

# Pricing and Optimal Resource Allocation in Next Generation Network Services

Michael G. Kallitsis  
Department of Electrical and  
Computer Engineering  
North Carolina State University  
Raleigh, NC 27695  
kallitsis@ncsu.edu

George Michailidis  
Department of Statistics  
University of Michigan  
Ann Arbor, MI 48109  
gmichail@umich.edu

Michael Devetsikiotis  
Department of Electrical and  
Computer Engineering  
North Carolina State University  
Raleigh, NC 27695  
mdevets@ncsu.edu

**Abstract**—In this paper, we introduce a pricing model that ensures efficient resource allocation that provides guaranteed quality of service while maximizing profit in multiservice networks. Specifically, a dynamic allocation policy is examined that relies on online measurements while each service class operates under a probabilistic bound delay constraint. We present a rigorous analysis of the properties of the policy that provides insights into its workings as well as its sensitivity to various parameters. Finally, its performance is evaluated through an extensive numerical study.

## I. INTRODUCTION

Recent technology advances have led to dramatic changes in the communications arena. The use of fiber optics and the increased performance of integrated circuits have brought to the forefront diverse types of networks, such as broadband, wireless ad hoc and mesh networks, and next generation cellular systems.

However, these new technologies are not sufficient by themselves to guarantee business success. Added value and service differentiation need also be considered, in order for the service providers to be profitable. Hence, the trend towards networks providing some degree of value added services has emerged. Specifically, Service Oriented Networks is an evolving architecture that would allow for a priced based differentiated choice of network services [16]. Along the same lines is the triple play network architecture, a user-centric approach in which customers are confronted with a variety of applications like Voice over IP, IPTV and Video on Demand and high speed internet services [18].

These emerging network services require enhanced and diverse quality-of-service (QoS) guarantees. Thus, the development of scheduling algorithms that provide differentiated service guarantees to various classes of traffic is of great interest. Such generalized schedulers and on demand routers should dynamically allocate the desired network resources (e.g., bandwidth, buffer size, CPU capacity) since a static allocation may result in significant under-utilization.

In this paper, we propose a service pricing model that ensures efficient allocation of resources in a dynamic manner in the aforementioned multiservice networks. The scheme requires close on-line monitoring of the incoming traffic. We

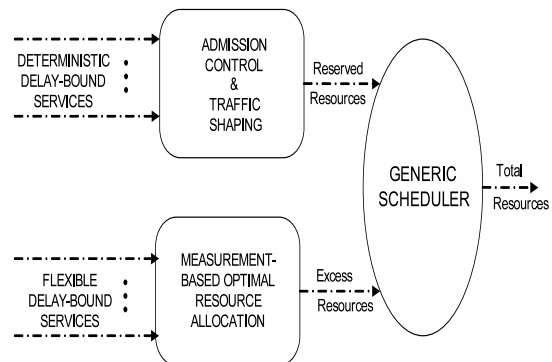


Fig. 1. Depiction of the proposed framework: traffic is divided into two categories; deterministic constraint and flexible constraint services. The system allocates the excess resources to the latter set.

assume a Fractional Brownian Motion traffic model, because of its ability to adequately capture characteristics of real network traces, such as self-similarity and the presence of heavy tailed marginal distributions [17].

Optimal resource allocation is also studied in [1], [9], [20]. Specifically, Peng *et al.* propose a measurement-based resource allocation scheme based on a linear pricing model and average queue delay guarantees. This scheme has the disadvantage of not being scalable to large number of service classes. Moreover, average queue delay is not always an appropriate QoS constraint. In [9], they perform a maximization over a utility function provided from the network users and resources are shared based on the solution of that optimization problem. In [20], the authors study the problem of resource allocation with dynamic pricing in which the network administrator controls the price of the resources that users demand; based on the demand the prices are dynamically changed over different time periods so as to maximize the revenue of the administrator. Finally, measurement-based resource allocation has also been studied in different contexts in [10]–[12].

