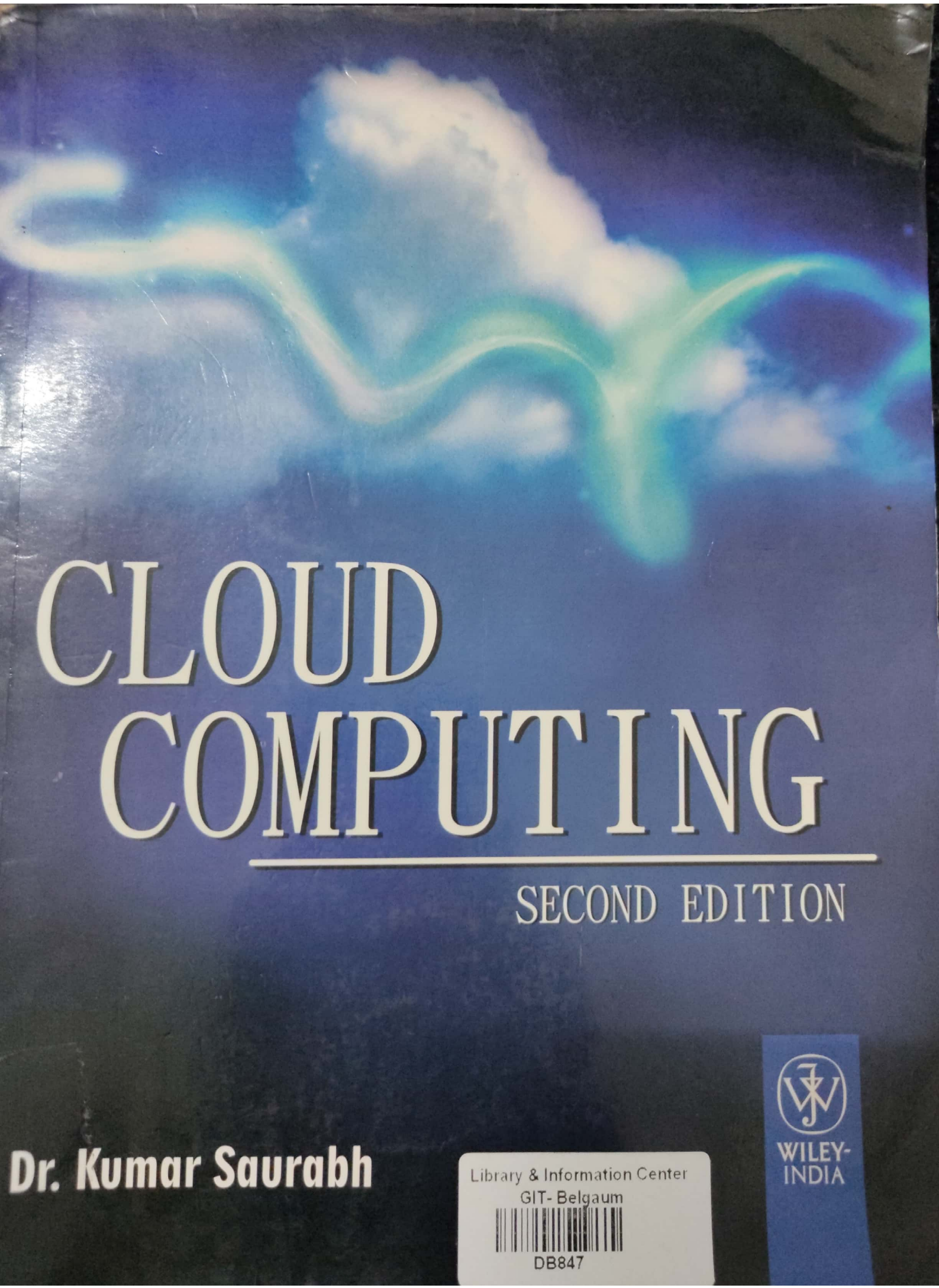
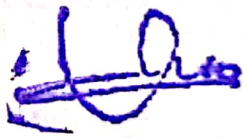


# CLOUD COMPUTING

SECOND EDITION



Dr. Kumar Saurabh

WILEY

# Contents

CHAPTER 1

# FIRST DRIVE

## 1.1 INTRODUCTION

One of the latest drift in small and medium businesses and enterprise-sized IT is the need for a significant transformation of the IT environment. Cloud computing provides a major shift in the way companies see the IT infrastructure. This technology is primarily driven by the Internet and requires rapid provisioning, high scalability, and virtualized environments. It provides the abstraction for the business and is handled by the actual owners of the infrastructure experts. In this demanding world, the raison d'être to adopt cloud computing over standard IT deployments is flexibility, stability, rapid provisioning, reliability, scalability, and green solutions. Cloud computing can trace its intellectual roots back to grid computing, but it is often confused as the outcome of grid computing advancements and research during recent period, and that is not totally true. Grid computing paves the path for the evolution of the cloud computing concept. While these may be examples of applications of cloud computing for IT infrastructure, they are not the only ingredients of it. So, before going into the details of cloud computing, let us have a cursory glance at grid computing that gives you an immense computing grid to tap into as you need it, and scale up and down as per the requirement.

