Technology Reports (Special Articles) Billing Management Cost AWS
Special Articles on Use of Public Clouds

Cloud Cost Optimization Measures

Innovation Management Department Tetsuo Sumiya

Today, public cloud services are extensively used by businesses in Japan and other countries. However, major public cloud systems are billed on a pay-as-you-go basis and if they are deficient in their ability to manage resources such as virtual machines, they can cause unexpected costs by activating resources unnecessarily. It is therefore necessary to consider resource management and cost optimization measures from the initial design stage. This article describes the know-how and cloud cost optimization measures that NTT DOCOMO has cultivated to date.

1. Introduction

In recent years, more and more companies have been using public clouds to provide their services. Unlike the structure of conventional systems in data centers, public cloud services make it easy for users to set up virtual machines*1 and other resources from a management console with just a few clicks. This greatly speeds up system construction and contributes to the enhancement of corporate competitiveness. However, because of their ease of use, many enterprises still struggle to manage and control the costs of public clouds. Major public cloud services are fundamentally billed on a pay-as-you-go basis, and if they are deficient with regard to the management of resources such as virtual machines, they can cause unexpected costs by activating resources unnecessarily. It is therefore necessary to consider resource management and cost optimization measures from the initial design stage.

To optimize costs, it is important to first visualize them to make the effects of cost management measures visible, and to perform repeated cost

*1 Virtual Machine: A computer (e.g., a server) that is implemented virtually in software.

reduction measures and verify their effectiveness. Continuous checks are performed on the usage of resources with visualized costs. Depending on their usage, it is necessary to study the use of payment plans and review the capabilities of the virtual machines being used, such as Elastic Compute Cloud (EC2)*2 instance*3 types in Amazon Web Services (AWS)*4 (**Figure 1**).

This article describes the key concepts of cloud cost optimization initiatives and discusses our specific know-how.

2. Cost Optimization

Cost optimization is considered with the following priorities (**Figure 2**).

STEP 1: Review the constructed architecture

The most effective way to reduce costs is to review the entire architecture and, where possible, to consider using managed services that are available off the shelf from cloud providers rather than building them yourself. Since it is difficult to alter systems during live operation, it is important to design

