Robust Cross-Layer Network Optimization for Diverse QoS Requirements: Work in Progress

V. Marbukh

***Abstract*—Conventional approach to cross-layer network optimization assumes elastic users adjusting their bandwidth requirements in response to the resource congestion prices. This assumption leads to Network Utility Maximization (NUM) framework with Lagrange multipliers associated with resource capacity constraints playing role of the congestion prices. However, often users can more naturally quantify their preferences in terms of the rate and high-level Quality of Service (QoS) requirements rather than networks level end-to-end bandwidth requirements. This paper suggests that replacing resource capacity constraints with constraints on the feasible QoS parameters may lead to cross-layer network optimization framework with elastic users adjusting their diverse QoS requirements directly in response to the QoS-sensitive prices. We illustrate the proposed framework on examples of end-to-end bandwidth allocation subject to the worst-case scenario and average end-to-end delay.**

***Index Terms*—cross-layer system optimization, quality of service, pricing.**

I. Introduction

Aggregate utility maximization as a goal for balancing competing user-level requirements has been proposed in [1]. Distributed, price-based Network Utility Maximization (NUM) framework for achieving this goal in a case when users express their utilities in terms of the end-to-end network bandwidth has been developed in [2]. Framework [2] assumes that elastic users respond to resource congestion prices represented by Lagrange multipliers associated with the capacity constraints. Later, framework [2] was extended to cross-layer optimization of wire-line as well as wireless networks [3]-[4]. However, these and other extensions retain basic assumption of [2] that user utilities are expressed in terms of the end-to-end bandwidth.

In many situations it is more natural for users to quantify their preferences in terms of the high-level Quality of Service (QoS) parameters rather than low-level parameters such as end-to-end bandwidth, packet loss, etc. This paper suggests that distributed, pricing-based framework [2]-[4] can be extended to this more general case by assuming that the user utilities are functions of both network and user level parameters, and the goal is to maximize the aggregate utility which is the sum of the individual user utilities. Replacing capacity constraints with constraints on the feasible QoS parameters may allow direct pricing of the QoS parameters in terms of the Lagrange multipliers associated with the QoS parameter constraints.

Manuscript received March 17, 2011. This work was supported by NIST funding from the Comprehensive National Cybersecurity Initiative.

Author is with the Applied and Computational Mathematics Division, National Institute of Standards and Technology, Gaithersburg, MD 80305 USA (301-975-2235; fax: 301-975-3553; e-mail: marbukh@nist.gov).

To demonstrate flexibility of the proposed framework we consider two instances of the QoS-sensitive utility maximization. The first instance is end-to-end bandwidth allocation subject to the worst-case scenario end-to-end delay. The allocation is based on Generalized Processor Sharing (GPS) [7]-[8]. The second instance is end-to-end bandwidth allocation subject to the average end-to-end packet delay. This cross-layer optimization framework relies on $M/G/1$ conservation laws for each link and assumes statistical independence of the queues on different links [9]. Both instances assume users sensitive to average transmission rate and end-to-end packet delays, which explicitly enter the user utility functions. The first instance is inherently simpler than the second since the worst-case scenario delay for GPS-controlled traffic is only determined by GPS parameters and does not depend on the packet-level traffic structure. The proposed cross-layer optimization increases the aggregate utility as compared to the conventional NUM by taking advantage of the different user QoS requirements.

The paper is organized as follows. Section II introduces QoS-sensitive user utilities. In a case of users concerned with the worst-case scenario end-to-end delay, user utility is expressed in terms of the GPS parameters: token arrival rate and token buffer size. In a case of users concerned with the average end-to-end packet rate and delay, user utility is hybrid since while the average rate is determined by the flow-level model, the average packet delay is determined by the packet-level model. Section III demonstrates how QoS-sensitive pricing leads to distributed cross-layer aggregate utility maximization. Finally, Conclusion briefly summarizes and outlines directions of future research.

II. User Utilities

Subsections A and B introduce user utility of having average end-to-end bandwidth subject to the worst-case scenario and average end-to-end delay respectively.

