Piotr Krawiec, Wojciech Burakowski
Institute of Telecommunications
Warsaw University of Technology, Warsaw

How to reduce signaling traffic in the future QoS Internet

Guaranteeing the expected QoS for traffic generated by such applications as VoIP or VoD requires appropriate resources allocation between communicating and systems. However, individually managing each flow, especially in transit domains, has scalability limitations result from, among other things, high amount of signaling traffic. In this paper, we present an approach for reducing amount of signaling traffic traversed transit domains. The proposed scheme, called STPF (SomeTimes Per Flow), assumes division of available link capacity into two main parts: pre-reserved part for handling calls without need of any signaling in transit domains, and part dedicated for handling calls in traditional per-flow manner, what leads to achieve high resource utilization thanks to the multiplexing gain. Presented simulation results demonstrate the benefits of using the STPF.

1. Introduction

The motivation for introducing effective signaling system into the Internet is to allow communication user-user and network-network and, in this way, to handle the QoS (*Quality of Service*) requests emitted by an application for making adequate resource reservations. This feature is vital for providing absolute QoS guarantees what is required for guaranteeing effective transfer of multimedia traffic flows generated by such applications as VoIP, VoD, VTC etc. The discussed approach for signaling in the Internet assumes two levels of signaling, where one level corresponds to user-user communication and is proceeded by SIP (*Session Initialization Protocol*) protocol and the second one that is needed for network signaling and is proceeded by NSIS (*Next Steps in Signaling*). Such approach is currently enforced and tested by the IST EuQoS [1] project. However, introducing signaling system in the Internet may cause some additional problems that are not recognized well yet. One of expected barriers may be related to the presence of signalling traffic in the network which if it is of high volume may then lead to scalability problems of solution. In particular, performing per flow signaling along the whole network, similarly as using RSVP (*ReSerVation Protocol*) in IntServ architecture, is not a desirable approach, especially for multi-domain connections [2].

In this paper we present an approach for reducing amount of signaling traffic in the EuQoS system that assumes, as it has been mentioned above, two levels of signaling. In fact, we focus on network-network signaling level only. The traditional approach (as in PSTN network) is to perform PF (*Per Flow*) signaling via the whole network. It is obvious, that the PF is the most effective for getting high level of resource utilization but requires to handle relatively high amount of signaling traffic. On the opposite pole, is to make the resource pre-reservations in transit domains for each pair of ending domains. In such solution, named PRO (*Pre-Reservations Only*), any signaling in transit domains is needed but it leads to low level of resource utilization since multiplexing gain is lost. In this paper we propose an intermediate approach named STPF (*SomeTimes Per Flow*).

The STPF is aimed at performing per flow signaling in the ending domains, making pre-reservations at the transit domains and using per flow reservations in transit domains only sporadically. In the paper we prove that by using STPF, the volume of signaling traffic in transit domains is radically reduced comparing with PF while we maintain similar level of resource utilization.

The rest of the paper is organized as follows: Section 2 briefly describes the signalling system currently applied and tested in the EuQoS system. In Section 3, we describe the STPF approach. We report preliminary simulation results in Section 4, and we draw conclusions in Section 5.

2. EuQoS scenario

The objective of the EuQoS (*End-to-end Quality of Service support over heterogeneous networks*) project is to find a solution for assuring QoS in the multi-domain and heterogeneous network environment. The capability to provide QoS on a per-flow basis implies two different behavioural subsystems (see Figure 1), i.e. the application layer and the (virtual) network layer, where we can distinguish two sub-layers: Network Technology Independent (NTI) sub-layer and Network Technology Dependent (NTD) sub-layer. The application layer is responsible for user-user signaling (to agree on the same set of multimedia devices, i.e. on a set of compatible codecs) and bases on enhanced SIP and SDP protocol. NTI sub-layer is responsible for QoS negotiation and reservation the QoS-path between end systems. The NTI control entities, called RMs (*Resource Managers*), communicate between themselves using NSIS [3]. The NTD sub-layer performs physical allocation of requested resources, using the most appropriate (thus different) solution in any of the different internet access networks and domains. The NTD control entity, called RA (*Resource Allocator*s), receives the guidelines for resource allocation from relevant RM. Communication between a RM and, associated to it, RAs is achieved using the Common Open Policy Service (COPS) protocol [4], specified by the IETF.

