Whitepaper

Bending the Curve Optimizing Cloud Costs through DevOps Automation

Subir Grewal
Global Head - AWS Practice

TABLE OF CONTENTS

Problem Statement

For many enterprises, migration to the cloud has been driven by artificial timelines, usually due to pre-determined data center exit dates. This has left most of them with a mix of legacy and cloud-native architecture. Cloud cost growth in such environments outpaces revenue growth, and senior executives begin to blame cloud environments for this.

Executive Summary

Rapid cloud cost growth rates lead to arbitrary mandates from senior budget owners, such as “reduce cloud spend by 30%.” Periodic fire drills or quarterly notices to right-size instances can be appropriate for smaller organization but are most times sub-optimal for large ones.

Cloud pricing is consumption-based, and organizations that rely on procurement to control technology spend perpetually chafe against the nature of cloud costs, unless they adopt cloud-native cost management strategies.

A reactionary monthly or quarterly “nastygram” sent by cloud cost czars will reduce some wasteful spending, but only months after the cost has already been incurred. Such an approach ignores the root cause of wasteful spend, and these attempts to reduce inefficiencies in running workloads months after they first appear is futile. Worse, it pits the finance and technology groups against each other.

In contrast, proactive investment in improved cloud architecture and automation **avoids wasteful spend in the first place.** We characterize this approach as improving the assembly line on which cloud workloads are built, rather than relying solely on quality control after the product has been assembled.

Organizations should recognize that rising cloud spend is not problematic in and of itself, but it can unnecessarily erode profits if it is not aligned to business value. Finance, Technology, and Business/Product teams need to collaboratively work to together to create real-time measures that align cloud costs with business value.

In addition, enterprises with cloud spend of 10 million USD or more per year can bend their cloud cost growth curve by using the cloud's scalability as it was designed to be used, in response to demand. Investing in DevOps automation and standardizing responsive scaling architectures produces a positive ROI when done correctly.

The Long-Term Cost of "Lift and Shift"

Each day, thousands of workloads are migrated from on-prem to public cloud platforms. Shockingly, most of these migrations follow a "lift-and-shift" migration pattern, where technology leaders seek to replicate in the cloud exactly what is built in their on-prem datacenter. They hesitate to modernize and leverage cloud-native patterns, because they incorrectly assume that the cloud will cost more if they make changes. It is not uncommon to hear a CTO say,

"Let's move it to the cloud first, and then we can modernize and reduce costs later," but this inherently leads to unnecessary and prolonged wasteful spend, because it causes over-provisioning of the cloud environment.

This approach offers tactical benefits and often makes it possible to exit data centers quickly when no other path is possible. However, lift and shift migrations leave organizations with more technical debt than when they started, with largely static workloads and a limited ability to control costs.

Such migrations fail to take advantage of some of the cloud's most powerful capabilities:

Elasticity

Ability to grow or shrink compute capacity real-time in response to changes in compute demands

Managed Services

Reduction in high personnel costs associated with managing underlying infrastructure

Modernization

A wide array of modern technologies and tools that can be customized to provide the ideal technology stack for workloads

Lift and shift migrations treat cloud platforms like moving to a cheaper datacenter, with a "pay as you go" cost structure. For large organizations, lift and shift migrations are sub-optimal unless followed by a determined application modernization effort that takes advantage of cloud native features.

The Gap Between Awareness, Intention and Execution

Technologists generally do not intentionally waste compute resources. When engaged, most cloud practitioners will successfully recite many of the approaches we outline here. Yet these practices remain aspirational for most, and realizing them in practice requires an intentional program to develop and mature cost optimization and cloud cost management processes.

