The Effects of QoS Level Degradation Cost on Provider Selection and Task Allocation Model in Telecommunication Networks

Hasan Huseyin Turan
University of Yalova, Department of Industrial Engineering
Research Assistant
Merkez, 77100, Yalova, Turkey
hasanturan@sabanciuniv.edu

Nihat Kasap
Sabanci University, Faculty of Management
Assistant Professor
Tuzla, 34956, Istanbul, Turkey
nihatk@sabanciuniv.edu

Berna Tektas Sivrikaya
Istanbul Technical University, Faculty of Management
Research Assistant
Maçka, 34367, Istanbul, Turkey
tektasbe@itu.edu.tr

—Abstract—
Firms acquire network capacity from multiple suppliers which offer different Quality of Service (QoS) levels. After acquisition, day-to-day operations such as video conferencing, voice over IP and data applications are allocated between these acquired capacities by considering QoS requirement of each operation. In optimal allocation scheme, it is generally assumed each operation has to be placed into resource that provides equal or higher QoS Level. Conversely, in this study it is showed that former allocation strategy may lead to suboptimal solutions depending upon penalty cost policy to charge degradation in QoS requirements. We model a cost minimization problem which includes three cost components namely capacity acquisition, opportunity and penalty due to loss in QoS.

Key Words: Penalty Cost Policy, QoS Level, Heuristic, Break-even-point Analysis

JEL Classification: C69

1. INTRODUCTION

In the real life firms do businesses in an environment in which they use data networks to perform and support their business operations, which we call tasks. A network resource can be characterized by its capacity (bandwidth and duration) and Quality of Service (QoS) levels. In the network environment, the firm firstly lease the capacity (bandwidth) from the providers (bandwidth broker firms or network operators) then allocate its planned tasks into that capacity. The firm generally uses networks in order to do works such as video conferencing, voice over TCP/IP and data applications, which are named as tasks (Kasap *et al.,* 2007; Tektaş and Kasap, 2008). As stated in Kasap *et al.,* (2007) and Tektaş and Kasap (2008), generally, two types of tasks are performed using data networks. A task is time-fixed (real-time, size compressible) if its size can be changed without disrupting its completion but the transmission time cannot be compressed or extended.

For telecommunication networks QoS offered by providers are measured in terms of delay, jitter, lost rate and latency (Pan *et al.*, 2009). Delay specifies how long it takes for data to travel across the network from source to destination (Comer, 2001; Gupta *et al.*, 2006; Mao, 2005). Jitter represents the variance in delay. Packet loss represents data dropped or irrecoverably damaged by the network. Specifications of these measures are described in Service Level Agreements (SLA). An SLA is an electronic contract between a service user and a provider, and specifies the service to be provided, QoS properties that must be maintained by a provider during service provision, and a set of penalty clauses specifying what happens when service providers fail to deliver the QoS agreed (Rana *et al.*, 2008).

Each task is affected differently by network reliability and speed. Therefore, major network providers have already started efforts to accommodate the QoS demand generated by these applications. Also different providers charge differently for the capacity they lease out. Therefore, it can be said that there are various billing models for quality-differentiated network capacity leasing as a function of bandwidth, traffic volume, applications, and pricing structure.

Despite the different pricing strategies in the literature, still operators manage the excess capacity by selling it in bulk, albeit at reduced prices. We believe, until the demand catches up with supply the all-you-can-send pricing will remain a popular option especially for the network operators. That’s why, in our model we consider

the all-you-can-send pricing scheme (Courcoubetis *et al.,* 2000) in which the firm pays a fixed price for a fixed bandwidth available for a fixed duration. We also assume that the providers can offer any quality of service and some capacity at competitive prices.

In optimal allocation scheme, it is generally assumed each operation has to be placed into acquired capacity that provides equal or higher QoS Level. In previous works, there are two types of costs associated with using data networks. The first one is the resource (i.e., bandwidth, capacity) acquisition cost. The second is the opportunity cost incurred due to insufficient transmission rate in performing certain tasks such as video conferencing. In this study, we include an extra cost component depending upon the firm's penalty cost policy to charge degradation in its QoS requirements. Therefore, the problem that the firm has to solve is a cost minimization problem that reflects a tradeoff between the cost of acquiring resources, the opportunity cost of degradation in realized quality of tasks performed due to reduction in transmission rate and the penalty cost of degradation in QoS levels considering delay and jitter. It is in this setting that we wish to model the problem from the firm's point of view. We consider the firm's cost minimization problem in an environment in which the firm can lease network capacity at competitive prices from different providers with different service quality.