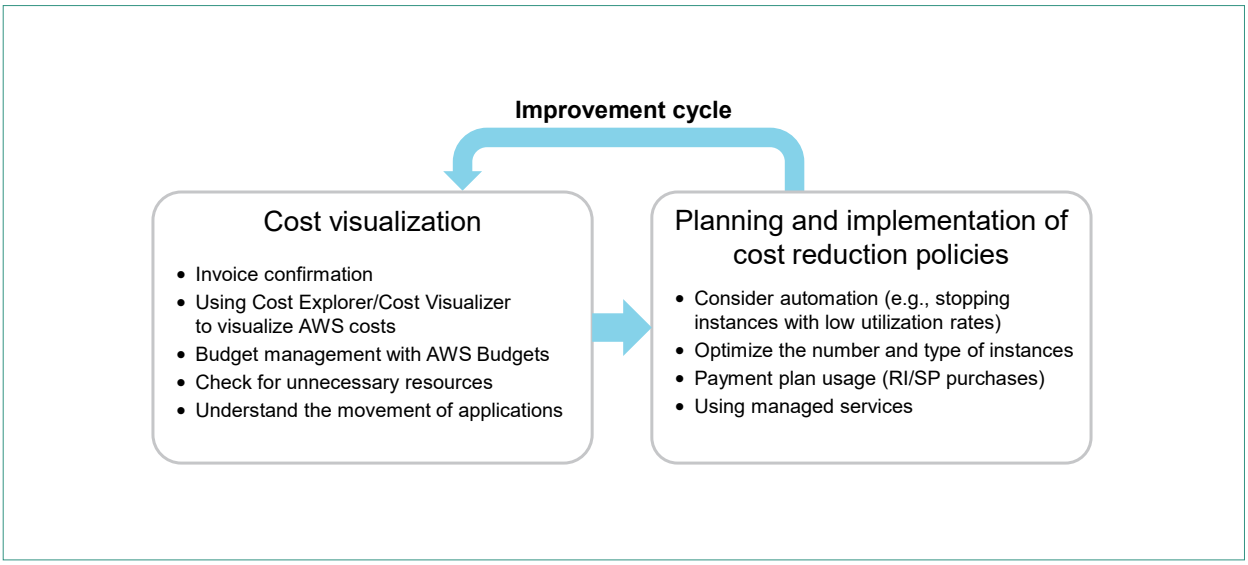

Figure 1 Cost optimization improvement cycle

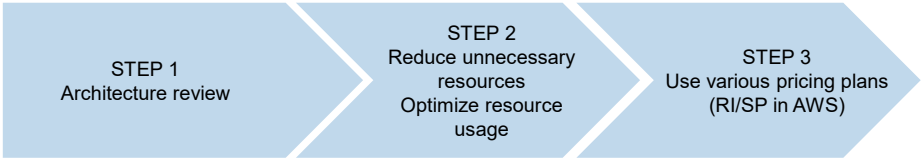

Figure 2 Cost optimization study process

*2 EC2: An IaaS offering provided by AWS. Provides services on virtual machines.
*3 EC2 instance: A virtual machine provided by AWS.
*4 AWS: A cloud computing service provided by Amazon Web Services, Inc.

systems with a firm awareness of costs during the initial design and system renewal stages.

STEP 2: Delete resources that are unnecessary and optimize the use of those that are necessary

When accounts are active for a long time, they sometimes accumulate unused resources, and can end up using excessive resources that were not originally needed. It is possible to reduce costs by incorporating regular resource reviews into operations to ensure that nothing is wasted.

STEP 3: Use various pricing plans

Major public cloud providers offer pricing plans whereby users can reduce their usage fees by committing in advance to a minimum level of resource usage for a specified period. For example, AWS offers pricing plans called Reserved Instance (RI) and Saving Plans (SP) that allow users to lower their fees by committing to one or three years of usage. Google Cloud Platform*5 offers a fixed usage discount, and Microsoft Azure*6 offers a similar pricing plan in the form of Reserved Virtual Machine Instances. If the service is inevitably going to be needed for a certain period of time on an ongoing basis after optimizing resources, this type of pricing plan can be used to reduce costs. However, when committing to use resources, it is difficult to implement major changes to the system configuration and review the resource usage within the period of this commitment, so the usage should be considered after conducting the studies of steps 1 and 2.

3. Cost Visualization

In consideration of the above points, it is essential to have a continuous grasp of the cost structure based on a visualization of the current costs so as to ascertain which parts of the system are incurring costs.

Major public cloud providers offer cost visualization tools. In this section, we will describe the Cost Explorer tool provided by AWS. We will also describe the Cost Visualizer tool, which we developed before the launch of Cost Explorer, and which is used throughout NTT DOCOMO.

3.1 Understanding Usage with Cost Explorer

Cost Explorer is a standard AWS tool that allows users to view a breakdown of their billing status in graphical form (**Figure 3**). This makes it possible to subdivide costs in various ways, such as by service and by member account*7. By default, it can output multiple reports on aspects such as RI/SP utilization and coverage.

As a precaution, Cost Explorer should be used by creating and accessing an Identity and Access Management (IAM) user*8 with only the minimum necessary privileges (e.g., only the ability to view costs).

3.2 Using Cost Visualizer to Ascertain the Usage Status

Cost Visualizer is a cost analysis tool developed and provided by NTT DOCOMO. Since the aforementioned Cost Explorer was not available when NTT DOCOMO began using AWS on a large scale in 2012, we developed Cost Visualizer in-house

*5 Google Cloud Platform: A cloud computing service provided by Google LLC.
*6 Microsoft Azure: A cloud computing service provided by Microsoft Corporation.
*7 Member account: An account that is not a manager account and belongs to an organization that consolidates multiple AWS accounts.
*8 IAM user: A user created with the IAM service who is permitted to access the AWS environment.