A. Flow-level Utility

Network Utility Maximization (NUM) quantifies each user $s \in S$ preference for the end-to-end bandwidth x by

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

monotonously increasing, and typically concave, utility function

$$u_s = u_s(x_s) \,. \tag{1}$$

For example, a widely used weighted (α, w) - fair rate allocation [10] is based on utilities

$$u_s(x|\alpha, w_s) = \begin{cases} w_s\, x^{1-\alpha}/(1-\alpha) & if \quad \alpha \neq 1 \\ w_s \ln x & if \quad \alpha = 1 \end{cases} \tag{2}$$

where $\alpha, w_s > 0$ are parameters. When $w_s = 1$, the cases $\alpha \to 0$, $\alpha \to 1$ and $\alpha \to \infty$ correspond respectively to an allocation which achieves maximum throughput, is proportionally fair, and max-min fair. Case $\alpha = 2$ describes bandwidth allocation by TCP-Reno protocol.

Typically, bandwidth x_s is associated with user s end-to-end rate. However, in many cases it is impossible to guarantee certain bandwidth to the users due to traffic burstiness or variability of the link capacities. Scheduling disciplines, which take advantage of statistical multiplexing, are vulnerable to load fluctuations due to traffic burstiness or Denial of Service attacks when malicious users "grab" a disproportional amount of bandwidth by injecting extra traffic to cause performance deterioration for legitimate users. In a case of a wire-line network, link capacity variability may be due to limited reliability of the network elements. In a case of a wireless network, in addition to limited reliability, link capacity variability may be a result of wireless channel impairments and node mobility. In all these cases the user end-to-end bandwidth is likely to fluctuate. Since user queuing performance depends not only on the average available end-to-end bandwidth but also on this bandwidth variability, the problem is how to account for user preferences with respect to the end-to-end delay.

The rest of this Subsection discusses an approach to accounting for these preferences, which is suitable for mitigating effect of load fluctuations due to traffic burstiness or Denial of Service (DoS) attacks. This approach is based on Generalized Processor Sharing (GPS) discipline, which provides certain protection against misbehaving users. Given fixed set of S users sharing bandwidth C, each user $s = 1,.., S$ receives bandwidth

$$c_s = C\, \phi_s \delta_s \Big/ \sum_{i=1}^{s} \phi_i \delta_i \tag{3}$$

where $\delta_s = 1$ if user s traffic is backlogged and $\delta_s = 0$ otherwise. Fixed parameters ϕ_s in (3), where $\phi_s \geq 0$, $\sum \phi_s = 1$, $s = 1,.., S$, determine the guaranteed user bandwidth

$$g_s = C\phi_s \tag{4}$$

Bandwidth (4), which is guaranteed to user $s = 1,.., S$ regardless of (mis)behavior of other users $i = 1,.., S$, $i \neq s$, determines the worst-case scenario for the user QoS.

The following relation exists between user s average transmission rate $x_s < g_s$, maximum burst size B_s, and maximum delay τ_s:

$$\tau_s = \frac{B_s}{g_s - x_s} \tag{5}$$

Equation (5), which can be rewritten as follows:

$$x_s = g_s - B_s/\tau_s \tag{6}$$

also holds with τ_s being the average delay and B_s being the burstiness of user s traffic. Assuming that user s is elastic with respect to the average rate x_s with fixed parameters B_s and τ_s, user utility of having *average* end-to-end bandwidth x_s: $u_s(x_s)$ translates into the following utility of having guaranteed end-to-end bandwidth g_s:

$$\tilde{u}_s(g_s) = u_s(g_s - B_s/\tau_s) \tag{7}$$

B. Hybrid Flow/Packet-level Utility

In a store-and-forward wireless network due to link impairments and packet scheduling, instantaneous end-to-end bandwidth fluctuates in time. Adequate characterization of delay-sensitive user requirements can be obtained in terms of the available end-to-end bandwidth averaged over time interval of the order of user-specific tolerable delay. Thus, end-to-end bandwidth may have different meanings for different applications. To overcome difficulty of dealing with different time scales one may consider using a hybrid user utility function

$$u_s = u_s(y_s, \tau_s) \,, \tag{8}$$

which explicitly incorporates the user end-to-end flow-level transmission rate y_s as well as packet-level delay τ_s. Function $u_s(y_s, \tau_s)$ is typically assumed to be increasing in $y_s \geq 0$ and decreasing in $\tau_s \geq 0$.