Grid computing approach starts with the breaking of the silos by inserting an additional layer on each server included in the grid. The main function of this additional layer is to create logical servers that distribute over different physical servers the computational needs (job, tasks) required by the different applications they are virtually executing. In this way, it is possible to decouple the applications from the physical systems on which they were running, and at the same time, it is possible to dynamically increase or decrease the computational power of the logical servers as per application needs.

### *1.1.1 Grid Computing*

A grid is made up of a number of resources and layers with different levels of implementation (Figure 1.1). As said, there are different types of grid that are usually organized according to this taxonomy. Starting from the layer at the bottom – virtualization, which involves only physical resources – we may have then:

- **Information grids:** These are aimed to provide an efficient and simple access to data without worries about platforms, location, and performance.
- **Compute grids:** These exploit the processing power from a distributed collection of systems.
- **Services grids:** They provide scalability and reliability across different servers with the establishment of simulated instance of grid services.
- **A mix of them:** Each of these have specific sets of characteristics that are peculiar of the hybrid characteristics of compute and service grids.

Conceptually, we can imagine three layers, the lower being the physical one where we have the servers, storage devices, and the interconnecting network. In the second layer, we see the different operating systems, mapped one-to-one with the servers. The upper layer is the application one where we map different applications supporting the enterprise

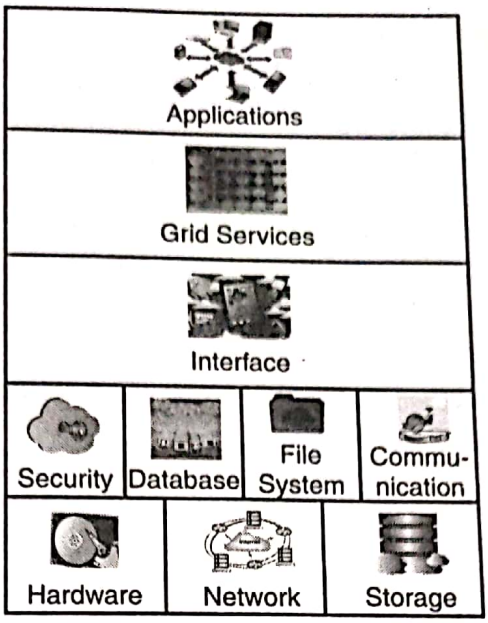



**FIGURE 1.1** Simple grid architecture.

Grid computing is an evolution of distributed computing that utilizes open standards to allow you to see independent and physically scattered computing resources as though they were a unique large virtual computer.

With these concepts in mind, we can consider a 'compute grid', where the grid's goal is to exploit the processing power from a distributed collection of systems. Main functionalities of a compute grid are to manage the resources' workload, apply utilization policies and security rules, schedule and execute parallel tasks across distributed resources, and provision (reserving, adding, removing) resources according to the scheduling needs. It is a special kind of compute grid where resources – typically distributed all over the world, but could also be within an enterprise – are used by the grid only when idle, which means provisioning and scheduling policies are very 'relaxed'.

Information grid provides transparent and efficient access to data independent of their location, type, and platform, and allows end-users secure and transparent access to any information source regardless of where it exists. It supports sharing of data for processing and large-scale collaboration, and provides logical views of data without having to understand where the data is located or whether it is replicated. It manages data cache or data replication automatically to get the most efficient and secure access.

Information is usually defined as 'meaningful data' from the perspective of the end-user. An information grid provides an abstraction over disparate and distributed information sources, such as a Database Management System (DBMS), flat files (for example, comma-separated files), structured files (for example, XML documents), or a Content Management System (CMS).

An information grid also has the ability to federate or integrate data and information from heterogeneous resources into a unified virtual repository. The whole idea is to present a single view of the information.

### 1.1.2 Grid – The Way to Cloud

The concept of cloud computing can trace its roots to grid computing that provides rapid provisioning of resources. It is not mandatory that grid computing should be in the cloud; actually it depends on the type of users, whether they are consumers or administrators. Grid computing requires software that can be divided and computed or serviced on a single or multiple systems. This creates a problem of non-functioning of the overall solution if one of the components fails because of the internal dependency. With the advent of Internet, computing crossed geographical boundaries and networks and has given us the chance to exploit services and computing globally over the Internet.

Both cloud and grid services provide scalability as a functionality. This is achieved through load balancing and high availability instances of the applications running either on variety of operating systems or a single one. Both services provide on-demand services for the instances, users, storage, networks, and data transferred at a particular time, and can be de-allocated when they are not required. These computing involve the multi-tasking environments available on single or multiple instances based on single or multiple servers.

Optimization is a grid type where the primary focus is optimization of underutilized IT resources in an organization. Grids require a different way of thinking about how to deliver IT datacenter services, and resistance to changing behaviour is always the toughest hurdle to overcome in technology adoption. Lack of industry standards is a barrier to widespread adoption, as clients perceive the risk of not-protected technology investment. Security will have to be proven over time to potential customers at a number of levels for grids to be considered for adoption in shared workload environments. The cost of computational power (both CPU and storage) continues to decline, which may erode part of the financial benefits of grids. To exploit grid advantages fully, physical resources across heterogeneous systems can be virtualized building a single resource image.

