2

WHY ARE BUSINESSES LOOKING AT GRID COMPUTING?

Corporations today are looking at and investing in grid computing not because it is a "cool" technology but rather because it answers core business needs and stringent financial requirements. It also offers a high-performance compute infrastructure at low cost. The technology combines commodity, throwaway hardware with ever-increasing network bandwidths, and self-administration software, to promote

- Significantly lower operational costs compared to those of today's data centers
- Significant return on investment and return on asset

Grid computing is no accident, and its future is very predictable. History provides a clear view of its adoption today and its path in the future. It offers a practical solution to fundamental requirements ranging from operations to business development, to corporate fiscal pressures.

HISTORY REPEATS ITSELF

History repeats itself twice. Corporations are looking at grid computing today for the same reasons that originally prompted the evolution of this technology in the first place. The future of grid computing is predictable; the same engineering principle that has driven the evolution of the telecommunications industry will evolve computing into a utility service.

Distributed Data Management for Grid Computing, by Michael Di Stefano

Early Needs

The 1990s were an exciting time to be in the business of the computer technology and information technology fields. The excitement surrounding the Internet and the possibilities that opened up beyond it seemed endless. Some business ideas were well founded, some not; but the number of technologies that quickly sprang up to support the new business models was staggering. The euphoria within the investment community to fund the exploration of both business and technology seemed as endless as the ideas that it financed.

During that same time period, universities, typically strapped for cash, needed to support their own business of research, which relied on computers to perform increasingly complicated, highly computational tasks, but lacked the budget or the unlimited venture capital (VC) funding that was afforded to the private sector. Universities had to figure out a way to support their research business with modest budgets. Their solution was to leverage the brain trusts of professors and students alike to create a method of networking inexpensive machines, so they acted as one large supercomputer: grid computing.

With few exceptions, commercial industry—fueled by limitless money and hardware—paid little attention to the developments in grid technology. This is not the situation today; the burst of the Internet bubble brought an abrupt halt to the days of free spending and the universities; grid computing projects are today laying the foundation for the next round of technology spending in corporate America. Perhaps the people in business today once attended those universities and participated in the creation of a powerful computer platform from inexpensive machines. Perhaps they recognized the parallels of the business need and financial drivers of universities in the 1990s, with those IT organizations in corporate America's face today. The business/financial environment of the university in the 1990s was very similar to that of today's corporate America. One reason why corporate America is looking at grid computing today is that the students who were involved in grid research in universities in the 1990s are now in the workforce, seeing the similarities and thus serving as an influential voice in pushing grid technology into corporations.

The converging forces of business drivers, downward financial pressures, world events, and a mature technology are ushering in a disruptive force that will change the fundamental way computing is done and create new business opportunities that otherwise would not exist (see Figure 2.1). Had it not been for the burst of the technology bubble in 2000, it would be safe to say that the wide adoption of grid computing that we are experiencing today would not be occurring.

We are now going to look at the business drivers from the prospective of the financial controller, the business manager, and the IT department, and examine how grid computing is uniquely positioned to address their disparate needs.

Artists and Engineers

Grid computing is the beginning of the shift of computing control out of the hands of the artist and into the hands of the engineer. Today, compute environments and