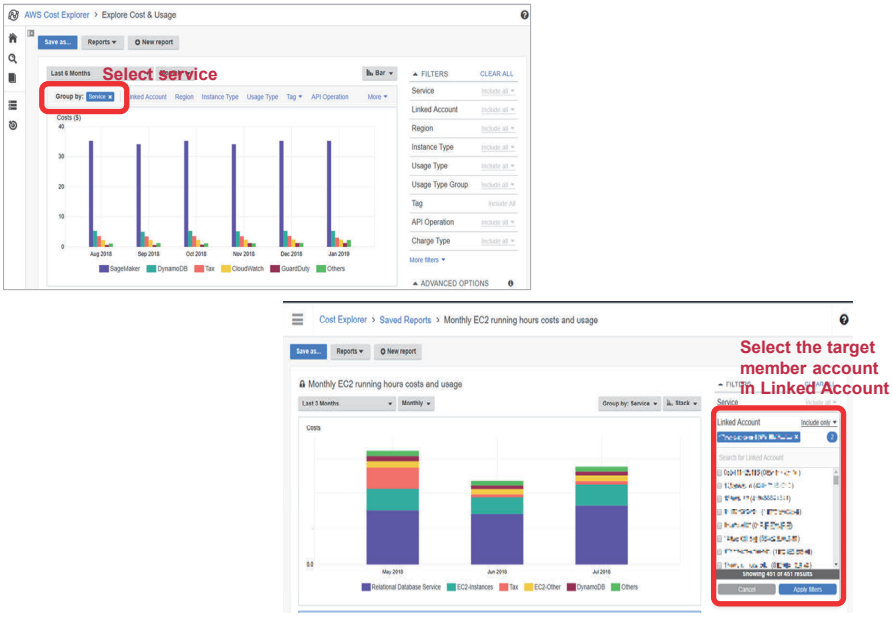

Figure 3 Cost Explorer screenshot

due to the need for cost management.

Cost Visualizer can be used regardless of the AWS support level or billing information access privileges because its privileges are set separately from AWS and the account management is performed separately. It also supports features that are not supported by Cost Explorer, such as the ability to manage privileges in greater detail, and functions for displaying pie charts and data groupings (**Figure 4**).

The system architecture of Cost Visualizer is shown in **Figure 5**. The Cost and Usage Report (CUR) provided by AWS is automatically stored in Amazon Simple Storage Service (Amazon S3)*9. This data is extracted and transformed by AWS Glue*10, which is an ETL (Extract, Transform, Load) service that loads data into a database and makes it available for use, and is then loaded into a database on the virtual machine running Cost Visualizer. We opted to set up a database on a virtual machine instead of using the Relational Database Service managed by AWS because it is internally designed to continuously process queries*11 on large quantities of data so as to minimize the delays until graphs are drawn. Outside the virtual machine, we use a configuration that combines AWS Glue with managed services such as AWS Lambda*12 (a serverless computing platform) and Amazon Dynamo*13 (a key-value store*14 service) in order to reduce costs as much as possible.

*9 Amazon S3: A storage service provided by AWS.
*10 AWS Glue: A PaaS offering provided by AWS. Capable of performing processing for data classification and manipulation.
*11 Query: A database query (processing request).
*12 AWS Lambda: A type of FaaS provided by AWS that provides an execution environment for application code so that the user need only register created source code to run the application.
*13 Amazon Dynamo: A PaaS offering provided by AWS. A highly reliable, high-performance non-relational database service.
*14 Key-value store: A data store with a simple structure that combines keys and values.