The following sub-sections will help us understand the benefits of grid computing when deployed for infrastructure management and extended to cloud computing arena (Figure 1.2). This is also discussed in detail later in the chapter.

#### *Storage/Data/Information*

- Provides logical views of data without having to understand where the data is located or whether it is replicated.

#### *System Management*

- Defines, controls, configures, and removes components and/or services (could be physical) on a grid using automated or physical methods.

#### *Metering, Billing, and SW Licensing*

- Provides tools to monitor and distribute the number of licenses while using licensed software.
- Provides metering and billing techniques, such as utility-like services, so that the owners of the resources made available are accurately compensated for providing the resources.

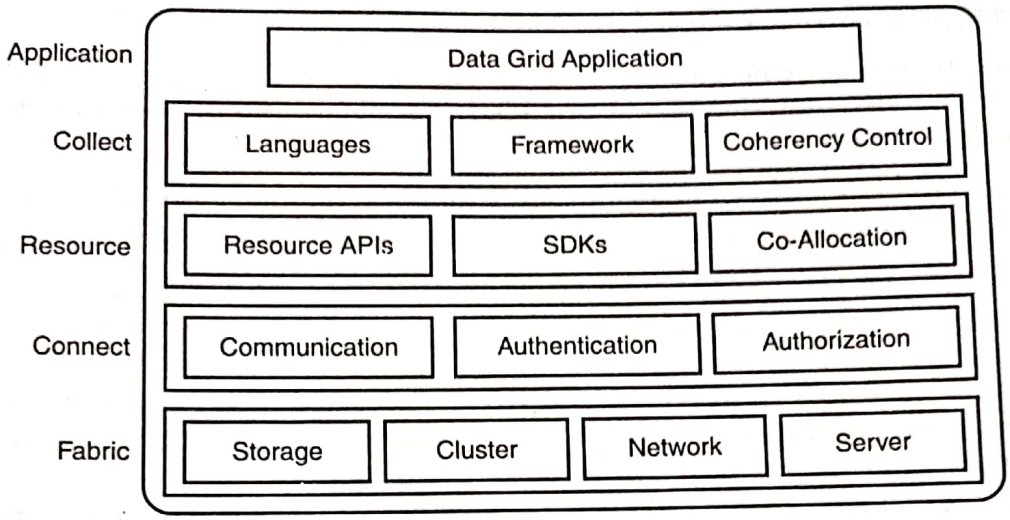



**FIGURE 1.2** Standard grid architecture.

*Security*

- **Authentication:** The grid has to 'be aware' of the identity of the users who interact with it.
- **Authorization:** The grid has to restrict access to its resources to the users who are eligible to access it.
- **Integrity:** Data exchanged among grid nodes should not be subject to tampering.

Differing grid solutions may hit differing stages, but majority of the grid marketplace is transitioning from the 'early adoption' to the 'early majority' phase. Over the past few years, the market has evolved from specialist customers – predominantly in the academic and research sectors – using grid to accelerate internal simulations to a stage where corporate users are starting to apply grid and virtualization in a meaningful way that delivers clear business benefit (risk and portfolio analysis, seismic applications, clash analysis, etc.).

Organizations are now starting to use grid and virtualization technologies to unleash idle computing capacity to accelerate critical business processes and to optimize and improve resiliency of their IT infrastructure.

## 1.2 ESSENTIALS

Cloud computing is a term that describes the means of delivering any and all Information Technology – from computing power to computing infrastructure, applications, business processes and personal collaboration – to end-users as a service wherever and whenever they need it.

The *Cloud* in cloud computing is the set of hardware, software, networks, storage, services, and interfaces that combine to deliver aspects of computing as a service. Shared resources, software, and information are provided to computers and other devices on demand. It allows people to do things they want to do on a computer without the need for them to buy and build an IT infrastructure or to understand the underlying technology.

### 1.2.1 Emerging Through Cloud

*Cloud computing is an emerging style of IT delivery in which applications, data, and IT resources are rapidly provisioned and provided as standardized offerings to users over the web in a flexible pricing model.*

*Cloud computing is a way of managing large numbers of highly virtualized resources such that, from a management perspective, they resemble a single large resource.*

*There is greater need for IT to help address business challenges and cloud computing can help you do all of these:*

- ***Doing more with less:*** Reduce capital expenditures and operational expenses.
- ***Higher quality services:*** Improve quality of services and deliver new services that help the business to grow and reduce costs.
- ***Reducing risk:*** Ensure the right levels of security and resiliency across all business data and processes.
- ***Breakthrough agility:*** Increase ability to quickly deliver new services to capitalize on opportunities while containing costs and managing risk.