The quality of service (or the lack of it) of the resource affects the customer in two ways. First, size-fixed tasks might be delayed beyond acceptable deadlines. Second, the realized quality of a time-fixed task such as a videoconference might be unacceptable creating an opportunity cost for the customer. We assume that a real-time task incurs an opportunity cost if its transmission rate falls below a desired level because of the network provider. In general, opportunity cost reflects the importance of a task. The more important the task is, the higher the penalty for not achieving desired targets (such as picture or sound quality). Consequently, what matters is the magnitude of the opportunity cost relative to other tasks and relative to the unit cost of a resource. Also, we assume that if the customer allocates tasks into resources with lower QoS levels than required ones, it incurs the cost of quality degradation. A cost minimizing decision maker will consider the trade-off between these costs when assigning tasks to resources.

The rest of the paper is organized as follows. In Section 2, we present an overview of related research in the literature. Section 3 describes our assumptions and the

mathematical model. Section 4 is devoted to discussion of proposed heuristic algorithm. Finally, we conclude with summary and future research.

2. LITERATURE REVIEW

Existing literature on QoS can be classified into two groups. The first group considers the design and implementation of network infrastructure and operating policies including QoS issues. The main concern addressed in this line of research is guaranteeing a promised level of quality through bandwidth allocation, buffer management and scheduling. There is extensive work on these issues in electrical engineering and computer science fields that describes the necessary infrastructure (including buffer management tools and packet scheduling algorithms) and service models to support applications with varying QoS parameters (Lee and Kim, 2002; Mao, 2005; Verma and Pankaj, 1998; Yuang, *et al.*, 1998).

Wu *et al.* (2006) present an analysis of the bandwidth allocation problem with a penalty cost and derive the optimal bandwidth capacity. They propose a stochastic model for macro-level bandwidth management from the viewpoint that emphasizing the randomness and risk aversion and their impacts on the network's performance. They introduce a penalty cost in the optimization model for network bandwidth allocation for unsatisfied traffic demand.

Doğan and Özgüner (2006) investigate the problem of scheduling a set of independent tasks with multiple QoS needs, which may include timeliness, reliability, security, data accuracy, and priority, in a heterogeneous computing (HC) system. They formulate the QoS-based scheduling problem by using utility and penalty functions, where a utility function associated with a task is used to measure how much the owner of this task will benefit from a given scheduling decision, while penalty functions associated with resources are used to provide incentives to users to set their QoS requirements in accordance with their needs.

The second group in the literature includes supply-side strategic issues such as pricing and QoS. Gupta *et al.* (2006) argue that a pricing mechanism can be used as a tool to manage the residual uncertainty in QoS. In Bouras and Sevasti (2004), prices are also selected to maximize customers' utility, while utility is a function of delay as well as bandwidth, and prices are decomposed into two parts accounting for buffer and bandwidth, respectively.

3. PROBLEM DEFINITION AND MATHEMATICAL MODEL

The objectives of proposed mathematical model is to: (i) analyze effect of penalty cost on task allocation problem when SLA specifications are not met; and (ii) specify to what extent taking penalty cost in other words, fail to deliver agreed QoS deliberately is more optimal (reasonable). For this purpose, mathematical model which is proposed by Kasap *et al.* (2007) is modified in order to cope with QoS degradation cost. Original model only takes into account two types of cost namely capacity acquisition and opportunity cost. Moreover, it enforces each task to be allocated into resources that offer at least equal or higher QoS level. Therefore, zero QoS degradation cost (w) occurs. In this study previous restriction is relaxed and quality degradation cost term is added to objective function.

3.1. Problem Notation

The following parameters and decision variables are used while presenting formulation and solution procedures throughout the text.

Parameters:

I, J : The ordered index set of resources, and tasks respectively.

A_T, A_S : The index set of tasks with fixed transmission time and fixed size respectively where $A_T \cap A_S = \varnothing$.

α_i : Transmission efficiency, calculated as one minus the packet loss rate of resource i.

β_i, L_i : The bandwidth and duration of resource i. $L_i = min$ (length of contract, planning horizon)

c_i : The total cost of resource i for specific β_i, L_i.

c_j^o : The opportunity cost of missing the target transmission rate for task j.

w_j : The quality degradation cost of not meeting required QoS level of task j

$\overline{r}_j^U, \overline{r}_j^L$: Target and the minimum transmission rate of task j at the receiving node, respectively.

Δt_j : Estimated scheduled transmission time for time-fixed task $j \in A_T$.