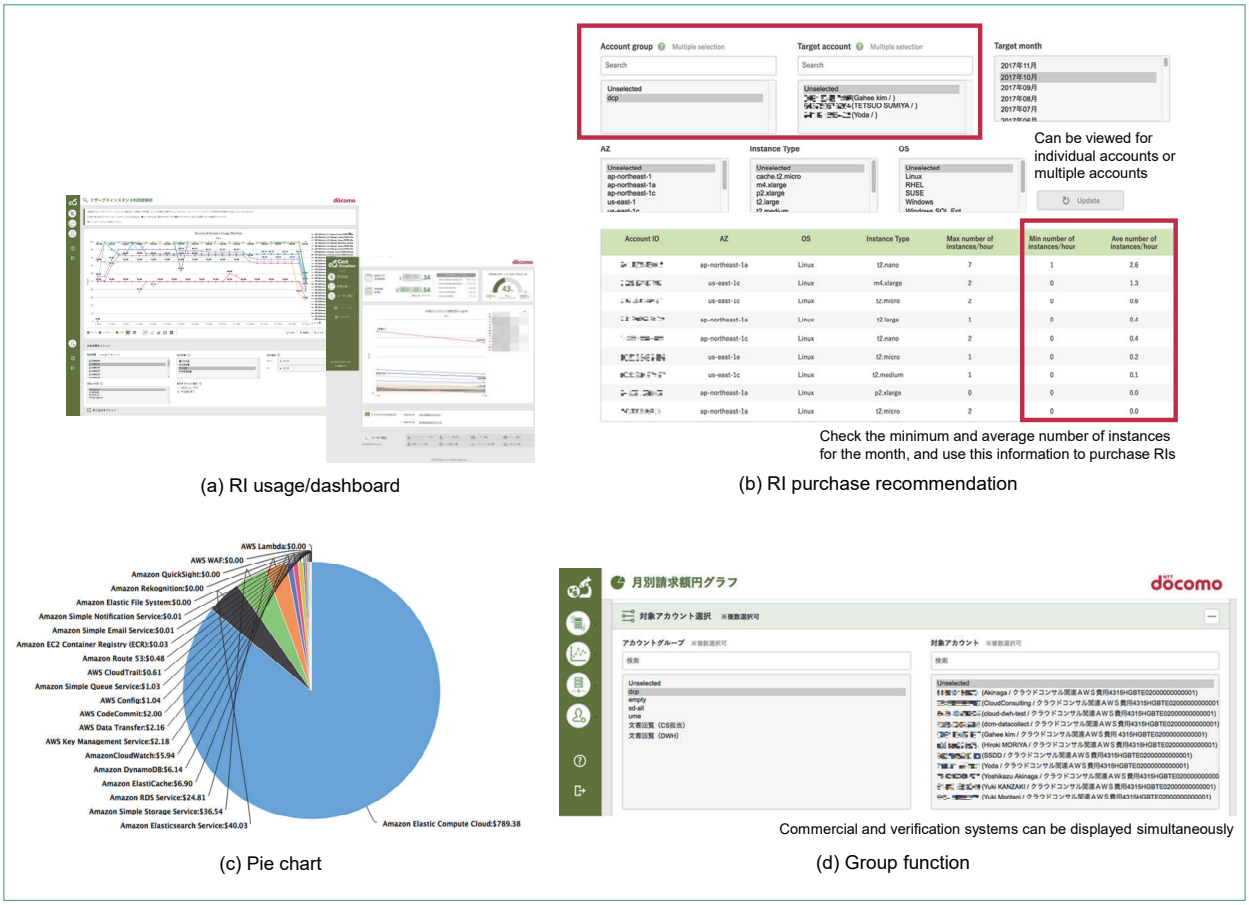

Figure 4 Cost Visualizer screenshot

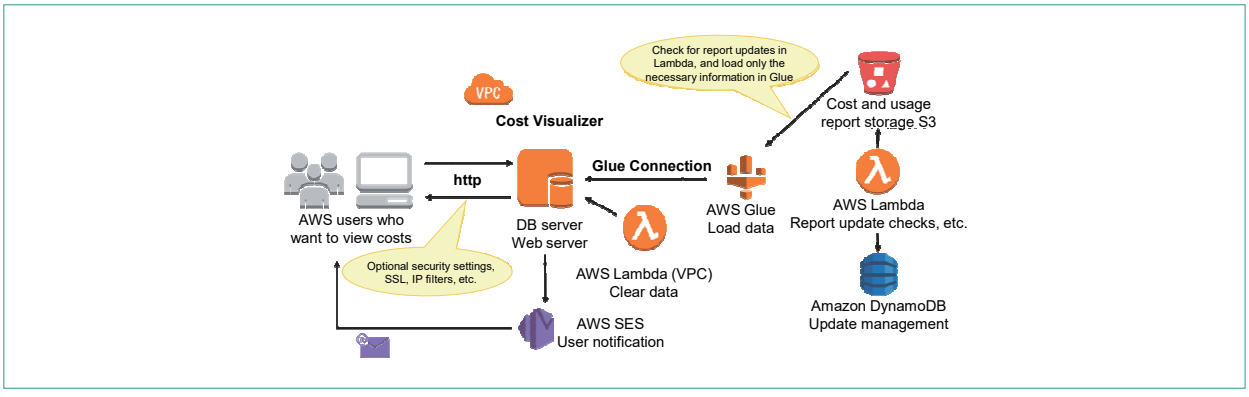

Figure 5 Cost Visualizer system architecture

An example of cost visualization using Cost Visualizer is shown in **Figure 6**. For an architecture centered on virtual machines that does not use managed services, the EC2 costs will dominate in this

