Chapter 7

Service Fulfillment

Upon completing this chapter, you should be able to understand the following:

- Cloud service fulfillment (cloud service provisioning) using ITIL processes
- Steps involved in cloud service provisioning on an end-to-end basis
- Service orchestration/service automation
- Cloud functional reference architecture

This chapter describes the details of cloud service fulfillment, also referred to as cloud service provisioning. Service fulfillment is responsible for delivering products and services to the customer. This includes order handling, service configuration and activation, and resource provisioning. In Chapter 6, "Cloud Management Reference Architecture," two cloud reference architectures are covered from a management perspective. This chapter will provide details on cloud service fulfillment and an end-to-end logical functional architecture for managing clouds. The end-to-end logical functional architecture is built based on the Tele-Management Forum (TMF) eTOM (enhanced Telecom Operations Map)[1] and the Information Technology Infrastructure Library (ITIL) V3 life cycle.[2]

Cloud Fulfillment Using ITILV3

ITIL V3 provided the IT life cycle processes: service strategy, service design, service transition, service operate, and Continuous Service Improvement (CSI). Applying these processes is a good way to establish service provision processes for data center/virtualization (DC/V) and cloud provisioning. Figure 7-1 shows cloud service provisioning flow based on ITIL Version 3.

Building data centers that are capable of providing service on demand, at scale, and with multitenancy requires principles of cloud computing and transformation from the current operational environment (current operational state) to the cloud environment (target operational state). Cloud management should be seen as a life cycle rather than an IT

providing product support. These principles are well articulated in ITIL V3. Figure 7-1 shows the five phases of the cloud ITIL V3 service life cycle:

1. Service strategy
2. Service design
3. Service transition
4. Service operate
5. Continuous Service Improvement (CSI)

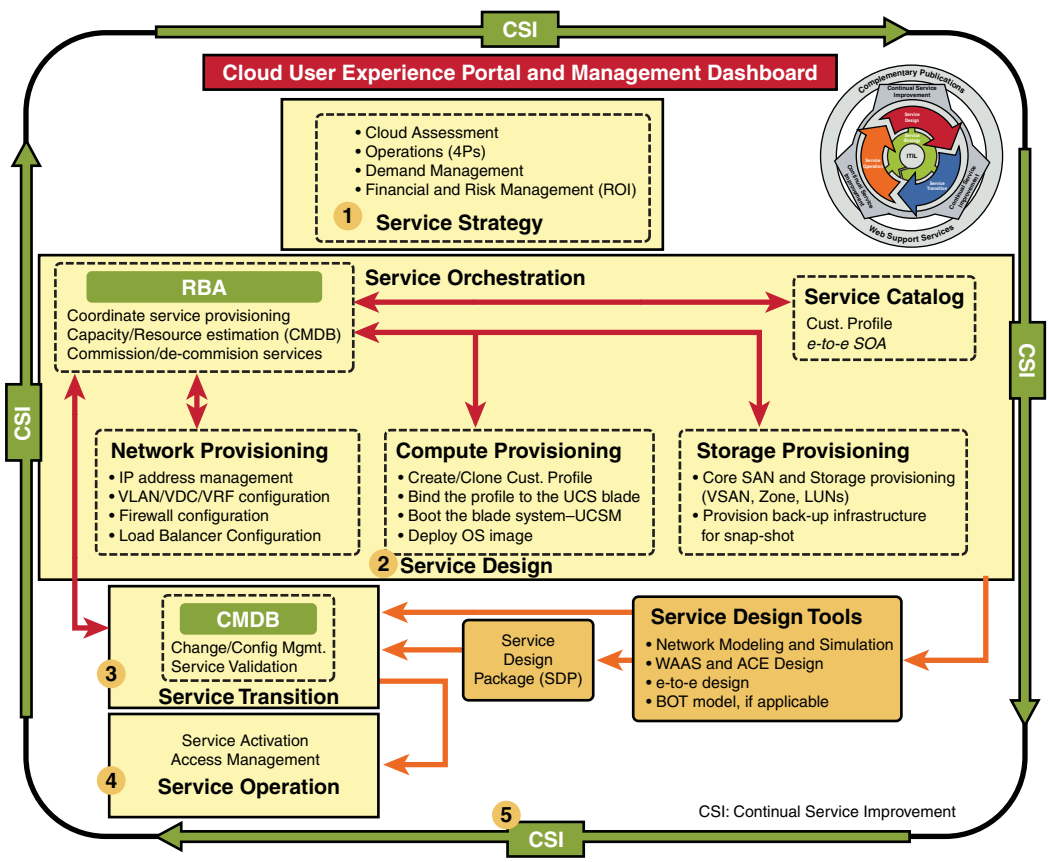

Figure 7-1 *Cloud Service Provisioning Flow Based on ITIL V3*

Figure 7-1 also shows some of the items that need to be considered in each of the five phases of the cloud service life cycle to provision a cloud. More details are provided in the following sections describing the ITIL V3 phases.

Service Strategy Phase

Data center/virtualization (DC/V) and cloud-computing technologies can have a significant impact on IT service delivery, cost, and continuity of services, but as with any transformative technology, the adoption is greatly influenced by the up-front preparedness and strategy.

In a dramatic change from just a few years ago, the role of CIO has changed from keeping the lights on to becoming a strategic thinker and transforming the IT department from a cost center or a commodity center to a strategic value provider. The CIOs are being invited into board rooms for strategic planning, and with the increased visibility comes increased responsibility to bring both top-line and bottom-line business value. When CIOs lack voice in the boardroom, their job becomes more of a keep-the-lights-on position and eventually becomes victim to budget cuts and outsourcing many of the IT functions. IT governance can help the CIOs to become the agent of change and be an active partner in laying out the company's strategy. The IT organization's success factors include the following:

- Technology decisions driven by a business strategy (not the other way around)
- Sustaining the IT activities as efficiently as possible
- Speed to market
- Technology architecture aligning with the business initiatives, and technology is at the heart of value proposition

With the preceding principles in mind, we will develop high-level tasks that are needed in each of the five areas of ITIL V3. During the service strategy phase, the following items should be considered at a minimum to be successful in cloud services:

- Cloud architecture assessment
- Operations (people, processes, products, and partners [the 4Ps])
- Demand management
- Financial management or value creation (ROI)
- Risk management

These items are discussed further in the sections that follow.

Cloud Architecture Assessment

Enterprises today are facing challenges to meet increasing demand for automating the services and how best to meet the demand as the business evolves. The cloud architecture assessment looks into current architecture and technology choices (with a focus on the future) to determine the most appropriate cloud strategy, architecture, and operations management. The architecture assessment analyzes current tools and new tools required for automation, current operating practices with an eye for improving the operations of

cloud management, demand management, financial management, and risk management. This architecture assessment helps make sure that the customer's updated infrastructure meets reliability and capacity goals and can scale to meet future requirements. The salient points of cloud architecture assessment are as follows:

