A Lagrangean Relaxation Based Scheme for Allocation of Bandwidth

Indranil Bose
Room 730 Meng Wah Complex School of Business
The University of Hong Kong, Pokfulam Road, Hong Kong
Phone: (852)-2241-5845, Fax: (852)-2858-5614
e-mail: bose@business.hku.hk

Abstract: In this paper we study the bandwidth packing problem in the presence of priority classes. The bandwidth packing problem is defined as the selection and routing of messages from a given list of messages with prespecified requirements on demand for bandwidth. The messages have to satisfy delay constraints and have to be routed over a network with given topology so that the revenue generated from routing these messages is maximized. Messages to be routed are classified into two priority classes. An integer programming based formulation of this problem is proposed and a Lagrangean relaxation based methodology is described for solving this problem. Several numerical experiments are conducted using a number of problem parameters such as percentage of messages, ratio of messages of lower to higher priority, capacity of links and high quality solutions to the bandwidth packing problem are generated under the different situations.

Keywords: Optimization under uncertainty, Bandwidth, Heuristic, Lagrangean relaxation.

I. Introduction

Networks of today often suffer from congestion problems due to the tremendous increase in traffic in recent times as well as irrational allocation of bandwidth to support this increased traffic. One of the fundamental problems related to design of networks is determination of which messages to route among a given list of messages and determination of the routes to be used for delivering messages between communicating nodes so that the revenue generated from routing these messages is maximized. This is known as the bandwidth packing problem. The objective of this research is to use an optimization based approach for solving the bandwidth packing problem for messages belonging to multiple service classes. This will involve selection of a target group of messages from a list of messages with different delay requirements provided by the users, and determination of the best paths for routing these messages. Usually the topology of the network, the capacities of the links, the revenues to be generated by routing the messages, and the demand requirements of the messages are specified prior to the start of network design. The messages are listed in the form of a message table. In this table, the messages are prioritized based on the demand requirements. Since the network capacity is usually insufficient to route all messages, a selected group of messages are routed during a given period of time. This is also known as the static bandwidth packing problem (as opposed to dynamic bandwidth packing where the demand requirements of the messages change over time) and is studied in this paper. Our goal in this paper is to find an appropriate message selection and routing scheme that provides an efficient resource allocation mechanism and maximizes the revenue generated from the usage of the network.

Various versions of the bandwidth packing problem have been studied in the literature. This includes research conducted by Amiri and Barkhi [1], Amiri et al. [2], Anderson et al. [3], Cox et al. [4], Laguna and Glover [5], Park et al. [6], Parker and Ryan [7], and Rolland et al.[8]. Most of these papers strive to maximize the revenue earned by routing the messages subject to some service related constraints. A notable exception is the paper by Amiri et al. [2] where the objective is to maximize revenue as well as minimize the delay cost associated with the use of the network. Various methods are used in these papers including tabu search [3,5], genetic algorithms [4], column generation [6,7], and Lagrangean relaxation [1,2,8]. However, the available research on bandwidth packing considers only a single message class. This assumption is not realistic as users may use the networks for running different applications. Some of these applications may be delay sensitive but others may be not. So it is more realistic to model the bandwidth packing problem in the presence of multiple priority classes. To the best of our knowledge, this is the first paper that addresses the priority bandwidth packing problem. As opposed to the existing literature this problem is considerably more difficult because exact analytical expressions for the average delay of the priority classes are difficult to obtain and this is turn complicates the formulation of the problem.

II. Problem Formulation

We introduce the following notation for developing an integer programming model for the priority bandwidth packing problem:

N	the set of nodes in the network
E	the set of undirected links (arcs) in the network
M	the set of messages
M_1	the set of messages with lower priority
M_2	the set of messages with higher priority

$M = M_1 \cup M_2$

Proceedings of the Fifth International Conference on Electronic Business, Hong Kong, December 5-9, 2005, pp. 167 - 171.