Note the hybrid nature of utility (8), which incorporates both flow-level deterministic rate y_s as well as random packet-level delay τ_s. Due to this randomness, utility (8) is also random. The hybrid nature of utility (8) also reflects separation of time scales: while *individual packet* delays τ_s are mostly affected by comparatively "fast" throughput fluctuations, the average transmission rate y_s characterizes comparatively "slow" changing rate of the *stream of packets*. Assuming that distribution of end-to-end packet delays $P(\tau_s \leq \tau)$ changes on the "slow" time-scale, it is natural to characterize aggregated, flow-level satisfaction of user $s \in S$ by the utility

$$U_s = E[u_s(y_s, \tau_s)] \tag{9}$$

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

obtained by averaging utility (8) over distribution $P(\tau_s \le \tau)$. A hybrid flow/packet-level utility (9) can be derived from a packet-level utility model. In the rest of this section we briefly describe two such packet-level utility models.

The first model assumes that end-to-end delivery of a source s packet generates for this source utility $v_s(\tau_s) = v_{s0}\omega_s(\tau_s)$, where τ_s is this packet latency, utility of an instantaneously delivered packet is $v_{s0} = v_s(0)$, and lost in packet utility resulted from packet delay is $\omega_s(\tau) = v_s(\tau)/v_s(0)$, where function $\omega_s(\tau)$ monotonously decreases as τ increases from zero to infinity.

Consider K consecutive packets transmitted by source s during time interval $[t, t+\Delta)$. Defining the aggregate utility accumulated by source s as a result of transmitting these packets to be the sum of the utilities of individual packets, we obtain the following formula for the utility accumulation rate:

$$U_s = v_{s0}\frac{1}{\Delta}\sum_{k=1}^{K}\omega_s(\tau_{sk}) \tag{10}$$

where τ_{sk} is the end-to-end delay of the packet $k = 1,..,K$.

We assume that K is large: $K >> 1$, but not too large, so that the end-to-end network conditions do not significantly change during transmission of these packets, and thus random packet delays τ_{sk} have the same probability distribution for all $k = 1,..,K$. We also assume "sufficiently weak" correlation between delays τ_{sk} for different packets $k = 1,..,K$ to ensure applicability of the strong law of large numbers to the sequence of random end-to-end delays τ_{sk}. These assumptions allow us to approximate accumulation rate (10) as follows: $U_s \approx (K/\Delta)v_{s0}E[\omega_s(\tau_s)]$. Assuming that if packets are not delayed, i.e., $\tau_s = 0$, utility $U_s \approx (K/\Delta)v_{s0}E[\omega_s(\tau_s)]$ is consistent with flow-level utility $u_s(y_s)$, results in the following hybrid utility (8): $u_s(y_s, \tau_s) = u_s(y_s)\omega_s(\tau_s)$.

Second model assumes that while flow-level utility $u_s(y_s)$ quantifies user utility with negligible packet delays, user accumulates negative utility at a constant rate for each delayed packet, and thus the negative utility is proportional to the rate of the aggregate delay accumulation by source s packets. These assumptions result in the following particular case of average utility (9)

$$U_s(y_s, T_s) = u_s(y_s) - \gamma_s y_s T_s \tag{11}$$

where parameter γ_s characterizes user delay tolerance. A natural generalization of utility (11) is given by the following function:

$$U_s(y_s, T_s) = u_{s1}(y_s) - u_{s2}(y_s)T_s \tag{12}$$

III. System Utility Maximization

Consider a network, which consists of a set of sources $s \in S$ and set of links $l \in L$ with capacities c_l. Each source $s \in S$ identifies a single origin-destination pair and a set of feasible routes $r \in R_s$, each route r being a collection of links $l \in r$. A link-centric version of Network Utility Maximization (NUM) framework intends to maximize the aggregate user utility over flow vector $x = (x_{sr}, r \in R_s, s \in S)$, where $x_{sr} \ge 0$ is user s flow rate on a route r, and thus

$$x_s = \sum_{r \in R_s} x_{sr} \tag{13}$$

Subsection A considers a comparatively simple situation of flow-level utilities when SUM can be achieved with flow control and routing. Subsection B discusses a situation of hybrid flow/packet-level user utilities when system performance can benefit from packet scheduling which accounts for different user end-to-end delay requirements. Since this situation is more difficult to analyze, Subsection B demonstrates that at least in a particular case of hybrid user utilities, which deteriorate linearly with packet delay, the optimal scheduling can be determined explicitly.

