Network Traffic Modeling for Load Prediction: A User-Centric Approach

Aleš Švigelj, Radovan Sernec, and Kemal Alič

Abstract

Nowadays, networks have to be able to cope with ever increasing traffic demands in order to deliver the desired quality to end users. Thus, proper network planning is essential in order to preserve telecom revenues by reduced income per bandwidth unit. This article addresses a user-centric approach to network and user traffic modeling that has been validated and used in the process of introducing, optimizing, and planning new services at the Slovenian national telecom operator and service provider, Telekom Slovenije d.d. The proposed approach is based on the end users and their user group profiles that are founded on real measurements from the observed telecommunication network consisting of more than 1000 MSANs and more than 300,000 subscribers. The proposed approach has been successfully validated, showing that for the observed period the modeled link load deviates less than 5 percent from the measurements. Furthermore, in the presented case study the proposed approach is used successfully in the process of introducing the Fast Channel Change service.

Studies on Internet traffic trends show staggering growth rates of between 70 and 115 percent per year [1]. This remarkable expansion in growth is the result of new applications, modifications in user habits, and changing trends:

- New industries and operator types such as data- center providers and cloud operators are present in the market.
- Devices that enable "non-stop" content generation and consumption, such as tablet computers and smartphones, have found their way into our everyday lives.
- New technologies have enabled a rapid change in user behavior patterns (e.g., instead of watching linear broadcast content via classic television, the same or personalized and self-selected content can be accessed via YouTube).
- DVD stores or rental outlets seem to be forgotten as video content can be accessed through online streaming services such as Netflix.

At the moment mobile Internet traffic represents 15 percent of total Internet traffic. However, the trends in 2011 and 2012 suggest that mobile Internet traffic is increasing at 50 percent per year, and that tablet computer users generate 2.5 times more traffic than those using smartphones. In addition, the migration to the fourth-generation (4G) mobile communications technology standards such as Long Term Evolution (LTE) promise 19 times more traffic than we had to cope with in the 3G mobile communications technology standards. By 2017 this will result in 70 percent of traffic having its origin in mobile devices, with mobile traffic increasing by 13 times and that of video by 16 times. For these reasons, telecoms are being pushed into the development of systematic methodologies and tools that enable them to cope in a timely and cost-effective way with such abrupt changes in network traffic demands [2].

Efficient network design and capacity dimensioning have to be employed to provide network resources that meet the traffic demands. However, a balance between investments maintaining network-level quality of service (QoS) and providing end-user satisfaction (i.e., a good quality of experience, QoE) even in partial network failure situations has to be found in a competitive business environment. Since service providers are introducing new services and billing schemes, which can heavily influence the traffic load on a regular basis, it is of paramount importance that such changes are analyzed appropriately in a simulation environment in order to omit network and server congestion, or even fallouts, and maintain the desired QoE.

The modeling of network traffic [3] for load prediction is a necessary tool for modern telecom operators. It allows them to close the life cycle control loop as follows: network development — deployment — operations — optimization — upgrade of expensive equipment. Simulation models allow the close monitoring of planned network upgrades and their performance evaluation. "What-if" simulation scenarios are especially appealing for planned (not yet deployed) network elements and services, since we can predict the behavior of the entire network by taking into account the behavior of various users and the network load conditions.

In this article we present a traffic model that adopts a user-centric approach and is employed in a large-scale-network

Aleš Švigelj and Kemal Alič are with Jozef Stefan Institute and the Josef Stefan International Graduate School.

Radovan Sernec is with Telekom Slovenije.

0890-8044/15/$25.00 © 2015 IEEE

planning framework. It is primarily developed for network planning and the optimization of real large-scale networks, in particular when introducing new services. The approach provides us with a better understanding of the Internet traffic's behavior, which is essential for the planning and management of existing networks, as well as for designing next-generation networks [4]. An important aspect of the described approach is that we can evaluate the behavior of users and make accurate decisions about short- and long-term network planning operations. Traffic is investigated to the point that each source-demand pair is defined per service per user. However, users are not modeled individually, but are aggregated into groups.

We use this approach on a full-scale real network (on the national telecom operator scale), where the load prediction and network redesign were made for specific network solutions that enhanced the performance of IPTV, that is, Fast Channel Change (FCC), Retransmissions (RET), and Catch-up TV (CupTV).

The article is organized as follows. In the next section the network-planning framework architecture is described, while the user-centric traffic modeling approach is discussed after that. The case study on using the complete framework and developed traffic model is then described and evaluated. The final section concludes the article.

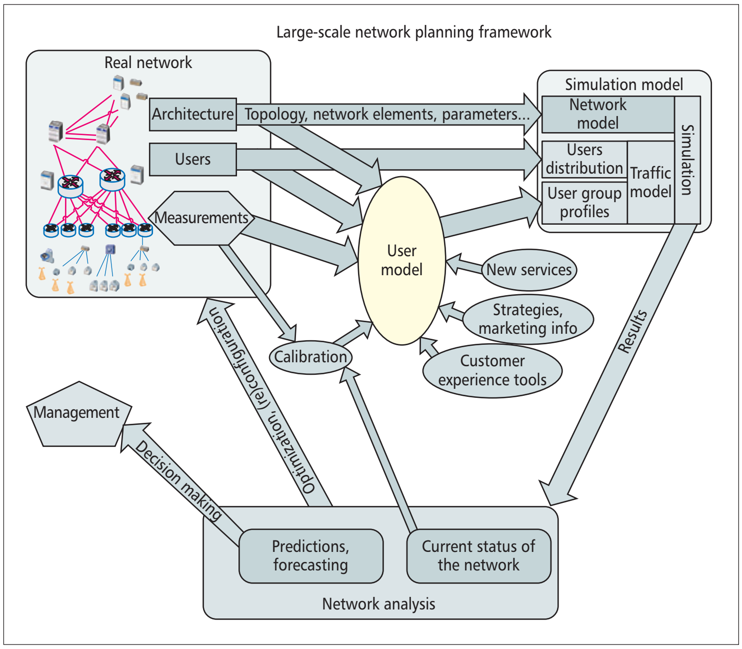

Figure 1. Network planning framework architecture.