$1/\mu_1$ average length of messages with lower priority

$1/\mu_2$ average length of messages with higher priority

d^{m_1} the demand for message $m_1 \in M_1$

d^{m_2} the demand for message $m_2 \in M_2$

r^m the revenue from message $m \in M$

$O(m)$ the source node for message $m \in M$

$D(m)$ the destination node for message $m \in M$

Q_{ij} the capacity of link (i,j)

δ_1 the upper limit on number of messages of lower priority in the network

δ_2 the upper limit on number of messages of higher priority in the network

$L_{ij}^{m_1}$ number of messages of lower priority on link (i,j)

$L_{ij}^{m_2}$ number of messages of higher priority on link (i,j)

The decision variables are:

$$Y_m = \begin{cases} 1 & \text{if call } m \text{ is routed} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij}^m = \begin{cases} 1 & \text{if call } m \text{ is routed through a path that uses link } (i,j) \\ 0 & \text{otherwise} \end{cases}$$

$$W_{ij}^m = \begin{cases} 1 & \text{if call } m \text{ is routed through a path that uses link } (i,j) \\ & \text{in the direction of } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

In order to obtain a mathematical formulation of the bandwidth packing problem we have to make several assumptions. These are listed below:

- The nodes have infinite buffers to store messages waiting for transmission
- Arrival process of messages entering the network follows a Poisson distribution
- Length of messages follows an exponential distribution
- Propagation delay in the links is negligible
- The average message length for each type of messages is used instead of using individual message lengths
- The link and message system is studied as a preemptive priority queue, i.e., the routing of lower priority of messages can be interrupted by higher priority messages

Based on the above assumptions, the telecommunication network is modeled as a network of M/M/1 queues. In this network, links are treated as servers with service rates proportional to the link capacities. The messages are treated as customers waiting to be routed at a particular node. If we measure elay in terms of messages, then the delay is defined as the total number of messages to be routed on a particular link. Since we want to limit the upper bound of delay, we can use a relaxed formulation of the delay in terms of maximum number of allowable messages on a link for priority 1 and priority 2 messages. We denote them by $L_{ij}^{m_1}$ and $L_{ij}^{m_2}$ respectively. The nature of the problem imposes certain restriction upon the characteristics of higher priority messages. Usually, they are shorter in length than low priority messages, i.e., $a = \mu_1 / \mu_2 \leq 1$. In addition, the higher priority messages generate more revenue while they are not as tolerant to queuing delay as the lower priority messages (i.e., $\delta_1 \leq \delta_2$) With the notations defined and queuing delay formulated, we can now model the bandwidth packing problem as follows:

Problem P

$$\text{Max} \sum_{m \in M_1, M_2} r_m Y_m \tag{1}$$

Subject to:

$$\sum_{j \in N} W_{ij}^m - \sum_{j \in N} W_{ji}^m = \begin{cases} Y^m \text{ if } i = O(m) \\ -Y^m \text{ if } i = D(m), i \in N \text{ and} m \in (M_1, M_2) \\ 0 \text{otherwise} \end{cases} \tag{2}$$

$$W_{ij}^m + W_{ji}^m \leq X_{ij}^m \;\forall\, (i,j) \in E \text{ and } m \in (M_1, M_2) \tag{3}$$

$$\sum_{m_1 \in M_1} d^{m_1} X_{ij}^{m_1} + \sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2} \leq Q_{ij} \;\forall\, (i,j) \in E \tag{4}$$

$$\sum_{(i,j) \in E} L_{ij}^{m_1} \leq \frac{(Q_{ij} - \sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2}) \sum_{m_1 \in M_1} d^{m_1} X_{ij}^{m_1} + a \sum_{m_1 \in M_1} d^{m_1} X_{ij}^{m_1} \sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2}}{(Q_{ij} - \sum_{m_1 \in M_1} d^{m_1} X_{ij}^{m_1} - \sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2})(Q_{ij} - \sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2})} \tag{5}$$