$\overline{x}_j$: The (fixed) length of task $j \in A_s$ in number of bits.

Q_i :QoS level of resource i.

q_j :Minimum required QoS level for task j.

Decision variables:

v_i : 1 if resource *i* is selected, zero otherwise.

r_j : The transmission rate of task *j*.

y_{ij} : 1 if task *j* is assigned to resource *i*, zero otherwise.

y_{ijt} : 1 if task *j* is active (transmitting) at time *t* on resource *i*, 0 otherwise.

t_j : Start time of task *j*.

ξ_{ij} :1 if task j is assigned to resource i which does not meet QoS requirement, zero otherwise.

3.2. Mathematical Model

$$Min \quad z = \sum_i c_i v_i + \sum_i \sum_{j \in A_T} \left(\overline{r}_j^U - r_j \alpha_i \right) c_j^o y_{ij} + \sum_i \sum_{j \in A_T \cup A_S} y_{ij} w_j \xi_{ij} \tag{1}$$

Subject to

$$\sum_{j \in A_S} \overline{x}_j y_{ij} + \sum_{j \in A_T} \alpha_i \Delta t_j r_j y_{ij} \le \alpha_i \beta_i L_i \qquad \forall i \in I \tag{2}$$

$$\xi_{ij} \ge y_{ij}(Q_i - q_j) \qquad \forall \mathrm{j} \in \mathrm{J} \text{ and } \forall \mathrm{i} \in \mathrm{I} \tag{3}$$

$$(t_j + \Delta t_j) y_{ij} \le L_i \qquad \forall j \in A_T, \forall i \in I \tag{4}$$

$$\sum_{j \in J} r_j y_{ijt} \le \beta_{\mathrm{i}} \qquad \forall i \in I, t = 1..L_i \tag{5}$$

$$\sum_{i \in I} y_{ij} = 1, \qquad \forall j \in J \tag{6}$$

$$\sum_{i \in I} \sum_{t \le L_i} y_{ijt} = \Delta t_j, \qquad \forall j \in A_T \tag{7}$$

$$\sum_{i \in I} \sum_{t \le L_i} r_j y_{ijt} = \overline{x}_j, \qquad \forall j \in A_S \tag{8}$$

$$y_{ij} \le v_i \quad \forall i \in I,\ j \in J \tag{9}$$

$$y_{ijt} \le y_{ij}, \forall i \in I, j \in J, t \le L_i \tag{10}$$

$$r_j \alpha_i \ge y_{ij} \overline{r}_j^L \qquad \forall j \in A_T,\ \forall i \in I \tag{11}$$

$$r_j \alpha_i y_{ij} \le \overline{r}_j^U \qquad \forall j \in A_T,\ \forall i \in I \tag{12}$$

$$r_j, t_j \ge 0$$

$$v_i, y_{ij}, y_{ijt}, \xi_{ij} \in \{0,1\} \quad (13)$$

Proposed model considers the tradeoff between the cost of acquiring capacity, the cost of not meeting target transmission rates in real-time tasks, and penalty cost for not meeting requested QoS levels. It is assumed that available capacity is purchased at a fixed price for a specific bandwidth and duration. Discrete QoS levels (as high and low) for both providers and tasks are used. By definition quality penalty cost occurs if and only if a task j is allocated a resource where $q_j < Q_i$ holds.

Constraint set (2) guarantees that we can use only up to available capacity. Constraint set (3) ensure that if the resources do not satisfy the minimum QoS requirements of the jobs that are assigned to it than positive amount of quality penalty cost occurs for that task. Within determined confidence limits. Constraint (4) ensures that all time-fixed tasks assigned to a resource are completed when the resource is available. Constraint set (5) prevents using more bandwidth than available at any time (bandwidth dimension). Constraint (6) along with (13) ensures that a task is assigned to only one resource and all tasks are assigned. Constraint sets (7) and (8) guarantee that the jobs are actually allocated the required amount of time slices. Constraint set (9) guarantees that a network resource is selected only if at least one task is assigned to it. Constraint set (10) ensures that a task is assigned to a network resource only if it occupies a time slice on it. Constraint set (11) states that transmission rate for a time-fixed job *j* should be high enough to satisfy the minimum transmission (reception rate) at the sink node. Constraint set (12) enforces the target transmission limitation for all jobs.