The remainder of this paper is structured as follows. The proposed modeling framework is described in Section II, while the optimization problem for a single network element based on a nonlinear pricing model is formulated in Section III. Some numerical results illustrating the model's performance

are presented in Section IV, while some concluding remarks are drawn in Section V.

## II. MODELING FRAMEWORK

The employed modeling framework was introduced in [1], [2] and is depicted in Figure 1. In its present form it represents a single network element, which can be either a traditional network component, such as switch or a router, or a modern network “service center”, like IBM’s DataPower Service Oriented appliances [3] or CISCO’s Application Oriented Network (AON) message routing system [19].

It is assumed that the network element serves two categories of traffic classes; deterministic delay-bound classes and flexible delay-bound ones. Due to the fact that deterministic delay-bound classes have strict requirements, their service level agreement (SLA) can be satisfied only by traffic shaping and admission control schemes [13], [14]. Thus, an amount of resources is dedicated to them and these classes are excluded from subsequent analysis. Examples of these inelastic classes of service include teleconferencing, remote seminars, real-time distributed computation/simulation and high-precision medical imaging.

Therefore, the proposed system is responsible for optimally allocating the excess resources to the remaining flexible delay-bound classes. These classes enter the Measurement Based Optimal Resource Allocation (MBORA) system proposed in [2] and shown in Figure 2. The MBORA system consists of a *measurement* module, an *optimization* module and a *resource orchestrator* module. The statistics of the arrival traffic are measured by the measurement module. It is assumed that the traffic can be accurately approximated by a Fractional Brownian motion model, which can account for the burstiness and long-range dependence observed in real traffic traces. Such a model can be fully described by the following parameters: the *Hurst* parameter  $H$ , the *mean* arrival rate  $\bar{\alpha}$  and the *variance*  $\sigma$ . An algorithm for on-line measurement of these parameters is discussed in [4].

The optimization module receives the traffic characteristics of each class and calculates the optimal allocation of resources by solving the optimization problem discussed in Section III. It should be noted that the optimization problem is solved only when there is a significant change in traffic characteristics. The optimal solution is fed to the resource orchestrator which dynamically updates the allocation of resources for each traffic class and forwards the packets (or, more generally, the messages, for example XML) toward their destination.

## III. PRICING MODEL AND OPTIMIZATION PROBLEM FORMULATION

We start by introducing the pricing model, whose solution yields the optimal allocation of resources to the network service node we described in the previous section.

### A. Non-Linear Pricing Model

Suppose that the node can provide  $K$  different types of services. The proportions of these services to be allocated are

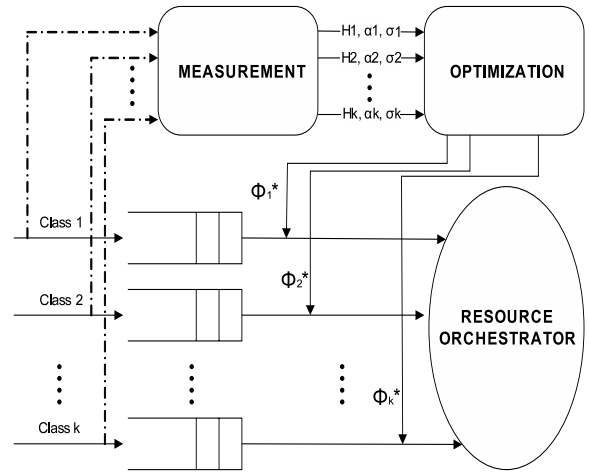


Fig. 2. The MBORA system: the optimization module receives input traffic measurements and calculates the optimal resource allocation.

denoted by  $\phi = (\phi_1, \dots, \phi_K)$ . According to [7], the profit of a provider is the difference between the revenue  $r(\phi)$  that is obtained for providing these services, and the cost  $c(\phi)$  that incurs from producing them. The aim of this provider is to maximize the profit function

$$\pi = \max_{\phi} \{r(\phi) - c(\phi)\} = \max_{\phi} \sum_{k=1}^K (r_k(\phi_k) - c_k(\phi_k)), \quad (1)$$

subject to the feasibility constraints:  $\phi_k \geq 0$ ,  $k = 1, \dots, K$ ,  $\sum_k \phi_k \leq 1$ .

