THE STATE OF THE AMAZON EC2 SPOT MARKET

RESEARCH, CONCLUSIONS AND OPPORTUNITIES

www.spotinst.com
Amazon Spot instances differ from on-demand instances in that they are acquired through a bidding process in which customers can specify the maximum price per hour that they are willing to pay. The purchased Spot instances are available as long as the Spot price is lower than their bid price.

However, with the possibility that a Spot instance can be terminated with a two-minute notification, corporate decision-makers are understandably concerned that AWS Spot instances are hard to use and require resources to maintain uptime assurance levels.

**Workload Types Running On EC2 Spot Instances**

The common perception was that Spot instances can only serve workloads such as batch processing. That was true years ago. Today it is much more common to see cloud deployments that leverage Spot resources as an essential part and best practice for their production environments. This includes application tiers that work with a load balancer, cache clusters (e.g., Redis, Memcached, etc.); container platforms (e.g., Amazon ECS, Kubernetes); NoSQL clusters (e.g., Elastic, AeroSpike, Cassandra); and CI/CD pipelines and big data workloads (e.g., Hadoop, Spark).

In average, it takes approximately 3-5 weeks to a company to adopt the Spot market. Where at the very beginning users usually express skepticism, and willing to test several dozens of Dev / QA environments and a small portion of their production workloads. Then, at around 3-4 weeks, trust is being gained dramatically and we see an exponential increase to around 60% their entire EC2 production workloads.
If we break the usage of Spot by environment to a chart, we can see that the vast majority of our customers run their Spot workloads in production and mission-critical environments with no doubt, using technologies like Spotinst.

The **Spotinst** core technology is based on an in-house prediction algorithm. This algorithm chooses the most effective and most likely available EC2 Spot instance options for Amazon customers costs. We decided to leverage the data we collected and use this engine to conduct research and shed light on various aspects of the AWS Spot instance market. These include instance availability, regional capacity and cost-saving opportunities.

Looking at the current market players, we believe that we have the technology, knowledge, statistics, experience and historical data required to conduct this research and provide valuable insights to the market.
TABLE OF CONTENTS

The Research ................................................................. 5
The Spotinst Forecast Engine ........................................... 6
Instance Availability ....................................................... 7
   By Region ................................................................. 8
   By Instance Type ...................................................... 9
   By Family ............................................................... 10

GPU Instances Are The Least Available .......................... 11
   By Size ................................................................. 12

Cost Savings ................................................................. 14
   Savings by Family and Size ......................................... 14
   Savings by Type and Region ....................................... 16

Final Note: The Real Cloud .............................................. 18

About Spotinst .............................................................. 19
The data we collected over the past few years has helped us generate the necessary research benchmarks and insights. Every minute, our systems record capacity per region, availability zone, and instance family and size - as well as uptime and the hourly cost of each Spot instance throughout its lifetime.

Prices are recorded every hour for each Spot resource, and uptime is measured from the moment the instance was provisioned until it is forced to be terminated. Our research deals with the instance availability and cost savings of the Spot resource in comparison to its on-demand parallel instance.

The findings and insights provided here are based on an analysis across the lifetime of over 40,000 EC2 Spot instances over a six-month period, from March-September 2016.

Disclaimer: This report is not based on anecdotally-reported usage. The research represents an accurate review based on real data. AWS users should not treat our findings and insights as rules that decide their bidding strategy. The research findings and conclusions should not conflict with instructions, recommendations or any other information provided by Amazon to its users.
Over the last two years, we developed our Mathematical Forecast Engine, which runs quality forecast research results, including the availability and potential savings that Spot instances can provide.

The forecast engine uses several mathematical models - such as the multiplicative damped trend and Holt’s linear trend - which support forecasts based on trending data. Our system aggregates and uses Spot prices and uptime history data together with these forecast algorithms to make informative decisions on the bid price we place on behalf of our clients. The combination of multiple models creates a quality forecast of available compute capacity and price in each Amazon region, as well as across the entire public cloud.

Our customers have leveraged this to automatically create and maintain their blended clusters composed from on-demand, reserved and Spot EC2 instances. In order to optimize clusters, our system automatically balances between the three price models, and as part of that maintains and manages the Spot instance purchasing operations. This includes setting the bid and selecting the instance and region. These are based on multiple parameters including available capacity, potential savings and user preferences.
Spot prices fluctuate periodically, depending on the supply and demand of each specific region. This makes each AWS region an independent market with its own instance prices (not only for Spot). It also means that the region is a crucial parameter when considering Spot instance purchasing.

As shown in the chart below, the regions with the largest pools of Spot instances are in us-east-1 (N.Virginia) and eu-west-1 (Ireland). It might be somewhat surprising that these hold spare compute capacity, considering that they are probably the most popular AWS regions. Based on our experience, we can safely say that users bidding in these regions enjoy hundreds of Spot instances available for weeks before automatically being terminated. This means less price fluctuations and higher certainty.
We found that the least utilized regions with the smallest Spot instance pools are eu-central-1 (Frankfurt), the new India region and the Singapore region. From our research we learned that in comparison to the most utilized regions, these data centers hold tens of Spot instances that are available for only several days, before they are automatically terminated.

