Generation and Analysis of Service-Based Traffic Flows

Alvaro Bernal, **Marc Ruiz, and Luis Velasco***

Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain

**e-mail: lvelasco@ac.upc.edu*

ABSTRACT

The rapid availability of new services makes that network operators cannot exhaustively test their impact on the network or anticipate any capacity exhaustion. This situation will be worse with the imminent introduction of the 5G technology and the kind of totally new services that it will support. In this paper, we present CURSA-SQ, a methodology to analyze the network behavior when the specific traffic that would be generated by groups of service consumers is injected. CURSA-SQ includes input traffic flow modelling with second and sub-second granularity based on specific service and consumer behaviors. The methodology allows to accurately study traffic flows at the input and outputs of complex scenarios with multiples queues systems, as well as other metrics such as delays.

Keywords: service-based traffic generation, logistic queue model, aggregated traffic models.

1. INTRODUCTION

The advent of 5G networks impose enormous challenges for network operators and vendors, since new services will require stringent quality of service (QoS) from the network. In fact, before 5G deployment and service commercialization, the impact on the traffic injected to Multiprotocol Label Switching (MPLS)-over-optical metro and core networks need to be considered so they can be adequately planned. Nonetheless, no real monitoring data is available for the targeted networking scenarios. Incipient services to be supported by 5G network technologies limit the availability of real monitoring data to only what it can be obtained from test-beds, which, in most of the cases, do not represent those realistic scenarios that autonomic networking pursues. To overcome the lack of real monitoring data, analyzing *synthetically generated traffic data* becomes a requirement to validate network design before they enter into operation.

Trying to replicate the observed self-similarity and long-range dependency in packet network traffic, several theoretical models have been based on stochastic processes. These models can be used within *discrete-event simulators* to generate discrete random input (*packet*) traffic propagated by a queue system that models the network under study. However, traffic generation based on discrete stochastic processes requires a set of parameters to be fit, which entails having real traffic traces. In this paper, we propose a fast, accurate, attainable, and scalable *service-centric traffic flow analysis methodology* based on statistical flow characterization, named CURSA-SQ. Starting from the packet traffic generated by single service consumers, CURSA-SQ generates synthetic network traffic, as well as other related traffic variables resulting from the activity of consumers and providers of 5G services for a wide range of use cases.

2. SERVICE-CENTRIC TRAFFIC FLOW ANALYSIS

In this Section, we present a general overview of the CURSA-SQ methodology. Without loss of generality, let us consider a scenario where a network operator provides connectivity between *service consumers* and *service providers*. Figure 1 illustrates the scenario, where service is requested by the consumers; the *upstream* traffic arrives from service consumers in a network node that aggregates and forwards it toward the selected service provider, whereas in the *downstream* direction, such node forwards the traffic coming from a service provider (in response to service requests) to the specific service consumer.

We are interested in studying and generating traces of the aggregated traffic flows as a function of consumers traffic flows (hereafter, *input traffic*) and the characteristics of the network node (e.g., link capacity). To reduce the number of input traffic flows, we group consumers of the same type of service and with the same characteristics. Finally, a consumer group can be served from one or more locations of the same provider.

We will use different traffic flow generators for upstream and downstream traffic. Those generators will generate traffic flows, in terms of bitrate, with granularity T fine enough to study flows (in the order of hundreds of milliseconds) but several orders of magnitude higher than those typical times and sizes of packet-based traffic generation (Fig. 2a). In the upstream direction, one single flow generator per consumer group will be used to produce the traffic flow for all the active consumers in the group; this flow generator will be located at the consumer group location and will target one or more service provider's sites. In the downstream direction, each service provider's site will contain a flow generator to produce the traffic flows toward the consumer groups.

The generation process is summarized in Fig. 2b; it is based on first characterizing each service (labeled 1 in Fig. 2b) to find the upstream and downstream traffic characteristics (2) for one single service consumer. Then, the traffic flow bitrate is generated by scaling the traffic characteristics to the number of active consumers forecasted for a given time period (3), while transforming the characteristics from the discrete to the continuous domain (4). The following groups of characteristics have been identified: 1) *Consumer behavior*: these characteristics capture the behavior of the consumers of a specific service. 2) *Data exchange*: these characteristics focus on how the service generates the data to be transferred according to consumers' activity.