The revenue is given by a linear function, while the cost by a nonlinear one. Specifically,  $r_i(\phi_i) = p_i \cdot \phi_i$  while the cost function has the form  $c_i(\phi_i) = b_i \cdot D_i(\phi_i) \cdot \exp[\beta(D(\phi_i) - d_i)]$ . The coefficient  $p_i$  corresponds to the price that the provider charges for service  $i$  and the parameter  $b_i$  is the amount that the provider has to reimburse the users whenever the SLAs are not met. A higher priority class  $u$  requires better service than a lower one  $v$  and thus it is charged more (i.e.,  $p_u > p_v$  and  $b_u > b_v$ ). The parameter  $\beta$  controls the steepness of the cost function, while  $D(\phi_i)$  denotes the value of the performance metric experienced by users of service  $i$  and  $d_i$  the target level under the SLA. Hence, if  $D(\phi_i) > d_i$  the users are not receiving adequate resources from the provider, which would incur a cost, until the situation is rectified. This function is monotone in  $D_i(\phi_i)$  and is shown in Figure 3. The steep increase in the cost observed beyond the desired by the users SLA value of  $d_i$  would force the provider to adjust the allocation of resources (if possible), in order to satisfy the QoS requirements and maximize profit.

*Probabilistic Delay Constraints:* We employ stochastic delay bounds as the metric for QoS considerations. Specifically, we adopt the approach used in [5], [6], where traffic is treated as Long Range Dependent (LRD) and is characterized by the Hurst parameter  $H$ , the mean  $\bar{\alpha}$  and the variance  $\sigma$ . It is shown that the queue length at any given time  $t$  is bounded by a value  $q_{max}$  with probability  $\epsilon > 0$  related to the desired QoS. It is

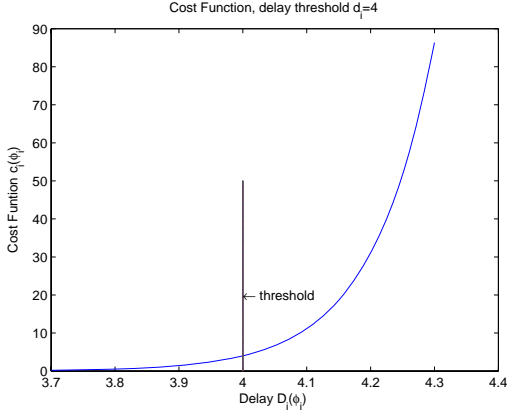


Fig. 3. Our cost function. Notice that even a small increase of 2.5% above the delay threshold yields an increase above 100% in the cost function. In this case parameter  $\beta = 10$ .

shown that for a specific class the following holds:

$$Pr(Q(t) > q_{max}) \approx \epsilon \quad (2)$$

and

$$q_{max} = (C - \bar{\alpha})^{H/(H-1)} (k\sigma)^{1/(1-H)} H^{H/(1-H)} (1-H) \quad (3)$$

where  $C$  can be interpreted as the resources (e.g., bandwidth) dedicated to this particular class,  $\epsilon$  is the required QoS and  $k = \sqrt{-2ln\epsilon}$ .

Thus, since the queue length and expected delay are related, we have the following probabilistic delay bound:

$$Pr(D(t) > D_{max}) \approx \epsilon \quad (4)$$

and

$$D_{max} = \frac{(C - \bar{\alpha})^{H/(H-1)} (k\sigma)^{1/(1-H)} H^{H/(1-H)} (1-H)}{C} \quad (5)$$

This delay bound is used in the cost function.