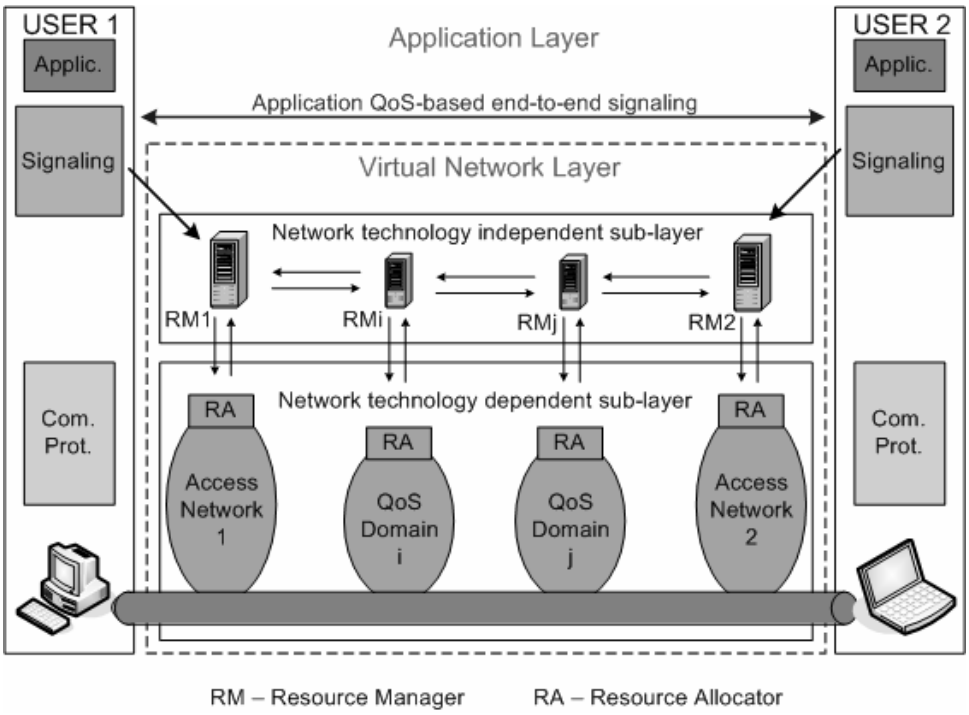

Figure 1. The global EuQoS architecture.

For now, we assume that accepting the flows (traffic streams – the stream of packets with the same both source and destination IP addresses, port numbers, etc.) is based on per flow operations (PF approach). So, the QoS request is generated by a source (end user side) to the network by using

the signaling system, named EQ-SSN (*EuQoS Signaling System in the Network*), as it is depicted in Figure 2. When the first RM, i.e. RM1 in the Access Network 1, received QoS request, it checks if it exists a suitable QoS-path between source and destination regarding the requested QoS. When RM1 finds an appropriate QoS-path, it performs resource checking for its own part of the QoS-path, that is AC (*Admission Control*) algorithm is performed. If the QoS can be met, QoS enforcement information is sent by the RM1 to the device nodes it controls (only those which need to be configured) through the RA. Next, RM1 forwards the QoS request to the next RM on the QoS-path. The AC process and the necessary resource reservation is repeated hop-by-hop at each of the consecutive domains belonging to the QoS-path. Finally, the connection is established only if AC decisions for each part of the network are positive.

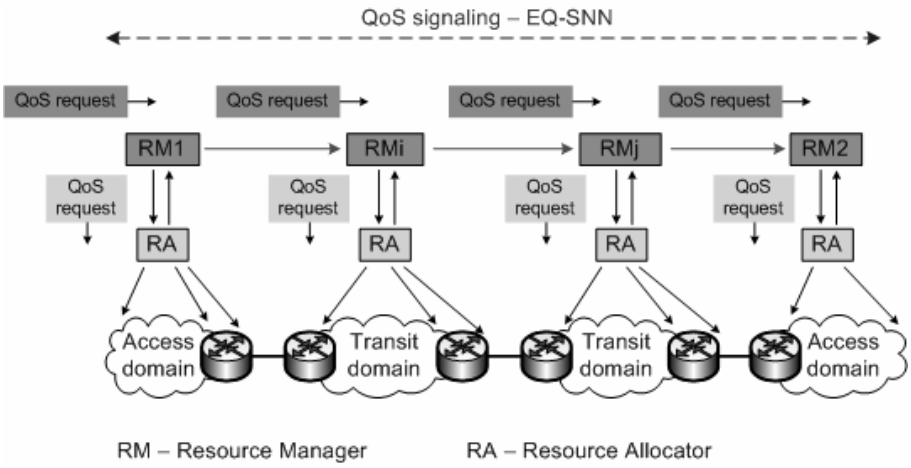

Figure 2. EuQoS scenario from the point of view of the network layer.

