Top Percentile Network Pricing and the Economics of Multi-Homing

Joseph Levy,
Teva Pharmaceutical Industries, LTD, Netanya, Israel

Hanoch Levy,
School of Computer Science, Tel-Aviv University
Tel-Aviv, Israel

Yaron Kahana*
Intel - Broadband Wireless Division,
Petach-Tikva, Israel

Author for correspondence:
Hanoch Levy
School of computer science
Tel-Aviv University
Tel-Aviv, ISRAEL
Email: hanoch@cs.tau.ac.il
Tel: +972.54.876.276

* Part of the results of this work was reported in Levy, Levy and Kahana (2003). The work was done while J. Levy and Y. Kahana were with Comgates Ltd. and H. Levy was partially with Comgates Ltd

Abstract

Top-Percentile pricing is a relatively new and increasingly popular pricing policy used by network providers to charge service providers. In contrast to fixed cost pricing and to pure per-usage pricing, top-percentile pricing has not been studied. Thus the efficient design and operation of networks under top-percentile pricing is not well understood yet. This work studies top-percentile pricing and provides an analysis of the expected costs it inflicts on a service provider. In particular we use our analysis framework to investigate the popular multi-homing architecture in which an Internet Service Provider (ISP) connects to the Internet via multiplicity of network providers. An ISP that uses multi-homing is subject to extra charges due to the use of multiple networks. Important questions that are faced by such an ISP are what is an efficient routing strategy (as to reduce costs) and how large the costs are. We provide a general formulation of this problem as well as its probabilistic analysis, and derive the expected cost faced by the ISP. We numerically examine several typical scenarios and demonstrate that despite the fact that this pricing aims at the peak traffic of the ISP (similarly to fixed cost), the expected bandwidth cost of multi-homing is not much higher than that of single-homing.

Keywords: Pricing, Traffic-Engineering, Multi-Homing, Top-Percentile pricing.

Top-Percentile pricing is a relatively new and increasingly popular pricing policy used by network providers to charge service providers Odlyzko (2003). To apply top-percentile charges, the network provider measures the amount of data sent at fixed intervals (say 5 minutes). It then evaluates these values for all the intervals over the charge period (8640 intervals per month, in the case of 5 minute intervals) and selects the traffic volume of the *top q-percentile* interval as the basis for computing the cost. For example, if top 5-percentile pricing is used (which may be called in the industry as 95-percentile pricing) then the cost is based on the traffic volume of the top 432*th* interval.

Traditionally, two types of pricing methods have been dealt with in the telecommunications industry and research: 1) *Fixed price,* and 2) *Variable (per usage) price.* In the former the buyer pays a fixed price for the month *regardless* of the amount of traffic shipped de-facto over the month. In the latter the buyer pays for the actual traffic shipped over the month. These two traditional policies have been thoroughly investigated in past studies. The fixed price[1] policy can be found mainly in the literature of network design in which the network designer pays fixed price for the "pipes", at the stage of network construction. Examples go back many years, e.g., Kleinrock (1976), and the studies referenced there. Recent examples include Herzberg and Shleifer (1999) in the context of designing reliable networks. The variable price policy is being dealt with mainly in routing problems (where one attempts to route the traffic as to minimize cost) or in the pricing of services/calls. Some examples of the former are Altman et. al.(2000), Wang and Schulzrinne (2001). Examples of the latter can be found in Paschalidis and Tsitsiklis(2000) where a charge is done on a call by call basis where price of a call is determined as function of the current state of the network and the call type and in Kelly (1997) where an application is charged, at real time, based on the call duration, the amount of bandwidth and the connection. Other studies such as Gibbens and Kelly (1998) offer even more fine-grain pricing where an application is charged by charging each packet individually as a function of the temporal network congestion. Other studies of per-usage can be found in Mackie-Mason and Varian (1995) where prices must be associated with packets and recorded in them. Pricing based on priorities can be accommodated under both fixed pricing (e.g the Paris-Metro Pricing proposed by Odlyzko (1997)) and per-usage pricing (see, e.g. Cocchi et. al. (1993, 1991) and Gupta, Stahl and Whinston (1997)). While most pricing approaches can be classified to either of the two categories (fixed price or per usage price), Shenker et. al. (1996) claim that the distinction between the two categories should not be sharp and that there is a continuum between these extremes. An example is Clark (1997) where the users are charged per their expected usage (which they must declare ahead of time) and policing methods must be used to control the actual usage. Tutorials that review various pricing policies (in which the reader may find more references) are Falkner, Devetsikiotis and Lambadaris (2000), Songhurst (1999) and Shenker et. al. (1996).

In contrast to the vast research conducted on those policies, very little research has been conducted on top-percentile pricing. As a result, the issues of network design and traffic engineering may be very hard to conduct under the environment of top-percentile pricing.

The aim of this work is to start addressing and study the top-percentile pricing and examine how efficient use of resources can be done under this policy. In this paper we choose to investigate it in the context of the Multi-Homing environment, which we believe, demonstrates the important tradeoffs associated with top-percentile pricing. Multi-homing (Orda and Rom (1990) is a popular architecture used by *Internet Service Providers* (ISP's) to connect to the Internet via multiple *network providers* (backbones). This connectivity improves the network reliability and quality of the ISP, since when one of the networks fails and its quality degrades the ISP can use the alternate network. While multi-homing improves the ISP's experienced QOS, it increases its inflicted costs and thus may make it economically inefficient. Top-percentile pricing resembles somewhat the fixed price policy as it charges for one of the largest volume intervals, and thus if one does not use the network for 94% of the time, one still pays as if one used it for all the time. As such, top-percentile may increase the incurred costs drastically and there is a question whether multi-homing is at all economically viable under this pricing.

In the context of Multi-homing we aim at examining the bandwidth costs inflicted on an ISP and the economical viability of the multi-homing concept. This, as mentioned previously, depends on the cost structure used by the network providers. Under the traditional *fixed-cost* pricing policy , in which the customer pays for a fixed capacity regardless of how much of it is being actually used, the cost of dual-homing (a special case of multi-homing with two network connections) is twice the cost of a single connection and thus may be too expensive. Under the pure *per-usage pricing,* there is hardly extra cost inflicted on the multi-homing architecture, since on each of the networks the customer pays only for the bytes transferred. There is, thus, an open question to which of these two, top-percentile pricing resembles. In other words, is dual-homing economically viable under top percentile or not. Note that while the multi-homing design problem is relatively simple in the context of fixed pricing or per-usage pricing, it is much more complicated in the context of top-percentile pricing. In fact, even the "simple" viability question seems to be non-trivial and is not well understood yet.

