ATLANTIS PRESS
Advances in Computer Science Research, volume 44
3rd International Conference on Wireless Communication and Sensor Network (WCSN 2016)

A Performance Optimization Model for Voice over IP Networks

Yao Lu
Dept. management engineering, Naval University of engineering, Wuhan, China
E-mail: yaolu79@126.com

Rui Wang
Library, Dept. training, Naval university of engineering, Wuhan, China
E-mail: kingwis@163.com

***Abstract*-In order to provide the same or better service quality in the Internet than traditional circuit-switched telephone network, there exist a number of issues to be dealt with that have hampered it in the Internet. These need good network planning and capacity management algorithms. The network performance optimization is modeled as nonlinear non-convex combinatorial mathematical formulations. The objective of the model is to minimize the total bandwidth consumption of IP telephony systems and subject to QoS and capacity constraints. Solution procedures based upon Lagrangean relaxation are proposed for the optimization formulation. In the computational experiments, the algorithm is tested to verify its effectiveness and efficiency.**

Keywords-Voice over IP networks; Performance optimization; Quality of service; Bandwidth allocation

I. INTRODUCTION

One major subject in the NII (National Information Infrastructure) is to construct one broadband network to provide integrated services. It is no doubt that the Internet will become the integrated-service platform for the next century. There are various kinds of applications to be appeared in the Internet. IP telephony is the most promising application to be deployed by telecommunication companies or promoted by network equipment vendors [1, 2]. IP telephony was first used as a simple way to provide point-to-point voice transport between two IP hosts, primarily to replace expensive international phone calls. This could reduce much of the communication cost and is the opportunity for many companies to enter the telecommunication industry. As the trend toward multimedia applications in the Internet, the scope of IP telephony has expanded to integrate video and data services.

It is generally accepted that Internet telephony and traditional circuit-switched telephony will coexist for quite some time [3]. The IP telephony architecture must deal with inter-working between IP networks and PSTN, so we need gateways between the two worlds. There are four possible models of IP telephony [4]. They are PC-to-PC, Gateway-to-Gateway, PC-to-Gateway, and Gateway-to-PC models. 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), to Quality-of-Service aware routing (QoS routing) [5], signaling protocol, resource reservation, internetworking between IP networks and PSTN, QoS-aware network management and billing protocols [6]. ITU-T defined H.323 to provide multimedia communication in packet networks. IETF proposed its own architecture for IP telephony. They both use RTP to transport voice and video data.

Real-time service requires the availability of resources in the network to meet its service requirement. This necessitates the use of performance optimization tool for service providers to adjust their traffic characteristics and service requirements to the network [7]. In this paper, we want to develop the performance optimization mathematical models for IP telephony systems. We minimize the total bandwidth consumption under users' QoS requirements, the network topology and the network capacity.

For IP telephony service, the traffic rate is constant bit rate so the codec delay equals the payload size divided by the source information rate [8]. In this paper, because codec delay is fixed in voice delay, we don't take them into consideration. The buffer delay does not do anything with the voice delay, it only reduce the delay jitter. We consider the transmission delay and delay variance in the Internet to guarantee the overdue probability of user requirements. For gateway models, the delay generated in the local PSTN is almost fixed, so we do not consider it, too.

II. PERFORMANCE MANAGEMENT MODEL

N denotes the set of network nodes. G denotes the set of user group g. L denotes the set of network link l. Lj denotes the set of incoming links to network node j. R denotes the set of source nodes for all user groups. Rg denotes the source node of user group g. LoRg denotes the set of outgoing links of the source node of user group g. Dg denotes the set of destinations of user group g. δpl denotes Indication function, 1 if path p uses link l, and 0 otherwise. Cl denotes Capacity of link l. Tgd denotes Time threshold for destination d of user group g. Kgd denotes End-to-end overdue requirement for destination d of user group g. N(bgl) denotes If bgl is 0 then N(bgl)=0, otherwise N(bgl)=1. Hgd denotes the max number of hops for destination d of user group g. Pgd denotes the set of paths destination d of multicast group g may use. $\hat{B}_l$ denotes the set of possible allocation bandwidth types for link l. $\hat{B}_l$ denotes the upper bound of possible allocation bandwidth types for link l. $\underline{B}_l$ denotes the lower bound of possible allocation bandwidth types for link l. λg denotes Equivalent bandwidth for user group g. Au denotes an upper bound of Agd. Bu denotes an upper bound of Bgd. Mgl(bgl,λg) denotes mean delay measured on link l for user group g given bandwidth reserved bgl and mean rateλg. Vgl(bgl,λg) denotes delay variance measured on link l for user group g given bandwidth reserved bgl and mean rate Mgl(bgl,λg).

