WHITE PAPER

ESSENTIAL STRATEGIES FOR MANAGING CLOUD COSTS

Putting collaboration and automation to work

Flexera
Inform IT. Transform IT.™

Essential Strategies for Managing Cloud Costs

Putting collaboration and automation to work

Summary

As companies have increased cloud adoption over the past few years, public cloud spend has now become a significant portion of enterprise IT budgets. Easily accessible cloud infrastructure has enabled a DevOps model that's unleashing increased agility and speed. Although cloud adoption started from a decentralized approach to provisioning resources, enterprise IT is now doubling down on the benefits with cloud-first strategies.

The cloud necessitates a new approach

While the benefits of cloud adoption are clear, the on-demand nature of cloud use often results in uncontrolled cloud costs. Legacy pre-approval processes used for on-premises data centers or outsourcing contracts can conflict with the speed to market benefits that cloud delivers. As a result, CIOs and enterprise IT organizations need to develop new approaches and processes to manage and optimize cloud costs.

The lack of cost optimization processes is resulting in significant waste in public cloud spend. (Figure 1) RightScale has measured waste among enterprise cloud users and identified that, on average, 35 percent of cloud spend is wasted. This amounts to more than $10 billion in wasted cloud spend across just the top three public cloud providers (AWS, Azure, and Google). This white paper provides a framework for thinking about cloud costs and optimizing your cloud spend.

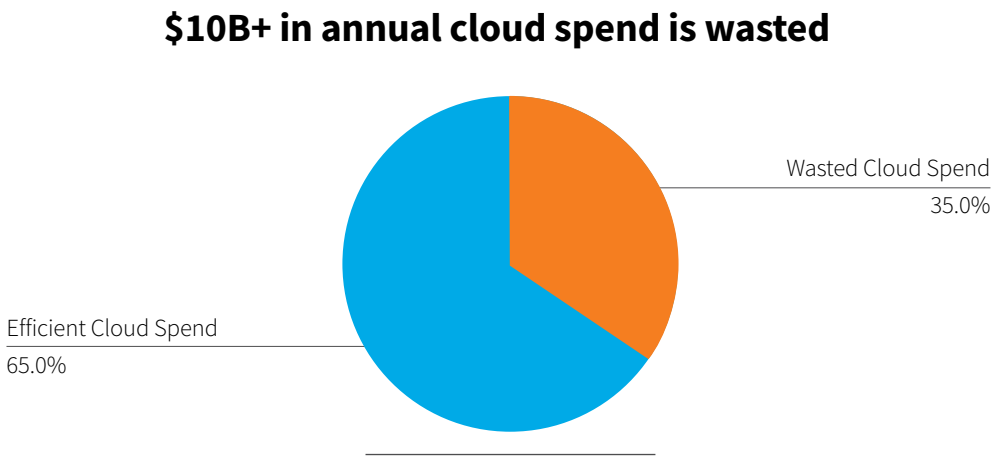

Figure 1: Annual Cloud Spend

A New Model for IT Spend

Barriers to efficiency of on-premises

On-premises spend requires companies to pre-buy capacity in order to meet imperfect projections of future demand. Overprovisioning is often rampant in order to meet peak periods of load. When hardware and software purchases are made months or years in advance of actual need, there's reduced motivation to make efficient use of resources since there's no near-term savings opportunity.

The opportunity of "down" and "off"

In contrast, public cloud provides the opportunity to realize immediate savings in operating costs from any waste that can be eliminated. The two most important words in cloud are "down" and "off." Rightsizing, scaling down, and eliminating idle or unused resources will quickly pay dividends when next month's cloud bill goes down.

Optimization is not "once and done"

The value of cloud is in its on-demand nature. This frees up developers and IT to instantly gain access to resources to solve business problems. This benefit of cloud creates challenges in governing costs when new resources are being continually provisioned, scaled, and de-provisioned. Achieving the goal of efficient cloud use must be more than a one-time event or once-a-quarter focus. IT teams, finance staff, and business users must become proficient in continuous, automated processes to monitor and optimize cloud spend.

Collaboration is key

Cloud computing has enabled more decentralized provisioning and ownership of IT infrastructure. Unlike traditional data centers that are owned by centralized IT groups, cloud resources are often owned and controlled by application teams and business units. Therefore, as centralized IT teams are being tasked with enterprise-wide cloud governance and cost control, they need to collaborate closely with resource owners in order to effect change and implement optimization processes.

> *"Cloud users are increasingly concerned about the money they might be wasting in their public cloud spending, but only a small percentage of them are doing something about it."*
>
> **InformationWeek**

The 4 whys of waste

In addition to the decentralized nature of cloud use, there are several characteristics of cloud use that lead to waste.

1. Complexity of cloud pricing options

While cloud pricing can seem simple on the surface—an hourly cost for a cloud instance or a cost per GB-month for storage—the reality is that there is a dizzying array of options from which to choose. There are tens of thousands of prices just for virtual machines across the three leading cloud providers. Instances can have significant price differences based on the region in which they run. Older version of instance families can be more expensive than their replacements. Even instances with similar amounts of CPU power and memory can vary greatly based on other add-on characteristics.