- **Current tools and architecture:** This depends on whether the customer is planning to move to clouds from a greenfield or a brownfield. If it is a greenfield, it would be easier from a product vendor perspective because the product vendors do not have to deal with the legacy equipment and all the integration issues with the legacy equipment. For a brownfield, more work is involved. It is possible that the customer infrastructure might have only parts of the network (compute and storage) or all the parts, and it is important to see how the customer is operating various parts of the infrastructure. Operating the infrastructure as silos using disparate systems is not only inefficient technically, but also operationally. So, it is possible that sufficient savings can be achieved in Operational Expenditure (OPEX) with changes in the management and operations architecture. Also, operating the infrastructure as silos is very inefficient because it would require so much coordination between the siloed operations centers that it would take much longer to isolate faults in the network, compute, and storage, all of which affect the service-level agreements (SLA) offered to the customer. These SLAs are always based on service and not individual parts of the cloud infrastructure.
- **Cloud-provisioning tools:** The provisioning assessment of tools should include, at a minimum, service portal, service catalog, Configuration Management System/Configuration Management Database(CMS/CMDB), service automation, and domain configuration tools. Depending on the size of the company and the type of the company (enterprise or service provider), investigate whether tools should be purchased or leased. Also, consider whether to host in the enterprise data center, in the enterprise private cloud, in the service provider cloud, or in Software as a Service (SaaS) provider clouds.
- **Replacement of tools:** This should be done carefully and can be achieved only by cooperatively working with the operations personnel. Operations personnel are used to legacy systems and would resist changes because they require learning and changes in operations. So, care should be taken to recommend replacement only if it is necessary.
- **Addition of new tools:** The need to add new tools to address gaps in the current architecture and new cloud services should be addressed. The new systems selection should be based on several factors, including ease of integration with the existing systems, open APIs, implementation costs, license costs, integration costs, and support from the vendor after it is implemented.
- **Security:** Security is a big concern for enterprises and service providers who want to move to the cloud. The business requirements are very important here during the assessment phase. Identity management (IDM) in cloud computing is a nebulous application for most enterprises. Although cloud efforts promote cost savings and management efficiencies, it all boils down to trust. Federating identity management

might make sense in a cloud environment where users might be logging in to applications within and outside of firewalls. Authenticating every user will come at a cost, because it would involve constant password resetting and more calls to the help desk. Security is more than identity management and is discussed in detail in Chapter 4, "IT Services," and Chapter 5, "The Cisco Cloud Strategy."

- **Identity management (IAM):** Some SaaS applications might be more cost effective for identity management because they are designed for efficiency, rapid time to value, and minimal disruption, and they can be rightsized. Also, a full-blown suite might not be necessary when all you need is a small subset of functions. This would keep client costs down from the perspective of both monthly services and professional services. For example, having an in-house IAM product could cost a total of $700,000 to $1 million (acquisition + license and connectors + implementation costs + administrator costs + infrastructure [servers and so on]).

All the preceding items should be documented in Current State Architecture (CSA) and Target State Architecture (TSA) documents. The CSA and TSA documents that document a customer's architecture become a blueprint of the customer architecture, as they provide documentation for all the changes and the rationale for making those changes.

Operations People, Processes, Products, and Partners (4Ps)

The move to cloud computing is well under way, and companies are investing for rolling out cloud services. The basic value proposition for cloud computing is straightforward: Users can leverage a wide range of computing resources without the capital investment or maintenance infrastructure necessary to build and maintain these services internally. However, companies considering a move to the cloud must understand whether and how cloud services might fit into their IT strategy and operations. IT operations is responsible for delivering the agreed-upon level of IT services to the business and to maintain the SLAs, even if the infrastructure is in an external cloud. Although not owning infrastructure provides Capital Expenditure (CapEx) benefits which is one off-benefit in CapEx, making operations more efficient improves the Operational Expenditure (OPEX), which is a benefit enjoyed year over year. As a part of a strategy, service providers should look into IT operations and make changes in operations to support cloud-based services. The following list provides some insight into operations with an eye toward cloud management:

- Current operations processes and products might need improvement based on new services and support. Many of the customer's organizations are established in silos. Operations processes need to be improved so that the organizations are more agile and inter-connected to operate new cloud-based services. Speed to market and speed to react are essential for service delivery and service support.
- Current methods and procedures should be checked and improvements/adjustments made to provide service delivery and service support for cloud services. Figure 7-2 shows the transformation to a virtualized infrastructure management. Today, some of the IT organizations might be organized in silos to support network, compute, and storage services. ITIL V3 teaches us that products and services should be not only fit

for purpose (utility) but also fit for use (warranty). In the cloud context, many organizations might be providing services "fit for purpose" for network, compute, and storage silos, as shown on the left side of Figure 7-2. For cloud management, it is essential to operate in a holistic way with the overall business in mind, as shown on the right side of Figure 7-2. Cisco's operations assessment service reviews current operations, and provides recommendations and roadmap for transforming the current operation into future-state operations required for managing cloud based services.

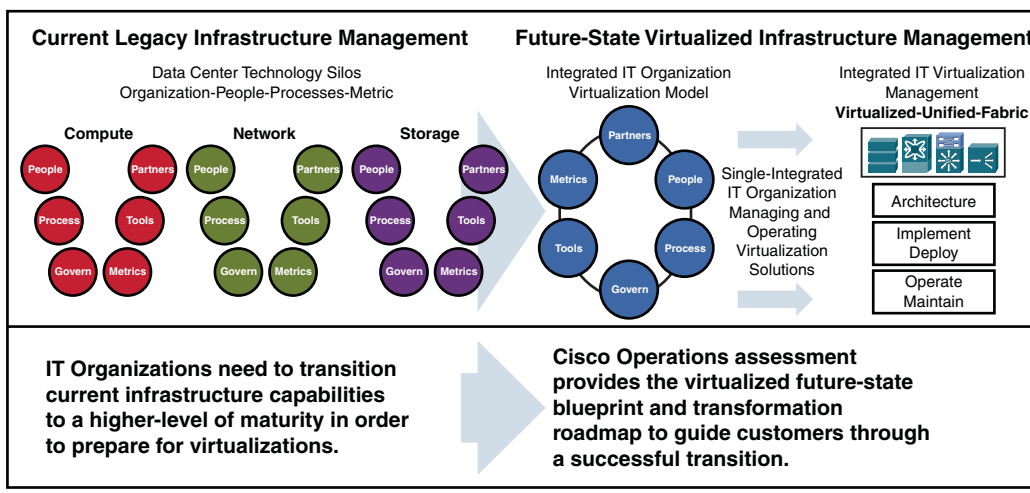

Figure 7-2 *Transformation to Virtualized Infrastructure Management*

- Figure 7-3 shows how the functional areas of the enterprise (people, processes, organization, and governance metrics) might be addressing products/services in a siloed and nonvirtualized manner. The Cisco operations assessment reviews these enterprise functional areas and provides recommendations to transform from a nonvirtualized to a virtualized environment.
- A partner's capability should be checked for its capability to manage the new services, and any gaps should be addressed through training, or replace the current partner with a new partner that has the required capabilities. The SLAs offered depend not only on contracts with the internal organization, but also with the external partners.

Functional Areas	Non-Virtualized		Virtualized
People: • Skills • Roles • Responsibilities •Training	• Technology Specific Silos • Deep Domain Expertise • Point Technology Staffed • Silo Career Path Training • Limited Cross-Technology Collaboration	**Transform To**	• E2E Services Focused • Consolidated I/O–FCoE • Compute-Network-Storage Skills • New Role–Virtualization Architects • New Role–Virtualization Engineers
Process: • Operation Management • Availability Management • Performance Management • Testing and Deployment • Architecture Planning	• Box Based Provisioning • Poor Process Integration • Silo Point Technology Driven • Not Well Documented–Understood • Throw-Over Wall Approach • Point Technology Defines Tools	**Transform To**	• Intelligent Software Based • Well Orchestrated Procedures • Integrated SLA and OLA • IP Management (Repository) • Shared Services Driven • Integrated Fabric Tool-Suite
Organizations: • Structures • Departments • Charters • Reporting • Culture	• Hierarchical • Department Technology Specific • Technology Funded–Silo • Technology MBOs–Metrics • Poor External Communication	**Transform To**	• Services Centric • Shared MBOs • Shared Services • Cross-Functional • Highly Collaborative
Governance–Metrics: • Architecture Standards • Virtualization Patterns • Security Policies • Financial Management • Metrics Monitoring	• Ad-Hoc • Limited Documentation • Silo-Not Shared • Limited Compliance • Not Measured/Monitored	**Transform To**	• Virtualization Standards • Consolidated I/O Guidelines • Cross-Functional Managed • Well Define Metrics • Architecture Review Board

Figure 7-3 *Transformation of Functional Areas: Nonvirtualized to Virtualized Environment*

Demand Management

Service management is faced with the task of finding a continual balance between consumption and delivery of resources. Demand management calculates this demand for service capacity and controls the necessary capacity with the expected flexibility. Demand management is a critical aspect of service management. Unlike goods, services cannot be manufactured, stored, and sold at a later time. Poorly managed demand is a source of risk for service providers because of the uncertainty in demand. Excess capacity generates cost without creating value, and customers do not like to pay for idle capacity unless it has value for them. Demand management includes the following important aspects:

- At a strategic level, cloud demand management involves determining Patterns of Business Activity (PBA). A PBA is a workload profile of one or more business activities, and it helps the service provider to understand and plan for the different levels of business activity. Understanding the customer's PBA, such as when he watches cable TV and when he accesses the Internet, would be important activities of the demand management process.
- At a tactical level, cloud management can involve different charging mechanisms based on service levels, and also encourage users to use service at less-busy times and provide incentives for using it. This is similar to how the telephone companies charge for telephone calls, often charging lower rates during off-peak hours on weekdays and weekends and charging more during business hours on weekdays. The charging

methods vary between private and public clouds, and Chapter 9, "Billing and Chargeback," provides more details on billing and chargeback for public and private clouds.