*Cloud computing is the provision of dynamically scalable and often virtualized resources as a service over the Internet (public cloud) or intranet (private cloud).*

## 1.3 BENEFITS

As an emerging IT delivery model, cloud computing can significantly reduce IT costs and complexities. The buzz surrounding cloud is based mostly on a new kind of user experience – particularly in the consumer Web space – for search, social networking, and retail. From the consumer perspective, cloud computing is a means of acquiring services without needing to understand the underlying technology. Many of us use cloud delivery models everyday without knowing it when we share photos online, download music, or access bank accounts using our mobile phone.

From a technology perspective, cloud computing is loosely defined as a style of computing where dynamically scalable resources (such as CPU, storage, or bandwidth) are provided as a service over the Internet. The process is typically automated and takes minutes. Cloud computing can be considered as a massively scalable, self-service delivery model that lets you access processing, storage, networking and applications as services over the Internet. Enterprises adopt cloud models to improve employee productivity, deploy new products and services faster and reduce operating costs – starting with workloads that are ripe for this environment. These typically include development and test, virtual desktop, collaboration, and analytics.

## 1.4 WHY CLOUD?

A cloud typically contains a significant pool of resources, which could be reallocated to different purposes within short time frames, and allows the cloud owner to benefit significantly from

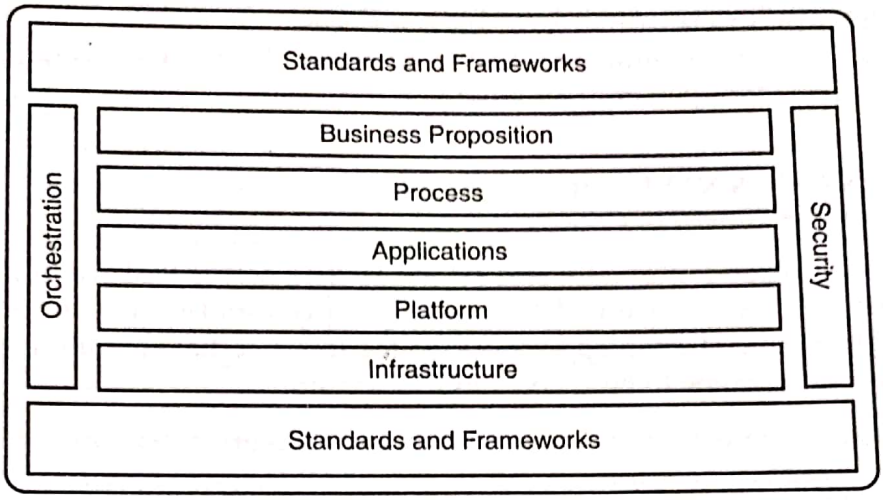



**FIGURE 1.3** Basic cloud computing model.

economies of scale as well as from statistical multiplexing (Figure 1.3). The entire process of requesting and receiving resources is typically automated, and is completed in minutes.

Cloud services today are delivered in a user-friendly manner and offered on an unprecedented scale. The payment model is pay-as-you-go and pay-for-what-you-use, eliminating the need for an up-front investment or a long-term contract. This presents a less disruptive business opportunity for businesses with spiky or unpredictable IT demands, as they are able to easily provision massive amounts of resources on a moment's notice and release them back into the cloud just as quickly.

There are different reasons for adopting the cloud:

- Massive, Web-scale abstracted infrastructure.
- Dynamic allocation, scaling, movement of applications.
- Pay per use.
- No long-term commitments.
- OS, application architecture independent.
- No hardware or software to install.

This results in business- and IT-aligned benefits:

- Accelerate innovation projects that can lead to new revenue.
- Make IT an enabler of, not a barrier to, rapid innovation.
- Provide an effective and creative service delivery model.
- Deliver services in a less costly and higher quality business model, while providing service access ubiquity.
- Create a sustainable competitive differentiation.
- Rapidly deploy applications over the Internet and leverage new technologies to deliver services when, where, and how your clients want them – before your competitors do.

- Lower IT barriers to launch new business services.
- Build and integrate modular services – in record time – by leveraging 'rentable' IT services capabilities, pay only for what you use.

## 1.5 BUSINESS AND IT PERSPECTIVE

Businesses are now looking internally and saying to themselves that we need to deliver this same level of end-user experience with our own IT for our end-users – employees, partners, and customers. Delivering IT-enabled services via the Internet that are built for the end-user to be in control is what has come to be called 'cloud computing'.

Cloud computing is an emerging consumption and delivery model that enables the provisioning of standardized business and computing services through a shared infrastructure, where the end-user is enabled to control the interaction in order to accomplish the business task.

Computing resources such as processing power, storage, databases, and messaging are no longer confined within the four walls of the enterprise. Instead, a tightly woven fabric of abstract – or virtual – resources are tapped into whenever they are needed. Essentially, everything needed from a computing resources standpoint is provisioned by the cloud – much like the electrical power grid we all tap into.

## 1.6 CLOUD AND VIRTUALIZATION

Virtualization has been around for 30 years. Yet, how many have really truly virtualized at all the layers of the stack? You really cannot expect cloud to produce what a cloud is expected to produce if it is not virtualized, standardized, and automated, because people expect scalable services.