In one of our experiments, we launched hundreds of C3, C4 and R3 EC2 instances, from the 8xlarge instance size pool of multiple AWS accounts in parallel (to eliminate the account-specific capacity limitations). We noticed that the price was growing in a linear fashion, while these regions filled up quickly. This might be due to the lack of supply, and to regional immaturity (in the case of India, which was launched in June 2016).

In addition, two new regions were recently announced: one in Canada (ca-central-1), and one in Ohio (us-east-2). It is highly recommended to purchased Spot capacity in these regions.
For this part of our research, we focused on understanding demand in comparison to each Spot instance type availability. We hope that the information we reveal here can help users switch between instance types in order to get better uptime for a better price.

Based on our sample and as shown in the chart below, we found the following:

- The most available Spot instance sizes are 2xlarge and large.
- The most popular Spot instance type is m3.2xlarge.
- The least available Spot instance size is 4xlarge.

Figure 3: Instance Capacity vs. Instance Type
The Winners: M3, C4 and C3

From the chart above, we see that the M3 family has good available Spot instance capacity. M3 is a general purpose instance with an expanded memory, and is well-suited to run mid-size databases. Amazon recommends using it to run backend servers of enterprise applications such as SAP and Sharepoint.

m3.2xlarge should be strongly considered when looking for a stable Spot instance.

The instance type chart above also shows that m3.2xlarge in particular is the most available. This is not surprising considering that it is a strong and well-balanced resource in terms of its compute power (8 vCPU) and memory (30GB). It should be strongly considered when looking for a stable Spot instance.

In addition to M3, the C4 family is also highly available throughout the various regions. This compute-optimized family supports the scalability required to run workloads such as web applications, containers and customer-facing APIs.
We saw no potential savings when trying to bid on the g2.2xlarge type.

In 2013, Amazon introduced the G2 series, the first using Nvidia’s GK104 Kepler GPUs. As you can see in figure 5 below, in almost all cases we saw no potential savings (the market price is higher than on-demand) when trying to bid on the g2.2xlarge type. This implies a significant lack of capacity for these instance types.

However, as the market matures, we can expect the overall availability of this instance family to increase. At the last re:Invent conference, Amazon unveiled their P2 instance, which is a powerful multi-vCPU processor and multiple, well-connected GPU instance. Furthermore, Amazon was the only cloud vendor with a GPU instance, until Azure recently announced their N-Series, their own GPU compute resource.
In addition, at re:Invent 2016 AWS announced new R4 instances, I3 instances and C5 instances. They also announced new Elastic GPU instances: eg1.medium (1 GiB), eg1.large (2 GiB), eg1.xlarge(4 GiB), and eg1.2xlarge (8 GiB). We believe that the availability of GPU instances will improve significantly in the next several months.

The most available instance size: 2xlarge and large

As shown in the chart in Figure 3 above, 2xlarge is the most popular size. Since Amazon is able to dynamically allocate the size of their instances, it can support the great demand for this Spot resource size. In our sample, for example, we observed an i2.2xlarge that was available for 76 days.
The least available instance size: 4xlarge

In the same chart, you can see that in comparison to more than 40K of 2xlarge provisioned instances, there were no more than 10K 4xlarge provisioned instances. From the chart we can conclude then that users prefer the smaller size Spot instances such as xlarge and 2xlarge. This impacts the supply of other larger size instances allocated by Amazon. We can assume that this is likely the case with on-demand instances as well.

In addition to these findings, we should mention the large and xlarge size instances that we saw available for many days in popular regions, as shown below:

Figure 6: Uptime Duration in Days by Instance Type and Region
The following is the cost savings data we found in our research. This data is based on hundreds of thousands of Spot price history records.

78.4%: Average actual savings in Linux servers, in comparison to on-demand instances.

57.3%: Average savings in Windows servers, in comparison to on-demand instances.

The potential cost savings below are calculated based on the price history we recorded during the sample period, across all regions.

<table>
<thead>
<tr>
<th>Family</th>
<th>M3</th>
<th>M4</th>
<th>C3</th>
<th>C4</th>
<th>R3</th>
<th>D2</th>
<th>I2</th>
<th>X1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost savings</td>
<td>84%</td>
<td>81%</td>
<td>74%</td>
<td>75%</td>
<td>72%</td>
<td>71%</td>
<td>79%</td>
<td>79%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Family</th>
<th>Large</th>
<th>Xlarge</th>
<th>2xlarge</th>
<th>4xlarge</th>
<th>8xlarge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost savings</td>
<td>82%</td>
<td>78%</td>
<td>81%</td>
<td>72%</td>
<td>76%</td>
</tr>
</tbody>
</table>

“Users should consider the most available ones - the M3 and M4 families - which on average together generated more than 80% in cost savings.”
As you can see, the average savings across all Spot instances families in AWS can lead to significant cost reductions. In particular, users should consider the most available ones - the M3 and C4 families - which on average together generated more than 80% in cost savings. In addition, based on our sample, 2xlarge instance sizes should also be preferred based on their potential savings and availability.

