Intent-based Service Management to Improve Resource Design Efficiency for Cloud Services

Chao Wu, Shingo Horiuchi, and Kenichi Tayama

Abstract

As cloud services have expanded, an increasing number of service providers are implementing various kinds of services and functions such as web services and machine learning in the cloud environment. To provide the cloud service promptly and improve customer satisfaction, a cloud service provider needs to efficiently design the resources in accordance with the service requirements. This article presents an intent-based service management framework that enables the needed cloud resources to be derived in accordance with the service provider's service requirements, cloud environmental conditions, and operation policies.

Keywords: intent-based service management, cloud, virtual machine

1. Current practices of cloud resource design

The cloud has become a popular choice for service providers (SPs) to allocate new service workloads or migrate existing ones. The SP, when requesting new cloud services or asking to scale existing ones, is concerned about the service requirements such as the functionality of the service, the levels of security and availability, and the ability to handle workloads. In contrast, the cloud service provider (CSP) needs to know the composition of resources and the amount of resources to be allocated to fulfill the service requirements from the SP (**Fig. 1**). Therefore, the CSP needs to analyze the service requirements from the SP, and on the basis of the results needs to design cloud resources. Current practices of designing the cloud resources include:

(1) Cloud-consultant approach

The SP is assisted by cloud consultants from the CSP who collect the SP's service requirements and determine resource details accordingly. This approach results in a high operating expenditure for the CSP and takes them a relatively long time to deliver services.

(2) Self-service approach

The SP is provided with a management interface to manage cloud resources and service needs in order to determine their resource requirements. This approach requires the SP to have IT (information technology) expertise and may be a barrier to SPs wishing to enter the market.

In both approaches, the transfer from service requirements to resource requirements relies heavily on an individual human's decision-making process.

In response to these issues, to more efficiently design and operate cloud resources, we have been researching an intent-based service management (IBSM) framework that analyzes the service requirements expressed through various channels (e.g., natural language (i.e., a *human* language such as Japanese or English, rather than a computer command language), graphical user interface (GUI), etc.) and determines the composition and amount of resources accordingly. Meanwhile, the output of IBSM can be used as the input of a cloud resources orchestrator, for example, OpenStack Heat, thus enabling the automation of the process from receiving a service request to service delivery and operation.

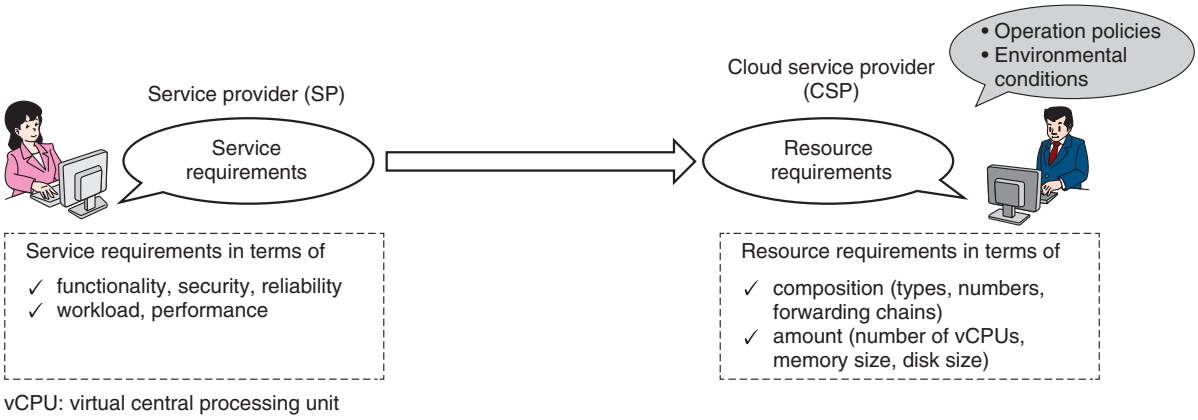

Fig. 1. Cloud resources are designed in accordance with service requirements.