In a cloud environment, people expect self-service, being able to get started very quickly, self-provisioning, or rapid provisioning. All of those things essentially demand that you do have these very important fundamentals in place.

The only way you are going to be able to get efficiency is by virtualizing, standardizing, and automating (Figure 1.4). And that's going to drive down costs and improve service. This is really a pretty simple equation and we are seeing organizations that are doing this achieve very real measurable business results. These results include:

- **Server/storage:**
  - IT resources from servers to storage, network, and applications are pooled and virtualized to help provide an implementation-independent, efficient infrastructure, with elastic scaling – environments that can scale up and down by large factors as demand changes.
- **Automation using:**
  - Self-service portal: Point and click access to IT resources.
  - Automated provisioning: Resources are provisioned on demand, helping to reduce IT resource setup and configuration cycle times.

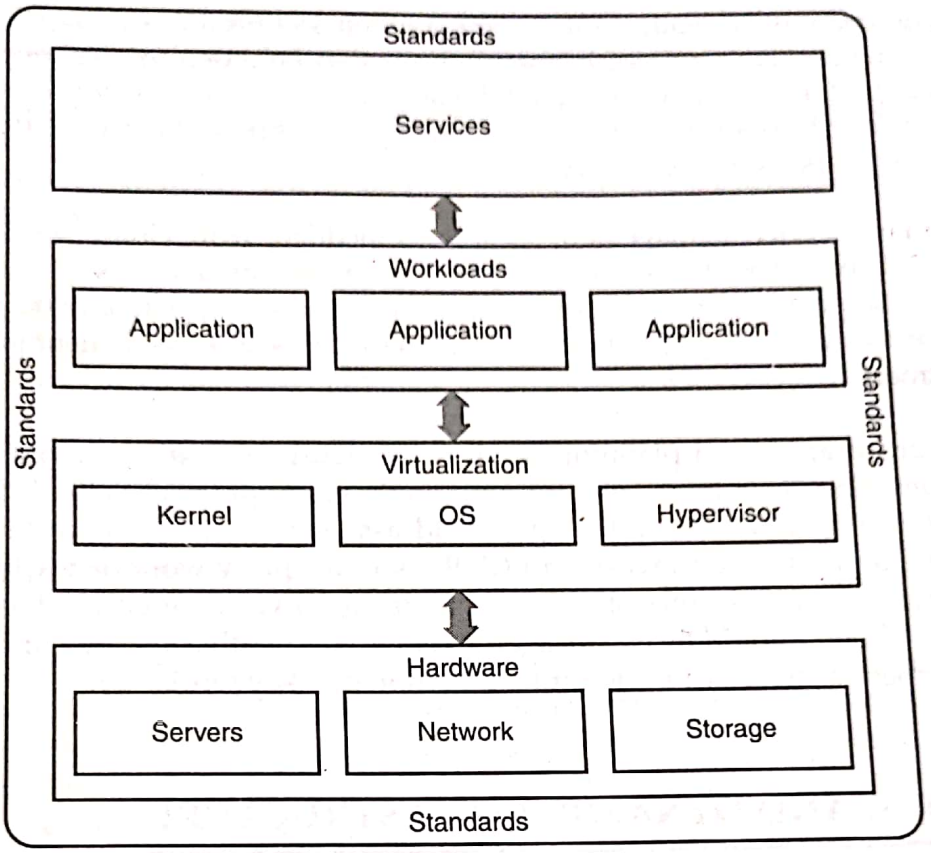



**FIGURE 1.4** Datacenter clouds.

- **Standardization through:**
  - Service catalogue ordering: Uniform offerings are readily available from a services catalogue on a metered basis.
  - Flexible pricing: Utility pricing, variable payments, pay-by-consumption with metering and subscription models help make pricing of IT services more flexible.

## 1.7 CLOUD SERVICES REQUIREMENTS

Cloud computing is being touted as the next best thing for cutting the cost of providing first-class IT services. You can decide which workloads are right for the cloud and which may not be through an examination of your workloads – uses of IT resources for particular activities or tasks. You can also decide which workloads can go on the vendor cloud (via the Internet or a virtual private network [VPN]) and which need to remain onsite (behind the organization's firewall). This focus on outcomes and delivery models presents a new opportunity to open up competitive accounts and expand the IT optimization conversation with existing clients.

Most cloud computing vendors offer point-solution and product offerings. In contrast, one should offer comprehensive, asset-based solutions that help deploy dynamic infrastructure, which is required for a cloud delivery model. These services along with workload solutions are designed to deliver business outcomes to our clients. Any approach to cloud computing should offer the following powerful advantages:

- A proven service management system embedded with cloud services to provide visibility, control, and automation across IT and business services.
- Services targeted at certain infrastructure workloads to help accelerate standardization of services, supporting significant productivity gains and rapid client payback on their investment.

