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 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 the recent period which is not totally true. Grid computing paves the way for the evolution of 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; at the same time, it is also possible to dynamically increase or decrease the computational power of 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 (Fig. 1.1). As said, there are different types of grids 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

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

Conceptually, we can imagine the following three layers:

1. **Lower layer:** This is a physical layer where we have servers, storage devices, and interconnecting network.
2. **Middle layer:** This layer represents different operating systems mapped one-to-one with servers.
3. **Upper layer:** This is an application layer in which we map different applications supporting enterprise business processes.

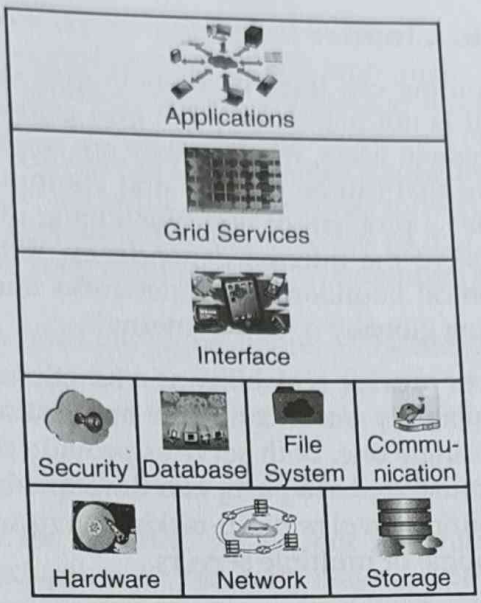

FIGURE 1.1 Simple grid architecture.

The concept of grid computing evolved from 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 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 these resources are 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 a 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 the 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 on a single one. Both services provide on-demand services for users, storage, networks, and data transferred at a particular time, and can be de-allocated when they are not required. These computing involve 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 behavior 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 fully exploit the grid advantages, physical resources across heterogeneous systems can be virtualized building a single resource image.

The following list will help us understand the benefits of grid computing when deployed for infrastructure management and extended to cloud computing arena (Fig. 1.2). These benefits are also discussed in detail later in the chapter.

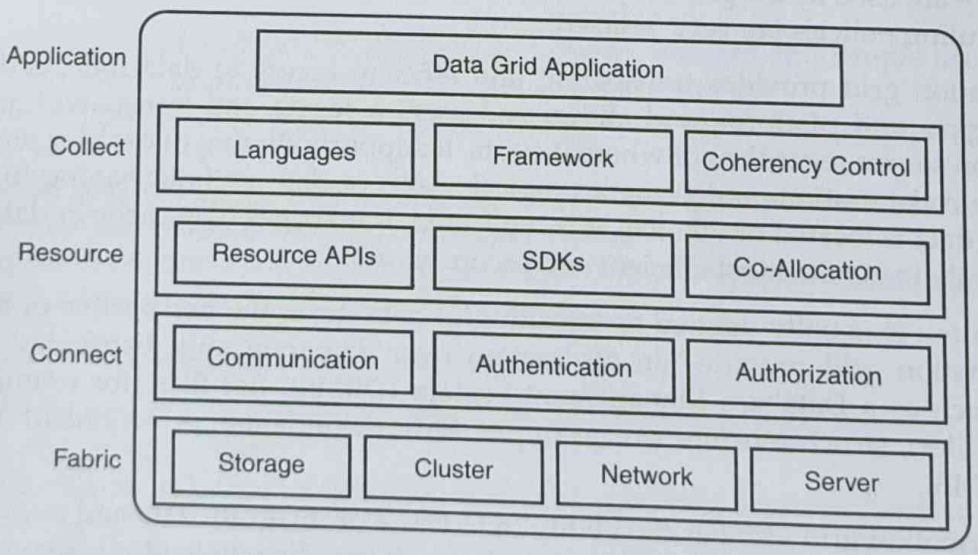

FIGURE 1.2 Standard grid architecture.

1. **Storage/data/information:** It provides logical views of data without having to understand where the data is located or whether it is replicated.
2. **System management:** It defines, controls, configures, and removes components and/or services (could be physical) on a grid using automated or physical methods.
3. **Metering, billing, and software (SW) licensing:** It provides tools to monitor and distribute the number of licenses while using licensed software. It also 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.
4. **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 benefits (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 IT – from computer applications, software, business processes, messaging, and collaboration – to end-users as a service wherever and whenever they need it.

The cloud refers to a group of hardware computing devices, software, storage devices, and application programming interfaces (APIs) that integrate and interface with each other to deliver the characteristics of cloud computing in a service model. 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 to buy and build an IT infrastructure or to understand the underlying technology.

1.2.1 Emerging Through Cloud

Cloud computing is a new paradigm for delivering IT where rapid provisioning is an important characteristic for computing resources, data, applications, and IT. This is offered as the highly standardized offering to the consumers over the Web portal via the Internet.

Cloud computing provides a way of managing large pools of servers that are virtualized. However, from the consumer point of view, it is regarded as a single, large resource pool.

Today it is the need of the business. This disrupting technology gives us the offerings to face the challenges in multiple ways by

1. Decreasing the capex and opex cost.
2. Enhancing the service quality and offer the next generation services.
3. Maintaining the desired and right level of security, compliances, regulations, and policies across the different functions of enterprise.
4. Rapid provisioning, agility, and business transparency for consistent self-service delivery.

Thus, cloud computing is the service and deployment model using large resource pool based provisioning of virtual or physical resources in a service model using the Internet (public cloud) or intranet (private cloud).

1.3 Benefits

Today cloud computing is emerging because it promises to reduce the IT complexities and costs. It is a new user experience especially over the Web traffic where the consumer is mostly related to social media, networking, and Web 2.0. This provides the mechanism to request and use the services with the abstraction of the technology. Today we are using the cloud in many ways without understanding the technology; for example, it can be experienced when we are sharing photos, streaming any video, using smart phones, transacting in banks, or even attending a meeting with Web-based Internet bridges.

When we see this concept as a disrupting technology, it is the methodology where the scalable resources such as computing resources, storage, and network are offered in a service-based model over the network. The entire self-service experience is highly automated and usually takes minutes to provision the services. It is based on self-service scalable delivery to use various offerings based on bundles of computing resources, middleware, operating systems, and applications. Organizations are adopting these disrupting technology models for faster services to their consumer to increase staff productivity, to decrease time to market, for economies of scale, and for effective workload management. These cases can be test and development, virtual desktop infrastructures, messaging and collaboration, and business analytics.

1.4 Why Cloud?

The 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 economies of scale as well as from statistical multiplexing (Fig. 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

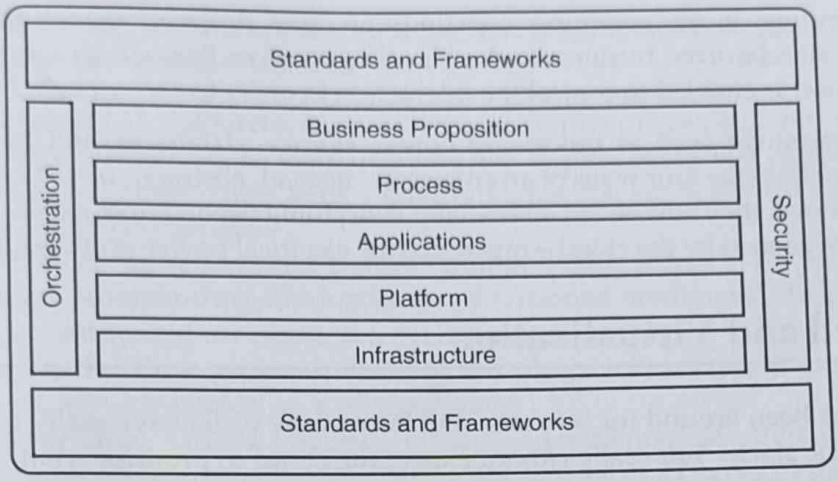

FIGURE 1.3 Basic cloud computing model.

to easily provision massive amounts of resources on a moment's notice and release them back into the cloud just as quickly.

