FACTA UNIVERSITATIS (NIŠ)
SER.: ELEC. ENERG. vol. 17, August 2004, 261-274

A Novel Approach for Providing Quality of Service in Multiservice IP Networks

Mirjana Stojanović and Vladanka Aćimović-Raspopović

Abstract: Considering specific operational and corporate service requirements, this article proposes an approach for providing quality of service (QoS) in multiservice networks based on the IP (Internet Protocol) technology. The basic elements of the proposed approach encompass DiffServ- aware QoS architecture, proposal for traffic classification, proposal for configuration of the data plane QoS implementation mechanisms, as well as the model of QoS control and management. Particularly, the proposed approach is applicable to private networks with specific operational and corporate needs (power utilities, traffic and transport, etc.), which mandatory require different levels of QoS in order to assure stable and reliable system operation.

Keywords: Internet protocol, quality of service, service differentiation, policing, scheduling, queue management, QoS control, QoS management.

1 Introduction

The next generation of broadband telecommunication networks comprises integration of heterogeneous services (multiservice network concept) in order to achieve cost efficiency, optimal use of network resources as well as unified network and service management. The IP (Internet Protocol) technology together with common resources shared among multiple users is widely accepted as a basis for future service integration. One of the key issues in multiservice IP-based networks concerns resolving problems of providing different quality of service (QoS) levels, in accordance with specific requirements of different applications and users.

Manuscript received March 7, 2004.

M. Stojanović is with Mihailo Pupin Institute, Volgina 15, 11050 Belgrade Serbia & Montenegro (e-mail: stojmir@kondor.imp.bg.ac.yu). V. Acimović-Raspopović is with Faculty of Transport and Traffic Engineering, University of Belgrade, Vojvode Stepe 305, 11000 Belgrade Serbia & Montenegro (e-mail: vaksa@eunet.yu).

In this paper, we propose a general methodology for providing different levels of QoS, considering architecture, traffic classification, specification of packet processing mechanisms, and QoS control and management. Particularly, we have in mind application of the proposed approach to private networks with specific operational and corporate needs (power utilities, traffic and transport), which mandatory require different levels of QoS in order to assure stable and reliable system operation [1]-[4].

Related research work comprises numerous aspects: network planning and design, development of conceptual QoS architectures, network equipment design, performance analysis, novel or improved algorithms for QoS implementation mechanisms.

Several EU projects in the FP5 dealt with scalable QoS architectures for the next generation IP networks. Recently published overview [5] of the most significant results indicated that all projects identified the Differentiated Services (DiffServ, [6]) architecture, MPLS (Multi-Protocol Label Switching, [7]) technology and the RSVP (Resource Reservation Protocol, [8]) as the most relevant elements for implementing QoS. The importance of research work in the scope of QoS control and management has been emphasized, with the objective to achieve interoperability through development of standardized solutions. Similar considerations stand for the process of QoS negotiation between the user and the provider (SLA - Service Level Agreement) and for QoS provisioning across multiple independent administrative IP domains.

The approach for implementing QoS at the Internet backbone, presented in [9], relies on traffic classification, applying suitable mechanisms for packet processing, and traffic engineering and restoration based on the MPLS technology features.

Heterogeneous variants of QoS implementing mechanisms have been proposed and verified and different combinations can provide satisfying performance, if input parameters are properly configured, regarding particular requirements, technical solutions and traffic load in the operational network. Besides, there is a number of proposals for new algorithms or improving of the existing ones.

The rest of the paper is organized as follows. Section 2 contains problem statement. In Section 3 we propose the methodology for implementing different QoS levels. Proposal for general traffic classification, followed by a representative example, is presented in Section 4. Section 5 contains proposals for configuring of queuing and scheduling mechanisms. QoS control and management have been addressed in Section 6. Finally, Section 7 contains concluding remarks.

2 Problem Statement

We adopt the IETF definition for IP QoS as a "set of service requirements to be met by the network while transporting a flow"[10]. In other words, we focus on the intrinsic QoS, which is determined by network design and provisioning of network access, terminations and connections. QoS is quantified by performance metrics like service availability, throughput, delay, jitter, packet loss ratio, etc.