Objective function is defined as below:

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

subject to:

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

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

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

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

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

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

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

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

$$\sum_{p \in P_{gd}} \delta_{pl} x_{gpd} \le f_{gld}, \forall d \in D_g, g \in G, l \in L \quad (9)$$

$$f_{gld} = 0 \vee 1, \forall d \in D_g, g \in G, l \in L \quad (10)$$

$$\sum_{p \in P_{gd}} x_{gpd} = 1, \forall d \in D_g, g \in G \quad (11)$$

$$x_{gpd} = 0 \vee 1, \forall p \in P_{gd}, d \in D_g, g \in G \quad (12)$$

$$b_{gl} \in \hat{B}_l, \forall g \in G, l \in L \quad (13)$$

$$b_{gl} \le y_{gl} \overline{B}_l, \forall g \in G, l \in L \quad (14)$$

$$y_{gl} \le b_{gl} / \underline{B}_l, \forall g \in G, l \in L \quad (15)$$

$$A_{gd} \ge 0, \forall d \in D_g, g \in G \quad (16)$$

$$B_{gd} \ge 0, \forall d \in D_g, g \in G \quad (17)$$

$$A_{gd} \le A_u, \forall d \in D_g, g \in G \quad (18)$$

$$B_{gd} \le B_u, \forall d \in D_g, g \in G \quad (19)$$

$$\sum_{l \in L^o_{Rg}} y_{gl} \ge 1, \forall g \in G, R_g \in R \quad (20)$$

$$\sum_{l \in L_{R_g}} y_{gl} = 0, \forall g \in G, R_g \in R \quad (21)$$

$$\sum_{l \in L} y_{gl} \ge GH(g), \forall g \in G \quad (22)$$

$$\sum_{l \in L} N(b_{gl}) \ge GH(g), \forall g \in G. \quad (23)$$

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 [9]. 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 Cl. 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(Agd, Bgd, Tgd), which is an end-to-end percentile-type delay objectives. We use normal approximation 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. Constraint (6) denotes the hop constraint, which requires the total number of hops each path traverses does not exceed the pre-defined threshold. Constraint (9) relates the routing decision variables x's to the auxiliary variables f's. The introduction of the auxiliary variables f's may facilitate the decomposition in the Lagrangean relaxation problem to be discussed later. Constraint (10) is the integrality constraint for each fgld. Constraints (11) and (12) require that exactly one path is selected for each destination d of user group g. Constraint (13) requires that bandwidth reversed for user group g be allowable. Constraint (14) requires that when link l is used for user group g then the bandwidth reserved for user group g is not exceed the upper bound of allocated bandwidth. Constraint (15) forces the ygl to be 0 when the link l is not used for transmitting the traffic of user group g. Constraint (15) helps us to prune the non-used tree branches. Constraint (16) to Constraint (23) are the redundant constraints. These redundant constraints help us get a tighter lower bound (dual problem solution). Constraint (16) is the theoretical lower bound of Agd. Constraint (17) is the theoretical lower bound of Bgd. Constraint (18) denotes the upper bound of Agd. Constraint (19) limits the upper bound of Bgd. Constraint (20) requires at least one outgoing link is selected for the source of user group g. Constraint (21) requires that no link incoming to the source of user group g is used for transmitting traffic for user group g. Constraint (22) requires that we must select at least GH(g) links for user group g. Constraint (23) does the same job of Constraint (22). Constraint (23) requires that we must select at least GH(g) number of bgl for user group g.

III. LAGRANGEAN BASED ALGORITHM

The basic approach to the development of the solution procedure to Formulation (IP1) is Lagrangean relaxation. Lagrangean relaxation is a method for obtaining lower bounds (for minimization problems) as well as good primal solutions in integer programming problems.