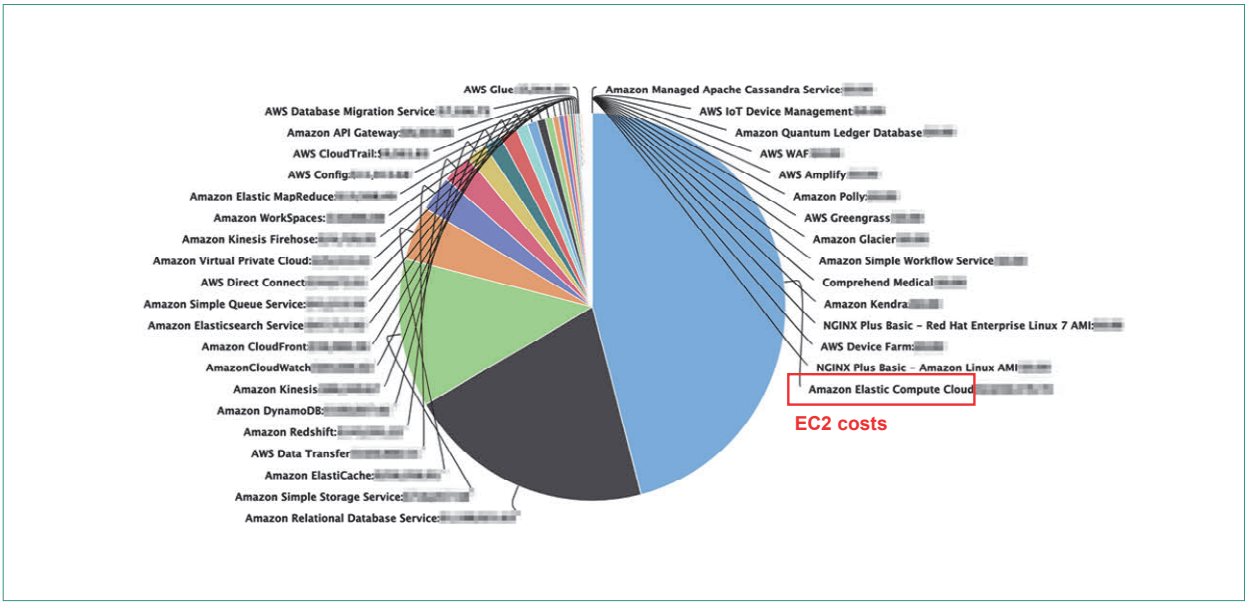

Figure 6 Example of cost visualization by Cost Visualizer

way, indicating that measures to reduce the EC2 costs will be necessary.

3.3 Budget Management Service

Major public cloud providers have budget management services that can send alerts by email or other means when costs or usage have exceeded set limits, or are likely to do so. AWS Budgets allow users to set limits for not only the cost but also for parameters including the quantity of AWS resources used and the RI usage rate. When used together with cost visualization tools, these can enhance the user's everyday cost awareness.

4. Planning and Implementation of Cost Reduction Policies

4.1 Use of Managed Services

Major public cloud providers offer managed services tailored to the characteristics of the processing to be performed, as well as computing resources such as virtual machines. In many cases, these managed services are billed according to the time spent using resources and performing processing, which can cost less than merely provisioning*15 the virtual machines necessary for building such a system.

With AWS, for example, it should be possible to achieve significant cost savings by taking the following steps.

- For services that run for a long time and have few requests, use Lambda instead of EC2.
- When it is necessary to run batch processing*16, consider using AWS Batch*17 instead of EC2. In AWS Batch, computing resources are dynamically scaled*18 according to the volume of batch jobs and the resources they require, thereby reducing costs.
- Switch to a serverless architecture using tools

*15 Provisioning: The process of securing and configuring resources such as servers and networks to run applications.
*16 Batch processing: A processing method where fixed quantities of data are collected and processed all together at fixed intervals.
*17 AWS Batch: A PaaS offering provided by AWS. This service facilitates simple and efficient large-scale batch processing.
*18 Scaling: The optimization of processing power by increasing or decreasing virtual machines that configure communications software whenever processing power is insufficient or excessive according to hardware and virtual machine load conditions.

such as Cognito, API Gateway, Lambda and DynamoDB, as shown in **Figure 7**.