The different reasons for adopting the cloud are as follows:

1. Very big, Web-scale infrastructure that is abstracted.
2. Dynamic allocation, scaling, movement of applications.
3. Pay-per-use.
4. No long-term commitments.
5. Operating system (OS), application architecture independent.
6. No hardware or software to install.

This results in the following business- and IT-aligned benefits:

1. More emphasis on innovation to launch new offerings.
2. IT as an enabler for innovation and rapid deployment.
3. Use of self-service based consistent delivery model.
4. Higher service quality and increased agility, ubiquitous computing.
5. Uniqueness via service models and competitive in the ecosystem.
6. Anytime anywhere computing for consumers over the Internet to deliver next generation technologies.
7. Reduced IT obstacles enabling launch of new offerings.
8. Building and integration of 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. So 'cloud computing' refers to delivering IT-enabled services via the Internet that are built for the end-user to be in control.

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 an enterprise. Instead, 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 users have really truly virtualized all the layers of the stack? You really cannot expect the cloud to produce what it is expected to produce if it is not virtualized, standardized, and automated, because people expect scalable services. The cloud environment is assumed as a self-service experience to quickly start the services with respect to on-demand services. This requires a fundamental platform to meet these demands.

The only way you are going to be able to get efficiency is by virtualizing, standardizing, and automating (Fig. 1.4). And that is going to drive down costs and improve service. This is really

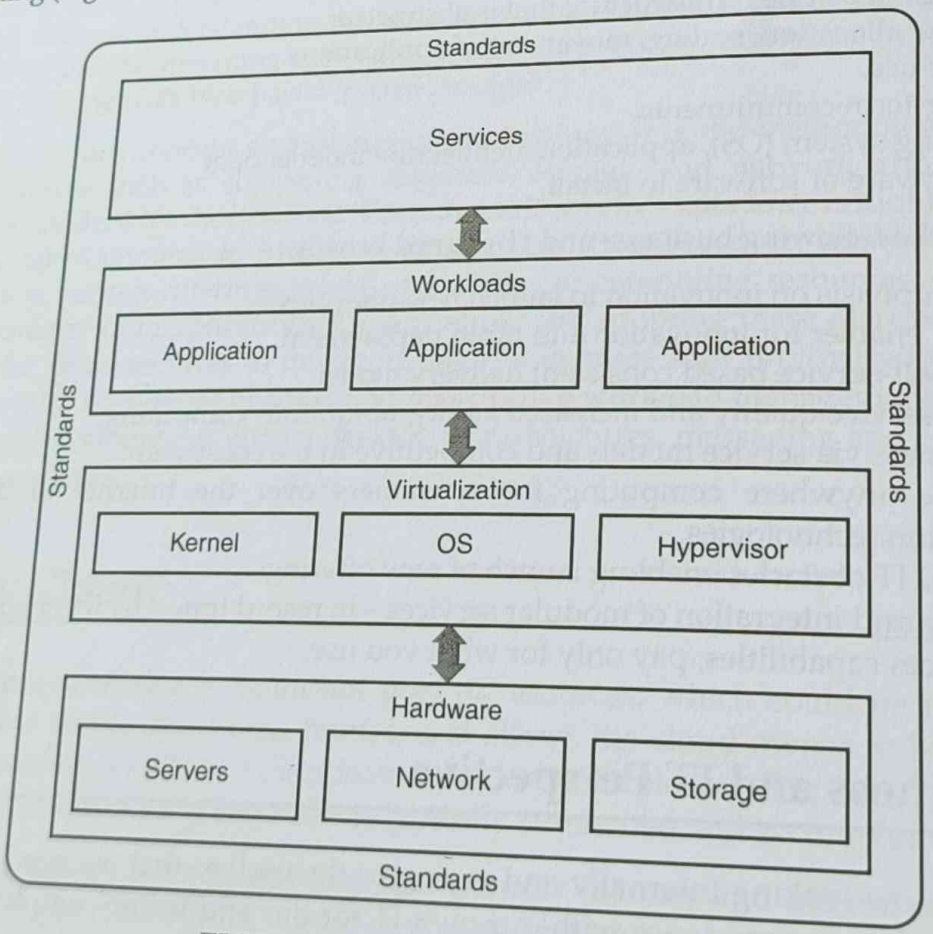

FIGURE 1.4 Datacenter clouds.

a pretty simple equation and we are seeing organizations that are doing this to achieve very real measurable business results. These results include:

1. **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.
2. **Automation using self-service portal:** Point-and-click access to IT resources.
3. **Automated provisioning:** Resources are provisioned on demand, helping in reducing IT resource setup and configuration cycle times.
4. **Standardization through service catalog ordering:** Uniform offerings are readily available from a service catalog on a metered basis.
5. **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 are not through an examination of your workloads – usage of IT resources for particular activities or tasks. You can also decide which workload 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 the clients. Any approach to cloud computing should offer the following powerful advantages:

1. A proven service management system embedded with cloud services to provide visibility, control, and automation across IT and business services.
2. Services targeted at certain infrastructure workloads to help accelerate standardization of services, supporting significant productivity gains and rapid client payback on the 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 road-map, 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 plan consulting engagement as well as a pilot implementation of a key workload.

1.8 Dynamic Cloud 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:

1. **Service management:** Offers business transparency and automation across the pillars of business for consistent delivery.
2. **Asset management:** Maximizes the value of critical business and IT assets over their lifecycle with industry-tailored asset management solutions.
3. **Virtualization and consolidation:** Reduce operating costs, improve responsiveness, and fully utilize the resources.
4. **Information infrastructure:** Helps businesses achieve information compliance, availability, retention, and security objectives.
5. **Energy efficiency:** Offers green and sustainable energy solutions for business.
6. **Security:** Provides end-to-end industry customized governance, risk management, and compliance for businesses.
7. **Elasticity:** Maintains continuous business and IT operations while rapidly adapting 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. Cloud computing uses a commodity-based software container system. For example, an instance or service can be migrated from one provider to other service provider with zero impact.

Cloud computing also requires a virtualization engine and an abstraction layer for the hardware, software, and configuration of systems. Cloud computing also has the multi-tenant feature 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 on the basis of the capacity needed.

1.9.1 Cloud Computing Hurdles

There are various hurdles in adopting the cloud for large-scale cloud deployment services. The first and foremost is security. Now because of new paradigm, the data security concern is more hyped for cloud, but in some manner it is same as securing the datacenter services, network, storage in hosted and utility-based solutions. As the services are opened and delivered over the network between the cloud service provider and the consumer, the security in this model is perceived at higher levels. Other inhibitors can be location-independent resource pooling where consumer does not know where his services are running or where his data is stored. It is also believed that multi-tenant models are somewhat less secure than dedicated

models. Limited service management and monitoring capabilities in the public cloud model also add to the complexities.