To address these questions we recognize that the costs inflicted by top-percentile pricing strongly depend on the statistical structure of the traffic demands. We therefore develop a probabilistic model that reflects the stochastic nature of traffic. Specifically, in Section 1 we build a general probabilistic model that accounts both for the stochastic volume of traffic streams and for the probabilistic nature of their placement on particular networks. The latter reflects the fact that network conditions (failures) and quality behave stochastically, which causes the routing decisions (that follow those conditions) to behave stochastically. Accounting for both we then provide the mathematical analysis of the model and derive the expected costs incurred on these networks. Then, in Section 2 we analyze the multi-homing problem. In particular we examine several strategies for routing the traffic in this environment. We propose to use special cases of the analysis developed in Section 1 (or slight variation on that analysis) in order to derive the expected cost incurred for these strategies under top-percentile pricing. We further derive upper and lower bounds on the cost incurred in the system. These can be used as reference points for evaluating the quality of the examined strategies. In Section 3 we use the analysis to examine several numerical examples. The examination reveals that the top-percentile pricing inflicts much lower costs than the fixed pricing. Accounting for bandwidth cost - the multi-homing cost inflicted on the ISP is higher than the non-multi-homing cost only by several percents (as opposed to doubling it under fixed cost structure). As such we conclude that multi-homing is economically viable (unless the cost structure contains significant fixed price components). The examples also provide insights into which type of routing strategy one should use in a multi-homing environment under top-percentile pricing. The analysis methodology developed in this work and the insight derived from the examples can be further used in more general traffic engineering and network planning frameworks. Concluding remarks are provided in Section 4.

1. Mathematical formulation

1.1 Preliminaries and notation

Let $L_1,\ldots,L_J$ be the set of network providers. Assume that the charge period of a network provider is divided into T intervals of equal length for the purpose of top-percentile charge calculation (for simplicity, assume that T is the same for all providers). The pricing policy is called a *top q-percentile* pricing, where $0 \le q \le 100$, if the network provider charges based on the q-percentile interval. Specifically, the network provider calculates the traffic shipped through the network during each of the T intervals. Then, the cost inflicted on the customer (normally a service provider) is determined by the volume of traffic shipped at the top q percentile interval. For example, in a typical situation each of the intervals 15 minutes long and the number of intervals in the month is 2880. If top 5% pricing is used (which is called commercially 95 percentile pricing) then the top 144$^{\text{th}}$ interval traffic forms the charge basis. In the broad framework, let q_i be the percentile used by network L_i and c_i be the cost per bit charged by L_i. Let $Y_i(t)$,$t=1, \ldots, T,$ denote the total traffic shipped over network L_i at interval *t*.. We will sort the vector $Y_i(t)$ in increasing order: Let $t_1^i, t_2^i, \ldots, t_T^i$ be a permutation on $1,...,T$, such that $Y_i(t_1^i) \le Y_i(t_2^i) \le ... \le Y_i(t_T^i)$. Let $r_i = q_i T/100$, and assume r_i is an integer. Using this formulation the monthly cost charged by L_i is $C_i = c_i Y_i(t_{T-r_i+1}^i)$.

A set of I traffic demands is a pair (X,π), where $X = \{X_i(t), i = 1,\ldots,I, t = 1,\ldots T\}$ is a collection of positive independent random variables, and $\pi = \{\pi_{ij}, i = 1,\ldots,I,\ j = 1,\ldots J\}$, such that $\pi_{ij} \ge 0$ for all i and j, and $\sum_{j=1}^{J} \pi_{ij} = 1$ for all i. The variables $X_i(t)$ represent the i[th] traffic demand at time interval t, and π_{ij} is interpreted as the probability (or proportion of time) that the i[th] traffic demand is routed via the j[th] network provider. Also let $U_{ij}(t)$ be a random variable that takes the values 1 and 0 with probabilities π_{ij} and $1-\pi_{ij}$, respectively.

Assume, without loss of generality, that the set of T intervals is partitioned into K subsets, where each subset consists of T_k consecutive intervals and $\sum_{k=1}^{K} T_k = T$, and that for$1 \le k \le, K$ $\left\{X_i(t),\quad t = \sum_{j=1}^{k-1} T_j + 1,\ldots,\sum_{j=1}^{k} T_j\right\}^2$ have a common probability distribution function $F_i^k(\cdot)$. The latter assumption stems from practical considerations by which typically one is not equipped with different statistical information for each $X_i(t)$ (for each short time interval) but rather with more general statistics (e.g the amount of traffic between the hours 8-12 or 12-16).

We will denote the combined traffic demand on network provider j at time interval t by $D_j(t)$, , , and $G_j^k(x) = P(D_j(t) \le x)$, where $\sum_{l=1}^{k-1} T_l < t \le \sum_{l=1}^{k} T_l$, is its probability distribution function. For fixed j, $D_{(r)}$ is the r-largest of $D_j(t),\quad t = 1,\ldots,T$.

$\Phi(x \mid \mu,\sigma^2)$ is the Normal distribution with mean μ and variance σ^2, $B(\cdot \mid n,p)$ is the cumulative distribution function of a Binomial random variable with parameters n and p.

1.2 Model and Assumptions

To account for a general stochastic model, we model “traffic demands” in the general form of (X,π), although a network structure is not assumed, and all "demands" effectively have the same set of routes available, thus any traffic can be routed across any provider. In some typical practical scenarios, a service provider faces one aggregate demand and splits it into several traffic demands, which can go across several available routes.

However, once the splitting is done, the model of “traffic demands” is valid. The splitting procedure and possible strategies for optimal splitting are beyond the scope of this paper.

1.3 The distribution of combined traffic flows

$D_j(t)$ can be expressed using the random variables $U_{ij}(t)$ and the demand variables $X_i(t)$, since the event $\{U_{ij}(t) = 1\}$ means that during time interval t traffic demand i is routed through service provider j. Using the independence of $U_{ij}(t)$, one can write $D_j(t) = \sum_{i=1}^{I} X_i(t) \cdot U_{ij}(t)$.

Assertion 1:

For all $x \ge 0$,

(Eq. 1) $$G_j^k(x) = \sum_{u_{1j}=0}^{1} \cdots \sum_{u_{Ij}=0}^{1} \left[\prod_{i=1}^{I} \left((2\pi_{ij} - 1) \cdot u_{ij} + 1 - \pi_{ij} \right) \right] \cdot \left(\mathop{*}_{i=1}^{I} u_{ij} F_i^k(x) \right),$$

where $*$ is the convolution operation.
Proof is straightforward, by conditioning on $U_{ij}(t)$.

Corollary 1:
Assume that for all i and k, $F_i^k(x) = \Phi(x \mid \mu_{ik}, \sigma_{ki}^2)$. Then, knowing that a linear combination of independent Normal random variables is also Normally distributed:

(Eq. 2) $$G_j^k(x) = \sum_{u_{1j}=0}^{1} \cdots \sum_{u_{Ij}=0}^{1} \left[\prod_{i=1}^{I} \left((2\pi_{ij} - 1) \cdot u_{ij} + 1 - \pi_{ij} \right) \right] \cdot \Phi\left(x \mid \sum_{i=1}^{I} u_{ij} \mu_{ik}, \sum_{i=1}^{I} u_{ij} \sigma_{ik}^2 \right).$$

1.4 The distribution of the traffic flow top-percentile

Our aim is at computing the *expected cost* of the customer. Thus, once the combined traffic demand function over a network provider is calculated, the expected value of its top percentile needs to be calculated.