Considering scalability as a fundamental requirement for multiservice IP networks, QoS provisioning will be analyzed in the context of aggregate architectures with service differentiation (DiffServ-aware). They have a common property that packets belonging to different traffic flows, but with similar QoS requirements, may be associated to the same traffic class and processed in the same manner at the network nodes. Since generic concepts of the DiffServ-aware architectures concern a single IP domain, additional mechanisms for providing end-to-end QoS over a single domain and across multiple domains will be proposed.

Starting from user requirements, heterogeneous applications and problems related with providing the required QoS levels, we propose solutions for implementing QoS. The following aspects are addressed: proposal of the traffic classification, analysis of QoS implementation mechanisms, proposal of the method for parameters configuring and the model for QoS control and management.

3 The Methodology of QoS Design and Implementation

The proposed methodology of QoS design and implementation is illustrated in Fig. 1. The design process begins with the analysis of QoS requirements, followed by the decision on service integration strategy (degree of integration, dynamics, etc.), which directly influences the specification of different QoS levels, i.e. traffic classes and their associated priorities.

QoS implementation encompasses a set of mechanisms distributed across three logical planes (the data, control and management planes) [11]. Data plane comprises mechanisms dealing with the user data traffic: classification, policing and shaping, packet marking, packet scheduling, and queue management. Control plane contains mechanisms dealing with the paths through which user data traffic is carried, e.g. admission control, resource reservation, QoS routing, etc. Management plane includes QoS mechanisms dealing with operation, administration and management aspects, e.g. traffic metering, network policy, service level agreement (SLA), traffic restoration etc.

Traffic engineering involves adapting of routing to network conditions in order to improve the overall network performance in the sense of increasing avail-

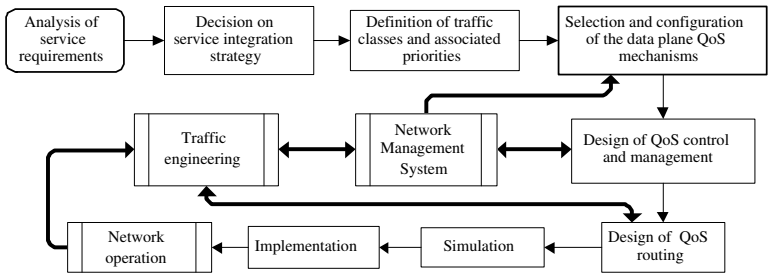


Fig. 1. The process of QoS design and implementation.

ability and throughput, minimization of packet loss and optimization of resource utilization. QoS routing or its upgrade with the other conditions when selecting paths (constrained-based routing) together with the traffic engineering may facilitate QoS assurance. Traffic engineering is a macro-control process (complementary to the DiffServ concept) that comprises both off-line network design and run-time operations. Implementing is company- or operator-specific and comprises control policies, measurement, performance analysis, and performance optimization.

The proposed methodology is applicable to multiservice IP networks with different switching and transmission technologies. Standard [6] or proprietary code values for different traffic classes and associated priorities may be used. The use of MPLS technology may additionally facilitate providing QoS, due to its suitability for traffic engineering, flow control, traffic routing and restoration, etc.

4 Traffic Classification

We adopt the following general model of traffic classification:

1. **Premium service** that provides high availability, guaranteed peak rate, and low delay and jitter variation.
2. **Assured service** that provides statistical guarantees for packet delivery with committed rate. Within this service, various classes can be defined, each with different priority levels.
3. **Best effort service** that does not provide any QoS guarantees, and corresponds to the service available in the present global Internet.

Different variants for definition of the assured service and for association of particular operational and corporate services to pre-defined traffic classes are possible. The choice of solution depends on the decision about service integration degree, granularity requirements, network administration policy and network equipment capabilities. However, we recommend the following general rules:

- Real-time services must be prioritized, particularly operational services;
- Differentiation between classes and associated priorities must be quantified by performance metrics like maximum end-to-end delay, maximum jitter, maximum packet discard ratio, etc.;
- A small set of distinguished classes should reduce implementation complexity and facilitate interworking with the other networks.