Storage, with its many different tiers and classes, can be just as complicated. Choosing storage classes beyond what is needed can result in significantly higher costs.

2. Difficulty selecting the appropriate instance sizes

As engineers and IT staff build and deploy applications, they need to decide which instances and sizes to provision. In many cases they may be unfamiliar with the performance characteristics of the cloud instances or of the applications they're deploying. When migrating instances from on-premises infrastructure, they may not know what the equivalent instance sizes would be. They will often take a "better safe than sorry" approach and select a larger size. Once the infrastructure is overprovisioned, it rarely gets downsized.

3. Resource owners don't have full visibility into cost implications

At the time of provisioning, resource owners often have little to no visibility into what their applications will cost in the cloud. The hourly cost of cloud instances can seem low, so they may not understand the full impact when they run instances for weeks, months, or years.

This limited visibility into costs can also be exacerbated during agile development processes, when teams are automatically provisioning and tearing down deployments for development and QA. If templates or automated scripts are used, instances can be repeatedly overprovisioned, resulting in a continual recurrence of wasted costs.

And once applications are running, resource owners may not receive reporting that enables them to see the cost implications of this overprovisioning.

4. Lack of automation to optimize workloads

Optimization is an ongoing challenge in the enterprise. Even after waste is identified and resolved, the dynamic nature of cloud use means that waste reoccurs. Automation is critical to dynamically monitor and respond to waste.

For cloud governance teams to ensure cost-efficient cloud use, they need automated tools that work across all of their cloud resource pools. They need to identify specific areas of waste and collaborate with resource owners to take automated action.

> *"RightScale found that many are passing up savings, and that overprovisioning resources is still widespread."*
>
> **Forbes**

Get recommendations to uncover waste

There are numerous areas where enterprises waste cloud spend. Running the majority of instances 24x7 and overprovisioning instances are common sources of waste. Thirty-nine percent of instance spend is on VMs that are running at below 40 percent of CPU and memory utilization, with the majority of those running under 20 percent utilization. This represents an opportunity to downsize and significantly reduce costs. Additionally, by not fully leveraging discounts offered by cloud providers and neglecting to clean up old storage data, enterprises are spending more than they need to. Following are several recommendations for reducing wasted cloud spend.

Recommendations for reducing wasted cloud spend	
INSTANCES	
IDLE INSTANCES	Instances that are no longer being used. Common with temporary instances for projects that have ended (dev, test, demo, training, experiments).
UNDERUTILIZED INSTANCES	Instances that have low utilization of CPU or memory and could be downsized or switched to a lower cost instance family.
PART-TIME INSTANCES	Instances (such as development) that are used only part of the time and could be scheduled to shut down during evenings or weekends.
SUPERSEDED INSTANCE FAMILIES	Instances from older instance families that have been replaced by newer lower-cost families.
HIGHER-COST REGIONS	Instances that are running in higher-cost regions that could be running in nearby lower-cost regions.
STORAGE	
UNATTACHED VOLUMES	Storage volumes that are no longer attached to instances and could be deleted.
OLD SNAPSHOTS	Snapshots that are beyond a snapshot retention policy.
OVERPROVISIONED STORAGE CLASS	Storage that has been provisioned as SSD and could be HDD. Storage that is provisioned as a higher class (hot, warm, cool, cold) than is needed.
OTHER	
UNUSED SERVICES	Services that have been left running but are no longer being used.
UNUSED ACCOUNTS	Accounts where services have been left running but the accounts are no longer being used.
DISCOUNTS	
NOT USING DISCOUNT OPTIONS	Not taking advantage of discounting options such as reservations or other volume commitments.
LOW COVERAGE WITH DISCOUNTS	Not purchasing enough coverage with discounts.
UNDERUTILIZED DISCOUNTS	Not using the appropriate resources to match already purchased discounts.

Collaborating for optimization

Because centralized cloud governance teams aren't the owners of the cloud resources, the biggest challenge they face is taking action on areas of waste. Once they've uncovered possible savings opportunities, they need to collaborate with the resource owners to take action. A typical collaboration process would include several steps by various stakeholders.

Collaboration steps

1. **Cloud governance team:**
   a. Identifies recommendations for savings
   b. Shares recommendations with resource owners
2. **Resource owners can then:**
   a. Identify recommendations that should be ignored
   b. Take action on recommendations
   c. Share recommendations further with other team members
3. **Cloud governance team:**
   a. Reviews results of actions
   b. Reports on savings

Resource owners can then determine whether any action should be taken. For example, a recommendation to downsize an underutilized instance may not be appropriate if it's used for a disaster recovery scenario.

> *"Roughly 35 percent of cloud computing spending is wasted via instances that are over-provisioned and not optimized, according to RightScale."*
>
> **ZDNET**

Tagging as a foundation

A consistent tagging foundation is needed both to allocate costs and better focus optimization efforts. These tags can then be used for automated policies that alert on or take action on waste. Some of the key tags needed for optimization include:

Environment

This tag can indicate whether resources are development and test and therefore can be scheduled. It can also be used to indicate instances that are temporary.

