O'REILLY®

Five Principles for Deploying and Managing Linux in the Cloud

With Azure

Sam R. Alapati

Five Principles for Deploying and Managing Linux in the Cloud

With Azure

Sam R. Alapati

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Five Principles for Deploying and Managing Linux in the Cloud
by Sam R. Alapati

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

Acquisitions Editor: Rachel Roumeliotis
Editor: Michele Cronin
Production Editor: Colleen Cole
Copyeditor: Shannon Wright

Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

August 2018: First Edition

Revision History for the First Edition
2018-08-09: First Release

978-1-492-04092-7

[LSI]

Table of Contents

Preface

Although it's common knowledge that the cloud is a cornerstone of computing environments, there's still an incomplete awareness of the available strategies for maximizing the benefits of a cloud architecture. This book serves as a guide for people who are either contemplating a major move to the cloud or who have already initiated one but aren't sure how to efficiently use the wide-ranging services and capabilities offered by cloud vendors. The book focuses, where relevant, on using Microsoft Azure, but it also refers to services and products from other cloud providers, such as Amazon Web Services (AWS).

When planning a move to the cloud or seeking to optimize your cloud environments, it's important to understand the key cloud-enabling technologies, such as virtualization, resource replication, cloud storage devices, and object storage. The book starts off by explaining these foundational cloud technologies. On-demand computing resources, dynamic scalability, load balancing, and resiliency are all hallmarks of a cloud-based architecture. As a cloud architect, administrator, or developer, you should know how these features work.

A key reason for an unsatisfactory move to the cloud is the failure to adequately assess an organization's cloud readiness. More than the pre-deployment and deployment-related tasks, the most critical steps in a successful cloud migration are the analysis of your current architecture, prioritizing the deployment of services, figuring out your cloud personnel needs, and determining the compliance and regulatory requirements. Instead of reinventing the wheel by trying to do everything from scratch, it's a good idea to capitalize on tools,

such as Azure Migrate, offered by cloud vendors to support your move.

High availability (through geographically disparate regions and multiple Availability Zones) and load balancing are two of the most common benefits offered by a cloud-based computing environment. Azure Virtual Machine Scale Sets (VMSSs) provide both high availability and scalability, and they support automatic scaling of server capacity based on performance metrics. Caching strategies and content delivery networks (CDNs) enhance the scalability of web applications in the cloud. You can adopt technology like Azure Storage replication to achieve high availability and durability.

Monitoring server and application health and performance in the cloud can pose many problems, as compared to traditional systems monitoring. Application performance monitoring is usually a key component of your overall efforts in this regard. Dynamic resource allocation means you have less visibility into how resources are being utilized in the cloud. To get a meaningful, unified view of your cloud infrastructure, you may need to reach beyond cloud vendor–offered tools, such as Amazon CloudWatch, Microsoft Azure Monitor, and Google Stackdriver. There are several excellent third-party tools (Datadog, for example) that you can effectively integrate with a cloud-based environment like Azure.

In the cloud, security is based on a shared responsibility model, where the cloud provider and the cloud user have specific security charges. The cloud provider is responsible for the security *of* the cloud, and the customer is tasked with security *in* the cloud environment. Shared IT resources in a public cloud are a natural cause of concern. A solid network security framework, practical configuration management tools, strong access controls, and virtual private clouds (VPCs) are some of the ways in which cloud consumers can strengthen their cloud security posture.

Effective cloud-based disaster recovery (DR) strategies differ from traditional DR strategies that rely heavily on off-site duplication of infrastructure and data. Cloud-based DR solutions offer features like elasticity and virtualization, which make it easier to offload backup and DR to the cloud. More likely than not, your backup and DR solution in the cloud will cost you less and be more dependable, with minimal downtime.

Finally, cloud environments pose special challenges in the areas of operational governance, legal issues, accessibility, and data disclosure regulations. The cloud service provider must satisfy the four fundamental requirements—security, compliance, privacy and control, and transparency—to effectively serve its customers in the cloud. Cloud consumers can adopt various strategies, such as role-based access controls, network controls, and hierarchical account provisioning, to enhance security and governance in a cloud environment.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

`Constant width bold`
: Shows commands or other text that should be typed literally by the user.

This element signifies a general note.

CHAPTER 1
Introduction

Cloud computing is the provisioning and maintaining of computing services, such as servers, networking, and storage, over the internet. A *cloud provider* supplies various computing resources and services, and it charges users based on their actual usage of those resources and services, just as a utility, such as an electricity provider does.

A succinct definition of *cloud provisioning* is: a computing model that remotely provisions scalable and measured computing resources and services.

Cost effectiveness and speedy provisioning of computing infrastructure and services are two important benefits of running your computing workloads in the cloud rather than in your own datacenter. Cloud computing is a different paradigm from the historical way of running private datacenters, although traditional datacenters and cloud computing can coexist. The cloud provider may provide an organization just the computing infrastructure, or it may offer additional services that run on top of the infrastructure, such as big data and analytics.

NOTE An IT resource can be a physical or virtual server, software programs, services, storage, or network devices.

A cloud provider owns the computing resources and is responsible for making those cloud resources and services available to cloud

consumers, according to previously agreed-upon Service Level Agreements (SLAs). The cloud provider provisions and manages the compute resources and owns the resources that it leases to the cloud consumers. However, it's possible for a provider to resell the resources it leases from even larger cloud providers.

Regardless of whether it's Amazon Web Services (AWS), Google Cloud Platform (GCP), or Microsoft Azure, all clouds consist of a set of physical assets that support virtual resources, such as virtual machines (VMs). These computing assets and resources run within datacenters located around the globe, in regions such as Western Europe or the eastern United States.

The distribution of computing resources across the globe offers redundancy in failure situations and higher speed (lower latency), by locating computing resources closer to users. Software and hardware both become services in a cloud environment. It's through these services that you gain access to the underlying resources.

Leading cloud providers, such as AWS, GCP, and Microsoft Azure offer a long list of services, such as computing, storage, databases, and identity and security, as well as big data and analytics services. You can mix and match the services to create custom computing infrastructures to meet your needs and then add your own applications on top of the infrastructure to build your computing environment.

Many cloud computing services let developers work with them via REST APIs, as well as via a command-line interface (CLI). All cloud vendors offer easy-to-use dashboards to control resources; manage billing, security and users; and to optimize your cloud usage.

How the Cloud Is Being Used

Cloud computing is being used for more things than many realize. Adobe Creative Cloud is based on Azure, and the Seattle Seahawks use Azure to power their customizable, technology-integrated Sports Performance Platform. Most of the popular movie, music, streaming video, games, and picture- and document-storing services use cloud computing to serve their users.

Many companies use a hybrid cloud environment, with some on-premises infrastructure running alongside, in concert with a public

cloud infrastructure. So the cloud is increasingly a venue for regular enterprise IT workloads.

Benefits of Cloud Computing

The immense popularity of cloud computing is due to its many benefits, including:

Agility
: You can implement a cloud environment very quickly. Traditional datacenters involve ordering and setting up hardware, provisioning power and cooling, and securing the premises, all of which involve considerable time and effort. Often, the projects take multiple years due the budgeting, contracting, and implementation work involved in running onsite datacenters. Cloud implementation, on the other hand, is extremely fast—you can spin up virtually unlimited servers and storage in a matter of minutes.

Pay-for-use billing model
: In a cloud environment, you lease computing resources, on a pay-for-use model. You are billed for only your actual usage of the IT resources. Obviously, this has the potential to reduce both your initial infrastructure investment and your operational costs, as compared to a datacenter-based computing model.

Cost
: Although you must be smart about how you utilize cloud computing and use all the deals offered by the cloud providers to reduce costs (such as spot pricing of compute instances), cloud computing doesn't involve the traditional capital expense of buying hardware and other components required for running a datacenter.

Elasticity
: The ability to quickly ramp up (and down, if needed) computing capacity is a hallmark of cloud computing and serves as a strong differentiator from traditional datacenter-based computing environments.

Reliability
: Traditional concerns, such as disaster recovery and data backups, become less worrisome since cloud providers offer built-in

resiliency by storing data in multiple, geographically separate from zones.

Security
: When you run workloads in a public cloud, you follow a shared responsibility model for security, in which you're responsible for application security and the cloud provider secures the computing infrastructure from external threats.

Performance
: Since a cloud provider can offer the very latest in computing hardware, as well as lower network latency, application performance is usually enhanced in a cloud environment.

Types of Cloud Services: Iaas, PaaS, and Saas

Cloud providers offer various types of services, depending on the depth and breadth of the computing stack they offer. Figure 1-1 illustrates the three broad types of cloud services.

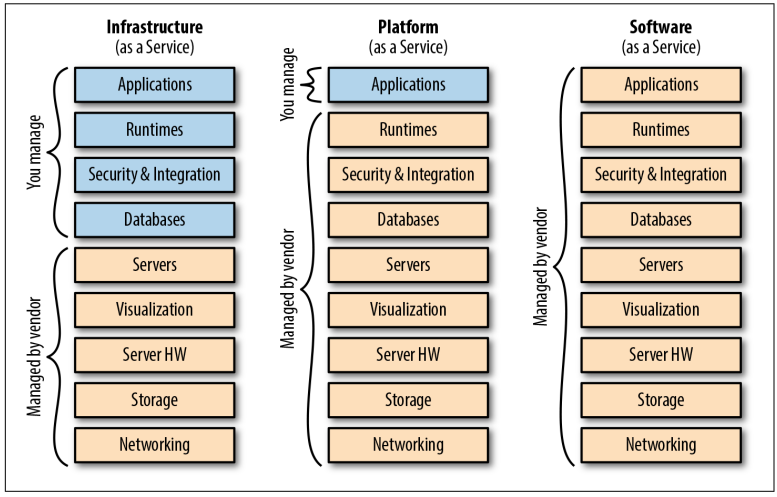

Figure 1-1. The three types of cloud services

Infrastructure as a service (IaaS)
: IaaS is the most common type of cloud service, and this is how most people use the cloud. Under IaaS, the cloud provider supplies the IT infrastructure, such as servers, storage, and networks, which you'll pay for based on usage. Most of the IT resources offered under the IaaS model aren't preconfigured,

meaning that the cloud consumer has a high degree of control over the cloud environment. The consumer must configure and maintain the bare infrastructure provisioned by the cloud providers.

Platform as a service (PaaS)

PaaS is a computing model in which the cloud provider provisions, sets up, and manages all the computing infrastructure, such as servers, networks, and databases, and you do the rest. PaaS is a ready-to-use computing environment since the resources and services are already deployed and configured. PaaS computing services include those that help you develop, test, and deliver custom software applications. Developers can quickly create their apps, and the cloud provider sets up and manages the underlying computing infrastructure. The cloud consumer can replace their entire on-premise computing environment in favor of a PaaS. Or they can use the PaaS to scale up their IT environment and/or reduce costs with the cloud environment.

Software as a service (SaaS)

SaaS is how a cloud provider delivers software applications on demand over the internet. In this mode, the provider manages not only the infrastructure but also the software applications, and users connecting to the application over the internet. The software program is modeled as a shared cloud service and made available to users as a product. Cloud consumers have limited administrative and management control, with a SaaS cloud delivery model.

A good example of a SaaS model is the transitioning of Adobe's well-known Creative Suite to a SaaS model called Adobe Creative Cloud. As Adobe migrates more products to this model, it has signed a strategic partnership with Microsoft to make Microsoft Azure its preferred cloud platform.

"Adobe is offering consumer *and* enterprise applications in Azure, along with our next-gen applications, like Adobe Cloud Platform," says Brandon Pulsipher, Vice President of Technical Operations and Managed Services at Adobe. "Our partnership with Microsoft demonstrates that cloud-native applications in Azure make great sense for large and small customers alike."

For more information about Adobe's use of the Microsoft Azure platform to successfully transition to the cloud through a SaaS model, please read "Adobe runs its vast open-source application portfolio in Azure".

Figure 1-1 illustrates how the three types of cloud services (IaaS, PaaS, and SaaS) differ from each other.

It's important to understand that cloud providers offer a mix of the three cloud services paradigms described here and their derivatives, including *functions as a service*, *containers as a service*, and others. Users often subscribe to more than one type of cloud service.

Types of Cloud Deployments

You can deploy cloud computing resources in three different ways: public, private, and hybrid.

Public cloud
: A *public cloud* is run by a third-party cloud provider, such as Microsoft Azure, AWS, or GCP. Users access the cloud publicly over the internet.

Private cloud
: In a *private cloud*, you maintain the computing infrastructure and services on a private network. Your organization owns the private cloud and helps you employ cloud computing technologies to centralize access to companywide IT resources by internal users/departments. An organization can run its private cloud in its on-site datacenter, or it can hire a third-party service to host it.

Hybrid cloud
: A *hybrid cloud* environment consists of two or more cloud deployment models. In a hybrid cloud, your private cloud and the public cloud share data and applications. Data can freely flow between the private and public clouds, or you may use a private cloud for hosting sensitive data and host other cloud services on the public cloud.

Cloud-Enabling Technology

The cloud owes its phenomenal growth over the past decade to several technological developments, of which virtualization (server,

storage, and network, among others) is but one. Other key innovations include various web technologies, service-oriented architectures, multitenant technologies, resource replication, cloud storage devices, and object storage. I briefly describe the main cloud enabling technologies in the following sections.

Virtualization

The largest cloud platforms, such as AWS and Azure, have set up a number of massive datacenters across the world, specifically designed to deliver services at a massive scale. By the end of 2017, Synergy Research Group, which tracks IT and cloud-related markets, estimated that there would be 390 hyperscale datacenters in the world. Each of the companies that fall under the large cloud platforms has at least 16 datacenter sites, on average, according to Synergy, with the biggest cloud providers (AWS, Microsoft, GCP, and IBM) operating the most datacenters.

Hyperscale virtualization is at the heart of cloud computing. A software called *hypervisor* sits on the physical server and helps abstract the machine's resources. Most of us are familiar with server virtualization, but in the cloud, other resources, such as storage and networks, are also virtualized.

Cloud computing relies on virtualization, but it's much more than simple virtualization. A cloud provider allocates virtual resources into centralized resource pools called a *cloud*. Cloud computing is the orchestration of these clouds of computing resources through management and automation software. In addition to virtualized resources, a cloud offers features such as self-service, automatic scaling, and enhanced security.

Virtualization is the process of converting a physical IT resource into (multiple) virtual resources. Cloud-based systems virtualize many types of IT resources, such as:

Servers
: Physical servers are the basis of virtual servers.

Storage
: Virtual storage devices or virtual disks are based on underlying physical storage.

Network
: Physical routers and switches can serve as the basis of logical network fabrics, such as VLANs.

Power
: You can abstract physical uninterruptable power supplies (UPSs) and power distribution units into virtual UPSs.

The best-known virtualization technology, of course, is server virtualization. In a nonvirtualized environment, the OS is configured for a specific hardware and you must usually reconfigure the OS, if you modify the IT resources. Virtualization translates IT hardware into emulated and standardized software-based copies. Thus, virtual servers are hardware independent. It's this hardware independence that enables you to move a virtual server to anther virtualization host, without worrying about the hardware-software compatibility requirements.

Virtualization, by allowing multiple virtual servers to share a single physical server, enables server consolidation, which leads to higher hardware utilization, load balancing, and optimization of computing resources. On top of this, virtual machines can run different guest operating systems on the same host. All these virtualization features support the hallmarks of cloud computing, including on-demand provisioning and usage, elasticity, scalability, and resiliency.

