Optimize Cloud-server Resources for Comfortable Web Conferencing

Hiroaki Kikushima and Chao Wu

Abstract

This article introduces an application example of the cloud-server resource-optimization technology, which is one of the technologies comprising the intent artificial intelligence (AI) mediator called Mintent, to a web-conferencing service. This technology recommends the minimum amount of resources necessary to meet the requirements (intents) of each service provider and user, reducing resource costs while ensuring quality of experience, by using AI that predicts performance in accordance with various service usage conditions. This helps to meet a variety of requirements, such as improving user engagement with services, reducing service providers' operating and resource costs, and reducing carbon dioxide emissions through more efficient energy use.

Keywords: intent, resource control, web-conferencing service

1. Introduction

Changes in social conditions, such as the rapid increase in remote work and the promotion of digital transformation, have accelerated the rapid change in demand for services and the diversification of user needs.

To keep up with changes and the diversity in these environments, various systems are becoming cloud-native, and various services and network functions are being provided on cloud servers with a shorter lead time. Therefore, server-resource control is becoming increasingly important; however, it is difficult to provide services that meet quality of experience (QoE) requirements and usage conditions with resource design based on experience and resource control based on system performance. Therefore, NTT Access Network Service Systems Laboratories developed a novel cloud-server resource-optimization technology that takes into account various user's QoE in addition to system-performance requirements.

The cloud-server resource-optimization technology automatically calculates optimal resources under various service-usage conditions while maintaining QoE, leading to reductions in resource design and operation and resource costs.

To show the effectiveness of this technology, this article introduces its application to a web-conferencing service.

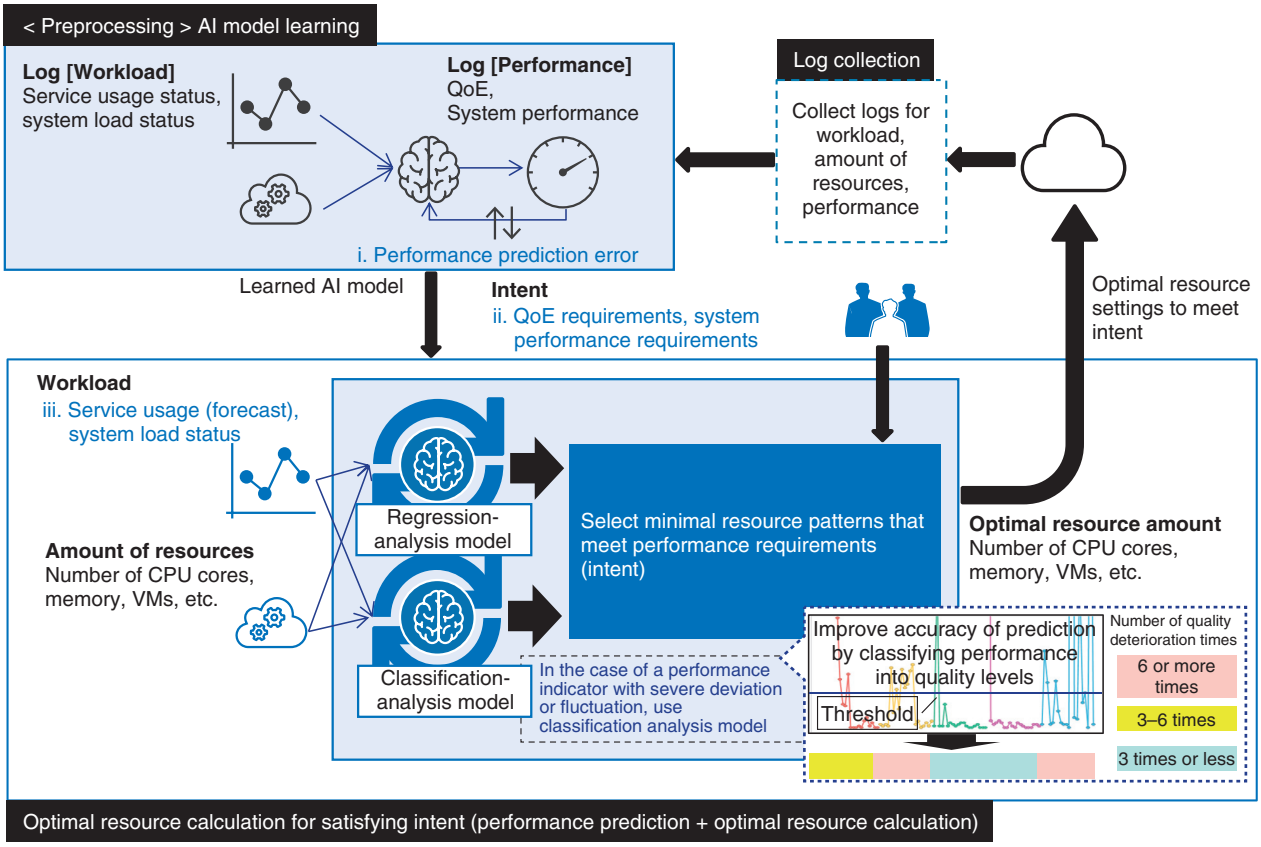
2. Cloud-server resource-optimization technology overview

