Journal of ELECTRICAL ENGINEERING, VOL. 54, NO. 9-10, 2003, 225–230

SLA CREATION IN RELATION TO RESOURCE MANAGEMENT

Srećko Krile * — **Slavko Šarić** **

While DiffServ architecture solves the scalability problem of QoS provisioning, it fails to be the solution for end-to-end provisioning. A combination of IntServ/RSVP signaling with aggregate traffic handling mechanisms could solve such deficiencies. To obtain quantitative end-to-end guarantees in DiffServ architecture, based on traffic handling mechanisms with aggregate flows, some kind of congestion control through negotiation process (in new SLA creation) is necessary. In this paper an efficient heuristic algorithm for end-to-end congestion control for N quality-of-service levels (service classes) is being developed. The problem is seen as an expansion problem of link capacities in given limits from a common source. QoS routing is in firm correlation with resource reservation process (resource management). If the optimal expansion sequence has any expansion value that exceeds allowed limits (link capacity) a new SLA cannot be accepted or must be redefined.

K e y w o r d s: quality of service in DiffServ networks, constrained-based path selection, end-to-end QoS routing, SLA creation, traffic handling mechanisms.

1 INTRODUCTION

The classification of the aggregated flows (on entrance in DiffServ/MPLS cloud) is performed according to the SLA (Service Level Agreement) signed between a customer and the network operator (ISP). Each SLA contract specifies how much traffic a user may send (service class - bandwidth, delay *etc.*) and defines a time period for utilization of that service. But very important element for efficient end-to-end QoS routing is good prediction of traffic demands that is defined with limited number of SLA agreements. So, in the process of SLA creation the problem of new SLA acceptance for network operator exists. Some kind of congestion control is necessary and can be a crucial element for efficient end-to-end QoS routing in DiffServ networks.

Service management (SrvMgt) and traffic engineering (TE) functions do not act in isolation. For example, TE functions provide the bases grounds on which SrvMgt functions operate, while SrvMgt functions set the traffic-related objectives for the TE functions to fulfill. Because of forecasted traffic demands (traffic matrix -TM), which specify anticipated QoS traffic demands between network edges, TE functions produce the resource availability matrix (RAM). In the same way SLA creation is in correlation with QoS routing, resource reservation mechanisms and admission control process (service invocation). It means that every new SLA acceptance directly influences on traffic handling mechanisms with other traffic flows (existing SLAs) because the network resources are limited and they are in strong correlation. It means that only through negotiation process a new SLA can be established, but it has strong influence on the subsequent SLA creation.

To provide explicit QoS support, MPLS makes use of certain elements in the IntServ and DiffServ approaches. Necessity of some combination of IntServ (management per-flow) and DiffServ (management with aggregate flows) clearly represents a trade-off between service granularity and scalability: as soon as flows are aggregated, they are not as isolated from each other as in IntServ architecture. In the moment of service invocation (explicit activation) the optimal routing sequence information for that traffic flow can be sent with RSVP (Resource Reservation Protocol) signaling protocol to MPLS routers, to ensure end-to-end guarantees. Sufficient resources must be available at any moment because former congestion control in the SLA negotiation process are made. For services activation without signing a SLA only QoS differentiation on DiffServ approach (on the best-effort basses) can be applied and wanted service can be established if enough resources are available at the time of service utilization.

The network operator or service provider has to check if any traffic congestion is present on the path between edge routers. We need very effective tool to check it, specially for huge and complex networks with many interior routers on the path. Some important papers about end-to-end QoS routing are: Giordano, Salsano and Ventre [1], Christin and Liebeherr [2] and Lu and Faynberg [3]. In papers of Biswas, Izmailov [5] and Bouillet, Mitra and Ramakrishnan [6] some proposed solutions for congestion control problem are given.