However, the recognized drawback of such approach is that we need to maintain PF signaling in each domain and this can lead to increasing both set-up latency as well as volume of signaling traffic we introduce to the network. Furthermore, such solution may not meet the scalability requirements. For this purpose, in the section we present approaches aimed at the reducing of signaling traffic.

3. STPF (SomeTimes per Flow) approach

As it has been mentioned above, for reducing signalling traffic in transit domains we can use the PRO approach that is based on the resource pre-reservations between each pair of ending domains. In this case, the RMs in transit domains play a role of transit signaling points while RAs are not engaged in the call handling process, as it is shown on Figure 3. However, as we will show in further part of the text, such approach does not guarantee effective bandwidth utilisation because the multiplexing gain in transit networks is lost.

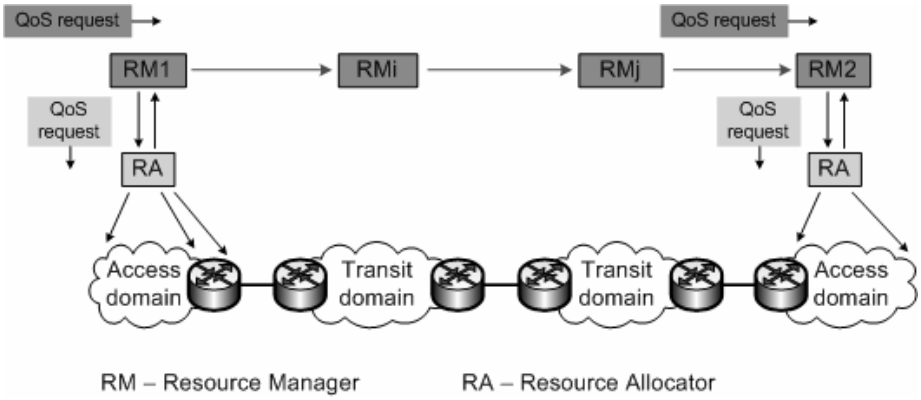

Figure 3. PRO scenario.

Taking into account the above drawbacks of the PF and PRO approaches, we propose to consider an intermediate solution, named the STPF. The STPF assumes that the available link capacity is divided into two main parts, where one part is reserved only for handling the calls on the basis of PRO scheme while the second part is handled by PF scheme. The resources belonging to the area of PF scheme can be seized only if no available resources in the part belonging to the PRO scheme. As a consequence, we expect that for the majority of calls we will use the PRO service with not necessary for exploiting the full reservation scheme while the full reservation process will be provided for a certain percentile of calls. In this way, we expect to get high resource utilization while required signalling traffic will be radically reduced.

4. Numerical results

In this section we present exemplary numerical results showing the effectiveness of the STPF approach. For obtaining the results we used our simulation tool (written in C++).

Figure 4 shows a sinking tree network scenario we investigate in the simulation experiments. In our example, the network consists of 9 nodes and among them the nodes named n11 … n16 are the source nodes that emit traffic to the destination node n4 via the transit nodes n21/n22 and n3. Furthermore, for call input process we assume Poissonian process with exponential service time distributions (normalised to 1). The calls arrive to each source node with the same arrival rate λ. Arriving calls request the same amount of bandwidth equal to 1 unit, so the capacity of each link, expressed in units, indicates the maximum number of simultaneously running connections. Each link has the reservation pool, described as "res x", which indicates the amount of capacity units designated to handle arriving calls in the PRO manner.