2.1 Technological superiority

The cloud-server resource-optimization technology uses an artificial intelligence (AI) model that predicts various types of performance on the basis of system load and service-usage conditions and calculates the minimum amount of resources that satisfy the requirements (intents) of service providers and users (**Fig. 1**).

It has the following advantages over conventional resource-control technology that focus only on system performance.

By combining multiple AI models that predict various types of performance, this technology can calculate the optimum amount of resources not only for system-performance requirements but also for intents including QoE requirements. The AI models used are trained to use different log data to predict different types of performance. Basically, a regression analysis



CPU: central processing unit
VM: virtual machine

Fig. 1. Technical overview.

model is used. However, some QoE indicators are difficult to predict numerically through regression analysis. In this case, by using a classification analysis model, the technology is able to predict the performance level with high precision (Fig. 1). These features make it possible not only to avoid quality failures and improve resource efficiency but also to provide services tailored to users' usage conditions and objectives.

Various efforts have also been made to train AI models to more reliably meet QoE requirements. The loss function is unique, for example, with modifications to lower intent violation risk (**Fig. 2**). The loss function is also defined separately not only for the degree of violation but also for the period and number of violations, which are important for performance management in network operation.

By using both the service (application) and system (infrastructure) workloads (iii. in Fig. 1) as input to the AI, it is possible to calculate resources in accordance with the service usage status (including predictions) in addition to the system load status. This makes it possible to proactively control the service in accordance with the demand forecast of the service provider. This advantage is especially effective in cases such as a web-conference service, where it is not easy to scale up the host server (virtual machine instance) of a conference or change the host server while providing the service once the meeting started. By calculating the optimal resource in advance on the basis of the meeting-reservation information, the system can scale out in advance when the meeting starts.

2.2 Application of cloud-server resource-optimization technology

This technology can be used not only in the resource-design phase when a service is introduced but also for dynamic resource control in the service-operation phase. This technology can be used not only for SaaS (Software as a Service) but also for