* Polytechnic of Dubrovnik, Department of Electrical Engineering and Computing, Faculty of Transport and Traffic Engineering,, Cira Carica 4, 20000 Dubrovnik, Croatia, E-mail: srecko.krile@vdu.hr

** University of Zagreb, Vukelićeva 2, 10 000 Zagreb, Croatia, E-mail: draganp@fps.hr

First version of this paper was presented at SympoTIC ´ 03, 26-28 October 2003, Bratislava, Slovakia

ISSN 1335-3632

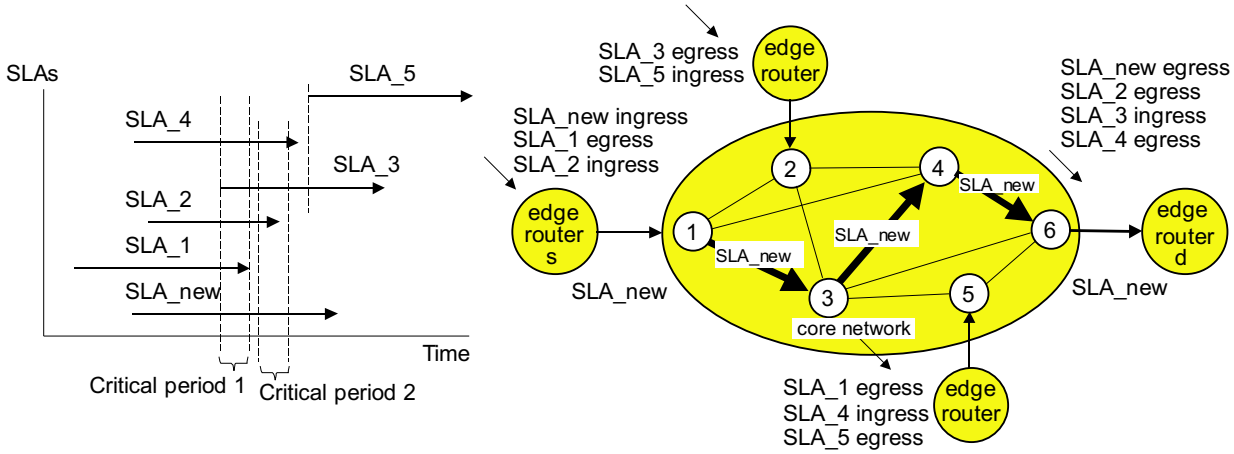

Fig. 1. An example of number of SLAs in definite period of time. The optimal routing sequence for new SLA need not to be the shortest path solution.

In section 2. The problem of new SLA creation and correlation with optimal resource management are investigating. Explanation of the mathematical model and heuristic approach for constraint-based path selection for new SLA creation is given in section 3.

2 NEW SLA CREATION PROBLEM

If network operator (ISP) wants to accept new SLA (between edge routers on the path) it has to be checked with congestion control algorithm related on limited link resources and predicted traffic (caused with former accepted SLAs). For each communication link in the network given traffic demands (consist of number of SLAs) can be satisfied on different QoS levels (*eg* used bandwidth). Optimal resource management can be seen as optimal link capacity expansion problem with expansion values in allowed limits (limited capacity).

If the optimal routing sequence has any link expansion with value that exceeds allowed limits (link capacity), it means that at time of service activation the congestion can be occurred and new SLA cannot be accepted or must be redefined through negotiation process. For example, if idle capacity on the link exists, conversion of traffic demands from one QoS level (service class) to another (only toward higher quality level) could be a good solution for the customer, but it might not be the optimal solution for the service provider (loss of capacity).

Predicted traffic demand increment (given in relative value) on input of each edge router, distributed to core routers on the path (for each QoS level), is the sum of all ingress and egress SLAs; see Fig. 1. and 2. and numerical test-example in Fig. 4. We suppose that link capacity between edge and core (internal) router is sufficient and it isn't matter of this work. In new SLA creation through negotiation process some congestion control must be done for traffic flow between edge routers (on each link), specially for definite period of time (critical moments); see fig 1. The optimal constraint-based path selection eliminates the possibility for traffic congestion. In contrary, new SLA cannot be accepted in that form.

