LEARNING MADE EASY

Nutanix Special Edition

Multi-Cloud Cost Optimization

for dummies®
A Wiley Brand

Understand cloud consumption

Answer your cloud management questions

Manage multi-cloud costs

Brought to you by

NUTANIX™

Scott D. Lowe

About Nutanix

Nutanix is a global leader in cloud software and hyperconverged infrastructure solutions, making infrastructure invisible so that IT can focus on the applications and services that power the business. Companies around the world use Nutanix Enterprise Cloud OS software to bring one-click application management and mobility across public, private, and distributed edge clouds so that they can run any application at any scale with a dramatically lower total cost of ownership. The result is that organizations can rapidly deliver a high-performance IT environment on demand, giving application owners a true cloud-like experience.

Multi-Cloud Cost Optimization

Nutanix Special Edition

by Scott D. Lowe

for dummies®
A Wiley Brand

Multi-Cloud Cost Optimization For Dummies®, Nutanix Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

ISBN 978-1-119-60026-8 (pbk); ISBN 978-1-119-60025-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

We're proud of this book and of the people who worked on it. For details on how to create a custom *For Dummies* book for your business or organization, contact info@dummies.biz or visit www.wiley.com/go/custompub. For details on licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

Some of the people who helped bring this book to market include the following:

Development Editor: Faithe Wempen
Project Editor: Martin V. Minner
Associate Publisher: Katie Mohr
Editorial Manager: Rev Mengle
Business Development Representative: Karen Hattan
Production Editor: Tamilmani Varadharaj

Table of Contents

Introduction

The public cloud came onto the scene in the early 2000s, offering businesses the option of moving away from on-premises data centers. Since that time, there has been a flurry of activity as new cloud providers popped up on the enterprise IT landscape and as industry analysts raced to coin new phrases to describe the emerging phenomena.

Public cloud was quickly followed by hybrid cloud, which has been followed by multi-cloud, but these aren't all. Companies such as Nutanix have coined their own cloud-centric terms, such as Enterprise Cloud, to describe a workload environment that provides the benefits of both public and private clouds.

In terms of the public cloud, hybrid cloud, and multi-cloud, each successive iteration results in new complexity that CIOs and IT staff need to grapple with in order to ensure that they're providing their companies with the right solution at the right time.

About This Book

In the pages herein, you learn about the rise of multi-cloud architectures and discover the critical challenges that you will face as you seek to adopt a wide array of services from different providers. By the end of this book, you will have insight into these challenges but, more importantly, you will be armed with knowledge of the tools you need to overcome these obstacles.

Foolish Assumptions

I assume that you know a bit about technology strategy, hyperconverged infrastructure, and the cloud. You don't need to be a pro, but you should be able to type your full name without having to look at the keyboard. This book is written primarily for IT executives and managers such as CIOs, CTOs, IT directors, and technical managers. If you aren't one of these people, that's okay. Read the book anyway because there's something for everyone!

Icons Used in This Book

Several helpful icons appear in the margins of this book. Here's a rundown of what these icons mean.

Anything that has a Remember icon is worth committing to memory.

The Tip icon points out helpful information.

The Warning icon alerts you to risks of various kinds.

Beyond the Book

There's only so much I can cover in a book of this size. If you're eager to learn more about managing your multi-cloud environment after reading this book, visit `https://www.nutanix.com/beam`.

Where to Go from Here

Like all *For Dummies* books, this book is designed to be read in any order you choose. Start with the chapter that interests you most, or read straight through. Hey, it's your book, so it's totally up to you.

IN THIS CHAPTER

- Comparing and contrasting local and cloud-based IT architectures
- Learning about public and private clouds
- Differentiating between hybrid and multi-cloud environments
- Discovering the key benefits and challenges of a hybrid or multi-cloud structure

Chapter 1
Introducing the Multi-Cloud

In the early days of IT, technology infrastructure and all its resources were kept snugly inside the four walls of data centers. It was safe there, and easy to manage, but that approach had some serious drawbacks too. Today, a new, decentralized operating paradigm is emerging that has the potential to reshape the way business is done: *cloud computing architecture*, often referred to as "the cloud."

However, reducing a discussion of enterprise architecture to something as general and generic as "the cloud" is overly simplistic and belies the complex technologies behind all that seemingly effortless processing and storage capability. A lot is going on in cloud architecture, particularly with the rise of multi-cloud.

In this chapter, you learn about various cloud-based IT operating methodologies at your disposal, as well as how the concept of multi-cloud fits into the current scheme of things.

Identifying Data Center and Workload Operating Models

Popular IT architectures have shifted through a series of operating models. A few decades ago, the mighty and bulky mainframe ruled the land with its highly centralized model. Not anymore! Like every other technology that undergoes continuous evolution, it ultimately gave way to the age of less monolithic systems like multi-tier client-server architectures.

In the following sections, I tell you about four workload operating models that you should know about in order to understand the best practices of today and the up-and-coming trends of tomorrow.

Three-tier infrastructure

For decades, three-tier infrastructure ruled the data center. The traditional three-tier environment, illustrated in Figure 1-1, consists of:

- **Compute:** You need servers on which to run your workloads.
- **Storage:** Those workloads need a place to store their processed data.
- **Networking:** You need some kind of communication between servers and storage to transmit and receive data.

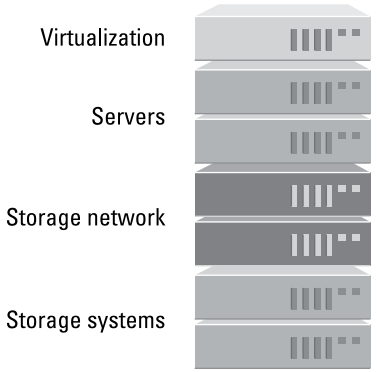

FIGURE 1-1: A traditional three-tier environment.

Three-tier environments typically require the enterprise IT teams to source and configure hardware from different vendors. These are stitched together to create an IT environment in which you can operate mission-critical business applications.

This piecemeal stitching together of components and expecting a unified whole can present challenges. Can you imagine buying spare parts from different shops to assemble a perfectly working car? That's what it's like when IT teams put together their own computing, networking, and storage solutions.

Each layer has a different vendor ecosystem, with Intel leading the computing layer, Cisco leading the network layer, and Dell/EMC leading the storage layer. The final integration is the responsibility of enterprise IT teams, who have to rely on highly specialized external network consultants to ensure the different layers work together as intended.

All these components must be integrated, tuned, maintained, upgraded, and managed separately. Each silo therefore requires separate attention and investment. Companies end up focusing more attention on the infrastructure than on the business outcomes that the infrastructure is supposed to be supporting.

To simplify the complexity of data center management inherent in the three-tier model, network virtualization solutions emerged. Virtualization tools enable administrators to configure and manage the different layers through a single management tool. Next-generation virtualization technologies such as *containerization* enable administrators to emulate multiple server instances on a single virtual network.

Although virtualization technologies simplify network management, the challenges of commissioning the network remain, because physical layers of compute, storage, and network still have to be set up individually. You may have anywhere from 1 to *n* management tools, where *n* is equal to the number of component types you've deployed. With such complex environments, you spend too much time turning tech knobs instead of business levers.

