

# G R I D

**I** often compare grid computing to an electrical power network in which multiple power-generation and storage facilities create a pool of electricity that consumers can access as necessary. Consumers don't care where the power comes from or how the grid is constructed. They just flip a switch or plug in an appliance, and only take notice of the power if it's not there when they need it.

This same concept can be applied to grid computing. Grid computing involves coordinating and sharing IT resources—such as servers, applications, storage devices and network resources—either across a network or within the context of a single data center.

In simplest terms, grid computing is the pooling of IT resources into a single set of shared services for all enterprise computing needs. The grid infrastructure includes management utilities that continually analyze demand for resources and adjust supply accordingly. The vision for grid computing is not worrying about where your data resides or which computer processes your request. You request information or computing power and have it delivered—as much as you want, whenever you want it.



# &BEARIt

Modernizing the  
Computing  
Infrastructure to  
Improve Productivity  
and Reduce Cost

By David Baum, Oracle



## Core Tenets of Grid Computing

Two key factors distinguish grid computing from other styles of computing: virtualization and provisioning. The theory behind virtualization is to pool computing resources such as processors and storage for on-demand use instead of calling on dedicated systems for individual applications. Not only does this boost efficiency, it also makes it easier to adapt the infrastructure to meet new requirements.

Provisioning determines how to meet the specific needs of each consumer while optimizing the performance of the system as a whole.

Virtualization and provisioning of infrastructure resources involve pooling and then allocating resources to the appropriate consumers based on policies. For example, one policy might be to dedicate enough processing power to a Web server so that it can always provide sub-second response time.

Treating infrastructure resources as a single pool and allocating resources on demand

saves money by eliminating underutilized capacity and redundant capabilities. At most companies today, server resources are only used between 20 percent and 40 percent of capacity. That's because systems are dimensioned to handle peak loads; dedicated servers are usually installed for each application and redundant systems are configured for testing, development, and backup purposes. As a result, systems keep getting bigger and system management more complicated.

Grids change this scenario, allowing enterprises to scale information systems incrementally, minimizing capital expenditures by only adding nodes when they are needed. Business applications can automatically harness the power of additional nodes as they are brought into the grid. Just as the local power company doesn't allocate electrical capacity for an individual house or business—rather, it creates distribution plans for a large set of consumers—grids are designed to service the aggregate

demand. Capacity planning is based on historical data revealing aggregate usage across hundreds of customers.

Grids also improve reliability. If a server malfunctions, processing can continue on the remaining servers in the grid, ensuring that data remains accessible and applications function without interruption. Spreading computing capacity among many different computers and spreading storage capacity across multiple disks and disk groups removes single points of failure so that if any individual component fails, the system as a whole remains available.

## History of Grid

The basic concepts of grid computing extend back to the early '70s when networks first linked computers and the idea of distributed computing was born. In many ways, grid computing represents the latest evolution of several other computing paradigms, including distributed computing, peer-to-peer networking, and virtualization. IDC calls grid computing the fifth generation of computing, after client-server and multi-tier, as shown in Table 1.

IDC analyst **Dan Kusnetzky** believes the adoption of grid computing as a platform for business-oriented applications will follow the same pattern as the previous four generations, beginning with applications that are relatively straightforward to segment into components or functions that can be distributed. As both networking technology and distributed computing software improve, this approach will be applied to more applications.<sup>1</sup>

Definitions of grid computing vary from company to company and analyst to analyst. Research firm **Gartner, Inc.**, defines grid computing as a way to solve computing tasks using resources that are shared by more than one owner and coordinated to solve more than one problem. According to Gartner analyst **Carl Claunch**, the phenomenon began in the scientific community as a way for universities and laboratories to share resources to tackle large computational problems—similar to what they used to do with supercomputers. For example, **NASA** might collaborate with **Lawrence Livermore Laboratory** to create a computing grid that is larger and more powerful than either organization could create on its own.

As the Internet became faster and virtual pri-

Managing hardware and software resources holistically reduces the cost of labor and the opportunity for human error while making it easier to add or remove processing capacity on demand.

