QoS Optimization Model and Algorithms for VoIP Network

Bin Zeng and Lu Yao

Department of Management Engineering, Naval University of Engineering, Wuhan, China

zbtrueice@163.com, yaolu79@126.com

Keywords: Voice over Internet protocol, Quality of Service, Optimization model, Network traffic.

Abstract. In order to provide the same or better service quality in the Internet than traditional circuit-switched telephone network, a number of issues have to be solved which have hampered it in the Internet. Therefore, in this paper the VoIP (Voice over Internet Protocol) network design problems are modeled as nonlinear non convex combinatorial integer mathematical programming problems. The optimization problems have great practice value for the network service providers. This problem is a constrained multicast QoS routing and is proved to be a NP complete problem. The total link capacity augmented cost should be minimized when we design a new VoIP network or when the original network could not serve all of the traffic demands. It is also known as a complicated problem.

Introduction

Originally, the Internet was designed to provide best-effort services for the data generated by computers. The timeliness of data communication is generally delay tolerant. Quality-of-Service (QoS) constraints are not as important as routing flexibility and connectivity. Hence, using IP to transport voice data is contradictory to the basic requirement of the voice service: a timely delivery of voice samples. Although IP was not initially designed to provide services for real-time traffic, recent technical progress has made IP have the capabilities to provide real-time services in near future.

In order to provide the same or better service quality in the Internet than traditional circuit-switched telephone network, we must deal with a number of issues that have hampered it in the Internet. Voice service requirements could be discussed from two perspectives: (1) application requirements such as end-to-end delay, jitter, packet loss and overdue probability; (2) user's perspective such as reliability, availability, and supplementary services [1][2]. Telephony service providers must guarantee the quality of service they provide e.g. maximum one way delay does not exceed 150ms.

Internet telephony requires a range of protocols, ranging from those needed for transporting real-time data across to the network e.g. Real-time Transport protocol (RTP) [3], to Quality-of-Service aware routing (QoS routing), signaling protocol, resource reservation, internetworking between IP networks and PSTN, QoS-aware network management and billing protocols. ITU-T defined H.323 to provide multimedia communication in packet networks.

From the perspective of network service providers, they want to optimize the network performance such as minimizing the total bandwidth consumption, maximizing throughput or total revenue [4][5] subject to user and application constraints. In this paper, we want to develop the mathematical model for VoIP network. We minimize the total bandwidth consumption under users' QoS requirements, the network topology and the network capacity.

Performance Optimization Model

It is generally accepted that Internet telephony and traditional circuit-switched telephony will coexist for quite some time. The VoIP architecture must deal with interworking between IP networks and PSTN, so we need gateways between the two worlds. There are four possible models of VoIP [6]. They are PC-to-PC, Gateway-to-Gateway, PC-to-Gateway, and Gateway-to-PC models. The architecture of VoIP is shown in Fig.1. The first model of VoIP is PC-to-PC architecture, which based

on the assumption that two or more users have access to multimedia computers that are connected to the Internet.