The inherent problems with three-tier environments have been partially responsible for the adoption of cloud solutions. But such issues have also given rise to innovative on-premises solutions such as hyperconverged infrastructure, which I tell you about next.

Hyperconverged infrastructure

Imagine buying components from different shops that all work together via a single centralized control system. That's the general idea behind hyperconverged infrastructure (HCI). HCI is an architectural style that uses a purely software-defined environment where all functional elements run on commercial, off-the-shelf (COTS) servers.

As shown in Figure 1-2, HCI eliminates the specialized hardware components that created data center complexity in traditional three-tier environments. HCI collapses the tiers into one with everything neatly encapsulated inside a set of applications. HCI also eschews separate silos and brings the hardware components under a single software umbrella.

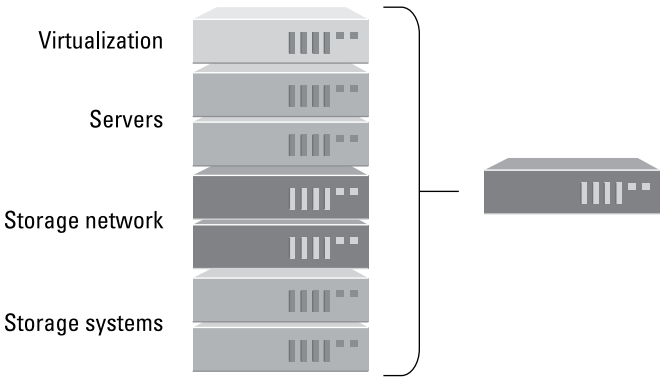

FIGURE 1-2: Hyperconverged infrastructure collapsing silos.

HCI is available as software that can run atop any supported hardware or as a preconfigured, turnkey appliance.

In a hyperconverged world, compute and storage are combined on a single server, with a hypervisor providing an abstraction layer for mission-critical workloads to operate.

On the networking front, you have no need for a storage network such as fiber channel or a separate management tool because the entire environment is managed as a part of the bundle.

REMEMBER

Hyperconvergence handily solves the complexity side of on-premises IT infrastructure. It enables you to focus squarely on the needs of the business rather than on the underlying technology.

Today, with companies like Nutanix providing powerful yet simple HCI solutions, hyperconvergence is an integral part of a broad and powerful platform that allows organizations to focus their attention more on business.

HCI has many benefits in and of itself, but it becomes even better when implemented as the on-premises part of mixed public and private cloud environments. I explain those cloud environments next.

Public cloud infrastructure

When your home or business needs water, you can drill a well, but it's often easier and less expensive to hook into a municipal water system and pay only for the water you use. That is a great metaphor for cloud computing versus owning your own infrastructure.

REMEMBER

A *public cloud* is an infrastructure for rental to the public, as the name implies. Large services like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud have invested billions in creating public cloud infrastructures in which other companies can rent compute, network, and storage capability. Customers and employees can run applications and store data in a secure, always-available environment, at a cost that may be lower than it would be to "drill your own well."

Applications running inside software-defined infrastructure environments do not care whether the physical server infrastructure is within the enterprise data centers or on a rented public cloud infrastructure. They provide the same experience for developers or end-users either way.

The cloud computing model enables organizations to quickly scale up new services without having to worry about internal computing resource constraints. The capabilities are unshackled from the internal infrastructure.

For workloads where it makes sense, IT can simply deploy a service in the public cloud and get it up and running right away. There are integration details to work out, of course, but once those are taken care of it's easy to get new workloads running.

Cloud architectures have also contributed to the ability of organizations to rethink how IT operates at a fundamental level. Gone are the days where IT had to either come up with a few hundred

thousand dollars for hardware or sign a huge lease every few years. Now, alongside the critical applications that need to live on-premises, IT can operate in a pay-as-you-go model with the use of a public cloud. This relative ease of scaling up and down IT resource consumption based on workload requirements is an attractive proposition.

REMEMBER

This trend from on-site to cloud-based infrastructure is often referred to as shifting from a *CapEx* focus to an *OpEx* one. The business restructures the IT budget so that there are fewer expenses in the capital column and more expenses in the operational column. CFOs and other financial types tend to love this approach because it makes IT costs more predictable and avoids overspending on infrastructure that isn't needed at the moment.

There's a downside, though. Without the right tools, public cloud usage can become frighteningly expensive. I go into the costs in more detail later in this book, and explain how to prevent overspending.

REMEMBER

Cloud adoption offers deployment flexibility with a level of scalability that can't always be matched by on-premises environments. It also has a cost structure that better aligns with *right now* needs and helps avoid pre-buying hardware that won't be needed immediately.

So if public clouds are so great, wouldn't more public cloud usage always be *more* great? Well, not necessarily. In some situations, public cloud usage is not your best choice. I explain what I mean by that in the next section.

Private cloud infrastructure

When enterprises make the decision to move application workloads to the cloud environment, they sometimes run into security complications. Certain applications require a high level of security as well as special access restrictions, particularly when the data being stored and processed is subject to government regulations. Although a public cloud environment has fairly good security, some applications need a higher level of privacy than any public cloud can provide.

To address these concerns, cloud providers came up with the concept of *private cloud*, cloud computing accessible only within a virtual private network (VPN) that still provides the ease of use and flexibility of the public cloud.

Many enterprises want to retain absolute control of some of their applications and want to run some of these applications on their on-premises platforms. These organizations have adopted a single public cloud to operate alongside their on-premises or private cloud environment, creating the *hybrid cloud* scenarios like the one shown in Figure 1-3.

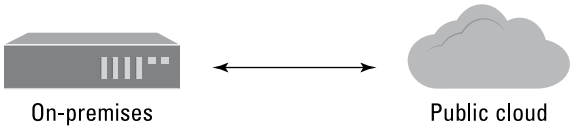

FIGURE 1-3: A hybrid cloud.

Multi-cloud infrastructure

Some enterprises have complex IT infrastructure needs that involve stitching together multiple cloud platforms — in other words, creating a *multi-cloud*. A multi-cloud infrastructure aggregates any combination of two or more public or private cloud environments. It can be as simple as one private cloud environment plus one public cloud platform, or it can combine multiple public cloud platforms, as shown in Figure 1-4.

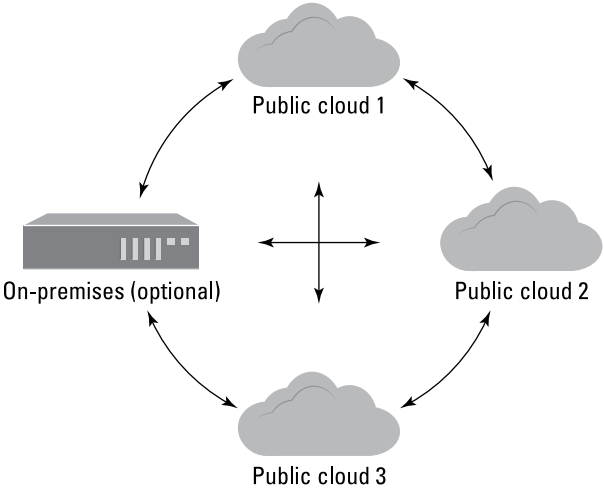

FIGURE 1-4: A sample multi-cloud architecture.

