Densify

The Definitive Guide to Optimizing Reserved Instances

What is a Reserved Instance?

As public clouds make up an ever-greater slice of the enterprise computing pie, companies are demanding solutions that help them gain control over spiraling costs. Major cloud providers, including Amazon Web Services (AWS), Microsoft Azure and Google Cloud, have responded by offering customers ways to realize cost savings by committing to a certain level of consumption. Amazon calls this Reserved Instances (RI), while Microsoft uses the term Azure Reserved VM Instances, and Google Cloud uses a similar concept called Committed Use Discounts (CUD).

The concept is simple. Reserved Instances offer substantial discounts compared to pay-as-you-go prices when you commit to a specified cloud capacity for a specified period of time—usually one year or three years. In some cases, they also provide guarantees that resources will be available to you in a specific hosting region.

As with normal instance purchases, you select a specific instance type and size (e.g. m4.large), platform (e.g. Windows) and region (e.g. US East 1) when purchasing the Reserved Instance. It's like buying a voucher that you can use to get a discount at any time during the reservation period. In most cases, this "voucher" can be shared among departments or divisions.

How to use RIs to reduce your cloud costs

There are some differences among the various RI offerings that offer varying degrees of flexibility (see Table 1 for a comparison of features). Some offer Convertible RIs, which allow you to change some attributes of an RI. As a general rule, longer term commitments offer the greatest savings, because they provide the cloud service provider with the highest level of predictability.

Feature	AWS	Azure	Google Cloud
What You Can Specify	Instance Type, OS, Region/AZ, Tenancy	Instance Type, Region	vCPUs and Memory per Region
Purchase Terms	1 or 3 Years	1 or 3 Years	1 or 3 Years
Purchase Options	All Upfront, Partial Upfront, No Upfront	All Upfront	No Upfront
RI Convertibility	Additional Cost	Free	N/A
Cancellation Option	Marketplace – 12% fee	Cancellation – 12% fee	None
RI Coverage Flexibility	Linux, Shared Tenancy, Region Scope	No	Yes

Table 1. Comparison of basic features of RIs from Amazon, Microsoft and Google.

How much can you save? Amazon claims its EC2 RI provides discounts up to 75% as compared to on-demand pricing. Microsoft Azure claims savings up to 72% (up to 82% with Azure Hybrid Benefit, which enables you to apply existing Windows Server licenses to Azure virtual machines at a discount). With savings of this magnitude, Reserved Instances deserve your attention.

However, there is a catch. A Reserved Instance is a "use it or lose it" proposition, and the financial benefit is lost every hour that an RI goes unused. To reap the full financial benefit of Reserved Instances, you need to maximize your use of the RIs that have been purchased, and when you go to buy more, make sure that you are purchasing RIs that match what you need. That requires analysis.

The instance selection challenge

The task of selecting the optimal Reserved Instance to meet application workload needs is extremely complex. On the "supply" side, cloud providers offer hundreds of possible virtual machine configurations—and these offerings are constantly changing, making it almost impossible to keep pace with what is available at any moment.

On the "demand" side, trying to analyze what an application needs and its workload patterns by studying raw utilization data is extremely difficult, especially for organizations that may be hosting large numbers of workloads in the cloud. Manual analysis methods are time-consuming and imprecise, which introduces risk into the equation. It is, therefore, not surprising that this analysis is performed infrequently or not at all. Yet it should be.

A common mistake people make is using bill reader tools to determine RI requirements. These tools look at instance usage hours, rather than how the instance is actually using its allocated resources. They can only provide minimal insight—or, worse yet, the wrong insights. The information they generate may lead you to purchase RIs based on the instances you are currently using, even if they are in the wrong instance type and size. *The result: You may end up reserving the wrong instance types and actually wasting money in the long run.*

To avoid this mistake, you should first optimize your cloud environment, finding the right match for your workloads, and then reserve. However, it requires a much deeper analysis. Given all the variables involved—including complex application workload patterns and continuously changing RI offerings—performing this analysis manually is just not possible. So how can you be sure you're selecting the right RI to maximize your savings, without putting your cloud-hosted applications at risk?