Financial Management and Business Impact

Before embarking on cloud services, one of the key steps is performing service value creation, service investment analysis, and service business impact analysis. The following steps will help in the financial management and value creation:

- **Service valuation:** This determines whether the service differentiation results in higher profits or revenue, lower costs, or better adoption of the services.
- **Service investment analysis:** This provides investment analysis for the stakeholders. Some SaaS applications might be more cost-effective because they are designed for efficiency, rapid time to value, and minimal disruption, and they can be rightsized. Consequently, companies face the challenge of determining between the two options: in-house development (packaged software deployed internally) versus the SaaS model (deployed and sourced by an external vendor).
- **Business impact analysis:** The cloud model introduces some pro and con business impacts, and that should be considered as well.
- **Some pro-business impacts include the following:**
  - Access to subject matter experts that are not available in-house
  - Automated upgrades
  - Ability to move service complexity off-site
  - Dynamically source and consume IT services
- **Some con-business impacts include the following:**
  - The boundary of service has moved from internal to external, which might result in support issues not being addressed properly and might affect customer support SLAs.
  - Cloud services can contribute to performance issues because of elasticity changes with demand fluctuations.
  - The capability of cloud providers to provide the same level of SLAs as the enterprise to which customers have become accustomed and expect.

Risk Management

The *risk* is defined as an uncertainty of outcome, whether a positive opportunity or a negative threat. A risk is measured by the probability of the threat happening. Risks can come from uncertainty in financial markets, project failures, accidents, natural causes, and disasters, as well as from deliberate attacks from an adversary. The strategies to

manage risk include avoiding the risk, reducing the negative effect of the risk, and accepting some of or all the consequences of a particular risk. Unfortunately, risks cannot be totally avoided. The main areas of risk in the cloud include security threats, failure of equipment, and the inability to provide services to customers. The security threats can be avoided by having security all over, and failure of equipment and, in return, failure of service can be avoided by having a strategy for business continuity through disaster recovery (BCP/DR). Risk management covers a wide range of topics, including

- Business continuity process (BCP) and disaster recovery (DR)
- Security
- Program/project risk management
- Operational service management

These topics need to be supported by a risk management framework that is well documented and communicated throughout the organization.

Service Design Phase

The service design provides guidance on the design and development of cloud services and for converting strategic objectives into a portfolio of services and service assets. It includes changes and improvements necessary to increase and maintain value to the customer over the entire life cycle.

During the service design phase, the following items should be considered, at a minimum, taking input from the service strategy phase:

- Service catalog management
- Orchestration
- Security design
- Network configuration and change management (NCCM)
- Service-level agreements (SLA)
- Billing and chargeback

The following sections describe these considerations for a cloud design in greater detail.

Service Catalog Management

Service catalogs have been around for decades, and ITIL books had them for many years. However, service catalogs have been used only by the service providers so that they can be paid for the services rendered to their customers. With the advent of cloud

computing, any cloud provider has instantly become a service provider and hence needs a service catalog. Amazon EC2 provides a service catalog to order virtual machines (VM) that can be provisioned in a matter of minutes. This saves lots of money in provisioning time. A good cloud service catalog should consider the following:

- **Elastic:** It allows increasing or decreasing the required capacity through a self-service portal and it is provisioned in minutes, not hours and days. One can order one instance, hundreds, and even thousands of server instances in minutes, all done at the click of button on a portal.
- **Self-controlled:** The user should have complete control and interact with the service catalog remotely using self-service portals (Web Services API).
- **Flexible:** It allows the user to select memory, CPU, and instance storage space. The operating system choice should include Linux, Microsoft Windows, and Solaris.
- **Reliable:** The service runs with proven network infrastructure and data centers and should offer a highly reliable environment where replacement instances can be rapidly and predictably commissioned.
- **Secure:** It offers an interface to configure firewall settings that control network access to and between groups of instances.

All the aforementioned features are offered today by Amazon EC2 cloud services. In addition, many service providers might require additional customizations in the service catalog and might need the following:

- **Firewall service options:** Some cloud providers might want to offer additional firewall services that can provide Low (L), Medium (M), and High (H) security options with various pricing options for users. In addition, some end customers might choose to configure the firewalls themselves, and cloud providers might want to provide that option in addition to the L/M/H security options. Note that a customer can choose Gold, Silver, and Bronze service levels, as mentioned in other parts of this book, and still be able to select L/M/H security options in the service catalog for each of the Gold/Silver/Bronze service levels.
- **Load balancing:** A cloud-based load-balancing service that allows the cloud provider to manage the content based on service delivery policies based on real-time conditions and user targets. This empowers the service provider to react to market-specific conditions without compromising availability, performance, and operational efficiency. The traffic management could be done dynamically so that the traffic can be moved based on the user requirements. For example, all traffic generated in the United States could go to servers in the United States, and all other traffic could go to servers in Europe and Asia.

Orchestration

Orchestration is important in service activation and interfaces to service catalog, CMS/CMDB, and the respective domain managers to activate a service. Many vendors, including Cisco, offer orchestration systems, and most of them allow making changes in the workflow to meet the customer's requirements. Orchestration need to ensure that the workflow is seamless and interfaces to all the required parts of the organization (tools, processes, and so on). More on orchestration is described later in this chapter.

Security

Security should be designed both in the network (firewall locations, access control lists, and port security) compute, storage, and access. Basically everywhere. The authentication and authorization should be designed as part of all service offerings so that the applications can only be accessed by the users that are authorized and entitled to access the services. Security is one of the most important areas in the cloud and is discussed extensively in Chapters 3, 4, and 5.

Network Configuration and Change Management

Network Configuration and Change Management (NCCM) plays a key role in the overall management. With DC/V and cloud, the role of NCCM has expanded to not only network devices but also to compute devices, storage devices, and applications. The NCCM systems should pay attention to some of the following areas:

- Configuration management plays a critical role in change management because detailed maps of the infrastructure devices and the configuration of each device and the connectivity between them are required.
- The topology views of the infrastructure are kept in Configuration Management Databases (CMDB) that contain detailed recordings of the configuration of each component and all updates or changes that have been made along the way.
- It would be ideal to have the CMDB updated automatically whenever changes are made to the infrastructure through audits or periodic polling. If this is not available, the operation would have to manually update the CMDB whenever a change is made.
- Compliance analysis is an important part of NCCM, and many of the tools available provide HIPAA (Health Information Portability and Accountability Act), PSIRT (Product Security Incident Response Team), and other audits and provide alerts whenever the configuration of the devices does not meet these standards. In addition, the device vendors provide field notices and configuration best practices that can be checked against the device configuration, and remediate whenever there is a discrepancy. Cisco SMART services audit and automate the changes without manual intervention.
- In addition to tools providing configuration and changes, the organization should be cognizant of the changes and should have a CAB (Change Advisory Board) and ECAB (Emergency Change Advisory Board) in place to address changes and compliance reports.