Assertion 2:
For all $y \geq 0$, $P\left(D_{(r)} \leq y\right) = 1 - \mathop{*}_{k=1}^{K} B\left(r-1 \mid T_k, G^k(y)\right)^3$.

Proof: Fix $y \geq 0$ and for each $1 \leq k \leq K$, define $N_k^y = \#\left\{ D(t) \leq y, \sum_{l=1}^{k-1} T_l < t \leq \sum_{l=1}^{k} T_l \right\}$. First observe that $D_{(r)} \leq y$ if and only if $\sum_{k=1}^{K} N_k^y \geq r$. Secondly, note that N_k^y is a Binomial random variable with parameters T_k and $G^k(y)$. Then

(Eq. 3) $$P\left(D_{(r)} \leq y\right) = P\left(\sum_{k=1}^{K} N_k^y \geq r \right) = 1 - P\left(\sum_{k=1}^{K} N_k^y \leq r - 1 \right) = 1 - \mathop{*}_{k=1}^{K} B_k\left(r-1 \mid T_k, G^k(y)\right).$$

QED.

Now, the expected value of $D_{(r)}$ is given by:

(Eq. 4) $$E[D_{(r)}] = \int_0^\infty y \, dP\left(D_{(r)} \leq y\right)$$

1.5 Implementation Considerations

If $T_1, \ldots, T_K$ are sufficiently large, the distribution of the Binomial random variables $N_1^y, \ldots, N_K^y$ can be approximated by the Normal distribution, i.e. N_k^y is approximately Normal with mean $T_k G^k(y)$ and variance $T_k G^k(y)(1 - G^k(y))$. Further more, since $N_1^y, \ldots, N_K^y$ are independent, their sum can be further approximated as Normal with mean and variance:

(Eq. 5) $$m(y) = \sum_{k=1}^{K} T_k G^k(y) \; ; \; v(y) = \sum_{k=1}^{K} T_k G^k(y)(1 - G^k(y))$$

Therefore, with standard continuity correction we have:

(Eq. 6) $$P\left(D_{(r)} \leq y\right) \approx \Phi(r - 0.5 \mid m(y), v(y)).$$

The expected value of $D_{(r)}$ can be calculated using this approximation and Equations 4, 5 and 6.
The complexity of carrying our these computations is $O((I2^I + K)|X|)$ where $|X|$ is the number of values for which $G(x)$ is evaluated.

2. Multi Homing, Operation Policies and Performance Bounds

Under multi-homing the service provider receives service from more than one networks; a special case, which is very popular and thus of interest, is the dual-homing where the service is received from two networks. Under this setting the service provider now faces the question of where to direct its traffic in order to reduce its expenses. An operation policy is a policy that decides where to direct the traffic.

A general way to model this problem is to assume that the provider faces K streams of traffic demands $X_1,\dots,X_K$ where X_i denotes the ith stream and $X_i(t)$, $i=1,\dots,K,\ t=1,\ \dots,T$ is the traffic demand of stream i at time interval t. The provider can route the demands through M networks, denoted $L_1,\dots,L_M$ .In our discussion below we will focus on $M=2$, which is the most common case for multi-homing. Further, under this assumption it also makes sense to limit the discussion to two traffic demand sets X_1 and X_2 (since larger number of demands can be treated by combining them into two traffic demands. Thus we deal with the problem of selecting the route L_1 or L_2 to each of the traffic demands $X_i(t)$, $i=1,2,\ t=1,\ \dots,T$. In this context, a *routing policy R* is a route assignment policy $R=\{R_i(t)\}$, $i=1,2,\ t=1,\ \dots,T$, where $R_i(t)\in\{1,2\}$ (where 1 and 2 denote networks L_1 and L_2, respectively). Let q_1, q_2 be the percentile pricing parameters used in L_1, L_2 respectively. Let us assume that $q_1 = q_2$ and denote it by q. We will assume that $r=qT/100$ is a whole integer.

Two special case routing policies, the *identical primary* policy and the *different primary* policy are presented in Section 3.4. Below, an idealistic optimal routing policy is presented in subsection 2.1. This is used to derive a lower bound on the cost of any policy. An upper bound for the cost is derived in Section 2.2, and the bounds are discussed in Section 2.3.

2.1 Optimal Assignment Under No Failures – A lower Bound on Cost

In this section we consider a situation where there are failures neither on L_1 nor on L_2, and examine what is the optimal routing policy. Specifically we assume that both L_1 and L_2 utilize top-q percentile pricing using the same parameter q and the same parameter T, and that $r=qT/100$ is an integer. Assuming that the values of the variables $X_i(t)$, $i=1,2,\ t=1,\ \dots,T$ are known deterministically and that there are **no failures**, the question is what is the route assignment of A, $R_i^A(t)$, $i=1,2,\ t=1,\ \dots,T$, such that the service provider's cost is minimized. Note that this is an offline optimization where the full knowledge of the values is known.
Let $X(t)=X_1(t)+X_2(t)$, $t=1,\ \dots,T$ be the total traffic at time t. Without loss of generality assume that $X(1)\le X(2)\le\dots\le X(T)$.

Let $Y_i^A(t)$ $i=1,2,t=1,\ \dots,\ T$, denote the total traffic shipped over network i at interval t, under the assignment policy A. Let C^A be the overall cost charged under policy A. Following the formulation given in Section 1.1, the cost is given by: $C^A = c_1 Y_1^A(t_{T-r+1}^1) + c_2 Y_2^A(t_{T-r+1}^2)$, where $\{t_i^1\}$ and $\{t_i^2\}$ are sorted according to the values of Y_1^A and Y_2^A respectively (note that the sequences $\{t_i^1\}$ and $\{t_i^2\}$ are dependent on *A;* to simplify notation we avoid adding a subscript *A* to them).

Assertion 3: There exists an optimal policy A^*, such that for all *i*, $T-r+1<i\le T$, $R_1^{A^*}(t_i^1)=R_2^{A^*}(t_i^1)=1$.
The proof is given in the appendix.

Assertion 4: There exists an optimal policy A^* (possibly different from A^* of Assertion 3), such that for all *t*, $T-r+1<i\le T$,. $R_1^{A^*}(t_i^2)=R_2^{A^*}(t_i^2)=2$.

The proof is similar to that of Assertion 3.

Theorem 1: Let A^* be an optimal assignment policy. Let S be the set of time intervals that do not belong to the sets $\{t_i^1\},\{t_i^2\}$, $T-r+1<i\le T$., that is $S=\{1,..,T\}-(\{t_i^1\}\bigcup\{t_i^2\})$. Let $\tau\in S$ be an interval for which $X(\tau)\ge X(t), \forall t\in S$. Then $C^{A^*}=\min\{c_1,c_2\}X(T-2r+2)$.

