3

SERVICE-ORIENTED ARCHITECTURE

In this chapter, we will discuss Service-Oriented Architecture (SOA) at a high level. Referenced within are various papers and articles that can go much deeper into the subject for interested readers. The objective is to explain what SOA is and how—along with business drivers and readily available technologies—it will lead to a paradigm shift in computing and compute service delivery. The technical reader can expect to be reacquainted with SOA and the technology behind it, while the business reader will learn about the economic forces driving SOA and how business (in terms of both current and future opportunities) will generate revenue streams sooner rather than later.

The importance to this shift of grid technologies and data grid in particular will be in supporting and enabling SOA. Data grid's role is less in how to implement SOA, but more in enabling SOA to deliver services to the customer when and where they are needed in a timely and cost-effective manner.

WHAT IS SERVICE-ORIENTED ARCHITECTURE (SOA)?

Service-Oriented Architecture is not a new concept in engineering and computer science; it touches our everyday lives in ways that we do not realize. Examples include television sets, DVD and CD players, the telephone, and electricity. In each case, the devices are interchangeable by make, model, and manufacturer, taking advantage of advancements in technology, but still offering the same respective service. In the software industry, SOA delivery paths have evolved. Examples of

Distributed Data Management for Grid Computing, by Michael Di Stefano

early attempts to deliver SOA are based on the evolution of middleware architectures such as Common Object Request Broker (CORBA) and Distributed Component Object Model (DCOM). Those of us who have delivered systems based on CORBA are all too used to such terms such as "CORBA Services," "locating the service," and "registering the service" and having the services available to anyone who needs them. Any "CORBA client" can "connect to" a "CORBA service" if it "knows" the service's Interface Definition Language (IDL).

Listen to the vocabulary, and you will quickly see why these early attempts at delivering SOA failed. The clients and services must be CORBA-based. Therefore systems built using any other middleware paradigm (of which there are many) could not leverage the services simply because of the tight coupling of connectivity, definition, and message (data). We will see that SOA must have "loose couplings" in order to be of value to a broad customer base.[2]

The concepts of loose coupling of services and the customer are very important in SOA. Systems built by leveraging the services of other systems in an enterprise have a "real dependency" on those services. However, to access the services, a number of human-made obstacles must be overcome. These are "artificial dependencies."[3] The CORBA example above is an example of an artificial dependency. While real dependencies cannot be eliminated, artificial ones can be reduced to a minimum. "Loose coupling" refers to the minimizing of artificial dependencies. Basic criteria are applied to achieve loose coupling of services and customers:

- Simple interfaces must be widely accepted by the community.
- Messaging (consisting of a schema) is self-describing in structure, both limiting vocabulary and enabling change for service versioning.
- Messaging does not have system behavior.
- Services need to be dynamically located by the customer.
- Services must be self-contained.

These criteria are purely technical in nature. To deliver a SOA, there are some nontechnical aspects of defining the service that need to be addressed Services must be "coarse on boundary." This implies that interfaces or boundaries of the service to the users of the service must describe the service from a business prospective, enough to describe what the service offers so that the consumer can make an educated decision. A service must provide a business function; it must make something happen. Services that simply move data are too fine-grained.[4]

Clients must be able to locate the service that best fits their requirements. There may be three billing services available, but the client must be able to reach all three and select the one that provides the best service to meet its needs. The latter is done via interface and messaging per the criteria listed above. Figure 3.1 shows how dynamic discovery can be achieved.[5]

There is a lot of buzz around Web Services and the tight correlation between Web Services and SOA. It is important to realize that Web Services is one way to

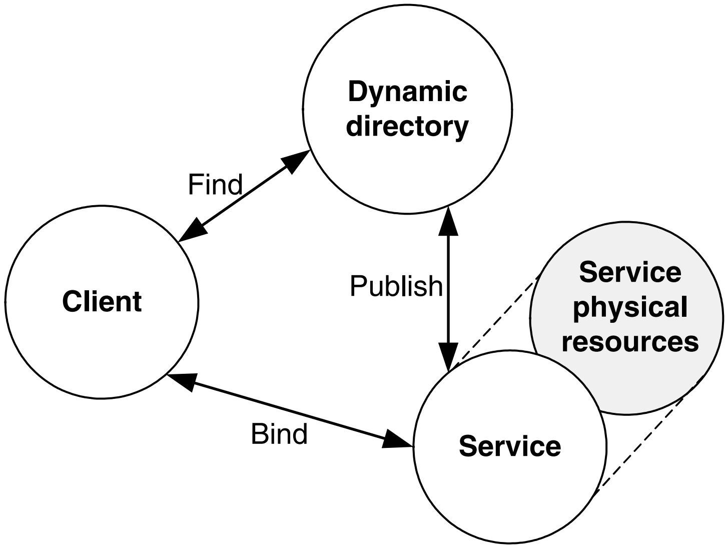

Figure 3.1. Dynamic location of services.