4.2 Identify Unnecessary Resources

Once provisioned, cloud resources incur costs even when they are not actually being used. It is therefore necessary to periodically check whether unnecessary resources are being retained. For example, these resources might include Amazon Elastic Block Store (EBS) volumes*19 that are not associated with a running EC2 instance, or old EBS snapshots*20 that are not even tagged.

Although these unnecessary resources can be checked from the console of the cloud service, it can be difficult to do so when dealing with a large number of resources. An efficient way of checking these resources is to use AWS Trusted Advisor. In Trusted Advisor, the items listed in **Table 1** can be checked.

As an example of actual cost reduction, in one project we used Trusted Advisor to check the resource status. As a result, we found that 22 out of 42 block storage devices were not attached. By removing these devices, we were able to reduce the computing cost by about 10%. We also achieved cost savings by deleting over 1,000 untagged snapshots that we found in other projects.

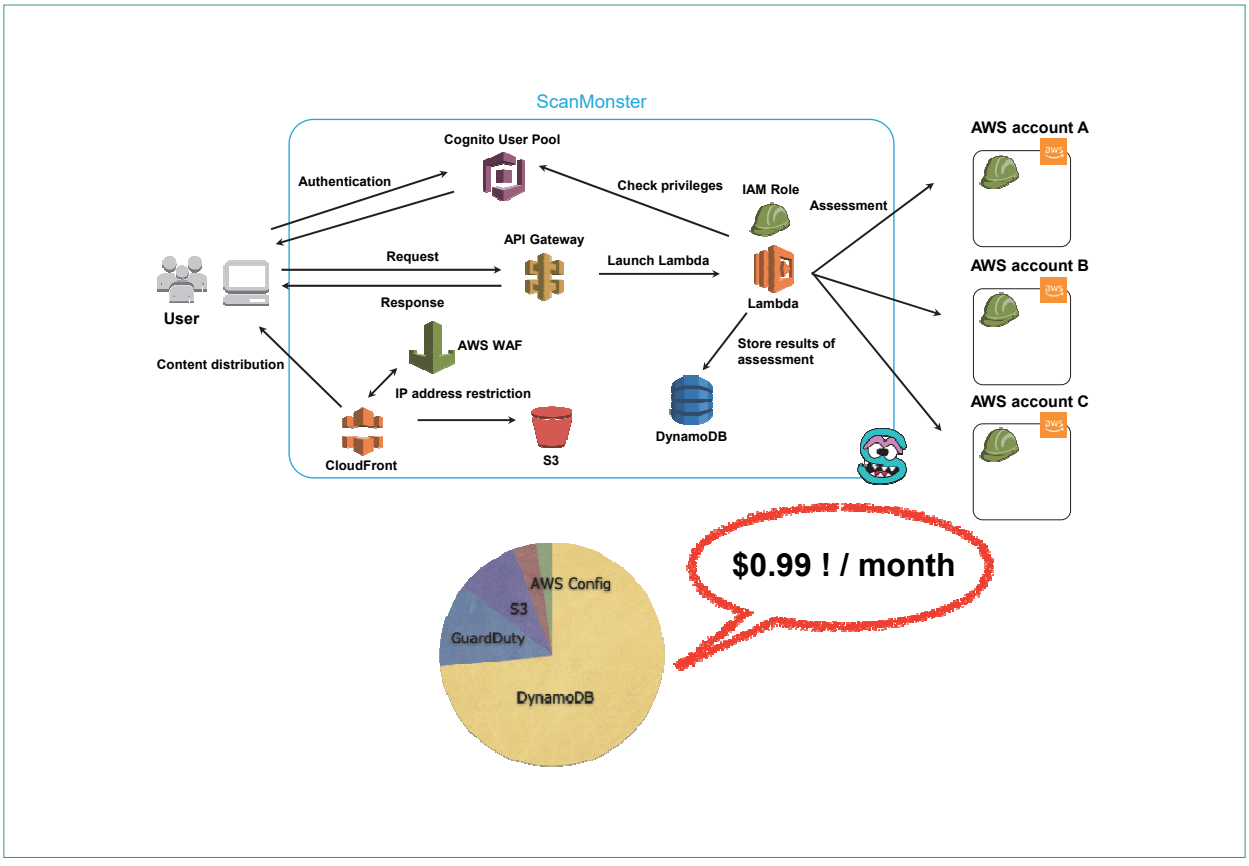

Figure 7 Example of a serverless architecture (ScanMonster)

*19 Amazon EBS Volume: A high-performance, highly available block storage service provided by AWS. Block storage refers to storage in which the recording area is managed by dividing it into units called volumes, and the interior of each volume is further divided into fixed-length units called blocks.

*20 EBS snapshot: Backup data for an Amazon EBS volume.

Table 1 Cost optimization points that can be checked in AWS Trusted Advisor

Item	Overview
Underutilized EC2 instances	Determine usage status based on CPU utilization and network I/O traffic
Idling load balancer	Determine usage status based on the number of requests to the load balancer and the number of associated EC2 instances
Idling RDS DB instance	Determine usage status based on the frequency of connections to RDS DB instances
Infrequently used Amazon EBS volumes	Determine usage status based on whether an EBS volume is not attached to an EC2 instance, or on the frequency of writes
Route 53 latency resource record set	Identify inefficiently configured latency record sets
Underutilized Redshift cluster	Determine usage status based on frequency of cluster connections to Redshift and on CPU usage
Unassociated Elastic IP Address	Check Elastic IP Addresses that are not associated with a running EC2 instance
RI expiration	Check RIs that have expired or will expire in the previous or following 30 days
Amazon EC2 RI optimization	View the recommended number of EC2 RI purchases