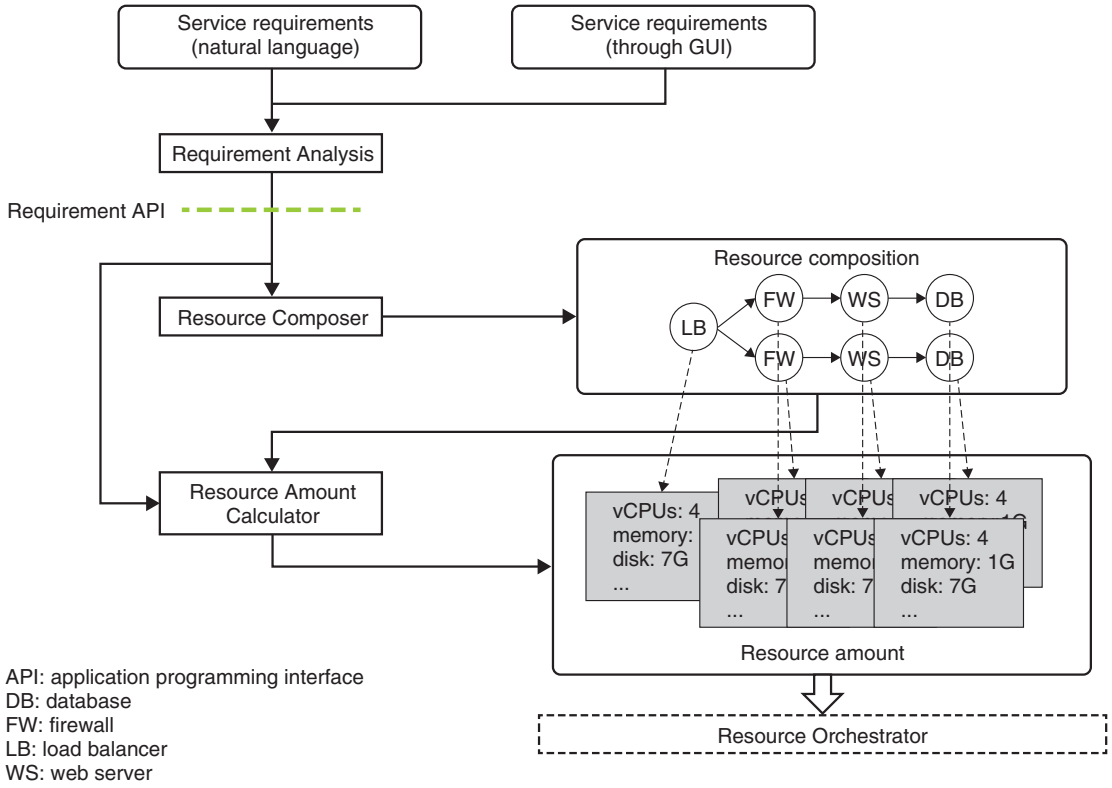

Fig. 2. IBSM framework.

2. IBSM framework

IBSM [1] consists of three main function blocks: Requirement Analysis, Resource Composer, and Resource Amount Calculator (**Fig. 2**). Below, we introduce each function and the approaches [2] to achieve them.

2.1 Requirement Analysis

SPs describe the service requirements through various channels such as natural language and GUIs. The Requirement Analysis block is responsible for