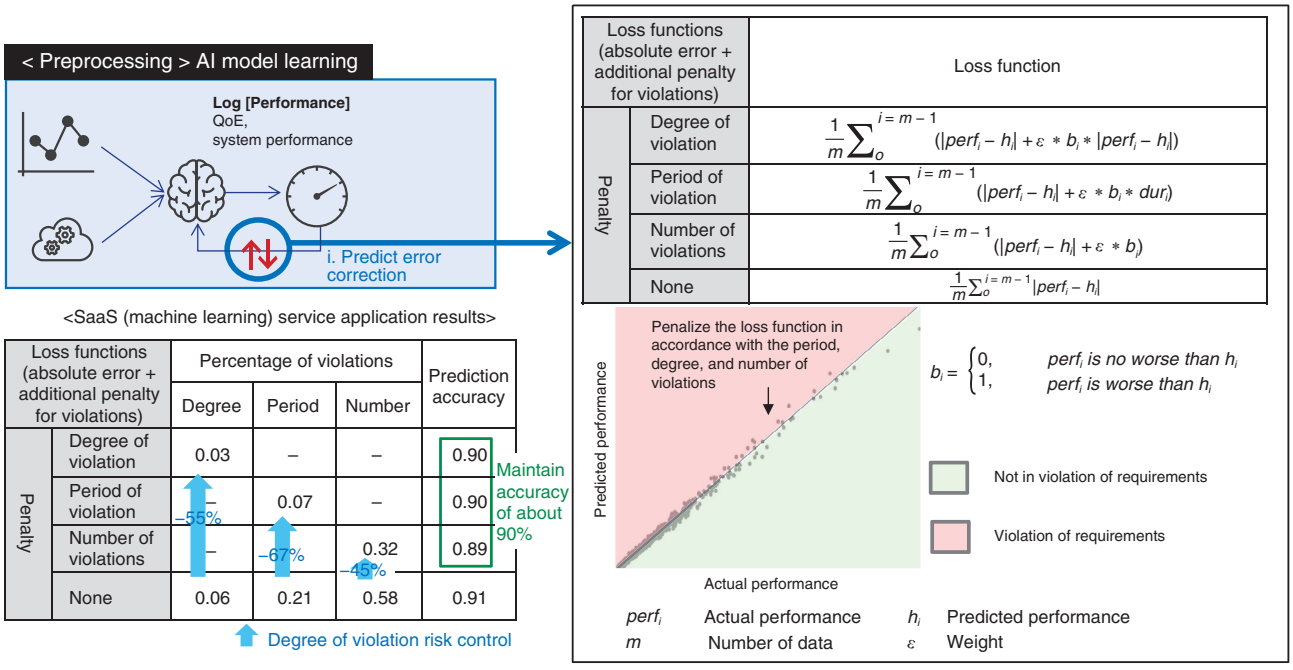


Fig. 2. Requirement-violation risk-control function.

IaaS (Infrastructure as a Service) infrastructure. In the future, we plan to expand the scope of application scenarios to virtual network services.

3. Application of cloud-server resource-optimization technology for web-conferencing services

For the application of this technology to web-conferencing services, we devised AI models and developed new functions on the basis of the issues and features of web-conferencing services.

3.1 Requirements with web-conferencing services and their solutions

Web-conferencing services have the following requirements due to the changes in social conditions described above.

(1) A web-conferencing service provider wants to optimize server resources and cost by allocating servers in accordance with server resource conditions and the amount of meetings.

(2) By clarifying that it has the technology for QoE maintenance, a web-conferencing service provider wants to improve the brand value for the service.

We can satisfy these requirements by applying this technology, as shown in **Fig. 3**.

With the conventional resource operation, it is common to prepare sufficient server resources in advance for the expected maximum service usage (number of users, etc.) on the basis of the experience of engineers and verifications. In this case, when there are only a few users, unnecessary resources are allocated, and when more users access than expected, QoE degradation occurs. It is also difficult to increase the number of central processing unit (CPU) cores while running a service such as web conferencing, which can lead to a fatal situation.

By combining this technology with the control functions of cloud services, it is possible to solve this problem by operating services with the minimum amount of resources when the number of users is small and scaling out in advance servers as the number of users increases.

3.2 Constructing a performance-prediction AI model for web conferencing

QoE indicators that are important for web-conferencing services are generally throughput and jitter, which are indicators of the cleanliness and stability of video and audio. Because these values fluctuate constantly, they are difficult to predict with high accuracy (**Fig. 4(a)**). Therefore, we focus on the fact that these