For Formulation (IP1), we dualize Constraints (1), (2), (3), (4), (5), (9), (14) and (15) to obtain the following Lagrangean relaxation problem (LR1):

$$\begin{aligned}
Z_{D1}(u,v,a,w,\alpha,\beta,\gamma,m) &= \min \sum_{l\in L}\sum_{g\in G} b_{gl} \\
&+ \sum_{l\in L}\sum_{g\in G}\sum_{d\in D_g} u_{gld}\Big(\sum_{p\in P_{gd}} \delta_{pl} x_{gpd} - y_{gl}\Big) \\
&+ \sum_{l\in L} v_l\Big(\sum_{g\in G} b_{gl} - C_l\Big) \\
&+ \sum_{g\in G}\sum_{d\in D_g} a_{gd}\Big(\sum_{l\in L} M_{gl}(b_{gl},\lambda_g) f_{gld} - A_{gd}\Big) \\
&+ \sum_{g\in G}\sum_{d\in D_g} w_{gd}\Big(\sum_{l\in L} V_{gl}(b_{gl},\lambda_g) f_{gld} - B_{gd}\Big) \\
&+ \sum_{g\in G}\sum_{d\in D_g} \alpha_{gd}\big(O(A_{gd},B_{gd},T_{gd}) - K_{gd}\big) \\
&+ \sum_{l\in L}\sum_{g\in G}\sum_{d\in D_g} \beta_{gld}\Big(\sum_{p\in P_{gd}} \delta_{pl} x_{gpd} - f_{gld}\Big) \\
&+ \sum_{l\in L}\sum_{g\in G} \gamma_{gl}\big(b_{gl} - y_{gl}\overline{B}_l\big) \\
&+ \sum_{l\in L}\sum_{g\in G} m_{gl}\big(y_{gl} - b_{gl} / \underline{B}_l\big)
\end{aligned} \tag{24}$$

According to the weak Lagrangean duality theorem, for any (u, v, a, w, α, β, γ, m) ≥ 0, the optimal objective function value of (LR1), ZD1(u, v, a, w, α, β, γ, m) is a lower bound on ZIP1. To find the maximum of ZLR1(u, v, a, w, α, β, γ, m), we solve the dual problem (D1). We would like to determine the greatest lower bound by

$$Z_{D1} = \max_{u,v,a,w,\alpha,\beta,\gamma,m\geq 0} Z_{D1}(u,v,a,w,\alpha,\beta,\gamma,m) \tag{D1}$$

There are several methods for solving (D1). One of the most popular methods is the subgradient method. Let a (|G|(2*|L|(|Dg|+1)+3|Dg|)+|L|) vector s be a subgradient of ZD1(u, v, a, w, α, β, γ, m). In iteration k of the subgradient optimization procedure, the multiplier vector, bk = (uk, ak, ak, wk, αk, βk, γk, mk) is updated by

$$b^{k+1} = b^k + t^k s^k \tag{25}$$

The step size tk is determined by

$$t^k = \delta \frac{Z_{IP1}^h - Z_{D1}(b^k)}{\|s^k\|^2} \tag{26}$$

Where Z_{IP1}^h is the primal objective function value for a heuristic solution (an upper bound on Z_{IP1}) and δ is a constant, $0 < \delta \leq 2$.

After optimally solving each subproblem of (LR1), we can use the information generated in the solution procedure to get primal feasible solution for (IP1). (LR1) provides us a lot of useful information to solve (IP1) to get good primal

feasible solution. One is the routing assignment for each destination of user group g. But the union of the routing assignment of each destination d of user group g is not necessary to form a multicast tree [10]. We have the difficulty to make these routing assignments to become a tree. In order to construct the multicast routing tree for each user group g efficiently, we use the multipliers in the solution procedure to find the routing tree. We sum up the multipliers (u, β and ν) as the weight of each link. Then run the Bellman-Ford shortest path algorithm to construct the multicast tree.

IV. COMPUTATIONAL EXPERIMENTS

In the computational experiments, we test the proposed algorithm for efficiency and effectiveness. The IP telephony performance optimization algorithm is coded in Java 2 language. The algorithm is tested on the traffic rate of each user group is constant bit rate 8 kb (G.729A).

There are several parameters to be varied. They are 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 |Dg| 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 [11], seven schemes are developed to allocate the end-to-end percentile delay objectives. 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 [12].

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 initial objective value (Zip) is set to the sum of link capacity in the network, but if we could set it to a tight upper bound, we could speed up the convergence rate of lower bound.

The computational results are shown on Table I.

TABLE I. COMPUTATION RESULT FOR THE TESTED NETWORK

Util.	# of user group	Upper limit of $\lvert D_g \rvert$	Z^h_{IP1} (Upper bound)	Z_{D1} (Lower bound)	E.D (%) $\frac{Z^h_{IP1}-Z_{D1}}{Z_{D1}}$
0.1063	100	1	2720	2720	0
0.1463	101	2	3744	3744	0
0.1613	89	3	4128	3824	7.95
0.1331	45	6	3408	3120	9.13
0.2331	92	5	5967	5376	11.01
0.2675	107	5	6848	5968	14.75
0.2987	95	6	7600	6944	9.45
0.2613	107	4	6688	5792	15.47
0.3925	107	8	10048	9232	8.84
0.4213	400	1	10784	10784	0
0.4613	97	11	11808	10960	7.74
0.44.625	229	3	11424	10464	9.71

The first column specifies the utilization of the tested network. The second column is the number of user groups. The third column is the upper limit of the number of destinations for each user group. The forth column shows the best objective value calculated by our proposed algorithm. The fifth column gives the tightest lower bound found in the (D1). The sixth column provides the percentage difference between Z^h_{IP} and Z_{D1}.