The second hurdle is of regulation and compliances. There is a need of data governance models to be established in the enterprises and federating data privacy. In large organizations, IT delivery is taken with the concerns of reliability, performance, and availability. There are different levels of maturities for organizations seeking different levels of service-level agreements (SLAs) but cloud service providers are not equipped to deliver the services. There is a need of stringent recovery point objective (RPO) and the recovery time objective (RTO) with the agreed number of mins/hours down-time. There is a very less chance to take any corrective actions after any impact and it does not cover the consequences happened due to outage. It is also difficult to cater to different type of SLAs for various tenants by the service providers as it is very difficult to tailor the SLAs based on the customer interest and requirements. Also it is observed that if the service delivery is for a complex application bundle, there is a risk of poor performance. There are few components in the cloud ecosystem that are beyond the control of a consumer or a service provider such as bandwidth.

Cloud migration is the next hurdle in cloud computing. This requires the property of powerful interoperability of platforms that should identify the appropriate application that can be migrated to the cloud. It is important to identify the interdependencies and integration points with standards and interfaces that are lacking today among service providers.

Cloud migration becomes more complex if the service bundles are integrated from multiple cloud service providers. This can also become the deal breaker or the reason for downgraded performance. There can also be the licensing problem in the cloud environment. Also there are issues related to multi-geography-based application platforms deployment and implementations in the cloud-based model and get some hits with respect to the desired levels of service for multiple offerings.

Last hurdle is the workload suitability for cloud. It is also a big question whether the workload is seasonal for the cloud deployment or not. Not all the applications are suitable candidates for the cloud. It depends on the function of the business, enterprise policies, application architecture, scalability, suitability, usage patterns according to pay-per-use-model, or infrastructure requirements in the service model.

1.10 Cloud Adoption

Business functions that suit cloud deployment can be low-priority business applications – such as analytics, against partner and field service-based functions – and other low-priority business functions. Cloud favors 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, and 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 such as e-mail, collaboration infrastructure, and collaborative business networks.

Based on technical characteristics, we can say that cloud adoption is suitable for applications that are modular and loosely coupled, isolated workloads, single virtual appliance workloads; software development and testing, and pre-production systems. It gels well with research and development 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 as those used almost solely during the business day, should be deployed through the cloud. Applications that need significantly different levels of infrastructure throughout the month, or that have seasonal demand, such as those used primarily during the end-of-the-quarter or during a holiday shopping season, are the best examples of cloud deployments. Applications for which the demand is unknown in advance, for example, a Web-based application for a start-up organization, will need to support a spike in demand when they become highly demanding and will need to reduce once all the users turn away from the workload.

Cloud adoption is not suitable for mission-critical and core business applications, transaction processing, and applications that depend on sensitive data normally restricted to the organization or requiring a high level of auditability and accountability as these processes cannot share the high-importance data, processing power, and hardware with the third party. Applications that run $24 \times 7 \times 365$ with steady demand and applications that consume significant amounts of memory – including applications dependent on large in-memory caches, databases, or data sets – are not suitable for the 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.

Cloud adoption is also not recommended for applications that require high-performance file system I/O which needs high-bandwidth interserver communications, for example, highly distributed applications. The cloud does not work well with applications that scale vertically on single servers, that is, by increasing a server's computational capacity rather than by 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

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

The 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:

1. **Resource aggregation and integration:** The cloud solution operates on the top of existing virtualization management, physical server provisioning, and system management environments. It retrieves inventory information about machines and softwa

templates from multiple locations, and aggregates this information into a central logical view of all resources in the infrastructure.

2. **Application services:** Rather than providing 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 the 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.
3. **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.
4. **Allocation engine:** Dynamic Resource Management (DRM) is an 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, and energy usage optimization. The use of sophisticated DRM helps to increase the utilization of cloud resources, reduces overspending by effectively using existing resources, and saves costs in terms of operations, power, and cooling.
5. **Reporting and accounting:** To close the loop and determine how the cloud is behaving, metering information on resource allocations as well as on actual usage is collected in an accounting database. Data are 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 an increase in the utilization of the existing resources at the datacenter of any customer.

1. **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

catalogs per user or group. The self-service portal is a Web interface also accessible in other ways, such as through a mobile client.

2. **Dynamic workload management:** With the cloud solution implemented, datacenters are enabled with automation and orchestration software that coordinates workflow requests from the service catalog 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 resource utilization.
3. **Resource automation:** Using cloud solution, administrators or engineering team members of the datacenter can 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.
4. **Chargeback, showback, and metering:** Using this feature, administrators can bring out the usage reports for cloud infrastructure service consumption, which serve as a basis for metering and billing system. Using this, administrators will be able to determine 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 the environment for the management to clearly see the usage and the dollar value associated to it and take decision-making steps.
5. **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 to enable the use of additional resources and should be managed through a single cockpit. It is also possible to meter the public cloud resource usage.
6. **Image pools:** The cloud solution should have a full-blown service catalog and support to most of the operating systems. It should be possible to vary the hardware configuration for the templates. The cloud should also integrate with existing templates and images used by the development and testing teams.
7. **Role-based access administration:** The cloud solution should have the capability to integrate cleanly with any of the existing Lightweight Directory Access Protocol (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.
8. **Virtualization:** The cloud should extend support to the virtualization layer. This implies that the cloud should support most of the industry-proven hypervisors. This enables the administrators and the engineering team of the datacenter to control the hypervisors 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 direct impact on the 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 just a few minutes. This makes an immediate impact on user efficiency as well as eliminates much of the manual labor previously required of both the IT

and the application team. In addition to this, ability to remove a physical or virtual compute, will have a similar performance efficiency. This once again allows computing power to be available for other users.

Public Cloud Interfaces

Cloud infrastructure with its policies should manage workload placement optimally by looking at several metrics. The cloud should also offer the capability to burst out to a public cloud or internal resources when needed and cut off that link when done. The cloud should also be able to meter the usage of the deployed instances in the public cloud. Customer datacenters could use resources in the public cloud for testing and development environment if there are no resources available on the premise, which will also help them to defer from 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 catalog to monitor an application or the entire environment and flex-up or flex-down with more resources.

Business Transparency

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

1.11.2 Benefits

The cloud brings a lot of benefits for any enterprise. These benefits are given in brief as follows and will be discussed in detail later in the book.

1. Increased agility on the IT datacenter resources and innovation.
2. Enabling of self-service portal and thus ensure virtual machines (VMs) in less lead-times.
3. Adherence of SLAs as the VM lead-times and down-times are significantly reduced.
4. Trial and error configuration tests can be done at ease.
5. Complete control over cloud usage for administrators is possible.
6. Scalability and flexibility allow the IaaS cloud to almost deliver the promise of unlimited IT services on demand.
7. Usage-based payment and not getting billed when the utilization decreases.
8. Significant reduction in the costs for IT datacenter.
9. Dynamic sharing of the resources available in IT datacenter through private cloud so that demands can be met cost-effectively.
10. Considerable increase in the utilization of resources of IT datacenter.

11. Increase in the operational efficiency of the resources in the IT datacenter.
12. Achieve a greener datacenter (server consolidation and virtualization enables over committed machines).
13. Support for heterogeneous hardware vendors. Avoids vendor locking.

