논문 00-25-11A-13 한국통신학회논문지 '00-11 Vol.25 No.11A

유틸리티 모델을 이용한 인터넷 서비스의 SLA 관리

정회원 백 종 욱*, 박 종 태*

SLA Management for Internet Service Using Utility Model

Jong-Wook Baek*, Jong-Tae Park* *Regular Members*

요 약

인터넷 서비스 품질 레벨의 효과적인 관리는 고객과 서비스 제공자 모두에게 점차 중요해지고 있다. 본 논문에서는 인터넷 서비스를 위한 서비스 레벨 협정 관리의 문제를 수학적으로 제시하고, 제시된 문제의 휴리스틱 해를 도출한다. 문제를 풀기 위해, 서비스 레벨 협정의 관리와 제어적인 측면을 나타내기 위한 유틸리티 모델을 제안한다. 유틸리티 모델을 이용한 SLA 관리의 문제 정형화 및 휴리스틱 해는 인터넷 서비스에 대한 승인 제어 및 품질 적응에 계산적으로 실현 가능한 해결책을 제공해준다. 마지막으로, VoIP 서비스 시뮬레이션을 통해 제시된 해의 성능을 평가한다.

ABSTRACT

The efficient management of quality level of Internet service is becoming increasingly important to both customers and service providers. In this paper, the problem of SLA management for Internet service is represented mathematically and a heuristic solution for the problem is presented. To solve the problem, we propose the utility model to capture the management and control aspect of SLA for Internet service. The heuristic solution provides a computationally feasible solution to do admission control and quality adaptation for Internet service. Finally, the performance of the solution is evaluated by simulating VoIP service.

I. Introduction

As competition is progressively introduced into all service provision markets, service providers are realizing the need to differentiate their service quality. Customers do not care how a service is composed, but the quality of service (QoS) is important to them. The QoS expectations are driving customers to negotiate with service providers that could meet their requirements for specific level of service. This is increasingly being done through a service level agreement (SLA) [1]. An SLA is a contract between the service provider and the customer that specifies the quality level of service that can be expected.

An SLA includes the expected behavior of the service and the parameters for quality of service and so on. The efficient management of an SLA is a new challenge and very important issue in the service provision markets.

There have been some research works on the SLA management. The QoS team in TeleManagement Forum has been working on the automation of the interface between service provider and customer for performance reporting with the SLA concept [1]. They have identified common terms and definitions, and created an industry-wide glossary for performance measurement and reporting. Bhoj [2] presented a Web-based SLA management framework to allow

* 경북대학교 전자전기공학부(park@ee.knu.ac.kr)
논문번호 : 00190-0601, 접수일자 : 2000년 6월 1일

easy inter-domain communication. He demonstrated how service providers could offer SLA monitoring capabilities to their customers for a number of services including email and network access services. Park [3] presented the support of QoS management using SLA concept, which was measured, monitored and controlled systematically in a multi-domain environment.

While these research works offer a good start for SLA management, there still remain unsolved problems. These works only focus on monitoring and reporting mechanisms of SLA. These works don't address how to manage and control the quality level of service provided to customers while utilizing efficiently the network resources. None of these research works presents a formalized solution of the problem.

In this paper, the problem of SLA management for Internet service is represented mathematically and a heuristic solution is presented. To solve the problem, we propose the utility model to capture the management and control aspect of SLA for Internet service. This model provides a computationally feasible solution to make admission control and quality adaptation for multimedia Internet service. Finally, the performance of the solution is evaluated by simulating VoIP service.

Ⅱ. Utility Model for SLA Management

We apply the utility concept used in microeconomic theory to capture the management and control aspect of SLA. The utility model is based on the concepts of quality profile, quality-to-resource mapping, resource constraints and utility function. The main concepts of the utility model for SLA management are illustrated in Figure 1.