3 MATHEMATICAL MODEL AND HEURISTIC APPROACH

Let $G(A, E)$ denote a network topology, where A is the set of nodes and E the set of links. The source and destination nodes (edge routers in domain) are denoted by s and d, respectively; see Fig 1. The number of QoS measures (*eg* bandwidth, delay) is denoted by z. QoS measures can be roughly classified into additive (*eg* delay) and nonadditive (*eg* available bandwidth). In case of an additive measure, the QoS value of the path is equal to the sum of the corresponding weights of the link along that path. For a nonadditive measure, the QoS value of the path is the minimum (or maximum) link weight along the path. In algorithm approach it is assumed that the network state information (*eg* traffic demands, link capacities) is temporary static.

Consider a network $G(A, E)$ where each link is characterized by z-dimensional link weight vector, consisting of z nonnegative QoS weights $(w_i(k,l), i = 1, \ldots, z, (k,l) \in E)$ as components. Given z constraints are denoted by $L_i, i = 1, \ldots, z$. Definition of the multi-constrained (MCP) problem is to find a path P from s to d such that

$$w_i(P) \,\mathrm{def} \sum_{(k,l)\in P} w_i(k,l) \le L_i \quad \text{for} i = 1, \ldots, z \tag{1}$$

In this paper we dealt about only one dimensional link weight vector, with only one constraint denoted by L.

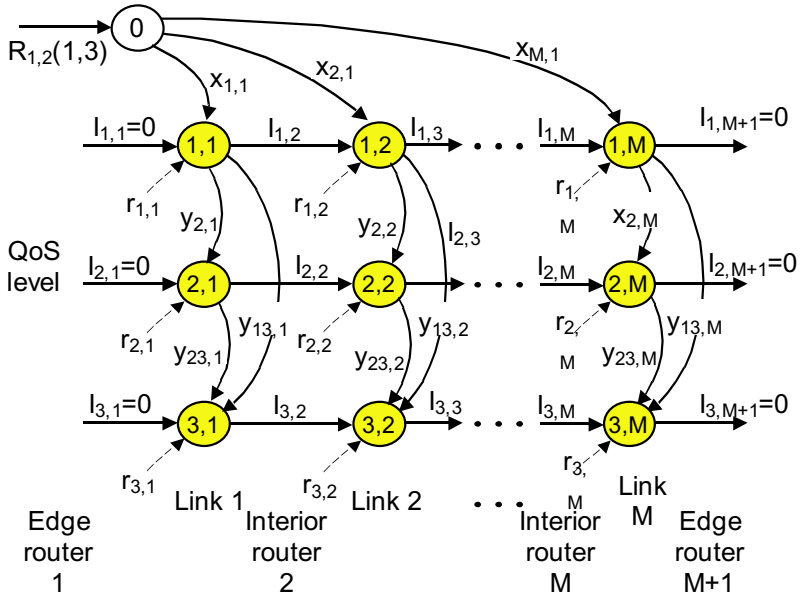

Fig. 2. A numerical flow representation of the CEP problem

Given constraints are limited bandwidth on each link on the path $L_m, m = 1, \ldots, M$. The link weight (cost) is the function of used capacity: lower used capacity (smaller bandwidth) gives lower weight (cost), taking care that given traffic demands can be satisfied. Nonlinear cost function is necessary if link weights are not positively correlated. The problem of the optimal QoS routing can be seen as the minimum cost network flow problem in the multi-commodity single (common) source multiple destination network. Such problem can be solved as the capacity expansion problem (CEP) without shortages. Partially expansions for each link are made from common source in given limits (maximal link capacity).