Web Technologies

Web technologies are used by cloud providers in two ways: as the implementation medium for web-based services, and as a management interface for cloud services. Well-known elements, such as Uniform Resource Locators (URLs), the Hypertext Transfer Protocol (HTTP), and markup languages, such as HTML and XML, are the essential components of the technology architecture of the web.

Web applications are distributed applications that use these web-based technologies, and their easy accessibility makes them part of all cloud-based environments. PaaS cloud deployment models help consumers develop and deploy their web applications by providing separate web server, application server, and database server environments. Many applications benefit from the cloud model, particularly from the elastic nature of of cloud infrastructure provisioning. Cloud providers themselves use a lot of web technologies for enablement, most notably REST APIs and JSON, among others.

Web services are the first popular medium for sophisticated web-based service logic. Web services are also called *SOAP-based*, since they rely on the SOAP messaging format for exchanging requests and responses between web services. The API of a web service uses a markup language called Web Service Description Language (WSDL), and the messages exchanged by the web services are expressed using the XML Schema Definition (XSD) language (XML Schema).

Along with the Universal Description, Discovery, and Integration (UDDI) standard for regulating service registries where WSDL definitions can be published, XML schema, SOAP, and WSDL are the essential components of early web service technologies. Later web service technologies (called WS-*) address other functional areas, such as security, transactions, and reliability.

Representational State Transfer (REST) services are based on a service architecture that operates according to a set of constraints to emulate the properties of the web. REST describes a set of architectural principles through which data is transmitted over a standard interface, such as HTTP. REST focuses on the design rules for creating stateless services. A client accesses the resources using unique URIs for the resources, and unique representations of the resources are returned to the client. With microservices or, at the very least, a proliferation of endpoints and applications, the cloud needs a lot of messaging and so all cloud providers have queues, buses, notifications, and other message passing and orchestration abilities.

Resource Replication

Resource replication is the creation of multiple instances of the same computing resource. Typically, virtualization strategies are used to implement the replication of the resources. For example, a hypervisor replicates multiple instances of a virtual server, using stored virtual server images. Most commonly, servers, cloud storage devices, and networks are replicated in a cloud environment.

Cloud Storage Devices and Object Storage

In a cloud environment, you can reference and store various types of data as web resources. This type of storage is called *object storage* and supports a wide variety of data and media types. *Cloud storage device* mechanisms implement the interfaces to object storage, and

you can access these object storage interface mechanisms via REST or web services.

For Linux system administrators, cloud storage represents new challenges that they're not used to with their local storage area network/network attached storage (SAN/NAS) storage systems. Cloud storage involves a lot of REST-based storage operations versus filesystem operations. Just like in Azure, you have blob storage, files, managed disks, and Third-party–provided NAS-like appliances. And that's just for files (blobs). Key-value pairs, secrets, document storage, and ultimately, database persistence are a whole different ball game.

Cloud Computing Architectures

Most cloud computing providers offer a set of common cloud features, as summarized in the following sections.

On-Demand Usage of Resources

A cloud consumer is completely free to provision any IT resources offered by a cloud provider. The cloud consumer doesn't need to interact with the cloud provider to provision and use any of the cloud-based services, thus establishing an *on-demand*, self-service usage pattern.

Measured Usage

Closely related to the ability to use computing resources on demand is the concept of *measured usage*. All cloud providers charge their consumers just for the IT resources used, rather than for the resources that are provisioned or allocated to the consumer. Measuring usage supports customer billing, as well as usage reporting.

Resource Pooling

Resource pooling is how a cloud provider pools a large amount of computing resources to service multiple consumers. The cloud provider dynamically allocates and deallocates virtual resources to cloud consumers according to fluctuations in demand. Multitenancy (multiple cloud consumers, unbeknownst to each other, sharing a single instance of a computing resource) supports resource pooling.

Dynamic Scalability (Elastic Resource Capability)

Dynamic scalability and *elasticity* refer to the ability of a cloud provider to transparently scale computing resources in response to the runtime conditions of a user's environment. Virtualization enables cloud providers to maintain large pools of computing capacity on hand to service the needs of their customers with minimum delays. One of the key reasons for migrating to the cloud is its built-in elasticity, which obviates the need to incur large capital expenditures on infrastructure, in anticipation of an organization's growth.

Resiliency (Servers and Storage)

Resiliency is a hallmark of cloud environments and is one of the biggest benefits offered by the cloud. Cloud providers frequently provide resiliency by locating redundant computing resources in different geographical areas, called Availability Zones in AWS and Microsoft Azure. The redundant implementation of cloud services means that the secondary (or standby) service can immediately and automatically take over the processing, in the event of primary services failure.

Load Balancing

Load balancing is how a cloud platform manages online traffic by distributing workloads across multiple servers and other computing resources. Load balancing can be automatic or on demand. The goal of load balancing is to keep workload performance at the highest possible levels by preventing overloading of the computing resources, thus enhancing the user experience.

Running Linux in the Cloud: The Role of Containers

As I explained earlier in this introduction, virtualization is a key enabling factor in the success of cloud computing. Microsoft Azure, for example, provides Azure Linux virtual machines running on Red hat, Ubuntu, or a Linux distribution of your choice. Azure provides its customers with the ability to run a Linux virtual machine in the cloud, whether it's Red Hat, Ubuntu, CentOS, SUSE, Debian, or other distributions, as well as the ability to bring their own Linux images.

Linux-based containers offer easier deployments through the maintenance of a secure registry of container images, and a more efficient use of resources. You also can manage and orchestrate sets of containers using dedicated orchestration tools, such as Kubernetes.

Although VMs are still the predominant way to run workloads in the cloud (and in on-premise datacenters), containers are becoming increasingly popular in cloud environments, with AWS offering the Amazon container services, and from Microsoft Azure, the Azure Container Service.

Container Use Cases

The three major use cases for running containers in the cloud include running microservices, batch jobs, and continuous integration and continuous deployment (CI/CD) of applications.

Running microservices

Containers are ideal for running small, self-contained applications that perform single tasks or run single processes. You can, for example, run separate containers for a web server, application server, or message queue, among others. Since the containers run independent of the other containers, it's easy to scale specific parts of the application up or down, as needed.

Running batch jobs

You can take advantage of one of the foundational principles of containers—isolation—to run batch and extract, transform, and load (ETL) jobs in containers. You can run multiple such containers on the same cluster, since they're all isolated from each other. Because containers start up very quickly, you can use them to handle spurts in demand, by launching more containers.

Continuous integration and deployment

Docker enables you to version your Docker container images, making it easy to use containers for continuous integration and deployment. An automated build process supported by a CI tool, such as Jenkins, can pull the latest code from the code repository and can build/package the code into a Docker image. Jenkins can then push the new Docker image to your Docker repository, where your

deployment process can pull the image, test the app, and deploy it to production.

You can achieve easily replicable, speedy, reliable, and manageable deployments by orchestrating the deployment of the containers you use for CI/CD, using Kubernetes in the Azure Container Service. Figure 1-2 shows a container-based CI/CD architecture using Jenkins and Kubernetes on the Azure Container Service.

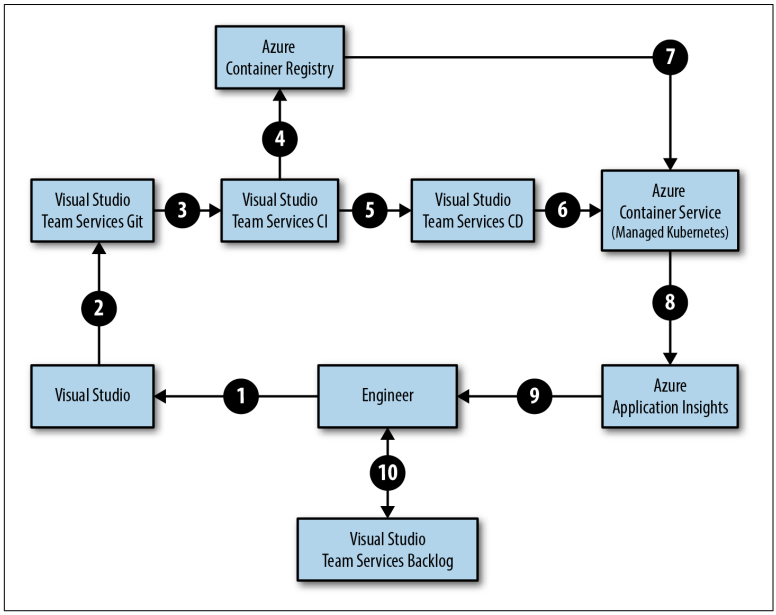

Figure 1-2. CI/CD with Jenkins and Kubernetes on the Azure Container Service

Running Containers in the Cloud

All cloud providers support containers, with Google Cloud Platform seemingly having embraced them earlier than its competitors. (Kubernetes, the most popular container orchestration system, was developed at Google.) However, both AWS and Microsoft Azure fully support containerization through dedicated container services that support the orchestration of containers.

Azure started out with DC/OS and Docker Swarm for managing containers and eventually added Kubernetes. However, Azure was first to the market with the launching of container instances, followed by AWS Fargate. GCP has no equivalent as of the time of this

writing. In summary, all major cloud providers (AWS, Azure, and GCP) now have a Kubernetes play when it comes to managing containers in the cloud.

Container Instances and Cloud Container Services

When you run containers in the cloud, you can run them on the VMs that you create. You can install Docker on the VM and download container images from any registry that you choose.

Many cloud providers, however, also offer a container service to facilitate the creation and management of the underlying infrastructure. So you can either spin up your own containers on VMs provisioned for you, or let the cloud provider create and manage them for you by subscribing to its container service. You may also choose to start with a container-optimized distribution such as Rancher or CoreOS. If you're going the PaaS route, you can start with a container-based PaaS, such as Tectonic, OpenShift, or Cloud Foundry.

Container instances and container orchestration

If you choose to run your own container cluster, you must have a way of managing the containers and launching applications on the cluster. Although you can launch and manage containers individually, with homegrown tools, you almost always use an orchestrator to automate the deployment of multicontainer workloads. Two well-known orchestration tools are Kubernetes and Docker Swarm.

Kubernetes helps you run a container cluster and deploy applications to the cluster and is quite popular in the container world. Docker Swarm is the other commonly used container orchestrator. You can use the three Docker-container related tools—Swarm, Machine, and Compose—together to put together a valid alternative to Kubernetes, although Kubernetes offers an easier way to get it all done.

Microsoft Azure offers various services to support your container needs, including:

Azure Kubernetes Service (AKS)
: For orchestrating containers using Kubernetes, DC/OS, or Docker. It may come as a surprise that AKS is fully Linux-based, showing that you can be all Linux in the Azure cloud.

Azure Container Instances (ACI)
: For running Docker containers on Axure VMs.

Azure Container Registry (ACR)
: For storing and managing container images.

NOTE

In addition to the Azure Container Service, you can use Azure Service Fabric to develop microservices and orchestrate containers on Linux. You can also use Azure App Service to deploy web applications on Linux using containers and Azure Batch to run repetitive compute jobs using containers.

Using a Container Service

Planning and deploying fully orchestrated containerized applications, even with a sophisticated orchestration tool, such as Kubernetes, isn't trivial. Setting up a Kubernetes managed cluster is complex and takes quite a bit of time.

A container service, such as Azure Container Service, lets you easily manage your containers without any expertise in container management. You can provision clusters very quickly and monitor/manage the cluster with a built-in operations console. Azure's Container Service offers a fully managed Kubernetes cluster, but you can deploy an alternative orchestrator if you wish, such as unmanaged Kubernetes, Docker, or DC/OS. But you must bring your own management and monitoring tools when you do this, whereas the fully managed Kubernetes cluster comes with everything already included.

The Azure Container Service aims to offer its customers the benefits of open source Kubernetes without the headaches and operational overhead. Azure also offers container monitoring, which shows all your running containers and the images they're running, as well as auditing information about the commands that are being run on the containers. Instead of remotely viewing the Docker hosts, you can troubleshoot containers by searching centralized logs.

When running a Kubernetes managed cluster within the Azure Container Service, you can use the open source packaging tool Helm (similar to Linux package managers, such as apt-get and yum) to install, deploy, and manage containers in the Kubernetes cluster.

Helm manages Kubernetes charts, which are packages of preconfigured Kubernetes resources.

In addition to Helm, Microsoft also offers the Brigade and Draft tools, all of which cater to developers, and container administrators. Brigade (event-driven scripting for Kubernetes) helps you to build any ordered workflow of containers in Kubernetes and to trigger the workflow by listening for arbitrary events. Draft makes it easy to develop container-based applications and to deploy them to Kubernetes clusters without worrying about Docker and Kubernetes (you don't even need to install them). Teams can focus on building applications with Kubernetes rather than on managing the container infrastructure.

Although you do pay for the Kubernetes nodes you use (as well as the underlying infrastructure resources you consume, such as storage and networking), the managed Kubernetes service offered by Azure is free, thus making the management of your Kubernetes cluster a free affair!

CHAPTER 2

Principle 1: Understand Which Linux VMs Are Adaptable to the Cloud

Running your services in the public cloud could mean lower maintenance costs, along with elasticity (the ability to quickly scale your infrastructure according to business demand), robust disaster recovery, and high-availability services. However, without thorough assessments and planning, a cloud migration effort can cost excessive time and money and can result in a paucity of required technological skills, along with security and compliance issues stemming from a lack of control over your cloud computing resources, bandwidth difficulties, and more.

In our discussion of how to migrate to the cloud, I chose not to dwell on the initial business case for a move to the cloud. I assume that a business case has been made for such a move, based on the cost, effort, potential pitfalls, long-term benefits, and the ease (or difficulty) of the migration and implementation. I focus on the technological implications of this move.

From a purely technical point of view, adopting the cloud is easier than setting up or expanding a datacenter-based computing infrastructure. However, it's also easy to flounder in your cloud adoption effort, if you don't educate yourself well. Often the failure to successfully adopt the cloud doesn't just leave you where you started—

you're actually likely to lose critical time and to waste your resources, which can put you behind your competitors.

There are many half-baked truths and pitfalls in cloud computing. To pick the best of cloud computing, it's important for you to understand which parts will work for you and how to plan and implement effectively for a migration to the cloud. In this chapter, I explain the importance of properly assessing and identifying your cloud readiness, along with the key phase of discovery during the migration to the cloud. Identifying the parts of the infrastructure that are cloud-capable is key step in migrating to the cloud. Identification of pilot applications and development of detailed plans to implement the pilot project come later in the journey to the cloud.

The Cloud RoadMap

To implement a cloud-first strategy, you must create a working cloud adoption road map. A well-built road map addresses the following concerns:

Benefits you can expect from the move to the cloud
: Benefits can flow from the optimization of the IT infrastructure and operations and from cost reductions.

What to move to the cloud
: Not all applications and infrastructure need to move. Evaluate which of your current infrastructure components and applications can benefit the most by making the move. Select the best candidates (service portfolio) for a cloud migration so that your initial cloud foray has a high chance of success.

Which technology to choose
: There are multiple cloud delivery and deployment models, along with various cloud providers. Establish the criteria by which to select the appropriate delivery/deployment models and vendors.