4. HEURISTIC ALGORITHM

Flow chart of the proposed heuristic is given in Figure 1. Underlying logic of heuristic algorithm is as follows. It starts with taking zero penalty cost by placing each task into resources such that QoS level of resource is always higher or equal to minimum requested QoS of task and it starts with highest possible capacity acquisition cost. This sub-heuristic is discussed in detail at Kasap *et al.* (2007) and named as Heuristic A. We use the same naming and use it as building block for suggested algorithm. Our approach tries to increase penalty cost but on the other hand looks for savings in acquisition cost. As given in Figure 1, the algorithm is divided into 3 sub-step for sake of simplicity for explanation. Each sub-step is briefly explained below:

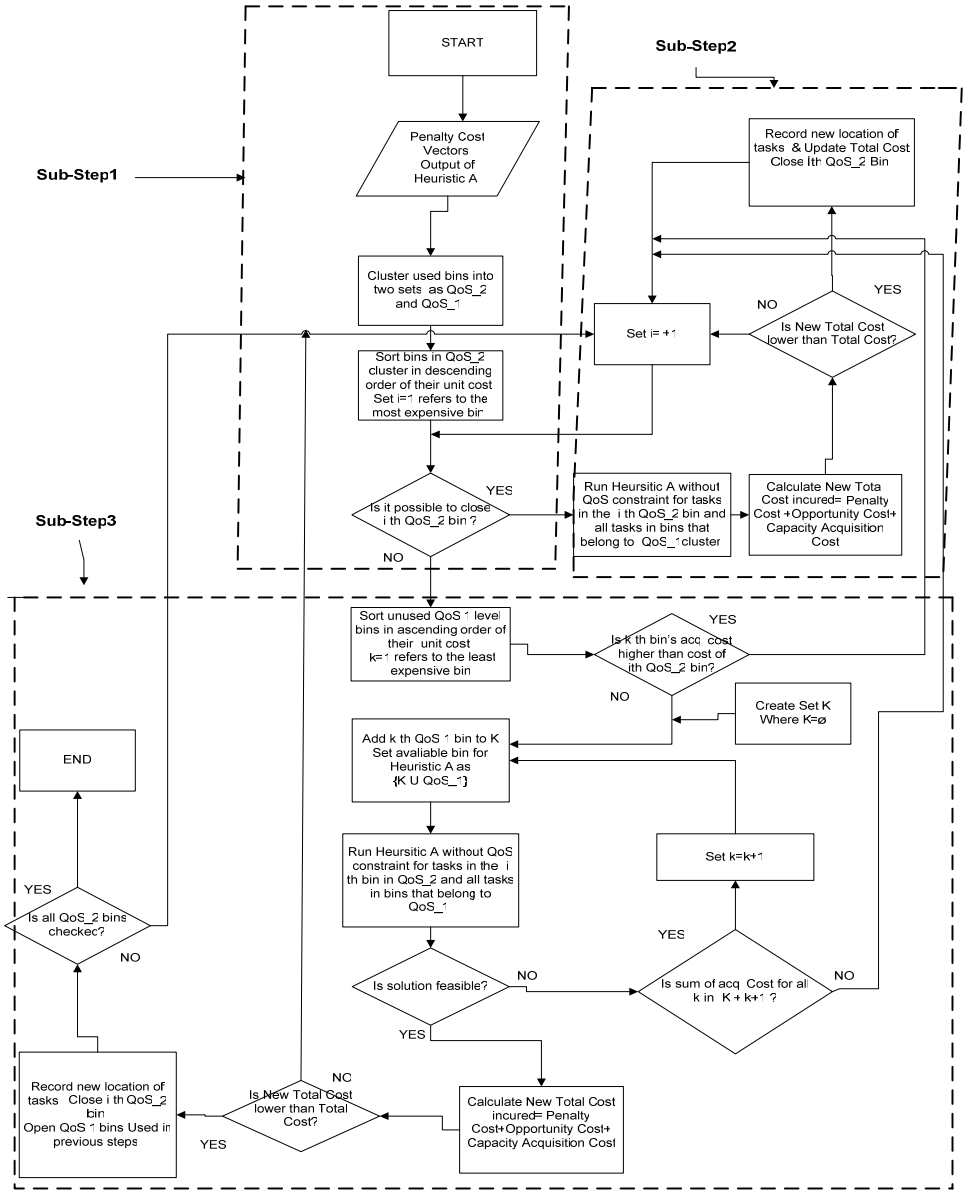

Figure 1: The Flowchart of Proposed Heuristic

Sub-step 1: This step mainly consists of reading data which are needed for Heuristic A and penalty cost vector for each task (w_j). Then, Heuristic A is run for given data sets and output of this algorithm is used as input for future sub-steps. Furthermore, selected resources are clustered into two disjoint sets as

QoS_1 and QoS_2. These two sets contain lower and higher QoS level resources respectively. All resources in these clusters are sorted by their unit cost ($c/(\beta xL)$ in descending order. The most crucial stage in this sub-step is to check whether it is possible to allocate tasks that are in most expensive resource into other used resources by taking penalty cost so that total capacity acquisition cost can be lowered. If this is possible sub-step 2 is executed otherwise algorithm jumps into sub-step 3.