Large-Scale-Network Planning Framework

The large-scale-network planning framework is depicted in Fig. 1 and consists of four major blocks: real network, user models, simulation model, and network analysis. A real network can be represented as a set of network elements, with links between the elements defining a topology. It is worth noting that all these elements are defined by numerous parameters (e.g., capacity), which influence the overall network performance. Thus, to build a proper simulation model we need to define the simulation network model, which is a map of the real network architecture, automatically distilled from the database of the telecom operator. Additional elements used for simulation purposes, such as control or in-depth analysis, can also be added.

The traffic model in the simulation model is very important. Different approaches to traffic representations exist [5], and they are usually based on flows [6, 7] and traffic matrices [8, 9]. The authors of [10] went a step further and represented traffic on a per-service basis. It is important that a selected traffic model for the backbone network take into account new services and that it is able to predict their future bandwidth requirements.

When building a traffic model, two basic problems need to be considered, as described in [11]. The traffic volume levels need to follow the spatial and temporal patterns observed in realistic networks, and the traffic volume level needs to be assigned to a particular node pair in a given topology. Our approach to building a network traffic model provides a solution to both of these issues. In addition to that, the proposed approach proved as simple, straightforward, and exact when introducing new services.

The most important parts of the network are the end users, and the network is built to serve their requirements. The traffic in the network (i.e., the traffic demand) is created by the end users, either directly or indirectly. Similar to this real network experience, we built the traffic model around the end users. By doing so we can extract, for every part of the network, the influence on the traffic intensity of the particular group of users at a particular location.

The traffic flows are based on the user group profiles (UGPs). The users are grouped into user groups for each observed service, and the user groups are assigned different profiles that take into account the traffic volume and its temporal distribution. The UGPs are based on information from the real network, such as load measurements (IP level), user locations, and distribution, as well as information about the new services, strategies, marketing plans, and so on.

Even though the initial UGPs are calculated, fine tuning is required. The traffic model is calibrated by comparing the real network measurements with the simulation model results representing the current status of the network. Thus, the UGPs are refined until the real measurements and simulation results are not adequately aligned. Then the traffic model can be used in the simulation model for network optimization and/or (re)configuration, predictions, and forecasting according to the company's strategies.

The simulation results can be used for better decision making of the management staff. The whole process of setting up the traffic model takes place in the user model block, which is broadly presented in the next section.

The proposed framework is most suitable in cases where the network and services are under the full control of a single operator, but it can also be used in the general application/user concept for a particular service provider or partial network operator. However, in this case the optimization can be performed only for the selected services optimizing the partial/rented infrastructure. In case services are provided from different operators, they can either be separated completely, thus modeled as transit traffic (i.e., the service provider just rents some "pipes") or use the main operator's infrastructure; thus, they can be modeled using the proposed user-centric approach even though the services are provided from different service providers.

Furthermore, in addition to user-centric traffic, there is a significant amount of traffic with different origins, such as transit traffic, which simply depends on the transit business agreement between the involved parties. This unique load can be accounted for by simply adding well defined traffic flows to the existing traffic model.

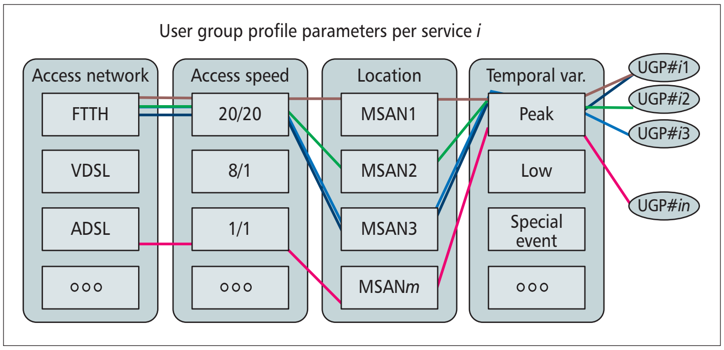

Figure 2. Definition of user group profile parameters per individual service.

User-Centric Modeling

We built the traffic model around the paradigm that the entire traffic load in the network has as its foundation the end users. Users generate all the traffic, coming either directly from them, such as in the case of peer-to-peer (P2P) applications or directly to them, as in video on demand (VoD); or they have a cumulative impact on the service, as in the case of a dynamic IP television (IPTV) multicast stream. Hence, the traffic in our model is caused by the end users (specific traffic, e.g., traffic that only depends on the transit business agreement, is considered separately). Users are grouped into user groups defined per service according to different criteria, where their profiles are described. It is worth noting that the word *user* is employed in a broader sense, meaning that it may represent many users on a common subscriber line (e.g., a common household). Although all the traffic is considered, only the dominant traffic and investigated new services are modeled in detail, while the rest of the traffic is modeled as aggregate other traffic, broken down into several subgroups depending on how it is distributed around the network. In our particular case we were able to categorize around 90 percent of the traffic (the value changes depending on the time of observation).