How to optimize
: Adoption of cloud computing must serve the purpose of optimizing your IT infrastructure. Key objectives behind moving to the cloud include lowering your long-term costs and reducing your capital outlays.

Cloud Readiness Assessments

Whether you're planning a complete migration to the cloud or you'd like to move a couple of applications over, a cloud assessment is your starting point. A good assessment takes your cloud goals and determines the best ways to achieve them, by understanding the changes you need to make and learning how the move impacts all areas of your business.

Cloud migrations tend to be more complex than some might estimate, and poorly done assessment prior to the move can make the migration even messier. A proper assessment reveals how ready your organization is, from both a technical and a business viewpoint. It should cover technology processes, the technology teams, and business elements. The assessment should set your expectations regarding the benefits you should reap and how to maximize the potential benefits from a move to the cloud.

Essentials of a Cloud-Readiness Assessment

A good cloud-readiness assessment must include an analysis of existing applications, a cost estimation, and explorations of cloud architectures, migration plans, and compliance regulations. The result of this assessment is a comprehensive report on your organization's cloud readiness. Main components of this assessment could include:

Shareholder Interviews

The main purpose of the shareholder interviews is to communicate the organization's vision for the cloud. The assessment team also gathers the expectations of the stakeholders regarding the potential performance of key enterprise applications in the cloud.

Current Infrastructure Analysis

The assessment team must also analyze the current datacenters, with a view to learning everything about the current computing, network, and other infrastructure components. In addition, the assessment must document all interfaces, file transfers, and dataflows that support current applications.

Workload, Application, and Database Analysis

The heart of any cloud assessment is the evaluation of the workloads that the organization plans to migrate. These can include business applications, external and internal websites, SaaS services, and email servers, among other workloads.

Business applications are the focus of most of an organization's computing resources, likes servers. Although some business applications are independent, most have dependencies with supporting applications. So a migration plan must review not just the primary business applications but also the vastly larger number of supporting applications and processes.

The review of current applications must include not only a study of the existing application design and architecture but also of the critical third-party dependencies and integrations. Understanding the usage patterns of different types of databases (such as relational, or NoSQL) is critical to the PaaS versus IaaS decision down the road.

Although the cloud may offer an organization enhanced flexibility, lower its costs, and increase its agility, not all applications are good candidates for a move. Hybrid cloud deployment models are thus pervasive—very few organizations choose a 100% cloud approach, in the face of this reality.

Prioritization

Select a set of noncritical applications and services to migrate for a new cloud infrastructure and service proof of concept (POC) exercise or to perform a risk analysis.

Cloud Architectures and the Cloud Deployment Model

Your assessment of the current infrastructure and the application and database analysis should provide you with sufficient knowledge to choose among IaaS, PaaS, and SaaS cloud delivery models.

Cloud Personnel Requirements

The move to a cloud environment reduces your need for traditional on-premise datacenter system administrators and probably for data administrators (if you use a PaaS cloud deployment model). However, you'll need a staff that knows how to get the most out of a

cloud-based system. This includes solution architects well versed in the cloud, as well as DevOps personnel who can work with the cloud vendor's application deployment and CI/CD tools.

Cost Analysis

The assessment team must find out the organization's cost expectations. For many organizations, a key reason for migrating to the cloud is that there are fewer capital outlays as compared to a traditional datacenter-based environment. However, if an organization doesn't spend the effort to learn how to optimize cloud resource use, the cloud may turn out to be more expensive than expected. For example, of spot purchases of surplus computing power is a powerful way to reduce the cost of running virtual servers in the cloud.

Compliance and Regulation Requirements

Sometimes stringent compliance and regulatory requirements make it harder for an organization to move to the cloud. The cloud assessment should check off all such requirements to ensure that the cloud provider can help the organization satisfy them.

Cloud Migration Plan

As part of the assessment plan, the migration plan should prioritize the applications that move to the cloud first. The assessment should list the applications in order of criticality. It should also estimate the code changes necessary to move the applications to the cloud.

The assessment report should be comprehensive, both in its analysis of existing applications and infrastructure and in its proposed cloud migration path. It should also specify the areas where the organization lacks expertise and suggest ways to build its teams, either through training or through hiring experts in these areas.

Cloud Migration Strategies

There are two basic strategies you can adopt when migrating to the cloud:

Lift and shift
: In this strategy, you move the entire current software stack, including the operating system, applications, databases, work-

loads, and other components to the cloud. You migrate your applications without fundamental changes in their architecture. In other words, this is the "old wine in a new bottle" application migration strategy, since you make little or no use of cloud-native features. This is usually an expensive option, since it doesn't deliver immediate cost savings.

Architect applications before migration
: This option positions your organization to take advantage of cloud-specific features. These include cloud APIs, built-in high availability, and elasticity (autoscaling of the computing capacity, for example). Obviously, there are more risks in this strategy, since you're simultaneously planning an infrastructure migration and upgrading the applications.

A key goal in migrating existing applications to the cloud is a zero-downtime, or at least a near-zero downtime migration. Most organizations that move to the cloud do so by first doing a POC-type migration of a noncritical application.

Cloud Migration Tasks

You must perform numerous preparatory tasks before migrating to the cloud. You can group these tasks into the following categories:

- Pre-deployment tasks
- Migration tasks
- Go-live tasks

Pre-Deployment Tasks

Pre-deployment tasks include broad brush tasks, such as understanding the scope of the migration and creating the cloud architecture.

Understanding the scope of the migration

If you've done your cloud assessment correctly, you should also have a good idea of the scope of your cloud migration. Some of your applications may be so old that they'd need to be rearchitected for a move, costing time and money. You can choose to leave these applications in the datacenter or to host them in the cloud without making any changes.

Creating the cloud architectures

Creating the cloud architecture involves selecting the types of cloud services that the organization must adopt, based on various criteria, such as business requirements, cost, performance, reliability, and scalability.

Setting up the cloud accounts

After you determine the cloud architecture, it's time to create cloud accounts and to onboard the teams by granting them access credentials and introducing the cloud architecture. Setting up identity and access management (IAM) precedes the creation of users and groups, and you can specify which resources users can access within their cloud accounts.

At this point, the organization is ready to perform the migration tasks.

Migration Tasks

Migration tasks include setting up the necessary networks, creating your computing infrastructure, deploying the applications and databases, and planning the cutover from on-premise systems to the new cloud-based systems.

Setting up the networks

The first major task is to set up a virtual private network (VPN) connection between your organization and the cloud. This step isn't mandatory but is common. A cloud account, such as Microsoft Azure, lets you create a virtual private cloud (VPC). You must create the VPC and the necessary subnets in your cloud account.

Creating your computing infrastructure

During this step, you create your computing infrastructure, such as the VMs, databases, and analytical services, in accordance with your architecture.

Deploying the applications and databases

Deploying application code and migrating data are the key steps in the deployment phase of the migration. You can migrate data using native database tools, like export/import or SQL dumps or using

specialized database migration tools, such as the Azure Database Migration Service, which I explain in more detail in the following sections.

Planning the cutover to the cloud

After the applications and the databases are deployed in the cloud, the on-premise databases and the cloud databases need to be synchronized, and you must get ready for cutting over to the cloud-based systems. There are two key steps before the cutover to the cloud:

Performance testing
: After you cut over your on-premise systems to the cloud, application performance is the main concern. Stress testing new systems and benchmarking execution are necessary steps to assure high performance.

Security assessment
: Ensure that the cloud systems are secure, by performing vulnerability assessments and penetration tests.

Go-Live Tasks

If all goes well during the performance testing and security assessment, perform the *go-live tasks* and cut over to the new systems. Careful, continuous monitoring of the new systems is critical, so you can quickly revert to the old systems if you run into unexpected glitches.

Using Tools for Migrating to the Cloud

Migrating to the cloud often involves risk and unexpected delays. For this reason, instead of trying to reinvent the wheel, it's better to use a formal migration strategy as well as tools and services developed explicitly for supporting such a migration. One such tool is the Azure Migrate Service, which, among other capabilities, can perform dependency mapping, to support the successful migration of multitier applications.

NOTE

Azure Migrate can discover up to 1,000 VMs in a single discovery.

Azure Migrate helps you primarily in the following ways:

Assesses your readiness for the Azure cloud
: You get an assessment about the suitability of your on-premise VMs to run in the Azure cloud.

Recommends the best sizes for your cloud VMs
: By default, Azure Migrate uses the performance history of your on-premise VMs to get appropriate size recommendations for the Azure VMs. This is very helpful when you've overallocated your on-premise VMs, compared to their utilization, and you'd like to fix this by correctly sizing the VMs in Azure, to save costs. You can also ask the service to size the VMs in Azure as "on-premise," without considering the performance history of the on-premise VMs.

Estimates your monthly costs
: The service provides an estimated cost of running your current set of on-premises VMs in Azure.

NOTE

Azure Migrate considers a buffer (comfort factor) during its assessment exercises. This allows you to provide a cushion to handle seasonal usage spurts and likely increases in future resource usage. The service applies the buffer on top of the server utilization rate for the on-premise VMs (such as CPU, RAM, I/O, and network bandwidth).

Figure 2-1 shows how the Azure Migrate service works by discovering information about your on-premise VMs.

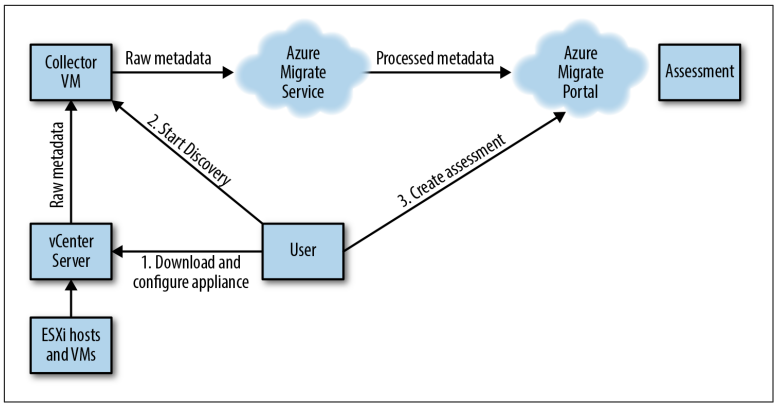

Figure 2-1. How Azure Migrate works

In addition, Azure Migrate creates groups of machines by visualizing the dependencies of the on-premise VMs that can migrate together to Azure, offering a high degree of confidence in the migration.

How reliable are the size recommendations provided by Azure Migrate? All Azure Migrate assessments have a confidence rating attached to them. The confidence rating ranges from one star to five stars (one star is the lowest, and five stars are the highest rating). The confidence ratings depend on the availability of data necessary to complete an assessment. The more data, the greater the confidence rating, and the more reliable the sizing recommendations. You can customize an assessment by changing its properties.

NOTE

Azure Migrate helps with the right-sizing of Azure Virtual Machines.

After you move to the cloud, you need to continuously push application changes to VMs. Figure 2-2 shows how you can set up an immutable infrastructure CI/CD using Jenkins and Terraform on Azure Virtual Machine Scale Sets (VMSSs). An Azure VMSS lets you create and manage a group of identical, load-balanced VMs. The number of VM instances automatically increases or decreases, based on demand or per a schedule that you define.

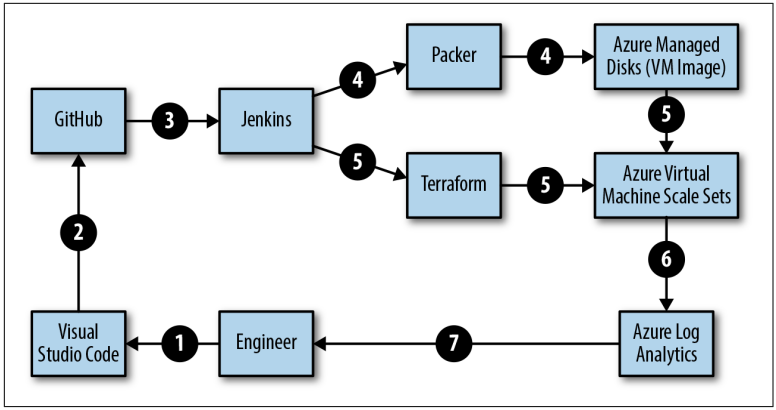

Figure 2-2. Immutable Infrastructure CI/CD using Jenkins and Terraform on Azure VMSSs

Whether you use Java, Node.js, Go, or PHP to develop your applications, you'll need a CI/CD pipeline to automatically push changes to the VMs that support those applications.

Here's an architectural overview of the immutable infrastructure shown in Figure 2-2:

1. Change application source code.
2. Commit code to GitHub.
3. Continuous integration trigger to Jenkins.
4. Jenkins triggers a Packer image build to create a VM and stores it as a VM image using Azure Managed Disks.
5. Jenkins triggers Terraform to provision a new VMSS using the Azure Managed Disks VM image.
6. Azure Log Analytics collects and analyzes logs.
7. You monitor the applications and make improvements.

For a CI/CD sample that uses a template and uses Jenkins and Terraform on Azure VMSSs, please read "CI/CD using Jenkins on Azure Virtual Machine Scale Sets".

Deploying VMs using a template based on Jenkins and Terraform on an Azure VMSS makes it simple for system administrators to deploy their infrastructure. Here's an example:

```
$ azure config mode arm
```

```
$ azure group deployment create <my-resource-group>
<my-deployment-name> --template-uri
https://raw.githubusercontent.com/azure/azure-quickstart-templates
/master/jenkins-cicd-vmss/azuredeploy.json
```

The Three Key Phases of Cloud Migration

You can divide most cloud migration project into three major phases: discovery, migration, and optimization (and modernization).

The Discovery Phase

The broad purpose of the discovery phase is to help an organization answer the following key questions:

- What is the nature of the current environment—applications, databases, and critical workloads?
- Will the application or workload run properly on the cloud provider's environment?
- What is the cost of running the current computing environment in the cloud?

The Azure Migrate service assesses your on-premise workloads for migration to the Azure cloud. The service maps the current environment to the Azure Virtual Machine instances. This mapping helps in figuring out the expected costs of running the infrastructure in the Azure cloud. The Migrate service also reports on potential compatibility issues, with guidelines for remediating them.

Azure Migrate is especially useful if you're not planning to redevelop or rearchitect your current applications but are instead setting up a lift-and-shift migration to the cloud. After Azure Migrate provides its assessment, you can use other services, such as Azure Site Recovery and Azure Database Migration Service, to migrate your VMs to Azure, as I explain in the following sections.

One of the hardest parts of a cloud migration is moving the data from an on-premise database to the cloud databases. Microsoft offers a tool, Data Migration Assistant (DMA), to help with database discovery and assessment. DMA scans your running databases to detect potential showstoppers, such as unsupported features that are currently in use in the on-premise databases.

NOTE Azure Migrate attempts to map every disk attached to an on-premise VM to a disk in Azure.

You can link DMA to the Azure Database Migration Assistant (Azure DMA), which partners with the Azure Database Migration Service (Azure DMS) for database discovery and migration. Acting in tandem with Azure DMS, Azure DMA can create migration workflows to move database schemas, data, users, roles, and SQL logins. Azure DMS helps migrate your on-premise Oracle, MySQL, and SQL Server databases to an Azure managed database in the cloud or to your own database running in an Azure VM.