Each customer specifies a quality profile which is the set of acceptable operating qualities for the service. A customer's operating qualities are assumed to be mapped uniquely to required resources. The service provider's system and network are subject to the system and network resource constraints. A customer's operating qualities are assumed to be mapped uniquely to

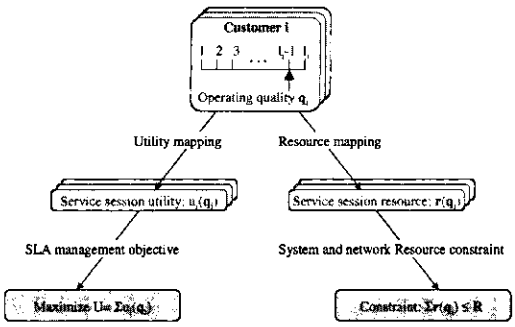

Fig. 1 The main concepts of the utility model

the service session utilities by utility function.

The service utility is the sum of all service session utilities. The problem of SLA management is to find the operating quality $\vec{q}_i$ of each customer i which maximizes the service utility under the system and network resource constraints. These concepts are explained in the following subsection.

1. Quality Profile

The quality profile specifies the quality preference of customers. It is a set of acceptable operating qualities for the service and specified through contract between a customer and a service provider. For simplicity, we consider three items in the quality preference: qa, qb, and qc. Then the operating quality of customer i can be expressed as

$$\vec{q}_i = (q_{ia}, q_{ib}, q_{ic})$$

We assume that the quality profile of customer i is a sequence of acceptable operating qualities from lowest operating quality to highest operating quality. Mathematically the quality profile of customer i can be expressed as

$$\vec{P}_i = (\vec{q}_{i1}, \vec{q}_{i2}, \cdots, \vec{q}_{il_i})$$

where li is the number of quality in $\vec{P}_i$, $\vec{q}_{i1}$ is the lowest acceptable quality and $\vec{q}_{il_i}$ is highest acceptable quality.

2. Quality-to-Resource Mapping

We assume that there exists a mapping from an

operating quality to the resources required to provide that quality. Let us assume only three resources for a service: d, e, and f. Then the required resources $\vec{r}_i$ of operating quality $\vec{q}_i$ can be expressed as

$$\vec{r}_i = (d_i, e_i, f_i)$$

where

$$d_i = d(\vec{q}_i) = d(q_{ia}) + d(q_{ib}) + d(q_{ic})$$

$$e_i = e(\vec{q}_i) = e(q_{ia}) + e(q_{ib}) + e(q_{ic})$$

$$f_i = f(\vec{q}_i) = f(q_{ia}) + f(q_{ib}) + f(q_{ic})$$

Here, d(·), e(·) and f(·) are quality-to-resource mapping operators for the resource d, e, and f respectively. In vector notation, the quality-to-resource mapping can be expressed as

$$\vec{r}_i = \vec{r}(\vec{q}_i)$$

In Figure 2, the quality-to-resource mapping is illustrated. Figure 2 (a) represents a quality profile. Figure 2 (b) shows how each operating quality of the quality profile is mapped to a resource profile. For example, operating quality $\vec{q}_{i1}$ would require resource profile $\vec{r}_{i1}$, i.e $\vec{r}_{i1} = \vec{r}(\vec{q}_{i1})$. The quality-to-resource mapping transforms the quality profile to the resource profile of a service session. The resulting resource profile is illustrated in Figure 2 (c). Intuitively, one may expect that better quality requires more resources. However the resource profile may not always show this monotonic behavior. For example, it is possible that the quality of service can be improved by using more di but fewer ei.

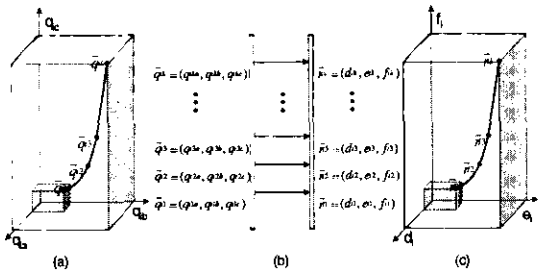

Fig. 2 Quality-to-resource mapping

3. Resource Constraint

For each resource related to the service, the sum of the quantities of the resource allocated to all the customers cannot exceed the total available quantities of the resource. Suppose that the available network resources are expressed as a vector $\vec{R} = (D, E, F)$. Then the resource constraint can be expressed as