User Group Service Profiles

In order to automate the classification of individual users in different per service UGPs, we used information that is already available in the databases of the telecom operators/service providers and is useful for characterization (e.g., what kind of terminal equipment the users have, like a smart TV box, what kind of subscription line they are using, to which services particular user is subscribed). In addition to the information available in the databases, we also used load measurements on particular links, again collected from the telecom operator's equipment, suchas a multi-router traffic grapher (MRTG).

By combining relevant and accessible information, the UGP for individual service contains the following (Fig. 2):

- The access network type is either fiber to the home (FTTH) or one of the digital subscriber line (xDSL) Internet access technologies.
- The access speed characterizes the information about the end user's Internet access speed. This only applies if the end user is an Internet customer.
- Temporal variations in the load generated per service used. The measurements represent samples at interesting moments throughout the day, or follow seasonal changes. The simulation can be conducted for any given set of points, although the most interesting are the peak hours.
- End-user location in terms of the access point. User habits also differ locally. The differences can be noticed for domestic users in different areas; however, the differences are most noticeable when comparing industrial and residential areas.

The UGPs are defined per individual service, where the services offered to the end user can range from classic IPTV, the Internet (we model different Internet services individually, that is, P2P, macromedia streams, etc.), VoIP, and so on, to relatively new ones, such as CupTV or FCC. We only model services that are demanding in terms of bandwidth consumption and have a large impact on the overall traffic load in the network or at particular servers. Attention is also devoted to services that have a low bit rate at the user side, but have a large impact when many users are using the same service at almost the same time (e.g., FCC, when a commercial break starts).

Using this approach we define the UGPs for each of the services enabled by the service provider, for example, an asymmetric DSL (ADSL) subscriber using an 8/1 Mb/s access speed located at multi-service access node (MSAN) *m* at a given time produces a VoD stream between the VoD application server and MSAN *m* with a traffic intensity of *y* b/s, represented as UGP#i3 in Fig. 2.

Users with the same access type but located on different MSANs generally have a different UGP for a selected service, though in practice it often happens that the profiles are the same (see also Fig. 2, that is, UGP#i1 is the same for MSAN#1 and MSAN#3), especially within the same geographical areas, which can be seen as the first aggregation level in the hierarchical topology. In practice, we have noticed that the UGPs for Internet services such as P2P are mostly dependent on the access network and access speed.

From the measurements, we found out that users with higher bandwidth access speed produce high cumulative traffic bursts (peaks) whether downloading or uploading, while users with lower bandwidth produce more constant cumulative traffic loads. For example, as shown in Table 1, a group of FTTH users can produce up to 100 percent more cumulative traffic load as the same size group of ADSL users in the downlink (DL). The compared groups produce even bigger differences in the uplink (UL), where for the given example FTTH users produce more than 30 times more traffic than ADSL users.

From the measurements we have not noticed any difference in television viewing habits for users using xDSL access technologies, while FTTH users often access more television channels simultaneously. For example, for the MSANs analyzed in Table 1 with only FTTH users, 269 IPTV users produced approximately 15 percent higher IPTV multicast stream than 282 ADSL users. It can be interpreted that FTTH users watch one channel while recording a second one, or have multiple TV sets in the same household. Similarly, on MSANs with users using a variety of access technologies (e.g., ADSL and FTTH), the number of multicast IPTV channels for a similar number of users is somewhere between xDSL and FTTH. Nevertheless, the number of multicast channels cannot exceed the number of all available channels (i.e., 195 in our case).

Network Observations

In order to set up and later on calibrate the UGPs, the traffic load measurements are performed throughout the network, as depicted in Fig. 3. The network under observation, for which the framework was developed, has a typical hierarchical structure. We model the network from the access points up. The MSANs are points where the end users are normally connected to the backhaul and also the place where the user profiles (denoted by *UGP#xy* in Fig. 3) are situated in the simulation model. Moving up the network, aggregation routers follow with the core in the center. The services are positioned on top

		Time of the day						
		8 am (Mb/s)	15 pm (Mb/s)	8 pm (Mb/s)	Average (Mb/s)	Average/user (Mb/s)	Average daily peaks (Mb/s)	Ratio: peak/average
MSAN "F": only FTTH (412 users)	UL	48	142	124	122	0.296	459	3.76
	DL	185	242	290	167	0.405	472	2.83
MSAN "A": only ADSL (408 users)	UL	6	8	7	4	0.010	15	3.75
	DL	162	183	203	95	0.233	221	2.33

Table 1. Traffic load measurements on two MSANs.

of the network on the service delivery point (SDP), which is usually a set of network components that provide main services delivery architecture for a different type of services. SDPs tend to be optimized for the delivery of a service in a given technological or network domain (e.g., web, IMS, IPTV, mobile TV). They provide environments for service control, creation, and orchestration and execution, as well as abstractions for media control, presence/location, integration, and other low-level communications capabilities. The number of SDPs in a network depends on the network size and the services it provides, and in our case there is one primary SDP. In addition to SDPs, there are also local services hubs denoted as service delivery hubs (SDHs) in Fig. 3, which has similar functionalities as SDP, but only for specific services (e.g., for FCC) and are generally located on the aggregation level. Please note that the dashed connections from the SDHs to the routers represent an optional placement of the SDH at a particular aggregation level and are highly dependent on the used services.