Machine learning to the rescue

Performing a deep analysis of cloud workload dynamics, and doing it rapidly and accurately, requires intelligent technology. Specifically, it demands machine learning. This is the only way to effectively automate the process of cloud instance selection by fully analyzing application workloads in depth, over time, and mapping these insights to available RI configurations.

On the supply side, this analysis monitors and aggregates cloud providers' on-demand and RI offerings, what they can do, and how much they cost based on the various types of pricing structures. When new services come online, it knows about those as well.

On the application demand side, machine learning technology can acquire a detailed understanding of your application workload dynamics, including:

- Each workload's CPU, memory and I/O demand patterns over time, including time of day, business cycles etc.
- Policy constraints and business preferences/priorities that impact workload requirements

In this way, machine learning technology establishes predictive demand patterns for each workload. This in-depth prediction of workload activity is essential for optimizing instance selection across the full range of workload scenarios.

Some workloads, such as batch processing jobs, have periodic high utilization but use little or no CPU capacity the rest of the time (see Figure 1). This workload might require a specific burst or scaling capability at specific times to support this activity. Only analyzing average utilization stats would not provide this insight.

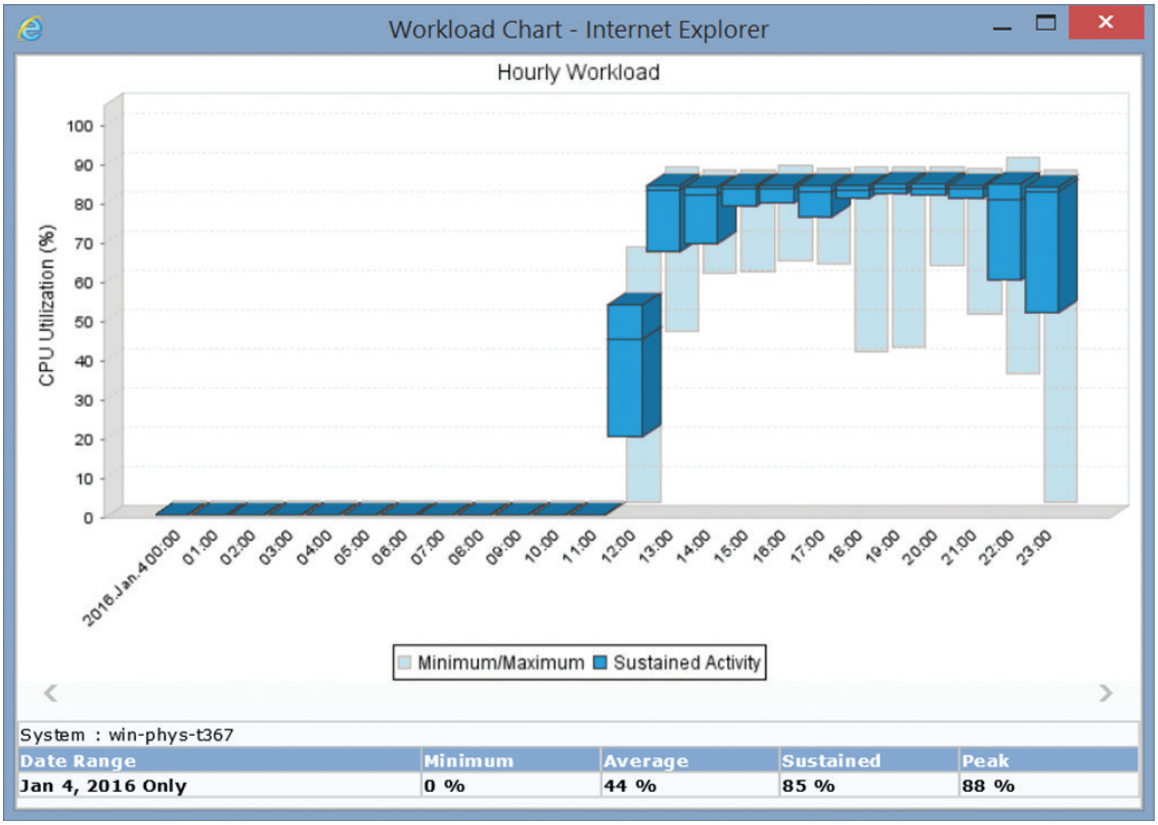

Figure 1. Typical batch processing workload.