Beyond Right-Sizing and Nastygrams

Picture the following scenario. The CFO is alerted to the fact that the company's AWS bill has tripled over the past month and immediately purchases a cost management tool that aggregates billing reports from multiple cloud vendors and provides a pivot analysis on this data. However, the CFO soon discovers that it cannot accurately identify the owners or cost centers for specific infrastructure and orders a resource-tagging exercise to properly identify this as a reactionary quick fix. This leads to a continuous and futile effort of "chasing the tail" in attempt to rectify the lack of clear cost allocation. This could have all been avoided if cloud cost management was considered **during the migration.**

Unfortunately, adaptive cloud cost management is often a secondary consideration during cloud migrations because they are typically driven by extraneous urgent deadlines. This leads to improper or incomplete tagging and an inability to accurately estimate and attribute cloud costs. Cloud spend spins out of control, and the finance team struggles to identify and reduce extraneous spend. The technology organization justifies spend as necessary, and senior executives are left with a vague sense that something seems off.

Several factors contribute to the confusion surrounding cloud costs and their drivers, including:

- Complex pricing at a very high level of granularity
- Lack of application automation, limiting scale-in capabilities
- Ease of cloud service provisioning
- Poor financial knowledge and complex billing
- Constantly changing cloud offerings

Effective cloud cost governance requires accurately identifying, measuring, and monitoring cloud spend. Native AWS tools that provide these capabilities to monitor spend and identify cost-effective alternatives include:

- AWS Billing Dashboard
- Cost Explorer
- Trusted Advisor
- Compute Optimizer
- Spot / Reserved Instances Management
- Budget Management and Notifications

In addition, there are other cost optimization tools available in the marketplace that perform the same basic functions. They collect usage data on cloud infrastructure and utilize it in two ways:

1. Recommend account changes to take advantage of discount/savings plans.
2. Identify specific resources with particularly low utilization and recommend removing under-utilized compute resources or archiving infrequently accessed data.

We characterize the second as tactical cost reduction, and it has several shortcomings:

- It identifies waste after it has occurred.
- It provides a shallow analysis of what constitutes waste.
- Discovery relies on manual processes, as does the response to address the issues.
- It does not respond to higher/lower utilization in real-time.

None of the cloud cost management or monitoring tools alone will bend the cloud cost growth curve, because they aim to treat the symptom and not address the cause.

To supplement cost management and monitoring tools, we outline three steps below to deliver long-term cost efficiencies and align cloud costs to enterprise value:

1. Building a consistent, cost-efficient architecture
2. Moving development and testing to a ride-share model
3. Implementing responsive scaling architectures for production workloads

These long-term cost-efficiencies will bend the cloud cost growth curve and align it to revenue growth.

Building a Consistent, Cost-efficient Architecture

The AWS Well-Architected Framework describes key concepts, design principles, and architectural best practices for designing and running workloads in the cloud.

Application workloads generally move through four stages during their lifecycle:

Discovery and Design

Implementation and Testing

Production and Growth

Retirement / Phase out

A mature cost-optimization strategy provides benefits at each stage of the lifecycle, but the second and third stages are the most impactful, as applications spend most of their time in a cycle of testing, followed by production deployment.

Moving Development and Testing to a Ride-Sharing Model

Historically, application development and test teams have jealously guarded their infrastructure since it was difficult to acquire and maintain. These patterns have survived migrations to the cloud despite a more efficient alternative being readily available.

The ideal state for development and testing workloads is to deliver truly on-demand test environments. Such environments are built when testing begins and are torn down once test results are produced for offline analysis, which provides a "just-in-time" delivery of test environments. To attain this ideal requires building easy-to-use test harnesses that automate on-demand provisioning of test infrastructure and data. Moving to such a model **requires an investment in tools and DevOps resources committed to making test environment provisioning as simple as a ride-sharing application.**

Relying on application development teams to build such automation is rarely a successful strategy. Most application development teams do not have the skills or expertise to build infrastructure automation, and they should focus on what they do best – developing and launching new product features. A dedicated DevOps team focused on improving infrastructure automation is better positioned to own such automation. Reference implementations and pipelines that deliver such automation while allowing for customization can bridge much of the gap between goals and reality here.