Transmission link capacities on the path between routers are capable to serve traffic demands for N different QoS levels (called facilities), for $i = 1, 2, \ldots, N$. Facility i is used primarily to serve demands for QoS level i, but it can be used to satisfy traffic demands for QoS level j. $(j > i)$. Re-routing of traffic demands towards higher QoS level is the same thing as facility conversion toward lower QoS level; see Fig. 2. Normally, it does not require physical modifications, it can be done by routing mechanisms and with no cost. Once converted from i to j, the amount of facility i becomes an integral part of facility j, but it can be rearranged to its original type at any time without cost. In this model conversion of traffic demand is permitted only in the direction toward higher QoS level. Usage of the new capacity (expansion), same as re-routing of traffic demands, can be used as stand alone strategies or can be combined together. If both strategies are necessary they are not substitutes but complements. The objective is to find optimal routing-policy that minimizes the total cost incurred over the whole path between edge routers (M interior routers and $M+1$ transmission links) and to satisfy given traffic demands. An example of the optimal expansion solution on the path with six interior routers for given traffic demands can be seen in Fig. 4. The optimal routing sequence on the path between edge routers could be the shortest path solution, but it is not necessary.

The flow theory enables separation of these extreme flows which can be a part of an optimal expansion solution from those which cannot be. With such heuristic approach we can obtain optimal result with significant computational savings. Fig. 2 gives a network flow representation of CEP for three QoS levels ($N = 3$) and M internal (core) routers included in the path. On that diagram the m-th row of nodes represents a possible link capacity state of each transmission link between routers for i-th QoS level. Link capacity values are positive only (idle capacity), and shortages are not allowed. Horizontal links between them represent the traffic flow between routers. Common node "O" is the source for used capacity (expansions), introducing the new capacity on the link. Vertical links represent re-routing of traffic demands (equal to capacity conversion of facility).

In the mathematical model of CEP the following notation is used; see diagram on Fig. 2.

i, j and k are QoS level. The N levels are ranked from $1, 2, ..., N$, and quality decreases with higher number.

m is the order number of transmission link on the path, connecting two successive routers. Path consists of M transmission links $(m = 1, ..., M)$ between $M + 1$ routers.

u, v are the order number of capacity points in the subproblem, $1 < u, \ldots, \nu < M + 1$. r_{im} is traffic demand increment for additional capacity of facility i (appropriate QoS level) on transmission link m. For convenience, r_{im} are assumed to be integers. The demands can also be satisfied by facility with higher QoS level (lower i).

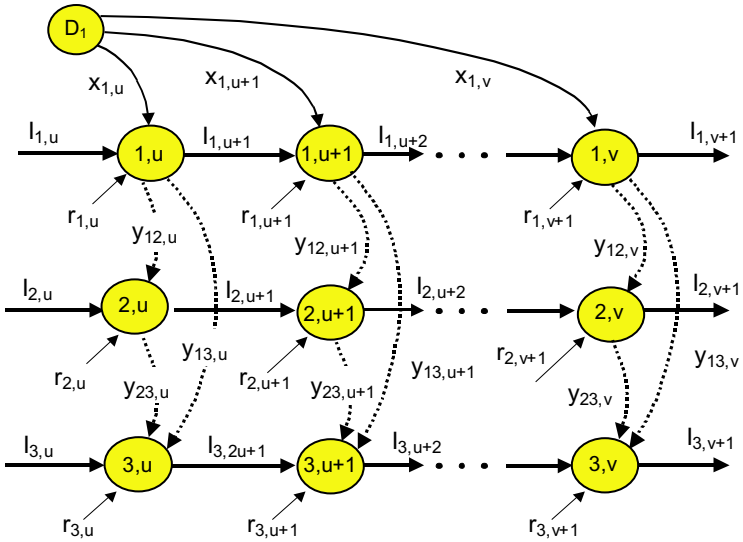

Fig. 3. A network flow representation of a sub-problem for N=3

I_{im} is the relative amount of idle capacity of facility i on the link m, related on the link before. We assumed that initially there is no capacity shortage between edge router and the interior router, that means that capacity is sufficient $I_{i1} = 0$, $I_{i,M+1} = 0$.‘

WI_{im} is the upper limit for capacity of facility i on link m.

kI_i is the lowest step of possible facility change for QoS level i.

x_{im} is the amount of used capacity for facility i on transmission link m.

Wx_{im} is the upper limit for allowed expansion for facility i on transmission link m (maximum for usage of capacity with no cost).

y_{ijm} is amount of re-routing capacity (conversion) of facility i on link m, redirected to satisfy the traffic demands of lower QoS level j.