X1 is a memory-optimized instance family designed to run large-scale, in-memory enterprise applications such as SAP HANA. Though we see a lack of X1 Windows Spot instance capacity, it can generate great costs savings.

In addition to the observed 79% average savings on X1 Spot instances, we also saw that in its peak the instance created 89.54% in cost savings. The on-demand instance price is $13.338/hr, and with this discount we can get it for $1.39/hr. On a monthly basis (30 days), that translates into $8,602 in cost savings for a single x1.32xlarge instance!

One of the more interesting cases is the T2 instance family, and their low costs. T2 on-demand instance prices started at $0.0065/hr, which makes T2 the lowest-cost Amazon EC2 instance family. According to Amazon, T2’s purpose is to serve workloads that don’t use their full CPU power consistently, and that can compromise on performance. These make them good options to support non mission-critical tasks such as development and testing.

Surprisingly, the T2 instance type is not supported as a Spot Instance. We assume that the reason Amazon doesn’t allow T2 Spot instance provisioning is because Spot instance users tend to take the instances
to their full capacity, and performance of other T2 tenants can be significantly harmed.

However, checking the below differences when using M or C families Spot instances makes the lower-performing T2 family a less desirable option. This is taking in mind that the M and C instances can be provisioned for the same price, if not lower with better performance.

The following are the price comparisons for an entire month:

<table>
<thead>
<tr>
<th>Type</th>
<th>T2.Small</th>
<th>T2.Medium</th>
<th>T2.Large</th>
<th>M4.Large (spot)</th>
<th>C4.Large (spot)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spec</td>
<td>1 vCPU (credit)</td>
<td>2 vCPU (credit)</td>
<td>2 vCPU (credit)</td>
<td>2 vCPU</td>
<td>2 vCPU</td>
</tr>
<tr>
<td></td>
<td>2 GB RAM</td>
<td>4 GB RAM</td>
<td>8 GB RAM</td>
<td>8 GB RAM</td>
<td>4 GB RAM</td>
</tr>
<tr>
<td>Monthly cost</td>
<td>$19</td>
<td>$38</td>
<td>$76</td>
<td>~$14</td>
<td>~$19</td>
</tr>
</tbody>
</table>

Figure 8: T2 vs M4 and C4 Spot Prices

The maximum discount in our sample was 96.49%. This was for the previous generation CG1 (GPU) instance type.

Our assumption here is that since users are encouraged to move to second generation instances, AWS is left with significant spare capacity of older instances - which makes their Spot instances an appealing option.

The next on our list was the m2.xlarge Spot instance, which generated 94.05% savings on August 7th in the EU1 region. In this particular occurrence, users paid $1.57/day instead of $26.40/day, which translates into hundreds of dollars in savings on a monthly basis.
Though there are potentially significant savings on Spot instances across all regions, we should not forget that each region has its own supply and demand, and should be considered separately as a different market.

The heat map below provides more information about the several popular instance types and their average savings by region. Note the negative values specifically in the ap-northeast-1 and us-west-1 region (N. California), which implies low instance capacity of the specific instance type.

![Heat Map](image)

**Figure 9: Cost Savings (%) by Region and Instance Type**
The Spot market today already supports multiple use cases such as big data advanced financial modeling, software load testing and DNA sequence analysis. This market creates the efficiencies that continue to lead to great new opportunities across a wide range of industries.

“We believe that the Amazon Spot market represents what we refer to as the “real cloud” - a dynamic environment that encourages diversification across regions, types and more. This helps support stable and robust workloads, at optimal, low costs.”

Amiram Shachar, Spotinst CEO

We hope that our research has helped you better understand how the Amazon cloud works, especially when discussing supply and demand. Most importantly, we hope that it will give you some valuable and insightful information for ‘better bidding’ the next time around.
Using advanced machine learning, Spotinst allows cloud customers to bid on unused public cloud capacity, including Amazon and Google. The Spot instances provision available through Spotinst enables cloud customers to optimize the cost and quality of their cloud computing deployments.

**Spotinst Elastigroup** facilitates balancing between on-demand, reserved and Spot instances, and auto-decides how to deploy the resources to optimize the balance between a cloud deployment cost and performance. It continuously analyzes and leverages the diversity of clouds, regions, zones and resource types - to ensure high availability of users’ clusters.

**Contact Us**

+1 415.848.9002
sales@spotinst.com
680 Folsom St, San Francisco, CA 94107.
[www.spotinst.com](http://www.spotinst.com)