Sub-step 2: The objective of this sub-step is trying to reallocate tasks which are already placed into expensive resources into QoS_1 level resources by taking penalty cost. At the end of this step new total cost incurred is calculated, and depending upon the result either new location of tasks are updated or not. This steps and Sub-step 1 runs back and forth till all QoS_1 resources are checked. The interaction between sub-step 1 and 2 is illustrated in Figure 1.

Sub-step 3: This step considers the possibility of purchasing lower level QoS capacity in order to unselect expensive resources that are already in use (selected). For this purpose, unused QoS 1 level resources are sorted by their ascending unit cost. Hence, heuristic A is executed by changing input available resource set by adding cheap QoS 1 resources. This step continue to looping until it can be sure that there is no possibility of unselecting any QoS_2 resource by taking penalty cost and as a final step all cost are calculated and added to check the improvement in solution in terms of cost.

5. SUMMARY AND FUTURE WORKS

We presented a novel formulation to solve the firm's network resource acquisition problem subject to QoS requirements, opportunity and quality degradation costs. We also proposed a heuristic to handle QoS requirements and quality degradation mechanism. As a next step we implement the proposed algorithm then perform sensitivity analysis and simulation. We will examine the results after analysis and try to show how different degradation cost policies affect the optimal behavior of the firm while selecting providers and allocating tasks.

ACKNOWLEDGEMENTS

This research is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK), Career Development Program (Grant No. 106 K 263).

BIBLIOGRAPHY

Bouras, C. and Sevasti, A. (2004). "SLA-based QoS pricing in DiffServ networks", *Computer Communications*, Vol.27, pp.1868 – 1880.

Comer, D. E. (2001). Computer Networks and Internets with Internet Applications, Third ed. Prentice Hall, New Jersey

Courcoubetis, C., Kelly, F. and Weber, R. (2000). "Measurement-based usage charges in communications networks", *Operations Research*, 48 (4), pp.535–548.

Doğan, A. and Özgüner, F. (2006). "Scheduling of a meta-task with QoS requirements in heterogeneous computing systems", *Journal of Parallel and Distributed Computing*, 66, pp.181-196.

Gupta, A., Kalyanaraman, S. and Zhang, L. (2006). "Pricing of risk for loss guaranteed intra-domain internet service contracts", *Computer Networks*, 50, pp.2787-2804.

Kasap, N., Aytug, H. and Erenguc, S. S. (2007). "Provider selection and task allocation issues in networks with different QoS levels and all you can send pricing", *Decision Support Systems*, 43, pp.375 – 389.

Lee, K. D. and Kim, S. (2002). "Optimization for adaptive bandwidth reservation in wireless multimedia networks", *Computer Networks*, 38, pp.631-643.

Mao, G. (2005). "A real time loss performance monitoring scheme", *Computer Communications*, 28, pp.150–161.

Pan W., Yu L., Wang S., Hua G., Xie G, and Zhang J. (2009), "Dynamic Pricing Strategy of Provider with Different QoS Levels in Web Service", *Journal of Networks*, Vol.4 No:4, pp.228-234.

Rana O., Warnier M. and Quillinan T. B. (2008), "Monitoring and Reputation Mechanisms for Service Level Agreements", *The Fifth International Workshop on Grid Economics and Business Models (GECON)*.

Tektas, B. and Kasap, N. (2008). "Time and volume based optimal pricing strategies for telecommunication networks." In Proceedings *CD IAMOT 2008*, Dubai: United Arab Emirates.

Verma, S. Pankaj, R. K. and Garcia, A.L. (1998). "Call admission and resource reservation for guaranteed quality of service (GQoS) services in internet", *Computer Communications,* 21, pp.362–374.

Wu J., Yue W. and Wang S. (2006), "Stochastic model and analysis for capacity optimization in communication networks", Vol.29, pp.2377-2385.

Yuang, M. C. and Haung, Y. R. (1998). "Bandwidth assignment paradigms for broadband integrated voice/data networks", *Computer Communications*, 21, pp.243–253.