A. Flow-level Utility

Consider the following version of link-centric NUM:

Maximize $$W(x) = \sum_{s \in S} u_s\left(\sum_{r \in R_s} g_{sr} - B_s/\tau_s\right) \tag{14}$$

subject to link l capacity constraints:

$$\sum_{s \in S}\sum_{r: l \in r} g_{sr} \le c_l \tag{15}$$

over $g_{sr} \ge 0,\ r \in R_s, s \in S$.

In a particular case of user utilities (2), corresponding to (α, w)-fair resource allocation, solution to optimization problem is

$$g_{sr} = \left(w_s \Big/ \sum_{l \in r} p_l\right)^{1/\alpha} + B_s/\tau_s \tag{16}$$

where Lagrange multipliers $p_l \ge 0$ are uniquely determined by capacity constraints (15).

In a particular case of hybrid flow/packet level user utilities (11) with all users having the same delay tolerance the aggregate utility is

$$W(x) = \sum_{s \in S} u_s\left(\sum_{r \in R_s} x_{sr}\right) - \gamma \sum_{s \in S}\sum_{r \in R_s} x_{sr} \sum_{l \in r} T_{srl} \tag{17}$$

where T_{srl} is the average delay experienced on a link l by a packet generated by source s and transmitted on a route r: $l \in r \in R_s$. Changing order of summation we can rewrite (17) as follows:

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$W(x)=\sum_{s\in S}u_s\left(\sum_{r\in R_s}x_{sr}\right)-\gamma\sum_{l\in L}y_lT_l \quad (18)$$

where the average link l load is

$$y_l=\sum_{s\in S}\sum_{r:l\in r}x_{sr} \quad (19)$$

and the average packet delay on link l is

$$T_l=\frac{1}{y_l}\sum_{s\in S}\sum_{r:l\in r}x_{sr}T_{srl} \quad (20)$$

Assuming that packets arrive at each link according to a Poisson process, all packet sizes are distributed exponentially with the same average and link capacities c_l and all rates are measured in packets per second we obtain the following expression for the average packet delay on a link l:

$$T_l=1/(c_l-y_l) \quad (21)$$

and thus the corresponding NUM takes the following form: Maximize

$$W(x)=\sum_{s\in S}u_s\left(\sum_{r\in R_s}x_{sr}\right)-\sum_{l}f_l\left(\sum_{s\in S}\sum_{r:l\in r}x_{sr}\right) \quad (22)$$

over $x_{sr}\geq 0,\ r\in R_s, s\in S$, where penalty functions are

$$f_l(y_l)=y_l/(c_l-y_l) \quad (23)$$

Average delay (23) realizes with First-In-First-Out (FIFO) packet scheduling on each link. Expression for the average end-to-end delay (23) implies approximate network decomposition into a system of jointly statistically independent M/M/1 queues on different network links. This decomposition has been verified by numerous simulations for networks with diverse routing where aggregate packet arrival on each link results from a large number of flows traversing this link [9].

In a particular case of user utilities (2), corresponding to (α,w)-fair resource allocation, solution to optimization problem (22) is

$$x_{sr}^*=\left(w_s\Big/\sum_{l\in r}p_l^*\right)^{1/\alpha},r\in R_s^*\subseteq R_s \quad (24)$$

and $x_{sr}^*=0$ for $r\notin R_s^*$, where set of minimum cost routes for source s is $R_s^*\subseteq R_s$, and the optimal link costs are

$$p_l^*=\left[\partial f_l(y_l)/\partial y_l\right]_{y_l=\sum_{s\in S}\sum_{r:l\in r}x_{sr}^*} \quad (25)$$

It can be shown that system of fixed point equations (24)-(25) has unique solution, which identifies optimal link cost p_l^* and flow allocation x_{sr}^* [10]. This solution can be found by a distributed, pricing-based algorithm.