Other workloads, such as transactional applications, can be far more complex. These may run continuously, with peaks and valleys throughout the day, creating irregular patterns (see Figure 2). Without a detailed understanding of these patterns over time, the likely outcome would be over-provisioning and over-spending.

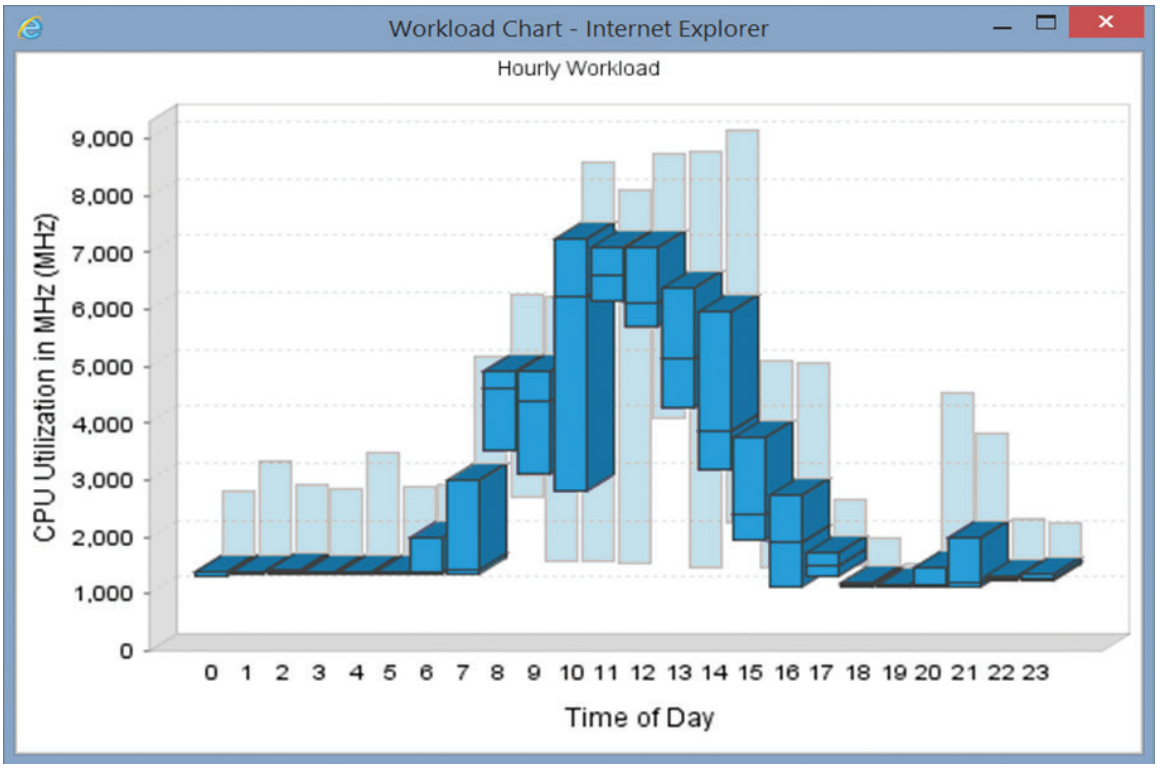

Figure 2. Complex, irregular workload pattern.