It will help the enterprises by

1. Reducing the number of administrators required to manage a more diverse IT resource pool.
2. Dramatically reducing in cycle times to provision new assets.
3. Realizing an infrastructure 'pay-per-use' model.
4. Reducing planned capital spending and maintenance.
5. Increasing end-user satisfaction with the help of IT services.
6. Reducing physical server count.
7. Consolidating enterprise application licenses.
8. Flexibility to meet future demands on infrastructure goals that can be leveraged.
9. Increasing the capacity on demand (pre-provision, automate).
10. Consolidated, streamline change control.
11. End-to-end application provisioning.
12. Allowing developers to provision development application environment autonomously.
13. End-to-end performance measurement.
14. Consumption-based charge back.
15. Plan for active/active datacenter operations.
16. Plan for increased datacenter density.
17. Separate production and development networks.

1.12 Summary

In this chapter, we explored cloud computing, its benefits, and its services. The chapter also gave deep insights into cloud computing models that are put into practice.

HAPTER

2

Cloud Deployment Models

CONTENTS

2.1 Introduction

Cloud computing is an emerging style of computing in which applications, data, and resources are provided to users as services over the Web. The services provided may be available globally, always on, low in cost, 'on demand', massively scalable, and on 'pay-as-you-grow' basis. Consumers of a service need to care only about what the service does for them and not on how it is implemented. Cloud computing is a technology that allows users to access software applications, store information, develop and test new software, create virtual servers, draw on disparate IT resources, and more – all over the Internet (or other broad network).

Cloud computing is a model-driven methodology that provides configurable computing resources such as servers, networks, storage, and applications as and when required over the Internet with minimum efforts. The cloud also indicates essential uniqueness, service models, and deployment models.

This chapter visualizes several models for cloud computing, including private clouds (where the deployment is within the organization's firewall) and public clouds (where application services and data are hosted by a third party outside the firewall). Consistent data availability and security are critical success factors for any cloud deployment. Businesses need to ensure that data are adequately protected and can be restored in a timely fashion following any disruption event.

When we deploy the cloud from the service provider's point of view, it requires a datacenter; but from the consumer's point of view, it is expected not to think about the datacenter. A datacenter is a facility house of servers, storage, security devices, networks, and telecommunications. It includes power supplies, redundant power backup options, communication setups, cooling, fire suppression, floor space, etc. with environmental controls.

All datacenters are designed on the basis of geographies and local requirements with redundant power, cooling, resources security, and environmental controls. However, clouds are not bound to location and possess the characteristics of location independency. The cloud provides the abstract view that it is not specifically tied to any datacenter for the self-service request of compute, storage, and disaster recovery options. Availability and redundancy of the datacenter come from the service provider, from which the resources are pooled and services are catered. In this way, the cloud is achieved and spanned from multiple datacenters, but there is no datacenter itself.

2.2 Cloud Characteristics

The cloud carries the basic infrastructure characteristics that are helpful to deploy the cloud service in a fast and cost-effective way (Fig. 2.1). The characteristics discussed in the following subsections set apart the cloud from other computing techniques.

2.2.1 Self-Service On-Demand

As a cloud consumer, users are privileged to request and provision computing capabilities bundled with services with or without approval process powered by automation and workflows.

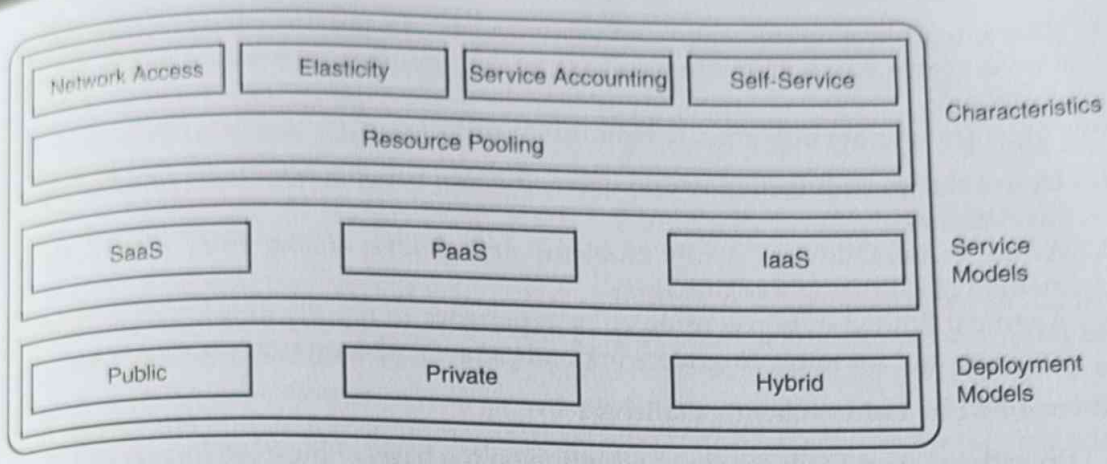

FIGURE 2.1 Cloud model.

2.2.2 Ubiquitous Network Access

This is the characteristic by which end-user and server computing devices can be accessed over the network even using the next generation heterogeneous devices such as smart phone, tablets, phablets, thin and thick clients.

2.2.3 Resource Pooling

This characteristic refers to the pooling of resources across multiple datacenters. These pooled virtual datacenters are then divided into multiple pools to provide their services to various consumers in a multitenant model. These pools can have both physical and virtual resources. Also the devices provided by this pool give the notion of location-independent compute (storage, servers, processing, network bandwidth, virtual machines, etc.), where the consumer does not have control or visibility about the service location and its geography.

2.2.4 Rapid Elasticity

This characteristic make the provisioning rapid and elastic. This provisioning can be automatic and can flex-up and flex-down on the basis of spikes of utilization. The consumer can view the infinite capacity available as a service, which can be bought at any point of time.

2.3 Measured Service Accounting

The cloud environment is optimized by effective workload management. This management requires measured service, monitoring, metering, and chargeback capability with the required abstraction, and optimization at user level. The usage of resources can be controlled, reported, monitored, charged, billed, and invoiced on actual from provider to consumers.

The measured service is directly proportional to standardization and to the economies of scale for operating expenses. The more the virtualization is practiced in the environment, the

more it helps to achieve optimization. Both standardization and virtualization help to reduce the cost investments while maintaining the required resources to meet the dynamic need of the infrastructure.

Now organizations are migrating to cloud computing to

1. Derive the greatest flexibility and cost-reduction benefits from their cloud computing investments.
2. Avoid vulnerability to costly problems and delays arising from a trial-and-error method of migrating workloads.
3. Augment limited in-house resource or experience to rapidly develop an optimization roadmap and smoothly migrate workloads to a cloud computing environment.

Cloud vendors can address client's challenges by

1. Prioritizing workloads for cloud adoption on the basis of business impact and risk.
2. Maximizing business return by identifying applications that are well suited for cloud computing and have high business impact.
3. Addressing problematic workloads to improve their propensity for cloud computing.
4. Avoiding costly implementation issues by identifying and addressing potential difficulties during the migration.
5. Mitigating the risk of costly implementation delays by identifying potential problems and addressing them before the migration.
6. Avoiding inadequate performance of highly complex and integrated workloads.
7. Leveraging the expertise to deliver an actionable roadmap to successfully migrate applications to a cloud computing environment.
8. Accelerating the cloud initiatives.

2.3.1 Cost Factor

There are a number of reasons why cloud computing is popular with businesses. One of them is the cost aspect. By virtualizing and standardizing your environment, you can deliver more services with fewer resources and drive up the utilization. By adding automation, you can reduce the labor cost which gives you an additional cost benefit. These advantages give you a lot of flexibility because you can access cloud workloads services without thinking about the location and the time of their execution. So cloud computing allows an organization to free up the budget so that money can be diverted to new innovations and development of new capabilities rather than just keeping the lights on and running the IT enterprise.