$$\sum_{i=0}^{n} \vec{r}(\vec{q}_i) \le \vec{R}$$

where n is the number of customers

4. Utility Function

The utility denotes the satisfaction of a service provider for the consumed network resources. This utility concept is used in microeconomics theory [4], and we have applied it for SLA management. The utility function is used to measure the amount of satisfaction that a service provider gets from a service. The utility function assigns each service session to some real number. The real number denotes the amount of utility obtained from service provision. It can be expressed as

$$u(\cdot) \rightarrow R$$

where u(·) is a utility function and R is a real number.

5. SLA Management Objective Function

If a customer i's utility value of a service can be obtained from its operating quality $\vec{q}_i$ by using the utility function, the objective of SLA management can be expressed as a service utility objective. The service utility objective is to maximize the service utility function U given by

$$U = \sum_{i=1}^{n} u_i(\vec{q}_i)$$

under the resource constraints in Ⅱ.3. The SLA management objective function may be modified by a set of operation policies based on issues such as profit, and priority.

Ⅲ. Genera Problem Formulation and Solution

Suppose there are n customers. The ith customer has li quality levels, and the amount of available resources is given by $\vec{R} = (r_1, r_2, ..., r_m)$. The quality level j of customer i has a utility uij, and requires resources $\vec{r}_{ij} = (r_{ij1}, r_{ij2}, ..., r_{ijm})$. Using these given definitions, the objective of SLA management for Internet service is to solve the following optimization problem.

$$U = max\left\{\sum_{i=1}^{n}\sum_{j=1}^{l_i} x_{ij}\, u_{ij}\right\} \quad (1)$$

such that

$$\sum_{i=1}^{n}\sum_{j=1}^{l_i} x_{ij}\, r_{ijk} \le R_k, \quad k = 1, 2, ..., m \quad (2)$$

$$\sum_{j=1}^{l_i} x_{ij} = 1, \quad i = 1, 2, ..., n \quad (3)$$

$$x_{ij} \in \{0, 1\}, \quad i = 1, 2, ..., n; \; j = 1, 2, ..., l_i \quad (4)$$

Equation 1 means that the problem of SLA management of Internet service is to find the operating quality $\vec{q}_i$ of each customer while maximizing the service utility under the resource constraint (Equation 2). Equation 3 and 4 mean that only one operating quality has to be chosen in the acceptable operating qualities of each customer.

The above problem is the variant of 0-1 knapsack problem. Since the 0-1 knapsack problem is known to be NP-hard [5, 6], the worst-case computation time of the optimal solutions grows exponentially with the size of the problem. This is not suitable for time-critical control and management such as dynamic resource allocation and admission control for multimedia Internet service. To cope with time-critical management and control, we present a heuristic solution. Figure 3 shows a heuristic solution for the problem. It starts with the operating quality with the smallest utility in each customer, and iteratively improves the solution by gradually replacing it as an operating quality with larger utility as long as the solution is acceptable. The decision criterion of upgrading operating quality is to maximize the extra resource savings (Δr) and the utility gain per unit of the extra resource (Δp).