Proof:

(1) First, we show that $C^{A^*}\ge\min\{c_1,c_2\}X(\tau)$. Let us examine the routing at τ: a) If $R_1^{A^*}(\tau)=R_2^{A^*}(\tau)=1$ then $X(\tau)\le X(t_i^1)$, $T-r+1<i\le T$ (by definition of $\{t_1^i\}$), thus the cost charged by L_1 is bounded from below by $c_1X(\tau)$ and the cost charged by L_2 is bounded from below by 0. The overall cost is therefore bounded from below by $c_1X(\tau)$. b) Similarly, if $R_1^{A^*}(\tau)=R_2^{A^*}(\tau)=2$ then the overall cost is therefore bounded from below by $c_2X(\tau)$. c) If $R_1^{A^*}(\tau)=1, R_2^{A^*}(\tau)=2$ then in a similar manner the cost charged by L_1 is bounded from below by $c_1X_1(\tau)$ and the cost charged by L_2 is bounded from below by $c_2X_2(\tau)$, and the overall cost is bounded from below by $c_1X_1(\tau)+c_2X_2(\tau)$. d) Similarly, if $R_1^{A^*}(\tau)=2, R_2^{A^*}(\tau)=1$ then the overall cost is bounded from below by $c_1X_2(\tau)+c_2X_1(\tau)$. Now, the expressions bounds derived in all four cases are bounded from below by $\min\{c_1,c_2\}(X_1(\tau)+X_2(\tau))$ from which the proof of (1) follows.

(2) Second, we show that there exists a policy A^{**} for which $C^{A^{**}}=\min\{c_1,c_2\}X(T-2r+2)$. Assume, without loss of generality that $c_1\le c_2$. Then A^{**} is constructed as follows:. For $1\le t\le T-2r+2$ and for $T-2r+3\le t\le T-r+1$ (which are consecutive sets of intervals) $R_1^{A^{**}}(t)=R_2^{A^{**}}(t)=1$, thus L_1 charges $c_1X(T-2r+2)$. For $T-r+2\le t\le T$, $R_1^{A^{**}}(t)=R_2^{A^{**}}(t)=2$, thus L_2 charges $c_2\cdot 0=0$, and the overall cost is $C^{A^{**}}=\min\{c_1,c_2\}X(T-2r+2)$.

(3) Next, we show that $X(\tau)\ge X(T-2r+2)$. Assume, for the sake of contradiction, that it does not hold, that is, $X(\tau)<X(T-2r+2)$. Let the rank of t be defined as $rank(t)=|\{t',X(t')\le X(t)\}|$. Now, since the set of time intervals consists of *{1,..,T},* then for every *i, j*_if $X(i)<X(j)$ then *rank(i)<j.* In particular, if $X(\tau)<X(T-2r+2)$ then $rank(\tau)<T-2r+1$. However, since τ obeys (by definition) $X(\tau)\ge X(t), \forall t\in S$ and since

$|S|=T-2r+2$, we must have $rank(\tau)\geq T-2r+2$, which forms a contradiction Thus, $X(\tau)\geq X(T-2r+2)$.

Lastly, the properties proved in (1), (2) and (3) together imply $C^{A*}=\min\{c_1,c_2\}X(T-2r+2)$.
QED.

Corollary 2: An optimal routing policy $A*$ is in the form of (or similar to) $A**$ presented in Theorem 1 item (2) and the optimal cost is given by $C^{A*}=\min\{c_1,c_2\}X(T-2r+2)$.

Corollary 3: The expected cost of *A**, namely $\min\{c_1,c_2\}E[X(T-2r+2)]$, forms a lower bound on the expected cost of any arbitrary strategy *A*.

Corollary 3 is implied directly from Corollary 4.

2.2 Worst Assignment – An Upper Bound on Cost

For the sake of completeness, having derived a lower bound, we next provide an upper bound on the cost incurred by an arbitrary strategy. The bound derived below is under the same conditions for which Theorem 1 is derived. Without loss of generality, we also assume that $c_1\leq c_2$.

Theorem 2:

1. Let *A* be an arbitrary assignment policy, then the cost charged to *A* obeys: $C^A\leq c_2X(T-r+1)+c_1X(T-r)$.
2. There exists a traffic pattern $X_1(1),...,X_1(T),X_2(1),...,X_2(T)$ and an assignment policy $A^{\#}$, such that $C^{A^{\#}}=c_2X(T-r+1)+c_1X(T-r)$.

Proof:

(1) The bandwidth by which network L_2 charges must be bounded by $X(T-r+1)$ since L_2 must disregard the $r-1$ highest volume intervals assigned to L_2 and under any assignment there can be at most $r-1$ intervals whose traffic volume is greater than or equal to $X(T-r+1)$. Similarly, L_1 must disregard the $r-1$ highest volume intervals assigned to L_1 and any interval for which all the traffic is assigned to L_2. These lead to **(1).**

(2) Consider the traffic patter where $X_1(T-r+2)=...=X_1(T)=x$ and $X_2(T-r+2)=...=X_2(T)=x$ and $X_1(T-r+1)+X_2(T-r+1)=x-\varepsilon$ and $X_1(T-r)+X_2(T-r)=x-\varepsilon$ and for all other intervals the traffic volume is smaller than *x/4*.

$A^{\#}$ is an assignment that assigns $X_1(T-r+2),...,X_1(T)$, $X_1(T-r)+X_2(T-r)$ to L_1 and $X_2(T-r+2),...,X_2(T)$, $X_1(T-r+1)+X_2(T-r+1)$ to L_2.

The resulting cost is $C^{A^{\#}}=c_2X(T-r+1)+c_1X(T-r)$.

QED

2.3 Discussion

The lower bound derived in Corollary 2 and Corollary 3 can be used in practice to evaluate how good is one's assignment algorithm. The optimal policy derived in Section 2.1, suggests that if the operator (service provider) has full knowledge of the traffic demands and no failures are expected, then the best policy is to direct the *r-1* largest traffic intervals to the higher cost network and all the other intervals to the lower cost network. This principle can be used in designing heuristic rules when low failure rates are expected (e.g., when the failure rate is significantly lower than *q)* and when the operator has a good estimate of the traffic volumes. Good estimates of traffic volumes can be expected for a large fraction of the segments (e.g. relatively low load at nights, peak volumes a mid-mornings) and can be used in applying such heuristics.

The values of the lower bound (Corollary 2 and Corollary 3) and the upper bound (Theorem 2) imply that $C^{A^*} \le 2C^{A^\#}$. In practice, the difference between the bounds is likely to be larger. This suggests that the cost reduction one can achieve in applying optimization algorithms is quite meaningful.

3. Numerical Examples

Below we use a typical scenario in order to evaluate the economical viability of operating in the multi-homing mode. We are considering a service provider (customer) which has to place two traffic demands (the customer can easily form two demands by classifying its traffic into two classes) over one or two networks, provided by different providers. The customer may use a single provider, in which case when the network fails the customer is subject to severe quality degradation. Alternatively, the customer may purchase service at two network providers, and use one network as a primary and one network as an alternate (to be used when the primary fails). We assume that the network providers divide the month to intervals of 15 minutes length (that is about 3000 intervals) and the charges are set as a function of the load on the top 5% interval. We consider a traffic requirement faced by the customer to consist of random variables that depend on time. For example the traffic volume in the morning is a random variable whose mean is much larger than the random variable of the traffic at a night hour.