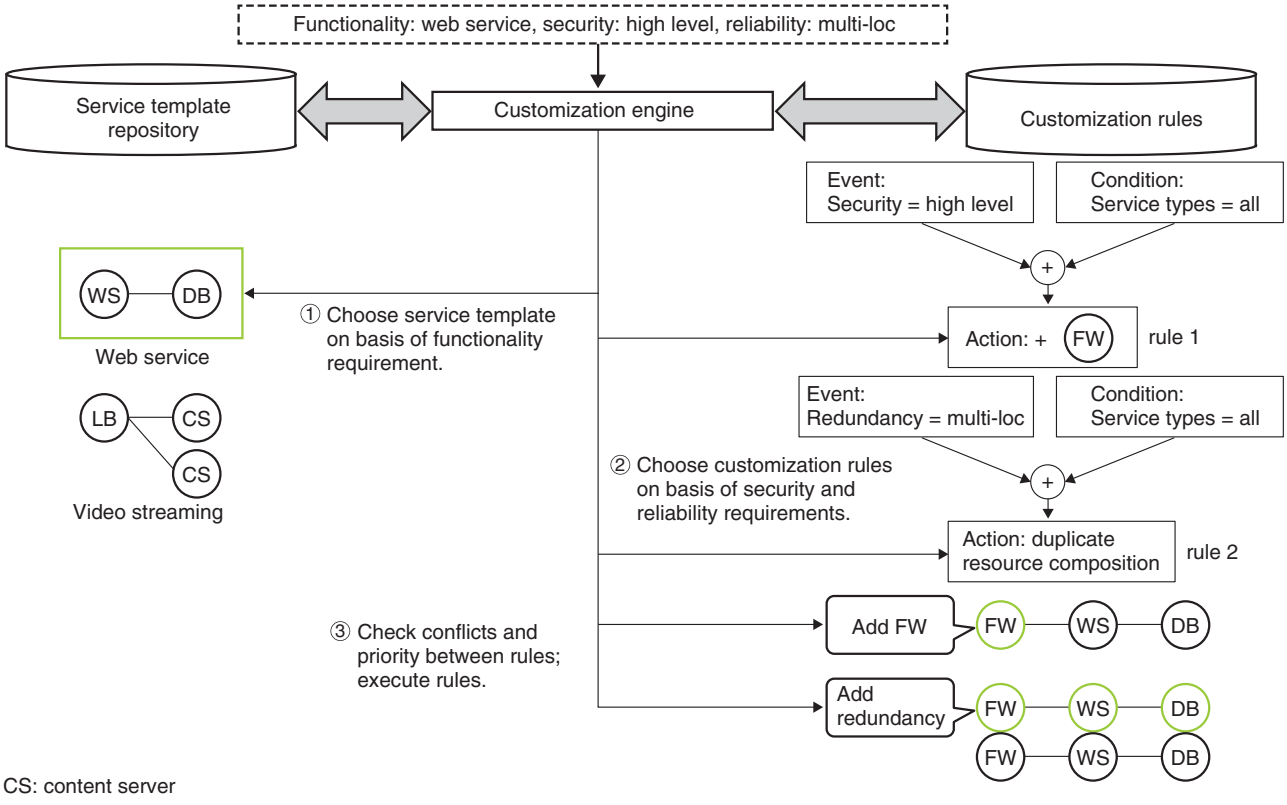

Fig. 3. Resource Composer.

parsing these service requirements and categorizing the requirements into atom requirements including the requirements of functionality, security, reliability, workload, and performance.

2.2 Resource Composer

The Resource Composer block takes the output of the Requirement Analysis block and enables the resource composition to be automatically determined in accordance with the service requirements, especially the requirements of functionality, security, and reliability. There is an existing approach to determine the resource composition in accordance with predefined service templates, but this approach leads to the problem of a dramatic increase in the number of service templates as the service variations increase. To solve this problem, as shown in **Fig. 3**, a Resource Composer is based on a small number of basic service templates that can be customized on the basis of service requirements.

2.3 Resource Amount Calculator

The resource Amount Calculator's roles and functionalities include:

- Upon delivery of a cloud service, it determines the amount of resources needed to satisfy the performance requirements. Besides the workload and performance requirements, environmental conditions and operation policies need to be considered to determine the amount of computation resources allocated to virtual machines (VMs).
- After the delivery of the cloud service, if there are changes in workload, environmental conditions, and operation policies, it adjusts the amount of computation resources allocated to VMs to ensure continuous satisfaction of the performance requirements.

In the following, we explain why the workload, performance requirements, environmental conditions, and operation policies need to be taken into consideration to determine the amount of resources.

(1) Workload and performance requirements

The SP processes the service workload in the cloud environment, and in most cases, requires the performance requirements to be met. In this article, *workload*

Table 1. Examples of workload and performance requirements.

	Workload			Performance requirements	
	Type	Features	Amount	Processing time restriction	Processing percentage restriction
(a)	Web server requests (get)	Web page size: 30 KB, etc.	10,000 requests per second	Keep average process time of requests under 1 second	Successfully process over 95% of requests
(b)	Neural network (training)	Layers, nodes of each layer, activation function, etc.	Training set: 32 x 32 pixels 256 color x 10,000	Train 1 epoch in less than 10 min	...