For a customer i, there may be (l_i -1) upgrade at worst case. The maximum iteration of the loop of line 4-23 is $\sum_{i=1}^{n}(l_i - 1)$. The computational complexity of the loop of line 6-19 is $\sum_{i=1}^{n}(l_i - 1)$. If we combine the computational complexity of two loop, we find the time complexity of the heuristic algorithm is $O\left(\left(\sum_{i=1}^{n}(l_i - 1)\right)^2\right)$. If we assume that the number of items in all groups is equal that $l_1 = l_2 = ... = l_n = l$, then the computational complexity of the heuristic algorithm is $O(n^2(l-1)^2)$. This is much smaller than the computational complexity of the optimal solution which is $O(2^{n/3})$ [5].

```
procedure HEURISTIC(n,m,l,Res,ResMax,Util)
// n: # of customers, m: # of resource
// l[n]: # of quality levels of customers
// Res[n][l][m]: resources usages
// ResMax[m]: maximum values of m resources
// Util[n][l]; utility values
1  integer Ans[n], i, j, i ′, j ′
2  real ResUsage[m], Δr, Δp, ΔrMax, ΔpMax
3  ResUsage[m] ← getResUsage(Ans, m)
4  loop
5    Ans[n] ← 1; ΔrMax ← 0; ΔpMax ← 0
6    for i ← 1 to n do
7      for j ← Ans[i]+1 to l[i] do
8        if checkUsageExceedMax (i, j, Res, ResUsage,
                        ResMax) then
9          continue endif
10       Δr ← getResSaving (i, j, Res, ResUsage)
11       if Δr > ΔrMax then
12         ΔrMax ← Δr; i ′ ← i; j ′ ← j endif
13       if ΔrMax ≤ 0 then
14         Δp ← getUtilityGain (i, j, Util, Δr)
15         if Δp > ΔpMax then
16           ΔpMax ← Δp; i ′ ← i; j ′ ← j endif
17       endif
18     repeat
19   repeat
20   if ΔrMax ≤ 0 and ΔpMax ≤ 0 then
       return Ans
21   Ans[i ′ ] ← j ′
22   ResUsage[m] ← getResUsage(Ans, m)
23   repeat
end HEURISTIC
```

Fig. 3 A heuristic solutioon

Ⅳ. Example: SLA Management for VoIP Service

In this section, we present the SLA management for VoIP service [7] using the utility model.

1. Utility Model for VoIP Service

Figure 4 shows that the proposed utility model is applied to SLA management for VoIP service. A VoIP service provider contracts with customers and makes the quality profiles of each customer which is a sequence of acceptable operating qualities from lowest acceptable operating quality to highest acceptable operating quality. The service provider maps each operating quality level to appropriate resource profile, and also maps it to utility value. The utility value of a quality level can be obtained using a utility function which is determined by the service provider's operation policy. A utility value is a real number and represents the amount of satisfaction that the service provider can get from a VoIP service. The details of the procedures are explained below.

1.1 Getting the quality profile

A VoIP service provider must be able to specify quality profile which expresses QoS requirements. This can be achieved using a static table of acceptable qualities. For instance, a simple quality profile for a VoIP session may have three discrete qualities: Bronze, Silver and Gold. The VoIP session's minimum acceptable quality is Bronze, and its maximum desired quality is Gold.

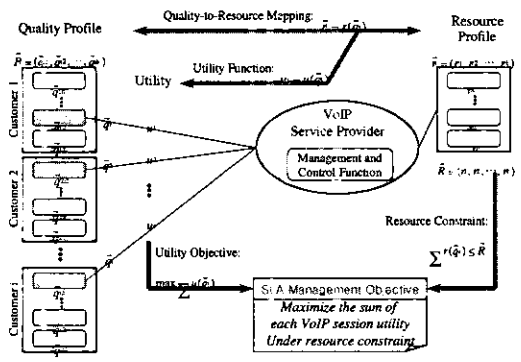

Fig. 4 The utility model for VoIP service

1.2 Quality-to-resource mapping

The utility model assumes the existence of an operating quality to required resource mapping. However, how such a mapping can be obtained is another research issue [8]. The resource allocation can be done by profit maximization, fair share policy, and priority policy. The resource allocation may be obtained using off-line experimental evaluation, but it is dependent on service provider's platform.

1.3 Quality-to-utility mapping

If the utility of each VoIP session represents a customer's bill, then the quality of the VoIP session can be mapped to a utility value using the following utility function,

$$u(x) = 1 - e^{-x/\tau} \tag{5}$$

where x is the cost of operating quality $\vec{q}_i$ and τ is a constant. In this case, the sum of each VoIP session utility means the service provider's profit. The management system uses these values for admission control and run-time quality adaptation. Since Equation 5 is the normalized exponential utility function, addition of a constant and/or multiplication by a positive constant leads to another strategically equivalent utility function, we could have written the exponential utility function in the form

$$u(x) = A - Be^{-x/\tau}$$

However, the given normalized form has the advantage that u(0) = 0 and the limit of u(x) as x goes to infinity is 1. The finite upper bound on utility as x goes to infinity can be regarded as the cost equivalent to infinity, but the utility assigned to it remains less than some finite value.