The growing complexity of IT systems and soon a trillion connected things demand that sprawling processes should become standardized services that are efficient, secure, and easy to access. In order to get the control, transparency, and automation of the system, a service management platform is required for consistent delivery. Self-service plus standardization will drive lower operational costs, unlock productivity, and ensure better security.

The cloud allows businesses to be smarter in services delivery. The first aspect of this is a self-service portal that allows your end consumers to see only the services they are allowed to have; however, it also allows them to initiate the process of getting those services. Behind that service request, you could put either a very light or no-touch approval process, or you could put a more complex one in which you may need multiple levels of signature.

This allows you to really fit what the business needs. In some cases, where you have high security, you have high-level service-level agreements (SLAs) – you really want to be able to control how those services are distributed. In other cases, let us say you do not have a research and development team and you want the ability to have as much flexibility as possible. It allows you to be very productive if you adopt the cloud platform which allows you to really make your infrastructure more dynamic and get the resources to the teams that really need them at any point of time.

Let us look at some of the major factors that are driving cloud computing economics. If you look at the infrastructure layer, first comes virtualization. By virtualizing workloads and being able to stack multiple workloads on a system to drive up the utilization, you can lower your capital requirements. In a number of cases, businesses have hundreds – if not thousands – of physical servers and unless they have used the virtualization and are really driving that utilization, the utilization could be as low as 10%. So, in a lot of cases, organizations that use cloud computing are able to drive the utilization up, lower the future capital requirements, and even retire the antiquated equipment and drive their costs down.

From the labor perspective, using a self-service portal allows your clients to help themselves. So there is less support and it makes the offering more available from a service availability perspective. In terms of automation, cloud computing takes the tasks which are manual and repeatable, and their automation reduces your IT operations cost. In a development or test environment, you need multiple skills – such as operating system skills, middleware skills, database skills, and application skills – to get that environment to the end-user. This allows you to define environment as a repeatable, deployable resource, and it drives down your labor cost. Of course, you need to standardize the workloads. Standardization has labor cost and quality benefits so that you can ensure consistency from one environment to another environment.

In many cases, you may want to use multiple models for different types of services that you want to deliver. Starting with private cloud services, the first model (which is also the most popular currently) is the private on-premise cloud. If the cloud is within the organization's datacenter, it is operated and managed by the organization itself.

One of the daunting task in the field of infrastructure management is to optimize the capex and opex cost. This task can be efficiently handled with the help of cloud computing. Cloud computing aims to guarantee the consistent service delivery with economies of scale with the infrastructure powered by the internet. The cloud gives the flexibility for a new model of business by optimizing the IT business while maintaining consistent self-service delivery.

2.3.2 Benefits

We can enjoy many benefits by adopting the cloud:

1. **Self-service capability:** When somebody deploys the cloud services, he becomes capable of self-service. Now testing teams do not have to buy computing services as they can enjoy the same services over the cloud. It also reduces the procurement process. Hence, the testing teams can concentrate on the testing services and efforts.
2. **Resource availability:** It is one of the most common benefits facilitated by the virtualization. It also helps to track and leverage the resource pool under the same umbrella of resource units.

3. **Operational efficiency:** Sometimes conventions and configurations followed by the test and operation teams may differ from those followed by the development teams. This difference may cause the application behavior to be different from what was intended as well as result in delayed services. The template-based approach, with its solution stacks of hardware, configurable applications, and operating systems, is more transparent and can help the teams to understand the environment better.
4. **Hosted tools:** In cloud computing, the developers and testers need not to install, configure, run, or maintain tools on their systems as they can log into the network maintaining these tools from any machine.

These four benefits help the developers and testers to concentrate on their core work and retain focus without worrying about other jobs. This increases quality and productivity; and therefore results in more developer innovation, increased test quality and coverage, which are beneficial for an organization.

There are a number of major challenges that developers face today in getting started and rolling out new applications and services faster. Innovative new products and services are the lifeblood of rapidly growing companies and represent a substantial portion of corporate sales and profits. In an environment of heightened competition, the inability to roll out new applications and services quickly means declining market share and lost revenue.

A growing application backlog leaves lines of business and end-users frustrated because they consider IT as a bottleneck and look for ways to work around IT to roll out new products and services more quickly. Testing backlog is often very long and a major factor in the delay of new application deployments.

A major reason why testing takes so long, on an average it takes weeks, is because of the time taken to set up the application environments for testing and QA as well as production. The testing phase becomes longer because of the time required to procure new hardware and software, and then schedule the time with the IT to configure and set up the systems. Configuration and setup are manual processes where errors are easily introduced. An average new application takes six to nine months to be deployed. The whole process application development is affected by a number of factors ranging from poor governance to poor collaboration between business users and development to inflexible infrastructure and tools. Almost 30% of the defects are caused by wrongly configured test environments. This is a result of following manual processes without any automation to replicate testing environment along with challenges organizations face in finding available resources to perform tests in order to move new applications into production. Test environments are seen as expensive and provide little real business value.

2.4 Cloud Deployment Models

Let us talk about cloud computing and different types of cloud deployment models and services that can be delivered using these models. Cloud computing is a style of computing in which business processes, application, data, and any type of IT resource can be provided as a service to the users.

Cloud delivery models can be briefly classified into the following three types (Fig. 2.2):

1. **Public:** In a public cloud, a business rents the capability and pays for what is used on-demand.
2. **Private:** In private clouds, a business essentially turns its IT environment into a cloud and uses it to deliver services to their users.
3. **Hybrid:** Hybrid clouds combine elements of public and private clouds.

Private cloud is good when the control and customization are bigger concerns to drive the efficiency. Today public clouds are also process-oriented and easily standardized with lower risk problems. There are many standardized functions that can be moved to public cloud such as collaboration, search, CRM, and sales force management.