Responsive Scaling for Production Workloads

Most organizations move to the cloud for elasticity and flexibility. At the most basic level, cloud technology allows them to provision additional resources without investing in physical infrastructure, and "leasing" it instead as needed. This is a welcome change for anyone who has struggled to acquire and install physical hardware at a large organization. It is also where many organizations stop when it comes to utilizing the elasticity of workloads.

The ability to provision resources of any size at any time is both a blessing and a curse. Unregulated ability to provision any number of resources will inevitably lead to some form of accidental over-provisioning. Therefore, organizations must erect preventative guardrails to avoid this, such as fine-grained budget control at the account level. However, theses guardrails only go so far, and organizations must implement and leverage cloud-native capabilities to properly align cloud costs with overall business value.

The most efficient workloads utilize native cloud capabilities to scale in and out based on need. They conserve capacity by provisioning or scaling up to additional capacity when required, and only when required.

Responsive Scaling allows you to:

- React dynamically to changes in load.
- Schedule for predictable increases/decreases in demand.
- Reduce over-provisioning.

Migrating legacy application workloads to responsive scaling architecture is not a simple or easy endeavor. The first step is to stop digging the hole, by ensuring new application workloads adopt responsive scaling architecture if at all feasible. This builds expertise and capability within the organization, which can then be applied to the efforts to migrate legacy workloads.

Measuring and Bending the Cloud Cost Growth Curve

Measuring the value of cloud spend can be done in two ways, depending on the type of activity being supporting.

Mature Products

For mature products, cloud spend should be aligned directly with revenue. We propose a metric called Revenue Enabling Expenditure (REE), a ratio that is calculated as

REE = % Growth in Revenue / % Growth in Cloud Spend

As an illustration, a 20% Growth in Revenue that is enabled by 10% Growth in Cloud Spend would yield an REE score of 2. The REE metric should be included as a company-wide Key Performance Indicator (KPI) that is forecasted, measured, and assessed regularly and together with the company's other KPIs.

Developing Products

For products or features that are still being developed, a different measure is required. One way to assess the value of cloud spend is to measure it as a portion of overall Capital Expenses (CapEx) or product development budget. This ratio is, however, highly depends on the type of R&D and product development being done.

For example, a data-intensive product development process that relies heavily on analysis/back-testing may incur large infrastructure costs roughly equivalent to the cost of personnel. Meanwhile, a software development project is likely to have higher spend on personnel, with infrastructure costs around 20%. Although the actual value of this metric will vary from organization to organization, it is a critical one that needs to be measured and reported against to ensure cloud spend is aligned with business value throughout the product development cycle.

Conclusion

Managing cloud spend is one of the tougher problems in an organization's cloud journey. Most are unprepared for the challenges of managing cloud costs and are put into a reactionary position when hit with unexpected high bills. Historically, enterprises have controlled technology spend by controlling IT procurement, but now the consumption-based pricing of the cloud make procurement-driven controls ineffective.

Rising cloud costs are acceptable, however, when they deliver business value. Finance, Technology, and Business/ Product functions must align to create a data-driven yardstick that measures the business value delivered by cloud spend, so that they don't inadvertently stifle necessary innovation to meet their business goals.

Author Profile

Subir Grewal

Subir Grewal, Global Head - AWS Practice at Ness Digital Engineering, has over 25+ years of experience building and delivering impactful technology solutions using Agile methodology in Capital Markets. For the past ten years, Subir has focused on delivering cloud mobility initiatives for our clients.

Full-Lifecycle
Digital Transformation Specialist
Discover. Envision. Engineered.

About Ness

Ness Digital Engineering designs, builds, and integrates digital platforms and enterprise software that help organizations engage customers, differentiate their brands, and drive profitable growth. Our customer experience designers, software engineers, data experts, and business consultants partner with clients to develop roadmaps that identify ongoing opportunities to increase the value of their digital solutions and enterprise systems. Through agile development of minimum viable products (MVPs), our clients can test new ideas in the market and continually adapt to changing business conditions—giving our clients the leverage to lead market disruption in their industries and compete more effectively to grow their business.

For more information, contact **NDE.Marketing@ness.com**