TOP THREE CHALLENGES IN RESERVED INSTANCE MANAGEMENT

Introduction

The cornerstone of cloud resource consumption is the on-demand model, where resources incur cost only when used, and priced by the hour in the case of virtual instances e.g. EC2 services. No upfront commitments required–pay only for what you use, when you use it.

AWS offers an alternative, unique pricing model for its EC2 (and other) services – the **Reserved Instance (RI)**. This pricing model guarantees users the capacity which they reserved, whenever they need it (during the RI duration), and offers significant price discounts over on-demand pricing. In return, users make an upfront commitment for the usage of a virtual instance, bound to a specific family, size, availability zone (AZ) and operating system (OS), over the commitment period (1 or 3 years). This allows AWS to efficiently plan future capacity as well as customer commitment to its services. AWS also offers three payment options for RIs, which are **all-upfront** - bulk sum at day 0, offering the highest discount; **no upfront** - in which the cost of RI is paid in monthly installments over the duration of the RI, offering the lowest discount; and **partial upfront**, in which ¼ - ½ of the price is paid up front, and the rest in monthly installments, with a discount rate which is lower, but close, to the all-upfront rate.

For the customer, using Reserved Instances holds great promise in reducing the cloud bill and streamlining IT operations. At the same time, it introduces new challenges which can make the task of managing your cloud deployment complicated and tedious. In this white paper, we will address the top 3 challenges in Reserved Instance Management and offer a solution for overcoming these challenges.

Challenge #1 Utilization of RIs

Once RIs are purchased, they are non-refundable. Even if partial/no upfront RIs are purchased, the user commits to pay for them until the end of their term, regardless of whether they are utilized or not. This means that every hour in which an RI is not applied to running instances, i.e. not utilized, the ROI on it will shrink, and you will effectively be paying for services not used.

After purchasing RIs, it is critical that you stay on top of them and make sure each RI is utilized to the maximum, i.e. that in every given hour there is an instance which matches the RI characteristics (family, size, AZ, OS). If you detect that an RI goes unutilized for an extended period of time (>1 month), then it's best to take action in order to utilize it.

Three options exist for utilizing unused RIs:

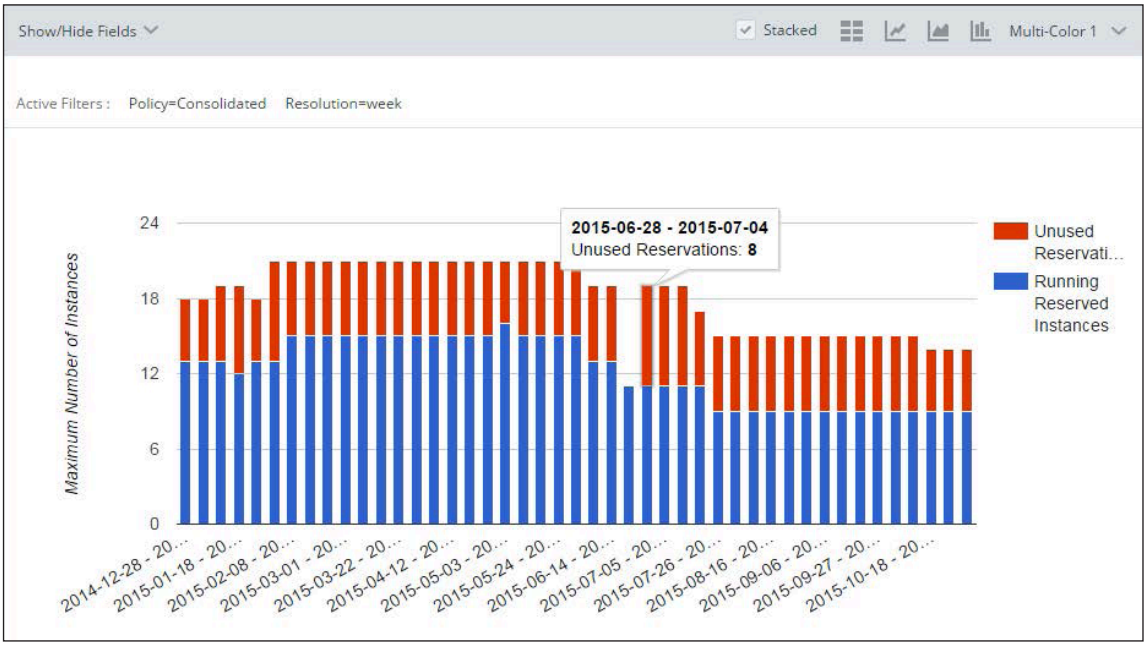
1 **Sell**: AWS holds an RI marketplace where customers can sell their reservations to other customers for the duration of their validity. This way, part of the RI cost is recovered.