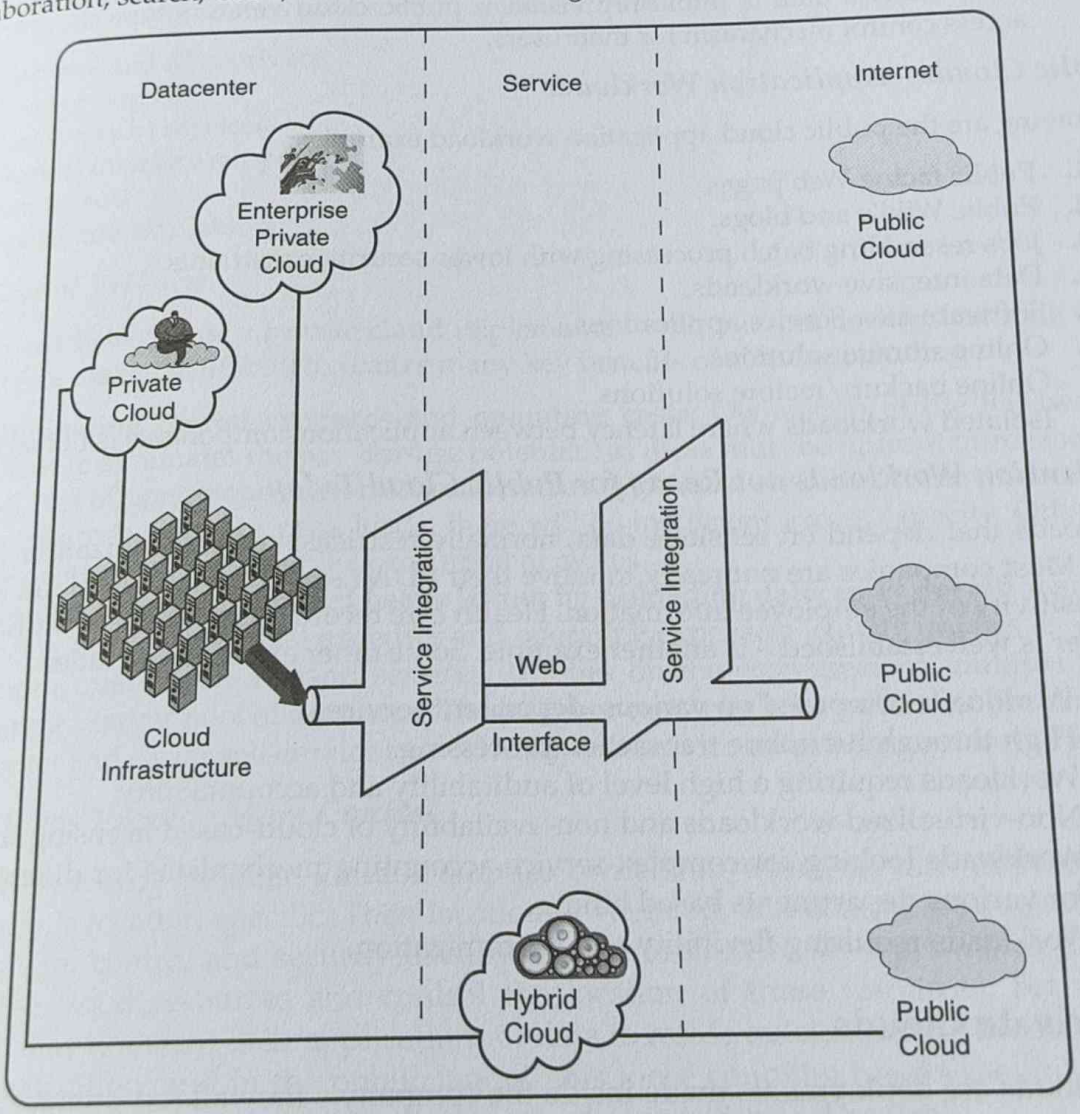

FIGURE 2.2 Private, public, and hybrid clouds.

There is no one-size-fits-all model; in a number of cases, businesses may end up using all these models eventually based on the business model for different services.

2.4.1 Public Clouds

Public cloud services are offered by third-party datacenter provider to end-user consumers over the Internet. Public cloud offers resource pooling, self-service, service accounting, elasticity, multi-tenancy to manage the solutions, deployment, and securing the resources and applications. Companies can use it on-demand and with the pay-as-you-use option, it is much like utility consumption. Enterprises are able to offload commodity applications to third-party service providers (hosters).

The term 'public' does not mean

1. That it is free, even though it can be free or fairly inexpensive to use.
2. That a user's data is publically visible – public cloud vendors typically provide an access control mechanism for their users.

Public Clouds – Application Workloads

Following are the public cloud application workload examples:

1. Public facing Web pages.
2. Public Wiki's and blogs.
3. Jobs resembling batch processing with lower security constraints.
4. Data intensive workloads.
5. Software-as-a-Service applications.
6. Online storage solutions.
7. Online backup/restore solutions.
8. Isolated workloads where latency between application components is not an issue.

Application Workloads not Ready for Public Cloud Today

Workloads that depend on sensitive data, normally restricted to an organization, are public today. Most companies are not ready to move their LDAP server to a public cloud because of the sensitivity of the employee information. Health care record – until the security of the cloud provider is well established – is another example. Some other examples include:

1. Workloads composed on various, dependent services.
2. High throughput online transaction processing.
3. Workloads requiring a high level of auditability and accountability.
4. Non-virtualized workloads and non-availability of cloud-based licensing strategy.
5. Workloads looking for complex service accounting mechanisms for different services for various departments based billing.
6. Workloads requiring flexibility and customization.

2.4.2 Private Clouds

Private clouds are deployments made inside the company's firewall (on-premise datacenters) and traditionally run by on-site servers. Private clouds offer some of the benefits of a public cloud computing environment, such as elastic on-demand capacity, self-service provisioning, and service-based access. Private cloud is suitable when the traditional requirements, such as control, security, and resiliency, are more emphasized by an organization with the restricted and designated user access and authorization.

Services in Private Cloud

This section highlights the services provided by a private cloud and the services consumed from public cloud, specifically:

1. Virtualization.
2. Government and management.
3. Multi-tenancy.
4. Consistent deployment.
5. Chargeback and pricing.
6. Security and access control.

The services consumed from a public cloud are as follows:

1. Security and data privacy.
2. Ease of access.
3. Discovery of services.
4. Restful interface support.
5. Lower cost.
6. Speed and availability.

High 'Cost of Privacy'

Many experts believe that a private cloud implemented with internal hosting/running of the infrastructure makes it difficult to realize many key benefits of clouds, including

1. **Eliminating capital expenses and operating costs:** Ownership of the hardware or software eliminates the pay-per-use potential, as these must be upfront purchases. The full cost of operations must be shouldered as there is no elasticity. If the private cloud hardware is sized for peak loads, there will be inefficient excess capacity. Otherwise, the owner will face complex procurement cycles.
2. **Removing undifferentiated heavy lifting by offloading datacenter operations:** Utility pricing (for lower capital expenses and operating expenses) usually implies an outside vendor offering on-demand services. It relies on the economies of multiple tenants sharing a larger pool of resources. These higher costs might be justified if the benefits of quicker and easier self-service provisioning and service-oriented access are large.

Private Clouds Provide More Control

In traditional security models, location implies ownership which in turn implies control when security is location-specific. Then location, ownership, and control are aligned. Strong requirements for control and security usually drive a preference for a private cloud, where they own the cloud resources and control the location of those resources. For example, government may not want their applications or data to reside outside certain borders. Clouds rely on virtualization; and in the public model, this loose coupling breaks the link between location and application, and reduces the perceived ownership and control.

When we talk about the information control, it is not related to fixed geography or total ownership of the information. One example is public key encryption – the ownership of the key means control over the information without owning the rest of the infrastructure. The information control can be managed over the infrastructure that is trustful on the basis of

the contracts, regulations, SLAs, standards, and imposition of the security mechanism on service providers. Compliance is difficult outside of traditional security models. As long as control through technology and contracts can be clearly demonstrated, it is possible to make a public cloud computing environment as compliant and as secure as a privately owned facility. Auditors and regulators are continuously adapting to new technologies and business models. Ownership can have multiple avenues as follows:

1. Full-implementation ownership.
2. Lack of full ownership.
3. Controlled ownership.

There are many possible approaches in between, such as partial control and shared ownership. There are also different levels of limited access – specific departmental access, industry-only access, and controlled partner access.

2.4.3 Hybrid Clouds

A hybrid cloud is a combination of an interoperating public and private cloud. This is the model where consumer takes the non-critical application or information and compute requirements to the public cloud while keeping all the critical information and application data in control. The hybrid model is used by both public and private clouds simultaneously. It is an intermediate step in the evolution process, providing businesses an on-ramp from their current IT environment into the cloud.