$$\sum_{(i,j) \in E} L_{ij}^{m_2} \leq \sum_{(i,j) \in E} \frac{\sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2}}{(Q_{ij} - \sum_{m_2 \in M_2} d^{m_2} X_{ij}^{m_2})} \leq \delta_2 \tag{6}$$

$$Y^m \in (0,1) \qquad \forall m \in (M_1, M_2) \tag{7}$$

$$X_{ij}^m \in (0,1) \qquad \forall (i,j) \in E \text{ and } m \in (M_1, M_2) \tag{8}$$

$$W_{ij}^m \in (0,1) \qquad \forall (i,j) \in E \text{ and } m \in (M_1, M_2) \tag{9}$$

The objective function (1) represents the total revenue earned from the routing of messages. Constraint set (2) represents flow conservation equations, which define a route for each message represented by a communicating node pair. Constraint set (3) links together the X_{ij}^m and W_{ij}^m variables. Actually, the problem can be correctly formulated with either X_{ij}^m or W_{ij}^m variables only. The constraint set (3) is redundant but useful for Lagrangean relaxation. Constraint set (4) guarantees that total flow does not exceed link capacities. Constraints (5) and (6) impose upper bound on the number of messages belonging to each priority class for each link. Constraint sets (7), (8) and (9) are the integrality constraints of the decision variables.

III. Solution Procedure

Problem P is a combinatorial optimization problem with non-linear constraints. It is known to be a NP-complete problem. So we propose a heuristic based on the Lagrangean relaxation by dualizing constraint set (3) using non-negative multipliers α_{ij}^{m} for all $(i,j)\in E$ and $m\in(M_1,M_2)$, then further dualizing constraints (5) and (6) using non-negative multipliers ψ_1 and ψ_2. The resulting Lagrangean relaxation of Problem P is further simplified by decomposing it into several message sub-problems and link sub-problems. Each message sub-problem resulting from the Lagrangean relaxation can be solved by solving the shortest path problem from *O(m)* to *D(m)* using the non-negative multipliers α_{ij}^{m} as the cost of the links. If the revenue from the message is greater than the cost of the shortest path, then the message is routed through that path, otherwise, the message is not routed and we set $Y^m = 0$ and $W_{ij}^{m} = 0 \ \ \forall(i,j)\in E$. On the other hand, for solving each link sub-problem we relax the integrality constraints and solve the continuous version of the problem using a greedy procedure. The solution obtained from solving each sub-problem is added up to give the upper bound of the optimal value of problem P. The feasible solution gives the lower bound of the maximization problem. The difference between the upper and the lower bound gives the gap in the solution and is a measure of how close the algorithm can approximate the optimal solution. The gap is usually calculated as a percentage gap. Like all relaxation procedures, the success of a tight lower bound depends heavily on the ability to generate good Lagrangean multipliers. In practice, the subgradient optimization method is used to obtain good values of the multipliers.

IV. Numerical Experimentation

In order to test the effectiveness of the solution procedure we conduct several numerical experiments. The experiments are conducted using networks where number of nodes is 10, 15, 20, and 25 respectively. The networks are generated in such a fashion that each node has a degree equal to 2, 3, or 4 with probability of 0.6, 0.3, and 0.1 respectively. The network is assumed to be made of OC4 links and has capacity of 192 Mbps. We perform some experiments using different values of link capacities as well. After the generation of the networks, we generate the message tables. The total number of messages in the message table is dependant on the number of nodes of the network and is equal to *k*N*(N-1)*, where *k* is the percentage of messages whose value is controlled by the network designer. For the base case, *k* is assumed to be 0.6. Hence, for the 10-node case, the total number of messages generated is equal to 54. We perform some experiments when the value of *k* is changed from 0.4 to 0.8 in steps of 0.1. The message table lists two types of messages and indicates the origin and the destination node for each node belonging to each class. The ratio of the number of messages belonging to lower priority to the higher priority is taken to be 80:20 in the base case. We perform experiments when the ratio of messages is changed to 90:10, 70:30, 60:40, and 50:50 as well. The messages belonging to the higher priority class are more demanding and the demand is assumed to be uniformly distributed between 30 Mbps and 40 Mbps. On the other hand, for messages belonging to the lower priority class the demand is uniformly distributed between 5 Mbps and 10 Mbps.