Here's a summary of the steps you perform when using Azure DMS to perform a database migration:

1. Create the target database in Azure.
2. Migrate the database schemas using Azure DMA.
3. Create an instance of Azure DMS.
4. Create a migration project by specifying the source and target databases, and the tables, to be migrated.
5. Initiate the target database load (full load).
6. Manually switch over the production environment to the Azure-based database.

Assessing the suitability of a machine to run in the cloud isn't a trivial issue. Doing this assessment without a sound migration tool, such as Azure Migrate, can leave you with a set of defined choices (migrate/don't migrate), but, the migration question is much more subtle in many cases.

When Azure Migrate assesses the on-premise VMs for their suitability to migrate to the Azure cloud, it categorizes the servers into the following categories:

Ready for Azure
: The machine can be migrated as is to Azure and boots up with full Azure support.

Conditionally ready for Azure
: The machine may have cloud-readiness issues that need remediation. The machine may boot in Azure, but it may not have full Azure support.

Not ready for Azure
: The machine won't boot in Azure and so it can't be hosted on Azure. For example, if a VM has a disk sized larger than 4 TB attached to it, it can't be moved to Azure. However, you can follow the remediation guidelines to fix this issue and move this server to Azure.

Readiness unknown
: These servers lack sufficient data in the vCenter Server for Azure Migrate to determine readiness.

The Migration Phase

Cloud migrations can potentially involve disruptive and costly downtimes. A migration tool, whether from a cloud vendor or a third-party provider, must be able to handle various types of data replication to ensure that a running database can be migrated to the cloud with little or no downtime.

How exactly does one migrate existing VMs, workloads, and applications from on-premise datacenters into the cloud? You can script some of the moves and manually move the rest of the infrastructure. However, the manual strategy isn't very useful when migrating large numbers of VMs and applications.

The smart way is to have custom tools do the migration for you. Some cloud providers, such as Microsoft Azure, have highly specialized migration tools, which you can use on their own or in concert with third-party tools. The third-party tools can also be independent tools. Following is a summary of how you could use multiple migration tools when moving to the Microsoft Azure cloud:

- Use the Azure Site Recovery tool to move Azure-compatible Linux machines that belong to any distribution. In Chapter 4, you learn about using Azure Site Recovery for business continuity and disaster recovery (BCDR). You can also use this tool to manage the migration of your on-premise VMs to Azure. Site Recovery uses replication technology. Therefore, you perform small differential updates after the initial upload to the cloud.

- Use the third-party tool CloudEndure to move a wider range of supported VMs to Azure. As with the Azure Site recovery tool, CloudEndure uses replication during the migration of the VMs to Azure.
- If speed of migration is a key requirement, use a tool such as Velostrata, which quickly moves on-premise VMs to Azure, by replicating just the VM's compute runtime to Azure and replicating the VM's storage slowly over time.

The Optimization Phase

The optimization phase follows the successful migration of your on-premise applications to the cloud. The crucial elements in this phase are performance management and cost optimization.

Optimization encompasses costs, service management, infrastructure, application management, and customer satisfaction.

CHAPTER 3

Principle 2: Define Your Workload's Required Availability

One of the most complex and risky tasks in a datacenter-based application environment is ensuring high availability. Availability could be threatened by fleeting issues, such as network outages or server crashes, or by more long-term issues, such as a datacenter disaster.

Organizations use a wide array of technologies and strategies to guard against the intermittent or long-term unavailability of their systems. Setting up these high-availability systems (such as Oracle Data Guard and Real Application Clusters for databases) isn't a trivial concern, and it can be hard to configure and maintain a bulletproof system.

A big advantage offered by cloud environments is the ease with which you can ensure high availability for your cloud-based computing infrastructure and services. Since all leading cloud providers employ the concept of geographically isolated Availability Zones and regions to run their computing resources, there's built-in capability to guard against disasters and against intermittent hardware and software failures.

Provisioning VMs and running workloads on them shouldn't be your first step when planning for the cloud. Cloud providers use various strategies to enhance the availability and reliability of VMs you run in the cloud. You must plan your VPNs, regions and Availability Zones, load balancers to handle traffic, and many other con-

siderations before you can start working in the cloud. Although some of these concepts are also present in on-premise environments, several of them may be foreign to Linux system administrators.

Azure, for example, uses a capability called *availability sets* to help deploy reliable VM-based solutions. An availability set is a logical grouping construct to ensure that your VMs are distributed across multiple hardware clusters, isolated from each other. Azure guarantees that your VMs in an availability set are placed across multiple physical servers, racks, storage units, and network switches. If a hardware or software failure occurs, your applications remain available to users since only a subset of your VMs is affected.

Availability sets are useful in scenarios, such as the following, which cause VMs to be unavailable or to go into a failed state:

- Unplanned hardware maintenance events
- Unexpected downtime (rare)
- Planned, periodic maintenance events

NOTE

In a cloud platform, failure is always a possibility, with hardware that fails (it can all fail at once, like a rack) and software that needs to be updated. Azure encapsulates this fact in a fault domain or an update domain. Things in a fault domain fail together, and things in an update domain get updated together. So, in an availability set, you want things to span a multiple of those domains.

The Azure platform assigns each VM in an availability set to a *fault domain* and an *update domain*. A fault domain is a set of VMs that share a common power source and network switch. If there's a failure in the fault domain, all resources in the domain become unavailable. After a set of VMs is placed inside an availability set, they automatically get an update domain.

In Azure, you create the availability sets before you deploy your VMs. Let's say you're deploying an application with eight frontend web servers and four backend VMs running a database. Azure lets you define two availability sets for your deployment—one for the

web tier and the other for the database tier. In this scenario, the availability set is distributed across two fault domains.

When you create your VMs, you specify the availability set as one of the parameters, so Azure can create the VMs in the availability sets you specify, isolated across multiple physical hardware resources. Should the physical hardware supporting a VM from the database or the web tier go down, your users don't encounter any issues.

Load Balancing and High Availability in the Cloud

Load balancing is how you spread incoming requests across multiple VMs in your environment. The VMs could be running a web server or a backend relational database. Load balancing enhances availability through the distribution of requests across multiple VMs. Load balancing in the cloud provides the following benefits:

- Helps scale your applications
- Supports heavy traffic
- Automatically detects and removes unhealthy instances and adds instances after they're healthy again
- Routes traffic to the nearest VM

All the major cloud providers offer built-in load balancing. There are two broad types of load balancers—application load balancers and network load balancers. Application load balancers sit between incoming application traffic and computing resources, such as VMs. They monitor the health of the registered targets and route traffic to only the healthy targets. You can add and remove targets, such as computing instances, dynamically from a load balancer, without interrupting the flow of requests to your applications. Network load balancers work at the fourth layer of the Open Systems Interconnection (OSI) model.

The Azure Application Gateway (Layer 7 load balancer) protects web applications against well-known web exploits. The Azure Security Center scans Azure cloud resources for vulnerabilities, such as web apps that aren't protected by the Microsoft Web Application Firewall (WAF), a feature of the Azure Application Gateway. Azure Security Center will recommend an application gateway WAF for

public-facing IP addresses associated with a network security group with open inbound web ports (80 and 443).

Adobe uses Azure Security Center to quickly find and respond to potential vulnerabilities. "Azure Security Center, coupled with Microsoft security and operations management services, allows us to utilize our existing security incident, event manager, and security operation center processes to triage events in Azure," Mike Mellor, Senior Director, Technical Operations at Adobe, says.

Adobe also implements limited automated response. For example, if an unapproved Azure Storage account is granted public access, the Adobe monitoring platform removes permissions until the proper approvals are in place.

Azure global security monitoring provides additional validation and support to further protect the Adobe environment and data. "We consider the Azure global security monitoring group a huge differentiator for Azure," Mellor says. "We point it out to customers. It's a second set of eyes, after our own security team, on our infrastructure."

For more details about Adobe's use of Microsoft Azure for transitioning it's well-known Creative Cloud to a SaaS model, please read "Adobe runs its vast open-source application portfolio in Azure".

AWS offers the Elastic Load Balancing feature, which automatically distributes incoming application traffic across multiple targets, such as its Amazon Elastic Compute Cloud (Amazon EC2) computing instances.

Provisioning the computing resources among multiple Availability Zones supports load balancing and increases the fault tolerance of your applications. You can configure the load balancer to distribute traffic across the registered targets in the Availability Zone where the load balancer is located, or across the required targets in all Availability Zones.

Azure offers the Azure load balancer, a Layer 4 network load balancer that provides high availability by monitoring all the VMs and distributing traffic only to the healthy ones. You define a frontend IP configuration with at least one public IP address, to enable the load balancer and your applications to be accessible across the internet. All VMs connect to the load balancer via a virtual network interface

card (NIC). The backend address pool contains the IP addresses of all the VMs connected to the load balancer via their NICs.

Although you can configure Azure to handle load balancing, it doesn't take care of clustering, which remains an in-OS (or sometimes an in-app) feature. A container orchestrator, such as Kubernetes, can handle the clustering aspects of load balancing, but on-premise setups are likely to be using Corosync, Pacemaker, Red Hat Cluster Suite, or a similar tool.

Although you can do all this in a cloud environment, there are some things you need to consider. In on-premise network infrastructures, since you own everything, you can depend on things such as User Diagram Protocol (UDP) multicast for cluster communications. Without workarounds, like tunneling, those things don't work in a cloud infrastructure, so you need to accommodate for that issue. Also, clustering eventually leads to shared storage. NFS is an option, but in on-premise infrastructures you might have a SAN via fiber channel and similar setups, but you won't have that in the cloud, so you need to examine your choices in this area, as well.

Microsoft Azure uses a VMSS to offer higher availability and scalability. A VMSS supports automatic scaling of instances based on performance metrics. Additional VMs are automatically added to the load balancer, as the load on the VMs increases.

VMSSs are mostly for scale-out. You define one image and then ask Azure to get you many instances of that image. After you do that at a reasonable scale, you no longer care about how those VMs are named, and maybe, even how to Secure Shell (SSH) into them. The VMs should therefore be smart enough to know they're being used like that. If they need to be aware of the other VMs in the scale set, you need to bake in some sort of cluster membership and discovery into them. You can certainly do that by using Azure Instance Metadata Service (IMS). But, of course, this works well for things like three-tier web apps, where you want the web workers to scale out. They get the application code from elsewhere, and they persist elsewhere (such as in a Cosmos DB or a managed SQL database).

NOTE

The Azure IMS provides information about running VM instances, such as SKU, network configuration, and upcoming maintenance events. This information can be used to manage and configure your VMs. IMS is a REST endpoint accessible to all IaaS VMs created via the Azure Resource Manager.

To increase the overall success rate when provisioning VMs, scale sets use *overprovisioning*. Under this strategy, the scale set creates more VMs than you ask for but deletes the extra VMs later. You can configure one or more large-scale VM scale sets behind a highly available IP address, and a health probe will manage and monitor the health and availability of all the instances in the scale set. Figure 3-1 shows how a large number of backend pool instances is serviced by a single load balancer.

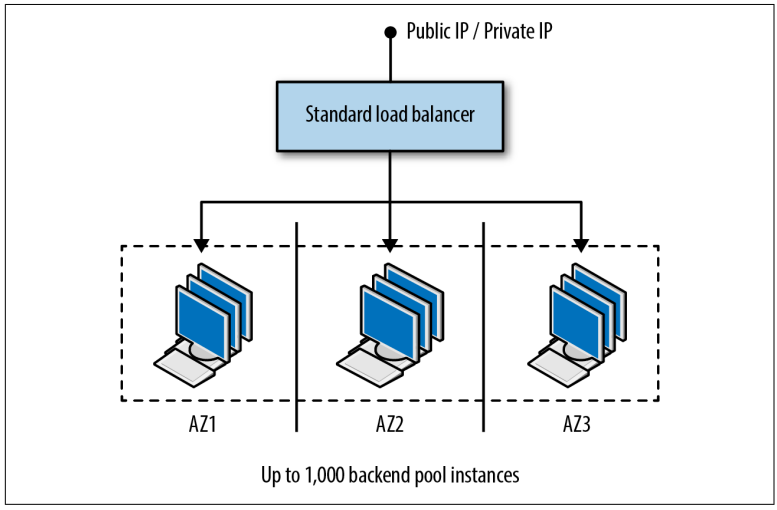

Figure 3-1. How a load balancer services a large number of instances in multiple Availability Zones

Azure load balancing users can tie their frontend IP addresses to a specific Availability Zone or use cross-zone load balancing, by enabling zone redundancy on their public frontends, using a single IP address.

Health Probes

The load balancer must send traffic only to healthy VMs, and thus it needs to monitor the VMs. The Azure load balancer uses health probes to monitor the status of the VMs. You can create a health probe based on a specific health indicator for your application or based on a protocol.

For example, you can create a custom HTTP probe by creating a health check page named checkhealth.js. If the health probe returns anything other than an HTTP 200 OK response, the load balancer will dynamically remove the VM from the load balancer's rotation. The following example shows how to create a TCP health probe named myHealthProbe:

```
az network lb probe create
--resource-group myResourceGroupLoadBalancer
--lb-name myLoadBalancer
--name myHealthProbe
--protocol tcp
--port 80
```

If a health check result is negative, the load balancer will remove the affected VM from its rotation and distribute incoming traffic across the remaining VMs that support your applications. After the maintenance of the VMs is completed, or when you need to expand capacity, you can add the VMs to the backend by connecting the virtual NICs to the backend address pool.

Load Balancer Rules

You control the traffic flow by defining load balancer rules for ports and protocols that map to the VMs. The load balancer rules determine how traffic is distributed to the VMs. In a load balancer, such as the one offered by Microsoft Azure, customers can enable high availability (HA) ports for load balancing on all ports on the frontend of an internal load balancer. This eliminates the need for a large number of individual load balancing rules and makes it simple to set up highly available active-active configurations.

In many cloud scenarios, you also consider a content delivery network (CDN), alongside load balancers, especially for public-facing web applications. Azure Traffic Manager helps enhance application availability by automatically directing traffic to alternative locations when there's a failure. To further protect key applications, Traffic

Manager can distribute traffic across multiple locations, such as multiple cloud services or multiple web apps. You can also integrate Traffic Manager with Azure CDN.

Running Linux VMs in Multiple Regions for High Availability

Most leading cloud providers support highly available applications with the provision of geographical regions. Each region is divided into multiple geographically separate Availability Zones. AWS and Microsoft Azure both provide resiliency through Availability Zones. Companies can choose their business continuity solution and the uptime they need, by placing their cloud computing resources in the appropriate Availability Zones.

An Availability Zone is a fully isolated location within a geographical region. Regions are typically named after the geographical area they represent, such as us-east-2 (AWS), and East US 2 (Microsoft Azure). AWS currently offers a total of 18 regions, and Microsoft Azure has the largest number of regions, 42 (at the time this writing) and, based on history, there will be more regions over time.

Availability Zones are separated from each other, with fast network links between them, and provide redundant power, cooling, and networking. Customers can run mission-critical applications with greater fault tolerance to datacenter failures by running them in multiple regions.

Cloud providers strive to offer faster, safer, private connections for their customers. Azure ExpressRoute, for example, is a high-performance network solution that helps you create private connections between Azure datacenters and infrastructure located in your datacenters or in colocated environments.