Depending on the location and the measurement tools available at a particular location, different types of measurement data can be obtained. In its simplest form, for each of the links we can obtain load measurements (temporal changes and intensities). The user group behavior has been recorded at selected MSAN locations where groups of users have been observed in terms of their load intensity and temporal variation. As an example, in Table 1 measurement from MRTG for traffic load for two MSANs, each providing access to approximately 400 Internet users, is presented. In this particular case we were investigating FTTH and ADSL user behavior; hence, we selected MSANs where only one type of access is available (the first MSAN serves only FTTH users, while the second one serves only ADSL users). Both MSANs are located in an urban area (within the same city). Average throughput measurements for UL and DL at three different characteristic weekday time periods (i.e., 8:00 a.m., 3:00 p.m., and 8:00 p.m.) have been averaged over four weeks. The highest traffic is measured in DL at 8 p.m.; thus, this period is used for peak traffic load evaluation. Long-term averages (four weeks) per user show that on average an FTTH user produces approximately 1.7 times more traffic than an average ADSL user in the DL direction and approximately 30 times more traffic in the UL. The ratio between average daily peaks and average traffic load, which is 2.83 in the case of FTTH and 2.33 in the case of ADSL users, shows that higher traffic bursts can be expected from FTTH users in DL. In the UL direction even higher bursts can be expected; however, the ratio is approximately the same (i.e., 3.76 and 3.75) for both types of users, although the total ADSL user peak traffic is 30 times lower than FTTH, hence having smaller total impact on the network. These types of results have also been used as a foundation for building UGPs, as explained below.

At a few locations the loads can also be classified according to the services used, and finally, there are links in the network that are occupied only by certain types of service, thus showing the exact aggregate service characteristics. These links appear mostly at the service hubs, although they can also be found inside the network.

Setting Up User Group Profiles

Based on all the information gathered from different sources, the UGP loads need to be set. The ultimate goal is to set the UGPs in such a way that the simulated traffic will be as close a match as possible to the real network measurements at different measurement points. The solution is not straightforward as load measurements cannot be obtained where we want them and do not contain all the information we would like to have. The initial UGPs are therefore based on short-term measurements conducted at specific locations and of selected groups of users. For example, we observe the lowest aggregation level end point where specific users are predominant or only one type of user is present (e.g., we observe an MSAN with only FTTH plugins in an urban area, or an

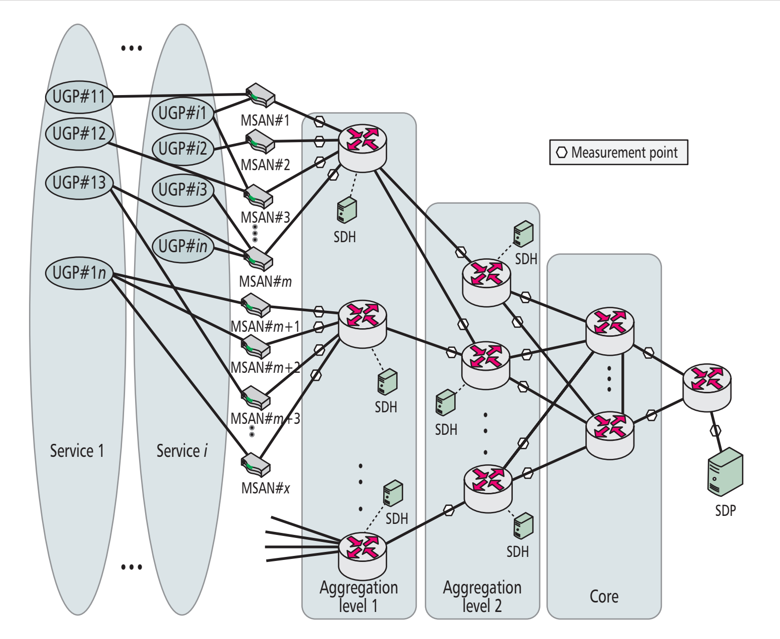

Figure 3. System architecture.

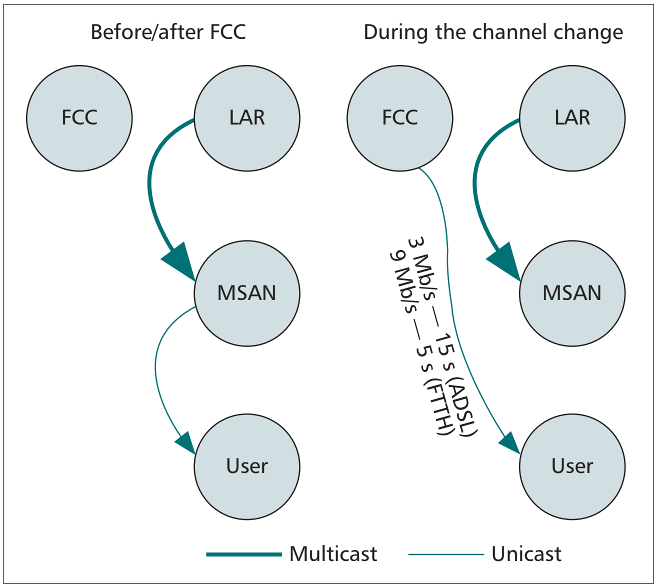

Figure 4. FCC service modeling (traffic stream).

MSAN with FTTH plugins in an industrial zone, or xDSL users in a rural area). Measurements are collected over several days. From these measurements we define the amount of traffic for a particular "access network" at a given "access speed" for all the services (as explained in the measurement example above and UGP example below). Modeled services included in specific UGPs were modeled at different levels of precision, and depended on the measurements we were able to obtain. Services modeled can be categorized into three larger groups:

- Internet traffic (P2P traffic, macromedia, HTTP, etc.)
- SDP services (VoD, VoIP, etc.) and specific traffic (IPTV multicast, time shift, FCC, retransmission, etc).