Amazon RedShift reserved node optimization	View the recommended number of Red Shift RI purchases
Amazon RDS RI optimization	View the recommended number of RDS RI purchases
SP recommendation	View the recommended number of SP purchases
Amazon ElastiCache reserved node optimization	View the recommended number of ElastiCache RI purchases
Amazon Elasticsearch RI optimization	View the recommended number of Elasticsearch RI purchases

Amazon ElastiCache: A fully managed in-memory data store service provided by AWS.
Amazon Elasticsearch: A managed service provided by AWS based on the Elasticsearch open-source search engine.
Elastic IP Address: A fixed IP address service provided by AWS.
Redshift cluster: A cluster of data warehousing services provided by AWS.
Route 53 latency resource record set: A combination of assets such as domains and EC2 instances that can be registered in Route 53 (a domain name service provided by AWS) to minimize latency from end users.
Load balancer: A device that equalizes the allocation of loads on a server. AWS provides a load balancer as a service.

4.3 Understanding the Movement of Applications

Sometimes, when an application is deployed[*21], it does not generate as much traffic as initially expected and ends up being over-provisioned. It is difficult to shrink resource allocations in on-premises[*22] systems, but in the cloud it is possible to shrink or expand resources as appropriate. Resource usage can be checked using services such as Cloudwatch[*23] for AWS, Cloud Monitoring[*24] for Google Cloud Platform, and Azure Monitor[*25] for Microsoft Azure. In addition to the services provided by cloud providers, there are also monitoring services available from companies such as New Relic and Datadog, and these services can be used to make appropriate changes to resources. AWS has a func-

*21 Deploy: Installing applications by placing them in their execution environments.
*22 On-premises: An environment in which a company owns, maintains, and operates the hardware making up its system.
*23 Cloudwatch: A monitoring service provided by AWS for AWS resources and applications running on AWS.
*24 Cloud Monitoring: A service provided by Google Cloud Platform that monitors Google Cloud Platform resources and the applications running on them.
*25 Azure Monitor: A monitoring service provided by Microsoft Azure for monitoring Azure resources and the applications running on them.

tion called the Compute Optimizer, which can identify idle instances and underutilized instances, and can recommend ways to reduce costs (**Figure 8**).

Also, the latest instance types always tend to be cheaper, so the possibility of switching to the latest instance type should always be kept in mind.