3) *Consumer infrastructure:* these allow adapting the data exchange to packet traffic since network infrastructure can impact the service. These service-related characteristics are not deterministic, but they follow statistical distributions. Therefore, by analyzing them, the packet traffic that every individual consumer introduces in the network can be modeled in terms of a few *random variables* capturing how bursts (and even packets) are generated by a single active consumer. The most relevant random variables are: *i*) *inter-arrival burst rate*, defined as the rate between consecutive bursts; *ii*) *burst size*, defined as the number of bytes transmitted in a burst; *iii*) *inter-arrival packet rate*, as the rate between consecutive packets in a burst; and *iv*) *packet size*, as the total amount of bytes (headers included) of a packet.

Once input traffic flows are generated in terms of bitrate for every period and every direction, they are used to generate aggregated traffic flows. To this end, a number of upstream input upstream traffic flows are aggregated, and the resulting flow feeds a queue system (Fig. 2c). The reverse process is followed in the downstream direction; the input traffic flows are aggregated (not showed in the figure) and the resulting traffic flow enters a queue system; at the output, a disaggregator separates the resulting flow into the defined traffic flows.

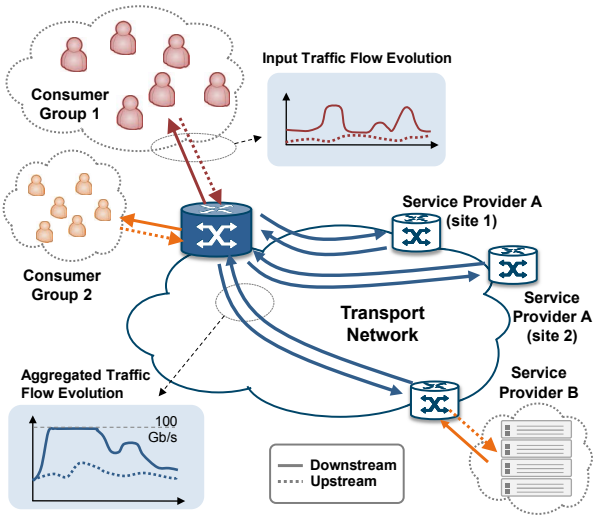

Figure 1. General overview of targeted scenarios.

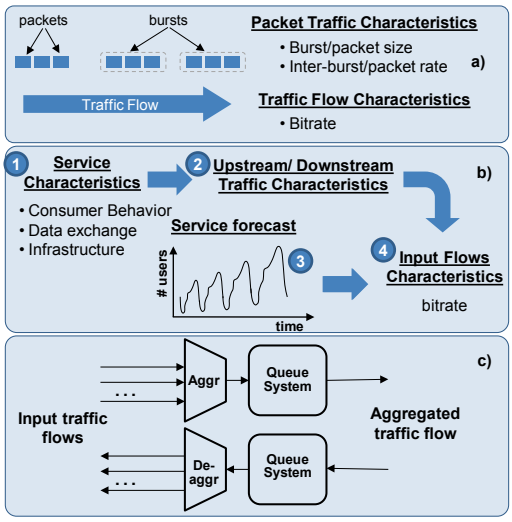

Figure 2. Overview Of The CURSA-SQ methodology.

3. INPUT TRAFFIC FLOWS AND TRAFFIC ANALYSIS

Since even simple studies entail generating input flows that aggregate many service consumers, a meaningful part of the CURSA-SQ methodology is devoted to reducing the computational effort of generating large amount of fine granular traffic flows while ensuring the required accuracy. Then, we first propose statistical and mathematical models to generate aggregated input flows feeding the queue systems in practical execution times, and next, the general CURSA-SQ methodology to generate traffic flows is detailed.