2 **Modify**: As mentioned earlier, RIs are purchased for a specific instance family, size, AZ and OS. AWS does, however there is an option to modify existing RIs in order to make them fit instance usage (for Linux-based instances only):

- Modifying the instance size, while keeping within the same instance family), and following a conversion table for size.
- Modifying the availability zone of the RI (while keeping within the same AWS region).
- Switching from "Classic EC2" (non-VPC) to VPC and vice-versa.

3 **Use**: While seemingly trivial, this is the optimal method for utilizing RIs. Let users in your organization know that RIs are available for certain type instances, and encourage them to choose this type when provisioning new resources. This way, the users get it for "free" (as it's already paid for), and the RI gets utilized, as opposed to the case where the organization pays twice: for an unused RI and for new instances provisioned on-demand.

EC2 Reservations over time



STRATACORE
DATA CENTER + CLOUD + CONNECTIVITY → BUY SMARTER

2320 2nd Avenue
Seattle, WA 98121

Phone: 206.686.3211
Fax: 206.456.5826

Web: stratacore.com
Twitter: @Strata_Core
Facebook: /StrataCore

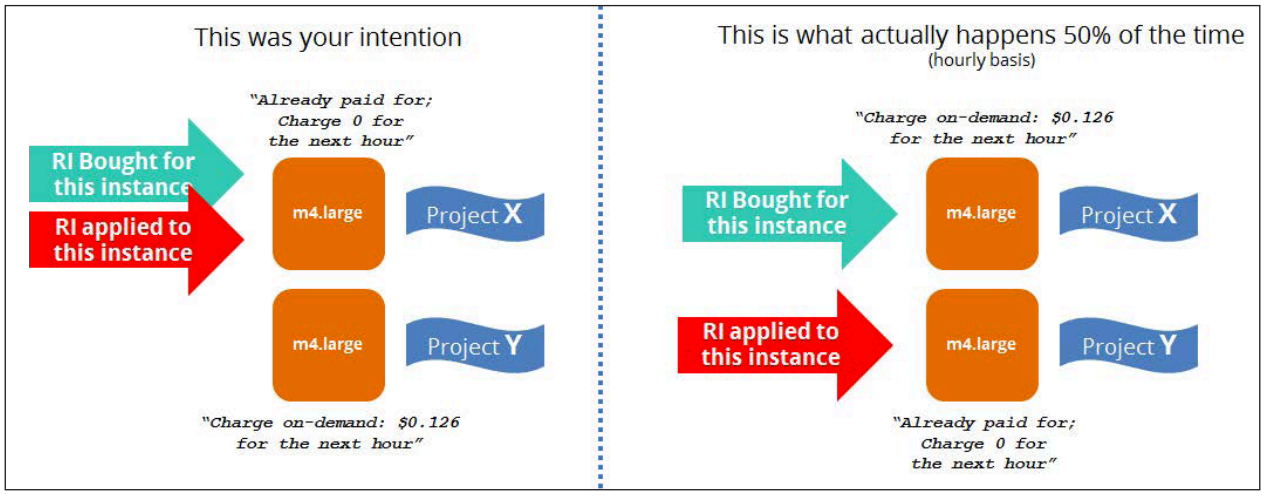
Challenge #2
Cost Allocation of RIs

It is important to understand that RIs are not attributed to a specific instance. Themanner in which AWS applies RIs to instances is as follows:

- First, AWS looks for matching instances (family, size, AZ, OS) in the account.
- If a single match is found, the RI is applied to it.
- If multiple matches are found, the RI is applied to one of them **at random**.
- If no matches are found, AWS will look for matching instances in any linked or consolidated account which is under the same billing family ("floating" the RI) and apply it to them.
- If no matches are found in "sister" accounts, the RI remains unutilized.