For the sake of the examples we assume that we are given 6 traffic representatives, representing say, the traffic volume of 9AM-1PM, 1PM-5PM, 5PM-9PM, 9PM-1AM, 1AM-5AM and 5AM-9AM. For each of these representatives there are 496 random variables (16 per day, for 31 days) all 496 are mutually independent and identically distributed. For ease of presentation we will assume that the number of random variables is 500 (all together 3000 per month).

We now consider 2 traffic demands, X_1, X_2, where $X_i(t)$ is a random variable denoting the traffic demand of source *i* at time *t,* and two networks L_1, L_2 where we assume that network L_i fails with probability p_i. We will evaluate the cost of various policies for placing the demands on the networks.

3.1 The traffic demands

We will consider a sample traffic demand that represents differences between day and night. The demand is given by: $X_1(1),...,X_1(500)$ [4] is *uniform* with M=20, S=6.66, where M is the mean and S is the standard deviation. All other demands are uniform: $X_1(501),...,X_1(1000)$ with M=30, S=6.66, $X_1(1001),...,X_1(1500)$ with M=50, S=13.33, $X_1(1501),...,X_1(2000)$ with M=70, S=13.33, $X_1(2001),...,X_1(2500)$ with M=100, S=20, $X_1(2501),...,X_1(3000)$ with M=120, S=20. An alternative traffic assumption used is the *normal distribution* in which we have: $X_1(1),...,X_1(500)$ with M=20, S=3.33, $X_1(501),...,X_1(1000)$ with M=30, S= 3.33, $X_1(1001),...,X_1(1500)$ with M=50, S=6.66, $X_1(1501),...,X_1(2000)$ with M=70, S=6.66, $X_1(2001),...,X_1(2500)$ with M=100, S=10, $X_1(2501),...,X_1(3000)$ with =120 , S=10.

3.2 A single demand over a single provider

Here we consider the cost of traffic demand X_1 when applied on a single network, say L_1. This is given by the top 5% of the traffic of a single source on a single network. This is a non multi-homing system, and thus the traffic may be subject to non-recoverable quality degradation. This evaluation is added for the sake of comparison.

3.3 No Multi-Homing: Two demands over a single provider

Here we consider the cost of running both X_1 and X_2 on a single network, say L_1. As in the previous sub-section this is a non multi-homing system, and thus the traffic may be subject to non-recoverable quality degradation. This evaluation is added for the sake of comparison.

3.4 Multi Homing: Two demands over two providers:

Under multi homing we consider two basic policies, *identical primary routes* and *different primary routes*.

3.4.1 Identical Primary routes

In this setting we consider a situation in which we are faced with two traffic demands, X_1 and X_2 which are statistically identical to each other. We assume that both demands are placed on L_1 as primary and on L_2 as secondary. We evaluate the cost of this solution as a function of p_1 the probability that L_1 fails (and where p_2 is the probability that L_2 fails). To analyze this system we need to compute $G_1^k(x)$ and $G_2^k(x)$ as presented in Eq. 1.. To this end note that a variation on Eq. 1 must be taken, yielding $G_1^k(x) = (1-p_1)F_1^k(x) * F_2^k(x)$, $G_2^k(x) = p_1 F_1^k(x) * F_2^k(x)$, where we have neglected the event that both networks fail (probability of $p_1 p_2$).

3.4.2 Different primary routes

This setting is identical to that of Section 3.4.1but where we assume that one demand is placed on L_1 as primary and on L_2 as secondary and the other demand is placed on L_2 as primary and on L_1 as secondary. For simplicity of presentation we also assume that the failure probabilities obey $p_1 = p_2$ and evaluate the cost as a function of p_1. As in Section 3.4.1 one need to adapt Eq. 1, yielding: $G_1^k(x) = (1-p_1)p_2 F_1^k(x) * F_2^k(x) + (1-p_1)(1-p_2)F_1^k(x)$ and $G_2^k(x)$ is symmetric. Note that this equation can be achieved, to a good approximation (whose error is in the order of $p_1 p_2$), if one uses Eq.1 with the substitutions $\pi_{2,1} = p_2, \pi_{1,1} = 1 - p_1$.

3.5 Results

The results for the normal distribution are provided in Figure 1. In the figure we depict the cost of running the demands, under various configurations, as a function of the probability of network failure. The line marked by squares (blue) represents the cost of a single demand times two (3.2). This is given as a reference. The curve marked by diamonds (green) represents the cost of running the two demands on a single network, under the assumption that there is no secondary network (no multi-homing, (3.3)). The curve marked by asterisks (red) represents a multi-homing scenario where the two demands are placed on an identical primary (and they also share the same secondary network, (3.4.1)). The curve marked by plus signs (light blue) represents the placement of the two demands on different primary networks where each of these networks serves as the secondary for the other demand (3.4.2). Finally, the curve marked by circles (purple) represents the lower bound (optimal assignment) derived in Section 2.1.

We can observe the following properties:

1. The cost of running two demands on the same network (no multi-homing) is less than double the cost of single demand. This is due to statistical multiplexing of the two demands (which are assumed to be independent of each other). The difference is in the order of 1.5%.
2. The cost of running two demands on the same primary network and using a second network for alternate (asterisks) is very low for low failure rates (up to 4% failure rate) but then increases quite sharply for high failure rates.

3. The cost of running the two demands on different primary networks and use the other networks for alternate(plus signs) is somewhat higher (than running them on the same network) for low failure rates but grows more modestly for high failure rates.
4. Overall – the extra cost due to applying the multi-homing mechanism is in the order of several percents.

Figure 2 demonstrates quite similar results for the uniform distribution traffic.

In Figure 3 we examine the sensitivity of the results to the variability of the individual demands, where we triple the standard deviation of each of the *X*'s. This means now that the value of *S* triples the value of *S* used in Figure 1. The results show that the extra cost roughly triples with comparison to the previous case, and thus still remain in the several percent extra cost range.

We next turn to evaluate the effect of the percentile pricing mechanism on the cost incurred. To this end we examine how the system will behave when top 10% pricing and top 2% pricing are used (instead of top 5% pricing). These results (top 10% pricing and top 2% pricing) are respectively depicted in Figure 4 and Figure 5. The figures demonstrate that in general, the cost decreases with the percentile of the top-percentile pricing. Thus, at 10% pricing the cost is very small while at 2% pricing it is much higher. Mostly sensitive to this change is the policy of placing two primaries on the same primary, which may end-up with 20% cost hike and more. However, this policy achieves very low cost as long as the failure probability is low. The figures demonstrate a rule-of-thumb in which this policy yields good results as long as the failure probability (denote it p) is somewhat lower than the top percentile parameter q, specifically as long as $p<0.8q$. .