Both hybrid and multi-cloud are comprised of multiple environments. These terms are easily confused, but the key difference is that hybrid clouds are *always* comprised of public plus private clouds. By definition they are a hybrid of two types. Multi-cloud environments, on the other hand, are simply environments with two or more clouds. Those clouds can be of any type (including two or more of the same type).

Another difference is the level of integration between the clouds. In a hybrid environment, some level of integration exists between the public and private components. This may not be the case in the multi-cloud world. With multi-cloud, cloud providers are chosen for specific tasks, so there may not be an intersection between them unless business cases or workloads call for it.

Discovering the Benefits of Multi-Cloud

As cloud computing architecture has become a proven model, many organizations have jumped onto the cloud bandwagon. Several "born in the cloud" companies now exist.

However, there are also numerous anecdotes from those that have made a significant jump into public cloud only to find that they went too far and too fast. To correct the problem, they've pulled some workloads back into on-premises or private cloud environments. At times, they've sought ways to simplify IT operations by using the HCI model for their on-premises workload needs. This section provides a realistic look at the pros and cons.

REMEMBER

Hybrid and multi-cloud-driven organizations gain significant efficiency and capabilities. Here are some of the key benefits:

- Improved agility and ability to deploy workloads practically in real time
- The capability to burst to one or more public clouds as capacity needs dictate
- Ease of scale whether on-premises (with hyperconverged infrastructure) or in public clouds
- The ability to choose the best location for a new workload — on-premises or public clouds

- Choice of cost model per application — CapEx for on-premises or OpEx for public clouds
- Improved reliability through the enablement of geographically dispersed services

You attain a great deal of workload flexibility in a multi-cloud world. This isn't to say that you'll bounce workloads around clouds like pinballs, but you do have more opportunity to align workload requirements with the right cloud.

You also avoid being locked into one cloud platform. No matter how many clouds you have, it still isn't easy or inexpensive to migrate away from a cloud provider, should you become dissatisfied with them. However, the first step in any plan to move away from a cloud provider is having an alternative place to go, and a multi-cloud strategy provides that flexibility.

Some other benefits of multi-cloud include:

- **Best of breed adoption:** You get to choose the best breed of provider for each workload need.
- **Cost:** You can choose a usage target based on the cost you will incur for that workload, allowing you to further optimize your cloud spending.
- **Increased reliability:** Public cloud providers already provide zones to help you distribute workloads, but they're still in the same provider environment. A multi-cloud strategy that uses different cloud providers to increase redundancy can help provide a reliability boost for mission-critical applications.

Understanding the Drawbacks of Multi-Cloud

There's always a downside, and multi-cloud is no exception, although many of the downsides can be mitigated with the right tools.

First, no matter how you look at it, more clouds equal more complexity.

Second, it can already be hard to find IT pros with cloud skills for *one* cloud provider, let alone two or more. A multi-cloud strategy may strain the available talent pool and make human capital costlier.

Third, every provider works differently, and this includes the security layer in each one. For every cloud you adopt, you need to become keenly aware of how its security apparatus works and make sure that each one stays in alignment with your company's policies.

These aren't insurmountable problems, but you need to plan adequately to address them.

IN THIS CHAPTER

- Understanding the differences between HCI and cloud consumption models
- Identifying the right cloud model for different application implementations
- Mapping a resource plan to a vendor offering

Chapter 2
Understanding Cloud Consumption

The ability to dynamically scale up and scale down resources, and thereby costs, is the defining feature of cloud models in general.

Each cloud model has its advantages and disadvantages, and these have a direct impact on costs. Considerations such as data security, application performance, and regulatory compliance are critical factors as well. It is up to you to choose the right model after considering various factors. In this chapter, I dive deep into various cloud consumption models to help you identify which model is best suited for your needs.

Comparing On-Premises and Cloud Consumption Models

This section explores how the availability of public and private clouds is challenging traditional IT consumption models.

Traditionally, companies have budgeted for a three-to-five-year leasing cycle for IT hardware. In other words, every three to five

years, the old IT hardware is replaced by newer and more capable equipment. That's all well and good if things go as planned. But as you're likely well aware, complications are nearly inevitable, and nearly always result in extra expense and hassle.

One such complication can occur when vendors tweak hardware specifications so that the new equipment rolled out at the specified interval is not compatible with existing resources. Another is that sometimes new or additional hardware is needed when it isn't time for it, and when there's no money left in the budget.

The hyperconverged infrastructure (HCI) model addresses the compatibility issue by ensuring that everything works well together. However, it is a capital-intensive approach. There's a big outlay of cash for leasing new hardware at regular intervals. Two common types of IT hardware leasing are:

- **Capital lease:** This is a way of financing a purchase. The buyer makes payments over time.
- **Operating lease:** This is more of a "true" lease, in that you don't get to keep the equipment. When the lease ends, the lessor takes back the equipment and often resells it.

Either way, *someone* is buying the hardware upfront. If it isn't you, then it's the company you are leasing it from.

A public cloud infrastructure offers a different kind of financing by enabling a pay-as-you-go (PAYGO) model. Typically, no huge upfront investment is necessary to procure public cloud resources.

The PAYGO model means that businesses buy only the resources that they need urgently. If an application requires more computing power, IT can easily add resources in the cloud management console.

Although scaling up through progressive addition of resources on-demand is not a major problem with HCI model, scaling *down* the resources can be a challenge because these purchased resources cannot be returned.

However, scaling up or down is relatively easy in the public cloud model. When unused resources are removed from the cloud management console, the billing for those resources stops. This flexible scaling capability has been the magic bullet for IT teams adopting public cloud.

Understanding How Applications Affect Cloud Consumption Models

The PAYGO model provides a definite advantage over an upfront payment model when the desired resource capacity is expected to change significantly over time. However, applications and workloads ultimately determine which cloud model is the best for your situation. In addition to end-user applications, the middleware software components like databases, messaging platforms, and even the virtualization platforms play a critical role in deciding the cloud model.

Comparing SaaS, IaaS, and PaaS

Cloud models can be classified according to the level at which the customer (your company) interacts with the cloud. The three key classifications are software as a service (SaaS), infrastructure as a service (IaaS), and platform as a service (PaaS).

In the software as a service (SaaS) model, the entire application code and all the data reside in the cloud. Because of its web delivery model, SaaS eliminates the need to download and install applications on each computer.

SaaS applications may not always have good IT governance models. Some business teams overlook data backup and security concerns for the sake of convenience. This has led to unsanctioned SaaS applications, called *shadow IT*, that can potentially compromise all of your data.

For example, someone in your organization may be using a cloud storage SaaS application to store your company's data instead of following the security protocols established by your security administrators, which require keeping all company data on your own private storage servers. Putting the data in a cloud-hosted SaaS application may make data retrieval easier wherever you go (think iCloud), but this method places your data at the mercy of the SaaS application's security governance model.

