Internet Service Providers (ISPs) provide service differentiation to support current and future development process of new Internet applications and services. The solution is to use a QoS architecture, one that would assure implicit service differentiation based solely on the traffic characteristics without any possibility of undesirable interference by ISPs or Internet users.

The net neutrality problem, which is one of the key issues in the Internet traffic policy, discriminates selected applications which may be moved to a more expensive tier. It may cause that the network performance will be available to only a fraction of users. The second scenario: “corporate bureaucracy nightmare” represents the possibilities for charging extra money by the large corporate broadband firms and telephone or cable companies. The former may require a special charge for the access to the upper tier, e.g., for a new adventuresome web site. At the same time the latter may insist that the users connected to that site are moved to a more expensive tier. It may cause that the network will become more expensive for the users. The third scenario, named “bad incentive nightmare” describes the situation where ISPs have their own services (like VoIP) and may block or discriminate the competition. The last presented scenario, called “less innovative content nightmare” involves worries that companies with a substantial market power may offer the higher tier, where the profit margins will be more lucrative.

The term ‘net neutrality’ is considered in a wide variety of ways. In [4], four nightmare scenarios for the net neutrality are presented. The first one, called “inequity nightmare”, assumes that companies with a substantial market power may offer the higher tier, where the profit margins will be more lucrative. The net neutrality followers show that the investment in the upper tier may lead to developing an advanced Internet that will be available to only a fraction of users. The second scenario: “corporate bureaucracy nightmare” represents the possibilities for charging extra money by the large corporate broadband firms and telephone or cable companies. The former may require a special charge for the access to the upper tier, e.g., for a new adventuresome web site. At the same time the latter may insist that the users connected to that site are moved to a more expensive tier. It may cause that the network will become more expensive for the users. The third scenario, named “bad incentive nightmare” describes the situation where ISPs have their own services (like VoIP) and may block or discriminate the competition. The last presented scenario, called “less innovative content nightmare” involves worries that the firms may produce new applications and services and protect their interests in those applications giving no chance for using and developing them by other providers. In [5] an additional nightmare scenario is presented. It is possible that ISPs will charge application providers twice (firstly to its users and secondly to its applications) for using and developing them by other providers. In [5] an additional nightmare scenario is presented. It is possible that ISPs will charge application providers twice (firstly to its users and secondly to its applications) for using and developing them by other providers. In [6] four nightmare scenarios for the net neutrality are presented. The first one, called “inequity nightmare”, assumes that companies with a substantial market power may offer the higher tier, where the profit margins will be more lucrative.
own ISP and secondly to the ISP of every single user who wants access to that application). It may begin to break the unique many-to-many nature of how information is linked in the Internet. The net neutrality problem and the complexity of its definition reflect a conflict of interests between application providers (APs), Internet users (IUs), and ISPs.

The groups of APs and IUs second the net neutrality in the most rigid version believing that any service differentiation whatsoever should be prohibited and that all traffic in the Internet should be handled in the best effort manner. They argue that current network links have capacities high enough to carry all traffic with the proper guarantees. In most cases it is really true because network resources are often overdimensioned, but we have to be aware of an enormous progress in the telecommunications, especially in the area of access networks. New applications and services grow rapidly and the number of Internet users rises significantly as well. One of the fathers of the Internet, Lawrence Roberts, predicts that in ten years time, the great majority of world’s population will be online [6]. It may cause that, in future, the network link capacities will not be sufficient to carry all traffic with a proper QoS. It is one of the key arguments raised by ISPs to allow for service differentiation and provide traffic priorities. They claim that the data transmission in the networks without QoS mechanisms may become unacceptable from the user point of view in the future. ISPs are also concerned that providing net neutrality may discourage network investments and development of new services and applications. The net neutrality followers may agree with differentiation of services, but without additional charging for it. They fear that some applications may be blocked or a traffic of the selected services might be poorly treated if appropriate fees are not paid to the ISPs. For example, it may be impossible to set up a VoIP connection or its quality may not be acceptable if a user does not pay more than for the basic service.

