ITC19/ Performance Challenges for Efficient Next Generation Networks 1195-1204
LIANG X.J. and XIN Z.H.(Editors)
V.B. IVERSEN and KUO G.S.(Editors)
Beijing University of Posts and Telecommunications Press

Network Capacity Optimization for Latency-Sensitive Traffic in Multi-service IP Networks*

Sanaa Sharafeddine[1], Anton Riedl[2] and Thomas Bauschert[3]

[1]Institute of Communication Networks
Munich University of Technology
Arcisstr. 21, 80290 Munich, Germany
sharafeddine@mytum.de

[2]Department of Physics, Computer Science and Engineering
Christopher Newport University
One University Place, Newport News, VA 23606
riedl@pcs.cnu.edu

[3]Siemens AG
Information and Communication Networks (ICN)
Hofmannstr. 51, 81379 Munich, Germany
thomas.bauschert@siemens.com

Abstract: In this paper, we address the problem of capacity assignment for latency-sensitive traffic in multi-service IP networks. The capacity assignment (CA) problem is formulated as an optimization problem with nonlinear constraints, where the total link capacity shares allocated to the given service are minimized subject to performance constraints in terms of end-to-end delay requirements. We introduce a new extension to traditional CA problems by setting a statistical bound on the maximum end-to-end delay experienced by all packets associated with the given latency-sensitive service. For a sample IP network scenario with VoIP traffic, we show that only 3% of additional capacity as compared to the classical approach is required to keep the probability of exceeding the maximum end-to-end delay threshold below 0.1%.
Keywords: capacity assignment, quality of service, latency-sensitive traffic in IP networks.

1. INTRODUCTION

Latency-sensitive services characterized by real-time communications are migrating to IP networks promising significant cost-savings, simplified administration and new revenue sources. Traditionally, real-time communications had their dedicated bandwidth throughout the circuit-switched network. This fact has provided a premium quality to the ongoing com-

* This work is funded by Siemens AG, ICN EN

munications in a reliable and robust environment. The IP network, on the other hand, is packet-switched and its philosophy has been centered on other types of services characterized by non-real-time communications. One crucial factor for the perceived quality of the newly emerging latency-sensitive services in IP networks is the end-to-end delay of traffic samples between the sender (speaker) and the receiver (listener). The one-way end-to-end delay value should be constrained below a certain threshold to allow for interactive communications.

In order to guarantee a deterministic delay threshold for all connections at any time, unreasonable amount of capacity is required leading to very low network utilization and, thus, waste of resources [1]. In this work, we consider a *probabilistic* bound for the *maximum* network delay that can occur from any ingress node to any egress node. This helps in saving great amount of capacity, while still achieving highly guaranteed quality of service (QoS). We define the maximum delay at one node as the time experienced by the packet that happens to be queued last in the buffer when the buffer is occupied the most. On an end-to-end basis (ingress-to-egress), the maximum network delay is then the maximum waiting time at all nodes along a path. The idea behind the concept of the maximum waiting time is: if we are aware of the packets that experience the maximum delay among all packets, we can practically protect all other packets of the active connections from extra delay and thus from quality degradation. Based on the probabilistic network delay constraints, we evaluate the amount of capacity required on each network link while minimizing the total network cost. In the literature, the network design problem addressing the optimum determination of link capacities is often called capacity assignment (CA) problem [2].

In packet-switched networks typically supporting connectionless non-real-time communications, traditional CA problems aim for minimizing the total cost of the network while satisfying certain bounds in the average packet delay which is the average time experienced by a packet traveling from source to destination. In circuit-switched networks typically supporting connection-oriented real-time communications, traditional CA problems aim for minimizing the total cost of the network while satisfying certain bounds in the maximum call blocking probability. However, with the migration of real-time services into IP networks, the most relevant performance criterion is the *maximum* (and not just the average) end-to-end delay in addition to the traditional maximum call blocking probability for these types of services. The novel aspect about our approach as compared to existing CA research on latency-sensitive services is that we include the *maximum* end-to-end delay constraints in the CA problem.

Previous studies addressing the same CA problem still focus on the average packet delay and specifically the queuing part of the delay [3,4]. In [4], CA is performed in a DiffServ network supporting two traffic classes where each class imposes different performance requirements (one has a non-real time nature and the second has a real-time nature). Performance constraints of both classes are determined by the average queuing delay of the priority queue, thus, guaranteeing an *average* delay for real-time traffic. Though the average delay is preserved, packets can undergo long delays causing severe deterioration in the communication quality making it intolerable.