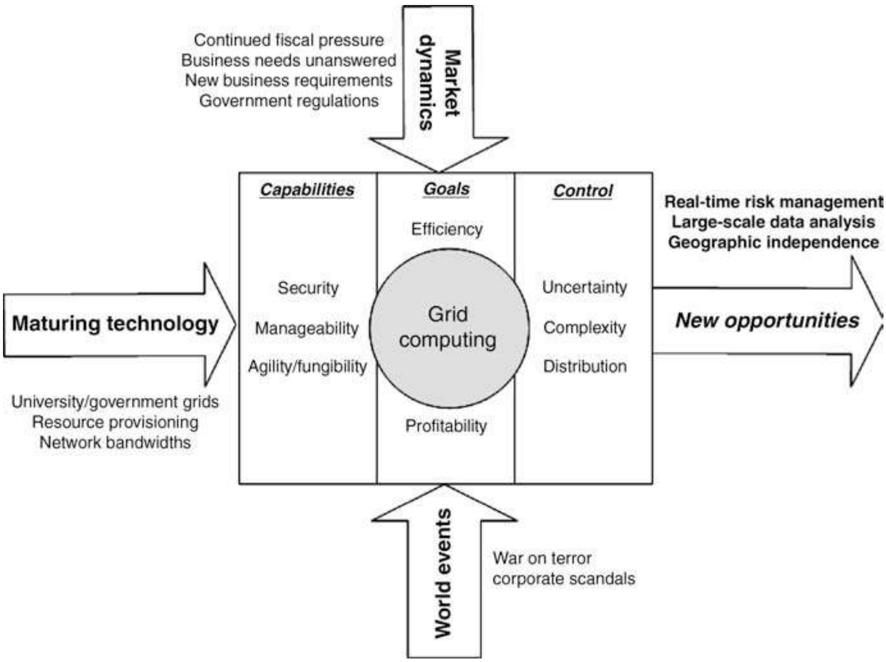

Figure 2.1. External forces, grid provisions, and new opportunities.

solutions are designed, integrated, developed, and operated by highly skilled individuals, the "artists." Grid computing opens a path to leverage the tried-and-true engineering and economic principles of utility services, meeting supply and demand curves of the customer. Thus, into the hands of the engineer.

Service-oriented network architecture (SONA) will be mentioned more than once in our discussions. SONA applies a combination of virtualization and orchestration to planetary-scale, distributed middleware. It describes the fundamental paradigm shift away from the client/server computing that the grid provides.

The same laws and principles that have enabled the information age will apply to the paradigm shift of grid computing, the proliferation of the network (see Figure 2.2). We will stand on the shoulders of Claude Shannon, Norbert Weiner, John Holland, and others and apply the all-too-familiar laws of Moore, Metcalf, and Amdahl to usher in the age of customer-centric information, content, and transaction standards of SONA.

It is the application of proven engineering techniques and methods that successfully moved a direct-wired telephone system of the early 1900s to the communication network utility that it is today. The same approach will change computing from a siloed data center to a grid utility that meets the economic principles of a free-market economy of supply and demand, and the reduction service of volatility.

The goal is to create a computer utility service that can be run and managed like a factory, with controlled costs, and the ability to increase output and change the

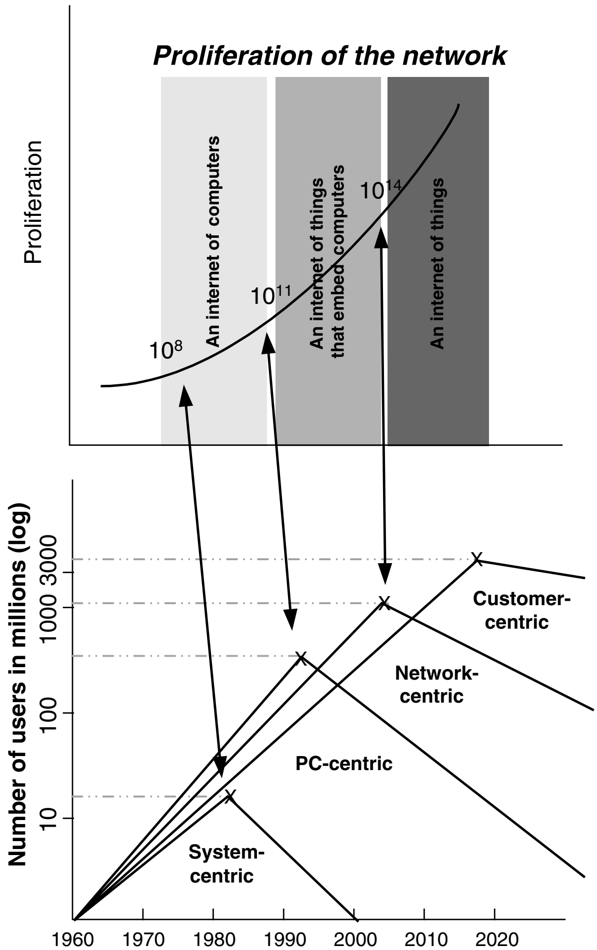

Figure 2.2. Proliferation of the network.[1]

production line as demand requires. This allows for better utilization of physical resources, which will drive down the operating costs.

The building blocks to achieve this start with the management of the physical resource for distribution of task—the compute grid—and must encompass:

- Data management techniques for the efficient movement of data
- Collection and use of metered data
- Application of feedback control logic, with metered data in, commands out
- The ability to provision your hardware quickly and efficiently
- Efficient administration without the need of an army of administrators.