There are also some proposals of how to cope with the net neutrality problem. The debate on how to guarantee the proper quality of a transmitted traffic in the IP based networks has been a hot topic for the last 15 years. We have to pay attention to the fact that some solutions are not scalable, do not work as they are supposed to, or allow for user misbehavior. The network architecture that will allow for implicit traffic differentiation and prioritization of a selected traffic without user or ISP intervention may be a desired proposal. It may ensure net neutrality with the awareness of QoS. In our paper, we propose such a solution based on the Flow-Aware Networking architecture.

III. MOTIVATION FOR FAN

Introducing QoS into IP networks seems inevitable in the future, since the traffic rate is constantly growing. The real problem is that each proposed QoS architecture, including the most mature, i.e., IntServ and DiffServ, has some flaws. IntServ’s problems with scalability are well described in the literature, e.g., in [7] and [8]. Although considered as a leading proposal, DiffServ is criticized mainly for its increased granularity and complexity. As both architectures may be considered orthogonal, they represent a trade-off between scalability and possible quality of service guarantees [9]. Although DiffServ capable routers are commonly available, operators rarely use this functionality, and even if so, only locally. Therefore, it is necessary to find a new architecture, one that could really be used in all IP networks in the future.

The other point of view concerns the net neutrality issue. Common QoS architectures, including DiffServ, provide means for the network operators to differentiate the service without any limitations. It is possible to discriminate traffic based on the application type, source or destination addresses, traffic volume, etc. It is also possible to implement a Deep Packet Inspection mechanisms [10] and police the traffic based on its mother application or content. However, since most of the differentiation actions are against the net neutrality, choosing such a powerful and complex solution is neither useful, nor cheap. The real goal, therefore, is to find a solution which could be used with the IP protocol, would be simple, efficient, scalable, and in conformity with the network neutrality. Thereby, we advocate Flow-Aware Networking [11] as a QoS architecture which fits perfectly into the mentioned boundaries. Consequently, in this section, we present the main difference between classic approaches to QoS assurance and Flow-Aware Networking. We also explain why FAN seamlessly blends into the net neutrality concept while introducing QoS awareness. In Section IV, the Flow-Aware Networking concept is presented in details.

![Diagram of QoS Approaches](image)

**Best effort**

**Standard QoS**

**Flow-Aware Networks**

Figure 1 shows a concept diagram of various approaches to assuring quality of service in IP networks. Originally, the
Internet was designed to be simple and efficient. However, mainly due to its simplicity, the traffic can be carried only in a best effort way (Figure 1a). It means that we do not control the quality of service, solely relying on the over-provisioning of the link resources.

Figures 1b and 1c present two different methods to achieve the QoS. The standard approach (Figure 1b) focuses on dividing the traffic into classes and treating different classes separately. The term “class of service” is understood differently in various proposals, yet, always the idea revolves around a separate treatment of different transmissions. Firstly, a flow must be identified by the classifier block and assigned to a certain class. The classification process is explicit, because nodes are pre-informed on how to recognize and treat a particular transmission.

Flow-Aware Networks (Figure 1c), on the other hand, operate quite differently. First of all, the flow identification process is implicit and its goal is not to divide flows into different classes, but only to create an instance on which the service differentiation will be performed. Then, all the flows that are currently in progress, i.e., are present on the Protected Flow List (PFL) are forwarded unconditionally, and all the new flows are subject to admission control. The admission control in FAN is measurement based (MBAC) which implies that the accept/reject decisions are made based only on the current link congestion status.

The main advantage of FAN, with respect to the net neutrality issue, is that it provides service differentiation, taking into account only the traffic characteristics of the ongoing transmissions. Therefore, it is not possible to discriminate certain applications or end-users. Moreover, instead of providing different treatment, FAN introduces fairness, which even enhances the current IP networks equality.