### B. Convex Optimization

Putting the profit and cost components together, the provider's profit problem becomes:

$$\max_{\phi} \left\{ \sum_{i=1}^k p_i \phi_i C - \sum_{i=1}^k b_i D_i(\phi_i) \exp[\beta(D_i(\phi_i) - d_i)] \right\} \quad (6)$$

subject to the feasibility constraints previously described plus the constraints  $\phi_i > \bar{\alpha}_i$ ,  $i = 1 \dots k$ .

In the above expression,  $D_i(\phi_i)$  is given from Eq. 5 by substituting parameter  $C$  with  $\phi_i C$ , since we are dealing with a network element with multiple input classes each of which is allocated a portion  $\phi_i$  of the total  $C$  resources. Note also that  $D_i(\phi_i)$  is actually  $D_{max,i}(\phi_i)$ . In addition, note the last constraint  $\phi_i > \bar{\alpha}_i$ : this should always stand true due to the fact that whenever  $\phi_i \leq \bar{\alpha}_i$  we have  $Pr[Q(t) > q_{max}] = 1$ . This implies that we are in an unstable case and the queue would never be able to accommodate the incoming traffic. This

constraint is introduced, in order to avoid for the network to operate in this undesirable from a QoS perspective regime.

In the over-provisioned case (i.e., when  $\sum_i \bar{\alpha}_i \leq 1$ ), the solution of the optimization problem exists since we deal with a convex optimization problem. An outline of the proof follows.

It is shown next that the profit function is convex, which combined with the convexity of the feasibility set guarantees the existence of a global maximum (maybe at a boundary point). It is easy to see that the feasibility set is a polyhedron in  $\mathbb{R}_+^K$  and hence convex. We establish next the concavity of the profit function. Some algebra shows that the second derivative of  $D_i(\phi_i)$  is positive in the feasibility set and therefore  $D_i(\phi_i)$  is convex on  $\mathbb{R}$ . The cost function  $c_i(D_i)$  is convex and nondecreasing (see Figure 3). The cost function is then convex as a composition of a convex and nondecreasing function with a convex function. Therefore,  $c_i(D_i(\phi_i))$  is convex on  $\mathbb{R}$ . Further,  $\sum c_i(\phi_i)$  is convex on  $\mathbb{R}_+^K$ , since each component of the sum is convex. Finally,  $\sum p_i \phi_i C$  is linear, which together with the previous result establishes the concavity of the profit function.

The optimal solution can then be found using standard algorithms, like the Newton method and its variations. Note that we are dealing with a constrained optimization problem, which implies that appropriate methods need to be considered (e.g., a penalty or barrier function to relax the constraints [8]). Moreover, we can take advantage of the Karush-Kuhn-Tucker (KKT) conditions that are necessary and sufficient for primal-dual optimality of a convex optimization problem. The primal problem is translated to an equivalent, but easier to solve dual problem. The primal problem has solution  $\phi^*$ , while its dual has solution  $(\phi^*, \lambda^*)$ . The result that KKT conditions give is that the optimal solution lies on the hyperplane  $\sum_{i=1}^k \phi_i = 1$ . The proof and the KKT conditions are given next:

$$\sum_{i=1}^k \phi_i^* - 1 \leq 0 \quad (7a)$$

$$-\phi_i^* + \bar{\alpha}_i < 0, \forall i \in \{1 \dots k\} \quad (7b)$$

$$\lambda_i^* \geq 0, \forall i \in \{1 \dots k\} \quad (7c)$$

$$\lambda_1^* \left( \sum_{i=1}^k \phi_i^* - 1 \right) = 0 \quad (7d)$$

$$\lambda_{i+1}^* (-\phi_i^* + \bar{\alpha}_i) = 0, \forall i \in \{1 \dots k\} \quad (7e)$$

$$\frac{\partial \pi(\phi)}{\partial \phi_i} - (\lambda_1^* - \lambda_{i+1}^*), \forall i \in \{1 \dots k\} \quad (7f)$$