From the perspective of a flow aggregating several individual active consumers, the effect of both packet size and inter-arrival packet rate variables can be neglected compared to burst size and burst inter-arrival rate. Since such traffic characteristics do not depend on the number of active consumers, the main source of input flow variations is precisely the evolution of consumers over time. Variations in the expected number of active consumers need to be modeled to capture any pattern, such as periodic behaviors (e.g., a daily pattern) or evolutionary trends (e.g., an annual increment). With these in mind, let us define the following random variables to model the traffic flow of a specific consumer group aggregating consumers of the same service:

ibr	Inter-arrival burst rate (s^{-1}), defined as the rate of consecutive bursts.	*bs*	Burst size (bits)
r	Consumer maximum flow rate (b/s)	γ	Traffic burstiness degree
$x(t)$	Bitrate (b/s) generated by a consumer group or service provider site	$u(t)$	Number of active consumers at time *t*
T	Traffic generation granularity (s)		

Since bitrate is expressed in b/s units, let us consider $T = 1$ s as a reference. We consider a modelling approach based on computing approximations of the expectation (E) and variance (V) of $x(t)$ based on the expectation and variance of *ibr*, *bs*, and $u(t)$; these can be easily obtained assuming prior knowledge on service traffic random variables distribution and active consumers models. Note that the product of *ibr* and *bs* results into a new random variable representing the bitrate generated by one single user. $E(x(t))$ can be approximated as the product of the expected number of users and the expected single user bitrate (eq. (1)). Regarding the variance and assuming that *bs* and *ibr* are independent, the variance of the individual user bitrate can be derived according to well-known expressions to estimate the variance of the product of two independent variables (eq. (2)). Then, $V(x(t))$ can be approximated as the sum of the variance of individual users. According to the definition of a consumer group and the independence assumption, $V(x(t))$ can be estimated (eq. (3)). The model in eq. (1) and eq. (3) allows generating random traffic flows with the selected *T*. To that aim, a pseudo-random generator

function ϕ following a given distribution, e.g., uniform, Gaussian, etc., can be used to generate random traffic x'(t) according to $E(x(t))$ and $V(x(t))$. See eq. (4), where $u(t)\cdot r$ is the maximum traffic that the consumer group can inject/receive due to access speed constraints.

$$E\left(x(t)\right) \approx E\left(u(t)\right)\cdot E\left(bs\cdot ibr\right) = E\left(u(t)\right)\cdot E\left(bs\right)\cdot E\left(ibr\right) \quad (1)$$

$$V\left(bs\cdot ibr\right) = V\left(bs\right)\cdot V\left(ibr\right) + E\left(bs\right)^2\cdot V\left(ibr\right) + E\left(ibr\right)^2\cdot V\left(bs\right) \quad (2)$$

$$V\left(x(t)\right) \approx E\left(u(t)\right)\cdot V\left(bs\cdot ibr\right) \quad (3)$$

$$x'(t) = \min\left\{u(t)\cdot r, \Phi\left(E\left(x(t)\right), V\left(x(t)\right)\right)\right\} \quad (4)$$

$$\gamma = \frac{bs/r}{bs/r + 1/ibr}. \quad (5)$$

$$x''(t,i) = \begin{cases} \min\left\{u(t)\cdot r, \gamma^{-1}\cdot x'(t)\right\}, & T\cdot\sum_{j=0..i} x''(t,j) < x'(t) \\ 0, & T\cdot\sum_{j=0..i} x''(t,j) \geq x'(t) \end{cases} \quad (6)$$

Although eq. (4) works fine generating random traffic flows for $T \geq 1$ s traffic flows with sub-second granularity need to be generated to estimate queuing delays. Such sub-second scale generation must reproduce the nature of a bursty traffic with on-off periods producing short intervals of high activity that fill queues up.

To this aim, a flow $x''(t,i)$ with sub-second granularity is generated from $x'(t)$; index i represents the i-th interval T within the one-second interval centered in t. To allow computing maximum expected delays, a worst case of traffic bursty behavior is considered, as sketched in Fig. 3. Specifically, the example of $x'(t)$ flow in Fig. 3a is used to produce the $x''(t,i)$ with T=100 ms. in Fig. 3b; every bitrate sample in x'(t) is transformed into 10 samples in $x''(t,i)$. Within every one-second interval, a first *on* period where bitrate can exceed that of $x'(t)$ is followed by an *off* period where bitrate is fixed to 0. Note that the summation of all samples in $x''(t,i)$ within one-second interval equals the bitrate in $x'(t)$. The number and magnitude of samples in the *on* period depends on the degree of burstiness γ of the traffic of the consumer group, and it is computed as in eq. (5), where γ thus, represents the proportion of time within a second where traffic is actually generated. Then, the generation of random traffic samples with sub-second interval is defined as eq. (6) Finally, it is worth noting that, if $T > 1$ s, $x'(t)$ can be easily computed by averaging random samples and $x''(t,i)$ do not need to be computed.