Due to their business criticality, some applications must be developed and deployed in the cloud. Infrastructure as a service (IaaS) provides a "virtual data center in the cloud" for enterprises. In the IaaS model, cloud servers are typically provided to

the organization through a dashboard or an application programming interface (API). IaaS users have complete control over the entire infrastructure and pay strictly for the use of the hardware or hardware related services that have been accessed. IaaS is the model of choice for organizations that want to create their own customized applications.

IT teams usually focus only on their enterprise applications and do not want to manage middleware tools like database systems, messaging platforms, and so on. To address this, cloud vendors provide platform as a service (PaaS), with preloaded operating systems and middleware. With PaaS, you don't need to hire specialized consultants to manage middleware tools and can leverage the PAYGO model offered by PaaS providers.


REMEMBER

Both PaaS and IaaS offerings are available in the public and private cloud models. The major difference between IaaS and PaaS is that IaaS provides the ability to customize even the virtualization layer, while most PaaS offerings provide limited choice in this regard.

Considering application demands

From a business perspective, although the workloads of most enterprise apps used by internal teams are stable and predictable, most of the customer-facing applications have fluctuating demands. Similarly, data analytics, machine learning, and artificial intelligence (AI) applications also have dynamic workloads.

Imagine running tax software that sees heavy demand when tax filing dates are near. Such "burstable" applications require dynamic scaling. As a result, hybrid and multi-cloud models are becoming the norm for cloud consumption models in enterprises across industries.

Choosing where to run your applications

One of the most important decisions you can make regarding cloud cost optimization is where to run your enterprise applications. Private cloud may seem costlier than public cloud at first, but offers more privacy and control. A hybrid model that includes public, private, and on-premises HCI infrastructures can be an

attractive compromise, but requires some thought and planning to determine how to best allot the workload to each.

TIP

When deciding where to host an application, answer the following questions:

- Does the application have latency restrictions between the workload and clients?
- Does it have relatively steady resource consumption requirements?
- Does it need to connect to many of your on-premises applications in the data center?

If the answer to any of these questions is "Yes," you should consider running the workload in your private cloud based on HCI infrastructure. For latency-sensitive workloads, it's better to keep those near the clients who will be consuming them. If those clients happen to be your employees, keeping the workload in-house makes the most sense. For applications that need to connect to on-premises systems, keeping it local may also improve security.

You can easily determine the cost to keep steady and predictable applications in-house. This information can be particularly helpful if the application runs continuously for an extended period. It's quite likely that, over time, the cost to move that application to the cloud will exceed the cost of keeping it local.

On the other hand, if you have "burstable" applications that have varying and unpredictable resource needs, aren't particularly sensitive to latency issues, and are self-contained, the public cloud might be a better option. In this situation, you can allow the public cloud to do what it's best at — providing on-demand hyperscalability of resources.

Right-sizing resources

Once the cloud model has been identified, the next step is to identify the resources required for your workloads. The resource plan should be dynamic in the cloud, specifying when the resources should be scaled up or down. This "sizing guide" is the key for cloud administrators to keep infrastructure availability at optimal levels at all times.

Mapping Your Resource Plan to a Vendor Offering

Each public cloud vendor has different product offerings and pricing tiers. Your challenge is to figure out which vendor's offerings match what you have determined that you need.

REMEMBER

You need to know which type of instance is ideal for your use case in order to choose the right option. It would be overwhelming for even an experienced architect to select the right component at all tiers.

Cloud service providers enable customers to choose the resources at all three tiers of the IT infrastructure — compute, network, and storage. There are multiple options for each of these three tiers in every major cloud platform. For example, in Amazon Web Services (AWS), the compute tier is called EC2. An Amazon EC2 instance can be configured with various combinations of memory, networking capacity, CPU, and storage.

WARNING

In a multi-cloud setup, the configuration is more complicated because major vendors like AWS, Microsoft Azure, and Google Cloud have separate terminologies for the resources. You will have to identify the equivalent terms for each of the resources to compare how one vendor's offering fares against the others.

In addition, resource sizing might be different for the same level of expected performance, or the performance might be different for the same resource sizing in different clouds.

For example, a compute instance with 8 vCPU and 16GB memory in AWS may not deliver the same level of performance as the same sized instance in Google Cloud.

The best way to avoid overpaying for underutilized resources is to continuously monitor resource utilization. You should identify over-allocation or under-allocation and size the resources accordingly.

Monitoring becomes more complicated in a multi-cloud environment because each cloud vendor provides administration panels only for deployed resources within *its* platform. Therefore, in a multi-cloud environment, you need a tool that can provide you with a unified view of all the resources deployed across various cloud platforms.

Cloud management tools such as Xi Beam by Nutanix provide this capability. Tools such as Beam help administrators to continuously monitor workloads across the multi-cloud and provide deep insights into cloud costs. These tools also provide recommendations on optimizing your cloud deployments to reduce costs.

IN THIS CHAPTER

- Understanding cloud cost management
- Learning how cloud management platforms can help manage costs

Chapter 3
Managing Costs in a Multi-Cloud Environment

Jumping into the cloud affects your IT economic model, but jumping into more than one cloud makes the "cloud cost" question much more difficult to answer.

This chapter begins with a look at the economic models that may be in play in your IT organization. It also looks at how cloud management platforms (CMPs) have emerged to ease operations, security, and cost management needs.

Understanding Cloud Cost Management

Financial comparisons between traditional data centers and cloud infrastructure are not simple to make. It is more than just comparing the compute, storage, and network costs between them. Many hidden overheads come into play and can become the deciding factor.

Many best practices are available to bring down your costs. Here are several to consider.

Leveraging vendor discounts

To manage costs effectively, you should have a buying strategy to get the best out of the pricing schemes that vendors offer.

Most cloud providers offer significant discounts when you pay the charges for computing resources upfront. For example, Amazon Web Services (AWS) provides "reserved instances," which are typically compute services that are available at significant discounts if you commit to them for a duration of one year or longer. If you commit to resources long-term, you lose the benefits of pay-as-you-go (PAYGO). Therefore, you should take a balanced approach between resources bought upfront and resources bought on-the-fly.

Most cloud providers also offer discounts if users buy and reserve server instances instead of consuming them in an on-demand basis. These reserve instances are usually available at discounts of 50–70 percent over on-demand pricing and are best used for production workloads or when you know that you will need that instance for an extended period.

Cloud vendors also offer their spare compute, storage, and network resources at a significantly discounted price-scheme called *spot instances*. For example, AWS sometimes offers its unused compute resources at discounts as high as 60–80 percent compared with on-demand pricing. These spot instances are perfect for dev/test workloads where you do not necessarily care about high availability and can work with whatever compute instance the cloud provider gives you.

You should have a buying strategy to get the best out of these pricing models. Why pay for a more expensive on-demand model? You may be able to either reserve the instances for an extended period for production workloads or use the available spot instances for dev/test workloads and save a huge amount on your public cloud costs.

Choosing storage types wisely

All cloud providers offer different types of storage, like solid-state drives (SSDs), hard disk drives (HDDs), tape drives, and so on.

These storage types come at different price points, with devices that have slower access times available at a discount. You might, for example, choose slower devices for secondary backup storage than for production data.