Since  $\frac{\partial \pi(\phi)}{\partial \phi_i} > 0$  and thus from (7f) it can be seen that  $\lambda_1^* - \lambda_{i+1}^* > 0$ . From the second complementary slackness condition (7e) we obtain  $\lambda_{i+1}^* = 0$ , since the constraint  $-\phi_i^* + \bar{\alpha}_i < 0$  always holds. Thus, we conclude that  $\lambda_1^* > 0$ , which together with the other complementary slackness condition (7d) gives that

$$\sum_{i=1}^k \phi_i^* - 1 = 0 \quad (8)$$

Thus, we can solve for  $\phi_k$  (or any other  $\phi_i$ ) and convert our constraint problem over  $k$  variables to a constrained one over  $k - 1$  variables, having eliminated the first requirement. The latter combined with a Sequential Quadratic Programming (SQP) algorithm implemented in Matlab releases the other constraints. The SQP algorithm is a generalization of Newton's method for unconstrained optimization. SQP finds the next step away from the current iterate after minimizing a quadratic approximation of the initial problem. For further details about SQP and its Matlab implementation the reader is referred to [15]. The advantage of using an iterative algorithm (like SQP) is that the proposed framework is scalable to any number of classes.

*Remark:* For the under-provisioned case, the problem is not particularly interesting, since the QoS constraints would be surely violated. Hence, the service provider would allocate resources according to the average traffic intensities; further, it is easy to see that the operation would not be profitable. Hence, this regime is not studied in this paper.

#### IV. PERFORMANCE EVALUATION

In this section, we evaluate our pricing model in the over-provisioned case with a numerical case study. It is assumed that there are two types of service classes and the profit function becomes:

$$\begin{aligned} \pi(\phi_1, \phi_2) = & p_1\phi_1C + p_2\phi_2C \\ & - b_1D_1(\phi_1)e^{\beta(D_1(\phi_1)-d_1)} \\ & - b_2D_2(\phi_2)e^{\beta(D_2(\phi_2)-d_2)} \end{aligned} \quad (9)$$

where

$$D_i(\phi_i) = \frac{(\phi_i - \bar{\alpha}_i)^{\frac{H_i}{H_i-1}} (k\sigma_i)^{\frac{1}{1-H_i}} H_i^{\frac{H_i}{1-H_i}} (1 - H_i)}{\phi_i}, i = 1, 2 \quad (10)$$

Hence, we have to solve the optimization problem:

$$\begin{aligned} \max_{\phi} \pi(\phi_1, \phi_2) \text{ subject to} \\ \phi_1 + \phi_2 = 1 \\ \phi_1 > \bar{\alpha}_1 \\ \phi_2 > \bar{\alpha}_2 \end{aligned} \quad (11)$$

The parameters of the profit function used in the study are shown in Table I. Its concavity over both arguments is shown graphically in Figure 4, while over the first argument in Figure 5, by substituting  $\phi_2 = 1 - \phi_1$ .

In Tables II, III the optimal solution is shown when the arrival rate and the price coefficients are varied. In Table II it can be seen that with equal arrival rates and all the other parameters the same, the optimal solution allocates the resource equally amongst the two classes, as expected. On the other hand, the class with the higher arrival rate is allocated a larger portion of the resources, especially if the system is not too stressed (see rows 2 and 3 in the Table). In that situation the profit does not also fluctuate much. Finally, when the system becomes stressed (last row in the Table) the class with higher

TABLE I  
PARAMETERS FOR TWO DIFFERENT CLASSES

	Class 1	Class 2
$p$ (cents/Mbps)	1	1
$b$ (cents/ms)	0.1	0.1
$d$ (in delay units)	0.01	0.01
$QoS(= \epsilon)$	$10^{-6}$	$10^{-6}$
$\bar{\alpha}$ (normalized to $C$ )	0.2	0.2
$\sigma$ (normalized to $C$ )	0.01	0.01
$H$	0.70	0.70