SLA

Service-level agreements (SLA) guarantee most aspects of service delivery, both technology and service aspects. Technology guarantees are concerned with system response time, system uptime guarantees, and error resolution time. The customer service guarantees are concerned with availability, support staff availability, and response time. The SLA should be designed through a collaborative effort between the marketing and technology groups. If it is only marketing, the technology and support staff might not deliver the SLAs offered, and if it is only technology, it will be filled so many loopholes and "it depends" that it would not be appealing to customers. Typically, service providers offer five 9s, or a 99.999 percent uptime guarantee. 99.9 percent uptime equals 8 hours, 45 minutes, and 57 seconds of downtime per year, while 99.999 percent uptime equals 5 minutes and 42 seconds of downtime per year. You should make sure that the uptime guarantees include only what the provider and the partner can cover and not what the customer can bring down. As long as you clearly define what you include in your guarantee, you can make aggressive uptime claims like 99.999 percent. However, it is important to make only promises that can kept and avoid making promises that are not in anyone's best interest.

Billing and Chargeback

Billing and chargeback are an important part of providing cloud services; Chapter 9 is dedicated to billing and chargeback. The following are some of the billing and chargeback considerations that should be kept in mind:

- Design of services in the service catalogue should pay attention to billing and charging capability. They should go hand in hand. There is no use in offering sophisticated services if the billing and charging systems cannot accommodate the new way of charging for the services.
- Ensure that proper data collection, metering, and charging systems are in place.
- The cloud providers typically break down their charges into various items such as servers in the cloud, storage in the cloud, applications in the cloud, bandwidth, space, heating, and cooling.
- The cloud pricing structure is based on many other factors as well, including service support, duration of the contract, security load balancing, disaster recovery, and additional charges hidden deep within the SLAs.

Service Transition Phase

The service transition phase implements the service design that was built in the design phase into production at the service provider location. Cloud computing is no shortcut to process maturity. It would require the same processes that are in existence today, but needs to make changes to adopt to cloud computing to deliver cloud services. The service providers must understand their goals and objectives, and without a clear understanding

of the business processes and supporting IT processes they have in practice today, and determining and adjusting what is needed for the future, success will be difficult to achieve. You need to know what to expect and where you hope to gain efficiencies and savings before you dive into the deep end. The following items are considered in the service transition phase:

- **Change management:** This process ensures that changes are recorded and then evaluated, authorized, prioritized, planned, tested, implemented, documented, and reviewed in a controlled manner. The objective of change management is to support the business and IT when it comes to outages and changes. A cloud service provider must ensure that all exceptions to normal operations are resolved as quickly as possible while capturing the required details for the actions that were taken. In a cloud-computing environment, network, compute, storage, applications, and middleware are all connected, and any changes can have an unintended consequence. CMDB is important here, because it provides the interrelationships among all the infrastructure resource configuration items (CI). Knowledge of CI relationships allows changes to be made with authority. This provides more visibility into the cloud environment for the change managers, allowing them to make more informed decisions, not only when preparing for a change but also when diagnosing incidents and problems.
- **Service asset and configuration management:** This process ensures that accurate configuration information of the current, planned, and historical services and infrastructure is available. Because IT organizations are accountable for the quality and SLA of the services offered to the end users, accurate information on the services contracts (SLAs, Operational Level Agreements [OLA], and Underpinning Contracts [UC]) is required and should be maintained.
- **Orchestration and integration:** The design and implementation of orchestration flows and integration into any legacy systems are important, if a provider is moving to cloud computing to offer cloud services. Many orchestration tools allow building process flows and should be built to meet the customer business flows.
- **Migration:** The operations should move from current state to target state (people, processes, products, and partners) to manage the cloud based services.
- **Staging and validation:** The systems and software are validated, and required testing should be completed to transition the systems/software to the provider organization.

Service Operate Phase

The service operate phase is where the service provider takes possession of the management of the cloud operations from the equipment vendors, system integrators, and partners, and will be taking service orders from its end customers. All ITIL V3 phases are important, but the service operate phase draws the most attention because 60–70 percent of the IT budget is spent in dealing with day-to-day operations. To reduce OPEX, it is important that attention is paid in the planning, design, and CSI phases. In the service

operate phase, the service provider not only takes the service orders for service fulfillment but also monitors and audits the service using the monitoring systems to ensure that the SLAs are met. The monitoring portion of operations is discussed separately in the next chapter under "Service Assurance." The following items are considered in the service operate phase:

- Service desk
- Incident management
- Problem management
- Service fulfillment
- Event management
- Access management

Service Desk (Function)

The service desk is an important function in the overall operations because it provides the first point of contact for the customer into the service provider organization. Figure 7-4 shows the service desk functions.

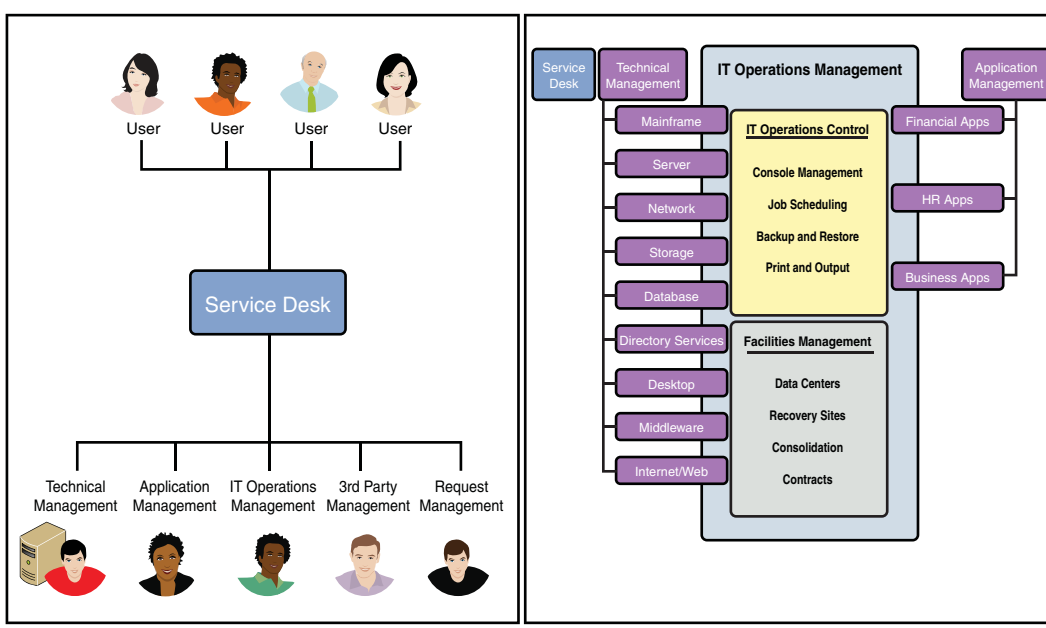

Figure 7-4 *Service Desk Functions*

Depending on the size of the company, service desk functions can be as follows: a local service desk for small organizations to meet local business needs, a centralized service desk for organizations having multiple locations, and a virtual service desk for organizations having multicountry locations. As shown in Figure 7-4, the service desk handles technical management, IT operations management, and applications management that are described in the following list:

- **Technical management:** Provides detailed technical skills and resources needed for the day-to-day operations of the IT infrastructure. The objective of technical management is to help plan, implement, and maintain a stable technical infrastructure to support an organization's business process.
- **IT operations management:** Provides day-to-day operational activities needed to maintain the IT infrastructure and provides service to meet the performance standards designed in the service design phase. Two key areas within the operations function are important:
  - IT operations control is responsible for console management, job scheduling, backup and restore, and reports.
  - The facilities management is responsible for data centers, backup sites, and contracts.
- **Applications management:** This supports and maintains operational applications and plays an important role in the design, testing, and improvement of IT applications that form part of IT services. The object of application management is to support the organization's business processes by helping identify functional and management requirements for applications software and to assist in the design and deployment of those applications and ongoing support and improvement of those applications.