B. Hybrid Flow/Packet-level Utility

We assume that set of sources $s\in S$ is comprised of J mutually exclusive subsets S_j, which include sources with the same delay sensitivity and represent Class-of-Service (CoS) $j=1,..,J$. A packet generated by source $s\in S_j$ losses its utility at rate γ_j with delay, where without loss of generality we assume

$$0\leq\gamma_1<..<\gamma_J \quad (26)$$

The aggregate utility is

$$W(x)=\sum_{s\in S}u_s\left(\sum_{r\in R_s}x_{sr}\right)-\sum_{s\in S}\gamma_s\sum_{r\in R_s}x_{sr}\sum_{l\in r}T_{srl} \quad (27)$$

where T_{srl} is the average delay experienced on a link l by a packet generated by source s and transmitted on a route r: $l\in r\in R_s$.

Changing order of summation we can rewrite (27) as follows:

$$W(x)=\sum_{s\in S}u_s\left(\sum_{r\in R_s}x_{sr}\right)-\sum_{l\in L}\sum_{j\in J}\gamma_j y_{lj}T_{lj} \quad (28)$$

where the average link l load by source $s\in S_j$ traffic is

$$y_{lj}=\sum_{s\in S_j}\sum_{r:l\in r}x_{sr} \quad (29)$$

and the average delay on link l for a packet from a source $s\in S_j$ is

$$T_{lj}=\frac{1}{y_{lj}}\sum_{s\in S_j}\sum_{r:l\in r}x_{sr}T_{srl} \quad (30)$$

Assuming (a) approximate network decomposition into a system of jointly statistically independent M/M/1 queues on different network links, and (b) flow control and routing operating on much slower time scale than scheduling allow us to decompose global scheduling into scheduling on each link:

$$f_l(y_{lj},j=1,..,J)=\min_{(T_{lj})}\sum_{j\in J}\gamma_j y_{lj}T_{lj} \quad (31)$$

subject to the corresponding conservation laws [9]. It is known [9] that solution to optimization problem (31) in the class of non-preemptive scheduling disciplines is as follows. Traffic within the same CoS, i.e., packets loosing utility at the same rate with delay, are scheduled for transmission according to First-In-First-Out discipline. Packets of higher class-of service j have higher scheduling priority.

It is known that the optimal average queuing delay for CoS $j=1,..,J$ on a link j is as follows [9]:

$$T_{lj}^*(x)=\left[1+\frac{\rho_l}{(1-\rho_l(j))(1-\rho_l(j+1))}\right]\frac{b}{c_l} \quad (32)$$

where link l utilization by CoS $j\geq i$ traffic is

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

$$\rho_l(i) = \frac{1}{c_l} \sum_{j \ge i} \sum_{s \in S_j} \sum_{r: l \in r} x_{sr} \tag{33}$$

and the aggregate link l utilization is $\rho_l = \rho_l(1)$.

Assuming that scheduling has been optimized on a fast time scale, the optimization problem for load allocation takes the following form:

$$\max_x \left[\sum_{s \in S} u_s \left(\sum_{r \in R_s} x_{sr} \right) - \sum_{l \in L} f_l(y_{lj}, j = 1,.., J) \right] \tag{34}$$

subject to constraints (29).

In a particular case of user utilities (2), corresponding to (α, w)-fair resource allocation, solution to optimization problem (34) for a source $s \in S_j$ is

$$x_{sr}^* = \left(w_s \Big/ \sum_{l \in r} p_{lj}^* \right)^{1/\alpha}, r \in R_s^* \subseteq R_s \tag{35}$$

and $x_{sr}^* = 0$ for $r \notin R_s^*$, where set of minimum cost routes for source s is $R_s^* \subseteq R_s$, and the optimal link costs are

$$p_{lj}^* = \left[\partial f(y_{l1},.., y_{lJ}) / \partial y_{lj} \right]_{y_{li} = \sum_{s \in S_i} \sum_{r: l \in r} x_{sr}^*} \tag{36}$$

It can be shown that system of fixed point equations (35)-(36) has unique solution, which identifies CoS-dependent optimal link cost p_l^* and flow allocation x_{sr}^*. This solution can be found by a distributed, pricing-based algorithm.