3.6 Discussion

The most important property we observe from the wide set of runs conducted, is that the extra expected cost incurred due to running multi-homing is quite small (a few percents), as long as the failure rate is not large. This cost can be further reduced if more sophisticated assignment algorithms are used.

The economical evaluation of multi-homing (in comparison to no multi-homing) must account also for the following two factors: 1) The fixed cost paid to each network, if any. 2) The extra reliability, and thus higher level of QoS guarantee, one gains due to multi-homing. With this respect we may comment that unless the fixed cost of the network is relatively large, the multi-homing solution seems to be very viable, since it drastically reduces the probability of failure (approximately, the probability gets squared) at the expense of very small increase (a few percents) in the bandwidth cost.

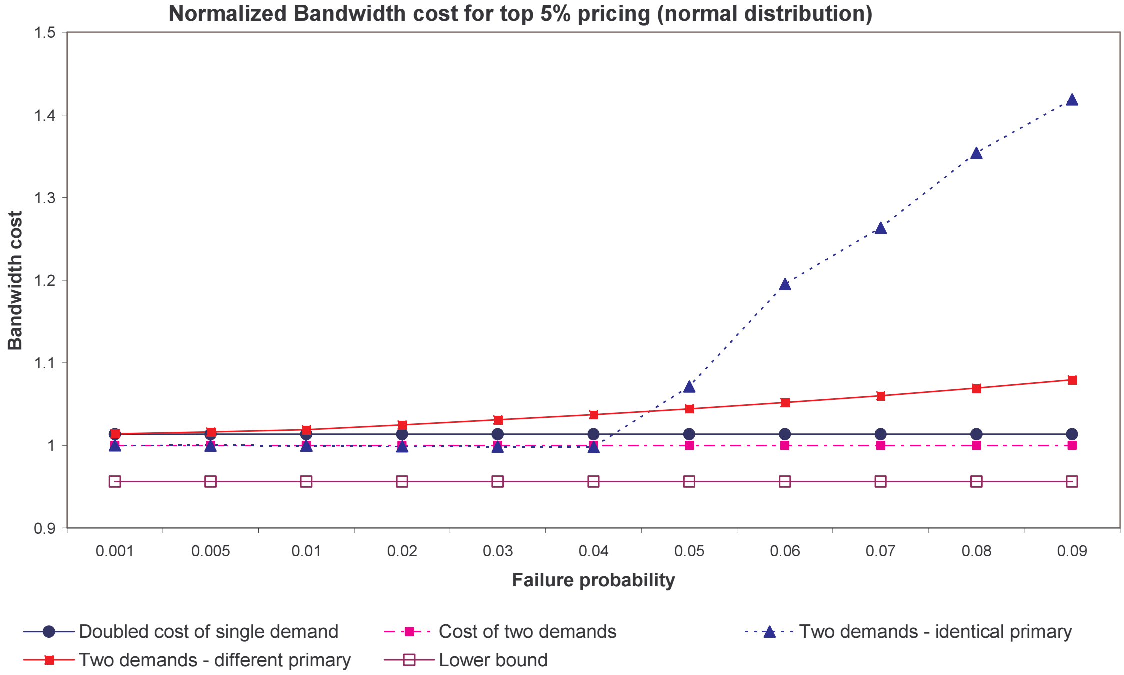

Figure 1: The relative cost of multi-homing for two traffic demands (normal distribution).

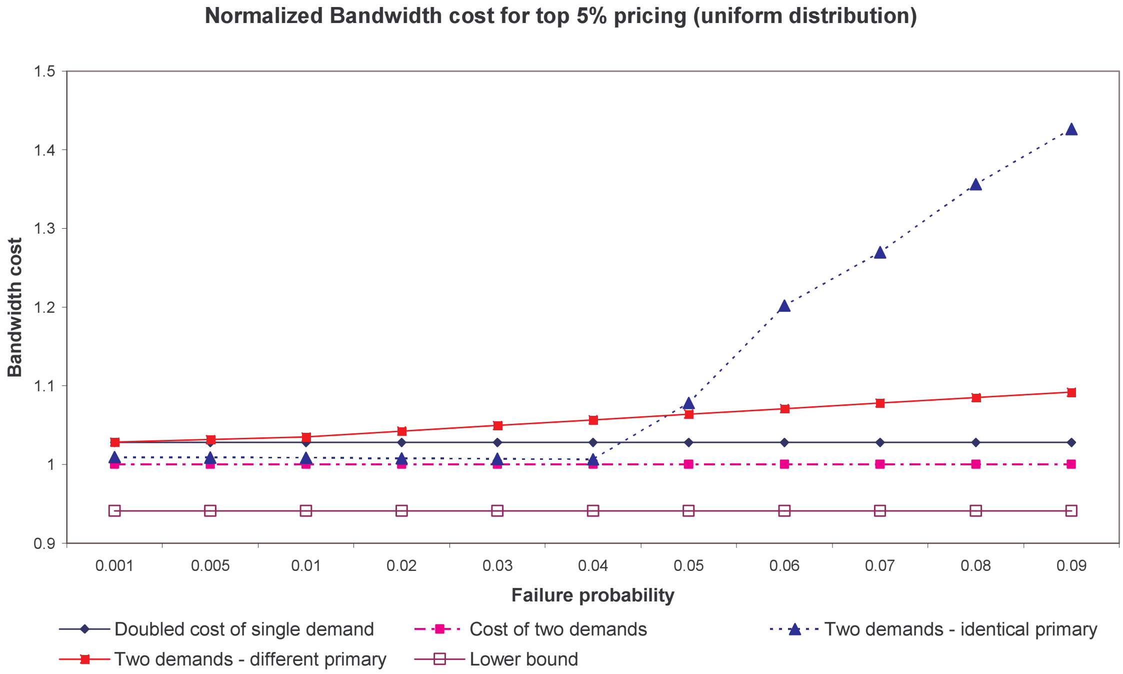

Figure 2: The relative cost of multi-homing for two traffic demands (uniform distribution)