Incident Management

An *incident* is any event that is not part of the standard operation of the service and which causes, or might cause, an interruption or a reduction of the quality of the service. The objective of incident management is to restore normal operations as quickly as possible with the least possible impact on either the business or the user, at a cost-effective price. When an incident happens, the service desk function (operator) follows the incident management process to resolve the incident. Incident management has three aspects to it:

- **Impact:** This gives an indication of the effect of the incident, whether it needs escalation, and whether it would have any effect on the SLA, time, and cost of resolving the incident.
- **Urgency:** This provides an indication of how long until an incident, problem, or change has a significant impact on the business. A high priority might not be urgent if it does not have a financial impact on the business.
- **Priority:** This indicates the relative importance of an incident, problem, or change on the business. Table 7-1 shows the relationship among priority, impact, and urgency.

The numbers 1 though 5 in Table 7-1 represent priority, as follows:

- **1:** Critical
- **2:** High
- **3:** Medium
- **4:** Low
- **5:** Planning

Table 7-1 *Priority, Impact, and Urgency Relationship*

		Impact		
		High	**Medium**	**Low**
Urgency	High	1	2	3
	Medium	2	3	4
	Low	3	4	5

A high-impact item could have a low urgency (say 3) and a medium-impact item could have a higher urgency (say 2). The resolution of the impact should be based on the urgency of the impact. Some high-impact items might not need to be attended to if they do not immediately impact the business.

Problem Management

A problem is a condition often identified as a result of multiple incidents that exhibit common symptoms. Problems can also be identified from a single significant incident, indicative of a single error for which the cause is unknown, but for which the impact is significant. The primary focus of Problem Management (PM) is to identify causes of service issues and commission corrective work to prevent recurrences. PM processes are reactive and proactive—reactive in solving problems in response to incidents and proactive in identifying and solving potential incidents before they occur. Problem management activities include

- Recording, managing, and escalating service problems as appropriate
- Analyzing historical data to identify and eliminate potential incidents before they occur
- Identifying the underlying causes of incidents and preventing recurrences

- Developing workarounds or other solutions to incidents
- Submitting change requests to change management as required to eliminate known problems

Service Fulfillment (Service Provisioning)

This process deals with taking orders from users/customers through service requests. A service request is generated by the user and might be asking for information, a standard change, or a service. The request can come through email, a telephone call, or a web interface. The service requests are handled by the service desk. In a cloud environment, most of the service orders come through web interfaces (self-service portals), and the service requests are provisioned seamlessly using many systems. The detailed step-by-step end-to-end service fulfillment/service provisioning for a cloud is discussed in the next section.

Event Management

An event is a change of state of a configuration item (CI) or an IT service, or an alert created by a CI, IT service, or a monitoring system, and can require operations people to take action. Events could be notifications that indicate a regular operation (such as scheduled workload uploaded, a user logged on to an application, and so on) or an unusual activity (such as a user attempting to log on to an application without the correct password). The event management process collects events from the infrastructure devices, makes sense of them, and determines appropriate action. More on this is discussed in the next chapter.

Access Management

This process allows authorized users the right to use the service while blocking unauthorized users and making sure that the policies and actions defined in security management and availability management are executed properly. The access methods that are applicable for cloud services include the following:

- **Identity management:** The identity of a user is unique to that user and is managed through identity management systems, such as Lightweight Directory Access Protocol (LDAP), LDAP over Secure Sockets Layer (SSL), and TACACS.
- **Access:** The level and extent of service functionality or data that a user is entitled to use.
- **Rights:** Also known as privileges, rights refer to the actual settings where a user is provided access to a service or group of services (for example, read/right/change/delete privileges).

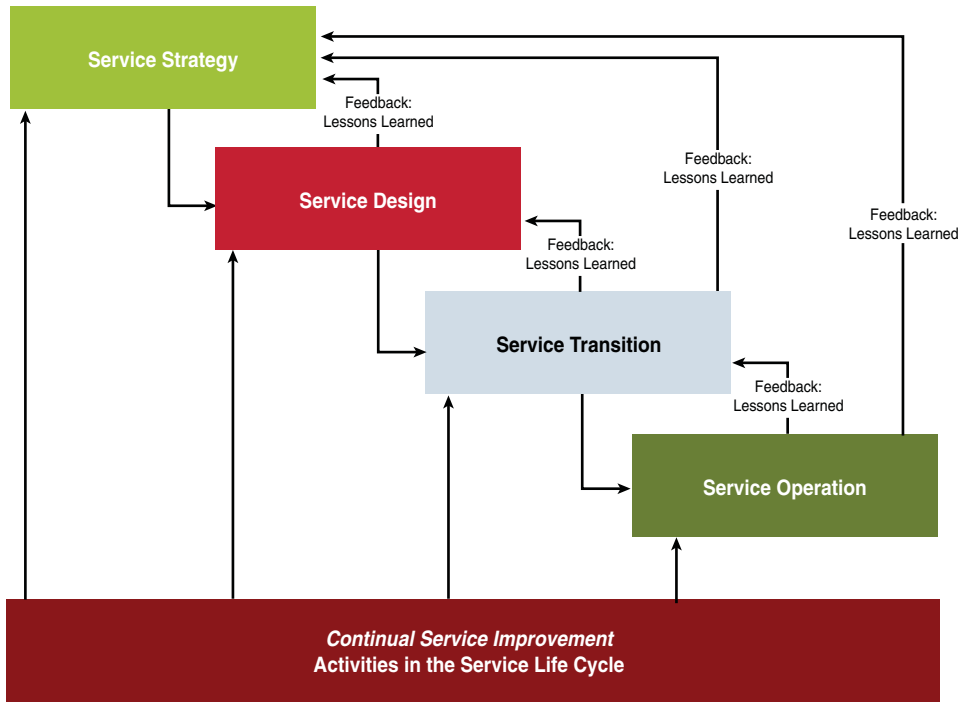

Figure 7-5 *Role of Continuous Service Improvement*

Cloud CSI (Optimization) Phase

The Continuous Service Improvement (CSI) phase is, as its name implies, an ongoing practice to improve the IT organization activities using best practices, as opposed to a reactive response to a specific situation or a temporary crisis. The CSI phase plays a role in all phases of the ITIL life cycle to provide improvements, as shown in Figure 7-5.

Figure 7-5 shows how CSI interacts with service strategy, service design, service transition, and service operation to help improve each of the phases through feedback loops with the lessons learned from each phase. Specifically, the following objectives are met through CSI:

- Review, analyze, and make recommendations on improvement opportunities in each life cycle phase: service strategy, service design, service transition, and service operations.
- Review and analyze service-level achievements.
- Review and improve effectiveness of IT delivery.

- Identify and incorporate best practices/good practices to improve IT service quality and improve the efficiency and effectiveness of enabling ITSM (IT Service Management) processes.
- Ensure that applicable quality management methods support continual improvement activities.

Some of the CSI/optimization-specific activities that can be related to DC/V and the cloud that can be quickly adopted include

- Audit the configurations in all the infrastructure devices in the network, compute, and storage areas (this includes core and access aggregation layers). There could be hundreds—even thousands—of devices, depending on the size of the network. The inventory and collection of data from all devices must be done to ensure the manageability of the cloud infrastructure.
- Compare the configurations against the best-practice configuration and make changes as appropriate to ensure that the infrastructure is up to date.
- The infrastructure devices should be compared with end of life (EOL), end of service (EOS), and field notices and make recommended changes to the infrastructure. Many times, business impact analysis is done to show the importance for the need to make the recommended changes.
- Fine-tune the management tools and processes based on best practices.
- Adding new products and services requires assessment to ensure that the new services can be incorporated into the current operating environment without sacrificing the quality of service to customers.

Cloud End-to-End Service Provisioning Flow

The previous sections discussed all ITIL V3 life cycle processes (service strategy, service design, service transition, service operate, and CSI) for service provisioning (also referred as fulfillment in TMF eTOM). This section will provide end-to-end service provisioning flow. Figure 7-6 shows end-to-end provisioning steps for a customer reaching the service provider through a self-service portal and ordering a web service, and then getting a confirmation from the service provider that the service is ready for use.