The CEP problem can be formulated as follows

$$min\{\sum_{t=1}^{M}(\sum_{i=1}^{N} c_{im}(x_{im}) + h_{im}(I_{i,m+1}) + \\ + \sum_{j=i+1}^{N} g_{ijm}(y_{ijm}))\} \qquad (2)$$

so that we have

$$I_{m+1} = I_{im} + x_{im} - \sum_{j=i+1}^{M} y_{ijm} - r_{im}$$
$$I_{im} = I_{i,M+1} = 0 \qquad (3)$$

for $m = 1, 2, \ldots, M$, $i = 1, 2, \ldots, N$, $j = i+1, \ldots, N$.

The total cost on the path from edge to edge router includes some costs: the cost for capacity expansion $c_{im}(x_{im})$, the idle capacity cost $h_{im}(I_{i,m+1})$ as penalty cost to force the usage of the minimum link capacity (prevention of idle capacity), and the re-routing cost of traffic demands $g_{ijm}(y_{ijm})$. For expansion of link in allowed limits we can set the cost to zero. Costs are often represented by the fix-charge cost or with constant value. We assume that all cost functions are concave and non-decreasing, reflecting economies of scale, and they can change for appropriate link. With costs parameters we can influence on the optimization process, looking for the most appropriate routing solution.

3.1 Definition of the Capacity Point

Generalizing the concept of the capacity state for transmission link m (capacity on the input of interior router) in which the capacity state of each link is known within defined limits and which at least one capacity state satisfies $I_{im} = 0$, we define as a *capacity point*. In (4) α_m denotes vector of capacities I_{im} for all QoS levels (facility types) on link m, and we call it capacity point.

$$\alpha_m = (I_{1m}, I_{2m}, \ldots, I_{Nm}) \qquad (4)$$

$$\alpha_0 = \alpha_{M+1} = (0, 0, \ldots, 0) \qquad (5)$$

Each column in the flow diagram from Fig. 2 represents a capacity point, consisting of N capacity state values. Equation (5) implies that idle capacities or capacity shortages are not allowed on the link between edge and interior (core) router.

3.2 Solving of the Sub-problem (CES)

Associated value between two capacity points, that represents minimum cost $d_{u\nu}(\alpha_u, \alpha_{\nu+1}$ we denoted as

CES (*Capacity Expansion Sub-problem).* In CEP we have to find many cost values $d_{u\nu}(\alpha_u, \alpha_{\nu+1})$ that emanate two capacity points, from each node (u, α_u) to node $(v+1, \alpha_{v+1})$ for $v > u$. Most of the computational effort is spent on computing the sub-problem values. Any of them, if it cannot be a part of the optimal sequence, is set to infinity.

Let the generalized capacity expansion sub-problem CES $(\alpha_u, \alpha_{\nu+1})$ be the expansion problem for N facilities $i = 1, 2, \ldots, N$ on the path between routers $u, u+1, \ldots, \nu$. Then $d_{u\nu}(\alpha_u, \alpha_{\nu+1})$ is as follows

$$min\{\sum_{m=u}^{\nu}\Big(\sum_{i=1}^{N} c_{im}(x_{im}) + h_{im}(I_{i,m+1}) + \sum_{j=i+1}^{N} g_{ijm}(y_{ijm})\Big)\} \quad (6)$$

where

$$I_{i,\nu+1} = I_{iu} + D_i - R_i(u, \nu) \quad (7)$$

$$R_i(m_1, m_2) = \sum_{m=m_1}^{m_2} r_{im} \quad (8)$$

$$D_i = \sum_{m=u}^{\nu} x_{im} - \sum_{j=1}^{N} y_{ijm} \quad i \neq j \quad (10)$$

for $i = 1, 2, \ldots, N$, $m = 1, 2, \ldots, M$ Let C_m be the number of capacity point values at router position m (link between core routers), $C_1 = C_{M+1} = 1$, and the total number of capacity points is

$$C_p = \sum_{m=1}^{M1} C_m \quad (11)$$

The total number of connections between capacity points is

$$N_d = \sum_{i=1}^{m} C_i \left[\sum_{j=i+1}^{M+1} C_j\right] \quad (12)$$

Links between two successive capacity points represents minimum costs $d_{u\nu}(\alpha_u, \alpha_{\nu+1})$. Suppose that all links are known, the optimal solution for CEP can be found by searching for the optimal sequence of capacity points and their associated link state values of interior routers. As shown in Figure 3. problem can be formulated as a shortest path problem for an acyclic network in which the nodes represents all possible values of capacity points. Than Dijkstra's algorithm can be applied.