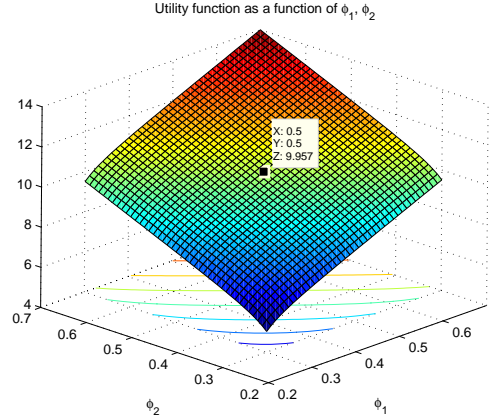


Fig. 4. Concavity of our Utility Function as a function of  $\phi_1, \phi_2$

arrival rate gets a higher proportion, but the overall profit for the provider decreases substantially, since violations of the SLA occur more often and therefore a large cost is incurred.

In Table III, the price coefficient varies, while all other parameters are held fixed (see Table I). Again, with equal prices we obtain equal allocations, while the allocation of resources exhibits a strong sensitivity to the price ratio  $p_1/p_2$ .

#### V. CONCLUSION

In this paper, we have studied a pricing scheme for next generation multiservice networks. An optimization problem based on a nonlinear pricing model was formulated, whose solution yields the optimal resource allocation in a network/service node, given the QoS requirements of each service class that the network element serves. Our non-linear pricing model responds well to changes of the characteristics in the input

TABLE II  
CHANGING THE ARRIVAL RATES  $\bar{\alpha}_i$

$(\bar{\alpha}_1, \bar{\alpha}_2)$	$(\phi_1^*, \phi_2^*)$	$\pi(\phi_1^*, \phi_2^*)$
(0.2, 0.2)	(0.5, 0.5)	9.96
(0.3, 0.2)	(0.5421, 0.4579)	9.93
(0.4, 0.2)	(0.5873, 0.4127)	9.89
(0.4, 0.5)	(0.4516, 0.5484)	6.72

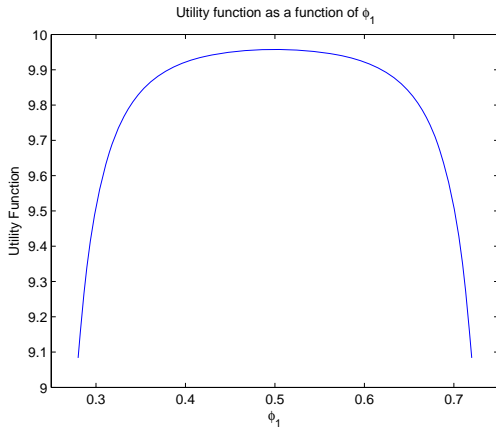


Fig. 5. Concavity of our Utility Function as a function of  $\phi_1$

TABLE III  
CHANGING THE PRICING FACTOR  $p_i$

$(p_1, p_2)$	$(\phi_1^*, \phi_2^*)$	$\pi(\phi_1^*, \phi_2^*)$
(1, 1)	(0.5, 0.5)	9.96
(2, 1)	(0.6917, 0.3083)	16.52
(4, 1)	(0.7183, 0.2817)	30.69
(4, 4)	(0.5, 0.5)	39.96
(1, 2)	(0.3083, 0.6917)	16.52
(1.5, 6)	(0.2739, 0.7261)	46.52
(4, 8)	(0.276, 0.724)	67.90

traffic, pricing parameters and QoS requirements. Further, the resulting convex optimization problem can easily and efficiently be solved using standard iterative methods and hence the proposed modeling framework approach is scalable to any number of service classes.

#### ACKNOWLEDGMENT

The authors would like to thank Dr. Peng Xu for useful discussions on the subject. The work of GM was supported in part by NSF grants CCR-0325571 and DMS- 0505535.