The good news is that all these technologies are converging. They are not bleeding-edge; they demonstrate immediate return on investment (ROI) and, within a reasonably short amount of time (3–5 years), will yield significant cost savings for the organization.

THE WHYS AND WHEREFORES OF GRID COMPUTING

Recent events provide a logical path culminating in the emergence of grid computing. Starting at the burst of the technology bubble, there are financial pressures to control costs and unanswered business demands to cope with the changing economy, causing stress on IT personnel to manage both. At the same time, various technologies have been quietly maturing, each springing from different needs; for example, grid technology for low-cost, high-performance computing, self-provisioning software for operational management, and infiniband and other high-performance networking technology. These forces are converging, like the "perfect storm," to create a fundamental change in how computing and compute services are developed, managed, delivered, and paid for.

Financial Factors

Corporate CFOs have, in the years since the technology bubble burst, endured the burden of keeping their companies financially viable in the most difficult of business environments. Like the blade of a double-edged sword, changing business models demand new support from information technology; the other side of the blade is represented by changes in revenue streams that continue to squeeze profit margins, thus requiring tight cost controls and reductions.

This has led to a fundamental shift in how IT projects are developed and maintained. The use of IT outsourcing for project development and operations—barely existent prior to the burst of the technology bubble—has become the rule of the day. Companies that survived have done well, restructuring their respective organizations in both IT and long-term operational cost reduction. Unfortunately, there is continued pressure to further reduce costs.

How does grid technology assist the CFO? Let us look at how projects are developed and maintained within organizations. There is development, QA (quality assurance), production, and sometimes a step between QA and production for preproduction staging. Each of the steps requires dedicated hardware and support personnel to keep the centers running. (True, the developers can maintain their own machines.) However, environments outside the development environment (QA, preproduction, production; see Figure 2.3) will each reside in a proper data center, requiring trained staff to administer the hardware, network, core services (databases, middleware, etc.) as well as the business applications that run on them. Each environment is not a shared facility but rather separate, siloed copies of each other, each forming a closed and controlled environment to ensure that the production systems behave in a well-known manner resulting from the rigorous

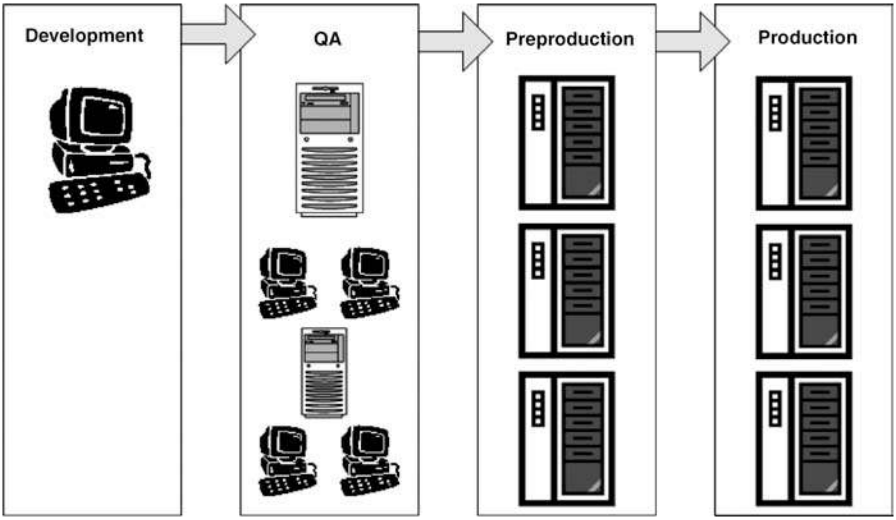

Figure 2.3. Dedicated and replicated environments.

testing done in both the QA and preproduction environments before being released to final production. This QA process and redundant physical setup is an expensive proposition in both physical resources (i.e. space, machines, and network) as well as in human resources.

The CFO looks to grid computing to reduce the long-term operational costs in a number of ways. First, each of these dedicated machines is not completely utilized. The rule of thumb is that the capacity for a machine should be twice what is required to handle the peak load, which may occur only at predictable, brief intervals throughout the day. In financial markets, these peaks occur at the market open and market close. Therefore, for most of the day, the dedicated and often expensive hardware is not completely utilized. When new projects come along, the purchase of additional machinery is necessary.