The distribution of traffic load between services is based on additional measurements. General Internet traffic is routed through different paths than the specific SDP application's traffic. Thus, in addition to being able to determine the amount of Internet traffic, we were also able to inspect Internet traffic, where we determined the sharers of different types of Internet traffic; for example, P2P traffic represents 66 percent of Internet traffic. and macromedia 13 percent. Based on the traffic load measurements and specific ways of routing the traffic, we were also able to evaluate the amounts of P2P traffic terminated within and outside the observed operator's network.

Services available from the network operator are observed at SDP and SDH points for each individual service. The traffic generated at the observed time was for initial profiles evenly distributed among all subscribers. Let us explain the approach using an FCC service example on how particular services were modeled. The basic behavior of FCC service is presented in Fig. 4. When the FCC enabled user switches between TV channels, a new unicast stream (lasting on average 5 s for FTTH users and 15 s for xDSL users) is established from an FCC SDH. After the transition time, the user receives the stream from MSAN multicast. Although the uniform usage of the service does not pose any major challenge to the network, the simultaneous requests in short periods, usually during commercials, significantly increase the traffic.

When the new service is introduced with as yet unknown user behavior, the problem is tackled in two phases. In the first phase heuristic assumptions are made, and preliminary simulations for selected case scenarios are performed in order to obtain basic understanding of the FCC influence on the network. In the second phase, measurements on the service provider endpoint are performed. In the FCC case the service was made available to around 10,000 test users. The measurements showed that at the peak time (i.e., around 7 p.m., when the commercials start) they generate more than 10,000 requests per minute. Note that when dealing with an existing service only the second phase is used. Based on the measurements, we produce different UGPs that are distributed in the network in different ways, also taking into account the available end-user terminal equipment that supports the service. The FCC UGPs based on test user group were used in simulations to position FCC SDHs and to evaluate their influence on the network, as shown in the results.

In our case the measurements from MSANs to aggregation level 1 are used to correct UGP, thus obtaining the local habits. By doing so the ratio between services in higher layers does not match the measurements sufficiently anymore, so the profiles need to be corrected accordingly. Traffic tuning is then performed through an iterative process by creating small

	Traffic flow intensity (kb/s/user)			
	ADSL 8/1 Mb/s user access type		FTTH 10/10 Mb/s user access type	
	DL	UL	DL	UL
HTTP	9.8	3.9	19.6	8.7
Macromedia	24.5	4.1	49.1	12.5
P2P in	28.4	4.6	56.8	115.5
P2P out	59.0	7.2	117.9	150.6
Other Internet services	4.9	0.5	9.8	9.3
VoD	4.8	0.0	6.2	0.0
Other SDP services	8.8	1.2	21.1	14.2

Table 2. Normalized UGPs examples on selected MSAN for two user access types at 8 pm.

Load deviation	< 1%	< 2%	< 5%
Link fraction	42.4%	57.6%	100%

Table 3. Traffic model calibration results.

	Direction	Average load deviation (%)	σ (%)
CORE <--> Aggregation level 2	-->	1.5	0.6
	<--	1.7	0.7
CORE <--> BRAS	-->	0.9	0.5
	<--	1.8	0.9
BRAS <--> Aggregation level 2	-->	1.7	1.3
	<--	1.5	1.5
Aggregation level 2 <--> Aggregation level 1	-->	2.6	2.3
	<--	2.3	2.2

Table 4. Normalized UGP examples on a selected MSAN for two user access types at 8 p.m.

modifications to the user profiles and a comparison of the simulation results and network measurements. We would like to point out that traffic measurements at the second and third aggregation level are not taken into account in the iterative process, but are used for the evaluation purposes; comparison results are presented and discussed in the case study section and in Tables 3 and 4.

After the initial tuning, the modifications to the profiles due to modified user habits and new services can be made quickly and efficiently. The main benefit of using the described approach is that changes in the network (re)configuration or additional users are taken into account in the traffic model automatically, since they are used in the process of setting up traffic flows as data related to these modifications is based on the databases of the telecom operator, where changes appear simultaneously with the changes in the real network. As with actual changes, it is simple to make case studies such as network (re)configurations. Added value of the presented work is that the traffic model is built around user groups and on modeling of their behavior, which is dependent on a number of different parameters and features. Some of them can be obtained from the operator's databases (e.g., access type, speed, equipment, prepaid services, location), while some parameters are measured (e.g. traffic intensity). In addition to observing the growth of the traffic in the network caused by "general Internet" growth or, in our case, daily profiles, our aim is to observe and study all application services as one distributed system with internal interactions. For example, the introduction of CupTV, which allows offline watching of broadcast, influences the IPTV multicast. Users watching CupTV at a certain time do not use IPTV multicast anymore, so the bandwidth required for the IPTV multicast is reduced.

UGP-Based Traffic Model Construction

Traffic flows for the network traffic model are generated using UGPs and the information on the number of users in the user profile group. Well-defined user profile groups define the amount of traffic generated by the individual service during the observed time period at a given user location used in the traffic flow matrix.