The parameter taken into account for making an evaluation of the discussed approach is the signaling ratio (sig_ratio), which is defined as the ratio of number of calls handled in a PF manner, i.e. complete reservation process in each domain is performed, to the total number of calls arrived during simulation. Moreover, the following parameters were measured: blocking probability, defined as ratio of total number of calls rejected by AC entity to the total number of calls arrived during simulation, and link L3 load, which indicates the average load of link between nodes n3 and n4, where flows are aggregated.

The simulations were performed respecting the following rules : (1) during each simulation at least 10^6 calls arrived to each source nodes, (2) each simulations were repeated 12 times to account for the random nature of the experiment, and (3) obtained results were statistically post-processed to calculate the intervals of confidence with the 0,95 confidence level.

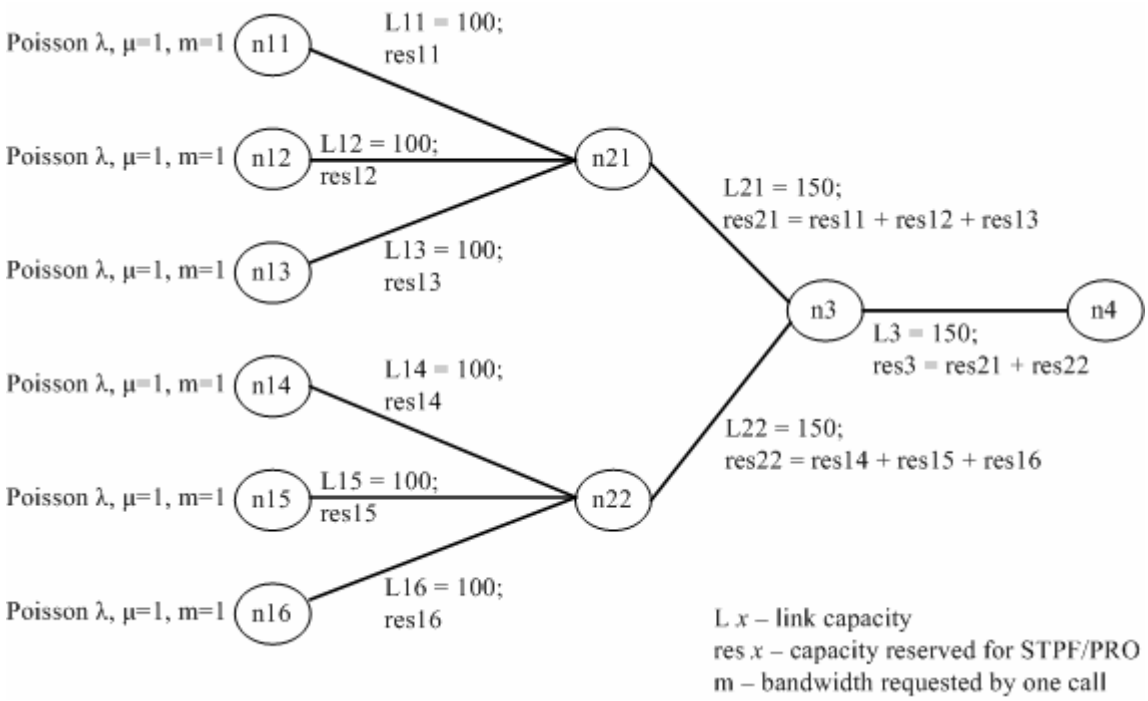

Figure 4. Simulation scenario.

Simulations were performed for three cases. In case#1, we study impact of reservation pool size on signaling ratio, blocking probability and links load when the offered load is fixed. In case#2 for the schemes PF, STPF and PRO, we measure parameters mentioned above as a function of the volume of offered load to the network. Case#3 concentrates on showing benefits of the STPF approach in comparison with PRO.

4.1. Case#1

Simulation results for the case of fixed offered load are presented in Table 1. This scenario assumes fixed links capacity whereas reservation pools res11…res16 varies from 0 to 25 units per one link between source nodes and transit nodes n21/n22. Capacity of links L11…L16 was set to 100 units. Capacity of links L21 and L22, where traffic is aggregated, was calculated as sum of capacity of descendent links divided by factor α, $L2x = \sum L1y / \alpha$, where $\alpha = 2$, what means, that capacity of link L21 (L22) is half as much as a sum of capacities of links L11, L12 and L13 (suitably L14, L15 and L16). In the same way we calculated capacity of link L3, i.e. L3 = (L21 + L22) / 2. Call arrival rate λ was obtained from Erlang B-formula under assumption, that for call arrival rate equal $6 \cdot \lambda$ the blocking probability on link L3 is equal 10^{-2} (losses on links L11…L16 and L21, L22 were not taken into consideration because of their very small values).