Azure ExpressRoute helps you create private connections between Azure datacenters and your on-premise infrastructure or a colocation environment. The connections are reliable and don't go over the public internet. They therefore offer faster speeds and lower latencies than internet-based connections. Due to low latencies, direct connections like ExpressRoute may also lower costs.

Load Balancing Under the IaaS and PaaS Cloud Models

Load balancing is done differently in the IaaS and PaaS cloud deployment models. When you're using a cloud vendor's infrastructure as an IaaS offering, you are responsible for configuring and running your own load balancing services. However, when you're deploying in the PaaS model, the cloud provider is responsible for load balancing your workloads.

Azure and AWS offer load balancing solutions for both IaaS and PaaS cloud models. Autoscaling is done differently in IaaS—you need to design for scale-out and intervene for scale up—versus a PaaS environment, which can autoscale based on application metrics, among other things. Since most scaling options are in the REST API or Azure CLI, you can also automate scaling with your own scripts and logic.

Storage Redundancy Through Replication

All cloud providers offer built-in replication to ensure durability and high availability for data storage. The cloud consumer's account settings determine the exact number of replicated copies. By default, both Microsoft Azure and Amazon S3 store three copies of every disk.

Azure Storage Replication

Let's see how Microsoft Azure replicates its storage. Data that you store in Microsoft Azure is always replicated to serve the twin purposes of high availability and durability. Customers can choose to replicate their data within the same datacenter, in multiple datacenters within the same region, or across regions, with the latter protecting them against a catastrophic failure at a single site.

Azure offers several storage replication options, based on the cloud consumer's needs. The lowest cost option, called locally redundant storage (LRS), still provides at least 99.999999999% (11 9s) durability for the objects you store.

> **NOTE**
>
> Asynchronous data replication have huge implications for data safety. These replications involve a delay, which means that the changes that haven't yet been propagated to a secondary region may be lost if there's a regional disaster and you can't recover data from the primary region.

LRS replicates your data within a storage scale unit, which is a collection of racks of storage units. Data replicas are spread across fault domains and upgrade domains, which represent groups of nodes that are a considered a physical unit of failure or are upgraded together. This ensures data availability, even if a hardware failure affects a single storage rack, or when you upgrade a set of nodes.

LRS however, doesn't offer protection against a datacenter-level catastrophe, such as a fire, in which all your data replicas might be lost. As a result, Azure recommends geo-redundant storage (GRS) replication for most of your applications. For scenarios such as transactional applications that can't accept any downtime, there's also a zone redundant storage (ZRS) replication, in which customers can access data even if a single Availability Zone is unavailable, since ZRS replicates data synchronously across multiple Availability Zones.

Hardware failures

When VMs fail due to hardware issues, cloud providers automatically move the VMs to a different location and restart them. For example, following hardware failure, Azure moves a VM to a different location and restarts it within 5–15 minutes. You can support a higher SLA by deploying two standalone nodes into an availability set.

Dynamic Failure Detection and Recovery in the Cloud

A cloud provider runs a vast array of compute, network, storage, and other resources, in addition to many supporting services such as security and IAM systems. Failure of one or more of the key supporting resources or services could create multiple failures that are beyond the ability of human monitoring and intervention.

In the cloud, traditional monitoring systems and procedures are, in most cases, unusable. Instead, cloud providers set up a dynamic failure detection and recovery architecture that monitors and automatically responds to a set of predefined failure scenarios.

The automatic failure detections system uses a resilient watchdog system which can resolve some issues itself by taking predefined actions in response to certain events and by escalating failure conditions that it can't fix by itself. The intelligent watchdog monitor can escalate an issue through various measures, such as sending console, text, and email messages, and even logging a service ticket to fix the issue.

Load Balancing of Virtual Server Instances

Virtual server instances constitute your compute service in a cloud environment, Cloud systems use a load balancing strategy to figure out the virtual server instances and their workloads, to distribute the processing across multiple physical servers. The cloud provider relies on a live VM migration system to move the virtual machines to the less-loaded physical servers.

Zero-Downtime Architectures

VMs are at the heart of many cloud environments. Since multiple VMs are run from a single physical server, the physical server is a single point of failure (SPOF) for all its guest VMs. Should a physical server crash or otherwise be affected, all the VMs running on that server can become unavailable. However, cloud VMs, in most cases, offer a very high level of uptime (11 nines).

The zero-downtime architecture is a failover system, in which the cloud provider moves virtual servers to a different physical server when the original physical host fails. You can place several physical servers in a server group managed by a fault tolerance system. The fault tolerance system automatically moves the virtual servers around while the VMs are live, thus avoiding any service interruptions.

In addition to strategies such as live migration of VMs, a cloud provider can also follow a strategy of resource replication by creating new VMs and cloud service instances automatically when the VM or a service experiences failure.

Enhancing the Scalability of Web Applications in the Cloud

Cloud providers offer enhanced scalability of web applications that you run in the cloud. Typically, web applications include a website and multiple REST web APIs. Cloud providers employ various architectural strategies and tools to support the enhanced scalability of web applications.

One basic strategy for enhanced scalability is to create the web applications and the web API as separate applications. This enables you to scale the application and the API independently.

Using Caching Strategies and CDNs to Enhance Scalability

Typically, cloud providers employ techniques such as caching and CDNs to enhance the scalability of web applications.

Caching

Caching is a common strategy to improve the performance and scalability of an application. Caching temporarily copies frequently accessed application data to fast storage. The closer the storage is to the application, the faster the application response times will be.

You can use either an open source caching solution, such as Redis, or a cloud provider–owned caching tool to implement a caching strategy that caches session state, semi-static transaction data, and HTML output. For example, Azure Redis Cache is an implementation of the open source Redis Cache that runs as a service in the Azure cloud. Any Azure application can access the Azure Redis Cache service.

Content delivery networks

CDNs serve cached content to users from edge servers that are located near the users, thus saving bandwidth and increasing application responsiveness. CDNs also help handle sudden traffic spikes and periodic heavy workloads, such as those an application may encounter during a major product launch, by reducing load on the application.

Both AWS and Azure offer their own CDNs to help you cache static content. If your applications consist mostly of static pages, you can cache the entire app on the CDN. Otherwise, you can store your static content, such as images, CSS, and HTML files, on a cloud storage, such as Amazon S3 or Azure Storage, and use a CDN to cache the static content.

Reference Architecture for Running a Web Application in Multiple Regions

You can run a web application in multiple geographical regions of a cloud provider to enhance high availability and to provide a robust disaster recovery infrastructure. A multiregional architecture in the cloud provides higher availability compared to deploying everything in a single region. The reference architecture shown in Figure 3-2 uses two regions—a primary region, and a secondary region to achieve high availability. The secondary region is meant for failover.

The Azure Traffic Manager routes requests among the regions. When the primary region isn't available for any reason, the Traffic Manager fails over to the secondary region. This architecture is also helpful when parts of a business-critical application fail. The reference architecture described here uses an active/passive approach with a hot standby. That is, traffic goes to the primary region while the other region is on standby, meaning that all the VMs in this region are always running.

Alternatively, you can have an active/passive configuration with a cold standby, which is cheaper, but the standby region takes longer to take over during a failover. You can also run the two regions in an active/active configuration and direct the incoming requests to both regions, thus balancing the workload. When a region becomes unavailable, you're left with the surviving regions.

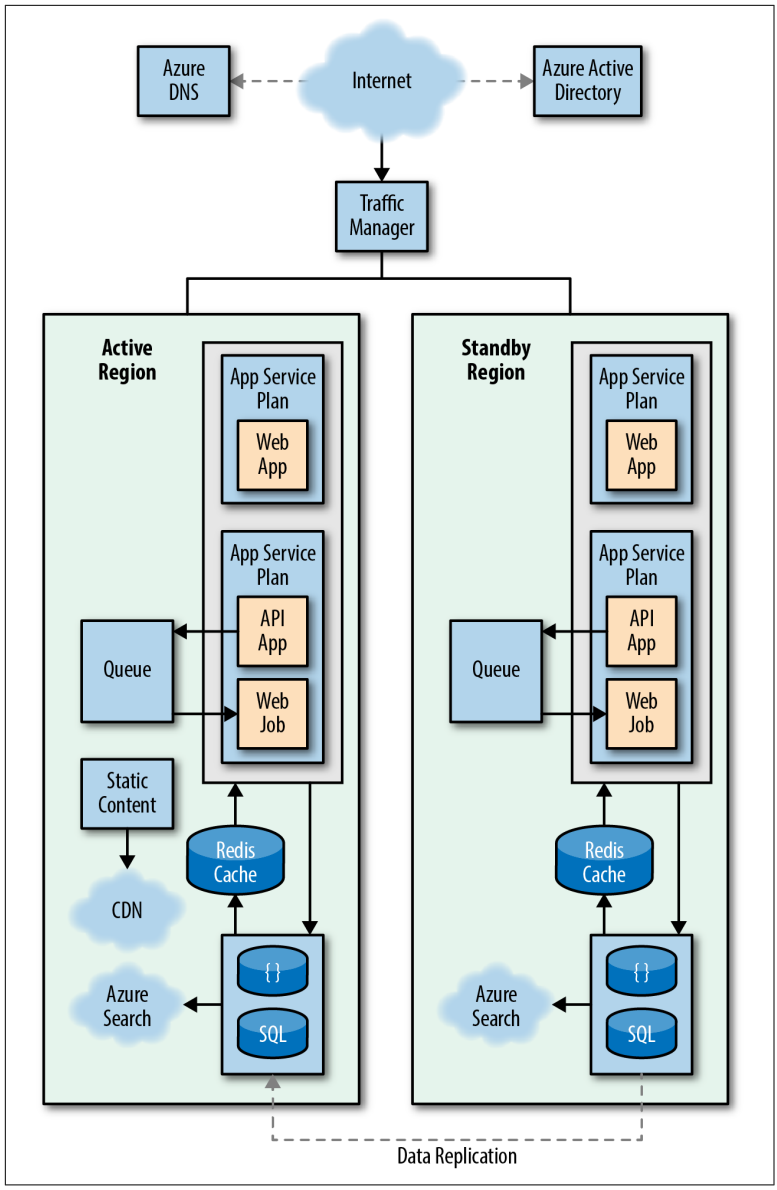

Figure 3-2. Reference Architecture for an Azure App Service application

CHAPTER 4

Principle 3: Monitor Your Applications Running on Linux Across the Entire Stack

Monitoring Linux servers in the cloud is vastly different from traditional monitoring. Cloud monitoring goes far beyond the monitoring of servers. You must not only monitor server performance metrics, such as the standard CPU, memory, I/O, and network performance, but also several other things. Application performance monitoring is at least, as important, if not more so, than server monitoring.

In modern environments, you must also monitor website performance, containers (which are increasingly prevalent in the cloud), and microservices. Monitoring the cloud involves monitoring of various types of cloud servers, beyond the servers that you host in AWS or Azure, or a different cloud platform. Here's a summary of the types of cloud services you may be monitoring in the cloud:

- Servers hosted by a cloud provider, such as Azure or AWS
- Serverless functions, like Azure Functions or AWS Lambda
- Cloud-based SaaS services, such as Office 365, Salesforce, or Adobe Creative Cloud
- Application-hosting services, like Azure App Service, Google Compute Engine, or Heroku

Application Performance Monitoring (APM) and the Cloud

Traditional IT monitoring is focused on monitoring the computing environment—servers, storage, and networks, among other pieces. However, most cloud deployments don't require you to perform these standard monitoring functions. The cloud provider monitors and manages the infrastructure that you're renting, so you don't need to worry about typical IT infrastructure issues, such as servers that crash, disks that fail, and networks that drop packets. All these traditional concerns are gone.

You may not even have any servers or other infrastructure when you are using a cloud environment. For example, you may use a service, such as Azure App Service, to deploy your applications to the cloud. And you may rely on Azure SQL databases and a hosted caching service, such as Redis.

Serverless computing (Azure Functions and AWS Lambda) is a relatively new phenomenon that promises to grow in importance. Serverless architectures help developers deploy applications as chunks of business logic. The cloud provider spins up the necessary computing infrastructure to process the requests for the functions. And it requires no servers at all, because the deployment unit is just code! You don't need to worry about provisioning the servers for running the functions, but you do need to know which requests are being heavily used and which requests are running slowly.

Monitoring your applications rather than your servers and other infrastructure components is key in a cloud environment. Application performance monitoring tools helps to monitor your end-user experience and to provide end-to-end visibility into your application stack. A good APM tool provides deep dive application component monitoring for your enterprise applications. It helps your development, middleware, database, and server experts to troubleshoot performance bottlenecks and to perform root-cause analysis across the cloud infrastructure.

APM tools replace guesswork and reduce your reliance on manual monitoring processes. They help managers understand how the IT services impact their business operations. By monitoring application performance end-to-end and providing insights into capacity utilization, they enable businesses to make sound decisions about

resource allocation. They also help the IT groups monitor how well the applications are meeting their SLAs, thus ensuring a good end-user experience.

Challenges of Monitoring Hybrid Architectures

Multicloud architectures are increasingly popular. The "RightScale 2018 State of the Cloud Report" finds that 81% of respondents have a multicloud strategy. Although enterprises still pursue hybrid cloud strategies (by combining public and private clouds), a clear majority of enterprises have a multicloud strategy.

There are several monitoring related issues that crop up in a multiple cloud and hybrid cloud architecture:

Multiple tools
: A wide assortment of tools adds to administrative overload.

Lack of visibility into resource utilization
: Cloud workloads change often and require you to forecast capacity to determine when you need more server resources.

Dynamic resource provision
: When supporting a dynamic environment, where the number of instances could vary according to changing workloads, monitoring every application and every server isn't a trivial task.

Monitoring Linux VMs and Containers in the Cloud

Monitoring your Linux servers in the cloud offers visibility into the utilization, health, and performance of the applications and workloads that run on those servers. Monitoring helps you proactively fix issues before your users get impacted.

Log Analysis

Collecting and analyzing your systems and application data can offer insights into your cloud infrastructure. Efficient log analysis helps you gain operational insights, with minimal time spent looking for anomalies across the cloud environment.

Linux Server Monitoring

Linux server monitoring helps give you visibility into your cloud deployments. Monitoring in the cloud certain capabilities, including:

- Monitor operational health, and trigger alarms when specified conditions are met.
- Provide application diagnostics.
- Provide diagnostic data to aid in troubleshooting.
- Provide resource utilization statistics.
- Offer a window into application performance.

Monitoring and Tracking API Calls

Most cloud services are offered as APIs, and thus, it's important to track API calls for various services. An API tracking tool, such as AWS CloudTrail, helps with troubleshooting operational issues, supporting security analysis, and contributing to your compliance efforts. AWS CloudTrail can do the following:

- Detect usage behavior patterns by tracking APIs.
- Track the creation, deletion, and modification of cloud resources, such as VMs, security groups, and storage volumes.
- Identify the most recent changes made to resources in the organization's cloud account.

Cloud Performance Monitoring

Organizations often use multiple cloud providers. A multicloud application contains a number of components, with workflows that travel through different cloud providers. All cloud providers offer tools for performance monitoring, but the tools vary in their features and focus.

Performance Benchmarks

To measure system performance effectively, you must compare performance metrics against valid performance benchmarks. Without

this comparison, it's hard for you to tell how current performance compares to the benchmarks, and you won't be able to gauge the severity of potential issues.