#### REFERENCES

- [1] Xu P., Devetsikiotis M., Michailidis G., *Profit-oriented Resource Allocation Using Online Scheduling in Flexible Heterogeneous Networks*. Telecommunication Systems, pp 289-303, 2006.
- [2] Xu P., *QoS Provisioning and Pricing in Multiservice Networks: Optimal and Adaptive Control over Measurement-based Scheduling*. PhD Dissertation, North Carolina State University, 2005.
- [3] Oltsik J., *Web Services Meet the Network*. [Online.] Available: <http://library.theserverside.com/detail/RES/114227381Q.440.html>.
- [4] Sun Q., Pleich R., Sauerwein R., *On-Line Measurement and Analysis of Fractional Brownian Traffic*. Proceedings of the IEEE Conference on High Performance Switching and Routing, pp 395-400, 2000.
- [5] Fonseca N., Mayor G. S., Neto C. A. V., *On the Equivalent Bandwidth of Self-Similar Sources*. ACM Transactions on Modeling and Computer Simulation, Vol. 10, No. 2, pp 104-124, April, 2000.
- [6] Norros I., *The Management of Large Flows of Connectionless Traffic on the Basis of Self-similar Modeling*. In Proceedings of IEEE International Conference on Communications, pp 451-455.
- [7] Courcoubetis C., Weber R., *Pricing Communication Networks*. John Wiley & Sons, 2003.

- [8] Rardin R. L., *Optimization in Operations Research*. Prentice Hall, 1998.
- [9] Kalyanasundaram S., Chong E. K., Shroff N. B. *Optimal Resource Allocation in Multiclass Networks with User Specified Utility Functions*. The International Journal of Computer and Telecommunications Networking, pp 613-630, April, 2002.
- [10] Chandra A., Gong W., Shenoy P. *Dynamic Resource Allocation for Shared Data Centers Using Online Measurements*. In Proceedings of ACM/IEEE Intl Workshop on Quality of Service, pp 381-400, 2003.
- [11] Knightly E., Shroff N., *Admission Control for Statistical QoS: Theory and Practice*. IEEE Network 13(2), pp 20-29, 1999.
- [12] Qiu J., Knightly E., *Measurement-based Admission Control with Aggregate Traffic Envelopes*. IEEE/ACM Transactions on Networking, pp 56-70, 1997.
- [13] Georgiadis L., Guerin R., Peris V., Sivarajan K. N., *Efficient Network QoS Provisioning based on Per Node Traffic Shaping*. IEEE/ACM Transactions on Networking 4(4), pp 482-501, 1996.
- [14] Breslau L., Jamin S., Shenker S., *Comments on the Performance of Measurement-based Admission Control Algorithms*. Proc. of IEEE INFOCOM, pp 1233-1242, 2000.
- [15] The MathWorks, Inc., *Sequential Quadratic Programming (SQP)*. [Online.] Available: <http://www.mathworks.com/access/helpdesk/help/toolbox/optim/ug/index.html?access/helpdesk/help/toolbox/optim/ug/t26622.html&http://www.google.com/search?hl=en&sa=X&oi=spell&resnum=0&ct=result&cd=1&q=sequential+quadratic+programming&spell=1>
- [16] Callaway R., Rodriguez A., Devetsikiotis M., Cuomo G., *Challenges in Service-oriented Networking*. GLOBECOM, 2006.
- [17] Leland W., Taqqu M., Willinger W., Wilson D., *On the Self-Similar Nature of Ethernet Traffic (Extended Version)*. IEEE/ACM Trans. Networking, pp 115, Feb, 1994.
- [18] Alcatel White Paper, *Alcatel Triple Play Service Delivery Architecture*. [Online.] Available: [http://www1.alcatel-lucent.com/tripleplay/docs/19786\\_TPSDA\\_wp.pdf](http://www1.alcatel-lucent.com/tripleplay/docs/19786_TPSDA_wp.pdf).
- [19] Cisco Systems, *Cisco AON: A Network Embedded Intelligent Message Routing System*. [Online.] Available: [http://www.cisco.com/en/US/products/ps6438/prod\\_bulletin0900aecd802c201b.html](http://www.cisco.com/en/US/products/ps6438/prod_bulletin0900aecd802c201b.html).
- [20] Savagaonkar U., Chong E. K., Givan R. L., *Online pricing for bandwidth provisioning in multi-class networks*. Comput. Networks, pp 835-853, 2004.