Grid technology allows us to run hardware closer to its full potential, typically at 70%, 80%, or even 90% utilization. This offers an organization a compute hardware environment that can support multiple business units according to their respective utilization requirements. No longer do their requirements need dedicated hardware for development, QA, and production. With grid technology there can exist a computer platform that is flexible and fungible to support the entire product development–production spectrum (see Figure 2.4).

Grid technology allows CFOs to buy inexpensive machines, and establishes a path to long-term operational cost reductions of an order of magnitude of today's data centers. The goal is to establish a compute utility service that obeys free-market economics, meets supply-and-demand (supply/demand) curves, and automates resource provisioning and commoditized hardware to the point where the value of a machine powered up and running in a data center is equivalent to its replacement sitting on a shelf in storage.

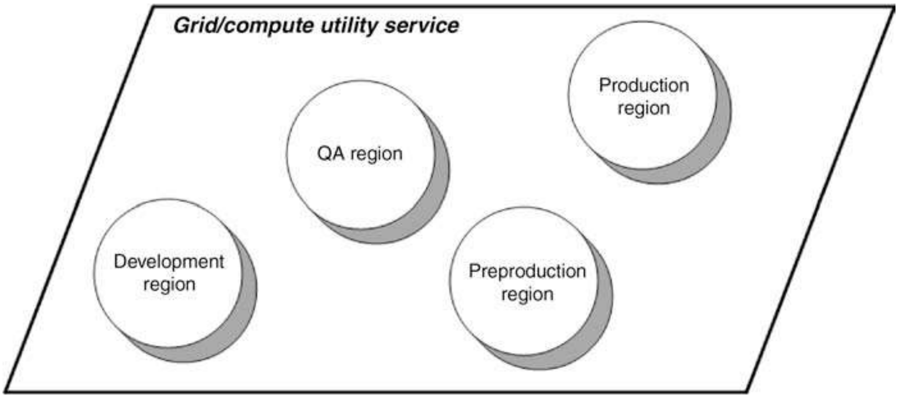

Figure 2.4. Grid: flexible and fungible compute environment.

Business Drivers

Business models have changed drastically since the technology bubble burst. The business units have had to be flexible and adaptive to quickly respond to external changes to continue to bring revenue into the organization. However, there is a lack of financial resources to spend on the technology needed to adapt to the new business models. As a result, the business's technology staff and the corporate IT organization supporting these business units have had to do more with less.

How does grid technology help? For the business manager, grid computing is a win–win scenario.

- *First Win*. As with their university counterparts, there are two opposing forces: business need and limited funds. Grid computing is a technology that offers computer power, flexibility, and high utilization at a low cost. A technological and financial model that supports the business and allows it to live within their respective budgets is available to the organization.
- *Second Win*. The use of grid technology creates new business opportunities. Grid technology enables new approaches to solving problems. This forces business heads, managers, and developers to start looking at things differently, thus creating new applications that would not exist without this technology. In Part III we will discuss practical business cases of grid technology and see how it ushers in a birth of new business and applications that otherwise would not be possible.

Technology's Role

We have discussed the business and financial pressures on the CFOs office to contain and reduce the ongoing operational costs, and the business's need to adapt quickly to changing business landscapes with limited technological resources. Now we will look at the people in the middle: the corporate IT organization.

IT finds itself caught between the pressures of demand from the business and supply in terms of limited resources from the CFO. IT is also under pressure to provide new services because simply maintaining existing services to the business is not an option. Grid technology is ideal for meeting both of these requirements and provides an inexpensive platform for the development and deployment of new applications, reduces operational costs, and offers higher resource utilization.

Grid computing represents a fundamental shift in how IT can support the business users when compared with a client/server environment. In a client/server data center, the business unit is financially responsible for the purchase of servers that are completely dedicated to that unit. The business is secure in its ability to support itself in peak environments but must live with additional hardware costs if demand increases and overpaying for resources should peak demand diminish. From the larger corporate perspective, the budget is used to purchase and operationally maintain hardware with low utilization levels.

The grid topology is a flexible infrastructure in which hardware can be used or shared by multiple groups depending on their usage demands. This implies cost structures different from that of a client/server data center. Grid technology is a decentralized flexible service that is owned and operated by IT, where business units' costs are shifted from a purchase/own/maintain-dedicated model to a "pay per use" model similar to that of a gas/electric utility service provider. Compute cycle usage must be metered and billed accordingly. The IT department must ensure that the resource is available when needed on a cost-effective basis, just like a gas/electric utility service provider.