Key Linux Server Metrics to Monitor

Monitoring VMs tells you which are overutilized or underutilized. You can then increase or decrease the number of virtual instances or resize the instances to match your workload requirements.

There are basic VM level metrics you must monitor in the cloud to ensure that your servers and services are functioning efficiently. The four most common server metrics to monitor are:

- CPU usage
- Disk I/O
- Memory utilization
- Network performance

CPU metrics

CPU usage has traditionally been the most common performance metric when monitoring Linux servers. You need to receive alerts when server CPUs are reaching their saturation point. The key statistic to watch is the percentage of time the CPU is in use. The raw CPU percentage doesn't tell the whole story—you want to dig deeper and find out what percentage of CPU usage is for running user applications (*CPU user time*) and what percentage is being used by the system (*CPU privileged time*).

I/O performance

Disk read and write metrics help you identify I/O bottlenecks. Cloud providers offer multiple instance types, with each one optimized for specific types of workloads. Some instance types are meant for high I/O-based workloads, and others, for heavy CPU usage-related applications.

If you're running applications that involve high amounts of writes and you notice I/O bottlenecks, you can switch to a different instance type that offers a higher number of input/output operations per second (IOPS).

Memory utilization

Monitoring memory usage is a crucial component of monitoring VMs in the cloud. A low memory condition adversely impacts application performance. Monitoring reveals the amount of used and free memory for the instances. Paging events occur when an application requests pages not available in memory.

In low memory situations, pages are written to disk to free up working memory. The application must then retrieve the page from memory. An excessive amount of paging drastically slows down an application. Spikes in paging indicate that the VM is unable to cope with the requests from the application.

Network performance

Network monitoring shows the rate at which network traffic is flowing in and out of a VM. Network metrics are shown in the statistic bytes per second (bytes received per second and bytes sent per second), indicating the volume of network traffic.

Getting a Unified View of Your Infrastructure

The special nature of cloud environments, especially multicloud and hybrid cloud environments, makes traditional on-premise Linux performance tools inadequate. Well-known Linux open source monitoring utilities, such as vmstat, iostat, top, and sar aren't enough to monitor your servers, because of the dynamic nature of server provision and the inability of these and similar tools to provide you an enterprise-wide, unified view of your cloud infrastructure.

Although cloud providers provide proprietary tools to help you monitor the cloud infrastructure, it may be a good idea to add a third-party tool, such as Datadog, to enhance your performance-monitoring capabilities in the cloud. You can integrate Datadog with Azure. Datadog helps you collect and view infrastructure-wide metrics. It helps you correlate the VM metrics with application-level metrics.

Datadog also helps you collect more metrics than you can access in the Azure portal. You can even integrate third-party tools, such as Datadog, with other third-party tools, such as PagerDuty and Slack, to get automatic alerts. A big benefit of using third-party monitor-

ing tools like Datadog is that they help you add metrics from multiple systems to performance dashboards, thus providing you a comprehensive view of your entire infrastructure, regardless of where the components live.

Serverless architectures, such as Azure Functions and AWS Lambda, use code as the deployment unit.

As mentioned earlier in this chapter, most enterprises (81%) use multiple clouds and hybrid clouds, rather than a single cloud provider. Monitoring these types of architectures presents additional monitoring concerns, such as maintaining application performance as you move them from on premise to the public cloud, understanding application dependencies in each infrastructure, and analyzing the root causes of performance issues. I recommend looking at performance-monitoring tools, such as SolarWinds, to gain an end-to-end visibility of your applications running in hybrid clouds and multiple cloud environments.

Cloud-Monitoring Tools

Although organizations typically pour a lot of effort and money into application development, they don't place a similar emphasis on cloud-monitoring tools. The truth is that, without the visibility and insight provided by powerful cloud-monitoring tools, you don't know exactly how your applications are performing and therefore you don't get the direction for improving the applications.

Cloud monitoring is a catchall phrase and includes a wide variety of tools. Service providers offer their own out-of-the-box monitoring tools, such as Amazon CloudWatch and Microsoft Azure Monitor. However, these tools may not be adequate for many cloud consumers, especially those with multicloud and hybrid cloud architectures.

Cloud-monitoring tools can be in-house tools offered by the cloud provider or tools offered by independent SaaS providers. Cloud monitoring is increasingly being offered as a fully managed on-demand service, with the service providing the tools for monitoring both cloud and on-premise infrastructures and web applications. The cloud monitoring is delivered through a SaaS-based software

that tracks performance across the entire cloud stack. Cloud administrators and development teams can review the performance statistics in a central dashboard, and they can get alerts about performance issues through email, and SMS, among other options.

Cloud proprietary and third-party monitoring tools can also work well together. There are specific advantages in using the two types of tools. Cloud provider monitoring tools are preinstalled and preconfigured, so they're ready to use, out of the box. SaaS monitoring tools have the advantage that they help monitor more than one type of cloud infrastructure, so they allow you to monitor all your applications and services from a single point.

New Relic, SolarWinds, and PagerDuty are some of the well-known cloud provider and third-party monitoring tools. All leading cloud providers offer built-in monitoring tools, as I explain in the following sections.

Amazon CloudWatch

Amazon CloudWatch us a tool offered by AWS that helps you monitor application metrics, log files, and react to changes in your AWS resources.

Google Stackdriver

Google Stackdriver offers monitoring and logging for applications that you run in the Google Cloud and in AWS. Although Stackdriver is natively integrated with Google Cloud Platform cloud products, it lets you aggregate data across cloud platforms.

Microsoft Azure Monitor

Azure Monitor is part of the overall Azure monitoring solution. Azure Monitor enables core monitoring for Azure services by allowing for the collection of metrics, activity logs, and diagnostic logs. It helps you track performance, maintain cloud security, and identify trends. In addition, Azure Advisor monitors resource configuration and usage telemetry and offers personalized recommendations based on best practices. Azure Application Insights help you to monitor the availability, performance, and usage of your cloud-based applications and to proactively identify and diagnose errors.

The Importance of a Comprehensive Monitoring Solution

An effective monitoring solution must help you do the following:

- Understand the detailed operation of your infrastructure components.
- Understand how your application components perform.
- Enhance the availability of your applications with proactive notifications about critical issues.
- Integrate with other tools to alert—and even fix—problems discovered by the monitoring.

No single tool can do everything in the monitoring space. You should integrate multiple monitoring services to deliver a comprehensive solution that helps you continuously assess the performance, availability, security, and health of your infrastructure and the applications that run on it. Figure 4-1 shows one such solution, offered by Microsoft Azure, in which multiple components work together to monitor Azure resources.

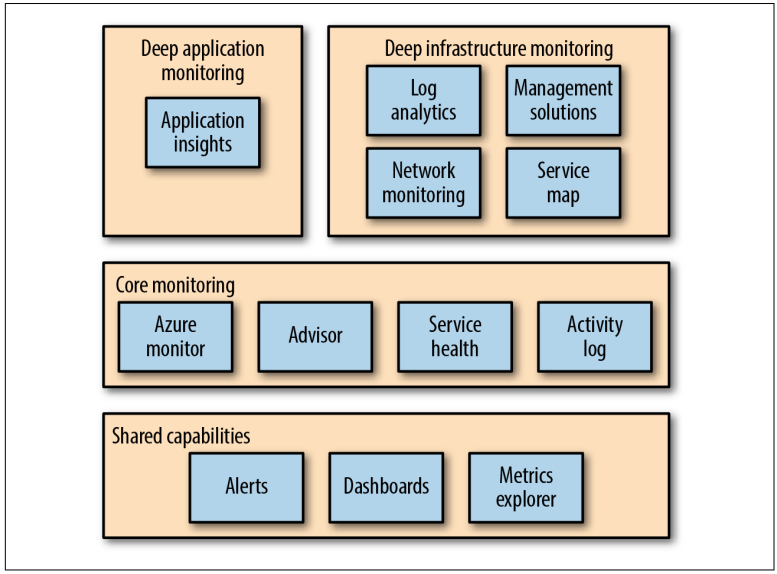

Figure 4-1. An example of a comprehensive monitoring solution

In a comprehensive monitoring solution, such as the one shown in Figure 4-1, the various services work together to collect, analyze, and act on telemetry from your cloud (and on-premise) infrastructure and the applications that you run on that infrastructure.

Best Practices for Cloud Monitoring

It's a good idea to follow some basic guidelines when monitoring the cloud:

Identify the right metrics
: There's no benefit in monitoring every activity in your cloud environment just because you can. You must monitor the key metrics that have a direct bearing on your goals.

Support autoscaling
: *Autoscaling*, which is the automatic adjustment of computing capacity to meet the changes in the cloud workloads, is a salient feature of cloud computing. The cloud-monitoring solution should be able to handle autoscaling, since it needs to monitor the continuously changing number of application instances.

Monitor the user experience
: The purpose of all your cloud activity is to provide a good user experience. By monitoring user-related performance metrics, such as response times, you can improve user experience.

Gather uniform metrics
: As mentioned earlier, many enterprises use a hybrid cloud environment, so they need to monitor both on-premise and cloud services. Uniform metrics help when displaying performance data from a variety of sources in a single location, such as a performance dashboard, providing you a comprehensive view of performance.

Monitor cloud service usage and costs
: A good monitoring solution should help you to track your resource usage in the cloud and to reduce costs.

CHAPTER 5

Principle 4: Ensure Your Linux VMs Are Secure and Backed Up

Probably the most important system management task is securing your data from threats. Following security is data backup, which is the backbone of a successful disaster recovery (DR) solution in the cloud. There are multiple ways to set up such a DR solution, depending on your recovery objectives.

This chapter explains the shared responsibility security model in cloud environments, the principles of backup and recovery in the cloud, and DR strategies in a cloud environment.

Security in the Cloud

A question asked by many potential cloud customers is, "How does the cloud provider help me ensure the security of my data?" Security in the cloud is somewhat different than security in an on-premise datacenter. When you move to the cloud, you share the security responsibilities with the cloud service provider.

Until recently, security concerns about cloud-based deployments have kept some potential organizations from moving to the cloud, but things have come full circle. Today, one can make a strong case that a key reason for moving to the cloud is the enhanced security provided by cloud deployments.

NOTE

According to the "RightScale 2018 State of the Cloud Report," security is a challenge for 77% of respondents. It is the largest issue for enterprises starting out with the cloud. For intermediate and advanced users, cloud costs are the bigger challenge.

A Shared Responsibility Security Model in the Cloud

Cloud providers, such as AWS and Microsoft Azure, follow a risk-management model of shared responsibility with their customers. Under this security model:

- The cloud provider is responsible for the infrastructure platform and strives to provide a cloud service to meet the security, privacy, and compliance needs of its customers. A cloud provider, such as Azure or AWS, is responsible for securing the underlying infrastructure that supports the cloud.

  Most cloud providers offer their global infrastructure in separate geographical areas, called regions, within which they create multiple Availability Zones. In addition, the cloud provider also maintains edge locations near its users, to enhance user experience and reduce latency. The cloud provider must secure all the datacenters in the various regions, Availability Zones, and edge locations.

  Both AWS and Azure offer several cloud-based managed services, such as database and analytical services. The cloud providers are responsible for securing all their fully managed services, along with the compute, storage, and networking products.

- As the cloud customer, you are responsible for the environment, after the cloud service is provisioned, and for all data that you place in the cloud. The customer identifies the necessary controls for their business and then implements and configures those controls to satisfy security and compliance with all applicable regulations. The shared security responsibility model that all cloud vendors follow diminishes your operational burden in several ways. It often improves your default security posture, with minimal extra work on your part.

Figure 5-1 demonstrates the typical shared security responsibility model in the cloud. The following sections explain how the cloud

provider and the cloud customer divide up the security responsibilities.

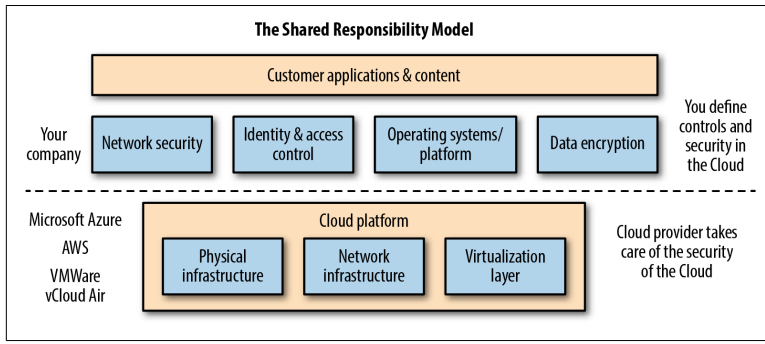

Figure 5-1. The shared security responsibility model in the cloud

Security in the Cloud

The customer is responsible for security inside the cloud environment. This includes:

- The security of the customer data
- The platform, applications, and IAM
- Operating systems, network, and the configuration of network firewalls
- Data encryption (client side and server side)
- Security of the network traffic (encryption, integrity, and identity)

The amount of work that the cloud customer must do in the security area depends on the specific cloud services that the customer chooses and on the sensitivity of their data. Regardless of the specific cloud services that you use, you'll always be responsible for security features, such as user account management and SSL/TLS of transmitting data, among others.

Security of the Cloud

The cloud provider is responsible for the security of the overall cloud environment, such as:

- Service level security—the computing infrastructure, storage, databases, and networking

- Global infrastructure, such as regions, Availability Zones, and edge locations

Let's review the two areas of security that fall under the cloud provider's domain—service security and global infrastructure security.

Service Security

In addition to being responsible for the security of the cloud infrastructure, the cloud provider must securely configure all the managed services (such as databases and big data analytical services) that it offers its users. The cloud provider handles all security tasks for the managed services, such as hardening the operating system, patching the databases with the latest security updates, configuring firewalls, and setting up DR plans. The cloud customer only needs to resolve two things with the managed services:

- Configure logical access controls for the cloud resources it uses.
- Secure its account credentials.

Global Infrastructure Security

The cloud provider is responsible for securing the global infrastructure that underlies all the services it offers. Infrastructure includes the hardware, software, networking, and all the physical facilities that support the cloud services. The cloud vendor must provide its customers reports from third-party auditors who have verified the provider's compliance with a set of required security standards and regulations.

Security Concerns Due to Shared IT Resources

When you move to the cloud, the concept of data security changes right away. You immediately start sharing the responsibility for securing your data with the cloud provider. Since your IT resources are being used remotely, you (as a cloud consumer) must expand the trust boundaries to include the public cloud. Establishing a security architecture that spans a large trust boundary introduces various security vulnerabilities.

These vulnerabilities can be avoided if the cloud consumers and the cloud providers support compatible security frameworks. But in a public cloud, you can't ensure this.

The cloud provider always has privileged access to the data you store in the cloud, since the trust boundaries overlap in the cloud. The security of the data depends on the security policies and access controls enforced by both the cloud consumer and the provider. Since most cloud IT resources are shared among users, it creates a potential source of data exposure to malicious cloud consumers.

NOTE

SLAs and cloud provider reliability

One of the key things a cloud customer must evaluate is the reliability of the cloud provider—how well does the cloud vendor maintain the guarantees they offer in its cloud service SLAs? Since your cloud solutions rely on the cloud services, an unreliable cloud vendor could jeopardize your business. Remember that the farther the cloud consumer is from the cloud provider, the greater the number of network hops. This means a greater potential for latency that fluctuates, in addition to possible network bandwidth limits.