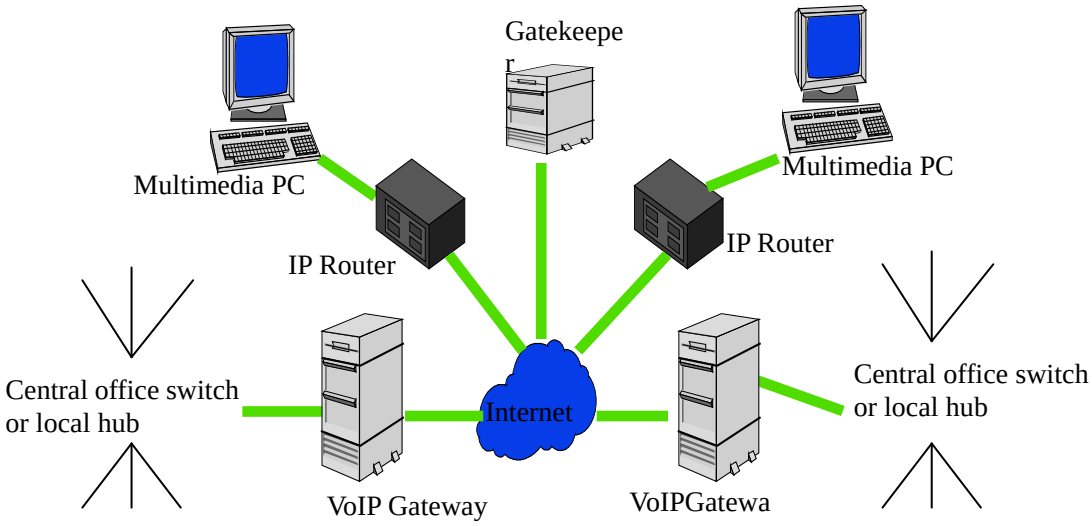

Fig.1 VoIP Architecture

The VoIP system is modeled as a graph, where the hosts, VoIP gateways, and switches are represented by nodes and communication link sets are represented by links. Let $N = \{1, 2,\ldots, n\}$ be the set of nodes and L be the set of links in the graph (network). Let G be the set of all user groups. An user group g is a voice communication session requesting for transmission in the network. An user group g may be an unicast from the source to a destination or a multicast from the source to multiple destinations. For each user group g, the traffic is transmitted exactly over one tree. D_g represents the set of destinations of user group g. λ_g is the required bandwidth of the voice transmission for each user group g. Below is a verbal description of the VoIP system design problem we considered.

Given :

Network topology

Capacities of network links

Equivalent bandwidth of each multicast/unicast group

Time threshold and tolerable overdue probability for each multicast/unicast group

To determine :

The minimum overall bandwidth consumption

Routing tree for each multicast/unicast group

Overdue probability of each multicast/unicast group

Objective :

To minimize the total bandwidth consumption

Subject to :

End-to-end QoS (overdue probability) constraint

Tree constraint

Multi-commodity flow constraint

Capacity constraint

Integrality constraint

Hop constraint

Table 1 and 2 are the notations we use in this paper .

Table1 Given Parameters of Performance Optimization Model

Notation	Description
N	The set of network nodes
G	The set of user group g
L	The set of network link l
L_j	The set of incoming links to network node j
R	The set of source nodes for all user groups
R_g	The source node of user group g
$L^o_{R_g}$	The set of outgoing links of the source node of user group g
D_g	The set of destinations of user group g
δ_{pl}	Indication function, 1 if path p uses link l, and 0 otherwise
C_l	Capacity of link l
T_{gd}	Time threshold for destination d of user group g
K_{gd}	End-to-end overdue requirement for destination d of user group g
$N(b_{gl})$	If b_{gl} is 0 then $N(b_{gl})$ =0, otherwise $N(b_{gl})$ =1
H_{gd}	The max number of hops for destination d of user group g
P_{gd}	The set of paths destination d of multicast group g may use
$\hat{B}_l$	The set of possible allocation bandwidth types for link l
$\overline{B}_l$	The upper bound of possible allocation bandwidth types for link l
$\underline{B}_l$	The lower bound of possible allocation bandwidth types for link l
λ_g	Equivalent bandwidth for user group g
A_u	An upper bound of A_{gd}
B_u	An upper bound of B_{gd}
$M_{gl}(b_{gl}, \lambda_g)$	Mean delay measured on link l for user group g given bandwidth reserved b_{gl} and mean rate λ_g
$V_{gl}(b_{gl}, \lambda_g)$	Delay variance measured on link l for user group g given bandwidth reserved b_{gl} and mean rate λ_g

Table 2 Decision Variables of Performance Optimization Model

Notation	Description
b_{gl}	Bandwidth allocated to user group g on link l
$O(A_{gd}, B_{gd}, T_{gd})$	The overdue probability for destination d of user group g
A_{gd}	End-to-end aggregate delay of user group g destined for destination d
B_{gd}	End-to-end delay variance of user group g destined for destination d
x_{gpd}	1 if path p is selected for user group g destined for destination d, and 0 otherwise
y_{gl}	1 if link l is selected for user group g, and 0 otherwise
f_{gld}	1 if link l is selected for user group g destined for destination d, and 0 otherwise