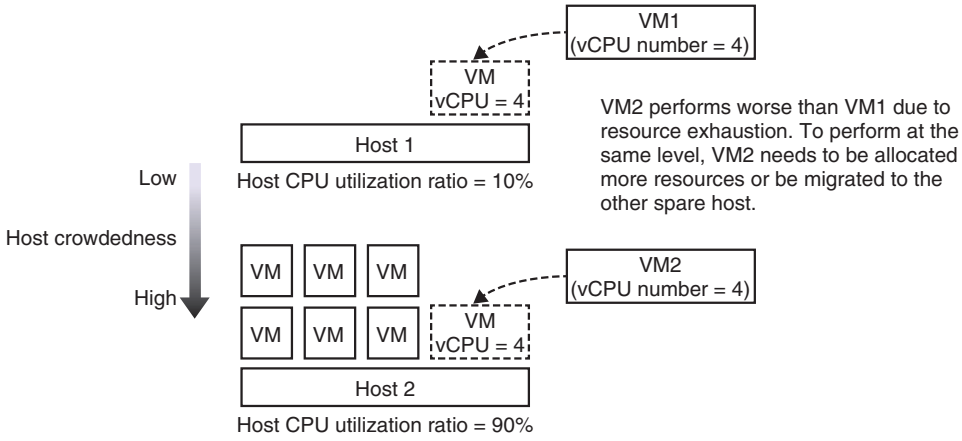

Fig. 4. Example of how environmental conditions affect VM's performance.

refers to the type, features, and amount of processing. The performance requirements can be divided into two main categories: the processing time restriction and the processing percentage restriction. Two examples of workload and the corresponding performance requirements are given in **Table 1**. For a certain workload, the computation resources allocated to the VMs directly affect the processing time and percentage achieved. Thus, workload and performance requirements must be considered when deciding the amount of resources allocated to VMs.

(2) Environmental conditions

Environmental conditions in this work are the conditions of the physical host to which the VM is allocated. Static environmental conditions include the central processing unit (CPU) clocks and memory architecture of the host; dynamic conditions include the resource utilization ratio of the host, also referred to as host crowdedness in this work. Given the consideration that the static environmental conditions are of relatively low variation and are not subject to change for a relatively long time span for a given CSP, we focus on dynamic environmental conditions in this work. An example of changes in environmental conditions affecting the performance of VMs is shown in **Fig. 4**. To satisfy the performance requirements, environmental conditions clearly need to be considered when determining the resource amount.

(3) Operation policies

Operation policies that need to be followed while providing a cloud service can be decided by the SP or CSP. The policies restrict the resource usage within a desired range. For instance, putting restrictions on the resource utilization ratio inside the VM, for example, 50–80%, prevents resources from being overused or underused, thus improving the resource efficiency and enhancing the satisfaction of service requirements. Furthermore, for a service composed of multiple VMs, setting the utilization ratio restrictions in the same range prevents bottlenecks in the service chain from occurring. The amount of resources allocated to VMs is a crucial factor in meeting the