It is very important to reduce that number of capacity points and that can be done through imposing of appropriate capacity bounds or by introduction of adding constraints.

3.3 Single Location Expansion Problem

The approach described above requires solving repeatedly a certain single location expansion problem (SLEP). Let $\mathrm{SLEP}_i N\,(m, D_i, \ldots, D_N)$ be associated with router location m for facility $i, i+1, \ldots, N$ and corresponding values of capacity change value, see (10), denoted as $D_i, D_{i+1}, \ldots, D_N$.

Solving SLEP_{13} for three different QoS levels we have many changing solutions. Changes of all three facility types depend on each other because the channel resources are in firm correlation (channel speed). Traffic re-routings to higher QoS level is allowed (or capacity conversion to lower QoS level), that can be seen from Fig. 4.

Many different expansions and re-routing solutions can be derived, depending on D_i polarity, see [7]. Lot of them are not acceptable and are not part of the optimal sequence, that is the key for the heuristic approach. The heuristic algorithm in all test-examples can achieve the best possible result (near-optimal expansion sequence), but requires the computation effort of $O(M^3 N^4 R_i^{2(N-1)})$. The required effort for one sub-problem is $O(N^2 M)$. The number of all possible $d_{u\nu}$ values depends on the total number of capacity points. If there are no limitations on capacity state (WI_{im}) and expansion amount (Wx_{im}) the complexity of such heuristic approach is pretty large and increases exponentially with N.

4 CONCLUSIONS

The problem of new SLA creation in correlation with QoS resource management on the path in DiffServ/MPLS-based networks was investigated. At any moment traffic can be routed through LSRs (Label Switching Routers) without congestion only if existing link capacities are sufficient, or predicted traffic demands (SLAs) are not too large. In the process of new SLA creation possible congestion can be checked with proposed heuristic algorithm, based on mathematical model for the capacity expansion problem (CEP); see fig. 4. It means that such heuristic approach can be successfully applied for congestion control in the SLA creation process, that is in firm correlation with resource reservation mechanisms and admission control process. In other words, it is in function of resource management. At the time of service invocation (explicit activation) the optimal routing sequence information for that traffic flow can be sent with RSVP (Resource Reservation Protocol) signaling protocol to MPLS routers. It will ensure end-to-end QoS routing guaranties, improving DiffServ granularity. Heuristic algorithm is able to find the best routing solution for traffic caused by new SLA creation, showing the significant complexity savings in comparison with algorithm based on exact approach.