An example of traffic classification with regards to applications and their requirements is presented in Table 1.

Table 1. An example of traffic classification.

Service class		Applications	Service requirements
Premium		Real-time, jitter sensitive, highly interactive (VoIP, signaling, multimedia conferencing)	End-to-end-delay$<$100ms, Jitter variation$<$10ms Maximum packet loss$\sim 10^{-7}$
Gold	High priority	Low delay and loss, interactive (transaction data)	End-to-end-delay$<$400ms, Maximum packet loss$\sim 10^{-5}$
	Low priority	Low loss only (video streaming, bulk data, short transactions)	End-to-end delay$<$1000ms, Maximum packet loss$\sim 10^{-3}$
Best effort		Traditional applications of default IP networks	Not specified

Within the assured service, only one service class is defined - gold service, with two priority levels, where high priority denotes lower packet discard probability under the conditions of congestion. Traffic flows for gold service are marked to lower priority in the traffic policing process at the ingress router, only if they do not conform to the pre-defined traffic profile. Otherwise, these flows should also be marked to high priority.

Several different approaches are also possible, depending on service requirements [2], [3].

5 Configuring of QoS Mechanisms in the Data Plane

5.1 Traffic policing

Traffic policing is a process at the ingress routers, which comprises discarding traffic that does not conform to the predefined profile or marking that traffic for another profile, with a lower priority. Traffic profile is defined through static or dynamic SLA, negotiated between the service user and the provider. For premium service, traffic with higher peak rate than a negotiated must be discarded, to assure service guarantees for the in-profile traffic. For the assured service with two or more priority levels, the out-of-profile traffic can be marked with a lower priority.

5.2 Packet scheduling

Packet scheduling enables sharing of the output link bandwidth between multiple traffic flows. It must be implemented in all network nodes. We propose a practical approach for scheduler parameters configuring, assuming two categories of service disciplines. The basic rule is that a separate physical queue should be assigned to each service class. If two or more priorities are defined inside a service class, the corresponding physical queue should contain an appropriate number of virtual queues.

The first category of disciplines relies on priority-based scheduling and assumes assignment of a certain priority level to each service class. The available amount of bandwidth for premium service should be reserved, but limited to prevent exhausting all resources to lower priority traffic. The reserved bandwidth should be slightly over-provisioned (peak arrival rate/ service rate < 1).

The second category of scheduling disciplines relies on queue serving based on the certain algorithm (Round Robin, Fair Queuing etc.) with assignment of the appropriate weighting factor to each physical queue. Thus, relative serving priorities for various traffic classes can be accomplished. Service rate of the physical queue i, $sr(q_i)$, should be determined from the following expression:

$$sr(q_i) = C\frac{w(q_i) \times ar(q_i)}{\sum_i w(q_i) \times ar(q_i)} \tag{1}$$

where C is the overall capacity of the output link, $w(q_i)$ is weighting factor for the queue i, and $ar(q_i)$ is packet arrival rate to the queue i.

A good trade-off can be achieved by combining the two approaches. The physical queue for premium class can be served with the absolute priority and reserved bandwidth, while weighted-based methods can be applied for the traffic marked for other service classes.

5.3 Queue management

Queue management should be applied in all network nodes, in order to avoid network congestion or to minimize its duration. Discarding of packets marked for premium service must be avoided by traffic policing at the ingress routers and bandwidth over-provisioning. For other traffic classes, we propose the use of Weighted RED (WRED) algorithm [12]. The basic RED algorithm [13] assumes probabilistic packet discarding, based on estimation of the average queue length and comparison with the two predefined thresholds: minimum, T_{MIN}, and maximum, T_{MAX}. When the queue length is below T_{MIN}, packets are not marked. If the queue length is

between T_{MIN} and T_{MAX}, packets are marked with the probability calculated as a function of the average queue length. The maximum allowed marking probability, P_{MARK}, must be pre-configured. If the queue length is above T_{MAX}, every packet is marked. Marked packets are stochastically discarded under the condition of congestion.