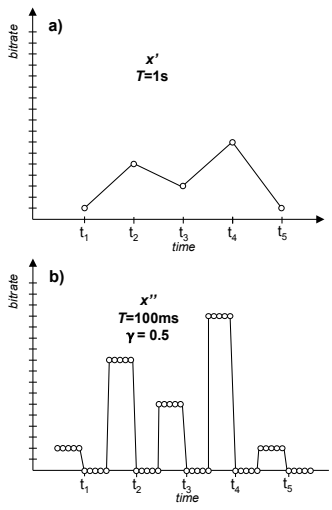

Figure 3. Traffic generated with second (a) and sub-second (b) granularity.

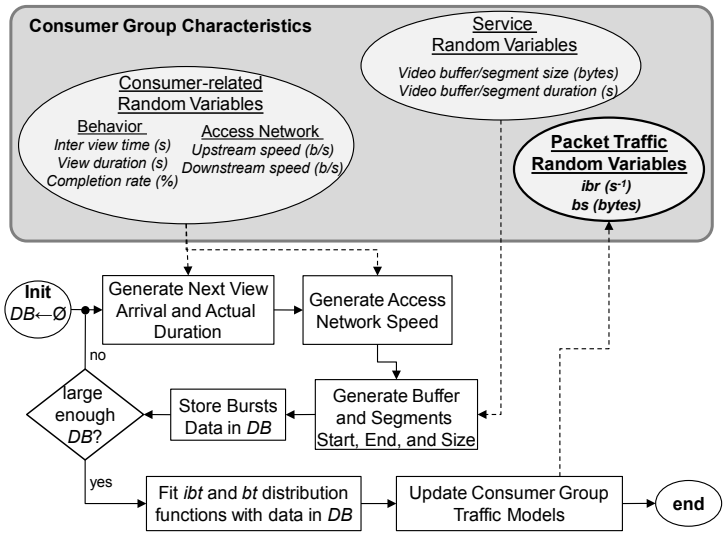

Figure 4. CURSA-SQ methodology applied to VoD traffic analysis.

4. NUMERICAL RESULTS

For the subsequent studies, we will consider three different services, namely: *VoD*, *Gaming*, and *Internet*. According to the CURSA-SQ methodology, relevant studies available in the literature providing consumer and service-related random variables characterization were used to characterize traffic sourced by consumer groups. Table 1 summarizes the expectation and variance of *ibr* and *bs* for these services.

Table 1. Services traffic characteristics.

Service	$E(ibr)$ (s^{-1})	$V(ibr)$ (s^{-1})	$E(bs)$ (MB)	$V(bs)$ (MB)
VoD	0.25	2.54e-5	3.84	1.21
Gaming	1.33	0.19	0.14	0.02
Internet	1.66	0.40	0.12	0.04

Let us detail the characterization of the VoD; regarding consumer behavior, according to the study presented in [1], the idle time y that an active user spends (e.g., deciding which content to watch) follows the power law probability distribution $p = \alpha \times y^{-\beta}$, with parameters $\alpha = 0.43$ and $\beta = 1.2$. On the other hand, the duration of the content selected by a user approximates an exponential distribution with a typical mean around 30 minutes and a reasonable maximum of 4 hours [2]. However, users usually stop a reproduction before its completion time. Completion rate depends on the content duration; the longer the duration is, the smaller the completion rate. A Weibull distribution with scale and shape parameters around 75 and 0.8 fits with a large variety of contents'

duration. Regarding service-related VoD characteristics, we adopt a typical on-off pattern consisting of an initial 10 – 20 s transmission of media contents, followed by a number of 2 s media segments, until the reproduction finishes. According to the previously defined statistical distributions, we simulated the activity of a single consumer and stored the time stamp and size of 10.000 traffic bursts. The analysis of this data lead to the VoD consumers traffic characteristics detailed in Table I, that indicates long spaced bursts of large number of bytes.

A similar procedure was followed to characterize gaming and Internet consumers' traffic from key statistical distributions detailed in [3]-[5]. The resultant traffic characteristics differ from that of VoD in both, the frequency of bursts (high *ibr*) and its size (small *bs*). Note that Internet traffic is the one that shows the highest variance in terms of *ibr*, which translates into a less regular traffic pattern.