3. QoS Management Function for SLA Conformance

In this subsection, we describe the admission

control and QoS adaptation functions for SLA conformance of VoIP service.

3.1. Admission Control of New VoIP Sessions

Admission control is necessary for service with quality guarantees because the system has to ensure that enough resources are available at run-time to meet the minimum quality guarantee. Suppose that the VoIP service provider has currently n sessions, and the current total utility is Un. When a customer requests a new session, the utility driven admission control can be processed as follows.

Step1: The management system checks whether any feasible solution of Equation 2 where the $n+1$ VoIP sessions can share the currently available resources. If such a solution does not exist the new session must be rejected.

Step2: If there exist a feasible solution with n+1 sessions, Suppose that the maximum service utility of $n+1$ VoIP sessions is U_n+1. If $U_n+1<U_n$, the new session should be rejected as unprofitable. Otherwise, the session should be accepted.

3.2. QoS Adaptation of VoIP sessions

Customers are more satisfied as the voice quality is better. However a service provider should guarantee a minimum service quality for all the customers although service provider's system and network condition are changing. In other words, service provider must be able to dynamically adapt the operating quality of each VoIP session when the quality is degraded or the network resource status is changed.

Figure 5 describes QoS adaptation function based on the proposed utility model. It is composed of monitoring function, assessment function and control function. The monitoring function plays the role of monitoring the performance of VoIP sessions and the network resource status. The assessment function decides whether the QoS violation occurs or the QoS

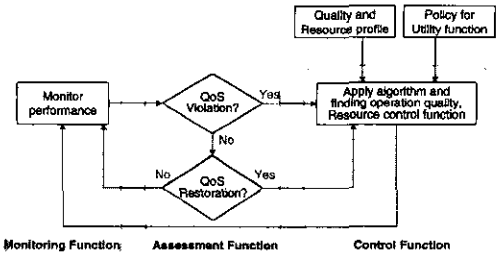

Fig. 5 QoS adaptation functional description

restoration is required. If required, the control function finds the new operating quality of each VoIP session using the heuristic solution in Figure 3 and reallocates the resources to guarantee the QoS of VoIP session within SLA.

V. Performance Evaluation

We compare the temporal variation of the service utilities provided by the three prototype implementations using random sequences of events for VoIP service: the utility model prototype using optimal solution(UMOP), the utility model prototype with proposed heuristic (UMHEU), and the simple reservation model prototype(SRM) for VoIP service.

We use the service utility provided by solutions under test as the main performance index. Since in the utility model, the goal of the adaptive SLA management system is expressed as a constrained maximization problem of the service utility, a higher service utility obtained by a solution implies better performance. We also compare the solutions in terms of service utility and computation time. The simple reservation model prototype is based on a reservation based QoS model with no adaptation for VoIP service.

We have implemented the event generation program GenEvent, which generates quasi-random events using four input parameters: number of events n, average size of time steps t_s, average size of utility steps u_s, and average duration of a session d_s. Here the parameter t_s is used to control the average time difference between two consecutive VoIP session requests, and the

parameter u_s is used to control the average difference between the utility values of two consecutive service levels of a session. The parameter d_s is used to control the average duration between the request and drop of a session. Events may be of two types: session request event or session drop event. The session request events are obtained from the input sequence of events, and if the session is admitted, a session drop event is inserted by the system.

Figure 6 shows the temporal variation of the service utility by three prototypes UMOP, UMHEU and SRM for the sequence of events generated by 'GenEvent 10 5 10 40' and for a value of 60 for each of the three resources.

In Figure 6, we note that the service utilities obtained by the two utility model prototypes are always higher than that of the simple reservation model prototype. We also note that the performance of the utility model prototype using the heuristic (UMHEU) is comparable to that of the utility model prototype using the optimal algorithm (UMOP). UMOP provided optimal service utility most of the time, and provided close to optimal service utility rest of the time.