Machine learning enables predictive analysis, providing the deep understanding required to match workload dynamics with specific instance requirements. Also, a critical part of interpreting the workload pattern is normalizing the data using benchmarks, so the actual work being done can be analyzed against the service offerings of the cloud providers. This enables the analytics to not only recommend the right instance size, but also recommend the best-suited instance family/type for a specific workload. You cannot accomplish this using percentages or summary statistic measurements.

Before you reserve—Match workloads to the right instance type and size

Using automated, predictive analysis before you purchase RIs is an essential first step. By identifying the optimal instance types for each workload, this eliminates guesswork and maximizes savings by specifying:

- **Right instance size.** Determine the optimal size for each instance to meet actual application workload requirements. *Example: Move from a t2.large to a t2.medium instance.*
- **Right instance family/type.** Determine the optimal mix of CPU, memory, I/O, and networking capacity and characteristics for each instance. For example:
  - Based on demand patterns and normalized models of cloud supply, machine learning analytics recommends moving a memory-focused workload from a Compute-Optimized Instance (c4.2xlarge) to a Memory-Optimized Instance (r4.large) for optimal performance and cost-effectiveness (see Figure 3).

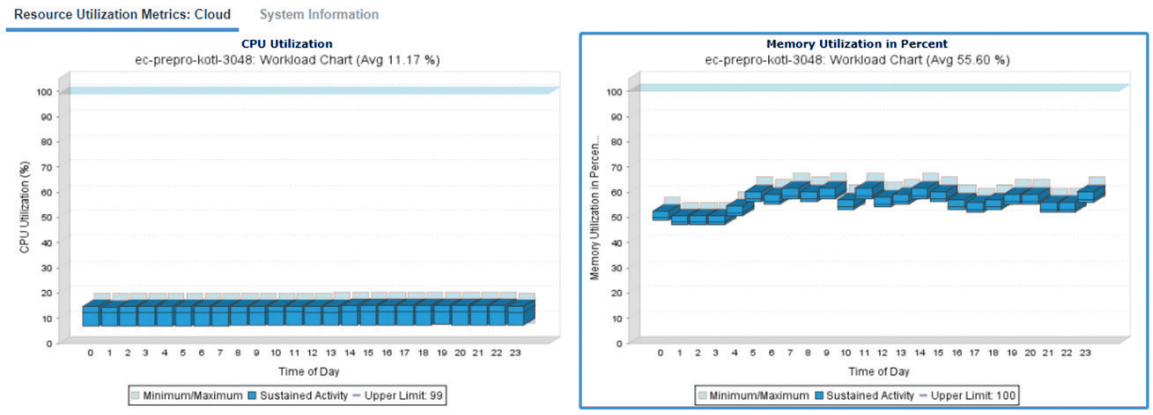

Figure 3. A memory-focused workload is suggested to move from c4.2xlarge to r4.large

- Based on demand patterns and normalized models of cloud supply, machine learning analytics recommends moving a bursty workload from a Fixed Performance Instance (m3. medium) to a Burstable Performance Instance (t2.medium) to better serve the bursty workload demand (see Figure 4).