4.4 Considering Automation

In verification environments that do not need to be kept running constantly, costs can be considerably reduced by shutting down overnight and during holidays. For example, by stopping a system for five hours every night on weekdays and altogether at weekends, its running cost can be reduced to 60% or less. If a system can be stopped this much, then it is likely to cost less than the discounted price for a system encumbered with a one-year usage commitment. It is difficult to manually stop a system every day when handling many resources, but this process can be automated to ensure that the system is stopped without fail. Costs can also be reduced by setting up the regular execution of backup scripts and generation management tools to automatically delete old versions.

4.5 Considering Fee Models

After implementing the cost reduction initiatives described above, further cost reductions can be achieved by using fee plans for resources that are absolutely necessary. AWS includes payment plans called RI and SP for computing resources. The RI payment plan makes it possible to reduce fees by committing for a fixed period to a system with specific attributes such as the OS, per-region[*26]/per-Availability Zone (AZ)[*27] deployment, or instance family[*28]. In contrast, SP relaxes these specifications (OS type, per-region/per-AZ, instance family, etc.) and commits the user purely in terms of the

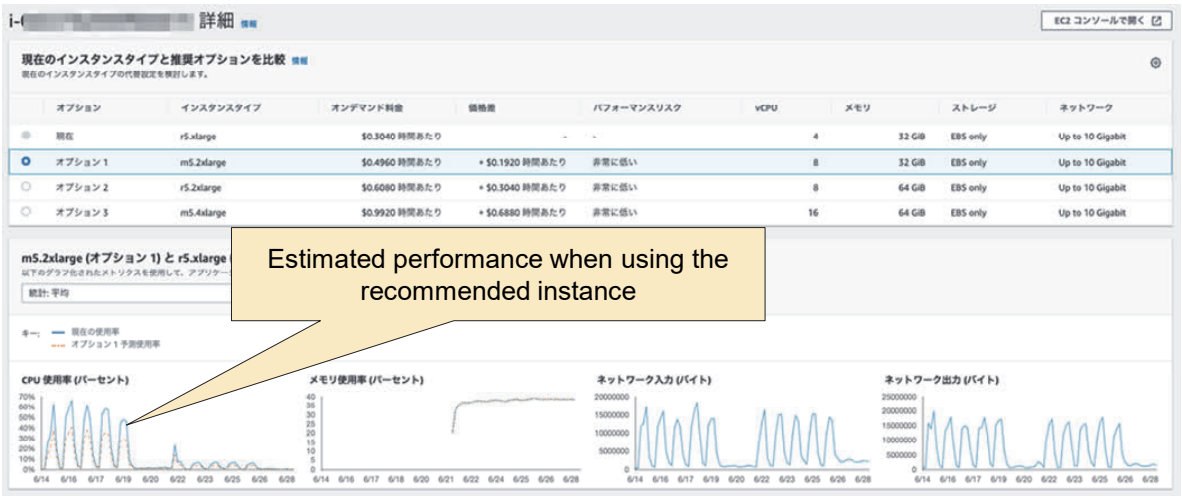

Figure 8 Screenshot of AWS Compute Optimizer

*26 Region: The region in which the data centers providing cloud services are located.
*27 AZ: A collective unit of data centers that are autonomous both physically and in software terms.
*28 Instance family: Instance types are classified by usage, such as "general-purpose," "computing-optimized," and "memory-optimized."

usage fee. As a result, the discount rate is lower for flexible purchases. However, since new instance families are announced from time to time, commitments made based on SP allow for operation with greater flexibility.

5. Conclusion

This article has described the key points of cloud cost optimization and our specific know-how in this field. Since cloud computing systems are billed on a pay-as-you-go basis, proper management of resources and their usage rates is important in terms of cost optimization, while repeated visualization and effectiveness verification is important for management. To optimize costs, it is important to keep cost effectiveness in mind throughout the design process from the initial system configuration, and even during the operational phase. It is also necessary to continue performing periodic checks of the usage status, and to consider the potential for making configuration changes, reviewing instance types, and using different fee plans as dictated by circumstances. In the future, we will consider reducing cloud costs for NTT DOCOMO as a whole through measures such as purchasing SPs with a representative account according to the RI and SP application rates.