Table 1 summarizes the differences between various approaches to QoS guarantees. The current best effort Internet, as proved over the last decade, has no problems with scalability and, by some net neutrality proponents, is seen as the only possible solution for the future. Unfortunately, best effort does not support service differentiation of any kind which is its obvious drawback. IntServ and DiffServ were chosen as the most common representatives for standard class-based QoS architectures. For the price of introducing explicit service differentiation, these architectures degrade the network scalability, significantly increase its complexity, and become net neutrality unfriendly. Flow-Aware Networks, on the other hand, retain high scalability, while introducing implicit service differentiation techniques. Additionally, FAN’s QoS methods do not interfere with the view of the neutral Internet. FAN, as opposed to IntServ and DiffServ, does not allow to provide an explicit differentiation by the ISPs. It is a very important advantage of this technique. Of course, ISPs may try to change a router’s software and provide a traffic classification which allows for packet queuing and servicing according to their rules. However, such behavior is opposed to FAN principles and, as so, it is more difficult to introduce than in, e.g., DiffServ.

IV. FLOW-AWARE NETWORKING

In this section, we present Flow-Aware Networking as a QoS architecture that is easy to implement, scalable, and above all, does not interfere with net neutrality issues. The concept of Flow-Aware Networking as an approach to assure quality of service in packet networks was initially introduced in [12] and, then, presented as a complete system in 2004 [13]. The goal of FAN is to enhance the current IP network by improving its performance under heavy congestion. To achieve that, certain traffic management mechanisms to control link sharing are introduced, namely: measurement-based admission control [11] and fair scheduling with priorities [13], [14]. The former is used to keep the flow rates sufficiently high, to provide a minimal level of performance for each flow in case of overload. The latter realizes fair sharing of link bandwidth, while ensuring negligible packet latency for flows emitting at lower rates.

A. Cross-Protect

FAN is supposed to be an enhancement of the existing IP network. In order to function properly, an upgrade of current IP routers is required. Figure 2 shows a concept diagram of a Cross-Protect router (XP router), the standard interconnecting device in FAN. FAN adds only two blocks to the standard IP router. They are namely: admission control block and scheduling block. The former is placed in the incoming line cards of the router, whereas the latter is situated in the outgoing line cards.

![Fig. 2. Concept diagram of a Cross-Protect router [13]](image)

Admission control is responsible for accepting or rejecting the incoming packets, based on the current congestion status.
If a packet is allowed, the flow associated with it is added to the protected flow list (PFL), and then all forthcoming packets of this flow will be accepted (see Figure 1c). The admission control block implements the measurement-based admission control (MBAC) functionality [15]. Therefore, admission control in FAN complies with net neutrality as it is not class-based or user-based: each new flow obtains the same treatment, and in case of congestion, all new flows are blocked. Such an approach may also be considered as “unfair” service differentiation as in terms of congestion, some flows are admitted and some are blocked. However, MBAC treats all the flows equally, i.e., a) the decision of accepting or rejecting the traffic affects all new incoming flows, not just a part of them, b) admission decisions are implicit, based only on internal measurements.

MBAC relies on congestion measurements that are performed within the scheduler. Two indicators are constantly monitored: fair rate (FR) and priority load (PL). Following [13], “fair rate is an estimation of the rate currently realized by backlogged flows”, and represents the amount of link’s bandwidth, which is guaranteed to be available for a single flow, should it be necessary. Similarly, “priority load is the sum of the lengths of priority packets transmitted in a certain interval divided by the duration of that interval”, and shows the amount of data that is prioritized. The manner of calculating both indicators is a feature of the proper scheduling algorithm, and is presented in [13].

The queue management in FAN is realized in the scheduling block of an XP router. Fair queuing ensures that link bandwidth is shared equally, without relying on the cooperative behavior of end-users. This is a different approach than in currently used IP routers, where, usually, the FIFO queue is implemented. The difference in both mentioned approaches is presented in Section IV-B.