WRED combines RED with different scenarios of traffic classification to achieve performance differentiation by a selective discarding of the lower priority traffic, to prevent congestion. However, it is quite difficult to derive analytical methods for WRED parameters configuring due to algorithm sensitivity under the various traffic loads. On the other side, practical recommendations, based on the empirical data in operational networks, have not been widely published. We propose a practical approach for finding suitable relationships between WRED input parameters assigned to different classes and determining sets of their optimal values to achieve satisfying performance differentiation. Assuming that N represents the total number of service classes and associated priorities to which WRED is applied, the following relations are supposed:

$$T_{MAX,i} = f_i(R_t) \times T_{MAX,i-1}, \quad i = 2,\ldots,N \tag{2}$$

$$T_{MAX,i} = C_i \times T_{MIN,i}, \quad i = 1,2,\ldots,N \tag{3}$$

$$P_{MARK,i} = g_i(R_p) \times P_{MARK,i-1}, \quad i = 2,\ldots,N \tag{4}$$

where index i denotes service priority level (the highest level is 1), R_t and R_p are real positive variables, f_i is a positive non-increasing function of R_t, g_i is a positive non-decreasing function of R_p, C_i is a real positive constant, $T_{MAX,1}$ and $P_{MARK,1}$ are pre-defined.

The objective is to find suitable sets of $\{f_i, g_i\}$ and to determine the range of values (R_t, R_p) that can provide satisfying performance differentiation for various traffic classes and associated priorities.

5.4 Simulation and results

Extensive simulations have been carried out to explore the influence of data plane QoS mechanisms to performance differentiation. For that purpose, the network simulator **ns-2** and network animator **NAM**, have been used [14]. Results affecting jitter and delay have been obtained by means of the **Trace Graph** analyzer [15].

The 10-node network topology has been assumed, with three core nodes and seven edge nodes (Fig. 2). Traffic classification has been performed according to the example from Table 1. Two priority levels for the gold service have been defined by means of the token bucket policer. The experiments concern paths with bottleneck links, e.g. $E_4 - C - E_1$, $E_4 - C - E_5$, $E_6 - C - E_2$.

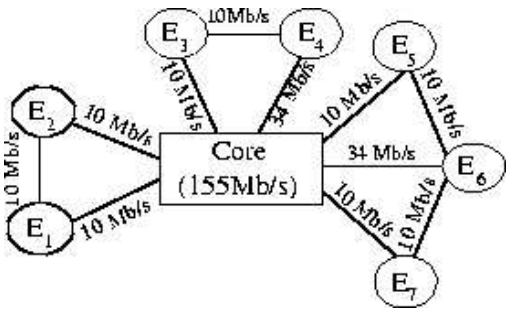


Fig. 2. Simulated network.

In the first group of experiments, we analyze maximum queuing delay and jitter at the core node, which forwards packets over a bottleneck output link. We assume 5 flows per aggregate for each traffic class. FTP traffic sources have been considered, with packet sizes 160By and 1000By, and the subscribed access rate 0.5Mb/s. The TCP transport protocol, with the window size 25, has been applied. The size of each physical queue (for premium, gold and best effort traffic) equals 10000By.

The following scheduling disciplines have been studied:

- Priority-based scheduling, where each class has available 1/3 of the total link bandwidth;
- Weighted Round Robin - WRR(6,3,1), with weighting factors 6, 3 and 1 for premium, gold and best effort service, respectively and
- WRR(4,3,3) - with weighting factors 4, 3 and 3 for premium, gold and best effort service, respectively.

If n is the overall number of transmitted packets, $R(i)$ is the time packet i was received, and $F(i)$ is the time it was forwarded by the core node, then the queuing delay (which in our simulation actually represents packet processing time at the observed core node) is calculated as

$$q\_delay(i) = F(i) - R(i), \tag{5}$$

and the maximum delay is $q\_delay_{MAX} = \max\{q\_delay(i)|_{1\le i\le n}\}$.

Jitter of forwarded packet i is calculated as

$$J_f(i) = |(F(i+1) - R(i+1)) - (F(i) - R(i))|, \tag{6}$$

and the maximum jitter of forwarded packets is $J_{fMAX} = \max\{J_f(i)|_{1\le i\le n}\}$.