implement a SOA. Other technologies and methods can be used to implement SOA, provided they meet the criteria listed above. Web Services addresses many of the SOA criteria today. It is a technology that not only provides for loose coupling but also is in wide use by the industry. The correlation of the ability of Web Services to deliver SOA is

- *Interface*—HTTP, FTP, SMTP (Simple Mail Transfer Protocol)
- *Message*—XML (eXtensible Markup Language)

These enablers are in place and proven and give rebirth to SOA. It is the convergence of available technology in production and widespread acceptance that will quickly enable the delivery of services to a wide variety of customers.

DRIVING FORCES BEHIND SOA

As in the movie *Perfect Storm*, there are forces converging at the right time and place under the right conditions that are resulting in a fundamental paradigm shift in computing. As shown in Figure 3.2, these forces are market dynamics, maturing technology, and world events. Within each is its own miniature "perfect storm." It is interesting to see how these seemingly unrelated events and advances in technology are converging, opening the window for new opportunities for business and information technology.

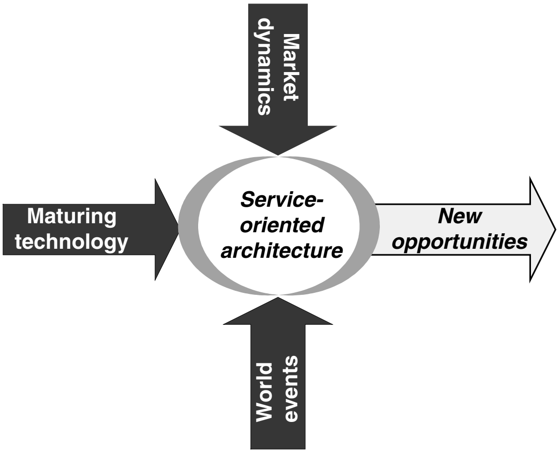

Figure 3.2. The SOA "perfect storm."

The opportunities emerging range from new business to improvements in information technology operations.

- New business
  - Real-time risk management
  - Large-scale data analysis
  - Utility services
- Operational improvements
  - Dramatic lower operational costs
  - Increase in compute utilization
  - Geographic independence

In-depth examples of these new opportunities are discussed in Part III of this book.

Maturing Technology

The advances in technology leading up to this point do not represent a single, revolutionary breakthrough. Rather, it is an evolution that is taking place in four areas of the IT world: Web Services, networking, distributed (grid) computing, and operational resource provisioning. Each has made advances in their respective areas that, on the surface, have no correlation to the others. However, advances in networking are the primary enabler for the rest.

Networking. Starting with the Ethernet, networking speeds have steadily increased since the late 1980s, 10 megabits per second (Mbits/s), giving way to 100 Mbits, giving way to 1 gigabit (Gbits). Enter Fiber Channel at 2 Gbits. Infiniband technology—first appearing as device interconnects for use in things such as storage area network (SAN) solutions—is now making its way into data center local-area networking with network speeds of $\geqslant 10$ Gbits/s.

Distributed Computing (Grid). Grid technology takes two forms: computational or compute grid, and data grids. Each has its origins in academia and research, where complex analytical methods operating over large data sets were becoming the norm. However, the norm was also the limited budget to purchase supercomputers to do the work. The result was utilization of existing idle compute resources on a network, either private or public (the Internet), and distributing and coordinating work, thus yielding a seemingly limitless compute backbone for little to no financial cost.

As grid computing gains favor in commercial industry, further advances are being made to continue to evolve both the compute and data grids. Early problems involved large static data sets. As commercial applications have come online (via SOA), the data sets are becoming more dynamic in nature, thus requiring an evolution in data grid and distributed data management in grid.

Resource Provisioning. Resource provisioning, or data center automation, addresses issues for lowering the cost of administration of data centers by automating the mundane, manual-labor tasks found in most data centers. Tasks such as software installation and machine configuration can be automated, thus reducing cost of administration, improving the consistency of server profiles, and increasing data center uptime or reliability. Spinoff benefits include further cost savings from leveraging idle systems by reprovisioning them on the basis of usage demand.

Web Services. The evolution of SOA and Web Services has been discussed. The benefits of SOA via Web Services are

- Improved return on investment (ROI) via
    - Lower infrastructure and operations costs
    - Development focused on business problems, not service delivery
- Reliability due to
    - Better quality assurance testing
    - Improved service maintainability
    - Improved service uptime
- Opening up business to a new and broader customer base

Business

Things always happen for a reason. But sometimes it can be difficult to see what those reasons are. For those who have lived through the technology and Internet boom of the late 1990s and seen the bubble burst, it was hard to understand the difficulties in the economic climate that occurred in the years that followed. Now that we have gotten past this, we can see that without the boom and bust of that bubble, the economic and business climate would not have existed to promote the elevated interest in grid computing and SOA that we are experiencing today.