Links capacity: L11 = L12 = L13 = L14 = L15 = L16 = 100, L21 = L22 = 150, L3 = 150 λ = 21.93 calls/s				
Reservation pool for links L11…L16	0 (PF)	10 (STPF)	20 (STPF)	25 (PRO)
Sig_ratio	1	0.574	0.207	0
Blocking probability	0.010	0.013	0.031	0.083
Link L3 load	0.869	0.866	0.851	0.805
Link L11,..L16 load	0.217	0.216	0.213	0.201
Link L21, L22 load	0.434	0.433	0.426	0.402

Table 1. Simulation results for different scheme: PF (resource reservation = 0), STPF (res11…res16 = 10/20) and PRO (res11…res16 = 25).

It is worth mentioning that when reservation pool is 0 (no resources reserved for STPF) we consider PF scheme. On the other hand, when the reservation pool amounts to 25 units, we consider PRO approach – reservation of 25 units on links L11 … L16 means that the whole capacity of link L3 is reserved (6·25 = 150) and there is no free capacity to handle QoS requests in per-flow manner.

As it was expected, increasing the amount of reserved resources allows us to reduce the signaling load up to sig_ratio = 0 for PRO scheme (Figure 7), but it leads to lower resource utilization (Figure 6) and higher blocking probability (Figure 5). However, STPF with moderate reservation pools (e.g. 10 units in our simulation – Table 1) helps us to maintain similar blocking probability and links utilization in comparison with PF simultaneously reducing by 40 percent the signaling traffic needed to handle arriving QoS requests (Figure 5, 6 and 7).

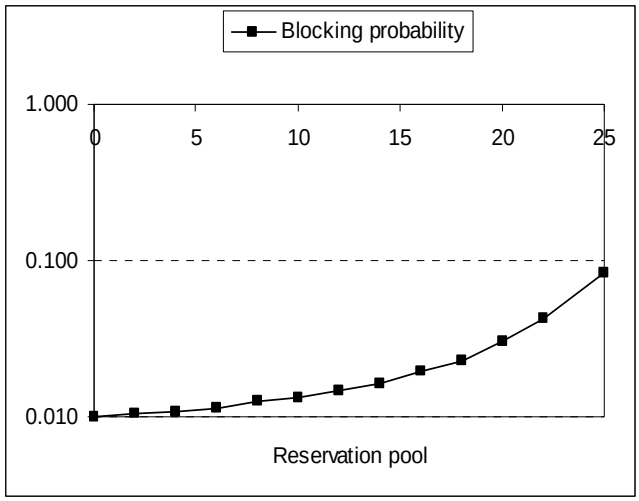

Figure 5. Blocking probability vs. resource reservation for different scheme: PF (reservation pool = 0), STPF and PRO (res11…res16 = 25).

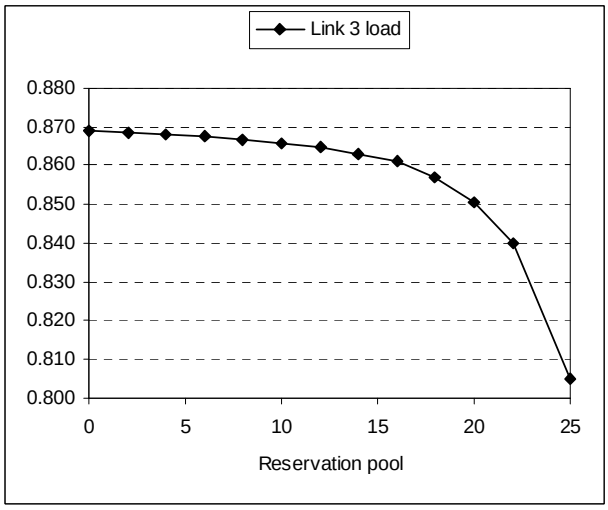

Figure 6. Link L3 load vs. resource reservation for different scheme: PF (reservation pool = 0), STPF and PRO (res11…res16 = 25).