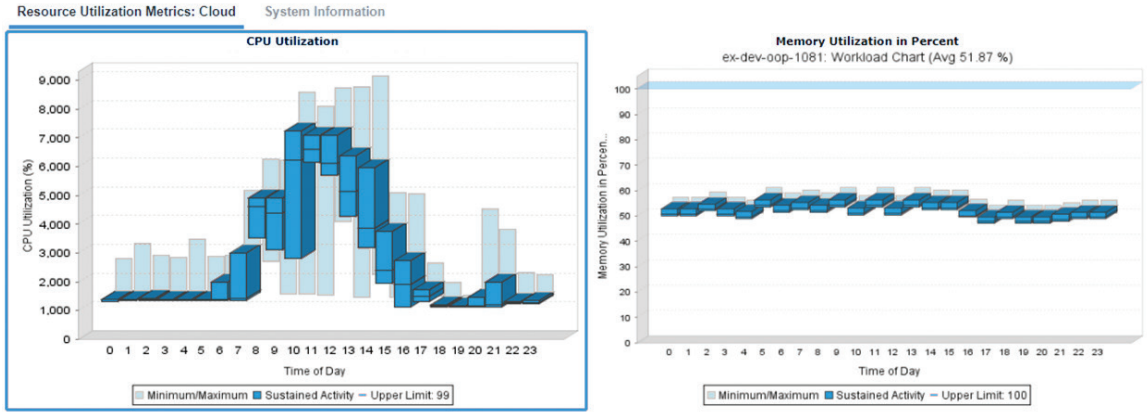

Figure 4. A workload with bursty pattern is suggested to move from m3.medium to t2.medium

- **Right scale group configuration.** Along with the optimal node family and size, the ideal scaling characteristics for Auto Scaling Groups should also be determined based on actual workload patterns. For example, upscale or downscale your Auto Scaling Groups to ensure the optimal minimum and maximum group size.
- **Right RDS instance.** Recognize the special requirements of Relational Database Service and determines the optimal RDS instances to meet your needs based on the workload patterns, the type of database, and regional variations in the sizes that are available for specific database types. *Example: Downsize from a db.m4.2xlarge to a db.m4.large.*

In fact, machine learning analytics can provide multiple layers of optimization, achieving additional optimization and savings with each layer (see Table 2).

	Optimization Level
Basic Operations	**Read the bill** and assign costs to users/line of business **Right-size instances** and identify deadwood
Advanced Optimization	**Optimize instance families** and database configurations based on normalized workload analysis
	Optimize scale groups to align with actual demand patterns
	Reserve instances based on optimized state
	Stack on bare metal with hypervisor, dedicated hosts, etc.
	Stack containers to optimize workload density and elasticity

Table 2. Machine learning analytics enables advanced optimziation, perfectly matching apps' demands with optimal cloud resources.

Advanced machine learning tools also analyze available Reserved Instance offerings, using a deep permutation analysis to match the demands of each workload with the precise cloud instance that optimally satisfies those demands. This eliminates guesswork, providing you with a clear choice of which instance is the best option to meet each workload's needs. Relying only on bill reading or basic right-sizing is likely to lead to RI decisions that lock you into the wrong things with higher spend than needed.