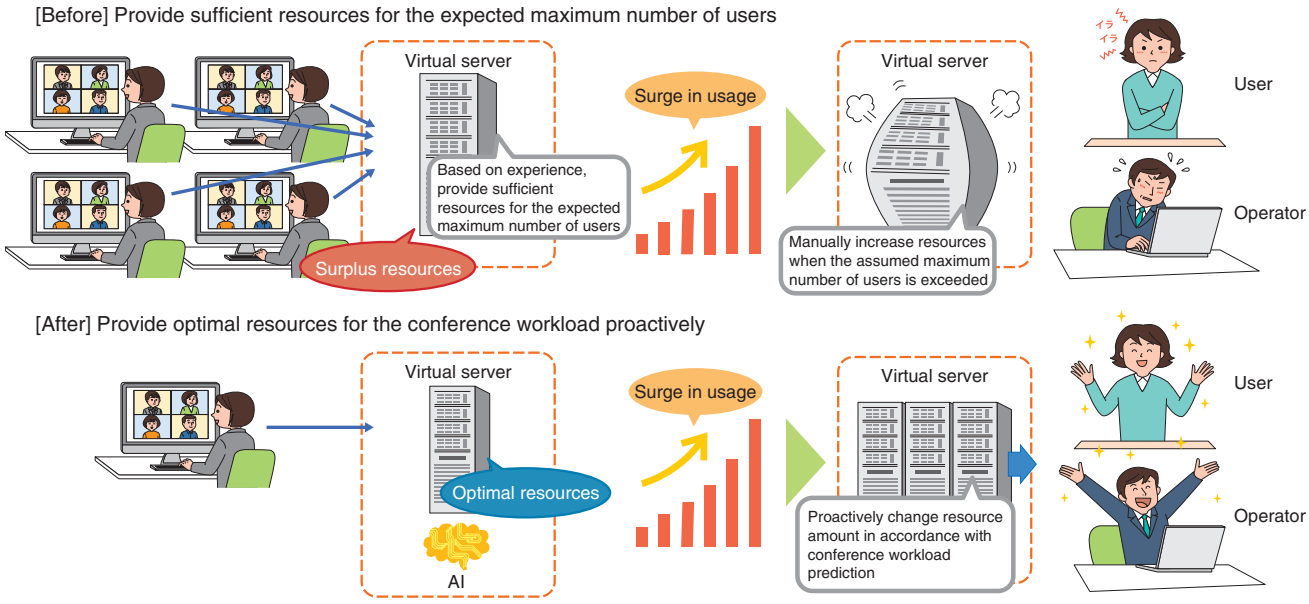


Fig. 3. Image of problem solving.

values do not deteriorate when the CPU usage is below a certain level, and we can derive the optimal resource amount to maintain QoE by using an AI model (**Fig. 4(b)**) that predicts the CPU usage from the workload and resource amount.

3.3 Peripheral function

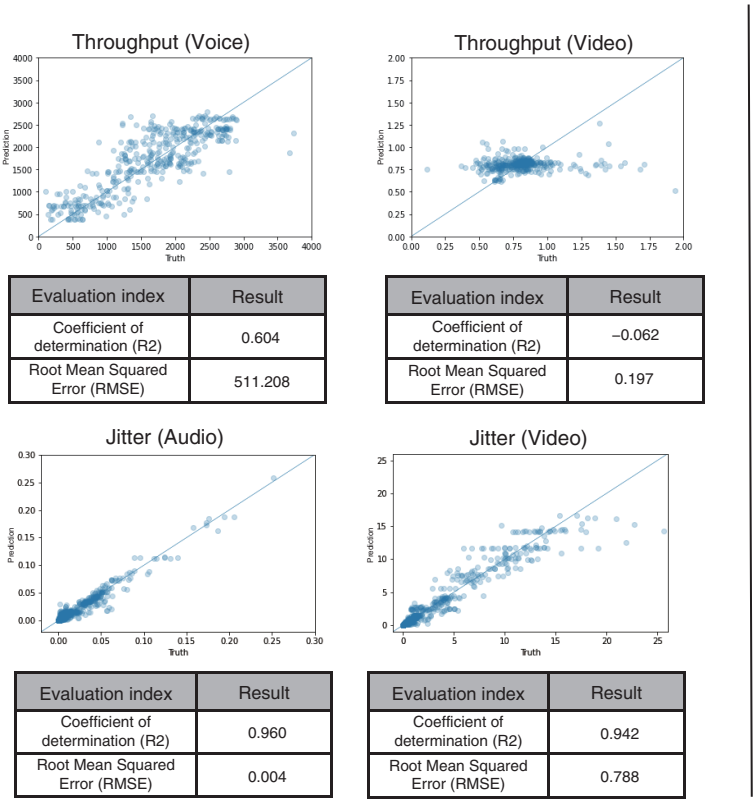
As mentioned above, to provide this technology to a web-conferencing service, it is necessary to have the function of determining the amount of resources required and the allocation of the conference among various server instances before the conference starts. Therefore, it is necessary to provide a function to supplement the workload information (the number of video and audio streams) from the meeting-reservation information before an actual meeting occurs (**Fig. 5(a)**). A function is also required to predict the QoE indicator of each meeting combination in accordance with the supplemented workload and determine whether it can be accommodated in the existing instances with predetermined instance type (number of CPU cores) and execute scale-out in advance (**Fig. 5(b)**). By operating this resource-calculation function on the basis of these usage conditions (predictions) periodically (from a few minutes to a few hours), proactive and automatic resource control can always be achieved during service operation, and the maximum effect can be obtained in both maintaining QoE and improving resource efficiency.