Aiming at validating the CURSA-SQ methodology including the aggregated input traffic flow model and the logistic queue model, we developed a *packet-based* simulation environment for benchmarking purposes. Specifically, a packet input traffic generator produces packets streams creating of a fixed size creating 1500-byte Ethernet frames, according to the specific mean and variance of *ibr* and *bs*; a packets stream is generated independently for each individual user. Then, the aggregated packets stream is sent to a simple queue system with one *discrete* queue, which processes packet by packet. This combination of packet-based traffic generation and discrete queue simulation provides the baseline performance for comparison purposes.

The CURSA-SQ methodology and the discrete simulator were implemented in Python 2.7. For each defined service, we considered a scenario with a single consumer group configured with a constant number of users. We run several executions with incremental number of users. Every execution generated a random flow of one day long and T = 1 s according to eq. (4) that was used for input flow comparison purposes. Then, a sub-second flow with T = 50 ms was generated according to eq. (6) to evaluate the performance of the logistic queue model; both discrete and logistic queues were configured with a 10 Gb/s server.

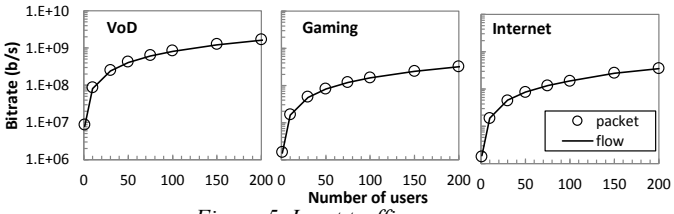

Figure 5. Input traffic vs. users.

Table 2. Relative errors of aggregated traffic flows.

	VoD		Gaming		Internet	
users	*mean*	*max*	*mean*	*max*	*mean*	*max*
10	6%	57%	4%	14%	4%	15%
50	5%	34%	2%	5%	3%	4%
100	4%	15%	2%	2%	2%	3%
200	4%	10%	1%	1%	2%	2%

Fig. 5 shows the average bitrate of the traffic flows of each consumer group against the number of users, using flow-based and packet-based generation. As shown, flow-based generation accurately matches the correlation between generated bitrate and number of users that packet-based generation produced. A detailed accuracy analysis is presented in Table II, where mean and maximum errors of flow-based generation w.r.t. packet-based generation are detailed for every service and different number of users. Mean errors are not higher that 6%, whereas maximum error remarkably decreases with the number of users, reaching no more that 15% in the worst case (for the VoD service) when 100 or more users are considered. Note that gaming and Internet services experience maximum errors not higher that 15% even with 10 users. In light of these results, the accuracy of the proposed statistical methodology to generate aggregated input flows is validated assuming scenarios with a medium/high number of consumers per group.

5. CONCLUSIONS

The CURSA-SQ methodology has been proposed to generate accurate synthetic traffic flows based on service characteristics and consumers behavior, and to analyze its impact on the network infrastructure. Input traffic flow modelling was statistically formulated aiming at producing traffic models of flows aggregating a number of consumers, where second and sub-second granularities were considered.

ACKNOWLEDGEMENTS

This work was partially supported by the EC through the METRO-HAUL project (G.A. nº 761727), from the AEI/FEDER TWINS project (TEC2017-90097-R), and from the Catalan ICREA Institution.

REFERENCES

[1] L. Huang *et al*., "Analysis of user behavior in a large-scale VoD system," in *Proc. NOSSDAV*, 2017.
[2] Y. Choi, J. Silvester, and H. Kim, "Analyzing and modeling workload characteristics in a multiservice IP network," *IEEE Internet Computing*, vol. 15, pp. 35-42, 2011.
[3] W. Feng *et al*., "A traffic characterization of popular on-line games," *IEEE/ACM Transactions on Networking*, vol. 13, pp. 488-500, 2005.
[4] D. Drajic *et al*., "Traffic generation application for simulating online games and M2M applications via wireless networks," in *Proc. WONS*, pp. 167-174, 2012.
[5] X. Wu *et al*., "Packet size distribution of typical Internet applications," in *Proc. ICWAMTIP*, pp. 276-281, 2012.