Infrastructure strategy and planning services for cloud computing should be designed to help companies plan their infrastructure workloads via appropriate cloud delivery model. Specific assistance includes cloud strategy, cloud assessment, design and development of a cloud roadmap, and return on investment (ROI) assessment by workload. Cloud leaders can help clients identify the right mix of public, private, and hybrid cloud models for infrastructure workload. Clients should be encouraged to get started with a strategy and planning consulting engagement as well as a pilot implementation of a key workload.

## 1.8 CLOUD AND DYNAMIC INFRASTRUCTURE

Through cloud computing, clients can access standardized IT resources to deploy new applications, services, or computing resources rapidly without re-engineering their entire infrastructure, thus making it *dynamic.*

Cloud Dynamic Infrastructure is based on an architecture that combines the following initiatives:

- **Service Management:** Provide visibility, control, and automation across all the busines and IT assets to deliver higher value services.
- **Asset Management:** Maximize the value of critical business and IT assets over the lifecycle with industry tailored asset management solutions.
- **Virtualization and Consolidation:** Reduce operating costs, improve responsivene and utilize resources more fully.
- **Information Infrastructure:** Help businesses achieve information compliance, ava ability, retention, and security objectives.
- **Energy Efficiency:** Address energy, environment, and sustainability challenges a opportunities across the business and IT infrastructure.
- **Security:** Provide end-to-end industry customized governance, risk management, compliance for businesses.
- **Resilience:** Maintain continuous business and IT operations while rapidly adap and responding to risks and opportunities.

## 1.9 CLOUD COMPUTING CHARACTERISTICS

Cloud computing uses commodity-based hardware as its base. The hardware can be replaced any time without affecting the cloud. It uses a commodity-based software container system. For example, a service should be able to be moved from one cloud provider to any other cloud provider with no effect on the service.

This also requires a virtualization engine and an abstraction layer for the hardware, software, and configuration of systems. It has the feature of multi-tenant where multiple customers share the underlying infrastructure resources, without compromising the privacy and security of their data. Clouds implement the 'pay-as-you-go' pattern with no lock-in and no up-front commitment and are elastic as the service delivery infrastructure expands and contracts automatically based on the capacity needed.

### *1.9.1 Cloud Computing Barriers*

IT organizations have identified four major barriers to large-scale adoption of cloud services. The first one is security, particularly data security. Interestingly, the security concerns in a cloud environment are no different from a traditional datacenter and network. However, since most of the information exchange between the organization and the cloud service provider is done over the web or a shared network, and because IT security is entirely handled by an external entity, the overall security risks are *perceived* as higher for cloud services. Some additional factors cited as contributing to this perception are (a) limited knowledge of the physical location of stored data, (b) a belief that multi-tenant platforms are inherently less secure than single-tenant platforms, (c) use of virtualization as the underlying technology, where virtualization is seen as relatively new technology, and (d) limited capabilities for monitoring access to applications hosted in the cloud.

The next one is *governance and regulatory compliance*. Large enterprises are still trying to sort out the appropriate data governance model for cloud services, and ensuring data privacy. Quality of service (availability, reliability, and performance) is still cited as a major concern for large organizations. Not all cloud service providers have well-defined service-level agreements (SLAs), or SLAs that meet stricter corporate standards. Recovery times may be stated as 'as soon as possible' rather than a guaranteed number of hours. Corrective measures specified in the cloud provider's SLAs are often fairly minimal and do not cover the potential consequent losses to the customer's business in the event of an outage. Inability to influence the SLA contracts is another issue. From the cloud service provider's point of view, it is impractical to tailor individual SLAs for every customer they support. The risk of poor performance is perceived higher for a complex cloud-delivered application than for a relatively simpler on-site service delivery model. Overall performance of a cloud service is dependent on the performance of components outside the direct control of both the customer and the cloud service provider, such as the network connection.

*Integration and interoperability* is the third concern. Identifying and migrating appropriate applications to the cloud is made complicated by the interdependencies typically associated with business applications. Integration and interoperability issues include a lack of standard interfaces

or APIs for integrating legacy applications with cloud services. This is worse if services from multiple vendors are involved. It also includes software dependencies that must also reside in the cloud for performance reasons, but which may not be ready for licensing on the cloud. There are worries about how disparate applications on multiple platforms, deployed in geographically dispersed locations, can interact flawlessly and can provide the expected levels of service.

The next concern is whether the workloads are suitable (or not) for cloud deployment. Not every application is a suitable candidate for moving to a cloud computing environment. Whether or not a particular application is a good fit depends on a combination of the nature of the business functions it provides, the capacity characteristics it requires (some processing patterns will be more cost-effective than others in a pay-as-you-use model), and technical aspects of the application or its infrastructure requirements.

## 1.10 CLOUD ADOPTION

Business function that suits cloud deployment can be low-priority business applications, for example, business intelligence against very large databases, partner-facing project sites, and other low-priority services. Cloud favours traditional Web applications and interactive applications that comprise two or more data sources and services and services with low availability requirements and short life spans; for example, enterprise marketing campaigns need quick delivery of a promotion that can just as quickly be switched off. It is also helpful when high volume, low cost analytics and disaster recovery scenarios, business continuity, backup/recovery-based implementation are required. It is like a boon to one-time batch processing with limited security requirements, record retention, media distribution, and mature packaged offerings, like e-mail, collaboration infrastructure, collaborative business networks.