In the next section, the network model along with the assumptions taken in this work is presented. Section 3 formulates the CA problem indicating its objective and constraints and identifies its complexity. We propose a reformulation of the problem based on link decomposition and describe the corresponding solution approach. In section 4, capacity assignment is carried out first on a simple two-link network and then on a realistic mesh network scenario. Finally, Section 5 concludes the paper.

$\hat{W}_l =$ maximum waiting time of traffic on link $l \in L$

$\hat{K}_l =$ maximum number of active connections on link $l \in L$

$\pi_s =$ set of links constituting the path of demand pair $s \in \Psi$

$D_l =$ delay budget allocated to link $l \in L$

$D_{\text{e2e}} =$ requested end-to-end delay threshold

$P_{\text{out}} =$ outage probability indicating how often the delay threshold is violated by the packet experiencing the maximum waiting time

$r =$ bitrate of one connection

Our CA problem (P) can be stated as follows:

$$Z = \min\left(\sum_{l \in L} C_l\right), \tag{2}$$

subject to:

$$\mathrm{P}\left\{\hat{W}_{(s)} > D_{\text{e2e}}\right\} \le P_{\text{out}} \qquad \forall s \in \Psi, \tag{3}$$

$$C_l \ge \hat{K}_l \cdot r \quad \forall l \in L. \tag{4}$$

In problem (P), (2) denotes the objective function that aims for minimizing the total network capacity over all links. Constraints in (3) represent the performance criterion that requires the maximum end-to-end delay $\hat{W}_{(s)}$ of all packets belonging to traffic demand pair s exceed the given threshold D_{e2e} only with probability P_{out}. Furthermore, the required capacity C_l of each link l has to be at least equal to the bitrate sum of the maximum number of connections traversing l as the system might become instable otherwise.

In [1] it is shown that the distribution of the maximum waiting time $\hat{W}_l$ for VoIP traffic on a single link l is a function of various parameters as shown below:

$$\mathrm{P}\left\{\hat{W}_l > D_l\right\} = f\left(\hat{K}_l, C_l, r, D_l\right). \tag{5}$$

The maximum end-to-end waiting time $\hat{W}_{(s)}$ of demand pair s can be computed as

$$\hat{W}_{(s)} = \sum_{l \in \pi_s} \hat{W}_l, \tag{6}$$

where $\hat{W}_l$ is the maximum waiting time experienced at link l of path π_s. Based on the assumption of mutually independent hops [2], the end-to-end maximum waiting time distribution of the target flow can be obtained by performing convolution of the maximum waiting-time experienced at each link along the flow path, i.e.:

new optimization variables are the individual delays allocated to each of the links in the network. For each D_l, C_l is calculated by solving (5) numerically for $\mathrm{P}\left\{\hat{W}_l > D_l\right\} \le P_{\text{out}}$, where $\hat{W}_l$ is the maximum waiting time experienced at link l. If the same P_{out} were used in both problems, as an end-to-end value in (P) and as a per-hop value in (P′), (P′) would yield an end-to-end outage probability less than P_{out}. Thus, (P′) represents a sub-optimal solution yet a more conservative one as compared to (P). Constraints (9) constitute a set of constraints, each corresponding to one traffic demand. As some traffic demands may follow completely independent paths from those followed by other traffic demands, (P′) may be divided into different sub-problems whose solutions make up the final solution of (P′). This will yield significant reduction in running time especially when multiprocessing is supported.

At this point, (P′) is reformulated into an unconstrained optimization problem using Multiplier and Lagrangian method that belongs to penalty function techniques and that constructs the new objective function in the form of Generalized Lagrangian function L_G expressed as:

$$L_G\left(\Gamma,\Lambda\right)=\sum_{l\in L}C_l-\frac{1}{4\alpha}\sum_{s\in\Psi}\left(\lambda_s^4-\left(\lambda_s-\alpha\left(D_{e2e}-\sum_{l\in\pi_s}D_l\right)\right)_+^4\right), \tag{11}$$