IV. Conclusion and Future Research

This paper suggests a possibility of NUM extension to explicitly include user QoS preferences. To demonstrate the benefits of optimized priority scheduling over FIFO scheduling consider a simple example of a single link of fixed capacity C serving two flows comprised of Poisson streams of packets. One flow is inelastic, i.e., has fixed rate x_1, and is comprised of delay-sensitive packets which can tolerate average delay τ_1 at most. Another, file transfer flow is elastic, i.e., can vary its rate x_2. Assuming that the average delay constraint for the delay-sensitive traffic is satisfied, compare utility $w_2 \log_e x_2$ derived by the elastic flow under FIFO and the optimal priority scheduling. For simplicity we assume that packets have exponentially distributed length with average one. Thus system ability to accommodate at least the inelastic flow imposes the following constraint $1/(C - x_1) \le \tau_1$.

Under FIFO scheduling, the delay constraint for the inelastic flow $1/(C - x_1 - x_2) \le \tau_1$ yields the maximum rate of the elastic flow $x_2 = C - x_1 - 1/\tau_1$ resulting in utility $w_2 \log_e (C - x_1 - 1/\tau_1)$. It is easy to see that the optimal scheduling gives non-preemptive priority to the inelastic flow. Under this priority scheduling, the maximum rate of the second flow is $x_2 = C - x_1$ resulting in utility $w_2 \log_e (C - x_1)$. Thus, the utility gain due to the optimal priority scheduling is

$$\Delta u = -w_2 \log_e \left[1 - \frac{1}{(C - x_1)\tau_1} \right] \ge 0 \tag{37}$$

In the extreme case of both flows being delay-insensitive: $\tau_1 \uparrow \infty$, packet scheduling does not improve performance: $\Delta u = 0$. In a case of delay-sensitive inelastic flow: $\tau_1 < \infty$, packet scheduling gain is positive: $\Delta u > 0$. If the inelastic flow occupies increasing portion of the link bandwidth: $(C - x_1)\tau_1 \downarrow 1$ the gain infinitely increases: $\Delta u \to \infty$.

Future research should evaluate accuracy of the proposed extension of the network decomposition from a case of FIFO scheduling [9] to a case of more general scheduling disciplines. If the accuracy is established, it would allow optimization of non-linear user level utilities with respect to the average end-to-end delays since generally this optimization yields dynamic scheduling. A distributed user-level utility maximization should be based on direct pricing of QoS requirements, which includes numerous research issues including a connection to Smart Market approach [11].

References

[1] S. Shenker, "Fundamental design issues for the future Internet," *IEEE JSAC*, 13 (1995) 1176-1188.

[2] F.P. Kelly, A.K. Maulloo, and D.H.K. Tan, "The rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, pp. 237-252, vol. 409, 1998.

[3] M. Chiang, "Balancing transport and physical layers in wireless multihop networks:jointly optimal congestion control and power control," *IEEE J. Sel. Areas Comm.*, Vol. 23, pp. 104-116, 2005.

[4] X. Lin, N.B. Shroff, R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Comm.*, Vol. 24, Issue 8, pp. 1452-1463, 2006.

[5] D. Tse and S. Hanly, "Multi-Access Fading Channels: Part I: Polymatroid Structure, Optimal Resource Allocation and Throughput Capacities", *IEEE Transactions on Information Theory*, v. 44, No. 7, Nov., 1998, pp. 2796-2815.

[6] S. Hanly and D. Tse, "Multi-Access Fading Channels: Part II: Delay-Limited Capacities", *IEEE Transactions on Information Theory*, v. 44, No. 7, Nov., 1998, pp. 2816-2831.

[7] A.K. Parekh, R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Transactions on Networking (TON)*, Vol. 1, Issue 3, pp. 344-357.

[8] A.K. Parekh, R.G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple-node case," *IEEE/ACM Transactions on Networking (TON)*, Vol. 2, Issue 2, pp. 137-150.

[9] L. Kleinrock, Queueing Systems, Volume II: Computer Applications, John Willey, 1976.

[10] Mo and J. Walrand, "Fair end-to-end window-based admission control," *IEEE/ACM Trans. on Networking*, No. 8, pp. 556-567, 2000.

[11] J. K. MacKie-Mason and H. R. Varian, *Pricing the Internet*, Kahin, Brian and Keller, James, 1993.

ISBN: 978-988-18210-6-5
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)