Continuous monitoring and optimization

Optimal cloud management requires continuously monitoring resource utilization such as CPU, memory, disk usage, and so on. In order to avoid waste, cloud administrators should implement policies to resize resources that are too bulky for what they are being used for. For example, if you have a compute instance that is continuously using less than 2 percent CPU, you may be able to downsize it and save on costs without compromising performance.

Leveraging auto-scaling tools

Most cloud providers offer auto-scaling tools that help you define usage limits and policies for various resources. These tools can automatically enforce your usage policy so that costs always remain within predefined limits.

Automating scheduling

One of the biggest drivers of public cloud costs are zombie resources left running in the cloud but that you may not be using anymore. For example, you may have compute and storage instances up and running 24/7 in your dev/test environment, but you may not need them overnight or on weekends. Why not use an automated scheduling policy that shuts them down when they are not needed and brings them up when you are ready to use them again? Turning off resources overnight and on the weekends can significantly add to your cloud cost reduction.

TIP

Many of these best practices can be automated, and a good cloud management platform (CMP) can help with that, as well as with cost management in general, covered next.

Understanding How CMPs Save You Money

CMPs give you granular resource usage visibility so you can better govern your cloud-based operations. CMPs help automate most of the best practices of cloud management, ensuring that you can maintain continuous control over your multi-cloud environment.

CMPs that control and orchestrate all assets such as the underlying infrastructure, applications, and services, from a single user interface, can help you with effective workload migration between public and private clouds.

Most hybrid CMPs provide the following features:

- A unified view of public and private cloud consumption
- Access control across clouds, data centers, and tenants
- Automated resource deployment and orchestration in a self-service manner
- Monitoring of infrastructure and application health along with alerting when issues arise
- Customizable policies to automate governance of costs, operations, security, and compliance

Typically, lack of visibility over cloud consumption results in rapidly escalating costs because of over-provisioning or underutilization of resources.

CMPs optimize cloud costs by eliminating unused resources, rightsizing underutilized resources, purchasing the appropriate reserved instances, and automatically migrating to spot instances.

If you are choosing a CMP for cost optimization, look for the following features:

- A single dashboard for the multi-cloud, including public and private clouds
- Automated actions to immediately realize cost savings when an optimization opportunity arises

- Proactive budget alerts that alert cloud administrators before costs exceed a certain threshold
- Resource purchase recommendations based on historical usage patterns
- Automated migration to more cost-effective instances without downtime

In Chapter 4, I take a deep dive into the actions you can take to optimize your cloud costs.

IN THIS CHAPTER

- Discovering the essential features of a cloud cost optimization solution
- Establishing a cloud financial governance framework
- Centralizing cloud financial governance

Chapter 4
Selecting the Right Cost Optimization Solution

Enterprises getting started with the cloud typically focus on managing features that provide operational efficiency. However, if you do not implement strong financial governance upfront before starting on your cloud journey, it may result in quickly spiraling costs.

When enterprises start moving most of their IT infrastructure to the cloud, they realize that poor cloud financial management causes them to overshoot their IT budgets. As the volume of workloads grows, simply *managing* cloud costs is not enough. They have to be *optimized* — the core of what this chapter covers.

Essential Features of a Cost Optimization Solution

A great cloud cost optimization (CCO) solution is one that saves you a lot of money, right? Simple enough. But behind "saving you money" is a series of very specific, practical strategies.

Real-time visibility and analytics

Real-time visibility and analytics are critical in helping you see the areas where you may have too much or too little in the way of resources. Either condition can be costly, especially when system performance suffers and customers are inconvenienced.

CCO tools use advanced machine learning to perform a time-series analysis of the resource consumption log data to identify anomalous trends and usage patterns. The best CCO tools can identify when cloud costs start to spike by looking for unexpected patterns in the billing data.

For, example, you should look for which users are consuming which services in which region at what time of day. Anomalous deviations should be immediately flagged when the spending goes above a certain deviation from normal patterns. Machine learning-based automated cost anomaly detection can act as an early warning signal.

Automated budget tracking and chargebacks

Another feature you want in your CCO tools is the ability to automate budget tracking and chargebacks. CCO tools can help to optimize spending through budget tracking and chargeback by answering these questions:

- How much are you spending across your organization?
- Which team is spending more?
- Who — or what project — is overshooting the allocated budget?

CCOs provide budgeting tools through which enterprises can apportion the cloud costs to different business units or cost centers. This could be for a team, a project, a region, or a department.

CCO tools also provide automatic real-time budget tracking so that cost-saving actions can be taken before it is too late.

Tagging for effective cost optimization

CCOs have to rely heavily on the tags associated with cloud resources in order to track spending patterns against allocated budgets.

A *tag* is a label that you assign to a cloud resource. By applying tags, you add metadata to logically organize resource groups using a taxonomy.

Each tag consists of a key and an optional value, both of which the user defines. AWS and Azure call them *tags*, and Google Cloud calls them *labels.*

CCOs empower financial leaders in an organization to create cost centers and associate tags with them. CCOs also help you associate budgets with the cost centers and then automatically aggregate all resource costs associated with tags for that particular cost center. This helps you track your spending against allocated budgets using flexible definitions that can be changed by altering the tags associated with a cost center.

TIP

All your cloud resources must be tagged in order to ensure that a CCO is tracking your cloud spending accurately. Also, ensure that you add the appropriate tag-key in the definition of the cost center.

You may have several cloud resources that are untagged or untaggable. In this case, it is important to manually charge back those resources to the appropriate cost center to ensure that the spending amount associated with each cost center is accurate.

REMEMBER

You may not be able to tag everything, but tag everything that you can.

Data and application optimization

Data and application optimization actions offer more granularity than resource optimization. They can add significant cost savings to your overall cloud spending. For example, different storage classes are priced differently, so storing objects in the appropriate class can help reduce costs.

CCOs provides tools that can make use of different storage classes for different stages of the data during the application's lifecycle.

Some CCOs allow you to define rules that can move objects to different storage classes or permanently delete them. This capability can greatly reduce your storage bucket costs.

Establishing a Cloud Governance Framework

As organizations measure their product teams on speed and time-to-market, the need to adapt from traditional compliance checks and security processes increases significantly. Would you be able to innovate at the pace desired if every resource deployment had to be audited by your security team? Conversely, if you create storage buckets and deploy databases in the cloud without anyone to look over your shoulder, would you be confident that you are not making a mistake that could lead to a security breach?

Businesses need to continuously evolve without compromising basic security needs. Cloud administrators need a way to perform automated integrity checks without hindering the speed of application development. This dynamic of business applies to the underlying IT systems as well.

Without governance controls in place, you will quickly find your cloud environment spiraling out of control. Of course, you can manually implement such governance controls, but you will likely end up spending hours on a process that can be automated. Ensuring that cloud management best practices are being adhered to requires a multi-cloud governance strategy that relies heavily on real-time detection and automated remediation of security risks.

REMEMBER

Cloud governance is the ability to align IT operations to business strategy, track performance, and make adjustments to ensure that organizational objectives are met through optimal consumption. All this should happen without breaching the parameters of risk tolerance or compliance obligations.