Objective function:

$$Z_{IP1} = \min \sum_{l \in L} \sum_{g \in G} b_{gl} \qquad \text{(IP1)}$$

subject to:

$$(1) \sum_{p \in P_{gd}} \delta_{pl} x_{gpd} \le y_{gl} \qquad \forall d \in D_g, g \in G, l \in L$$

$$(2) \sum_{g \in G} b_{gl} \le C_l \qquad \forall l \in L$$

$$(3) \sum_{l \in L} M_{gl}(b_{gl}, \lambda_g) f_{gld} \le A_{gd} \qquad \forall d \in D_g, g \in G$$

$$(4) \sum_{l \in L} V_{gl}(b_{gl}, \lambda_g) f_{gld} \le B_{gd} \qquad \forall d \in D_g, g \in G$$

$$(5) O(A_{gd}, B_{gd}, T_{gd}) \le K_{gd} \qquad \forall d \in D_g, g \in G$$

$$(6) \sum_{l \in L} \sum_{p \in P_{gd}} \delta_{pl} x_{gpd} \le H_{gd} \qquad \forall d \in D_g, g \in G$$

$$(7) \sum_{l \in L_j} y_{gl} \le 1 \qquad \forall j \in N, g \in G$$

$$(8) y_{gl} = 0 \vee 1 \qquad \forall g \in G, l \in L$$

The objective function is to minimize the total bandwidth consumption in the network. Constraint (1) ensures that if l is not used by group g then the path $p \in P_{gd}$ can not use link l. Constraint (7) is referred to as the tree constraint. By using Constraints (7) and (8) we can avoid the inefficiency of pre-stored candidate tree method in [7]. Constraints (1), (7) and (8) ensure that the union of the selected path(s) for the destinations of user group g forms a tree. Constraint (2) is referred as the capacity constraint, which ensures the aggregate bandwidth reserved on link l does not exceed the link capacity C_l. Constraints (3), (4) and (5) are the QoS constraints, which require the end-to-end QoS requirement for each source-destination pair of user group g to be satisfied. Constraint (3) denotes the aggregate delay on the path p for destination d of user group g. Constraint (4) denotes the jitter constraint. The Constraints (3) and (4) are based on the assumption that the delay and variance generated on each link in the network are mutually independent. The end-to-end delay and delay variance could be calculated by summing up the delay and delay variance of each link on the path p. Constraint (5) denotes the packet overdue constraint and the function $O(A_{gd}, B_{gd}, T_{gd})$, which is an end-to-end percentile-type delay objectives. We use normal approximation [8] to model the end-to-end delay distribution. Then we could compute the overdue probability for destination d of user group g using the normal distribution approximate function by given the end-to-end delay, end-to end-delay variance and a predetermined time threshold. Constraint (5) ensures that the end-to-end overdue probability to be satisfied for each destination d of user group g. Below is the algorithm to calculate overdue probability.

Algorithm: Cal_Overdue_Probability

Set the mean delay and standard to the values calculating from Constraint (3) and (4) to tm_{gd}, ts_{gd} respectively for each destination d of user group g. And the time threshold for each destination d of user group g is tr_{gd}. Then determine the overdue probability o_{gd} by the following normal approximation equation where three intermediate steps to calculate Z_{gd}, t_{gd} and F_{gd}.

1. Compute $Z_{gd} := (tr_{gd} - tm_{gd})/ts_{gd} \ge 0$ then use it in the subsequent equation as is. However if $Z_{gd} < 0$ then drop the negative sign.

2. Compute $t_{gd} := 1/(1+0.2316419* Z_{gd})$. Note that Z_{gd} here and next step are always nonnegative.

3. Compute $F_{gd} := 0.3989423/e^{z*z/2}$

Finally compute overdue probability

(if $Z_{gd} \ge 0$):

o_{gd} := F_{gd}*(0.319382*t_{gd} - 0.356564*t_{gd}^2 + 1.781478* t_{gd}^3 – 1.821256* t_{gd}^4 + 1.330274* t_{gd}^5). else if $Z_{gd} < 0$, however, the overdue probability is $1 - o_{gd}$.

Computational Experiments

In the computational experiments, we test the proposed algorithm for efficiency and effectiveness. The VoIP performance optimization algorithm is coded in Java 2 language and run on an IBM compatible PC whose CPU is Pentium IV. The algorithm is tested on five networks: GTE (12 nodes with 50 directed links), OCT (26 nodes, 60 directed links), PSS (14 nodes, 42 directed links), SITA (10 nodes, 56 directed links), and SWIFT (15 nodes, 40 directed links) under different traffic loads. The traffic rate of each user group is constant bit rate 8 kb (G.729A).

There are several parameters to be varied. They are link capacities, the number of user groups and the number of destinations of each user group. We assume the link capacities in the network are homogeneous i.e. the same value for each link. The user groups and the number of destinations of each user group are obtained using random value generator provided by Java 2 language. Internet telephony service is interactive that means n-way communications. The network to be optimized is composed by directed links. For each user group g, we need to generate additional $|D_g|$ user groups so that the n-way communication could proceed.