If grid computing and SOA only provided a better and faster way to run applications, no one would be interested. The interest in grid computing and SOA is due to their collective financial benefits and their ability to address the climate that exists in business today, of tight fiscal spending and continued pursuit of ways to control costs and operational expenses. The added bonus of grid computing and SOA is that they both offer an improved way of running current business applications while offering an avenue for creating new business opportunities using a technology that is not bleeding-edge. None of these technologies is new. They have been explored and tested, and have been running in a production environment for many years. As such, the business—the CIOs, CTOs, and CFOs—do not view them as experimental. Rather, they see this as an application of an existing technology that can be quickly leveraged to put together IT infrastructures that not only support current business and promote new business but also reduce long-term operational costs.

It also becomes apparent to the business that widespread adoption of a SOA, implemented using Web Services and grid technology, fundamentally changes the way in which data centers are operated, moving them away from the dedicated silo-based structures that they are today to assume the appearance of a factory. As services are used by a growing number of customers—both internal or external—the price that can be charged for those services and the costs to the manufacturer begin to follow basic economic principles of supply and demand. It would not be economically feasible to deliver services given the current silo-based structure of the data center; it is not economically feasible to run a factory when your equipment is only 30% utilized. Grid technology addresses issues of delivery of service and economic costs by allowing data centers to run at efficiency/capacity rates of 50%, 60%, 70%, and higher using commoditized hardware components. Provisioning software allows the shift of physical resource to meet a service need when the demand now dictates. Below, we will discuss the business fundamentals of this paradigm shift in computing and tie them to the basic economic principles of supply and demand—in other words, market dynamics. In addition, we will see how this shift can leverage the same tried-and-true economic and engineering principles that govern utilities such as telephone companies to transform the silo-based data center into a compute utility service.

World Events

The third element in the "*perfect storm*" we described above is represented by the world events occurring during this same post-bubble-burst time period. We have seen a new war on terror, which is forcing an unprecedented level of communication and information sharing between the various law enforcement and intelligence agencies, and the judicial communities. It becoming apparent that the most important weapons we have in this war are information collected from a wide range of sources and the ability to analyze and correlate the data as quickly as possible. It is also part of the as-yet uneffected offense. How to foster this level of information sharing has become a driving factor in the advancement of grid computing, data grid, and SOA among the various government agencies fighting this war.

This same thirst for information sharing is being driven elsewhere by investigations relating to the seemingly endless series of corporate scandals afflicting American business. In this case, rather than tracking terrorists' movements or the diversion of illicit funds through money laundering schemes, law enforcement officials, and regulators are joining forces to share information on the timing of stockmarket trades, emails, and telephone calls.

In both cases, grid computing, data grid, and SOA have a role to play in achieving the level of interagency communications required to deal with the threat at hand.

ENTER BASIC SUPPLY–DEMAND ECONOMICS

In the movie *Trading Places*, an elaborate plot by Randolph and Mortimer Duke to corner the frozen orange juice market was foiled by Billy Ray Valentine and Louis Winthorpe III once they "acquired" the farm reports from Clarence Beeks. How did this happen?

1. The true farm reports stated a "good" or abundant supply of oranges.
2. Valentine and Winthorpe gave the Dukes a false report stating the opposite: a bad orange crop, thus a shortage of oranges.
3. The Dukes having the farm report in advance of the government announcing it to the public (just a few SEC rules were broken here alone), went on a buying spree to purchase as much of the available orange crop at a low price, knowing that once the shortage was announced, their orange crop holdings would be worth more that what they paid for them.
4. This caused a feeding frenzy that caused others to "follow" the Dukes' lead of buying the orange crop. The result of all this buying caused the price to rise.
5. Valentine and Winthorpe, knowing the content of the true farm report, sold orange crops at what was becoming an artificially high price due to all the Dukes (and others') buying. Technically, they were "selling short," selling something that they did not own, knowing that at some point in the future they would have to deliver the oranges they sold. They were not worried, for they knew that soon they would have all the orange crops they would need to meet these commitments.
6. Once the government announced an abundance of oranges, the value of the orange crops decreased; therefore the price started to go down as well.
7. As the price went down, Valentine and Winthorpe started buying back the same orange crops they had just sold but now at a much lower price. Buy low and sell high, and you make a profit.
8. The Dukes bought high and sold low, so they lost a lot of money (not to mention all the legal trouble they were in for insider trading).