4. Optimal resource-calculation results and effects

On the basis of the actual usage trends of commercial web-conferencing services, we generated user workloads assuming that up to 100 users would use the service and verified the effectiveness of our technology. First, we used this technology to predict CPU usage for an instance type with four CPU cores and confirmed that the result was almost the same as the actual measurement when the same number of workloads were actually generated (**Fig. 6(a)**). On the basis of these predictions, the system can change the web-conferencing server to an instance with two CPU cores in advance when resources are predicted to be in surplus and change the web-conferencing server to an instance with eight CPU cores in advance when QoE is predicted to degrade. This enables service provision while maintaining CPU usage below the threshold and efficient use of resources (**Fig. 6(b)**).

From these results, we found that by using this technology to control resources in accordance with the expected amount of workload, we could reduce the amount of server resources by approximately 37% while maintaining QoE compared with when constantly operating excess resources with eight CPU cores.

By greatly improving the efficiency of server resources in this manner, it is possible to reduce

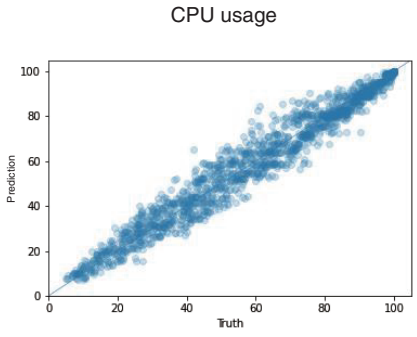


Evaluation index	Result
Coefficient of determination (R2)	0.604
Root Mean Squared Error (RMSE)	511.208

Evaluation index	Result
Coefficient of determination (R2)	–0.062
Root Mean Squared Error (RMSE)	0.197

Evaluation index	Result
Coefficient of determination (R2)	0.960
Root Mean Squared Error (RMSE)	0.004

Evaluation index	Result
Coefficient of determination (R2)	0.942
Root Mean Squared Error (RMSE)	0.788

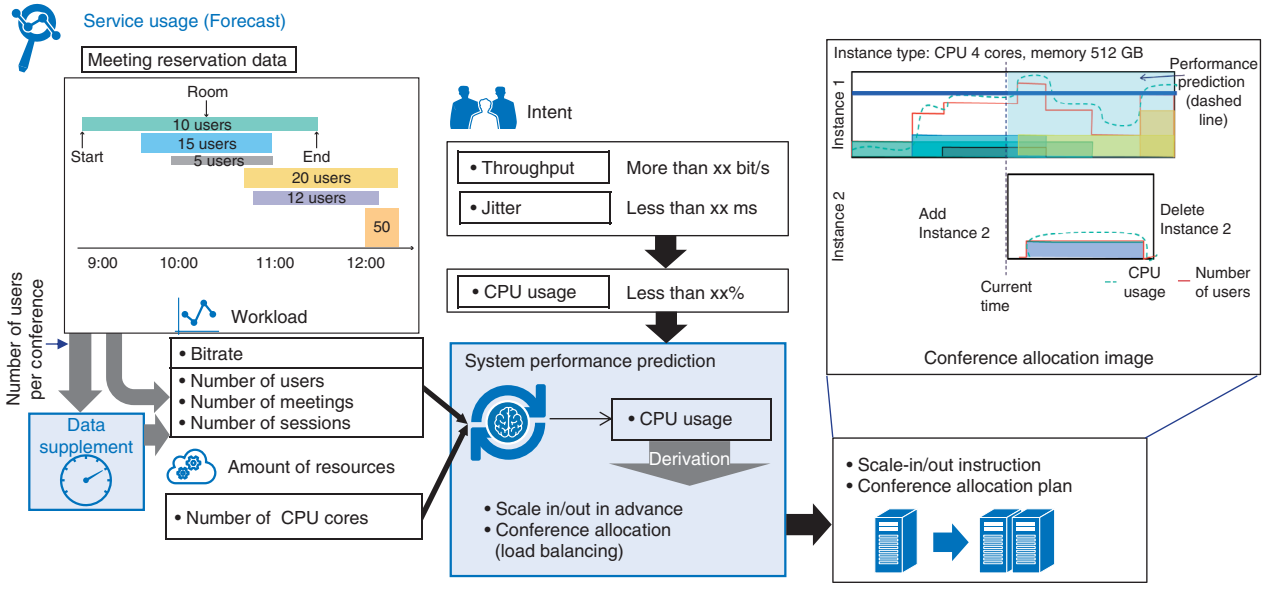


Evaluation index	Result
Coefficient of determination (R2)	0.970
Root Mean Squared Error (RMSE)	5.039