Cloud Security Tools and Mechanisms That Contribute to Better Security

There are many tools and mechanisms to help cloud consumers secure their resources, meet security standards, and strengthen their security posture. Following is a brief description of the key security tools and mechanisms in the cloud.

Strong Network Security

Both AWS and Azure offer built-in firewalls (security groups) that enable cloud consumers to control network access to their computing instances. The cloud provider may also offer a private or dedicated connectivity option to connect the cloud consumer's office and on-premise environments with the cloud environment. Encryption of data in transit is common when you transmit data to and from the cloud.

Configuration Management Tools

Some configuration management (CM) tools in the cloud work like Chef and Puppet, which are well-known CM tools. Both Chef and Puppet help you to automate infrastructure tasks. Other tools track changes in configurations made by users to the cloud resources.

- Inventory and CM tools help identify and track resources.
- Template definition and management tools help users create standard, preconfigured virtual servers.
- Deployment tools manage the creation and decommissioning of resources according to an organization's standards.

Access Control

Cloud providers generally excel in the provision of strong access control mechanisms. Most cloud providers rely on centralized IAM to manage users, security credentials (like passwords, access keys, and permissions), and authorization policies that control which resources and services users can access.

Virtual Private Clouds

VPCs allow cloud users to provision a logically isolated section of the broader cloud where they can launch their own resources in a private network. The organization that creates a VPC has full control over its virtual networking environment and can select its own IP address ranges, create subnets, and configure its own route tables and network gateways. Both AWS and Microsoft Azure let you leverage multiple security layers to protect your resources in the cloud. You can create security groups and network access control lists to control access to the cloud server instances that live in the private subnets.

Disaster Recovery in the Cloud

DR is about preparing for and recovering from a disaster that could affect your IT operations. You can classify any event that adversely impacts your organization's business continuity as a disaster. Disasters can include hardware or software failures, power outages, and

physical damage caused by a fire or a flood. A disaster could also be the result of human error, such as accidental deletion of data.

Elastic and speedy provisioning of computing power is a primary reason for the success of cloud computing. Just as you modernize your server inventory with low-cost, virtual servers that you provision on demand, and just as you provision low-cost cloud storage of various types (object, block, or archive, for example), you must also modernize your backup solutions to take advantage of the elasticity and flexibility offered by a cloud environment.

Since a cloud environment usually requires you to transfer and store large amounts of data, cloud users must learn to leverage the provider's elastic cloud computing features to efficiently perform the data transfer and storage.

Recovery Time Objective and Recovery Point Objective

In the cloud, as in a local datacenter, you must be aware of two key terms when planning for a disaster:

Recovery time objective (RTO)
: This is the time it takes to restore a business process to its established operational service levels, following a disaster. If, for example, your RTO is four hours, and disaster strikes at 4 p.m., you should be able to restore your business processes to acceptable service levels by 8 p.m.

Recovery point objective (RPO)
: This is the acceptable amount of data loss, as measured by time. That is, if your RPO is one hour, and a disaster occurs at 2 p.m., your system must be able to recover data all the way up to an hour before disaster struck—in this case 1 p.m. Your data loss will, at most, be over the span of the one hour (between 1 p.m. and 2 p.m.).

An organization must determine the acceptable RTO and RPO levels, based on the business impact of a system downtime. Following this, the organization plans a DR solution to provide system recovery within the RTO and RPO that it chose.

Traditional DR Strategies Versus Cloud-Based Strategies

Traditional DR strategies depend heavily on off-site duplication of both the infrastructure and data. The companies set up their critical business services and infrastructure offsite and test the DR systems at regular intervals. It is standard DR practice to locate the DR solution at a considerable geographical distance from the primary site of operations, to isolate the DR site from the disaster that may impair the functioning of the primary business systems.

Setting up and maintaining offsite DR solutions can be challenging. Your DR infrastructure should virtually mirror your production systems and include the following:

- Infrastructure facilities, such as buildings, power, and cooling
- Physical security of the DR site
- Agreements with internet service providers (ISPs) for the provision of internet connectivity for the DR environment for an indefinite period
- Sufficient processing and storage capacity to last until you can get back to the restored primary infrastructure

Types of Disaster Recovery Solutions

There are three main types of DR solutions—real-time data replication, off-site storage with tape vaults, and a secondary appliance in the DR site.

Real-Time Data Replication

Real-time data replication is for critical workloads, and you configure a failover from the primary production application to a secondary DR site, where a mirror image of the production system is on standby to take over at short notice.

This solution offers the shortest RPO and RTO. However, it is expensive, since it requires personnel with expertise in complex disaster recovery and standby systems. In addition, you must maintain an additional facility with all the (idle) standby resources ready to take over from the primary site in the event of an outage.

Off-Site Storage with Tape Vaults

An off-site tape storage service-based DR strategy is vastly less expensive than maintaining a standby DR site, but this solution offers much lower RTOs and RPOs. Following a disaster, you must set up your alternate infrastructure and configure it before you can restore the data on tape storage, all of which takes considerable time.

Backup Target Appliance in the DR Site

Several organizations replace tape vaults by backing up their data to a secondary appliance on their DR site, instead of using tape storage. The deduplication solutions come with a longer RPO than real-time replication, but if you don't need a real-time RPO, it's a cost-justifiable DR solution. The RTO here is much higher than what's offered by a tape vault-based DR solution, but you get this at a much higher cost than that of a simple tape storage solution.

The prevalence of the IaaS cloud has led more organizations to consider cloud storage as an alternative to traditional tape-based DR strategies and to the use of a standby secondary deduplication appliance.

Why the Cloud May Offer Better DR Solutions

Although one might think that it's harder to set up a DR solution for your cloud-based systems, these systems come with several built-in advantages in this regard. Following are some of the standard cloud features or services that enable effective disaster recovery in the cloud:

Elasticity

Elasticity is one of the calling cards of a cloud environment. It's very easy to provision compute, storage, and networking services in the cloud. You can provision, configure, and start up huge amounts of these services within a few minutes. The elasticity offered by cloud compute and storage is ideal for the infrequent requirements for infrastructure in a DR solution.

Virtualization

A cloud computing system relies on virtual servers, which are easy to copy and back up to off-site DR datacenters, and you can quickly

spin them up in a few minutes on their new hosts. This helps to dramatically reduce the RTO time as compared to a nonvirtualized conventional DR solution, where you must load the servers with the OS and ensure that all patches are applied before you can restore data to the new servers.

Using the same cloud datacenter as a DR site for the production infrastructure and for the backup infrastructure helps in faster site recovery.

Advantages of a Cloud-Based DR Solution

Unplanned downtime can devastate your reputation and revenue, and it can put you of out compliance with key regulations. Even short downtime periods can cost an organization significant amounts of revenue. Legacy local backups and secondary site replication in complex hybrid cloud environments make it harder to perform valid DR drills, and they expose you to considerable risk. It's common to experience loss of data and to encounter application errors when you failover to a standby datacenter during a disaster recovery.

Instead of relying on conventional backup and DR solutions, you can employ a modern data protection platform that can keep your business running with zero downtime, by moving all backups and DR to the cloud. In complex IT environments, you need to worry not just about recovering data but also about ensuring that all your web applications and business services living in multiple on-premise datacenters and cloud locations are available. Cloud-based backups and DR solutions are fast replacing traditional tape-based solutions.

You can ensure application availability with cloud-based recovery solutions such as Azure Site Recovery. BCDR solutions, such as Site Recovery, reduce application downtime during an IT server interruption by offering three major benefits: simpler of management, lower costs, and reduced downtime.

Simpler Management

One of the big advantages of using a cloud-based DR solution is that you don't need to perform any patching to bring the DR site up-to-date. The DR site is automatically updated, and you can minimize

recovery issues by sequencing the order of multitier applications running on multiple VMs. You can test your DR plans without adversely impacting your production workloads.

Lower Costs

A cloud-based BCDR solution allows you to lower your infrastructure cost by using the cloud as the secondary site for running your business during outages. You can avoid datacenter costs by moving to the cloud and taking advantage of the geographical regions offered by cloud providers and setting up DR between those regions.

Infrastructure costs are lower for a DR solution in the cloud because you pay just for the resources you need to run your applications in the cloud during outages. In addition, you can use automatic recovery to the cloud to keep your on-premise applications available during outages.

Reduced Downtime

Cloud-based BCDR solutions, such as Azure Site Recovery, can offer best-in-class RPO and RTO. Cloud-based BCDR solutions should offer very highly dependable SLAs. Instead of waiting weeks to replicate the infrastructure and recover data following a disaster, you can expect immediate recovery—at no additional cost.

How the Cloud Shifts the DR Tradeoffs

A cloud-based DR solution offers a lower RTO compared to that offered by a dedicated DR site but at a greatly reduced cost for servers and storage. Infrastructure cost is lower because the compute and storage are powered on only during disaster recovery.

There is an inherent trade-off between the RTO and RPO you choose and the cost of achieving that RTO or RPO. The lower the RTO, the higher the cost, in general, since you need to maintain sufficient appropriate infrastructure at the DR site, usually in a warm or hot standby mode, ready to take over on short notice. The readier the DR site needs to be, the higher the cost. Cloud computing can deliver a faster recovery time (shorter RTO) at a much lower cost than conventional DR strategies based on the maintenance of offsite DR solutions. Figure 5-2 shows how the DR trade-off curve

shifts to the left in the cloud, giving you faster recovery times at a significant reduced cost.

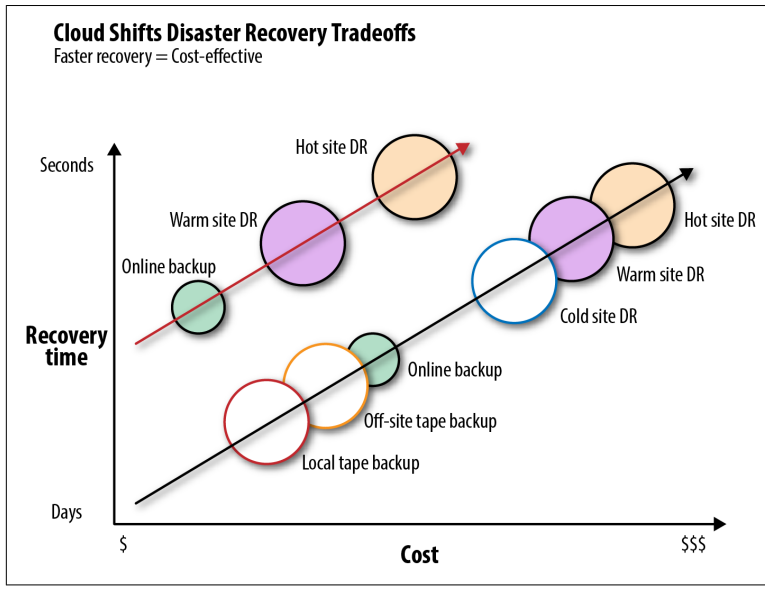

Figure 5-2. How the cloud shifts the DR trade-off

Large numbers of server instances and a wide variety of operating systems make it difficult to protect against unplanned downtime in the cloud. This can also wreak havoc on your business reputation, with regard to regulatory compliance issues.

Legacy backups within the datacenter and secondary site replication for DR can't adequately protect all your data in the cloud, since you need to backup and secure so many data sources. This variety of data sources also means that you need to use multiple data backup and DR solutions. Traditional tape-based backups have been steadily eclipsed by cloud-based data protection for backups and DR. Modern cloud-based data protection strategies provide a wide range of data protection functions, beyond merely backing up data, including replication, deduplication, and centralized management

Using a Backup Service

Instead of backing up your cloud-based VMs yourself, you can take advantage of a cloud-based backup service (SaaS). Instead of building expensive backup storage systems, you can simply outsource all

backups to the service. Managerial costs are lower (no tape costs, for example) as are management costs when you contract with a backup service.

Backup services don't merely back up your data. They offer several powerful features, such as ease of use (no scripts to back up your data), central dashboards to manage the backups, and the ability to export backup reports, to help in your compliance efforts. The most important protection offered by a good cloud backup service is the security of the data that's backed up. The service allows you to set up access controls to ensure that only authorized users perform backup operations.

Microsoft Azure offers Backup to protect your data. Azure Backup is a pay-as-you-go service that offers you flexibility as to the data you want to protect and as to the length of time for which you want to retain the backups. Backup helps you restore your VMs and physical servers, both in the cloud and on-premise, free of cost.

If Azure Backup detects unauthorized deletions, it retains the data for several days, allowing you the opportunity to recover the deleted data.

You can use Azure Backup along with Azure Site Recovery to store backups in Azure and to replicate the workloads to Azure rather than to a secondary site. Using these two services together simplifies the building of DR solutions. Both tools support hybrid environments. Azure Site Recovery helps you replicate workloads to Azure rather than to a secondary site, eliminating the need for a dedicated secondary datacenter. It offers a better solution than that of running a secondary site which replicates data to the cloud.

Azure Site Recovery stores the replicated data in Azure Backup, and when a failover is required, it creates Azure VMs based on the replicated data. You can set your own RPO thresholds to determine how often Azure creates data recovery points. Azure Site Recovery reduces RTOs, with its automated recovery process. You can test failovers to support DR drills without adversely affecting your production systems.

CHAPTER 6

Principle 5: Govern Your Cloud Environment

Although cost efficiencies, elastic resource provisioning, and speed of deployment maybe driving forces of cloud adoption, security and regulatory compliance are often major concerns around cloud deployment. A strong governance framework ensures the review of service levels, manages risk effectively, and certifies that your critical business data is secure—and that you comply with legal requirements and business-specific certifications and attestations.

The essential challenge of cloud compliance is that the customer places vast amounts of sensitive data into the hands of the cloud provider who controls the facilities, thereby trusting the provider to safeguard their data. The customer must do this while also being subject to stringent regulatory requirements (for example, the health industry's HIPAA and the financial industry's PCI DSS) and security standards.

NOTE

Traditional IT management followed industry best practices, such as the Information Technology Infrastructure Library (ITIL), which were developed prior to the emergence of cloud computing as a fundamental way of doing business. Standard ITIL processes, such as Service Catalogs and Service Design, require a lot of adaptation when you move to the cloud.

Data encryption and auditing of the provider's datacenters might seem to be obvious solutions to the cloud consumer's quandary. But the nature of the cloud environments, where encryption may even hinder processing activity, and the infeasibility of consumer auditing of the cloud service provider's datacenters mean that these solutions aren't practical in most cases.

Governance and Compliance in a Cloud Environment: The Issues

IT governance in the cloud, just as governance in your own datacenter, is the application of policies relating to the use of your cloud resources. It requires you to define the organizing principles and rules that you must adhere to while working in a cloud environment. IT governance in the cloud aims to ensure that:

- Your cloud infrastructure and applications are implemented and used according to specific policies and procedures.
- The cloud assets are adequately controlled and managed.
- The cloud assets support your organization's business priorities.

Cloud environments pose special challenges around operational governance, multiregional compliance and legal issues, accessibility, and data disclosure regulations, as described in the following sections.

Reduced Operational Governance in the Cloud

In the cloud, the customer is usually given a lower level of governance control than what they are accustomed to in on-premise datacenters. Of course, this leads to the question of the potential risk, which depends directly on how securely the provider runs its operations in the cloud. Additional risk stems from the extra connections that you need to set up between the cloud provider's infrastructure and your organization.