Table 1

Computer Generation	Characteristics
First (Host-based Computing)	<ul style="list-style-type: none"> <li>■ Dumb terminal</li> <li>■ Single server</li> <li>■ Monolithic applications</li> </ul>
Second (Remote Access)	<ul style="list-style-type: none"> <li>■ Single client supporting only terminal emulation functions</li> <li>■ Single server</li> </ul>
Third (Client/Server)	<ul style="list-style-type: none"> <li>■ Single client supporting rules processing as well as user interface</li> <li>■ Up to two servers</li> </ul>
Fourth (Multi-tier)	<ul style="list-style-type: none"> <li>■ Single client supporting rules processing as well as user interface</li> <li>■ More than two server tiers</li> </ul>
Fifth (Grid Computing)	<ul style="list-style-type: none"> <li>■ Virtual environment where all systems are considered a pool of resources</li> <li>■ N-tier</li> <li>■ Service-oriented architectures</li> </ul>
<small>TABLE 1, FIVE GENERATIONS OF DISTRIBUTED COMPUTING. (IDC, 2004)2</small>	

vate network (VPN) technologies made communication more secure, grid architects realized that the computers in a grid could be located almost anywhere. The Search for Extraterrestrial Intelligence (SETI) grid is a good example. As part of this scientific experiment, volunteers provision idle processing cycles on their PCs to run a free program that downloads and analyzes radio telescope data. These Internet-connected computers apply the power of grid computing to analyze massive amounts of data from researchers all over the world. Claunch says we also see examples of this type of shared grid in the investment banking industry, when market researchers borrow capacity on traders’ workstations to study complex market scenarios.

One type of grid solution is focused on clus-

tering databases and applications within a data center. Claunch calls this type of grid computing “scale-out clustering.” Dividing databases among multiple machines yields tremendous up-time advantages and simplifies capacity planning.

In most cases, implementing grid computing does not require a massive paradigm shift. Most businesses can adopt grid technologies with minimal investment. Of course, not all applications and workloads are appropriate for database clustering. According to Claunch, it all depends on the type of data and how it is used. If the pattern for database access is well defined and known in advance—such as is the case with many transaction processing applications—then it is generally a good fit for grid computing as part of a database cluster. Since these applications consist of >>

1 IDC, 2004, “ORACLE 10G: PUTTING GRIDS TO WORK”  
2 IDC, *OP. CIT.*



many discrete transactions, it's easy to add additional computing resources to handle a growing load.

This type of distributed processing scenario is boosting adoption of Web services—a complementary set of standards for implementing distributed information systems as part of a service oriented architecture (SOA). Today's software developers are adopting these standards to make it easier for applications to request information from each other, even when those applications reside on different types of platforms or are separated by wide-area networks.

“In some cases, SOA and grid go hand in hand,” confirms Claunch.

“If a large process can be designed as a bunch of discrete services, then a service-oriented architecture is ideal. But scaling out applications and dividing them into discrete services are not necessarily the same thing.”

Claunch says standards bodies such as the **Global Grid Forum**, **W3C**, and **OASIS** have made it easier to create Web services for grid environments, resolving some of the conflicts between these two independent but coincidental trends—SOA and grid.

When you have a lot of computers, the ability to manage on an exception basis—to look across all of them and immediately see if there's a problem—is incredibly powerful.

## Holistic Management Practices

Using grid management tools, IT professionals can group multiple hardware nodes, databases, application servers, and other targets into logical entities—making it easier to support distinct segments of the grid for individual customers. Managing hardware and software resources holistically reduces the cost of labor and the opportunity for human error while making it easier to add or remove processing capacity on demand.

Grids are all about standardization, as companies deploy small individual hardware components, such as blade servers and low-cost storage resources. This enables incremental scaling and reduces the cost of each individual component. Capacity upgrades are much easier

and faster, because, in most cases, one or more nodes with similar or identical configurations can be added to an existing cluster, instead of using completely new nodes to upgrade information systems.

For database applications, customers commonly rely on a clustered database architecture to gain fault tolerance by spreading processing over multiple nodes. Since the physical nodes run independently, the failure of one or more components typically does not affect the other nodes in the cluster. This type of architecture also allows a group of nodes to be taken off-line for maintenance while the rest of the cluster remains online.

By executing jobs, enforcing standard policies, monitoring performance, and automating tasks across a group of targets instead of on many systems individually, we can manage and scale grids smoothly. Thanks to advanced management features, the existence of many small computers in a grid infrastructure does not increase complexity.