The obtained simulation results are plotted in Figures 3 and 4. Large packets experience larger absolute delay and jitter, because of longer processing at the node. However, we observe both metrics normalized to packet-time at flow access rate, i.e. Tp=(8×packet_size)/flow_access_rate. Related to packet-time, both metrics are more critical for small packets. In a real situation, an aggregate can be a mixture of packets of various sizes, which is worse for small packets.

Simulation results indicate that the benefit for premium service can be achieved, on the count of performance of other service classes, with priority-based scheduler or with the WRR scheduler with the large corresponding weighting factor. However, with the WRR, when differentiation is made between the gold and best effort service, i.e. WRR(6,3,1), a better jitter performance for gold service is achieved, in comparison with the priority-based scheduler. Results also verify the need for reserving only a small part of the overall link bandwidth to premium traffic in the case of prority-based scheduling, in order to avoid blocking of the traffic marked for other classes. Fair utilization of the overall network resources (i.e. all physical links) may be accomplished by other mechanisms, like traffic engineering and QoS routing.

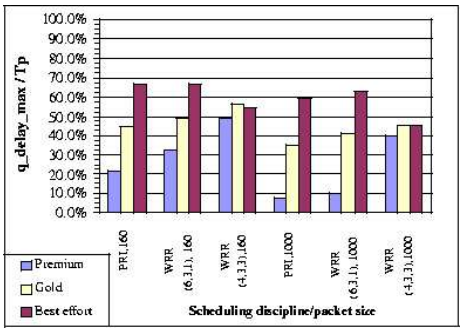


Fig. 3. Maximum normalized queuing delay at the core node.

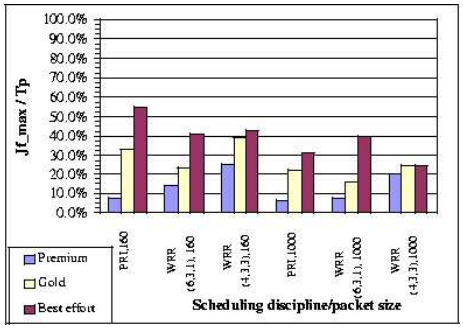


Fig. 4. Maximum normalized jitter of forwarded packets at the core node.

The second group of simulation experiments concerns congestion avoidance. Five FTP sources per service class have been considered, with packet size 160By, and the TCP protocol has been applied. The packet scheduling discipline is WRR (6,3,1). According to the approach suggested in the section 5.1, packets marked for the premium service are not discarded at the outgoing queues, since the traffic policer allows admission to the network only to packets conforming to a pre-defined profile.

First, we suppose that the access rate of each flow from the gold and best effort services equals 1.15Mb/s. The subscribed rate for the gold service is 0.75Mb/s (i.e. the overload factor OL equals 1.53), which means that a proportional amount of

packets is marked to the lower priority level. WRED algorithm has been applied for both, gold and best effort service. The size of each physical queue equals 10000By. Error-free transmission is supposed. Considering equations (2)-(4), we examine two groups of power law functions $\{f_a, g_a\}$ and $\{f_b, g_b\}$, defined in the following manner

$$f_{ai}(R_t) = R_t^{-1}; \quad g_{ai}(R_p) = R_p, \quad \text{for } i = 2,3; \tag{7}$$

$$f_{bi}(R_t) = R_t^{-(i-1)}; \quad g_{bi}(R_p) = R_p^{i-1}, \quad \text{for } i = 2,3; \tag{8}$$

$$C_{ai} = C_{bi} = 2, \quad \text{for } i = 1,2,3; \tag{9}$$

where $i = 1,2$, and 3 denotes gold class with high priority (GH), gold class with low priority (GL) and best effort (BE) service, respectively. $P_{MARK,1}$ equals 0.01, $T_{MAX,1}$ equals 8000By, factor R_t takes values from the set $\{1, 1.5\}$, and factor R_p takes values from the set $\{1, 10\}$.