The service under observation and the end-user group location define the flow source and destination. Traffic flow sources or destinations are where the end users normally connect to the backhaul, that is, at the access nodes (MSANs), while the amount of traffic load is aggregated to fit the real number of users in the user group. The amount of traffic caused by users for a particular service always depends on the number of users. In the case of Internet traffic such as P2P traffic, the flows can simply be summed on all aggregation levels, while in the case of, for example, IPTV, the more users using IPTV, the more likely it is that a larger number of television channels will be multicast from the dedicated application server. The IPTV multicast traffic is affected at different levels of the backhaul and the backbone in a different manner, as in the real network not all the TV channels are multicast to a particular MSAN, but only those channels that are in use by the users. In addition, when users request a particular channel that is not included in the multicast stream, at the beginning the unicast channel is streamed (to have faster response to user request), and only later is the channel added to the multicast stream. In practice, although there are more than 150 TV channels, only about 30–50 are running concurrently in the multicast stream at the MSAN.

Example UGPs for two access types at the peak hour (i.e., 8 p.m.) on a selected MSAN with 548 Internet and 283 IPTV users normalized to one user are presented in Table 2. UGPs for services that are available to all end users regardless of their terminal equipment and subscription are listed. The values in the table are normalized per one user. For example, the UGP for VoD traffic on this MSAN for each ADSL user with access speed 8.1 Mb/s at the peak hour is on average 4.8 kb/s. This value is then multiplied with the number of users in order to obtain the estimation of VoD traffic on MSAN for a particular access type user group. Only when normalized values are multiplied with the number of users do the obtained values also have practical meaning. The same is true also for other services. Let us also point out that for each service we not only define the source destination pairs but also paths for the traffic that generally also differ (e.g., Internet traffic including macromedia and P2P is routed through different alternative paths than SDP traffic such as VoD).

Case Study

The described system has been developed for and used by Telekom Slovenije d.d., the national telecom operator. It is the largest Internet provider in Slovenia and can be considered a small-to-medium-sized telecom operator on a global level. It has more than 300,000 Internet subscribers and more than 100,000 IPTV subscribers, which are connected to more than 1000 MSANs. The whole telecom operator's IP/multiprotocol label switching (MPLS) aggregation network and all the users were considered in the case study. The planning framework was used for the optimization purposes of the server(s) placement (e.g., the SDH) for the deployment of new services and planning the network link capacities.

In general, the whole simulation process was made in two phases: the traffic load calibration, and the exploitation of the model for optimization and planning purposes.

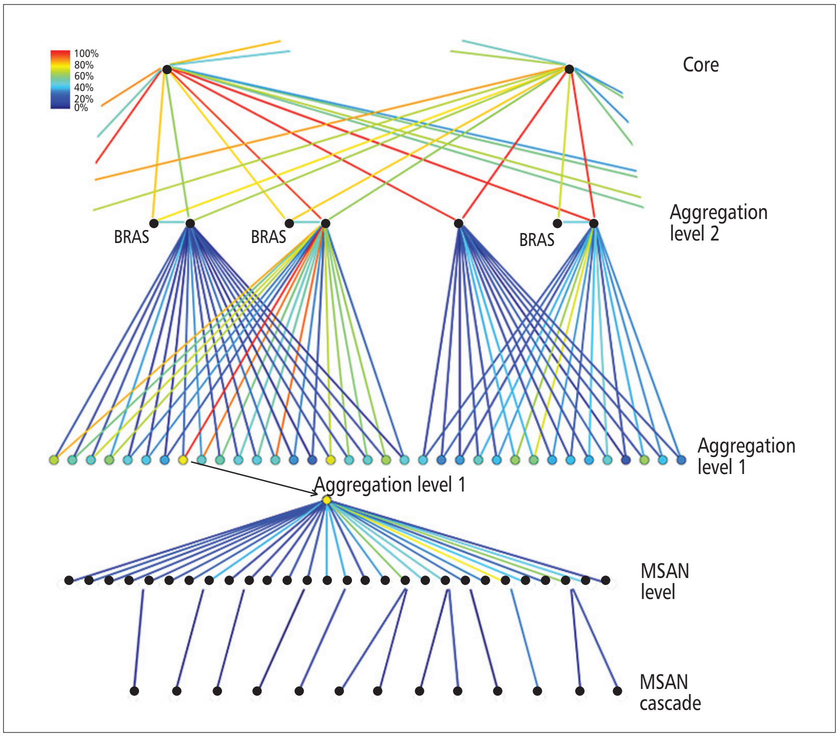

Figure 5. Predicted network load after the introduction of FCC service to all users.

In the first phase the described user-centric approach was used to appropriately model the user traffic in order to match the current traffic conditions in the network for the particular observation point (e.g., peak load). Iterative user group model adaptations were made to the point where we estimated the user-generated traffic as being matched to the real network sufficiently well. Then comparison at the aggregation 2 level was performed. The correctness of the simulation tool was evaluated by comparing the simulation results and real measurements. A sample of such evaluation results is given in Table 3, where we compare the measurements from the real network with the simulation results. The simulation was conducted using for this purpose a developed simulation tool based on *MATLAB* and *C*. We compare the 20-minute average measurements at 8 p.m. over four weekdays in the middle section (between aggregation levels 1 and 2) of the network. The deviation was calculated as the ratio between the absolute difference between the two measurements and the link capacity. The proportion of links is the share of all the links that meet the deviation requirement. We can see that we have been able to model the UGPs to the extent where the modeled traffic is a close match to the real network measurements (i.e., the load deviation is less than 5 percent on the aggregation level 1 and above links). However, detailed analysis has been performed on links and aggregation points where the simulated traffic and measured traffic differed most. In such particular cases, we have always identified specific traffic sources connected directly to the routers on aggregation level 1 (and not on the MSAN as is the case with general end users), such as the source point for the streaming VoD service provider. These specific users are in the later phases modeled as individual specific users, or individual traffic flows have simply been added to mimic their behavior.