From the computational results we have the following observations:

- When the number of destinations is small (1 to 3), our algorithm could get near optimal solution.
- The utilization of tested network does not affect the error difference.
- 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.
- Different traffic rates do not affect the result of our algorithm.
- The redundant Constraint (23) is important to the lower bound. We found the dual solution is very sensitive to Constraint (23). If we could improve the efficiency of the redundant constraint, which means the number of links we must select close to the optimal number of links to cover the destinations. The optimal number of links could be thought as the optimal solution of the special case of Steiner tree problem when the link weight in the network is 1 for all links. We found the least cost multicast tree for each user group.

V. CONCLUSION

In the paper, we model the network performance optimization as nonlinear non-convex combinatorial mathematical formulations. The objective of the model is to

minimize the total bandwidth consumption of IP telephony systems and subject to QoS and capacity constraints.

After completing the work, there are still other issues to be done for IP telephony systems. For example, reliability is a very important issue to telephony service. During the twentieth century, traditional telephony networks has been deployed on the earth. It provides strong reliability even when some catastrophes occur e.g. earthquake. IP telephony still could not provide the same reliability as circuit-switched telephony service. It is important to consider the reliability of IP telephony systems.

REFERENCES

[1] H.Sinnreich, A.B.Johnston, Internet communications using SIP: Delivering VoIP and multimedia services with Session Initiation Protocol. John Wiley & Sons, 2012.

[2] E.Karthikeyan and R. Shankar. "VoIP Packet Delay Techniques: A Survey," Global Journal of Computer Science and Technology vol.14, Mar. 2014, pp. 1-6.

[3] H. Naomi, E. Gillen, and A.Hines. "TCD-VoIP, a research database of degraded speech for assessing quality in voip applications." Proc. 2015 Seventh International Workshop on. IEEE Quality of Multimedia Experience (QoMEX), 2015, pp. 67-85.

[4] H. Assem, D. Malone, J. Dunne, et al. “Monitoring VoIP call quality using improved simplified E-model,” Proc. 2013 international conference on. IEEE Computing, networking and communications (ICNC),, 2013, pp. 927-931.

[5] P. Peter, H.Melvin, and A. Hines. "An analysis of the impact of playout delay adjustments introduced by voip jitter buffers on listening speech quality." Acta Acustica united with Acustica, vol 101, Mar. 2015, pp. 616-631.

[6] T. Jelassi, K. Sofiene, "Quality of experience of VoIP service: a survey of assessment approaches and open issues." IEEE Communications surveys & tutorials,vol 14, Feb. 2012, pp. 491-513.

[7] D. Villacís, F. R. Acosta, R. A. L. Cueva, “Performance analysis of VoIP Services over WiFi-based systems,” Proc. 2013 IEEE Colombian Conference on IEEE Communications and Computing (COLCOM), 2013, pp.1-6.

[8] C. Carvalho, L. Silva, and D.S.M. Edjair, "Survey on application-layer mechanisms for speech quality adaptation in VoIP," ACM Computing Surveys (CSUR), vol. 45, Mar. 2013, pp. 36-49.

[9] H. P. Singh, S. Singh, J. Singh, et al. “VoIP: State of art for global connectivity−A critical review,” Journal of Network and Computer Applications, vol .37, Dec. 2014, pp. 365-379.

[10] M. I. Tariq, M. A. Azad, R. Beuran, et al. “Performance analysis of VoIP codecs over BE WiMAX network,” International Journal of Computer and Electrical Engineering, vol. 3, Mar. 2013, pp. 345-362.

[11] G. Asante, J. B. Hayfron-Acquah, K. Riverson, “Leveraging VOIP on Local Area Network using Java Media Framework,” International Journal of Computer Applications, vol. 113, Oct. 2015, pp. 423-435.

[12] K. Nisar, A. Amphawan, S. Hassan, et al. “A comprehensive survey on scheduler for VoIP over WLAN,” Journal of Network and Computer Applications, vol. 36, Feb. 2, 2013, pp. 933-948.