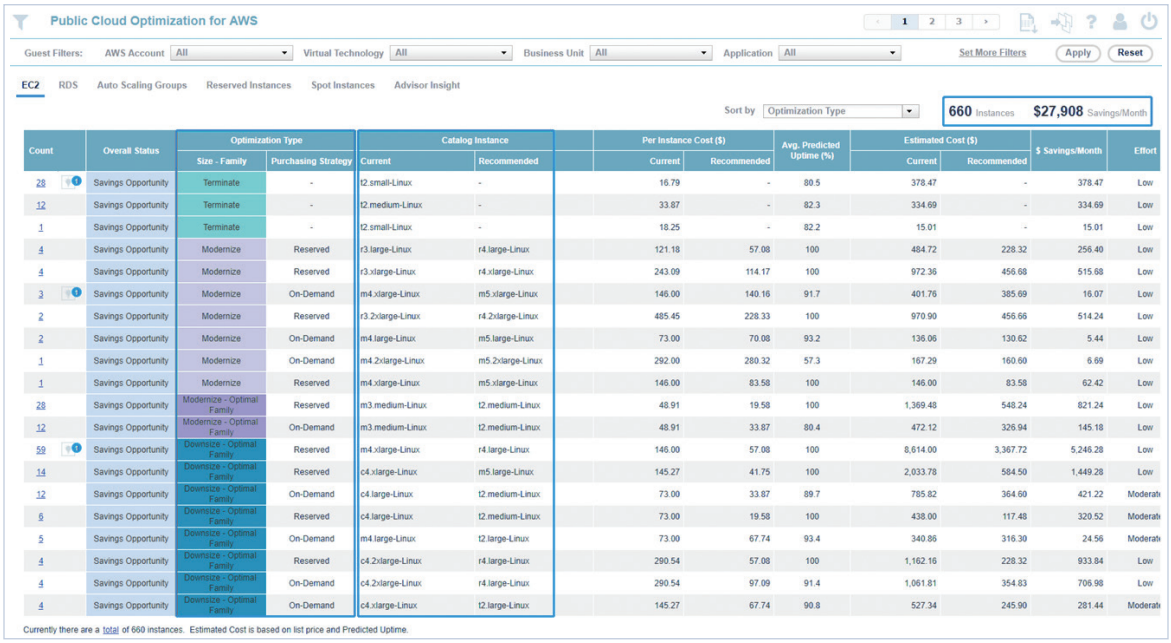

Figure 5. Example of optimization recommendations and associated cost savings.

Existing RI analysis and lifecycle planning

What if you already have workloads running in Reserved Instances? This complicates your efforts to optimize your RI spend. Yet here, too, machine learning analytics can help.

First, it can perform predictive optimization, analyzing workload patterns and dynamics, policy constraints, and instance benchmarks to match workloads to optimal instance size, family, scaling, etc. as explained previously.

Then, based on the analyses of your current set of Reserved Instances—including their resource utilization, uptime, and optimal instance size and family—analytics can determine the right action to either buy, sell or keep your existing Reserved Instances, with the corresponding RI changes required to make sure you have the optimal set of RIs. By treating the RIs as a portfolio you can globally optimize the purchasing strategy, and not just focus on isolated purchases.

This *Reserved Instance Purchasing Strategy* enables you to maximize the benefits and minimize the risks specific to RIs. For example, based on actual workload patterns—such as an application that only runs once a week to perform accounting functions—the analysis may recommend against reserving an instance for that workload, determining that an on-demand instance would be more cost-efficient.

What if you are already locked in the suboptimal set of RIs? Machine learning can analyze subscription end dates and remaining capacity for existing RIs, providing a clear implementation plan defining when to change instance types in order to achieve the optimum savings. This analysis is especially valuable when a cloud portfolio contains RIs that are convertible or can be resold, determining whether it makes sense to exercise those options in the context of applied fees and, if so, when. It is a lot like a very sophisticated game of musical chairs, where the workloads that are consuming the RIs dynamically change as the analytics optimizes the RI composition.

Taking the guesswork out of managing RIs

Densify has harnessed the power of machine learning to transform the process of selecting RIs to maximize efficiency and savings, based on an in-depth analysis of application workload patterns. We call our analysis engine **Cloe** (Cloud-Learning Optimization Engine). A key advantage of Cloe is its high degree of automation. There is no complicated software to install, configure and learn.

While algorithms are essential for performing cloud optimization, so is human expertise. We have found that having an experienced cloud optimization advisor working with you is a critical factor for long-term success. Having such an expert work with you to understand your environment and your cloud computing needs, and then manage the process—from data collection and tuning policies to running the analyses and reviewing the results with you—is tremendously valuable in helping assess specific RI pricing options and the opportunity cost of reserving instances.

Helping you make the best RI decisions

Reserved Instances can provide substantial levels of efficiency, application uptime, and cloud cost savings, if they are used intelligently. Densify's multi-layered approach to cloud optimization, powered by Cloe, can lead to significant positive results.

By automating the process of optimization, Densify also enables significant time savings. Cloe eliminates tedious, time-consuming and imprecise manual methods of analyzing cloud utilization. And by making optimization recommendations using sophisticated, predictive analytics of workload patterns over time, Densify dramatically reduces the risk of under-provisioning services, safeguarding application performance and availability.

Densify's typical user has a return on investment of about three months. This means, pure financial savings gained from RI savings pays for the Densify service after only 3 months, and beyond that you could gain pure RI savings in terms of dollars back. This does not account the benefits of time saved, value of analytical automation and the improvements in application performance which are all gained in the first 48 hours.

Densify offers its machine learning service free for select size organizations to try at: ***www.densify.com/try***.

Densify

The Definitive Guide to Optimizing Reserved Instances

www.densify.com