It is worth noting that the deviations at a single MSAN could be much higher, particularly if there is a small number of users, as the models are in this case statistically irrelevant. However, a practical investigation of such links shows that they are not loaded heavily, and thus do not give us any reasons for finer tuning. Furthermore, since there are few users, the traffic load origin from there is low, thus meaning that the overall MSAN contribution of uncertainty to the overall network performance is low.

In addition to aggregation level 2 evaluation results (Table 3), in Table 4 simulation results and measurements at 8 p.m. are compared on different aggregation levels in both directions (DL (—>) and UL (<—)) (please also refer to Fig. 5). The 20-min average measurements at 8 p.m. over four weeks of weekdays are used as reference measurements. As expected, our model produces better results at higher aggregation levels, where higher numbers of users are considered, and thus the evaluation for average user is more precise, and anomalies in particular user behavior have smaller impact on traffic model.

In the second phase, the simulation framework was used for different purposes, but mainly for an evaluation of the new services to be deployed in the network:

- Retransmission (RET) of the deteriorated IPTV stream: a multicast stream is repaired for xDSL users.
- Fast Channel Change (FCC): supporting faster television-program switches. The unicast stream from a dedicated server to the end user is created, for the time the change is taken into account in the multicast stream.
- Catch-up television (CupTV), which allows viewers to watch the shows outside the broadcast schedule. This content is not stored locally, but rather at the sourced hub offices located higher in the network hierarchy.

The complete workflow of the described framework as used in the decision-making process and the network-planning process is as follows:

1. The initiative for the new service comes from the product-management department. They are interested in the feasibility from the network point of view, possible additional investments, and so on. The inquiry is usually broad ranging, from offering a new service to all the users to offering it to only a specific segment of them.
2. In cooperation with the technical team we define new UGPs for the additional services. In the first step, intuitive decisions based on engineering expertise, past experience and marketing predictions are made for the new user-profile creation. For example, 30 percent of IPTV users that have the service enabled and are currently watching the television will access the CupTV service. This data also affects the IPTV multicast stream since there are fewer viewers following the live broadcast and the stream is reduced accordingly.
3. The adapted UGPs, and thus the new traffic model, are used in the simulation model for the definition of the number and positioning of the application server, testing link failures, observing the load increase and detecting possible bottlenecks caused by introduction of the new service.
4. Marketing and strategists analyze the results provided by the technical teams. The decision on for whom to enable the service, price positioning, and marketing strategies are made. The plans for end-user short- and long-term service penetration are made.

5. Given the new data on the expected number and distribution of users, the final application server positioning is made, and the expected network behavior is examined in greater detail. In our case stress tests were also made.
6. When new equipment is set up and the service is running, a representative user test group is selected. They are observed carefully, and their behavior with respect to the new service is measured.
7. With the new measurements and thus better knowledge of the users' behavior, the UGPs can be adapted, and new simulation results are obtained. Possible investment plans in the network infrastructure can be refined.
8. When the first real users begin using the service their behavior is observed, and possible new findings are used in the UGPs and thus in the upcoming analyses. Worth mentioning is that new users normally show different habits compared to already established users. The new service requires testing and using if for no better reason than because it is there. The established users take the service for granted and already show habits that can be observed nicely at the group level.
9. Even when the service is already running for a longer period of time, it is important to keep track of user behavior. This is because user habits change. For instance, CupTV could become an excellent way to avoid commercial breaks during shows, as users tend to login late and follow the live stream throughout the show, catching up just at the end of the show. This modified user behavior has to be detected, and the influence of such a change has to be taken into account in the UGP, affecting the traffic models, simulation results and long-term load predictions.

In order to show an illustrative example of the presented framework, in Fig. 5 we depict the predicted impact of FCC service on the observed network. The dots represent the aggregation points or routers, which are connected with colored links showing its utilization (dark blue 0 percent -> red 100 percent). Please note that links which also ensure redundancy (e.g., above MSANs) are considered "full" (i.e., 100 percent) if they are "utilized" 50 percent. For clarity and non-disclosure agreement reasons only a representative portion of the network is presented, and the real names of MSANs and aggregation levels elements are blurred. For visualization purposes, as it is difficult to display all the MSANs (there are more than 1000 of them) at once, the color of the aggregation level 1 dots indicates the traffic load intensity of the link with the highest link utilization at the MSAN level of the particular aggregation point. Thus, the color of the eighth point is the same (yellow) as the color of the link with the highest utilization (sixth link from the right in the depicted case) between MSAN and observed aggregation point.

In the case where no FCC is used, there are no DL links with utilization higher than 20 percent even during peak time. In Fig. 5 traffic impact on the network (DL direction) is shown for the scenario where all IPTV users with terminals supporting the service are using FCC in prime time, when the commercials start. It is clearly seen that without an upgrade of particular links or restricting the FCC service, several links have too small capacity (denoted by yellow and red), so disruption of other services can be expected. Congestion is expected at the MSAN aggregation level and also in the higher aggregation levels. The first can easily be solved by adding additional capacity, while the second can be solved by introducing additional SDH in the network on the lower levels or increasing link capacity.

The use of the described framework for the introduction and optimization of new services as described above proved to be a useful tool, not only for network-capacity planning, but also for a better understanding of the whole network and user behavior.