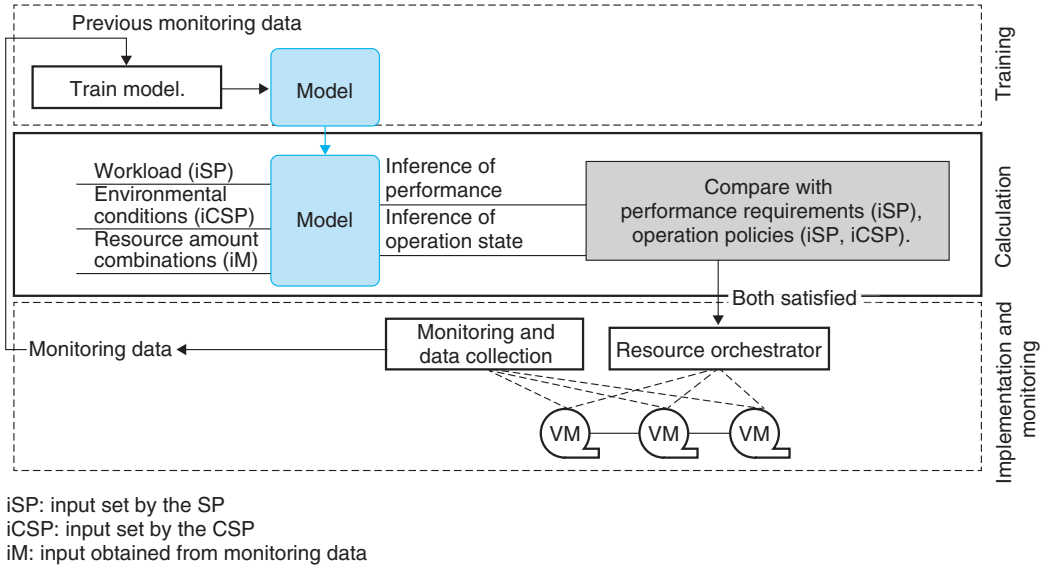

Fig. 5. Resource Amount Calculator.

restrictions of the utilization ratio inside the VM.

The structure of the Resource Amount Calculator is shown in **Fig. 5**. The Resource Amount Calculator uses a model between the workload, environmental conditions, and resource amount and the performance and resource utilization ratio inside the VM, as shown in the figure. The model is trained based on the log data collected during the previous service provision period. In the calculation phase, given the workload requirement, the current environment conditions, and the combinations of the number of vCPUs (virtual CPUs) and memory size, the performance and operation state are inferred by using the trained model. The input parameters can be set by the SP or the CSP or obtained from the monitoring data as shown in the figure. Next, if the inferred performance and operation state satisfy the performance requirements and the operation policies, the corresponding combination of resource amounts is output as the feasible solution for the Resource Amount Calculator.

3. Application scenarios of IBSM

IBSM can be used to assist in the consultation, design, and operation phases in cloud service delivery. For example, in the consultation phase for an SP that plans to migrate services implemented in an on-premises environment into a cloud environment, IBSM is used to show the performance that can be achieved after the migration and the needed cloud resources and cost. In the design phase, IBSM enables the automatic design of resources. Thus, service design time and human labor can be expected to be reduced. In the operation phase, IBSM is able to adjust the resource composition and amount in accordance with the changes, thus ensuring the continuous satisfaction of service requirements, which contributes to higher customer satisfaction.

4. Future plans

This article introduced an automatic service design technology for a cloud service under development at NTT Access Network Service Systems Laboratories. Our plans in the near future are to verify the effectiveness of the technology for representative workloads in the cloud service, identify the specific functions needed for different phases of cloud service delivery, and conduct trial experiments to enhance commercial use of the technology.

References

[1] C. Wu and S. Horiuchi, "Intent-Based Service Management -- To Decrease Complexity of Virtualized Network Management," IEICE Tech. Rep., Vol. 117, No. 305, ICM2017-28, pp. 41–46, Nov. 2017.

[2] C. Wu and S. Horiuchi, "Intent-based Service Management," Proc. of the 21st Conference on Innovation in Clouds, Internet and Networks (ICIN 2018), Paris, France, Feb. 2018.

Chao Wu
Research Engineer, Access Network Operation Project, NTT Access Network Service Systems Laboratories.

She received a B.E. in engineering from Zhejiang University, People's Republic of China, in 2009 and an M.E. in engineering from Waseda University, Tokyo, in 2013. In 2014, she joined NTT Access Network Service Systems Laboratories, where she has been researching and developing management mechanisms for telecommunications. She is also involved in standardization efforts of the next-generation operations support system in the European Telecommunications Standards Institute Experimental Networked Intelligence (ETSI ENI) and TM Forum.

Shingo Horiuchi
Senior Research Engineer, Access Network Operation Project, NTT Access Network Service Systems Laboratories.

He received a B.E. and M.E. in engineering from the University of Tokyo in 1999 and 2001. In 2001, he joined NTT Access Network Service Systems Laboratories, where he has been researching and developing access network operation systems. He has also been involved in standardization efforts for operations support systems in TM Forum as a member of the Open Digital Architecture (ODA) Project since 2014. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Kenichi Tayama
Senior Research Engineer, Supervisor, Access Network Operation Project, NTT Access Network Service Systems Laboratories.

He received a B.E. and M.E. in electrical engineering from the University of Electro-Communications, Tokyo, in 1993 and 1995. He joined NTT Optical Network Systems Laboratories in 1995. He also worked at NTT EAST's IT Innovation Department and NTT-ME's Network Operation Center, where he researched and developed network operations support systems. He is a member of IEICE.