The time threshold is 125ms for one way. The overdue probability requirement for the round-trip is normally 0.05. How to efficiently allocate the end-to-end delay objective is important. Simply allocate half of the required overdue probability on one way is not a good scheme. In the computational experiments, the one way overdue requirement is calculated by $1-\sqrt{0.95}$ about 0.02532. The delay performance model in the computational experiments is M/D/1.

Our model could serve any kind of delay performance model as long as providing the mean delay and delay variance on each link. Choosing M/D/1 is just for demonstration purpose. The mean delay and the delay variance of M/D/1 model are below:

$$D = \bar{t} + \frac{\lambda \bar{t}^2}{2(1-\rho)}$$

$$V = \frac{4\bar{t}^2\rho(1-\rho) + 3\bar{t}^2\rho^2}{12(1-\rho)^2}$$

where $\bar{t}$: the mean packet service time. The mean traffic rate for G.729A is 100 packets/s (1s/10ms=100) and the mean packet service time is the function of reserved bandwidth. The utilization is the production of the mean traffic rate and the mean packet service time. The maximum iteration we run the algorithm is set to 200 by default. The step size control parameter δ is initially set to 2 and halved whenever the objective function value does not improve in 20 iterations.

The first column represents the tested network. The second column is the capacity for each link in the tested network. The third column specifies the utilization of the tested network. The fourth column is the number of user groups. The fifth column is the upper limit of the number of destinations for each user group. The sixth column is the CPU time to get the upper and lower bound.

Table 3 Traffic Loads and Results for Tested Networks

Networks	Link Capacity	Util.	# of user group	Upper limit of $\|D_g\|$	CPU time (sec)
GTE	512	0.1063	100	1	132.39
GTE	512	0.1463	101	2	210.07
GTE	512	0.1613	89	3	250.38
OCT	512	0.0833	40	1	137.70
OCT	512	0.1224	47	2	107.15

OCT	512	0.1466	39	3	131.10
PSS	512	0.1131	60	1	84.18
PSS	512	0.1429	63	2	132.75
PSS	512	0.1749	62	3	187.90
SITA	512	0.0804	100	1	139.97
SITA	512	0.0898	87	2	184.44
SITA	512	0.1038	76	3	219.00
SWIFT	512	0.1203	70	1	97.75
SWIFT	512	0.125	51	2	11021
SWIFT	512	0.1492	59	3	160.24

The computational results are shown on Table 3. From the computational results we have the following observations:

1. When the number of destinations is small (1 to 3), our algorithm could get near optimal solution.
2. The utilization of tested network does not affect the error difference.
3. The error difference is larger when the number of destinations increases. And the error difference decreases when the number of destinations approaches the number of nodes in the network.
4. The CPU time is much larger when the number of destinations becomes large. The reason is the number of potential trees to cover the destinations is increasing very fast as the number of destinations increases, so the algorithm needs much more time.
5. Different traffic rates do not affect the result of our algorithm.

Summary

In this paper we propose three mathematical models for design and planning of VoIP systems. Firstly we consider the first problem that is performance optimization of the VoIP system. We minimize total bandwidth consumption under end-to-end QoS guarantees. The mathematical formulation and solution approach for this problem is discussed. We minimize the total capacity augmented cost in order to serve all of the user groups under QoS constraints. Normal approximation is the end-to-end delay objective allocation scheme. Although normal approximation can not guarantee QoS, it could provide close estimate on QoS.

References

[1] W. Mazurczyk, P. Szaga and K. Szczypiorski: Multimedia Tools and Applications Vol. 70(2014), p. 2139-2165

[2] M.A.Akbar and M.Farooq: knowledge and Information Systems Vol. 38(2014), p. 491-510.

[3] V. Subramanian and R. Dutta: Measuring SIP Proxy Server Performance Vol. 20(2013), p. 5-13

[4] W. Mazurczyk: Journal of ACM Computing Surveys Vol. 46(2013), p. 212-224

[5] A.R.Lakshman and K.M. Praven: International Journal of Advanced Research in Computer Science Vol. 5(2014), p. 23-27

[6] R.S.Naoum and M. Maswady: WCSIT Vol. 2(2012), p. 110-114

[7] C. Shivani and J.Preety: International Journal of Research in Engineering and Applied Sciences Vol. 4 (2014), p. 161-166

[8] A.Rahman and P. P. Amritha: Advances in Intelligent Systems and Computing Vol. 325(2015), p. 489-494