Summary

Users' expectations have increased in almost every field of products and services. The average user expects the latest services to be of high quality, easy to use, and unlikely to fail. Similarly, end users demand that customer service is able to answer questions quickly and resolve any issues in short order. In this competitive environment, service providers and telecom operators are looking to reduce the time and price required to introduce new products and services. The pressure to do so increases daily. The risk to a strong brand reputation from a poorly executed launch is high. Thus, it is of paramount importance that new services be well planned, the equipment cost optimized, and deployment tested in a simulation model. Beyond these rather obvious advantages of such a solution, its potential use can also be beneficial to capital and operational cost reduction in network operations, dynamic pricing of network infrastructure for interconnected service providers and telecom operators, a lower-energy-consuming network, and so on.

In this article we have presented a user-centric approach to traffic modeling in the concept of a network planning framework, which is suitable for shortening the time to market of different services, and large-scale-network optimization. As we put the user in the center, by creating UGPs, defining the general user behavior in terms of traffic load, the introduction of new services, which usually target user groups, is fast and efficient. In addition, by modifying the UGPs to reflect potential special events (e.g., sport events), telecom operators and service providers can prepare for increased demands on particular services. With the described framework we can also evaluate the network in terms of link/node failures. The framework is based on the network level (i.e., IP) traffic flows that are defined with the UGPs, and can also be used for large-scale simulations as shown in the case study, where we simulates the whole nationwide telecommunications network with multiple access points' granularity.

In the case study we showed the use of the network planning framework for new service introduction in a medium-sized national telecommunications network with 300,000 subscribers. The framework proved to be useful and efficient for both (i) the sales department planning the service penetration and (ii) the network planning department, which has to meet the market demands with an optimized investment cost.

Although we demonstrate the traffic modeling and load prediction on fixed access networks, the same methods and solutions are equally well applicable (although by imposing much more dynamics) to mobile networks (e.g., the network access point becomes the eNodeB), which will be considered in our future work.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20122017," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf, retrieved Sept. 2013.
[2] U. Sedlar *et al.*, "Contextualized Monitoring and Root Cause Discovery in IPTV Systems Using Data Visualization," *IEEE Network*, vol. 26, Nov./Dec. 2012, pp. 40–46.
[3] M. Roughan *et al.*, "Spatio-Temporal Compressive Sensing and Internet Traffic Matrices (Extended Version," *IEEE/ACM Trans. Networking*, vol. 20, June 2012, pp. 662–76.
[4] J. Yang *et al.*, "Characterizing Internet Backbone Traffic from Macro to Micro," *Proc. IEEE Int'l. Conf. Network Infrastructure and Digital Content '09*, 2009. Beijing, China.
[5] S. Stoev, G. Michailidis, and J. Vaughan, "On Global Modeling of Backbone Network Traffic," *Proc. IEEE INFOCOM '10*, San Diego, CA, 2010, pp. 196–200

[6] C. Barakat *et al.*, "A Flow-Based Model for Internet Backbone Traffic," *Proc. 2nd ACM SIGCOMM Wksp. Internet Measurement*, 2002, pp. 35–47.

[7] K. Shiomoto, I. Inoue, and E. Oki, "Multi-Layer Network Operation and Management for Future Carrier Backbone Networks," *Proc. IEEE GLOBECOM '08,* New Orleans, LA, 2008, pp. 1–5.

[8] D. Jiang and G. Hu, "GARCH Model-Based Large-Scale IP Trafc Matrix Estimation," *IEEE Commun. Lett*ers, vol. 13, Jan. 2009.

[9] A. Soule *et al.*, "Traffic Matrices: Balancing Measurements, Inference and Modeling," *Proc. ACM SIGMETRICS Int'l. Conf. Measurement and Modeling of Computer Systems*, 2005.

[10] E. Palkopolou *et al.*, "Traffic Models for Future Backbone Networks: A Service-Oriented Approach," *Euro. Trans. Telecommun.*, vol. 22, Dec. 2011.

[11] A. Nucci, A. Sridharan, and N. Taft, "The Problem of Synthetically Generating IP Traffic Matrices: Initial Recommendations," *ACM SIGCOMM Computer Communication Review*, vol. 35, July 2005, pp. 19–32.

Biographies

Aleš Švigelj (ales.svigelj@ijs.si) was awarded his Ph.D. from the Faculty of Electrical Engineering, University of Ljubljana, Slovenia, in 2003. He is a research fellow in the Department of Communication Systems at the Jozef Stefan Institute, Slovenia, and an assistant professor at the Jozef Stefan Postgraduate School. He has extensive research in modelling, simulation, and design of advanced telecommunications elements, systems, and services. His current work focuses on advanced networking technologies for wireless systems.

Radovan Sernec (radovan.sernec@telekom.si) was awarded his Ph.D. from the Faculty of Electrical Engineering, University of Ljubljana, in 2000. He works in Telekom Slovenia's R&D Department as a senior researcher and strategist. His research interests include network architectures and topologies of interconnection networks, also for data centers, sustainable renewable energy models for telco operators, and innovation management within enterprises.

Kemal Alič (kemal.alic@ijs.si) received his B.Sc. and M.Sc. degrees in electrical engineering from the University of Ljubljana in 2005 and 2008, respectively. He is a researcher at the Department of Communication Systems at the Jozef Stefan Institute and a Ph.D. student at the Jozef Stefan International Postgraduate School. His recent research interests are in the fields of network protocols, architectures, and cross-layer optimization for wireless and wired networks.