You can mitigate governance risk by combining legal contracts, SLAs, technology inspections, and appropriate monitoring. Different cloud delivery models offer you varying degrees of operational control. Cloud computing, regardless of whether it is IaaS, PaaS, or SaaS, follows the general cloud computing model of offering cloud

products "as a service." SLAs help you establish a cloud governance system. You, as the cloud consumer, must evaluate and track the service levels offered, along with the operational guarantees proffered by the cloud provider.

Shared Resources in the Cloud

In a cloud environment, you don't normally have exclusive access to your own dedicated physical infrastructure. You have little or no visibility about the way that the provider segments the physical resources. You may not know which security controls the provider has in place to secure your data in colocated cloud resources.

You may audit your own data in the cloud, but you don't have insight into the provider's security and compliance controls. Although providers claim that they provide adequate separation between the virtual instances that you use in the cloud, you really don't know.

Multiregional Compliance and Legal Issues

When a cloud provider hosts your data, you don't know the exact geographic location where it is stored. If your organization must comply with data privacy and data storage policies that prohibit the movement or storage of company data outside specific jurisdictions, you may easily fall out of compliance with the regulations. For example, there are laws in the United Kingdom which require that personal data belonging to UK citizens be stored only within the UK.

Accessibility and Data Disclosure Regulations

A key legal issue regarding the storage of and access to data are the regulations that govern access to and disclosure of data. Many countries have strict regulations requiring the disclosure of specific types of data to a government agency. If you store data that belongs to a European customer in the US, the US government agencies may access that data with greater ease than their governmental counterparts in the European Union countries.

NOTE

Many regulating agencies recognize the ultimate responsibility of the cloud consumer for the storage, security, and integrity of their data, eventhough the data is stored in a cloud vendor's infrastructure.

Mobility and Multitenancy

Unlike in an on-premise datacenter, cloud computing resources move around. Critical business data may also move through the cloud. Security and compliance policies also need to move along with the resources and the data, which poses special challenges in adhering to associated regulations.

Identity and Access Management in the Cloud Is Different

Depending on the type of cloud service delivery model (IaaS, PaaS, or SaaS) one adopts, providers support different IAM controls. For example, in a public SaaS model, you may have to accept the provider's authentication controls. Your own strong authentication mechanisms, such as digital certificates, may not be supported by the provider. In terms of authorization policies in the cloud, a cloud provider may not support the definition of detailed roles or fine-grained authorization policies.

Encryption and Compliance in the Cloud

Organizations that must meet industry compliance requirements, such as HIPAA compliance for healthcare, SOX for financial reporting, and PCI-DSS standards for ecommerce and retail business, must consider encryption as a best practice. Even if your risk of losing data is small, encryption can help. If you ensure that the encryption keys aren't stolen, the loss of the encrypted data itself isn't considered a reportable security event.

Cloud providers, such as AWS, offer managed services to simplify the creation, control, and management of your encryption keys. AWS Key Management Service (KMS) provides a centralized view of all the key users in the organization. It also uses a hardware security module (HSM) to enhance key security. KMS also integrates with AWS CloudTrail to provide a log that shows key usage across the organization, thus satisfying several key regulatory and compliance

requirements. The HSMs are designed for governmental and other standards that ensure secure key management. You can generate encryption keys and manage and store them such that they are only accessible by you. You can provision a cloud HSM instance in AWS within your own VPC, with an IP address that you provide.

NOTE

Securing your cloud environment isn't a passive task. Security and compliance must be actively managed by the cloud consumer.

The Fundamental Pillars of a Secure and Compliant Cloud Service

A cloud service provider must satisfy the following four fundamental requirements for its consumers.

Security

IT managers are concerned about potential vulnerabilities in the cloud compared to their on-premise security. The cloud provider must safeguard its customers' data with rigorous security controls and state-of-the-art security technology, including vulnerability assessments and data encryption.

Compliance

The cloud provider must enable its customers to satisfy a wide range of governmental and regulatory agency compliance standards, both domestic and international, as well as industry certifications and attestations.

Compliance road maps continuously evolve, and cloud users must be assured that the cloud provider's compliance strategies are also evolving over time to meet increasingly stringent standards and regulations.

Privacy and Control

Businesses worry about the unique privacy challenges of storing data in the cloud. They anticipate a loss of control over the storage, access, and usage of their cloud-based data.

Although the cloud provider has physical control of its customer data, the customer is the ultimate owner of its business data. Thus, the customer gets to determine the privacy levels of data access, by controlling access to the data.

Transparency

The cloud service provider must enable its customers to have full visibility into their data, such as the locations where the provider stores the data and how the provider manages it. Businesses must be able to independently verify the storage, access, and security of their data.

Strategies and Tools for Enhanced Governance in the Cloud

Businesses must feel comfortable that the cloud service provider runs a well-managed cloud environment that complies with all their internal policies and with external regulations. Cloud providers use several strategies and tools to enhance cloud governance, such as the identification of noncompliant resources in the cloud, in addition to security assessments. Following is a brief description of the various security and compliance–related strategies.

Cloud users must continuously monitor their resource configuration to ensure that it doesn't have any security weaknesses. A key requirement is the inventorying of all cloud resources and their configuration attributes. The ability to quickly identify recent resource configuration changes is critical to secure yourself in the cloud. Cloud providers offer managed services that help you inventory your cloud resources and audit the resource configuration history. These tools, such as AWS Config, also notify you in real time about any configuration changes made to your cloud resources.

Security Policies and Processes to Enhance Governance

Cloud providers can employ stringent controls, such as the following, to enhance security and governance in the cloud:

Role-based access control

Role-based access control (RBAC) is a well-established, fine-grained access management technique, which ensures that you give users

only the specific access privileges they need to do their job. Role-based security polices reduce the risk of exposing critical business data to security attacks by eliminating unrestricted access permissions to all users.

Networking controls

Network access in a hybrid cloud environment can include both internal and external (internet-based) network access. VPNs in the cloud, such as Amazon Virtual Private Cloud and Azure VN (VNets), logically isolate a part of the public cloud to help keep a business from the rest of it.

Network security groups are virtual firewalls that consist of rules that control the flow of network traffic by specifying how a cloud resource, such as a VM, can connect to the internet or to other subnets in a virtual network.

Hierarchical account provisioning

Defining account hierarchies is a core governance structure that limits the use of cloud services within the customer's business. For example, in Azure, enterprise customers can divide the cloud environment into departments, accounts, subscriptions, resource groups, and finally, individual resources.

Security Assessments in the Cloud

Continuous security assessments in the cloud are essential to mitigate vulnerabilities and reduce the probability of attacks from malicious actors. Amazon Inspector is an automated security assessment service that helps a cloud customer improve the security and compliance of the resources and applications that they deploy on AWS. The tool can automatically test applications for security vulnerabilities or for deviations from best practices. It can also produce remediation steps as part of its security assessment report.

Tools like Amazon Inspector employ a knowledge base of security rules, which is continuously updated by security researchers. The rules are mapped to common security standards, such as PCI DSS, as well as to formal security vulnerability definitions. For example, a rule may check whether remote root login is enabled. Another rule may check whether any vulnerable software versions are installed.

Using Geo-Specific Services

Legal and regulatory requirements, such as data privacy and sovereignty laws, mean that a business can unwittingly breach a regulatory requirement by sending its data across the globe. Providers can ensure that the cloud consumers satisfy the requirements by offering *geo-specific services*, that is, services where operations are confined to specific jurisdictional boundaries.

Ideally, the cloud provider must offer its customers an easy way to view the security status of their cloud resources and must provide automatic recommendations to help prevent security breaches. One helpful tool is Azure Security Center, which offers integrated security monitoring and policy management across your Azure cloud infrastructure. And it helps detect security threats.

Azure Security Center is a combination of best practice analysis and security policy management for your Azure cloud resources. It automatically collects and analyzes all the security data from your cloud resources and from other security solutions, such as firewalls and anti-malware programs. Security Center offers the following capabilities:

- Visibility into the cloud security status, such as event detection
- Centralized policy management
- Continuous security assessments and actionable recommendations
- Adaptive application controls
- Prioritized alerts and incidents
- Enabling of control and governance through policies

Trusting the Cloud Service Provider

When you work in a cloud environment, you aren't simply renting IT infrastructure and services. You're engaging a service to which you are entrusting the management of critical business assets and services, without complete visibility into the operations of the cloud provider. You must ensure that there is a satisfactory level of transparency in the provider's operations.

In on-premise or outsourced environments, you gain visibility through internal or third-party audits. In the cloud, traditional auditing isn't feasible, since you're dealing with an infrastructure that's spread throughout the world.

It's impossible for a cloud provider to allow thousands of its customers to inspect its datacenters and to audit its regulatory compliance. In fact, this would, itself, constitute a security risk that would adversely affect its customers. Therefore, you need alternative methods of gaining visibility into the security and control mechanisms that are in place. Cloud providers recognize the need to establish trust with their customers and are increasingly offering more visibility into their operations.

Instead of individual cloud customers auditing the provider's cloud facilities, cloud providers and consumers use a form of delegated trust in which independent third parties certify widely recognized formal security standards. Following are some of the ways in which cloud providers provide transparency using the delegated trust model. No one method is the best, and you usually use a combination of methods.

Independent Auditor Reports

Cloud service providers engage independent auditors to assess the design and operation of their security controls. The providers then make the audit reports available to their cloud users. In the US, these independent reports include the financial industry's SOC 1 and SOC 2 reports.

Certifications and Attestations

Independent auditor reports, as useful as they are, aren't sufficient to ensure compliance. A good way to compare cloud service providers is to evaluate the range of industry certifications, such as the following:

- International Organization for Standardization (ISO): ISO 27001/27002 (general IT security)
- ISO 27018 (protection of PII information stored in the cloud)
- Cloud Security Alliance (Cloud Controls Matrix 3.0.1)
- US federal government's FedRAMP

- Healthcare sectors' HIPAA
- Financial industry's PCI DSS

The ISO 27001 and 27002 certifications, for example, provide assurance that the cloud provider has implemented a set of specific security controls and a system of management practices to ensure that the controls function as they should.

In addition to the US standards, there are numerous regional or national standards, such as Europe's ENISA Information Assurance Framework and Japan's Cloud Security Mark. All these standards require rigorous annual visits to the cloud providers' facilities by accredited auditors.

Nondisclosure Agreements

Cloud providers naturally zealously guard proprietary information about their physical architecture and their security and control systems. However, the provider must be able to share certain aspects of its architecture and internal security controls with its customers, subject to the customers signing a nondisclosure agreement.

Summary

This book explained the strategies for deploying Linux environments in the cloud, with a focus on Microsoft Azure. There are multiple strategies for an organization to move to the cloud. Regardless of the cloud vendor one chooses, the key to success in cloud environments is to follow a set of guiding principles for cloud operations. Understanding how virtualization, and more recently, containerization, and serverless computing play a crucial role is also important to doing well in the cloud.

Planning a cloud migration is vital, since a poorly planned and implemented cloud effort can set an organization back. Before you start a cloud migration, it's important to create a working cloud adoption road map. Conducting effective cloud readiness assessments sets the tone for the ensuing migration. Accurate workload, application, and database analysis reduces the surprise factor when you make the move to the cloud.

Cloud migrations consist of distinct operations, such as a set of predeployment tasks to get you ready for the migration, the migration

tasks, and go-live tasks. Although you can perform all the tasks in an ad hoc manner, using a cloud migration tool, such as Azure Migrate, reduces the time and effort required to move to the cloud, and it enhances the likelihood of a smooth and successful move. Azure Migrate is especially helpful during the discovery phase of a cloud migration by helping you assess your on-premise VMs for their suitability for a migration to the Azure cloud. You can use additional tools, such as Azure Site Recovery, and third-party tools, like CloudEndure, to facilitate your move.

Following a move to the cloud, you can use technologies, such as Azure VMSSs, to set up an immutable infrastructure CI/CD platform. Doing so enables you to automatically migrate application changes to the VMs that are supporting the applications.

The availability of your cloud-based applications and services can be affected by intermittent outages and by the possibility of a datacenter disaster. A cloud environment offers advantages in the availability area, since it's built to quickly provision virtually unlimited compute resources. Azure employs the concept of availability sets, which contain a fault domain and an update domain, to provide enhanced resiliency in the face of physical hardware failures.

Load balancing in the cloud helps you to scale your applications and to automatically detect and remove unhealthy instances. Azure's VMSSs enhance application availability and scalability. Azure Traffic Manager helps enhance the availability of applications by automatically directing traffic to alternative locations when some VMs fail.

Cloud vendors, such as AWS and Azure, enhance high availability by provisioning compute power across geographic regions, which are further divided into separate Availability Zones. The Availability Zones provide greater fault tolerance for mission-critical applications. Storage redundancy is another feature offered by cloud vendors to ensure the durability and availability of data. Caching and using CDNs are common strategies to enhance web application scalability.

Monitoring Linux servers in the cloud is inherently different from doing so in a local datacenter. In addition to monitoring the uptime and the performance of the servers, you must also pay attention to application performance monitoring. APM tools are of great help, since they help you monitor the end-user experience and they pro-

vide visibility into your application stack, helping you troubleshoot performance issues in the cloud.

If you're running a multicloud or hybrid cloud architecture, out-of-the-box monitoring tools, such as CloudWatch or Azure Monitor, may not be sufficient. There are powerful monitoring services and tools offered by independent SaaS providers, as well as third-party monitoring tools, like New Relic, PagerDuty, and SolarWinds. Identifying the right metrics to monitor user experience, gathering uniform metrics for on-premise and cloud-based services, and paying attention to cloud service usage and costs are the key guidelines when monitoring a cloud environment.

Cloud environments employ a shared security model, in which the cloud vendor is in charge of securing the cloud infrastructure, and you are responsible for securing your infrastructure and applications. Configuration management tools, access control mechanisms, and VPCs are some of the ways you can enhance your cloud environment.

RTO and RPO are what determine an acceptable system downtime. Cloud environments make it easier to set up a DR solution for your systems, since they offer elasticity and virtualization of resources. Instead of relying on outmoded conventional strategies, you can move all your backups and disaster recovery solutions to the cloud and run your business with zero downtime. A cloud-based solution, such as Azure Site Recovery, offers a good RTO and RPO. Instead of using a homegrown backup system, you can take advantage of a cloud-based backup service (SaaS), such as Azure Backup, serviced to safeguard data.

Governance and compliance are two areas where a cloud environment poses special problems, due to multiple compliance, legal, accessibility, and data disclosure agreements. You can enhance security and governance in the cloud by using strategies such as RBAC, hierarchical account provisioning, and network security groups. Continuous security assessments in the cloud, through tools such as Amazon Inspector, are key to mitigating vulnerabilities and reducing the incidence of malicious attacks on your cloud environment. Azure Security Center offers centralized security policy management, continuous security assessments, prioritized alerts, and policy-based enablement of control and governance.

About the Author

Sam R. Alapati is a Data Administrator at Solera Holdings in Westlake, Texas. He is part of the Big Data and Hadoop team. Sam is an Oracle ACE, a recognition conferred by Oracle Technology Network. He is the author of *Modern Linux Administration* (O'Reilly, 2018), as well as over 20 database and system administration books. Sam has experience working with all three major cloud providers: Amazon Web Services, Microsoft Azure, and Google Cloud Platform.