Schedule

This tag can be used to indicate the schedule and days when an instance should be run, along with the relevant time zone. Example: schedule=GMT+2:00-7-to-7-M-F or schedule=24x7.

Expiration

This tag can be used to indicate the date when an instance should be shut down. For example, an instance used for development might need to run for one week from the date it was started. Example: expirationdate=27-jan-2019 or expirationdate=never.

Owner

This tag can be used to provide the email address or other ID of the resource owner. It can be used to notify the owner of any issues such as missing tags or to alert the owner to optimization opportunities. Example: owner=name@company.com.

For tips on establishing a global tagging strategy and to see a matrix of tag requirements from the major public cloud providers, download our guide, Tagging Best Practices for Cloud Governance and Cost Management.

Automating optimization

As companies get cloud costs under control, automation becomes critical to maintaining an efficient cloud footprint. While manual cleanup of wasted resources is important during the early phases of optimization, the dynamic and decentralized nature of cloud use means that waste reoccurs as new resources are added and changed.

Automation is key to identifying, fixing, and preventing waste. The first step is identifying areas to automate. The next is using automation to create alerting and reporting on waste. The third step is taking automated action to resolve waste.

Key areas for automation	
INSTANCES	
ALERTS ON IDLE/UNDERUTILIZED INSTANCES	Alert on instances that are underutilized or idle.
ALERTS ON INSTANCES WITHOUT SCHEDULES	Alert on instances that should be scheduled (such as those tagged for development), but that are not tagged with schedules.
AUTOMATED STOP/START PER SCHEDULES	Use schedule tags to indicate instances that can be shut down according to a schedule (such as development instances that are not needed nights or weekends). Automation can then be used to stop and start instances according to the schedule.
ALERTS ON INSTANCES WITHOUT EXPIRATION DATES	Alert on instances that should have expiration dates (such as temporary instances), but that are not tagged with expiration dates.
AUTOMATED TERMINATION ACCORDING TO EXPIRATION DATES	Use expiration date tags to indicate the date when temporary instances can be shut down. Automation can then be used to alert resource owners in advance and then to terminate resources on the expiration date.
STORAGE	
ALERT ON OR DELETE UNATTACHED VOLUMES	Provide alerts on volumes unattached more than a certain period of time or automatically delete.
ALERT ON OR DELETE OLD SNAPSHOTS	Provide alerts on snapshots that are older than a retention policy or automatically delete.
OVERPROVISIONED STORAGE CLASS	Storage that has been provisioned as SSD and could be HDD. Storage that is provisioned as a higher class (hot, warm, cool, cold) than is needed.
DISCOUNTS	
ALERT ON UNDERUTILIZED DISCOUNTS	Alert when existing commitment discounts (such as AWS/Azure Reserved Instances or Google Committed Use Discounts) are not fully utilized.
ALERT ON EXPIRING DISCOUNTS	Alert when existing commitment discounts (such as AWS/Azure Reserved Instances or Google Committed Use Discounts) are close to expiration.
ALERT WHEN DISCOUNT COVERAGE LEVEL FALLS BELOW THRESHOLD	Alert when percentage of instances covered by commitment discounts falls below a specified threshold.

Waste prevention

As companies mature in cloud adoption, it's important that they develop processes and guardrails to help prevent waste at the time that resources are provisioned or address it immediately after provisioning. There are several strategies that can be used to place controls in self-service provisioning catalogs that help with optimization, including:

- Controlling the options for clouds, regions, services, and VM/instance families that can be selected
- Automatically selecting the best instance sizes based on specifications for vCPU and GB of memory
- Automatically selecting the lowest-cost cloud option at launch time
- Building auto-scaling systems to make best use of resources
- Attaching automated schedules and end dates to workloads

Collaboration is enhanced with RightScale Optima®

Collaboration and automation are the keys to ongoing automation. They require central cloud teams, business units, resource owners, and finance teams to work together to prevent, identify, and eliminate waste. RightScale Optima, a solution for cloud cost management and optimization, can help with cloud cost aggregation, allocation, reporting, forecasting, and optimization. Users can get recommendations for savings and collaborate to take action. And cloud teams can implement automation to prevent waste before it happens, alert on it once it occurs, and take automated action according to rules and processes you define. With RightScale Optima, you can optimize your cloud spend and minimize waste.

NEXT STEPS

To see how RightScale Optima can help, contact us

LEARN MORE

ABOUT FLEXERA

Flexera helps executives succeed at what once seemed impossible: getting clarity into, and full control of, their company's technology "black hole." From on-premises to the cloud, Flexera helps business leaders turn IT insight into action. With a portfolio of integrated solutions that deliver unparalleled technology insights, spend optimization and agility, Flexera helps enterprises optimize their technology footprint and realize IT's full potential to accelerate their business. For over 30 years, our 1300+ team members worldwide have been passionate about helping our more than 50,000 customers fuel business success. To learn more, visit **flexera.com**

Flexera
Inform IT. Transform IT.