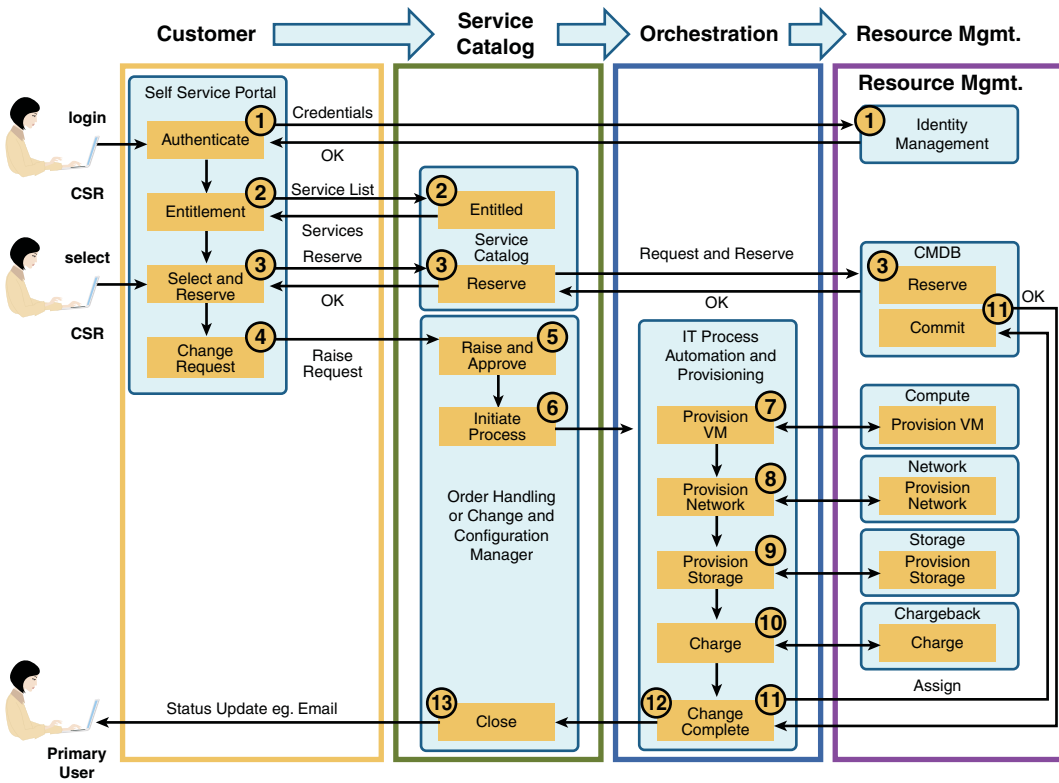

Figure 7-6 *Cloud Service Provisioning Flow Steps (End-to-End)*

The steps illustrated in Figure 7-6 are explained as follows:

Step 1. The user logs on to the self-service portal and is authenticated by the identity management.

Step 2. Based on the user's role and entitlement, the self-service portal extracts a subset of services that the user can order from the service catalog.

Step 3. The user selects a service to provision—perhaps a nonvirtualized web server, for example. Associated to this service is a set of technical requirements, such as RAM, processor, and so on, along with business requirements such as high availability or service-level requirements.

a. The self-service portal will query the Capacity Management System or CMS/CMDB to see whether these technical requirements can be met by existing resources. For a virtual resource, it's likely that more emphasis will be placed on capacity requirements—for example, can my ESX host support another virtual machine?—whereas physical servers will be more inventory based.

b. The resource details are passed back to self-service portal, which displays these resources to the user who selects one, which is then reserved in the Capacity Management System or CMDB.

Step 4. The self-service portal now raises a service request with the service desk which, when approved, will create a service instance in the service catalog and notify the self-service portal. It is assumed that the approval process is automatic and happens quickly; otherwise, the notification step might be skipped. The service request state is maintained in the service desk and can be queried by the user through the self-service portal.

Step 5. The service desk will now raise a request with the IT process automation tool to fulfill the service. The orchestration tool extracts the technical service information from the service catalog and decomposes the service into individual parts such as compute resource configuration, network configuration, and so on.

Step 6. In the case of our "nonvirtualized web server" running on UCS (Unified Computing System), we have three service parts: the server part, the network part, and the infrastructure part. The provisioning process is initiated.

Step 7. The virtual machine running on the blade or server is provisioned using the server/compute domain manager. (There are several domain managers available in this area from Cisco, EMC, BMC, CA, HP, IBM, and other vendors.) The following steps provide information on how to provision a VM for a customer:

- **a.** Create or clone a customer profile with the applicable parameters (UUID, MAC address, IP/subnet, WAN, VLAN, VSAN, adapter properties, and boot policy).
- **b.** Select the blade from the available pool and bind the profile to the blade.
- **c.** Boot the server (done by the Cisco UCS Manager).
- **d.** Deploy the OS image using standard tools.

Step 8. The network, including firewalls and load balancers, is provisioned using the network domain managers. (There are several domain managers available in this area from Cisco, EMC, BMC, CA, HP, IBM, and other vendors.) The following steps provide information on configuring the network and network services, such as firewalls and load balancers:

- **a.** IP address assignment and management
- **b.** VLAN/VDC/VRF configuration
- **c.** Firewall configuration for ACL, ports, and IP
- **d.** Load balancer configuration to map servers to the VIP
- **e.** Wide Area Application Service (WAAS) configuration, if required

Step 9. The storage is provisioned using the storage domain manager. (There are several domain managers available in this area from EMC, NetApp, and other vendors.) The following steps provide information on storage provisioning:

- **a.** Provision VSANs, zones, and Logical Unit Numbers (LUN).
- **b.** Provision backup infrastructure for snapshot.

Step 10. The change process for billing or chargeback is initiated.

Step 11. The service is committed by the cloud provider by committing the resources in the CMDB, and the resources are locked for this customer.

Step 12. The change process is completed.

Step 13. The customer is informed through email or other electronic medium.

Service Orchestration

Although service orchestration is not a separate phase in ITIL V3, it is described here separately because of its key role in the overall service management process and the benefits it provides in automating many tasks in cloud service provisioning.

As shown in the previous section, Steps 6 through 12 are automated through service orchestration. Service orchestration refers to coordinated provisioning of virtualized resources, as well as the runtime coordination of resource pools and virtual instances. Service orchestration also includes the static and dynamic mapping of virtual resources to physical resources, and overall management capabilities such as capacity, analytics, billing, and SLA.

Figure 7-7 illustrates some of the provisioning realities for cloud provisioning. The very essence of cloud computing is being able to provide services to customers on demand and at scale in the most efficient way possible using less human resources and more automation. In most of the IT organizations, 70 percent of the IT budget is being used in maintenance activities, with the remaining going toward strategic activities. For IT organizations and CIOs to be relevant, they need to be more strategic thinkers and make IT organizations more nimble. To this end, orchestration and automation can help in automating the repetitive tasks and using the IT staff for doing more strategic activities.