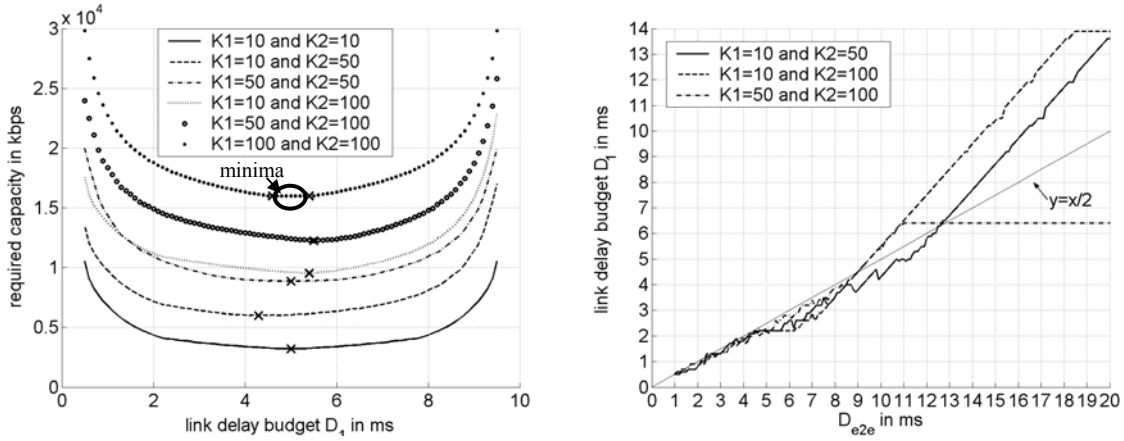
where Λ is the vector of Lagrange multipliers λ_s, $s\in\Psi$, α is a sufficiently large number and $\left(x\right)_+=\max\left(x,0\right),\forall x\in\mathbb{R}$. Note that constraints (10) are not included in Generalized Lagrangian function L_G since it is obsolete at this point: in (P′), the optimization variables are the link delay budgets which are fed into (5) to compute the link capacity; however, (5) assures implicitly that the link capacity is at least $\hat{K}_l\cdot r$, $\forall l\in L$. In (P), on the other hand, the solution approach was different due to the fact that the link capacities were the optimization variables and so the constraint $C_l\ge\hat{K}_l\cdot r$ has to be taken into consideration whenever a new set of link capacities is selected. Finally, (P′) is solved by minimizing $L_G\left(\Gamma,\Lambda\right)$ by means of the Simplex method of Nelder and Mead.

4. NETWORK EXAMPLES AND RESULTS

4.1 One-Path Network

In this section, we investigate the CA problem for one-path network to gain insight about its fundamental properties. Later on, these scenarios can serve as building components for larger-scale networks.

We first consider a one-path network with a single link having K active connections where K = 10, 50 and 100 active connections. Each connection carries voice traffic generated by G.711 coders with 80 kbps bitrate (at the IP layer). Figure 1 plots the required capacity in terms of the link delay budget for each value of K. As expected, link capacity can be reduced if higher link delays are acceptable. Especially for low delay values (i.e., for very strict delay requirements) a slight increase in acceptable delay can lead to a significant decrease in required capacity. However, once the capacity has been brought down to the sum of the bitrates of active connections, no further gains are possible. This is due to (10), which guarantees that



(a) (b)

Figure 2. (a) Capacity sum of links 1 and 2 versus D_1 **(b)** D_1 vs. D_{e2e}

Examining the figure, we can make the following observations. While these observations correspond to the curve of $K_1 = 50$ and $K_2 = 100$, they also apply to other combinations of K_1 and K_2 values in a similar way.

- At low D_{e2e} (up to 4.7 ms), the curve almost overlaps with $y = x/2$ line. Hence, the delay budget is allocated equally to both links irrespective of the number of connections.
- If D_{e2e} falls between 4.7 ms and 9 ms, the link with more connections i.e. link 2 is granted more delay. This is more obvious when $K_1 = 10$ and $K_2 = 100$.
- If D_{e2e} falls between 9 ms and 11 ms, link 1 is allocated more delay than link 2, i.e., the link with fewer connections is granted more delay. This is explained by the fact that link 2 is already capable of serving all its connections with the minimum capacity in 4.6 ms of delay. The slope of this part of the curve gets back to 1 again as all delay above the 4.6 ms is totally allocated to link 1 to reduce its needed capacity and thus the total capacity.
- If D_{e2e} exceeds 11 ms, then both links reach their minimum capacity. Hence, any delay budget above the 11 ms has no extra advantage in decreasing the total costs.