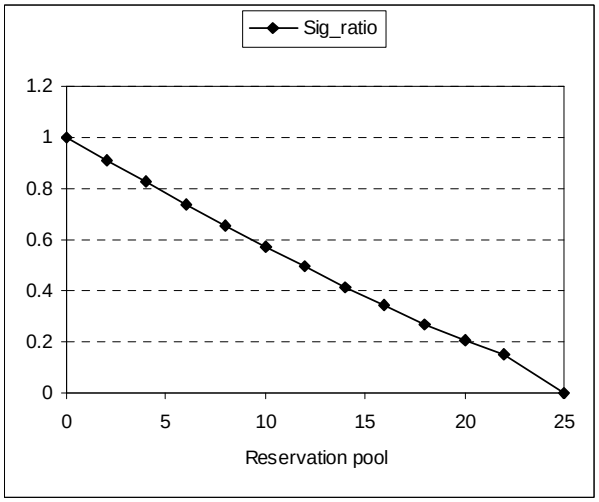

Figure 7. Signaling ratio vs. resource reservation for different scheme: PF (reservation pool = 0), STPF and PRO (res11…res16 = 25).

For the results presented in Table 2 we consider a scenario in which we want to achieve the same blocking probability for each scheme: PF, STPF with reservation pools equal to 10 and 20 units and PRO, assuming fixed average call arrival rate. To accomplish this goal, we increase the capacity of "bottleneck" link L3, where traffic is aggregated. As can be seen, the PRO requires

the highest amount of resources to keep blocking probability on the same level as we achieve for PF (the each link between source nodes n11 … n16 and transit nodes n21/n22 needs reservation pool equals 32 units, what causes necessity of increasing the link L3 capacity up to 192 units), during STPF needs only a small growth of link L3 capacity to pursue this aim.

average call arrival rate = 21.93 calls/s				
	PF	STPF	STPF	PRO
Sig_ratio	1	0.574	0.207	0
Blocking probability	0.01	0.01	0.01	0.01
Links L11 ... L16 capacity	100	100	100	100
Links L21 and L22 capacity	150	150	150	150
Link L3 capacity	150	152	158	192
reservation res11 … res16	0	10	20	32
Link L3 load	0.869	0.857	0.825	0.679

Table 2. Simulation results for different scheme: PF (resource reservation = 0), STPF (res11…res16 = 10/20) and PRO (res11…res16 = 32) with fixed blocking probability.

4.2. Case#2

Scenario for case#2 assumes the same links capacity as indicated in Table 1 and reservation pools res11…res16 equal to 10 units in case of the STPF approach. The average call arrival rate for each source nodes varied from 20 to 25 calls/s, i.e. the volume of load offered to the network increased. One can observe that when the average call arrival rate increase STPF, in contrast to PRO, it achieves nearly the same results as PF (Figure 8 and Figure 9), while it requires less than 60 percent of signaling load (Figure 10).

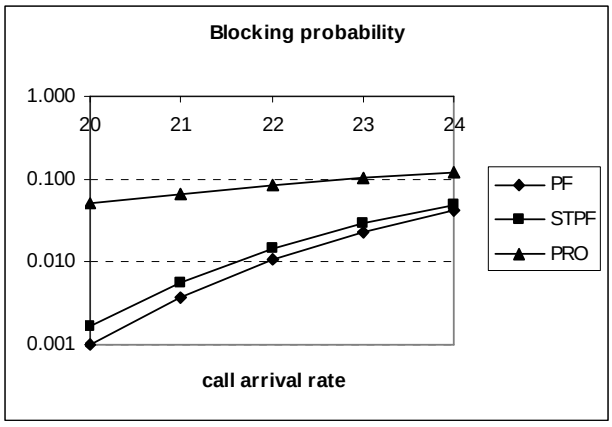

Figure 8. Blocking probability vs. average call arrival rate for PF, SPF and PRO strategy.