- This process is implemented for **each RI, every hour.** The random application of RIs creates a challenge in allocating the cost (payments) and benefit (discounts) to a specific instance or cost entity. To further explain, let's take a look at the following example:
- A company has two projects running on AWS. Resources for the projects are provisioned from the same AWS account.
- An internal chargeback process is in place, which charges each of the projects for their consumption of AWS resources. The instances are tagged with a "Project" tag, attributing each one of them to Project "X" or to Project "Y".
- Project manager X would like to lower his costs; he requests to buy an RI for an m4.large instance running in his environment.
- Ideally, he should be charged for the RI payments, and not be charged for hourly usage.
- In practice, this is not the case. As explained above, the RI is applied to a random matching instance every hour. This means that if project Y is also running matching instances, the RI could be applied to them as well on certain hours.

AWS does not enable attribution of RIs to specific instances or tags



EXPERIENCE THE DIFFERENCE. WORK WITH AN ADVOCATE.

2320 2nd Avenue
Seattle, WA 98121

Phone: 206.686.3211
Fax: 206.456.5826

Web: stratacore.com
Twitter: @Strata_Core
Facebook: /StrataCore

So, at the end of the month, Project X will receive a bill for a monthly RI payment, as well as on-demand charges for that same instance. Project Y on the other hand, will not be charged with RI payments, and will also not be charged for part of the hours in which its instance was running on-demand.

The straightforward approach to overcoming this challenge involves manipulating billing reports manually, and consumes precious human hours, while being prone to multiple errors.

Challenge #3 The Complexity in Large Numbers

The challenges mentioned above can be addressed and mitigated manually as described, only if a handful of RIs are purchased. Once you pass the barrier of ~10 RIs, manual management becomes a tedious, error-prone process. The complexity of managing and mitigating the negative effects of these challenges when RIs start adding up, is near impossible. Just think about keeping track of payments, utilization rates and cost allocation when you have tens, hundreds or thousands of RIs.

The Solution Automated Management

Any number of RIs, more than a handful, requires an automated solution which will:

- Provide recommendations for purchasing RIs based on actual usage.
- Manage all RI payments and expiration dates, making sure that expiring RIs are renewed (or not, in case they're not needed anymore).
- Keep track of unutilized RIs and suggest modifications according to actual usage.
- Provide accurate and reliable cost allocation of RIs, attributing both cost and benefits to the purchasing cost entity.

Such a solution would not be limited to management of RIs, but also to **monitoring** and **optimization** of your whole cloud deployment, across accounts, vendors, cost entities and services. **Inefficiencies** in cloud deployments lie in underutilized instances, unused or forgotten data storage, sub-optimal pricing models and much more. The automated nature of the cloud, and the transparency in which usage and billing data can be retrieved via APIs, allows for such solutions to monitor your deployment, detect and mitigate inefficiencies, provide **reliable cost allocation** and facilitate **cloud cost transparency and accountability** in your organization.

Summary

TReserved Instances is a unique pricing model offered by AWS. It encompasses an upfront commitment to use certain types of instances for a given period of time (1 or 3 years), in return for a discount on usage rates as compared to on-demand rates.

While holding great promise and opportunities, RIs introduce several challenges and risks, the most significant ones being:

- Utilization
- Cost allocation
- Complexity of managing large numbers

For more information on AWS RIs, and Cloudyn's solutions for managing your cloud deployment, contact us at **info@stratacore.com**

About StrataCore

StrataCore is the premiere Data Center, IT Infrastructure, Connectivity, and Cloud Services Agency in the Pacific Northwest. We partner with the industry's top service providers to save our clients time and money - while maximizing business results. We offer unbiased, custom solutions while maintaining a clear view of the competitive landscape to optimize contract terms and pricing. Our market intelligence, tools, and detailed vendor selection process provides clients with the necessary insight to make informed IT decisions.

For more information on how we help save our clients money, visit us at **www.stratacore.com**.

STRATACORE
DATA CENTER + CLOUD + CONNECTIVITY → BUY SMARTER

2320 2nd Avenue
Seattle, WA 98121

Phone: 206.686.3211
Fax: 206.456.5826

Web: stratacore.com
Twitter: @Strata_Core
Facebook: /StrataCore