Cloud governance ensures that cloud applications and services meet business expectations. Cloud governance can be automated through the use of cloud management platforms (CMPs) that can define and implement various aspects of cloud infrastructure, such as specifications for design, architecture, security, and privacy policies.

They should also help with the implementation aspects of cloud management such as monitoring, alerting, and remediation of incidents.

CMPs typically achieve the specification aspect of cloud governance by integrating policies, profiles, access control lists (ACLs), encryption, and other security features.

Once a set of policies is specified, a cloud governance solution must be able to continuously monitor and detect if the cloud infrastructure complies with the agreed policy. For example, if a cost governance policy mandates that temporary data backups should be deleted every month, the CMP would enable the administrator to specify that old snapshots should be deleted after 30 days. Similarly, governance policies could mandate that any unattached volumes should be spun down or compute instances be downsized if they use less than 5 percent CPU over a certain period.

Cloud governance tools should have three key components:

- Real-time inventory monitoring
- Validation against governance policies
- Task automation to establish a baseline

An effective cloud governance tool allows policies to be set at the granular level as well as the macro business objectives level. This approach helps to align IT operations with business goals.

Centralizing Cloud Financial Governance

Financial governance is a key business governance solution and involves a lot more than simply keeping costs under control.

The need to centralize control across multiple teams becomes critical as enterprises foray into multi-cloud and multi-application environments. Cloud vendors and business owners need a systematic way to map consumption to business units in complex environments.

CCO tools with governance capabilities delivered as a part of a CMP can help you visualize resources by groups and departments, empowering you to manage their usage and enforce policies based on allocated budgets.

This financial governance capability creates better accountability by identifying the teams, users, or projects that are driving the cloud spending. This approach enables business owners to make data-driven spending decisions that ultimately help you ensure that you get the maximum value out of your cloud investment.

IN THIS CHAPTER

- Understanding the importance of abstraction
- Learning about three key pillars of cloud governance
- Discovering how Beam automates cost optimization

Chapter 5
Maximizing Return on Your Investment

To be confident that you are selecting the right cloud cost optimization (CCO) tool, you need to understand how certain key cloud optimization concepts are implemented. (To learn how important it is to select the right CCO tools for a multi-cloud infrastructure, see Chapter 4.)

This chapter provides a walk-through of Xi Beam by Nutanix, a rich tool that provides visibility, optimization, and automated control over your multi-cloud deployments including both public and private clouds. By seeing what Beam can do, you will better understand what a CCO can do for you and what features to look for.

Understanding the Importance of Abstraction

The first benefit that Beam brings to the table is a single unified view of underlying public and private cloud resources. Its dashboard rolls up all the information on configuration files, access

logs, resource utilization and cost information, and so on into one easy-to-use interface, shown in Figure 5-1.

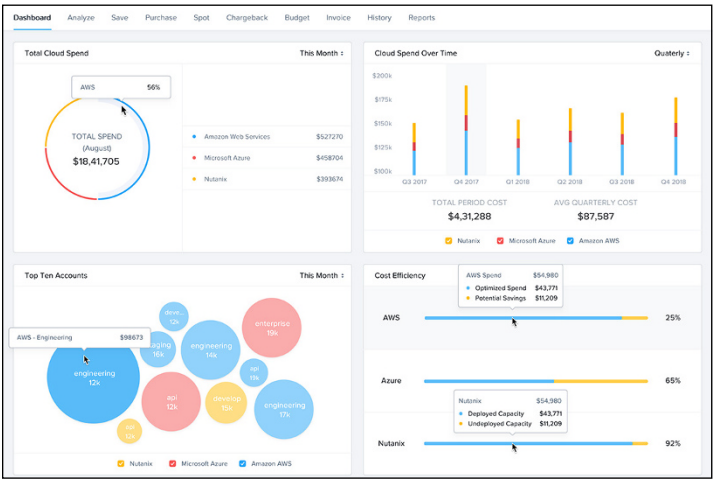

FIGURE 5-1: Multi-cloud cost governance overview in Beam.

Beam allows you to access your hyperconverged infrastructure-based private cloud consumption details, along with the public cloud ones, without having to shuffle between multiple consoles.

Three Key Pillars of Cloud Governance

As a multi-cloud governance tool, Beam organizes all the infrastructure information along three key pillars of cloud governance:

- **Visibility:** Beam provides deep visibility into multi-cloud consumption patterns at aggregate as well as granular levels to easily identify cost drivers.
- **Optimization:** Beam uses machine intelligence to continuously provide right-sizing recommendations and deliver more than 50 percent in cloud cost savings.
- **Control:** Beam helps you create automated policies that continuously take actions to reduce cloud costs. Beam also provides budget alerts so that you can take remediation actions before costs get out of control.

The next several sections explain how Beam uses these three pillars to provide you with the tools necessary to manage and optimize your cloud infrastructure spending.

Forecasting future cloud bills

Beam utilizes a machine learning-driven algorithm to project future cloud spend using historical consumption trends. Beam can project up to the next six months of cloud spending, as shown in Figure 5-2, thus allowing financial administrators to more accurately plan their budgets.

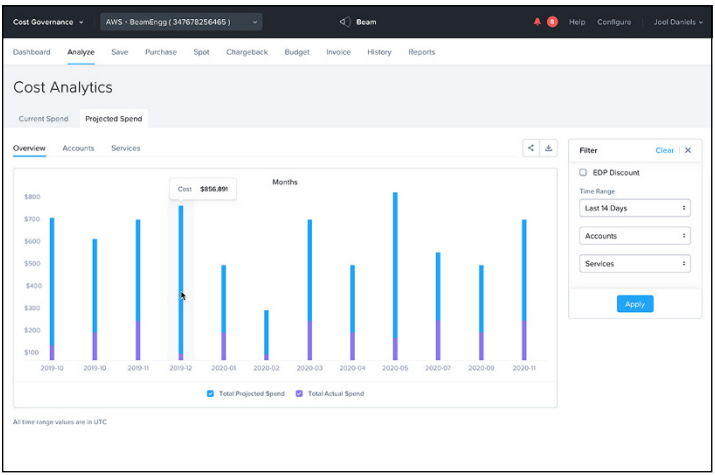

FIGURE 5-2: Beam's spending projection feature.

Automated cost anomaly detection

Beam also uses the machine learning-driven algorithms to detect spending anomalies that may be caused by out-of-pattern resource consumption patterns.

Beam's cost anomaly detection feature, shown in Figure 5-3, can help you identify spending spikes before they get out of hand. Cloud operators don't have to keep looking at the billing reports; Beam does that for you and alerts you when costs spike.

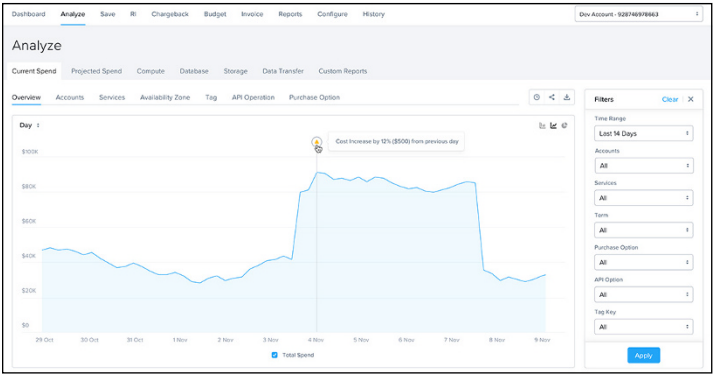

FIGURE 5-3: Automated cost anomaly detection in Beam.