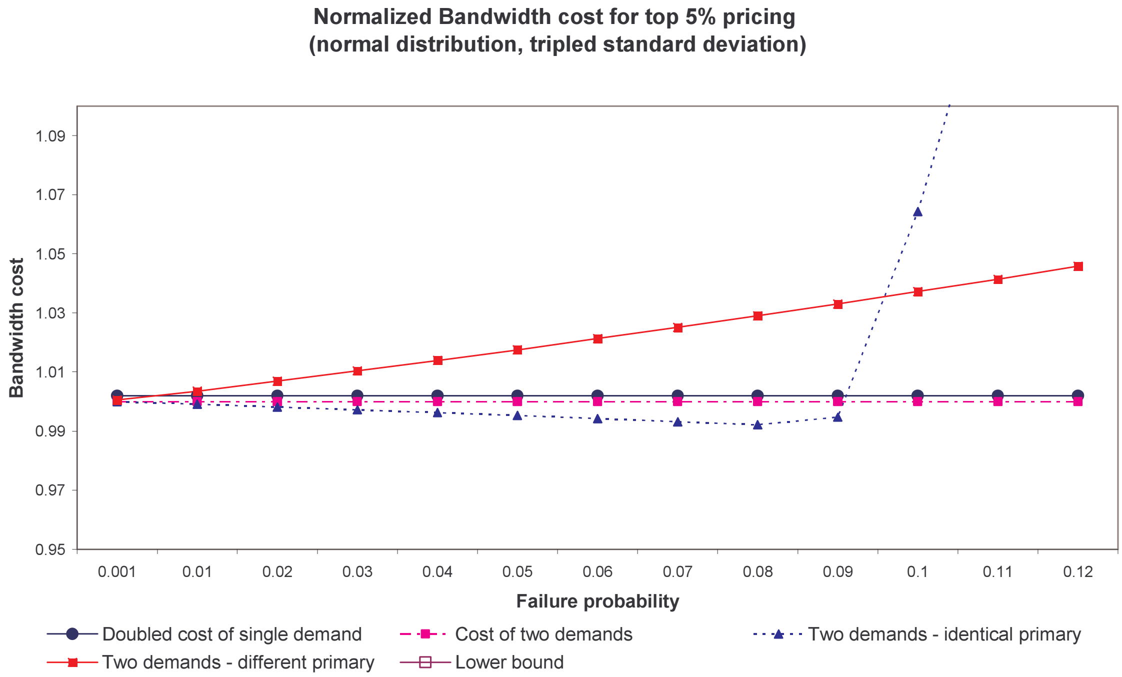

Figure 3: The relative cost of multi-homing for two traffic demands (normal distribution, triple standard deviation)

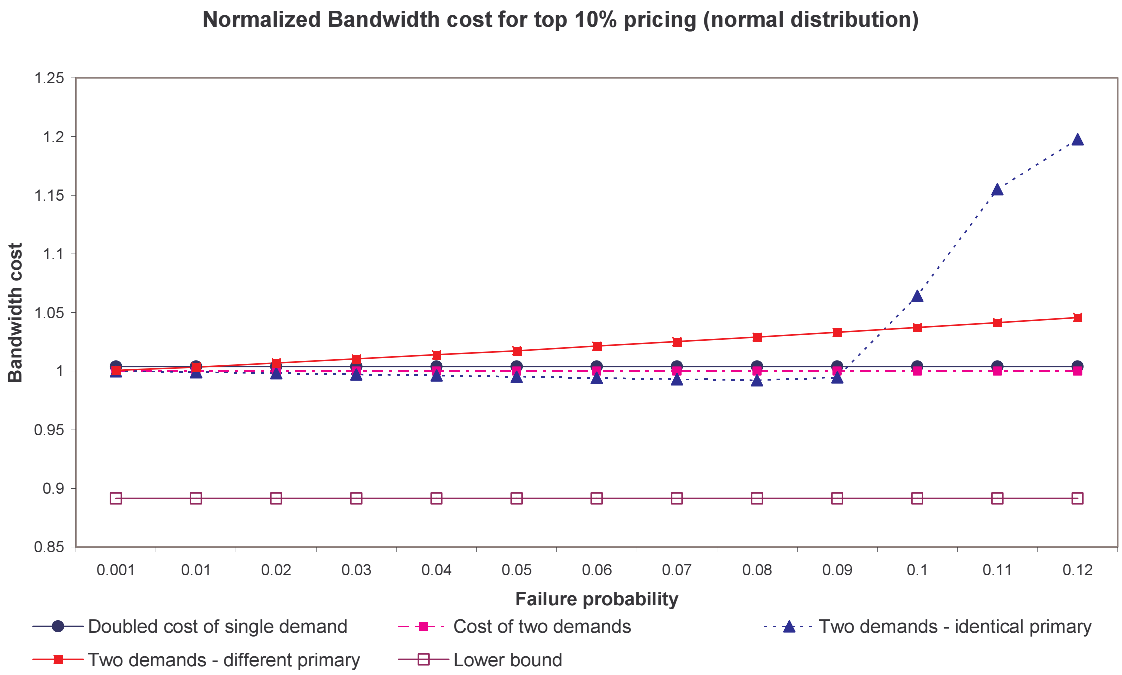

Figure 4: The relative cost of multi-homing for two traffic demands (normal distribution, top 10% pricing)

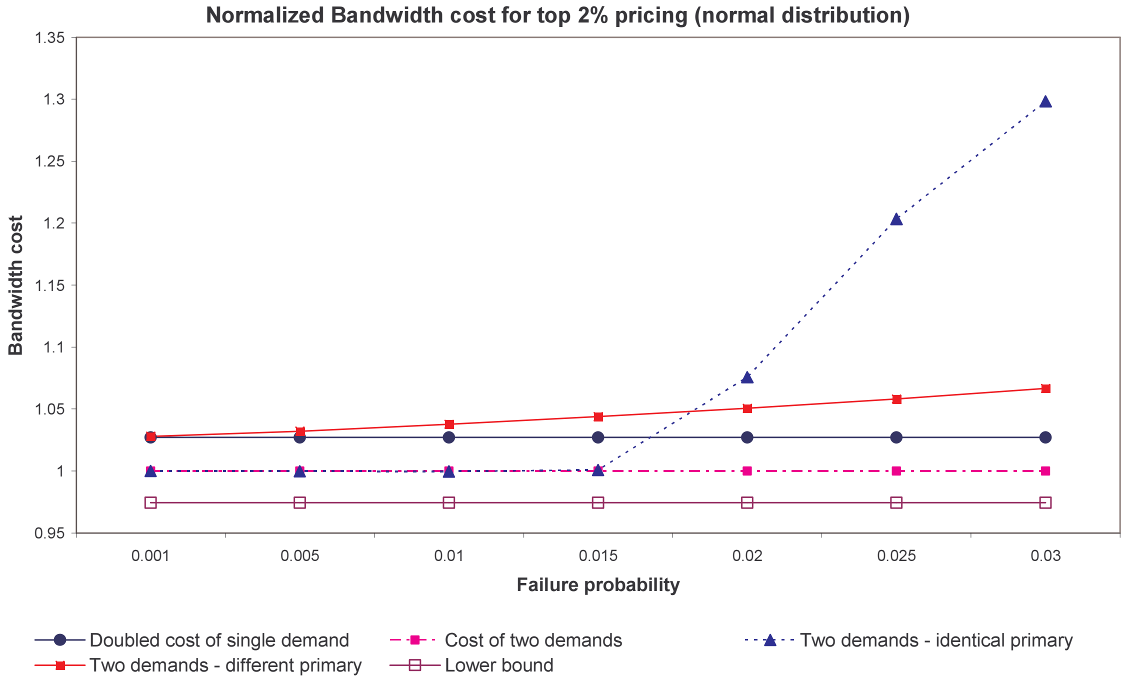

Figure 5: The relative cost of multi-homing for two traffic demands (normal distribution, top 2% pricing)

4. Concluding Remarks

Top percentile pricing, which is becoming increasingly popular, poses new challenges on network design and traffic engineering. The efficient operation of a network under the top-percentile paradigm has not been studied and is not well understood. In this work we proposed a model of this pricing and derived a mathematical framework that can be used for evaluating the expected cost of this pricing for general network structures. We specifically evaluated the efficient operation of the multi-homing architecture under the top-percentile pricing. Our analysis showed that if multi-homing is operated properly, then under wide set of conditions the extra cost incurred by it is relatively small. Yet, there are conditions where this extra cost can be significant. The mathematical model developed in this work can potentially be further used to evaluate other network design and traffic engineering problems under top-percentile pricing.