Naming FAN devices as “Cross-Protect routers” is a result of a mutual cooperation and protection, which exists between both discussed blocks. The admission control block limits the number of active flows in the XP router, which essentially improves the queuing algorithm functionality, and reduces its performance requirements. It is vital, that queuing mechanisms operate quickly, as for extremely high speed links, the available processing time is strictly limited. On the other hand, the scheduling block provides admission control with the information on congestion status on the outgoing interfaces. The information is derived based on, for example, current queues occupancy. The mutual protection contributes to a shorter required flow list and queue sizes, which significantly improves FAN’s scalability.

It has been proven in [16] that fair queuing is scalable since complexity does not increase with link capacity. Moreover, fair queuing is feasible, as long as link load is not allowed to attain saturation levels, which is asserted by admission control. Compared to other QoS architectures, FAN scalability, due to the lack of signalling and very low data handling complexity, is not matched by any other architecture [17].

Finally, migrating towards FAN is relatively inexpensive, compared to other QoS architectures. Building a “Cross-Protect router” is an easy task as most required functionalities are already developed and accessible by vendors. Fair queuing and admission control mechanisms are currently available, therefore, to implement FAN, only slight changes in software need to be done. Additionally, FAN is transparent to the current IP network and can be introduced gradually, starting from the most congested links.

B. Service differentiation

In FAN, admission control and service differentiation are implicit. There is no need for a priori traffic specification, as well as there is no class of service distinction. Both streaming and elastic flows achieve a necessary quality of service without any mutual detrimental effect. Nevertheless, streaming and elastic flows are implicitly identified inside the FAN. This classification, however, is based solely on the current flow peak rate. All flows emitting at lower rates than the current fair rate are referred to as streaming flows, and packets of those flows are prioritized. The remaining flows are referred to as Fig. 3. Flow rates under congestion in: a) classic FIFO and b) FAN.
elastic flows. As the prioritization process in FAN is based only on the current bit rate of each flow, no special treatment can be imposed by the ISP, hence, the net neutrality paradigm is maintained.

To explain how, in general, service differentiation in FAN works, some ns-2 simulations were performed, and the results are presented in Figure 3. The plot shows the obtained bit rates of 5 UDP flows, ranging from 100 up to 500 kbit/s on a bottleneck 5 Mbit/s link. The background traffic, which was started after the 100 second of the simulation, consisted of 300 TCP flows having on average 1 MB data to send, and starting on average 1.5 second after the previous one. Figure 3a) presents a classic IP link with a FIFO queue. Before the 100th second each presented flow realizes its desired bit rate, as the bottleneck link capacity is available. After that point, the congestion occurs due to the background traffic causing all the flows to reduce their obtained bit rate. The amount of the reduction is more or less proportional to the original flow bit rate.

The FAN link (Figure 3b), on the other hand, behaves differently. Low rate flows are prioritized, hence, they do not even notice the congestion. High rate flows are to fairly share the remaining bandwidth, which due to the PFQ algorithm, they do. The minimum FR value was set to 5% of the link capacity (250 kbit/s), therefore, MBAC admitted only a certain number of flows to achieve this limit. This threshold can easily be set to a different value. For instance, using 8% of the link capacity as a minimum FR value prioritizes 100, 200 and 300 kbit/s flows and keeps the elastic flow rates at approximately 375 kbit/s minimum.

The service differentiation provided by FAN, despite its obvious limitations, is very useful and practical. The priority is given to the low rate applications, including the Internet telephony service, low-rate video conferencing and streaming, gaming, remote desktop connections, and many other. Additionally, flows rated as elastic also obtain good-enough quality of service due to the admission control’s preservation of the fair rate. As nothing comes without a price, in case of congestion, MBAC has to temporarily deny the access of certain flows, however, the waiting time is justified by the promise that once a flow is admitted, it is to be served with at least decent QoS.

The mentioned connection waiting process may sometimes be highly unwanted, like in case of emergency VoIP calls. This issue is described in [18] along with a simple solution which may mitigate the problem. Moreover, a brief analysis of VoIP call transmission under the congestion state in the network link without QoS mechanisms and in a FAN link is presented in the next section.