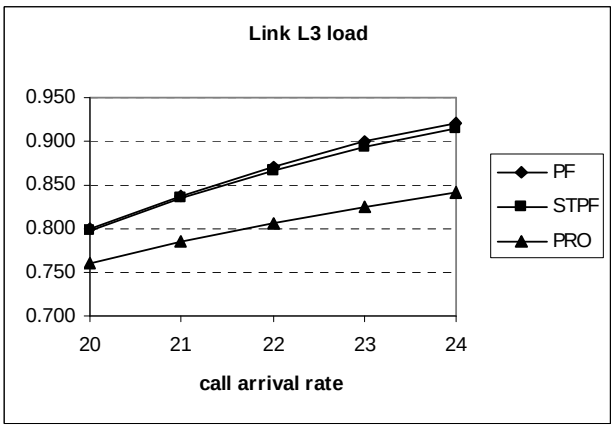

Figure 9. Link L3 load vs. average call arrival rate for PF, SPF and PRO strategy.

4.3. Case#3

In this case we study the benefits of using the STPF instead of PRO approach. Simulation scenario assumes the same links capacity as indicated in Table 1. First results were obtained for the PRO scheme (the whole link L3 capacity was reserved) assuming load offered to the network appropriate to achieve the blocking probability equals 10^{-2}. In the next simulations we investigated STPF scheme considering different part of bottleneck link L3 capacity dedicated to the per-flow operations. The results presented in Table 3 say, that allocation of not big part of link L3 capacity for handling calls in PF scheme helps us to get higher link utilization while the necessary signaling traffic constitutes a small part of signaling traffic required in case when only PF approach is used. For simulated scenario using the STPF with 90 percent of link L3 capacity pre-reserved and

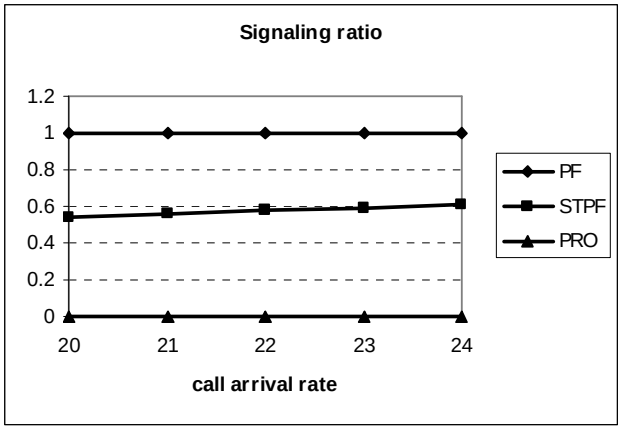

Figure 10. Signaling ratio vs. average call arrival rate for PF, SPF and PRO strategy.

10 percent designated for per-flow call handling increases link utilization by 20 percent in comparison with PRO scheme (from 0.639 to 0.767) whereas sig_ratio reaches merely about eight percent, i.e. only eight percent of arrived calls required full reservation process. Transferring the bigger amount of link capacity for handling calls in PF manner gives higher link utilization because it allows better multiplexing of the traffic, but it causes increase of signaling traffic introduced to the network, too.

	PRO	STPF	STPF	STPF	STPF	STPF
Resources for PF operations	0%	6.70%	10%	13.30%	16.70%	20%
Sig_ratio	0.000	0.056	0.083	0.111	0.140	0.169
Blocking probability	0.010	0.010	0.010	0.010	0.010	0.010
Link L3 load	0.639	0.743	0.767	0.785	0.799	0.808

Table 3. Simulation results for different amount of link L3 capacity dedicated for PF operations.

5. Conclusions

In this paper we have presented preliminary results showing the rationality for using the STPF scheme in the multi-domain Internet using the signaling system for making resource reservations, that is required for introducing QoS. The results say that by using this approach we can get high link utilization while the volume of signaling traffic is reduced, in some cases even radically.

The future work is concern on verifying the approach for other network scenarios and different traffic conditions in the network.

References

1. The IST-EuQoS (End-to-end Quality of Service support over heterogeneous networks) project, www.euqos.org
2. A. Mankin, et al., “Resource ReSerVation Protocol (RSVP) Version 1 Applicability Statement Some Guidelines on Deployment”, RFC 2208
3. Hancock, R., Karagiannis, G. Loughney J., van den Bosch, S.; Next Step in Signaling (NSIS): Framework IETF Working group. http://www.ietf.org/Internet-drafts/draft-ietf-nsis-fw-06.txt, July 2004
4. D. Durham et al. The COPS (Common Open Policy Service) Protocol. RFC 2748