TCO visibility

The total cost of ownership (TCO) is a configurable financial model that allows you to analyze the direct and indirect costs of owning, operating, and maintaining your cloud infrastructure.

A TCO-based costing view helps you to better understand how much you are paying to run your workloads in the cloud.

Beam's TCO-based costing analysis for the Nutanix private cloud includes the following cost heads:

- **Hardware:** Cost of OEM or third-party hardware
- **Software:** Cost of Nutanix or third-party software
- **Facilities:** Infrastructure costs incurred on your privately owned data center or a colocation
- **Telecom:** Ethernet/top-of-rack switch costs
- **Services:** Third-party services costs
- **People:** Salaries of IT administrative staff

For hybrid and multi-clouds, Beam provides such a TCO view for private as well as public cloud infrastructure, as shown in Figure 5-4. This TCO model is configurable at the individual cluster level, allowing you to set different costs for clusters running in different geographic locations or on different hardware and software configurations.

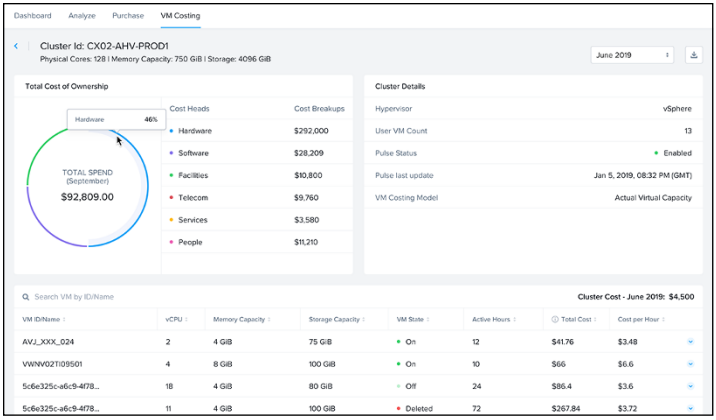

FIGURE 5-4: TCO model in Beam.

Beam also tracks the VMs running in each cluster and shows you the cost of individual VMs, which are also derived using the TCO model.

Optimizing cloud spending

To optimize multi-cloud governance, companies must adopt automation wherever possible. Here's a look at Beam's features for automating cost optimization.

Continuous rightsizing of resources

Over-allocation of resources is one of the most recurring reasons for spending waste. Continuous analysis of key metrics (CPU, memory, storage, and so on) — along with automated rightsizing to optimally size resources that may have lesser capacity — can drive cloud spending lower.

As shown in Figure 5-5, Beam proactively identifies idle and underutilized resources along with building automated policies to clean up idle resources or downsize those that have excess capacity.

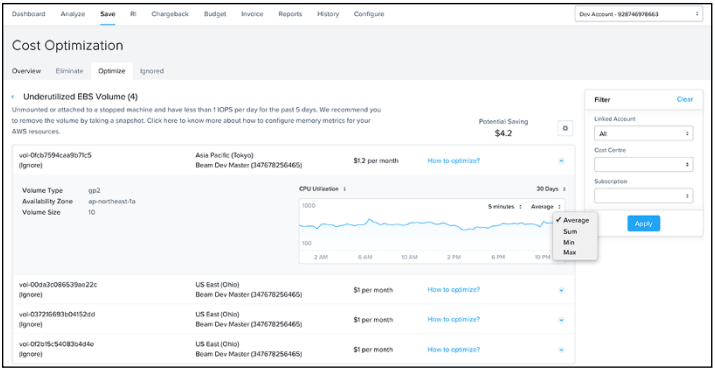

FIGURE 5-5: Rightsizing resources with Beam.

Intelligent consumption planning

Cloud providers offer multiple purchasing options that can yield significant savings when utilized effectively. You can purchase instances in an on-demand model, reserve them for a period of time, or utilize spare instances that are available on the spot.

Beam makes this planning process easy by analyzing your workload patterns and the available instances to make the most cost-effective purchase recommendations for your environment.

Additionally, cloud administrators should plan to automatically migrate their on-demand instances, as shown in Figure 5-6, to spot instances or vice versa with no downtime. Cloud administrators can realize 50 percent in cloud cost savings through all these cost saving actions.

Control over cloud spending

Automation is the key to identifying deep savings in your cloud environment. There are likely many tasks you perform on a daily basis that can be easily automated. For example — even such a mundane task as turning your virtual machines (VMs) on or off likely happens at least twice daily, if not more. It is human nature to sometimes miss out on performing these manual tasks. This is where automated governance comes into play.

Think about your dev or test environments. Do you use your VMs or databases in those accounts at all times? Likely not, unless your engineering team is running some QA or testing at 3 a.m. on a Saturday!

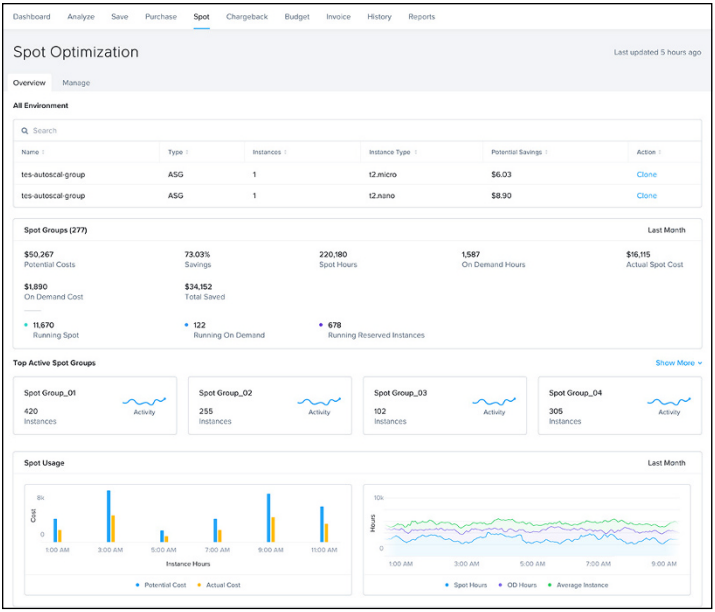

FIGURE 5-6: Planning for automated spot instance migration.

Beam makes it easy to create scheduling policies, as shown in Figure 5-7, that can automate turning various cloud resources on or off at predetermined times. These can be of your choosing, but typically they are overnight and on the weekends. Configure the policy once, and Beam does the automation. This may sound like a trivial thing to do, but it can easily add 25 percent or more to the cost savings for those VMs and databases.

Beam tracks cost consumption across all cloud resources at both aggregate and granular levels: per application workload, team, and business unit.

Beam allows you to define business units to map your organizational structure with your cloud resources, spanning boundaries across multiple cloud providers — public or private.