C. VoIP call transmission analysis

In this section we show how FAN may improve the network performance in the congestion state. We made 100 simulation runs in various conditions to show the packet loss (Fig. 4a), traffic rate (Fig. 4b) and packet delay (Fig. 4c) of flow representing a VoIP call between two nodes in the network with best effort service and in FAN. The simulation environment was the ns-2 simulator. We decided to provide the traffic pattern with
the exponential distribution for generating the time intervals between beginnings of the transmissions by those flows. The duration of each simulation run was set to 250 s. In FAN, the PL values were estimated every 50 ms while the FR values were estimated every 0.5 s. The maximum PL parameter was set to 70% what means that if the estimated PL values were higher than 70% of the link capacity the congestion was encountered. The minimum FR parameter was set to 5% which means that if the estimated FR values were lower than 5% of the link capacity the congestion was encountered. The flow_time_out parameter was set to 20 s what means that if a flow was inactive during 20 s, its identifier was removed from the PFL. In our simulation scenario the flow representing the VoIP call (in FAN, called streaming) began the transmission at 30 s and the number of the background flows, which had the traffic to send, was changed. 95% confidence intervals were calculated by using the Student’s t-distribution.

The values of the observed packet loss parameter increase with the increasing number of the background flows in the scenario with the best effort service (Fig. 4a). In the same scenario the bit rate of the VoIP call decreases with the increasing number of background flows (Fig. 4b). The values of the packet delay parameter for that flow do not change (Fig. 4c). The obtained results are as expected. In the network with the best effort service the increasing number of active flows causes the decreasing transmission rate and the increasing number of lost packets of each flow and does not affect the values of the packet delay parameter. In the analogous scenario with the FAN architecture the packets of a streaming flow are not lost if the identifier of that flow is written to the PFL. Moreover, the transmission rate of the streaming flow is constant and equal to the desired value independently of the changing number of the background elastic flows. The mean values of the packet delay in this scenario are also constant independently of the number of elastic flows and negligible if we assume that the streaming flow represents a VoIP call. It is very important from the net neutrality point of view. As one can see, in FAN the implicit traffic differentiation is realized. ISPs do not have any possibility to prefer the selected traffic based on the source or destination addresses or any other parameter except the traffic rate. However, there are means to give some kind of priority to emergency services [18]. We also have to note that a new streaming flow may not be accepted in the admission control block immediately in the congestion state. It may cause that the VoIP call may begin with a significant delay in opposite to the network with best effort packet service where it begins immediately. The mechanisms for resolving this problem are presented in [19] and [20].

V. CONCLUSION

The significance of the net neutrality problem force the researchers to propose and develop new solutions for QoS guarantees. The Flow-Aware Networking is a solution that meets net neutrality assumptions and allows for implicit service differentiation. Using this architecture, ISPs will not have to implement any traffic policies or explicit QoS mechanisms to guarantee proper traffic performance. Moreover, they will not be able to do it, and in consequence to charge extra money from Internet users. FAN, originally simple, is a viable proposal for the future Internet. It perfectly fits into both the followers and opponents of the net neutrality concept. The simulation results presented in the paper confirm the usefulness of FAN in this context. We are convinced that the solutions proposed in this paper will contribute to solving the net neutrality problem with satisfaction to any side.

ACKNOWLEDGMENT

Project operated within the Foundation for Polish Science Ventures Programme co-financed by the EU European Regional Development Fund.

REFERENCES

The First International Conference on Evolving Internet

INTERNET 2009

23-29 August 2009
Cannes/La Bocca, France

Editors
Eugen Borcoci
Shawn McKee
Alessandro Bogliolo
Oana Dini

Includes
HOWAN 2009: The First International Workshop on Hybrid Optical and Wireless Access Networks
NEUTRAL 2009: The First International Workshop on Neutral Access Networks