First, the influence of adjusting threshold levels (factor R_t) to packet discard ratio, PDR, has been studied, for different marking probabilities (factor R_p). PDR is determined as a ratio of the total number of discarded packets by WRED and the total number of packets arrived to the corresponding physical queue. Packets may be discarded due to overflow of the queue and due to activation of early detect WRED mechanism. Simulation results for four representative values of R_p have been plotted in Fig. 5.

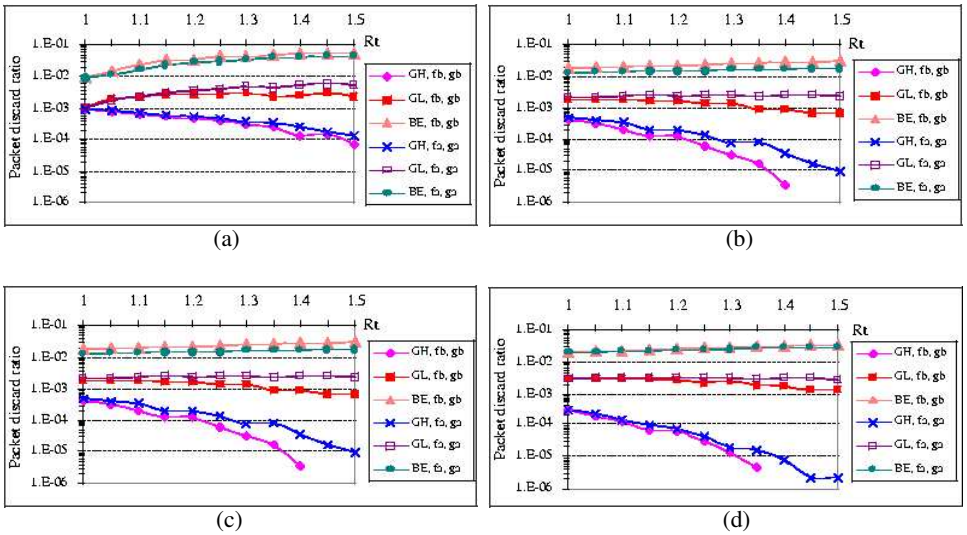
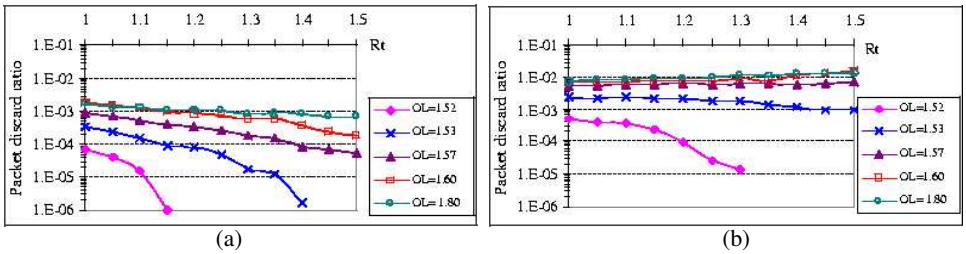


Fig. 5. Packet discard ratio vs. R_t for different marking probabilities and traffic priorities. (a) $R_p = 1$. (b) $R_p = 4$ (GH,f_b, g_b| PDR = 0, for $R_t > 1.4$). (c) $R_p = 7$ (GH,f_b, g_b| PDR = 0, for $R_t > 1.4$). (d) $R_p = 10$ (GH,f_b, g_b| PDR = 0, for $R_t > 1.1.35$).

In all cases, performance differentiation between traffic priorities can be noticed for $R_t \geq 1.3$, and more emphatic for sharper weighting functions $\{f_b, g_b\}$. For low values of R_p (e.g. $R_p = 1$, when all marking probabilities are equal and very low), the percentage of preventive, early packet drops is low; hence a higher packet discard ratio appears as a consequence of the overflow of corresponding physical queues. For medium values of R_p (e.g. $R_p = 4$), the superposition of moderate percentage of early drops and low percentage of drops due to queue overflow results in lower packet discard ratio for lower priority services, while preserving prioritization of gold service with high priority. For higher values of R_p ($R_p = 7$ and $R_p = 10$), the gold service with high priority is absolutely favoured, due to the very high percentage of early drops of lower priority traffic.