We have also compared the total computation time required by the implementations to process the sequences of events. For the sequence of 10 events, the computation times required by implementations UMOP, UMHEU and SRM are 28411, 1289 and 331 microseconds, respectively. For the sequence of 100 events, the corresponding numbers are 1856318, 24695 and 2947.

Table 1 compares the performance of UMOP, UMHEU and SRM in terms of service provider's utility and computation time. We note that the servicer provider's revenue provided by UMHEU is 99% of the servicer provider's revenue provided by UMOP, and the computation time of UMHEU is less than 5% of the computation time of UMOP. On the other hand, the servicer provider's revenue provided by SRM is significantly lower than that of either UMOP or UMHEU.

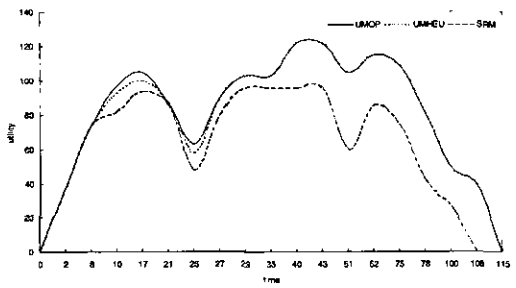

Fig. 6 Utility variation for three prototypes

Table 1. Performance of three prototypes

Sequence of 10 events		Sequence of 10 events	
utility	computation time	utility	computation time
1.00	1.000	1.00	1.000
0.99	0.045	0.99	0.013
0.74	0.011	0.80	0.002

Ⅵ. Conclusion

The problem of SLA management for Internet service is maintaining the quality level provided to customers within a pre-negotiated range while utilizing efficiently the resources of system and network. In this paper, we have presented the utility model to solve SLA management problem. The problem has been mathematically formulated and the heuristic solution has been presented. The proposed utility model can be used not only for resource allocation decisions, but also for quality adaptation and admission control for multimedia Internet service. We have also shown by simulation that the heuristic solution is proven to be effective enough for SLA management for VoIP service. For future work, we plan to apply the utility model to SLA management for other types of Internet service.

참 고 문 헌

[1] NMF, "Performance Reporting Definitions Document," NMF 701, June 1998.

[2] P. Bhoj, S. Singhal, and S. Chutani, "SLA Management in Federated Environments," Proceedings of 6th IFIP/IEEE International Symposium on Integrated Network Manage-

ment", Boston, MA, U.S.A., pp. 293 - 308, May 1999.

[3] J. T. Park and J. W. Baek, "Web-based Intranet/Internet Service Management with QoS Support," IEICE Transactions on Communications, Vol.E82-B, No.11, pp.1808~1816, November 1999.

[4] A. Mas-Colell, *Microeconomics*, Oxford University Press, 1995.

[5] E. Horowithz and S. Sahni, *Fundamentals of Computer Algorithms*, Computer Science Press, pp. 501-558, 1989.

[6] S. Martello and P. *Toth. Knapsack Problems: Algorithms and Computer Implementations*, John Wiley & Sons, Chichester, 1990.

[7] G. Held, *Voice Over Data Networks*, McGraw-Hill, 1998.

[8] TOMQAT Project, "Dynamic QoS Management Framework for IBC Networks," TOMQAT Deliverables 12, November 1995.

백 종 욱(Jong-Wook Baek) 정회원

1995년 2월 : 경북대학교 전자공학과 졸업
1997년 2월 : 경북대학교 전자공학과 석사
1997년 3월~현재 : 경북대학교 전자공학과 박사과정

<주관심 분야> 망관리, 인터넷서비스관리, 멀티미디어통신, 분산처리

박 종 태(Jong-Tae Park) 정회원

1978년 2월 : 경북대학교 전자공학과 졸업
1981년 2월 : 경북대학교 전자공학과 석사
1987년 2월 : The University of Michigan 공학박사
1989년~현재 : 경북대학교 전자 전기공학부 교수

<주관심 분야> TMN, IP기반 멀티미디어 서비스 제공기술, CTI, 이동통신 망관리