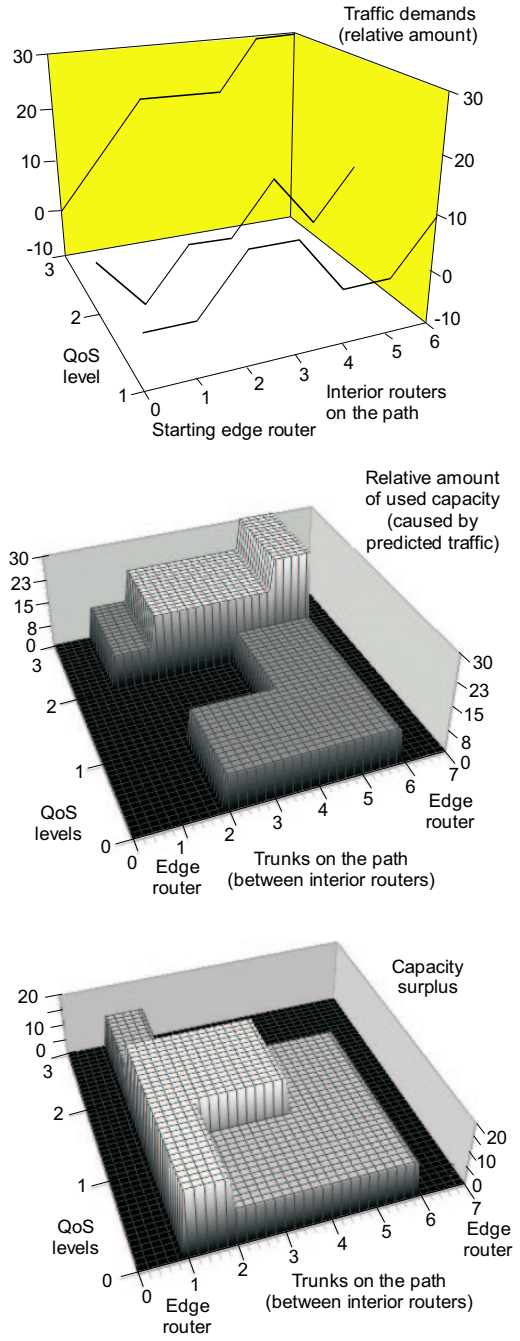

Fig. 4. With such resource management tool we can analyze the capacity situation on the path for new SLA creation. In example above used capacities don't exceed given limits (max. 30 channel units for each link), that means no congestion on the link exists

References

[1] GIORDANO, S.—SALSANO, S.—VENTRE, G.: Advanced QoS Provisioning in IP Networks, The European Premium IP Projects, Communications **41** No. 1 (2003), 30-36.

[2] CHRISTIN, N.—LIEBEHERR, J.: A QoS Architecture for Quantitative Service Differentiation, Communications **41** No. 6 (2003), 38-45.

[3] LU, H.—FAYNBERG, I.: An Architectural Framework for Support O QoS in Packet Networks, Communications **41** No. 6 (2003), 98-105.

[4] KRILE, S. Kos,M.—: A Heuristic Approach for Path Provisioning in Diff-Serv Networks, Proc. 7th ISSSTA (Int. Symp. on Spread-Spectrum Tech. & Application) – IEEE, Prag (2002), 692-696.

[5] BISWAS, K. Ganguly,—S.—IZMAILOV, R.: Path provisioning for service level agreements in differentiated services networks, ICC 2002 – IEEE International Conference on Communications **25** No. 1 (2002), 10631068.

[6] BOUILLET, E.—MITRA, D.—RAMAKRISHNAN, K. G.: The structure and management of service level agreements in networks, IEEE Journal on Selected Areas in Communications **20** No. 4 (2002), 691-699.

[7] LUSS, L. H.: Multifacility-type capacity expansion planning: algorithms and complexities, Opnl. Res. **35(2)** (1987), 249-253.

Received 3 October 2003

Srećko Krile (BSc, PhD) was born in 1957 in Dubrovnik, Croatia. He graduated from the Faculty of Electrical Engineering and Computing, University of Zagreb in 1980. in field of Telecommunication and information. In 1988, after postgraduate study on the same faculty, he won a master's degree with thesis "Algorithms for Optimal Capacity Expansion Planning of Telecommunication Network". In 2000. he took doctor degree on thesis: Optimal Capacity Sizing of Satellite Links in Mobile Maritime Networks. From 1983 to 1991 he worked on planning and maintenance of functional communication networks for government. From 1991. Till now he has been teaching on Polytechnic (Veleućilište u Dubrovniku) former Maritime Faculty. On this school he is senior lecturer on two departments (electrical engineering and computing and nautical department) and he was the first chief of the Electrical-Computing Engineering Department. He authored two books "Electronic Communications in Shipping",part I (ISBN 953-96858-4-2), and part II - Mobile satellite communications (ISBN 953-6705-13-3), further two booklets "GMDSS GOC" and "GMDSS ROC" for purpose of courses, and published many scientific papers.

Slavko Šarić. Biography not supplied.