

Un modèle de trafic adapté à la volatilité de charge d'un service de vidéo à la demande: Identification, validation et application à la gestion dynamique de ressources.

Shubhabrata Roy, Thomas Begin, Patrick Loiseau, Paulo Gonçalves

▶ **To cite this version:**

Shubhabrata Roy, Thomas Begin, Patrick Loiseau, Paulo Gonçalves. Un modèle de trafic adapté à la volatilité de charge d'un service de vidéo à la demande: Identification, validation et application à la gestion dynamique de ressources.. [Research Report] RR-8072, INRIA. 2012. hal-00734295v2

HAL Id: hal-00734295

https://inria.hal.science/hal-00734295v2

Submitted on 1 Oct 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

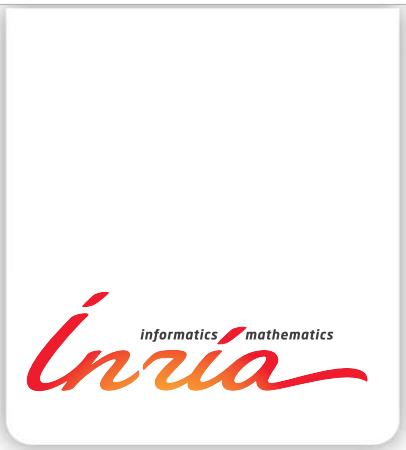


Un modèle de trafic adapté à la volatilité de charge d'un service de vidéo à la demande: Identification, validation et application à la gestion dynamique de ressources.

Shubhabrata Roy, Thomas Begin, Patrick Loiseau, Paulo Gonçalves

RESEARCH REPORT
N° 8072
September 2012
Project-Team RESO

ISSN 0249-6399 ISRN INRIA/RR--8072--FR+ENG



Un modèle de trafic adapté à la volatilité de charge d'un service de vidéo à la demande: Identification, validation et application à la gestion dynamique de ressources.

Shubhabrata Roy, Thomas Begin, Patrick Loiseau, Paulo Gonçalves

Équipe-Projet RESO

Rapport de recherche n° 8072 — September 2012 — 27 pages

Work described in this report has been supported by the EU FP7 project SAIL

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Résumé : La gestion dynamique de ressources est un élément clé du paradigme de *cloud computing* et plus récemment de celui de cloud networking. Dans ce contexte d'infrastructures virtualisées, la réduction des coûts associés à l'utilisation et à la ré-allocation des ressources contraint les opérateurs et les utilisateurs de clouds à une gestion rationnelle de celles-ci. Dans ce travail nous proposons une description probabiliste des besoins liée à la volatilité de la charge d'un service de distribution de vidéos à la demande. Cette description peut alors servir de consigne (input) à la provision et à l'allocation dynamique des ressources nécessaires. Notre approche repose sur la construction d'un modèle stochastique inspiré des modèles de Markov standards de propagation épidémiologique, capable de reproduire des variations soudaines et intenses d'activité (*buzz*). Nous proposons alors une procédure heuristique d'identification du modèle à partir de séries temporelles du nombre d'utilisateurs connectés au serveur. Les performances d'estimation de chacun des paramètres du modèle sont évaluées numériquement, et nous vérifions l'adéquation du modèle aux données en comparant les distributions des états stationnaires ainsi que les fonctions d'autocorrélation des processus.
Les propriétés markoviennes de notre modèle garantissent qu'il vérifie un principe de grandes déviations permettant de caractériser statistiquement l'ampleur et la durée d'évènements extrêmes et rares tels que ceux produits par les *buzzs*. C'est cette propriété que nous exploitons pour dimensionner le volume de ressources (e.g. bande-passante, nombre de serveurs, taille de buffers) à prévoir pour réaliser un bon compromis entre coût de re-déploiement des infrastructures et qualité de service. Cette approche probabiliste de la gestion des ressources ouvre des perspectives sur les politiques de *Service Level Agreement* adaptées aux *clouds* et servant au mieux les intérêts des opérateurs de réseaux, de services et de leurs clients.

