INTRODUCTION TO AWS SPOT INSTANCES

Amazon Web Services offers a variety of purchase options that allow customers to optimize their cost investments based on actual usage requirements changing in real-time or dynamically. Customers can purchase capacity, host infrastructure and instances on demand for short term use, schedule subscriptions for recurring demands or reserve instances for the long term. The cost for each option can vary based on capacity and usage statistics, schedule, AWS services and solutions employed.

AWS essentially offers a purchase structure for every type of organization and usage requirements. The company is continuously expanding its infrastructure and pursues a wider market to maximize profitability. This means that a range of AWS infrastructure resources are left unused at any given moment. Failing to sell the available resource capacity translates into lost business opportunity and potentially losses for the cloud vendor. In order to attract more customers, AWS therefore offers the purchase structure of Spot Instances, which allows customers to take advantage of the unused AWS capacity at a heavily discounted price.

How AWS Spot Instance Billing Works

AWS Spot Instance pricing structure offers up to 90 percent discount on the On-Demand instance service. The service is identical to the regular EC2 instance offering except that it can be interrupted when the pricing increases the amount specified by the customer or when the capacity is no longer available. The company evaluates the supply and demand trends to set a price for the Spot Instances and changes based on instance availability. The customer gives their maximum bid for the Spot Instance and continues to receive the market spot price as long as the latter doesn't exceed the customer's bid. The market spot price is determined at the beginning of the instance-hour and billed on a per-second basis to the customer. As a result, the pricing is usually stable for at least one instance-hour and only changes as long as demand-supply gap justifies the discounted price. The service gets interrupted differently based on the capacity availability of the particular instance type and Availability Zone, although AWS claims the average interruption frequency is less than 5 percent.

As a simple example, consider that an instance is billed at 1 USD per hour for a regular On-Demand pricing structure. With the Spot Instance pricing structure, assuming a low demand and high supply availability of the same instance, AWS sets the pricing for the same instance at 0.3 USD per hour. A customer willing to purchase the instance bids 0.6 USD per hour as part of the Spot Instance pricing structure. The customer will then be able to use this instance at the spot market price of 0.3 USD per hour until the instance is no longer available (due to higher demand from customers opting to pay the On-Demand tariff for it) or the number of Spot Instance customers for the same instance option increases and AWS adjusts its spot market price at higher than the customer's original bid of 0.6 USD per hour. Indeed, the customer could have bid the maximum On-Demand price and continue to take advantage of the Spot Instance discount as long as the spot market is lower, which will always be the case as long as the instance is sold within the Spot Instance billing model. This means that customers need to optimize their choice of Spot Instance bid based on their budget, value to business as well as the alternate pricing models available on the AWS store. Here's a list of best practices to quickly guide you through this goal:

- **Diversify Capacity Pool:** AWS offers infrastructure capacity pools containing instances of specific types for each AWS Region and Availability Zone. Spot Instance customers can choose launch specifications to set the resource allocation strategy for their apps. Select multiple capacity pools from different Regions and Availability Zones to reduce the impact of interruptions to your app. The AWS system can rebalance and distribute the workload across the capacity pools to meet specific Spot Instance bid requirements while accounting for the availability of instance on the Spot market. AWS Spot Instances can be used strategically along with On-Demand and Reserved Instances to optimize cost while maintaining the necessary performance requirements. Follow this AWS guide for valuable information on devising the optimal allocation strategy for your Spot Instances.
- **Improve Price-Awareness in the System:** Develop price-aware systems that account for the infrastructure resource allocation. Using appropriate storage and instance types, tracking and analyzing usage, and developing a mechanism to deal with interruptions can help optimize cost savings using the Spot Instance service.
- **Analyze Historical Pricing:** AWS provides information on the historical pricing for capacity pools for the last 90 days. This information can help evaluate price sensitivity and optimize budgetary decisions accordingly. Older generation of instances with lower price swings and interruptions can be expected to continue in the same way, as future spot market price is also determined by past supply-demand trends.
- **Choose Fewer, Larger Instances:** While there is no definitive answer to the multiple small instances versus fewer large instance debate, the case for AWS Spot Instance pricing might differ. The baseline price of instance classes is assigned depending upon demand and the relative operational cost of smaller instances may be higher to address the rising demand in the spot market. Also, lower demand on larger instances tends to reduce its spot market price, as seen through historical data. For the On-Demand service, the pricing structure tends to change linearly across instance types, sizes and categories. This is not the usual case with the AWS Spot Instance and further investigation may be required to identify the best pricing

conditions available for different instance options.

- **Use Cases:** Spot Instances are not suitable options for every type of workload. In fact, AWS customers typically use the service for two types of workloads: time-insensitive workloads that are not bound by SLA requirements; and peak-time time-sensitive workloads using On-Demand instances that may require additional supplement capacity to accommodate demand spike for a predictable time period.