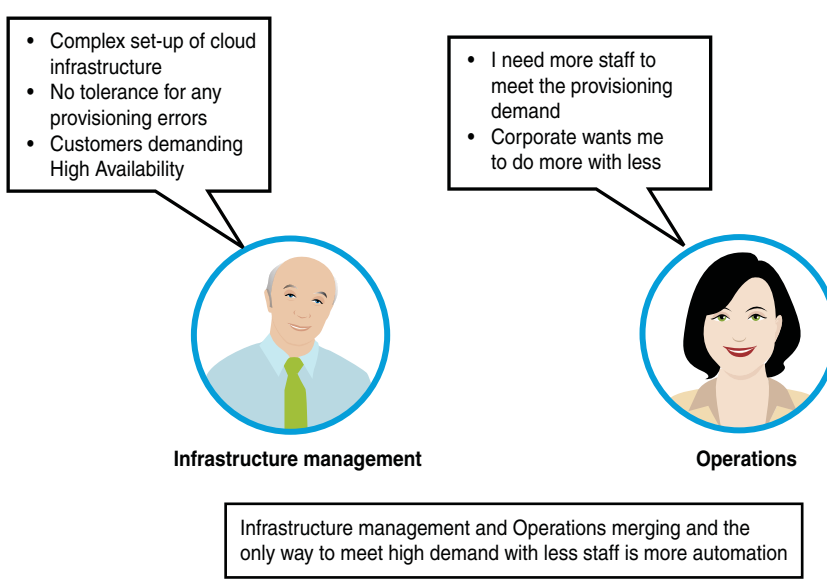

Figure 7-7 *Cloud Provisioning Realities*

The EMA (Enterprise Management Associates) showed from its research that there is clear evidence that virtualization delivers savings, and over 70 percent of the organizations queried reported that virtualization provided "real, measurable cost savings" in CapEx. However, the CapEx reduction is a one-off capital budget, but the Operational Expenditure (OPEX) benefit is year over year. The EMA research further concludes the following OPEX benefits because of automation:[3]

- **Reduction of service failures:** Fixing problems up to 24 times faster, eliminating up to 43 hours of downtime, and increasing the uptime to 99.999 percent.
- **Improved staff efficiency:** Automation improved the staff efficiency as much as 10 percent, reducing annual management costs as much as $1000/server, and each administrator was able to manage up to 1800 servers.
- **Faster service deployment:** Allowed new systems to be deployed 24 times faster and applications 96 times faster, saving almost $2000 per deployment while reducing downtime and improving time to market new products and services.

Service orchestration/automation is important for cloud provisioning because it automates some of the repetitive tasks and reduces the time it takes to provision a service. However, there is no gain to using automation when the task is not repetitive or requires human intelligence to perform. For example, if the customer has only ten servers, there are not many manual tasks involved and automation might not make sense; however, most medium to large service providers have thousands and even hundreds of thousands of servers, network elements, and storage device pools. Figure 7-8 shows the ROI factors for orchestration.[4] As service requests increase, the staff demands increase to fulfill those service requests that require many manual tasks. If the staff level remains the same or goes down, as is the case many times, the manual tasks take much longer to complete. This results in longer provisioning intervals and customer satisfaction going down, and ultimately the customer will switch to another service provider.

As shown in Figure 7-8, the orchestration value is medium when the staff demand is low and staff level is high, and as the staff level remains the same or goes down and the staff demand goes up with an increase in service requests, the need for orchestration increases. After a threshold point (see the dotted line), orchestration/automation is mandatory and the return on investment (ROI) for orchestration is high. After the threshold point, the demand for the labor increases exponentially, and automation/orchestration is not something nice to have, but it is mandatory. A lot of manual tasks that are done by talented staff need to be automated, and the talented staff can be used for doing other intelligent tasks. Another point for quality/consistency is that the errors are reduced, improving the quality, because machines tend to make less errors in the repetitive tasks.

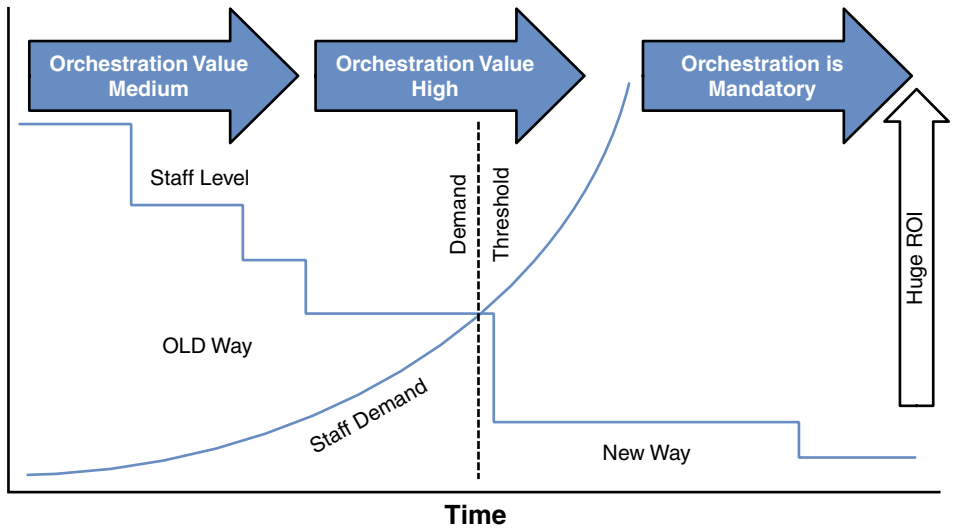

Figure 7-8 *ROI Factors for Orchestration*

Cloud End-to-End Architecture Model

To deal with complexity, end-to end cloud management is presented in a logically layered architecture (LLA) for showing all the functional blocks. The LLA is a concept for the structuring of management functionality that organizes the functions into a grouping called "logical layers" and describes the relationship between layers. A logical layer reflects particular aspects of management and implies the clustering of management information supporting that aspect. The grouping of management functionality implies grouping certain functions into layers. For cloud management, a hybrid approach is used with TMF eTOM and ITIL V3. This hybrid approach will allow grouping of all the cloud management required functions into a user experience that includes user experience functions, service and resource management functional groupings (from eTOM), and services life cycle from ITIL V3. In addition, we added the collection layer that can be used for discovery, collection of events, alerts, and so on, and an infrastructure layer that has all network, compute, and storage devices. Figure 7-9 shows this end-to-end logical architecture for cloud management.

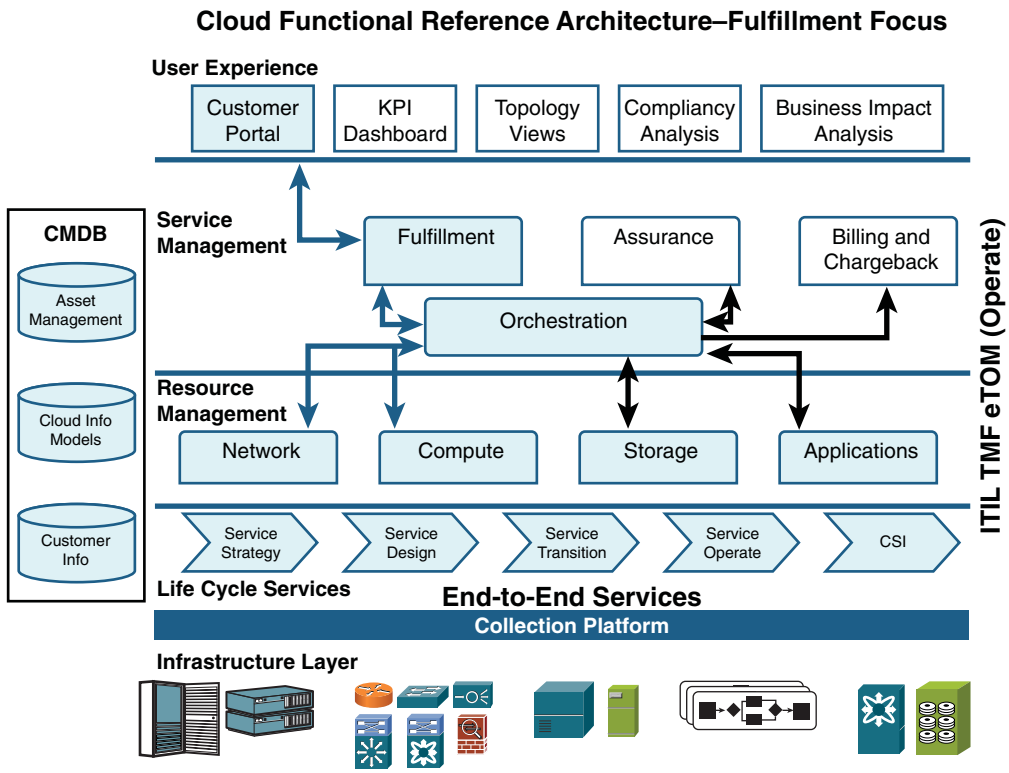

Figure 7-9 *End-to-End Cloud Management Architecture - Functional*

The functional groupings contained in each layer in Figure 7-9 are further explained here:

- **User experience layer:** The user experience function group contains the following:
  - **Customer portal:** Also referred as the customer or user self-service portal. This is where the customer/user orders services by selecting from the user interface screen. For cloud Infrastructure as a Service (IaaS), the customer could be selecting, for example, Gold, Silver, or Bronze service on the portal from the available services in the service catalog. In addition, the customer can select CPU, memory, and storage space based on the pricing options provided on the portal.
  - **KPI dashboard:** Key performance indicators (KPI) are the building blocks of many dashboard visualizations because they are the most effective means of alerting users as to where they are in relationship to their objectives. Chapter 8, "Service Assurance," provides some key KPIs for availability, capacity management, service-level management, and a relationship between KPI, KQI, and SLA is discussed. A KPI is simply a metric that is tied to a target, and often a KPI represents how far a metric is above or below a predetermined target. KPI dashboards usually show as a ratio of actual to target and are designed to instantly

show a business user whether he is on or off his plan without the end user having to consciously focus on the metrics.

- **Topology views:** Topology views are useful for IT operators to get a snapshot view of the entire network and in trouble resolution. Topology views generally provide a physical, logical, and services view of the entire infrastructure. They could also provide a view of the number of available ports and the number of hops to the host. The topology view is most useful when combined with the traceroute option, because it discovers the network path to a host.
- **Compliance analysis:** Cisco has created a suite of applications and methodologies to capture intellectual capital (IC). This IC can be leveraged to automate this knowledge so that vast numbers of customer devices can be matched against this knowledge and reports generated to inform the engineers and the customer about issues seen in the devices or network.
- **Business impact analysis:** Many times, operations would not want to make changes unless it is absolutely necessary. This method of operation leads to operating the network in a suboptimal mode. The purpose of the business impact analysis is to provide the impact of not making the required changes on the business. Cisco smart services can determine the changes required and apply those changes without any impact to the service.

- **Service management layer:** The service management is concerned with services offered to customer and interacts with functions within the same layers and other layers to perform the tasks expected at this layer. The services in this layers include the following:
  - **Fulfillment:** Fulfillment is responsible for delivering products and services to the customer. This includes order handling, service configuration and activation, and resource provisioning. A detailed explanation of this was provided in the section, "Cloud End-to-End Service Provisioning Flow."
  - **Assurance:** Assurance includes proactive and reactive maintenance activities, service monitoring (SLA or QoS), resource status and performance monitoring, and troubleshooting. This includes continuous resource status and performance monitoring to proactively detect possible failures, and the collection of performance data and analysis to identify and resolve potential or real problems. Service assurance is discussed in Chapter 8.
  - **Billing and chargeback:** These activities require the collection of usage data records, various rating functions for billing customers in the case of service providers, and chargeback/showback to the business units for enterprises. Chapter 9 covers this in detail.
  - **Orchestration:** Service orchestration/automation is important for cloud provisioning because it automates some of the repetitive tasks and reduces the time it takes to provision a service. Orchestration and the value of it to cloud provisioning was discussed earlier in the chapter.

- **Resource management layer:** The resource management grouping functions include management systems to manage network, compute, storage, and applications. The resource management layer should know what resources are available, how these resources are interrelated, and how these resources should be allocated based on the instructions received from the higher layer (for example, a self-service portal). Furthermore, this layer is responsible for the technical performance of the actual infrastructure and will control the available infrastructure capabilities and capacity to give the appropriate accessibility and quality of service.
- **Life cycle services:** The complete life cycle services is described as part of ITILV3, and the cloud tasks for all ITILV3 phases (service strategy, service design, service transition, service operate, and CSI) were described earlier in this chapter. The life cycle services layer ensures that the entire business service, along with its underlying components, cohesively assures that we are considering every aspect of a service (and not just the individual technology silos) to assure that we are delivering the required functionality and service levels (delivered within a certain time frame, properly secured, and available when necessary) to the business customer.
- **Collection platform/layer:** The collection layer enables device discovery and collection of data for postprocessing and display of data for the user experience, including the KPI dashboard, topology views, compliance analysis, and business impact analysis. The data collection schemes will include domains where a network device pushes the data to the Data Collection Service (DCS) or where the DCS will pull data from these network devices at a periodic interval. An example of a periodic pull model is where DCS will use Simple Network Management Protocol (SNMP) to pull data from the device at periodic intervals. The push model is used when the device pushes the data on a periodic interval to the DCS, for example, RADIUS (Remote Authentication Dial-In User Service) call detail records (CDR) and NetFlow records being sent by the network devices. In this push model, DCS waits for the data to arrive. The discovery of infrastructure includes the periodic collection of devices to ensure that infrastructure changes are captured and reconciled with the CMS/CMDB.
- **CMDB:** The Configuration Management Database (CMDB) is not shown in any layer and is shown separately. The CMDB is a critical part of running the traditional enterprise and touches many layers of the reference architecture. The CMDB contains
  - All the configuration items (CI)
  - A relationship model between the CIs
  - How the CIs are connected

  CMDB includes the services for the enterprise business and provides the links across ITIL V3 processes to tie it to the services. CMDB contains physical, logical, and conceptual data that are important and relevant to your business. Many of the ITIL processes require the data from CMDB to enrich the data that is collected from the devices. Also, it is important to keep the data up to date through autodiscovery and audit processes. CMDB also provides the interrelationships among all the

infrastructure resource CIs. Knowledge of CI relationships enables changes to be made with authority because it provides more visibility into the cloud environment for the change managers, allowing them to make more informed decisions, not only when preparing for a change but also when diagnosing incidents and problems.

For scaling purposes, it is expected that CMDB should be able to hold about 1 million CIs, and that number appears to be an industry standard. For cloud services, it is essential to build the relationship models for network, compute, and storage and place them in the CMDB so that the entire infrastructure CIs are used when determining the relationships.

A further discussion of the relevance of the CMDC/CMS and service inventory in cloud can be found in Chapter 10 "Technical Building Blocks of IaaS"

- **Infrastructure layer:** The infrastructure layer is the foundation of the cloud pyramid and is what cloud platforms and applications are built on. Cloud infrastructure providers such as Cisco provide network, compute, and storage devices and also assist in putting together all the facilities, cabling, and so on. Cloud vendors provide network devices (access, aggregation and core devices, plus firewalls and load balancers for services), compute devices (servers such as Cisco UCS, MS Windows, or Linux servers), and storage devices (online storage to store virtually unlimited amounts of data). In a cloud environment, the infrastructure layer (consisting of network, compute, and storage devices) is virtualized and offered to customers on demand and at scale through service portals. The virtualized resources are provisioned dynamically, and billing for the services is typically done on a usage basis and by the quality of service offered.

Summary

This chapter covered cloud service fulfillment (service provisioning) using ITIL V3 phases (service strategy, service design, service transition, service operate, and CSI) as a guide and detailed cloud service provisioning steps for provisioning network, compute, and storage. The chapter also presented the rationale for orchestration/automation for automating cloud-provisioning activities. In addition, you learned about the cloud functional reference architecture using eTOM and ITIL standards for a complete cloud life cycle.

References

[1] TMF document "Guide to Applying Business Process Framework (eTOM)," GB 921 Addendum G, version 0.10, March 2010, at www.tmforum.org/Guidebooks/GB921BusinessProcess/43162/article.html.

[2] IT Service Management - IT Infrastructure Library (ITIL) - ITIL V3, at www.best-management-practice.com/officialsite.asp?FO=1253138&ProductID=9780113310500&Action=Book.

[3] "Reducing Operational Expense (OPEX) with Virtualization and Virtualization Systems Management," an Enterprise Management Associates (EMA) white paper prepared for VMware, November 2009.

[4] Forrester presentation to Cisco on orchestration.