Based on technical characteristics, we can say that it is suitable for applications that are modular and loosely coupled; isolated workloads; single virtual appliance workloads and software development and testing; and pre-production systems. It gels well with R&D projects, prototyping to test new services, applications, and design models and applications that scale horizontally on small servers – that is, by adding more servers, rather than by increasing a server's computational capacity.

Applications that need significantly different levels of infrastructure throughout the day such those used almost solely during the business day, should be deployed through cloud Applications that need significantly different levels of infrastructure throughout the month, that have seasonal demand, such as those used primarily during the end-of-the-quarter close during a holiday shopping season, are the best examples of cloud deployments. Applicatio where demand is unknown in advance – for example, a Web start-up will need to support spike in demand when it becomes popular, followed potentially by a reduction once some the visitors turn away – can also be deployed using clouds.

It is not suitable for mission-critical and core business applications, transaction proce ing and applications that depend on sensitive data normally restricted to the organization, requiring a high level of auditability and accountability as these process cannot share the h importance data, processing power, and hardware with the third party. Applications that 24×7×365 with steady demand, applications that consume significant amounts of mem

including applications dependent on large in-memory caches, databases, or data sets are not suitable for cloud. Applications that take full advantage of multiple cores, such as those that do a significant amount of parallel processing, and thus benefit from many cores on a single server, are not recommended for cloud deployment.

It is not recommended for applications that require high-performance file system I/O needing high-bandwidth interserver communications, for example, highly distributed applications. Cloud does not work well with applications that scale vertically on single servers – that is, by increasing a server's computational capacity rather than adding more servers and applications dependent on third-party software, which does not have a virtualization or cloud aware licensing strategy.

## 1.11 CLOUD RUDIMENTS

Cloud delivers a software platform that will enable customer IT to build an Infrastructure-as-a-Service (IaaS) cloud. Cloud is built on the capabilities of existing virtualization management and physical server provisioning solutions to deliver application infrastructure to users that can be consumed in a self-service manner.

Cloud optimizes the usage of the physical and virtual infrastructure through intelligent resource allocation policies, and adds the ability to flex applications elastically based on demand. The high-level capabilities of any cloud include the following:

- **Resource Aggregation and Integration:** Cloud solution operates on top of existing virtualization management, physical server provisioning, and system management environments. It retrieves inventory information about machines and software templates from multiple locations, and aggregates this information into a central logical view of all resources in the infrastructure.
- **Application Services:** Rather than provide access to resources directly, cloud solutions' application 'Definitions' describes packages of machine capacity and software images that can be allocated by resource consumers. Applications can range from individual machines provisioned with an operating system image through to full multi-tier application environments that consist of collections of machines and software stacks provisioned in a specific order with network and storage dependencies handled through integration with third-party management tools. Application instances represent an agreement between the cloud provider and consumer to use capacity on a reservation or on-demand basis. Reservations allocate capacity in the resource inventory, guaranteeing that the capacity will be available to the consumer at some defined point in the future. On-demand allocations provide access to resources but do not guarantee availability. Reserved and on-demand capacity can be combined in an application, where a baseline of capacity can be elastically increased or decreased according to metrics and policies defined by the consumer.
- **Self-Service Portal:** An important principle of a cloud solution is to enable self-service access to resources with minimal IT involvement. It should support the notion of account owners signing up for contracts and then being able to delegate the use of the purchased capacity within their own groups or departments. Users can request machines

or entire multi-machine application environments and monitor and control them using a web-based self-service portal. The system will drive the workflows necessary to create the environment, and provide run-time environment management in order to support application elasticity.

- **Allocation Engine:** Dynamic Resource Management (DRM) is the automated allocation and reallocation of IT resources based on policies that express business demands and priorities. DRM is a key component of any cloud solution that maximizes the efficiency of the IaaS infrastructure. DRM policies should be applied both when initially placing applications onto machine resources and when selecting applications to migrate in order to preserve SLAs around application performance. Some of the allocation and migration strategies include advance reservation of resources, load-based placement and migration, application and resource topology constraints, energy usage optimization, etc. The use of sophisticated DRM helps to increase utilization of cloud resources, reduces overspending by effectively using existing resources, and saves costs in terms of operations, power, and cooling.
- **Reporting and Accounting:** In order to close the loop and determine how the cloud is behaving, metering information on resource allocations as well as actual usage is collected in an accounting database. The data is centrally available to create reports on inventory capacity, capacity allocated versus capacity used by contract, and usage-billing reports based on consumed resources.

The following are the *cloud features* that would help to bring in *agility* and *transparency* along with increase in the utilization of the existing resources at the datacenter of any customer.