In the next experiment, the influence of change of traffic load to packet discard ratio has been examined. We assume the set of relations $\{f_b, g_b\}$ according to the equation (8) and a moderately high value of factor R_p ($R_p = 7$). Assumptions on the traffic sources are the same as in the previous experiment, except that we vary the access rate of the flows marked for the gold service. Simulation results concerning packet discard ratio of the gold service, as a function of different threshold levels, are plotted in Fig. 6.



(a) (b)

Fig. 6. Packet discard ratio vs. R_t for different traffic loads, $\{f_b, g_b\}$, $R_p = 7$. (a) Gold service, high priority (OL=1.52 and 1.53 | PDR = 0, for $R_t > 1.15$ and $R_t > 1.4$). (b) Gold service, low priority (OL=1.52 | PDR = 0, for $R_t > 1.3$)

It should be noted that there is a range of values of the overload factor, OL, for which service differentiation is preserved, within the same range of values of R_t, as in the previous experiment. Further increasing of the overload factor over the certain value causes saturation of the packet discard ratio due to interaction of WRED and the embedded end-to-end TCP congestion control mechanisms.

Results of experiments concerning WRED clearly indicate that adjusting thresholds and packet marking probabilities and establishing an appropriate relationship between their levels for different service classes can achieve the required service differentiation, if traffic policers at the ingress routers are appropriately configured.

6 The Model of QoS Control and Management

We adopt the notion of the Network Resource Manager (NRM), which is a logical entity that decides about the admission of new flows and resource allocation. NRM entity is a part of the TMN system and it should be implemented as a software process in the control center from which the network configuration is controlled. Configuration management (CM) is performed through the TMN system mechanisms. End systems should request certain level of QoS for individual flows, from the NRM entity, through RSVP or some proprietary-based access signaling protocol. Due to DiffServ- aware QoS architecture, resource allocation is performed in accordance with the sender needs. In the backbone network, signaling messages are associated with the premium service and, in general, they don't have to be coupled with the user data paths. For interconnecting with the other IP domains, SLA should be negotiated, either statically or dynamically. Interface with the neighbouring domain encompasses mapping of the negotiated features to particular intra-domain solutions.

An example of QoS-enabled communication for user application distributed in the two independent administrative IP domains is illustrated in Fig. 7.

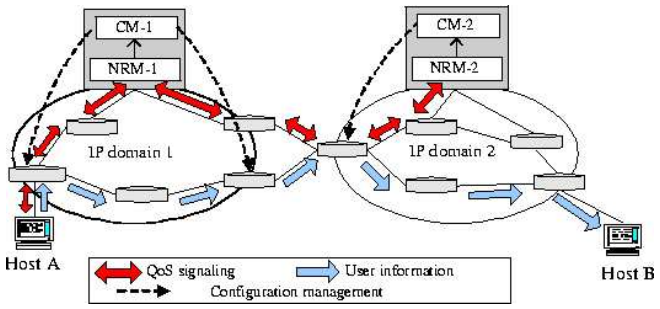


Fig. 7. An example of dynamic SLA negotiation between two neighbouring domains.

Host A requests certain level of QoS for its data flow, through a signaling message, which is transported to the ingress router. It associates the message with the premium service and transparently forwards it to the domain's NRM entity. If NRM-1 grants the admission of the flow, it sends the appropriate signaling message to neighbouring domain (NRM-2). If NRM-2 grants the request, it returns the acknowledgement to NRM-1, which further forwards it to the end-system. NRM entities indicate the admission to their associated configuration managers (CM-1, CM-2), which configure border routers, based on that information. Due to sender-oriented resource reservation, at the receiver's domain only the ingress router should be configured to perform QoS mapping and traffic conditioning. Note that similar operation must be performed in the opposite direction, for the data flow

from host B to host A. After that, the exchange of user information takes place.