## Demonstrating Savings from Grid Computing

As part of a recent ROI study, **Mainstay Partners** evaluated enterprise grid technology platforms currently in use at seven participating companies.<sup>3</sup> Their grids were being used for a variety of applications, including enterprise resource planning (ERP), decision support, customer relationship management (CRM), and supply chain management (SCM). The study revealed significant savings, with an average return on investment (ROI) of 150 percent and an average internal rate of return (IRR) of 43 percent.

Most of the savings were attributed to moving from large symmetric multiprocessor (SMP) servers to lower cost commodity server clusters. “Scaling out versus the old paradigm of scaling up yielded an average 43 percent savings in hardware cost avoidance,” notes **Amir Hartman**, founder and managing director at Mainstay Partners. “These customers are also experiencing greater performance and availability as they distribute their computing activity across multiple nodes, which increases computation power and throughput.”

Many of the customers in Mainstay Partners’

3 MAINSTAY PARTNERS, “AGGREGATE ROI STUDY RESULTS: MAKING ENTERPRISE GRID COMPUTING A REALITY WITH ORACLE IOG SOFTWARE”

study demonstrated a reduction in labor costs and better availability as they moved from a single-node or dual-node environment to a multi-node environment. Like most IT departments making a move to grid computing, these companies are focused on three basic tasks:

- Consolidation of hardware, applications, and information shared among one or more data centers to create pools of resources
- Standardization of servers, storage devices, and operating systems—anchored by common infrastructure services such as provisioning, identity management, and Web services
- Automation of day-to-day management tasks, enabling a single administrator to manage multiple servers in clusters via a comprehensive management console.

## Standardizing and Simplifying the Infrastructure

Before adopting grid components and concepts, most of the companies profiled by Mainstay Partners depended on mainframes or large-scale SMP computers running proprietary operating systems. Without exception, these older environments were more expensive, in part because the hardware did not scale efficiently, requiring significant upfront investments in excess capacity to allow for future growth. By contrast, the enterprise grids they rely on today show more flexibility, both in balancing current workloads and in their ability to scale incrementally and cost effectively. These customers also report substantial productivity savings, sometimes amounting to millions of dollars per year by taking advantage of centralized systems management practices bolstered by labor-saving management tools.<sup>4</sup>

Grid control is all about managing a large number of systems as if it were a single system. When you have a lot of computers, the ability to manage on an exception basis—to look across all of them and immediately see if there's a problem—is incredibly powerful.

## A Vision for Grid's Future

As today's business leaders search for new ways to take advantage of these powerful computing strategies, they are envisioning a raft of creative

uses for grid technology. Gartner's Carl Claunch believes these innovative ideas will boost acceptance for grid technology on a broad scale.

For example, if a pharmaceutical company discovers it can bring a product to market one year earlier by using a grid to discover candidate drugs more quickly, then they see the payback for deploying a grid. "Complex business intelligence is another good example—such as using the computing power of a grid to spot purchasing trends or detect fraud correlations," Claunch says. "If a grid enables you to more quickly track crime or turn products on a retail shelf, it becomes very easy to justify the investment. Every industry has similar opportunities, but few people have conceived of the business advantages. The difficult technical problems have been solved. Now it's just a question of using your imagination."

## Grid Computing at Oracle's Austin Data Center

Oracle depends on grid computing to improve availability and simplify management tasks at its **Austin Data Center** (ADC) in Austin, Texas. Oracle customers house their applications and data here as part of the Oracle On Demand program, in which Oracle hosts and manages key software applications on behalf of their clients. The Oracle On Demand Grid also hosts many of Oracle's internal >>

## Grid Deployment

As part of a large outsourcing deal, EDS manages a storage grid for a global bank based in the Netherlands, which includes about 500 Oracle databases. According to Roland Rosenbloom, DBA team leader at EDS, software known as Oracle Enterprise Manager 10g Grid Control makes it easy to track patches, install new patches, and verify system configuration metrics—dramatically simplifying the process of managing grid environments. Database administrators can use a single console to view the entire technology stack—including the e-business applications, application servers, and databases—and take advantage of a service dashboard to communicate system issues to business stakeholders at the bank.

Rosenbloom contends that these grid management tools not only help EDS monitor existing systems and components, but add additional capacity as necessary, a phenomenon which he refers to as agile computing. "The provisioning tools give us the opportunity to install new nodes in the grid very fast and very easily."