- **Self-Service:** This feature presents an interface for separate authenticated end-users – via role-based access controls (RBAC) – to select options for deployment. It should have unique policy controls per tenant and user role, and the ability to present unique catalogues per user or group. The self-service portal is a web interface also accessible in other ways, such as through a mobile client, etc.
- **Dynamic Workload Management:** With cloud solution implemented, datacenters are enabled with automation and orchestration software that coordinates workflow requests from the service catalogue or self-service portal for provisioning virtual machines. Also each provisioned virtual machine is enabled with a life-cycle for deployment expiration which increases the efficiency of utilization of resources.
- **Resource Automation:** Using cloud solution, Admins or engineering team members of the datacenter could control the heterogeneous environment on a single pane. This feature establishes secure multi-tenancy, isolates virtual resources, and helps prevent contention in the load aware resource engine which intelligently does the workload packing or load balancing across hypervisors automatically.
- **Chargeback, Showback, and Metering:** Using this feature Admins could bring out the usage reports for cloud infrastructure service consumption and these usage reports serves as a basis for metering and billing system. Using this Admins will be able to understand if the virtual machines are attached with appropriate resources. Enabling *chargeback*, *showback*, and *metering* in any organization would bring in transparency to the business and environment for management to clearly see the usage and dollar value associated to it and take decision-making steps.

- **Open Architecture:** The cloud should be integrated with existing third-party products that are already installed in the datacenter. It should also be integrated to a public cloud for using additional resources and should be managed through a single cockpit. It is also possible to meter the public cloud resource usage.
- **Image Pools:** The cloud solution should have full blown service catalogue and support to most of the operating systems. It should be possible to vary the hardware configuration for the templates. It should also integrate with existing templates and images used by the development and testing teams.
- **Role-Based Access Administration:** The cloud solution should have the capability to integrate cleanly with any of the existing, LDAP, or other authentication and identity mechanisms. These features are crucial for providing secure multi-tenancy. This would also bring in security to the self-service portal.
- **Virtualization:** The cloud should extend support to virtualization layer. This implies that it should support most of the industry-proven hypervisors. This enables the Admins and engineering team of the datacenter to control them over a single pane.

### 1.11.1 Cost Savings with Cloud

#### *Faster Time-to-Market (Missed Business Opportunity)*

Deploying new application environments quickly and reliably can have a directly impact on competitiveness enabling organizations to take market share. The cloud will enable automated delivery of application environments exponentially faster than current practices.

With the cloud model, teams could be delivering fully configured, multi-component application environments to users in some minutes. This makes an immediate impact on user efficiency as well as eliminating much of the manual labor previously required of both the IT and application teams. In addition to this, ability to remove a (physical or virtual) will have similar performance, and once again allows that compute power to be available for other uses.

#### *Public Cloud Interfaces*

Cloud infrastructure with its policies should manage workload placement optimally by looking at several metrics. Cloud should also offer the capability to burst out to public cloud or internal resources when needed and cut off that link when done. The cloud should also be able to meter for the usage of the deployed instances in public cloud. Customer datacenter could use resources in public cloud for test and development environment if there are no resources available on the premise which will also help them to defer from the new hardware procurement.

#### *Automated Scaling*

The cloud solution should provide an out-of-box functionality to flex-up or flex-down an application instance or resource based on performance metrics and should also flex-up and flex-down an environment automatically or manually. The cloud solution should offer policies that can be customized to look at any metric and take action based on the threshold. These policies

must be embedded in a service catalogue to monitor an application or the entire environment and flex-up or flex-down with more resources.

***Business Transparency***

Service Accounting helps to improve utilization of datacenter infrastructure with accurate visibility into the true costs of physical and virtualized workloads. It will enable decision makers to have full cost transparency and accountability for usage, metrics, roles and definitions. This would also help an Admin to understand whether a machine is equipped with right resources or not.

### *1.11.2 Benefits*

Cloud brings lot of benefits for any enterprises. Let us explore these in brief here. They will be discussed in detail in the later chapter also.

- Increase agility on the IT datacenter resources and innovation.
- Enable self-service portal and thus ensure VM in less lead-times.
- SLAs are met as the VM lead-times and downtimes are significantly reduced.
- Trial and error configuration tests can be done at ease.
- Complete control over cloud usage for Admins.
- Scalability and flexibility allow the IaaS cloud to almost deliver the promise of unlimited IT services on demand.
- Pay for only what they use and are not charged when their service demands decrease.
- Significant reduction in the costs for IT datacenter.
- Private cloud enables dynamic sharing of the resources available in IT datacenter so that demands can be met cost-effectively.
- Considerable increase in the utilization of resources of IT datacenter.
- Increase in operational efficiency of the resources in the IT datacenter.
- Achieve a greener datacenter (server consolidation and virtualization enables over committed machines).
- Support for heterogeneous hardware vendors. Avoids Vendor Locking.

It will help the enterprises by

- Reducing the number of administrators required to manage a more diverse IT resource pool.
- Dramatic reduction in cycle times to provision new assets.
- Realization of an infrastructure 'pay-per-use' model.
- Reduction in planned capital spending and maintenance.
- Increased user satisfaction with IT services.
- Reduction in physical server count.
- Consolidation of enterprise application licenses.
- Flexibility to meet future demands on infrastructure goals that can be leveraged.
- Capacity on-demand (pre-provision, automate).