4.2 Sample Mesh Network Scenario

The solution approach presented earlier is applied on a sample network scenario depicted in Figure 3 and denoted as *N11*. The network scenario represents an enterprise backbone network consisting of 11 nodes and 48 unidirectional links. At each node, 1000 users are connected where each user is expected to generate voice traffic load of either 0.1 erl or 0.2 erl in the busy hour. Each node communicates with five other nodes selected randomly.

Solving (P′) as described in Section 3, the capacity share required to service voice traffic on all links *l* of network *N11* with minimum costs is evaluated. (P′) can be compared to other approaches namely: *erlang* calculations, *per-hop* (P_{out} *= 0%)* calculations and *per-hop* (P_{out} *= 0.1%)* calculations. The former refers to the classical way of dimensioning used in circuit-switched networks and it is based on Erlang formulae which only considers the desired call blocking probability without any concerns about the connection traffic delay from source to destination. The *per-hop* approaches, on the other hand, refer to calculations based on per-hop QoS constraints. For each link, the capacity is determined separately according to (5) where a mapping procedure transforms the end-to-end QoS value into a per-hop value. We assume that the end-to-end delay constraint D_{e2e} is partitioned into equal per-hop delay constraints D_{hop}, (i.e. D_{e2e} is divided by the hop-count of the longest path in the network) and that the per-

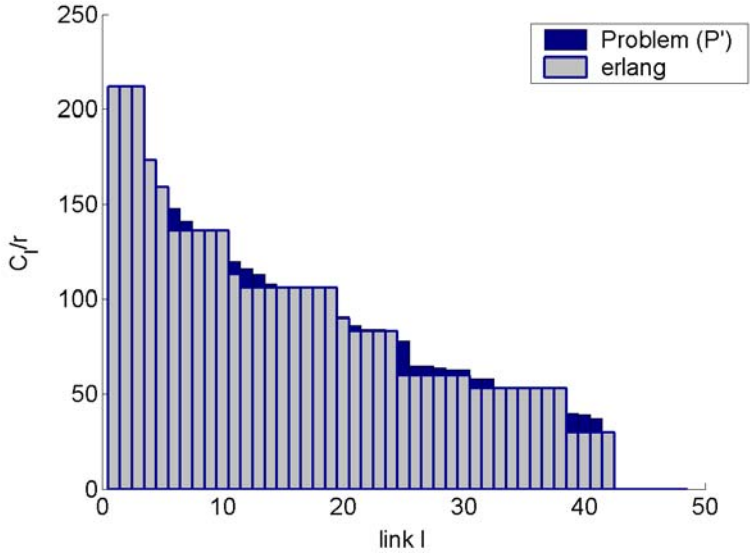


Figure 4. Maximum channel histogram

5. CONCLUSIONS

In this paper, we addressed the problem of capacity assignment for real-time services over IP networks and presented it as an optimization problem aiming at minimizing the total link capacities subject to nonlinear performance constraints (probabilistic end-to-end delay threshold). Because of its numerical complexity, the original problem is reformulated applying the link decomposition approach. The problem is then transformed into an unconstrained nonlinear optimization problem using the Multiplier and Lagrangian method and is finally solved by means of the Simplex method of Nelder and Mead. Applying our method to a sample network scenario we show that a moderate 3% increase in total capacity compared to *erlang* calculations keeps the probability of exceeding the desired end-to-end delay threshold below 0.1%.

REFERENCES

1. S. Sharafeddine, N. Kongtong and Z. Dawy, “Capacity Allocation for Voice over IP Networks Using Maximum Waiting Time Models,” in 11th International Conference on Telecommunications (ICT), Fortaleza, Brazil, Aug. 2004.
2. L. Kleinrock. Queueing Systems: Computer Applications , Vol. 2, John Wiley and Sons, New York, 1975.
3. T. Ng and D. Hoang, “Joint Optimization of Capacity and Flow Assignment in a Packet-switched Communications Network,” in IEEE Transactions on Communications, Vol. Com-35, No. 2, Feb. 1987.
4. K. Wu and D. Reeves, “Capacity Planning of DiffServ Networks with Best-Effort and Expedited Forwarding Traffic,” in IEEE International Conference on Communications (ICC), Anchorage, USA, May 2003.
5. S. Sharafeddine and Z. Dawy, “A Capacity Margin for IP Networks with QoS Constraints and Uncertain Demands,” in 9th IEEE Symposium on Computers and Communications (ISCC), Alexandria, Egypt, June/July 2004.
6. P. D. Bertsekas, Nonlinear Programming: 2nd Edition, Athena Scientific, U.S.A. 1999.