<sup>4</sup> MAINSTAY PARTNERS, *OP.CIT.*



production applications. For example, Oracle's Application Demo Systems (ADS) is used by its global sales organization to demonstrate Oracle products to prospects, customers, and partners. ADS runs the entire Oracle E-Business Suite—some 180 modules—in a single instance, with 450 copies of the environment maintained and supported for the Oracle sales force. ADS hardware includes 300 **Dell PE2650** servers running **Linux**, with dual **P4 Xeon** processors and 6 GB of memory and 75 terabytes of storage.

Additionally, Oracle runs its own business on Oracle E-Business Suite applications at the ADC to provide application services to more than 50,000 Oracle employees worldwide. The database consists of Oracle Database 10g and Real Application Clusters (three nodes) running on 3 Sun F12K machines with 104 GB of RAM and more than 3 TB of storage; separately, the middle tier includes 21 Dell 2650 machines and 6 GB of RAM running Linux. New development activities, **Global Education Services**, and demo software are also housed at the ADC, with multiple grids to serve these different groups.

"The Austin Data Center is what a power station of the future will look like," explains **Benny Souder**, vice-president of distributed database development for Oracle. "Not only is it massive; it's incredibly precise. Every computing resource is cabled, labeled, stacked, configured, and deployed in exactly the same way, over and over again. This standardization allows ADC administrators to quickly isolate and troubleshoot problems, eliminating downtime and ensuring reliable computing services."

Oracle Real Application Cluster (RAC) technology is the backbone for its grid architecture in the Austin Data Center, creating a scalable environment in which nodes can be added or removed as necessary. This computing environment makes it easy to allocate capacity among many different applications and customers. When a customer initiates a new application or service, Oracle provisions a "slice" of the grid for that customer's needs. As the implementation progresses and the applications are moved into production mode, system administrators establish grid control rules that allow the environment to dynamically respond to the need for additional processors or storage. Rather than attempting to size each customer

uniquely, they can measure the overall utilization of grid elements.

Gartner's Carl Claunch says this type of technology provides many of the advantages associated with expensive, high-end computing environments at a relatively low cost. "Even when you have a thousand servers running as one, Oracle has done a really good job of shielding administrators from that complexity with its grid control software for setting up and managing these clusters," he notes.

Claunch notes that technologies such as Oracle Automatic Storage Management (ASM) make it easier to add and manage storage devices to a cluster. Using ASM, data can flow automatically on to additional resources without stopping production. This allows administrators to manage data as a single entity even if multiple storage devices are involved—just as they can do with applications on the computers.

"One of our requirements is one-hour time to recovery (TTR), 90 percent of the time," explains **Chris Pohto**, senior director at the Austin Data Center. "Utilizing the grid, we always have excess capacity provisioned, with at least 30 systems and several terabytes worth of storage just for this purpose. If a server fails, our first task is service restoration—fixing the box is secondary. So we simply move the affected instance to a free server, and the corresponding systems are back up in 15 or 20 minutes. The infrastructure can be scaled simply by adding more servers to the configuration."

With an intensive focus on energy usage and mechanical efficiency, the Austin Data Center has not experienced any electrical or mechanical infrastructure outages since it opened in 2002. In the unlikely event of a primary site failure, a backup site can pick up the load in less than two hours, with all data and applications synchronized.

"We're proving that you can run very fundamental business operations on an outsourced basis," says Pohto. "The grid is designed to make it easy to allocate capacity on demand. It allows us to run our customer environments very securely, very reliably, and with very high availability—often for less than these customers can run these environments themselves." [S]

About the Author: David Baum is publishing editor for Oracle.

## Key Components of Grid Computing

### Infrastructure

Infrastructure resources include storage devices, processors, memory, and networks, as well as software designed to manage this hardware, such as databases, storage management utilities, system management programs, application servers, and operating systems.

### Applications

Grid applications consist of business logic and process flows—either from packaged applications or custom applications. “Virtualizing” these application resources enables grid administrators to provide information and resources to consumers as individual components or services. Some companies use service oriented architecture (SOA) to combine these services into powerful business flows.

### Information

Information resources include all data in the enterprise and all metadata required to make that data meaningful. Just as grids exploit the power of the network to allow multiple servers or storage devices to be combined toward a single task—and to combine applications into composite business processes—grids combine various information resources to exploit inherent relationships among various types of data, both structured and unstructured.