and network resources with a view to make all resources available to those applications and/or users who need them the most at any given instant in time.

What must Grid Computing be leveraged for?

The necessary outcome of grid computing is virtualization of resources. This should be leveraged to contain server sprawl, drive server consolidation and share resources and data. In the financial services industry, the computing power thus unleashed can help in dramatically reducing the time needed for completing complex and compute-intensive risk analyses and running sophisticated pricing models. Because grid computing frees up CPU cycles from underutilized resources (servers and desktops), compute-intensive, scenario-based market and credit risk simulations can be performed with more iterations and a greater accuracy. This in turn can enhance service levels and enable the firm to meet strict regulatory requirements more efficiently.

The grid's dynamic resource management features ensure that the failure of a particular node does not affect the system's performance and availability. In the event of a failure, the load is dynamically allocated to other available resources that serve the application demands. Resources are transparently made available to users, providing guaranteed higher uptime for mission-critical systems at a reduced cost of operation.

Are Grid Computing and Cluster Computing synonymous?

No. Grid and Cluster computing are two different implementations of the same vision of harnessing the collective power of resources to meet application demands however there are many differences.

A cluster is typically built of homogenous operating system and chip architecture, while a grid typically consists of heterogeneous resources – different operating systems, chips and storage resources. A cluster is built of tightly coupled interconnections, generally in a single location, which rely on very low network latency. Grid resources are distributed across LAN, MAN and WAN. Clusters complement a grid. Clusters can be one resource on the grid just like any other participating computer, storage and network resource.

What are the basic building blocks of grid?

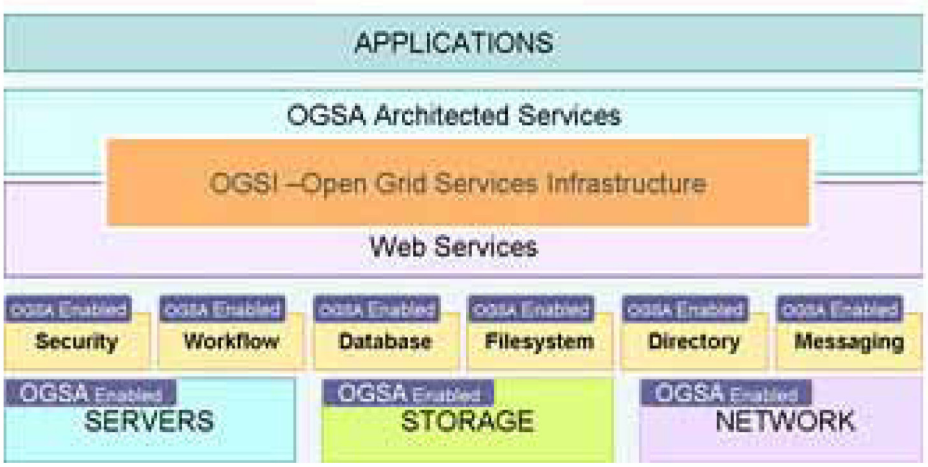
A grid has six building blocks:

- Grid Portal – A web-based portal to manage and monitor applications and grid resources
- Resource Management – To allocate jobs to specific resources and retrieve results
- Data management – To move data across applications and resources
- Grid Scheduler – To handle job queues and manage alternatives when encountering reserved resources
- Workload Management - Dynamic discovery of available resources
- Security – Authentication, authorization and encryption

Grid Standards

The formulation and compliance of standards is extremely important in a grid scenario given that grid comprises heterogeneous resources. Only common standards can provide integration and interoperability in a virtualized environment.

This is the basis for Open Grid Services Architecture (OGSA), specified by the Open Grid Services Infrastructure working group of the Global Grid Forum (GGF), which seeks to:



- Define open, published interfaces. OGSA is an open standard managed by the GGF standards body.
- Manage resources across distributed heterogeneous platforms.
- Deliver seamless Quality of Service (QoS) for robust, behind-the-scenes services such as authorization, access control, and delegation.
- Enable intelligent self-regulation and autonomic management of participating resources.
- Encourage industry-standard integration technologies like Web services

Turbo-charging Monte Carlo Simulations with Grid Computing: an example

Monte Carlo simulation is an important tool in risk management, used commonly in the financial services industry for performing randomized scenario analyses. Monte Carlo analysis helps gain a better sense of the tail of the bell-shaped distribution curve. However, doing this requires a sufficiently large number of iterations – typically a minimum of 10,000 different re-evaluations to get adequate curve plotting points.

Monte Carlo analysis involves immense amounts of computation, which means that when run on a single system, takes considerable time to compute the result. Moreover, the application code involves similar computations done on different sets of data (the number of data sets can run into millions).

Since the data sets are different and computation on these data sets common, we can compute sub-results for the entire computation on different systems (called compute nodes) and finally collate these sub-results to arrive at the final result. This requires the same application to run in parallel on different systems but work on different sets of data. Now, the same application can compute results faster, as each compute node employed for computation will do a part of the work, and finally the entire result will be collated.

Infosys has developed a framework using Open Grid Services Architecture and Globus toolkit to accelerate financial analytics like the Monte Carlo simulation.

The framework comprises:

- Grid Infrastructure Setup
- Grid Enablement Of Applications
- Application parallelization
- Grid performance tuning

Grid computing is no more confined to R&D laboratories. Its application is especially pragmatic in the current scenario where financial services companies sit on a pool of underutilized resources that represent millions of dollars of idle investment. Grid computing provides a viable, open standards-driven option to harness unutilized potential to achieve operational efficiency by means of a virtualized, flexible, responsive and resilient infrastructure.

About the Author

Pankaj Misra is a Group Solutions Manager (Technology) in the Banking and Capital Markets Solutions practice at Infosys. His responsibilities include solution architecture design and technology.