After defining your business units, you can create cost centers within each business unit. You can then allocate a budget at the entire business unit or the individual cost center level. For example, a business unit might be your engineering department and some cost centers within it might be for your QA, dev, or production environments.

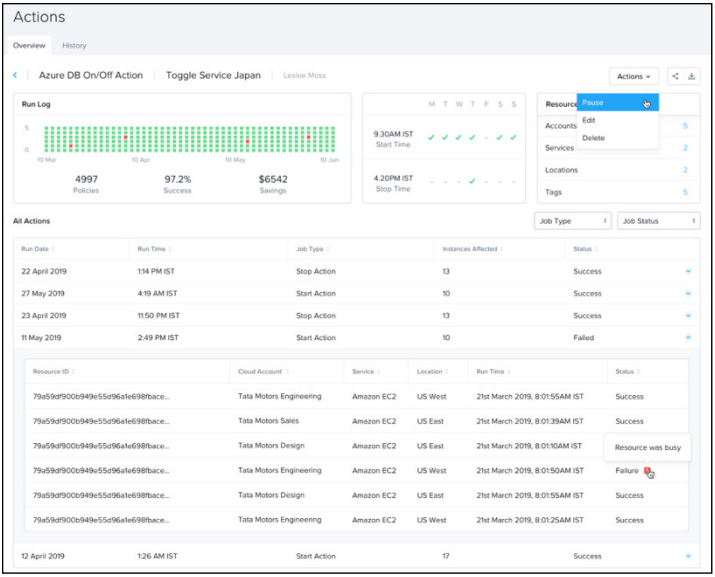

FIGURE 5-7: Scheduling policies in Beam.

Budgets are effective only when they are tracked and enforced automatically. Beam leverages the tagging feature provided by all cloud providers to implement budget tracking and alerting before they are exceeded.

You can set up budget alerts to notify the appropriate users who will need to implement cost-saving actions before budgets are exceeded.

The cost reports provided by Beam are extremely detailed and can also be used for chargeback purpose in your organization. Enabling automated chargeback and budget tracking helps to centralize and effectively implement your financial governance policies in the multi-cloud.

Unlocking digital transformation without huge cloud bills

To remain competitive in today's world, many companies are undergoing a customer-experience focused *digital transformation* that positions them to rapidly adapt to the digital world.

IT plays a leadership role in this digital transformation and is therefore a mission-critical element in propelling the business forward. IT policies and practices must align with the business's current and future needs and goals. But it's far easier said than done. Because of this, cloud governance has become a key concern for IT teams in order to meet those digital transformation challenges.

Xi Beam by Nutanix provides adequate tools for the key pillars of cost governance covered earlier in this chapter: visibility, optimization, and automated control.

Beam helps the business owners make data-driven decisions for ensuring cloud consumption remains within budget and aligns with business objectives. Meanwhile, cloud administrators can automate many of their mundane manual tasks using policies built into Beam.

Beam is a complete multi-cloud financial governance tool for public and private cloud environments.

IN THIS CHAPTER

- Understanding how enterprise cloud benefits businesses
- Making a case for a multi-cloud IT infrastructure

Chapter 6
Ten Reasons Why Multi-Cloud Is the Future of IT

Multi-cloud has a bright future in IT, and for good reasons. Here are ten of them:

- **A brand new economic model:** Your old legacy-based IT economic model is no longer sufficient. With a multi-cloud model, you can adopt the pay-as-you-go (PAYGO) characteristics of the public cloud while providing a common foundation upon which you can run both legacy and next-generation apps.
- **A focus on the end-user:** Your users are demanding new services, and you may not even know it. Until you discover why your users are deploying shadow IT systems, you may not understand their needs. A multi-cloud can help you focus your efforts on addressing the deficiencies that the end-users perceive.
- **Faster response from IT:** An IT team that utilizes multi-cloud architecture can instantly deploy new infrastructure, whether in the public cloud or private cloud. Response times of multi-cloud IT teams are much faster than those of traditional on-premises IT teams.

- **Refocusing IT on the business:** Multi-cloud can help you make this shift. You can tailor your IT department's services to activities that generate revenue rather than simply keeping the lights on.
- **Public cloud simply makes sense:** By using the public cloud, you get built-in economies of scale, instant deployment, and powerful management tools.
- **Private cloud simply makes sense:** You know that the public cloud doesn't always address issues such as data locality, security, and compliance in a way that works well for your company. However, the private cloud does so. By deploying a multi-cloud architecture, you get the best of both.
- **The trends are on your side:** Lots of trends came together to make multi-cloud viable. Public cloud providers are increasingly expanding their services into on-premises environments. Private cloud providers are increasingly building harmonious platforms that bring the best of both worlds to you. Multi-cloud is here to stay.
- **The choice is key:** Your organization needs a choice of where to run workloads. You shouldn't be forced into a single cloud. With the right multi-cloud foundation, you choose the cloud, and you don't end up with workloads trapped somewhere you don't want them to be.
- **You need to think beyond bimodal IT:** Bimodal IT — managing two separate set of policies for two different environments — isn't necessarily the best path forward. With multi-cloud architectures, you work with only one set of governing policies and don't have to worry about the inefficiencies of a bimodal model.
- **Users are smart:** Your users are far ahead of where they were a few years ago. Your infrastructure environment must reflect this fact by enabling user self-service and automation, both of which are better supported in a multi-cloud scenario.

ELIMINATE CLOUD COST LEAKS

- **Visibility into Cloud Consumption.** Get deep visibility into your multi-cloud consumption at an aggregate and granular level. Xi Beam automatically identifies cost anomalies to ensure cloud operators can immediately identify when spending deviations happen.

- **Optimization of Cloud Consumption.** Enjoy one-click optimization actions to easily right-size cloud resources. Xi Beam uses machine learning to continuously suggest optimal instance types – on demand, reserved or spot – that drive deep cost savings.

- **Control over Cloud Consumption.** Maintain high levels of cloud cost efficiency by automating various cost saving actions. You can also create budgets for various teams or projects, track the spending against allocated budgets and get alerts when a budget is exceeded.

NUTANIX™
YOUR ENTERPRISE CLOUD

Xi Beam

Get the best value from multi-cloud

Each new iteration of the cloud — public cloud, hybrid cloud, and multi-cloud — brings new complexity for CIOs and IT staff to grapple with in order to provide their companies with the right solution at the right time. This book is your guide to the rise of multi-cloud architectures and the critical challenges that you will face as you seek to adopt a wide array of services from different providers.

Inside...

- Understand workload operating models
- Compare public and private cloud models
- Drive out shadow IT
- Learn to right-size workloads
- Understand multi-cloud interoperability
- Establish a cloud governance framework
- Optimize your monthly cloud costs

NUTANIX

Scott D. Lowe is a former CIO and co-founder of ActualTech Media, a content creation and demand generation firm focused on creating content, conducting market research, and connecting technology companies with the right audiences.

ISBN: 978-1-119-60026-8
Not For Resale

for dummies® A Wiley Brand

Also available as an e-book

WILEY END USER LICENSE AGREEMENT

Go to www.wiley.com/go/eula to access Wiley’s ebook EULA.