The revenue generated from the routing of the two classes of messages are also assumed to be uniformly distributed between [10,25] and [30,50] for the lower and the higher priority class respectively. The upper bound on the number of higher and lower priority messages on each link is taken to be 800 and 400 respectively. In our experiments we obtain the feasible solution, the gap between the feasible solution and the Lagrangean solution, the maximum and average utilization of the links, and the number of higher and lower priority messages that are routed over the network.

In the first set of experiments, depicted in Table 1, the experiments are conducted for networks where number of nodes is 10, 15, 20, and 25. The networks have a link capacity of 192 Mbps. The ratio of higher priority to lower priority message is 80:20. The demand for higher priority message is U[30,40] Mbps and the demand for lower priority message is U[5,10] Mbps. For this experiment, the percentage of messages (*k*) is increased in steps of 0.1 from 0.4 to 0.8. As *k* is increased the number of messages in each network increases. For example, for the 25 node case, when *k*=0.8, the total number of messages is 270. Of these 96 messages belong to the higher priority class and 384 messages belong to the lower priority class. From Table 1, we can observe the following. The feasible solution increases with the increase in *k* across all networks. The percentage gap remains reasonably low for all cases expect when *k*=0.8. In those cases, the average utilization of the links increases a lot and there is not enough bandwidth to route all messages. It is also to be noted that for all cases where gap is greater than 5% the maximum utilization reaches 100%, which implies that for such a choice of *k*, at least one bottleneck link is obtained for the network. Since the messages belonging to the higher priority class has preemptive priority over messages belonging to the lower priority class, the algorithm preferably routes higher priority messages. Hence, we note that for high values of *k*, greater number of lower priority messages is dropped compared to higher priority messages, though higher priority messages are more demand intensive.

Traffic ratio	Feasible solution	Percent. gap	Maximum utilization	Average utilization	Routed calls for priority 1	Total calls for priority 1	Routed calls for priority 2	Total calls for priority 2
10 node								
90:10	1070.93	1.13	86.46	38.47	49	49	5	5
80:20	1164.87	1.05	100	45.73	43	43	11	11
70:30	1377.88	1.82	100	61.56	37	38	16	16
60:40	1329.4	4.89	97.92	73.75	30	32	21	22
50:50	1401.66	7.51	100	72.60	23	27	26	27
15 node								
90:10	2447.80	2.43	98.96	56.71	109	113	13	13
80:20	2712.57	1.73	100	74.10	99	101	25	25
70:30	2791.65	9.79	100	79.13	71	88	38	38
60:40	2732.05	16.73	100	81.31	62	76	42	50
50:50	2862.03	21.55	100	84.10	43	63	53	63
20 node								
90:10	4387.67	1.96	98.44	47.36	199	205	23	23
80:20	4853.35	4.81	98.96	58.64	170	182	46	46
70:30	5050.04	7.58	99.48	64.89	138	160	68	68
60:40	5001.76	16.84	100	71.97	98	137	85	91
50:50	5418.40	19.81	100	76.28	85	114	97	114
25 node								
90:10	7251.48	0.19	94.79	41.78	324	324	36	36
80:20	7663.60	1.74	100	44.36	280	288	72	72
70:30	8187.08	3.45	98.96	56.37	234	252	108	108
60:40	8626.35	5.29	99.48	65.10	188	216	143	144
50:50	9038.43	9.12	100	71.04	157	180	165	180