It offers the best of both cloud worlds – the scale and convenience of a public cloud and the control and reliability of on-premises software and infrastructure – and let them move fluidly between the two on the basis of their needs. This model allows the following:

1. Elasticity, which is the ability to scale capacity up or down within minutes, without owning the capital expense of the hardware or datacenter.
2. Pay-as-you-go pricing.
3. Network isolation and secure connectivity as if all the resources were in a privately owned datacenter.
4. Gradually move to the public cloud configuration, replicate an entire datacenter, or move anywhere in between.

2.4.4 Community Clouds

This is the cloud managed by groups of people, communities, and agencies especially government to have the common interests – such as maintaining the compliance, regulation, and security parameters – working on the same mission. The members of the community share access to the data and applications in the cloud.

2.4.5 Shared Private Cloud

This is a shared compute capacity with variable usage-based pricing to business units that are based on service offerings, accounts datacenters. It requires an internal profit center to take over or buy infrastructure made available through account consolidations.

2.4.6 Dedicated Private Cloud

Dedicated private cloud has IT Service Catalog with dynamic provisioning. It depends on standardized Service-Oriented Architecture (SOA) architectural assets that can be broadly deployed into new and existing accounts and is a lower-cost model.

2.4.7 Dynamic Private Cloud

Dynamic private cloud allows client workloads to dynamically migrate to and from the compute cloud as needed. This model can be shared and dedicated. It delivers on the ultimate value of clouds. This is a very low-management model with reliable SLAs and scalability.

2.4.8 Cloud Models Impact

Clouds will transform the IT industry and profoundly affect how we live and how businesses operate. Cloud computing

1. Offers the scalable compute model to be accessed from anywhere.
2. Simplifies service delivery.
3. Provides rapid innovation.
4. Provides dynamic platform for next generation datacenters.

Some say it is grid or utility computing or Software-as-a-Service, but it is all of those combined.

Public Clouds: Benefits

There are various benefits of public clouds. Following are some of the benefits of public clouds:

1. No big upfront investments.
2. Offer self-service for rapid-start development.
3. Deliver new pricing models for hardware, software, and service consumption.
4. Flex-up and flex-down capacity in short span.
5. Demonstrate proof of concept (POC), collaboration, workload management.

Internal Private Clouds Drive Cost Savings

There are significant cost savings in implementing an internal private cloud instead of using a usual traditional infrastructure. With a traditional infrastructure, each server typically runs a single application and the hardware is sized to meet peak demands. This setup leads to very low average hardware utilization and high software costs due to the number of servers that are deployed and the lack of resource sharing. The internal private cloud uses virtualization on larger servers and leverages advanced service management capabilities to drive efficiency. The servers can be dynamically provisioned to adjust to workload changes and end-users can request the services they need through self-service portals, which drive automation.

Significant cost savings can be achieved by leveraging these capabilities to automate test and development environments. Automation drives down IT labor cost by automatically responding to the changes in the environment and taking action before the problems occur. Virtualization coupled with service management greatly improves server utilization and reduces software license costs since fewer machines need licenses. Automated provisioning and standardization

allow systems to be provisioned within minutes by scripting the install process. In addition, end-users can now interface with IT through self-service portals to request services similar to ATMs that are leveraged to improve banking service. So the virtualization can

1. Reduce IT labor cost by configuration, operations, management, and monitoring.
2. Reduce license cost by utilizing the capital effectively.
3. Lower administrative costs.
4. Reduce end-user IT support costs.
5. Reduce provisioning cycle times from weeks to minutes.
6. Provide benefits of cloud economics with security within your firewall.
7. Provide self-service for rapid-start development.
8. Provide consistency of application environments.

2.4.9 Savings and Cost Metrics

The use of virtualization in cloud computing consolidates systems, which drives reductions in hardware costs. This is often the initial appeal of funding virtualization projects.

Savings on manual efforts are also bigger. Today enterprises are involving human resources to provision the compute resources, which require longer cycles sometimes even week, hence they suffer cost. In this way, highly skilled resources are doing the administration work and not focusing on important higher value jobs. This can be avoided by the automation of the tasks that are highly repetitive. This automation can save good amount of labor cost while removing the errors, enhancing agility, and improving quality.

Cloud computing features two delivery models, namely, private cloud computing and public cloud computing. Private cloud computing exists behind the firewall, while public cloud computing is accessed through the Internet. Cloud vendors believe that these three models – traditional IT, private cloud services, and public cloud services – can co-exist as the part of an overall strategy, based on the application type and the business need that would dictate which model to use.

Hybrid clouds services, delivered to the end-user, are composed of both private and public cloud computing elements.

2.4.10 Commoditization in Cloud Computing

When businesses started taking advantage of IT, the first organizations to computerize their business processes had significant gains over their competitors. As the IT field matured, the initial competitive benefits of computerization fell. Computerization then became a requirement just to stay on a level of the playing field. In essence, there was an increasing amount of IT that operates as a commodity.

For example, a paper products company needs a certain amount of unique IT to run its business and make it competitive. But it also runs a huge amount of commodity IT. The commodity 'technology' takes time, money, people, and energy away from their business of producing quality paper products at a competitive price.

As executive management realizes that the company is operating a lot of commodity IT, which is not core to their competency, the debate shifts from whether cloud computing will

take hold on the enterprise to a debate on how much of the organizational IT will be left internal, on-premise. For this process, IT functions should be evaluated and a determination be made of what is a 'commodity' and what is not. Then it also needs to be determined where to place that function in the new IT organization.

2.5 Security in a Public Cloud

Let us now discuss some of the security concerns that should be considered for the cloud deployments.

2.5.1 Multi-Tenancy

As long as the cloud provider builds its security to meet the higher-risk client requirements, all of the lower-risk clients get better security than they would have normally. A bandage manufacturer may have a low risk of being a direct target of malfeasants, but a music label that is currently using file sharers could have a high risk of being targeted by malfeasants. When both the bandage manufacturer and the music label use the same cloud (multi-tenancy), it is possible that attacks directed at the music label could affect the bandage manufacturer's infrastructure as well. So the cloud provider must design the security to meet the needs of the music label – and the bandage manufacture gets the benefits.

2.5.2 Security Review

As the time passes, organizations become lenient with their security policies. In order to tackle this, cloud service provider should conduct regular audits, review, and assessments for the security. This should be done by security specialists who are able to identify the issues and fix them. The report should be provided to each client immediately after the assessment is performed so that the clients know the current state of the overall cloud's security.

2.5.3 Mutual Risk

There can be a situation where the cloud service provider may not be the cloud operator, but providing a value-added service on top of another cloud provider's service. Like somebody wants to offer the SaaS-based services, it is good to lease the infrastructure of an IaaS provider and offer the SaaS-based services instead of building the infrastructure from the scratch. In this way, the tiers of IaaS and SaaS are developed on top of each other. In this setup, there is a risk associated to each operator and service provider and it is shared among them. They share the security risks at different levels. Therefore, a holistic risk mitigation plan should be devised to suit the architecture of the cloud provider.

2.5.4 Employee Physical Screening

In this outsourcing world, it is common for the organizations to hire the contract services. Same thing works for the cloud service providers. Like regular employees, contract employee background verification should be done by a third party for cloud service provider.

Service provider should publish its policy to all type of employees and report should be generated for the employees once the background verification is accomplished. This screening establishes the trust between the user and the service provider.