Service-oriented economies follow the basic economic principles of supply and demand. Service-oriented architectures produce "business services" or product for anyone who desires them (consumers). The ability to supply a quality service to the customer will determine customer demand for the service.

In the example presented above, a supply of a product or service that falls short of demand forces the price up. This is what the Dukes were led to believe by the false report. So, no matter what price they bought the orange crops at, once the public knew of a supply shortage, the prices would rise even further, thus allowing them to sell at the higher price. But Valentine and Winthorpe knew that the opposite was true: an excess supply of oranges. In this case, oranges in excess of demand will force the fair market or equilibrium price down. The end result was that the Dukes lost money by buying high and selling low, while Valentine and Winthorpe made a profit by buying low and selling high.

Supply–Demand 101: Vocabulary

- *Desire* refers to people's willingness to own a good.
- *Demand* is the amount of a good that consumers are willing and able to buy at a given price.
- *Utility* is the satisfaction people get from *consuming* (using) a good or a service.
- *Supply* is the amount of a good that producers are willing and able to sell at a given price.

Figure 3.3 shows that "excess supply" forces the equilibrium price (*P**) down while "excess demand" forces the equilibrium price up. In a market economy, whether the product is oranges or billing systems, the ability of the producer to meet the market demand will determine the value of the product.

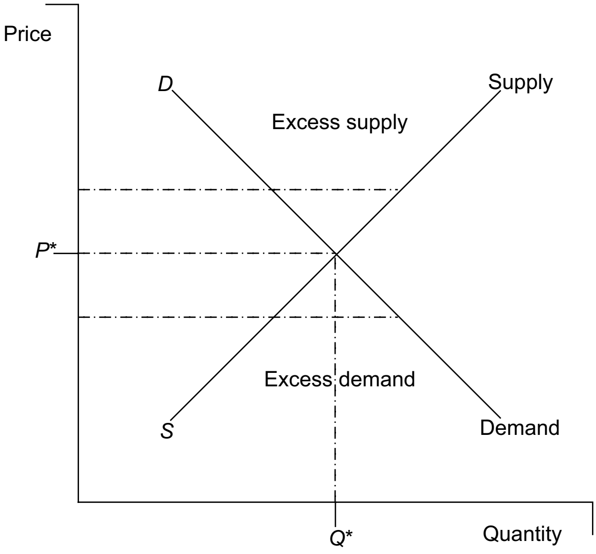

Figure 3.3. Basic supply–demand economics.

Service-oriented architectures, when successful, will package a product in such a fashion that renders it palatable or of a sufficiently high "utility" to encourage consumers to purchase the product rather that produce it themselves. As service offerings to customers increase, a true market economy will emerge that will be driven by the basic economic principles that guide a free-market economy.

As evidenced by Randolph and Mortimer Duke, the price of a product or service is determined in part by the ability to supply the product to the market. If the cost is too high, the size of the market is limited only to those who can afford to pay the higher price. There is little difference between a late winter frost destroying the orange crops, causing a supply shortage and a silo-based data center incapable of producing a quality service at a sufficient quantity and reasonable cost. Data centers need to run more efficiently to produce a quality product/service at a market cost indicated by the convergence of the supply/demand curves.

Distributed computing technology (compute grid, data grid, networking, provisioning, etc.) is an enabler that offers service-oriented architectures and shifts the operations of a data center from a custom "silo per product" offering to a factory producing quality product at affordable prices. The methods to do this are not new, for the telephone and electricity companies produce their respective products and distribute them across vast networks of devices to a broad consumer base. The quality of service is judged on ease of access, availability on demand, cost, and responsiveness to correct supply outages. Information technology can, via the evolutionary state of distributed computing, apply the tried and true principles of Shannon, Weiner, and others to transform the data center into a "compute utility service."

FUNDAMENTAL SHIFT IN COMPUTING

Technical communications in the latter half of the twentieth century were a function of computing innovation. Historically, communications relied on improvements in switching and resource allocation technologies. Switching technologies, in turn, have depended on the evolution of computing capabilities. This symbiotic equation, linking communications and computing, has become particularly important in the present environment of service-oriented architectures.

Today, a significant increase in computing power is driving a fundamental paradigm shift in technical communication.[6] The evolution of technical communication and its present-day drivers all point to a convergence point, a paradigm shift. This shift stands on the shoulders of history but also satisfies the elasticity, fungibility, granularity, and dependability needs of SOA. The shift is toward *Service-Oriented Network Architecture* (SONA). SONA is characterized as an overlay network, of Internet scale. It is architected to take advantage of virtualized hardware and policy-based dynamic resource allocation. It is multipurpose, and can be thought of as a learning system through its use of a continuous feedback loop for service improvements. The implementation of SONA is a nexus of grid computing and Web Services. We will discuss SONA and its relationship to grid computing and Web Services in Chapter 20 of this book.