(a) Throughput/jitter

(b) CPU usage

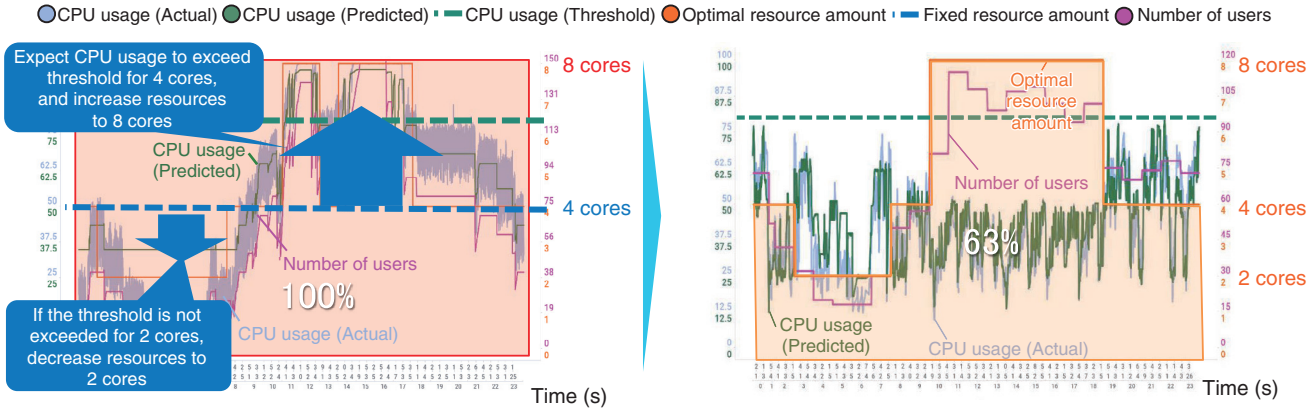
Fig. 4. AI-model evaluation results.



(a) Data supplement function

(b) Conference allocation control function

Fig. 5. Application to a web-conferencing service.



(a) CPU usage when fixed 4 cores are allocated to the VM

(b) CPU usage when cloud-server resource-optimization technology applied

Fig. 6. Optimal resource calculation results and effects by cloud-server resource-optimization technology.

energy consumption, and by adapting this technology to various use cases, it will be possible to further contribute to carbon dioxide reduction.

5. Conclusion

This article gave an overview of cloud-server resource-optimization technology and explained its effectiveness when applied to a web-conferencing service.

In the future, we will examine the technology's interface with the current cloud-resource management systems for commercial introduction to a web-conferencing service, improve its functions such as improving the speed of conference allocation, combine it with other control technologies that comprise Mintent, and consider expanding the application area to other use cases in view of the IOWN (Innovative Optical and Wireless Network) era. We will also promote the creation of new technologies to extract quantitative intent from more ambiguous intents of users and service providers and convert business-service-layer intent to resource-layer intent.

Hiroaki Kikushima
Research Engineer, NTT Access Network Service Systems Laboratories.
He received a B.E. and M.E. in electrical engineering from the University of Yamanashi in 1994 and 1996. He joined NTT Software Headquarters in 1996. He also worked at NTT COMWARE's AI Business Strategy Office, where he created various services. Since 2020, he has been researching and developing access network operation systems at NTT Access Network Service Systems Laboratories.

Chao Wu
Research Engineer, NTT Access Network Service Systems Laboratories.
She received a B.E. from Zhejiang University in 2009 and M.E. from Waseda University, Tokyo, in 2013. In 2014, she joined NTT Access Network Service Systems Laboratories, where she has been researching and developing intelligent management technologies for telecommunications in cloud and virtualization. She is also involved in standardization efforts of the European Telecommunications Standard Institute (ETSI) and TM Forum.