2.5.5 Multi-Geographical Datacenters

Disasters, whether manmade or natural, are part of life. They can be hurricanes, earthquakes, fire, or cable cuts. In practice, the cloud is a reliable model as it is not based on single or one location-based datacenter. Cloud datacenters are distributed and hence less prone to disasters. But sometimes organizations sign up the public cloud services for one location only. In this case, it is more important for the providers to test their disaster recovery option as they are heavily tied with SLAs and penalties. At the same time, organization as a consumer should also check and test the disaster recovery options with mock drills of failover.

2.5.6 Physical Security

Physical threats are also important to be analyzed when opting for cloud services from a provider. There are various points to be analyzed:

1. Whether all the facilities of the cloud provider have the same level of security?
2. Is it possible that only one site is secured and there is no information available for the data residency?
3. Whether datacenter is having the all the necessary physical security components such as biometric access, surveillance cameras, logbook, escorts, and automatic alarms?

2.5.7 Regulations

If any of the service provider says that they never had a security issue, it means they are either misleading or not aware of the consequences of the incidents. So all cloud service providers should have a special task force for any incident response based on the policies and regulations. These policies should be shared with the end customers also.

2.5.8 Programming Conventions

Whether it is IaaS, Saas, or PaaS, cloud providers still use their own software that may be prone to security threats and bugs. It is recommended to the cloud providers to use the secure coding and programming practices. It should be written based on standards that are well documented, reviewed, accepted, and adhered.

2.5.9 Data Control

Today in the security arena, an organization's greatest risk is data and information control. All governments and corporate organizations have laid down compliances and regulations to handle the situation.

Therefore, the cloud service provider should be able to adhere to the guidelines laid by the region or agency. The cloud provider should own the policies to meet the regulation and compliances. There should be strong encryption mechanism for the in-flight data. The cloud

provider should also call for a rigorous risk assessment of the data at least once a year. The report should be published periodically once the audit and assessment are completed. The cloud provider should maintain the security incident policy.

2.6 Public Versus Private Clouds

A public cloud is a shared cloud computing infrastructure that anyone can access. It provides hardware and virtualization layers that are owned by the vendor and are shared between all customers. It is connected to the public Internet and presents an illusion of infinitely elastic resources.

Initially, it does not require upfront capital investment in the infrastructure. For consumption-based pricing, the user pays for the resources used, allowing for capacity fluctuations over time.

Now the new request-based provisioning is done on the basis of self-service portal. If one applies economies of scale, service provider can save on cost. This operation cost is also included in the chargeback models. Here consumers do not directly have any say in the SLAs and the data also moves outside the organization firewalls. The service provider location distance can also create problems because of the bandwidth and latency issues.

A private cloud is a cloud computing infrastructure owned by a single party. It provides hardware and virtualization layers that are owned by or reserved for the business. It, therefore, presents an elastic but finite resource and may or may not be connected to the public Internet.

2.7 Cloud Infrastructure Self-Service

The cloud infrastructure has to be provisioned and paid up-front in private clouds. Users pay for the actual compute usage, allowing the spikes over the period of interval. Self-service provisioning of infrastructure capacity is only possible up to a point in private clouds. Standard capacity planning and purchasing processes are required for major increases. For a large, enterprise-wide solution, some cost savings are possible from providers' economies of scale. The enterprise maintains ongoing operating costs for the cloud, and the cloud vendor may offer a fully managed service (for a price). SLAs and contractual terms and conditions are negotiable between the cloud vendors and the customers to meet specific requirements.

The sensitive data and information stay behind the organization's firewall and there is no connectivity to other clouds. Private clouds can also be designed on the basis of preferred and supported operating systems, application, and usage and business use cases. It can be based on the geographic location, compliance, and regulations. Actually , there is no best model or right model for any organization. It depends on the use, organizational policies, application behavior, geographic location, and government compliances and regulations. Likewise public cloud can be a good option for the testing and development cloud as the provisioning requirement in the development and test environment is very rapid and for a shorter period of interval. It requires to decommission the service frequently. Also for this environment, SLAs are not very stringent. At the same time, the private cloud can be a good option when the environment requires lots of information and data controls, and the sizing is known with the capacity forecast and resource availability.

Cloud computing employs a structured technique to holistically leverage IT industry best practices to uncover areas of relative strength and weakness across multiple IT domains (strategic alignment, computing system and storage, applications and data, processes, organization, finance/environment, and network) to determine readiness for a cloud computing deployment.

Infrastructure strategy and planning for cloud computing gear the clients who are looking for assistance in understanding the business value that the cloud computing model can bring. It is designed to help the clients evaluate their readiness for cloud computing and possible cloud computing usages within their infrastructure. The goal is to develop a high-level vision strategy, value case, and roadmap for cloud computing.

2.7.1 Infrastructure Strategy and Planning Features

The strategy and planning has the following three major features:

1. Evaluation to know the gaps, readiness, and strengths of the existing environment.
2. Development of the value proposition for cloud computing in the enterprise.
3. Strategy, planning, and roadmap to successfully implement the selected cloud delivery model.

Cloud architected solutions have introduced new set of characteristics, like scalability and unique service delivery model, for consistent delivery. They help to reduce the capital investment and operational cost with meeting the high SLAs. It heavily relies on the services that make the environment highly available and redundant all the time in case of disaster whether natural or manmade.

It should be ensured by the administrators that minimum data should be impacted while taking the backup to face the challenges of hard numbers of recovery point objectives and downtime impact should be minimum for recovery time objectives after the outage recovery. Emerging models, where users can have access to applications or compute resources from anywhere with their connected devices through a simplified user interface (UI), are best suitable alternatives for the ease of use. These models help applications to reside in massively scalable datacenters where compute resources can be dynamically provisioned and shared to achieve significant economies of scale. The 'pay-as-you-go' usage model enables the users and companies to predict and manage expenses, reduce costs, and simplify operations better.

2.7.2 Cloud Computing Steps

The process from virtualization to cloud computing can have the following phases:

1. **Stage 1 – Server Virtualization:** Companies usually start virtualization as a consolidation attempt. The focal point tends to be reducing capital expenses (such as server, storage, and networks), reducing energy costs, and perhaps avoiding or delaying a datacenter build-out or move.
2. **Stage 2 – Distributed Virtualization:** Once companies start down the virtualization way and start to achieve capital expense improvements (such as server, storage, and networks), the next focus tends to be on elasticity, operational improvements, rapidity, and organizing downtime more efficiently.

3. **Stage 3 – Private Cloud:** Once the processes are designed for alacrity and the standards are in place to enable broad automation, the company is ready to look at introducing self-service capabilities based on the virtualization architecture.
4. **Stage 4 – Hybrid Cloud:** Private clouds will not be the only answer for any enterprise. The self-service portals and interface introduced by private clouds should enable the IT enterprises to leverage public cloud services when they make logic without affecting the end-users.
5. **Stage 5 – Public Cloud:** Virtualization is not the must thing or the stepping stone before companies use public cloud services. Actually, some companies will attempt with the cloud in the public cloud arena first, and use their lessons to establish private clouds for their enterprises.

2.8 Summary

We have discussed several models for cloud computing, including private clouds (where the deployment is within the organization's firewall) and public clouds (where the application services and data are hosted by a third party outside the firewall). Consistent data availability and security are critical success factors for any cloud deployment. Businesses need to ensure that data are adequately protected and can be restored in a timely fashion following any disruption.