Table 2: Impact of ratio of low to high priority traffic

The impact of capacity of links on the generated solution is reported in Table 3. Our prior experiments assumed that all links in the network have a capacity equal to 192 Mbps which corresponds to an OC4 link. Next, we experimented with two more link capacities – 96 Mbps which represents OC2 links and 500 Mbps which represents OC9 links. Table 3 clearly shows that OC2 link is not suitable for our choice of problem parameters which are kept exactly same as the earlier two experiments and with *k*=0.6 and ratio of low to high priority traffic fixed at 80:20. Our solution technique fails to generate high quality solutions for a link capacity of 96 Mbps and this leads to a large number of dropped messages. Between 192 Mbps and 500 Mbps it seems that the choice of OC9 links is overkill. When the link capacity is 500 Mbps, the average utilization remains less than 20%, which implies that a large amount of bandwidth is wasted.

Capacity	Feasible solution	Percent. gap	Maximum utilization	Average utilization	Routed calls for priority 1	Total calls for priority 1	Routed calls for priority 2	Total calls for priority 2
10 node								
96	972.55	21.11	100	67.08	37	43	8	11
192	1164.87	1.05	100	45.73	43	43	11	11
500	1164.96	1.04	41.20	17.67	43	43	11	11
15 node								
96	2147.13	28.94	98.96	69.73	80	101	16	25
192	2712.57	1.73	100	74.10	99	101	25	25
500	2736.83	0.49	48	19.29	101	101	25	25
20 node								
96	3423.32	47.76	98.96	59.07	107	182	36	46
192	4853.35	4.81	98.96	58.64	170	182	46	46
500	5046.42	0.251	50	16.13	182	182	46	46
25 node								
96	5181.45	50.20	100	61.23	172	288	49	72
192	7663.60	1.74	100	44.36	280	288	72	72
500	7770.06	0.16	90.4	17.98	288	288	72	72

Table 3: Impact of capacity of links

V. Conclusion

In this paper we provide an integer programming based formulation for the bandwidth packing problem with two priority classes. Since the problem is NP-complete, we use a Lagrangean relaxation based heuristic for solving this problem. We solve it numerically using networks of different sizes. Several experiments are conducted to check the impact of percentage of messages, ratio of low to high priority traffic, and capacity of links on the solution and the algorithm is able to generate high quality solutions under different situations. The experiments show how various factors affect the generated revenue and the routing of messages.

References

[1] Amiri, A., Barkhi, R. “The multi-hour bandwidth packing problem,” *Computers and Operations Research*, 2000, 27, 1-14.

[2] Amiri A., Rolland, E., Barkhi R. “Bandwidth packing with queuing delay costs: bounding and heuristic solutions procedures,” *European Journal of Operational Research*, 1999, 112, 635-645.

[3] Anderson, C.A., Fraughnaugh, K., Parker, M., Ryan, J. “Path assignment for call routing: An application of tabu search,” *Annals of Operations Research*, 1993, 41, 301-312.

[4] Cox, L., Davis, L., Qui, Y. “Dynamic anticipatory routing in circuit-switched telecommunications networks,” in: L. Davis, Editor, *Handbook of Genetic Algorithms*, Van Nostrand/Reinhold, New York, 1991.

[5] Laguna, M., Glover, F. “Bandwidth packing: A tabu search approach,” *Management Science*, 1993, 39, 492-500.

[6] Park, K., Kang, S., Park, S. “An integer programming approach to the bandwidth packing problem,” *Management Science*, 1996, 42, 1277-1291.

[7] Parker, M., Ryan, J. “A column generation algorithm for bandwidth packing,” *Telecommunications Systems*, 1995, 2, 185-196.

[8] Rolland E., Amiri A., Barkhi R. “Queuing delay guarantees in bandwidth packing,” *Computer and Operations Research,* 1999, 26, 921-935.