New SLA Creation and Optimal Resource Management

Srećko Krile
University of Dubrovnik
Department of Electrical Engineering and Computing
Cira Carica 4, 20000 Dubrovnik, Croatia
Tel: 385 20 445-739, Fax: 385 20 435-590
srecko.krile@unidu.hr

Abstract: 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. For $N$ quality-of-service levels (service classes) an efficient heuristic algorithm for end-to-end congestion control is being developed. The problem is seen as an expansion problem of link capacities in given limits from a common source. 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 through a negotiation process.

Index Terms: 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 (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 may be sent (service class, bandwidth, delay etc.) and defines a time period for utilization of that service. 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.

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).

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 congestion control in the SLA negotiation process are made former.

In section 2. The problems 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) 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 (for example used bandwidth).

The optimal resource management can be seen as the optimal link capacity expansion problem with expansion values in allowed limits (capacity). If the optimal routing sequence has any link expansion with value that exceeds allowed limits (link capacity), it means that new SLA cannot be accepted or must be redefined through negotiation process.

Traffic demand (given in relative value as increment) on input of each edge router, represents the sum of all ingress and egress SLAs; see fig. 2. Congestion control must be done for traffic aggregated flows 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. We need very effective tool to check such congestion possibility in the network. Some important papers about that problem are [1], [4] and [5]. In the paper [2] such algorithm is the part of service management architecture.
3. The 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 (e.g. bandwidth, delay) is denoted by $z$.

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

$$w_i(P) \sum_{(k,l) \in P} w_i(k,l) \leq L_i \quad (3.1)$$

For $i = 1, ..., z$.

In this paper we dealt about only one dimensional link weight vector, with only one constraint denoted with $L_i$. Given constraints are limited bandwidths for each link on the path: $L_m, m = 1, ..., M$; see fig. 2. The link weight (cost) is the function of used capacity: lower used capacity (smaller bandwidth) gives lower weight (cost). The main condition is that given traffic demands must 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 (link capacity).

Transmission link capacities on the path between routers are capable to serve traffic demands for $N$ different QoS levels (called facilities), for $i = 1,2, ..., N$. 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. In this model conversion of traffic demand is permitted only in the direction toward higher QoS level. 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 for six interior routers can be seen in fig. 3. 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 the optimal result with significant computational savings. Fig. 2 gives a network flow representation of CEP for three QoS

![Figure 1](image1.png)

**Figure 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.
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 traffic demands (equal to capacity conversion of link 1,2,..., $r$). Vertical links represent re-routing of routers. Common node “O” is the source for used 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 $k =$ QoS level. The $N$ levels are ranked from $1, 2, \ldots, N$, and quality decreases with higher $k$.
- $m =$ the order number of link on the path, connecting two successive routers. Path consists of $M$ links ($m = 1, \ldots, M$) between $M+1$ routers.
- $u, v =$ the order number of capacity points in the sub-problem, $1 \leq u, \ldots, v \leq M+1$.
- $r_{im} =$ traffic demand increment for additional capacity of facility $i$ (appropriate QoS level) on link $m$. For convenience, the $r_{im}$ are assumed to be integer.
- $I_{im} =$ the relative amount of idle capacity of facility $i$ on the link $m$, related on the link before. Initially there is no capacity shortage between edge router and the interior router, $I_{1m} = 0, I_{M+1m} = 0$.
- $W_{im} =$ upper limit for capacity of facility $i$ on the link $m$.
- $kI_i =$ the lowest step of possible facility change for QoS level $i$.
- $x_{im} =$ the amount of used capacity for facility $i$ on the transmission link $m$.
- $W_{Xim} =$ upper limit for allowed expansion for facility $i$ on the transmission link $m$.

$y_{ijm} =$ the amount of capacity of facility $i$ on the link $m$, redirected to satisfy the traffic of lower level.

The CEP problem can be formulated as follows:

$$\min \left( \sum_{i=1}^{M} \sum_{k=1}^{N} c_{im}(x_{im}) + h_{im}(I_{im+1}) + \sum_{j=i+1}^{N} B_{ijm}(y_{ijm}) \right)$$

So that we have:

$$I_{m+1} = I_{im} + x_{im} - \sum_{j=i+1}^{N} y_{ijm} - r_{im}$$

$$I_{im} = I_{i,m+1} = 0$$

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 $h_{im}(I_{im+1})$ 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-charges 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 $m$ 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 (3.1.1) $\alpha_m$ denotes the vector of capacities $I_{im}$ for all QoS levels (facility types) on link $m$, and we call it capacity point.

$$\alpha_m = (I_{1im}, I_{2im}, \ldots, I_{ Nim})$$

$$\alpha_0 = \alpha_{M+1} = (0, 0, \ldots, 0)$$
Each column in the flow diagram from fig. 2 represents a capacity point, consisting of $N$ capacity state values. (3.1.2) implies that idle capacities or capacity shortages are not allowed on the link between edge and interior router.

### 3.2. Sub-problem

Associated value between two capacity points, that represents minimum cost $d_{uv}(\alpha_u, \alpha_{v+1})$ we denoted as CES (Capacity Expansion Sub-problem). In CEP we have to find many cost values $d_{uv}(\alpha_u, \alpha_{v+1})$ that emanate two capacity points, from each node $(u, \alpha_u)$ to node $(v+1, \alpha_{v+1})$ for $v \geq u$. The approach described in [3] requires solving repeatedly a certain single location expansion problem (SLEP). 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.

Supposing that all $d_{uv}$ values are known, the optimal solution for CEP can be found by searching for the optimal sequence of capacity points and their associated values. On that level the problem can be formulated as the shortest path problem between edge routers for an acyclic network, which nodes represent all possible capacity states.

### Conclusions

In the process of new SLA creation possible congestion can be checked with proposed heuristic algorithm. Algorithm is based on mathematical model for the capacity expansion problem (CEP); see fig. 3. 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. It will ensure end-to-end QoS routing guarantees, improving DiffServ granularity. The heuristic algorithm in all test-examples can achieve the best possible result (near-optimal expansion sequence), but requires the computation effort of $O(M^2N^4R_{2N-1})$. The required effort for one sub-problem is $O(NM^2)$. The number of all possible $d_{uv}$ values depends on the total number of capacity points. If there are no limitations (for $W_{lm}$ and $W_{xlm}$) the complexity of such heuristic approach is pretty large and increases exponentially with $N$.

### References