5. References

1. E. Altman, T. Basar, T. Jimenez and N. Shimkin (2000), "Competitive routing in Networks with Polynomial Cost", *Proceedings of INFOCOM' 2000*, pp. 1586-1593, Tel-Aviv, 2000. http://www.ieee-infocom.org/2000/papers/128.ps.
2. D. D. Clark (1997), "Internet Cost Allocation and Pricing," *Internet Economics*, L. W. McKnight and J. P. Bailey, Eds., Cambridge, Massachusetts, 1997, MIT Press, pp. 216-252.
3. R. Cocchi , S. Shenker , D. Estrin , L. Zhang (1993), Pricing in computer networks: motivation, formulation, and example, *IEEE/ACM Transactions on Networking* (TON), Vol.1 No.6, Dec. 1993, pp.614-627.
4. R. Cocchi, D. Estrin, S. Shenker, L. Zhang (1991), "A Study of Priority Pricing in Multiple Service Class Networks"; *ACM Computer Communication Review*, Vol 21, No. 4, September 1991, pp 123-130.
5. M. Falkner, M. Devetsikiotis and I.. Lambadaris (2000), " An overview of pricing concepts for broadband IP networks," *IEEE Communications Surveys and Tutorials* 3(2), 2000, pp. 2-13.

6. R.J. Gibbens and F.P. Kelly (1998), "Resource Pricing and the evolution of congestion control", preprint, 1998. available at http://www.statslab.cam.ac.uk/~frank/evol.html.
7. A. Gupta, D. O. Stahl and A. B. Whinston (1997), "Priority Pricing of Integrated Services Networks," *Internet Economics*, L. W. McKnight and J. P. Bailey, Eds., Cambridge, Massachusetts, 1997, MIT Press, pp. 323-352.
8. M. Herzberg and F. Shleifer (1999), "Optimization Models for the Design of Bi-directional Self-healing Ring Based Networks", *International Teletraffic Congress (ITC)* 16, Edinburgh, United Kingdom, 1999, pp 183-194.
9. F. P. Kelly (1997), "Charging and accounting for bursty connections", in *Internet Economics*, Lee W. McKnight and Joseph P. Bailey eds., MIT Press, 1997.
10. L. Kleinrock (1976), *Queueing Systems, Vol. 2*, *Computer Applications*, John Wiley, 1976.
11. J. Levy, H. Levy and Y. Kahana (2003), "Top Percentile Network Pricing and the Economics of Multi-Homing", *Proceedings of INOC' 2003, Paris, October 2003.*
12. J. MacKie-Mason and H. Varian (1995), "Pricing Congestible Network Resources," *IEEE JSAC*, vol. 13, no. 7, Sept. 1995, pp. 114-149.
13. T. Monk (2002), "Inter-domain Traffic Engineering: Applications in Complex Networks," Ixia, presented at *NFOEC2002*, Dallas, TX, September 15-19, 2002. http://www.ixiacom.com/support/techinfo/presentations/Paper_IDTE_NFOEC2002.pdf.
14. A. M. Odlyzko (2001), "Internet pricing and the history of communications", *Computer Networks* 36 (2001), pp. 493-517.
15. A. M. Odlyzko (1997), "A Modest Proposal for Preventing Internet Congestion," Sept. 1997, AT&T Research Labs, Technical report, http://www.dtc.umn.edu/~odlyzko/doc/modest.proposal.ps.
16. A. Orda and R. Rom (1990), "Multi-homing in Computer Networks – A Topology Design Approach", *Computer Networks and ISDN Systems*, Vol. 18, February 1990, pp. 133-141.
17. I. Ch. Paschalidis and J. N. Tsitsiklis (2000), "Congestion-Dependent Pricing of Network Services", *IEEE/ACM Transactions on Networking*, vol. 8, 2000, pp.171–184.
18. D. J. Songhurst, *Charging Communication Networks: From Theory to Practice*, Elsevier Science, Amsterdam, The Netherlands, 1999.
19. S. Shenker, D. Clark, D. Estrin, S. Herzog (1996), "Pricing in computer networks: reshaping the research agenda," *Telecommunications Policy*, 20 (3), 1996, pp. 183-201.
20. X. Wang and H. Schulzrinne (2001), "Pricing Network Resources for Adaptive Applications in a differentiated Services Networks", *Proceedings of INFOCOM' 2001*, Anchorage Alaska, 2001. http://www.ieee-infocom.org/2001/paper/854.ps.

6. Appendix

Proof of Assertion 3: For the sake of contradiction assume that the claim does not hold. Consider an optimal policy A that does not follow the claim. If both $R_1^A(t_i^1) = R_2^A(t_i^1) = 2$ then we must have $Y_1^A(t_1^1) = \ldots = Y_1^A(t_i^1) = 0$ and thus one can change A to $A*$ by setting $R_1^A(t_i^1) = R_2^A(t_i^1) = 1$ which leads to an increase in the value of Y_1^A, that is $Y_1^{A*}(t_1^i) \geq Y_1^A(t_1^i)$. This off course does not increase the cost charged by L_1 since all values of $Y_1^{A*}(t_1^1) = \ldots = Y_1^{A*}(t_{i-1}^1)$ remain 0, in particular $Y_1^{A*}(t_{T-r+1}^1) = 0$. Also the cost charged by L_2 does not increases as a result of this change (traffic is transferred from L_2 to L_1). Thus $C^{A*} \leq C^A$, in contradiction to the assumption. In a similar manner suppose $R_1^A(t_i^1) = 1, R_2^A(t_i^1) = 2$. Then one can change A to $A*$ by setting $R_1^A(t_i^1) = 2$. This again will not change the cost charged by L_1 since $t_{T-r+1}^1 < t_i^1$, and does not increase the cost charged by L_2 (traffic shift from L_2 to L_1). Thus $C^{A*} \leq C^A$, in contradiction to the assumption. In a similar manner treat the case $R_1^A(t_i^1) = 2, R_2^A(t_i^1) = 1$. By the way of contradiction the proof follows.

QED.

---

[1] Note that at the design stage, one may still consider the price as being "per-usage", since one selects the "width of the pipes" based on one's estimate of traffic. Past the design stage the price is then fixed regardless of the actual traffic shipped.

[2] Where $T_0 \equiv 0$.

[3] For ease of reading, the network index is omitted.

[4] The reader should note that the indexing of the internals (1 through 3000) is done for mathematical convenience. The index is not necessarily related to the specific time of the interval (during the month or the day).