This approach assumes QoS-aware applications in the sense of capabilities to define QoS requirements, to determine traffic profile and specify it through a formal description of the traffic flow, and to forward to the network information on QoS requirements, by means of the QoS signaling protocol. The task can be facilitated by use of the unified, highly portable and extensible application programming interface (QoS API), in order to release applications from concern about the details on QoS solutions implemented in network elements.

Concatenation of several bilateral SLAs in order to achieve end-to-end QoS through several domains should be controlled by a common network service manager, which should coordinate the communication between individual NRM entities, by evaluating and merging SLAs and finding optimal paths for the user data flows (inter-domain QoS routing).

7 Conclusions and Future Work

Taking into account scalability requirements, QoS in multiservice IP networks is considered in the context of the DiffServ-aware architectures, with additional mechanisms that provide end-to-end QoS guarantees. The proposed methodology is applicable to multiservice IP networks with different switching and transmission technologies.

Traffic classification depends on the service integration degree and company-specific requirements, but it must provide quantitative performance differentiation, efficient implementation and suitability for interworking with the other networks. In order to assure service guarantees, the network administration policy should provide a proportionally low percent of premium traffic over individual links, together with absolute or relative processing priority. We have also proposed a practical approach for selection and configuring of queue management mechanisms, in order to preserve the desired level of service differentiation under the conditions of congestion.

The proposed model of QoS control and management relies on the Network Resource Manager, which decides about the admission of new flows and resource allocation, and dynamically cooperates with the network management system, as well as with the other IP domains. End-to-end QoS should be accomplished by means of the RSVP or some QoS access signaling protocol.

Future work should address security issues in the context of QoS environment, including the needs of QoS signaling protocols, preserving QoS guarantees in the case of rapid changes in resource availability, as well as the inter-domain security aspects.

References

[1] G. A. Giannopoulos, "The application of information and communication technologies in transport," *European Journal of Operational Research*, pp. 302–320, Jan. 2004.

[2] M. Stojanović, "A novel approach to design of telecommunication networks in the power utilities," *Journal of the Union of Yugoslav Electric Power Industry*, vol. 56, no. 3, pp. 74–86, Sept. 2003, also contribution to CIGRÉ WG D2.07.

[3] M. Stojanović and V. Aćimović-Raspopović, "Multiservice telecommunication networks in road traffic and transport," in *Proc. of the 6th Symposium TES 2004*, Sombor, Serbia and Montenegro, Apr. 2004.

[4] ——, "A framework for implementing QoS in power utility telecommunication networks," in *contribution to CIGRÉ WG D2.07*, Mar. 2004.

[5] S. Giordano, S. Salsano, S. V. den Berghe, G. Ventre, and D. Giannakopoulos, "Advanced QoS provisioning in IP networks: The european premium IP projects," *IEEE Communications Magazine*, vol. 41, no. 1, pp. 30–36, Jan. 2003.

[6] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," *RFC 2475, IETF*, Dec. 1998.

[7] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol label switching architecture," *RFC 3031, IETF*, Jan. 2001.

[8] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource reservation protocol (RSVP) - version 1 functional specification," *RFC 2205, IETF*, Sept. 1997.

[9] X. Xiao, T. Telkamp, V. Fineberg, C. Chen, and L. M. Ni, "A practical approach for providing QoS in the internet backbone," *IEEE Commmunications Magazine*, vol. 40, no. 12, pp. 56–62, Dec. 2002.

[10] J. Gozdecki, A. Jajszczyk, and R. Stankiewicz, "Quality of service terminology in IP networks," *IEEE Communications Magazine*, vol. 41, no. 3, pp. 153–159, Mar. 2003.

[11] H.-L. Lu and I. Faynberg, "An architectural framework for support of quality of service in packet networks," *IEEE Communications Magazine*, vol. 41, no. 6, pp. 98–105, June 2003.

[12] H. J. Chao and X. Guo, *Quality of Service Control in High-Speed Networks*. New York: J.Willey & Sons, 2002.

[13] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397–413, 1993.

[14] Network simulator ns2 and network animator NAM. [Online]. Available: http://www.isi.edu/nsnam

[15] J. Malek. (2003) Trace graph - network simulator ns trace files analyzer. [Online]. Available: http://www.geocities.com/tracegraph