Mots-clés : Réseaux, Cloud, Gestion probabiliste des Ressources, Modèles Epidémiques , Générateur de Charge, Estimation Statistique, Principe de Grandes Déviations, *Service Level Agreement*, Vidéo à la Demande, Buzz

A Versatile Model for VoD Buzz Workload: Identification, Numerical Validation and Applications in Dynamic Resource Management

Abstract: Dynamic resource management has become an active area of research in the Cloud Computing paradigm. Cost of resources varies significantly depending on configuration for using them. Hence efficient management of resources is of prime interest to both Cloud Providers and Cloud Users. In this report we suggest a probabilistic resource provisioning approach that can be exploited as the input of a dynamic resource management scheme. Using a Video on Demand use case to justify our claims, we propose an analytical model inspired from standard models developed for epidemiology spreading, to represent sudden and intense workload variations. As an essential step we also derive a heuristic identification procedure to calibrate all the model parameters and evaluate the performance of our estimator on synthetic time series. We show how good can our model fit to real workload traces with respect to the stationary case in terms of steady-state probability and autocorrelation structure. We find that the resulting model verifies a Large Deviation Principle that statistically characterizes extreme rare events, such as the ones produced by “buzz effects" that may cause workload overflow in the VoD context.
This analysis provides valuable insight on expectable abnormal behaviors of systems. We exploit the information obtained using the Large Deviation Principle for the proposed Video on Demand use-case for defining policies (Service Level Agreements). We believe these policies for elastic resource provisioning and usage may be of some interest to all stakeholders in the emerging context of cloud networking.

Key-words: Cloud Networking,Probabilistic Resource Management, Epidemic Model, Workload Generator, Statistical Estimation, Large Deviation Principle, Service Level Agreements, Video on Demand, Buzz

Table des matières

1 Introduction **4**

2 Related Work **7**

3 A VoD system and its modeling **8**
3.1 Markov Model to describe the VoD user behavior 9

4 Estimation procedure **11**
4.1 Parameters estimation 11
4.2 Numerical Validation 15

5 Validation of the estimation procedure against a real workload trace **17**

6 Large Deviation Principle and its interpretation **19**
6.1 Numerical Interpretations 20

7 Resource management policies **22**
7.1 Identification of the reactive time scale for reconfiguration 22
7.2 Link capacity dimensioning 22

8 Conclusion **24**

1 Introduction

In recent trend of data-intensive applications with pay-as-you-go execution in a cloud environment, there are new challenges in system management and design to optimize the resource utilization. Types of the application, deployed in a cloud, can be very diverse. Some applications exhibit highly varying demand in resources. In this paper we consider a Video on Demand (VoD) system as a relevant example of a data-intensive application where bandwidth usage varies rapidly over time.

A VoD service delivers video contents to consumers on request. According to Internet usage trends, users are increasingly getting more involved in the VoD and this enthusiasm is likely to grow. According to 2010 statistics a popular VoD provider like Netflix accounts for around 30 percent of the peak downstream traffic in the North America and is the “largest source of Internet traffic overall" [1]. Since VoD has stringent streaming rate requirements, each VoD provider needs to reserve a sufficient amount of server outgoing bandwidth to sustain continuous media delivery (we are not considering IP multicast here). However, resource reservation is very challenging in a situation, when a video becomes popular very quickly leading to a *flood* of user requests on the VoD servers. This situation, also known as a “buzz", demands an adaptive resource allocation strategy to cope with the sudden (and significant) variation of workload. Following is one example of “buzz" (see Figure 1) where interest over a video “Star Wars Kid" [2] grew very quickly within a short timespan. According to [3] it was viewed more than 900 millions times within a short interval of time making it one of the top viral videos. Such bandwidth volatility creates significant challenges to meet, namely, both the desired QoS and efficient resource allocation. A sensible approach to this problem is to help the providers in better understanding and capturing the underlying characteristics of their applications. For example, if the information diffusion process follows a gossip- (or epidemic-) behavior, the rate at which a viewer keeps gossiping about a video and for how long (in average).

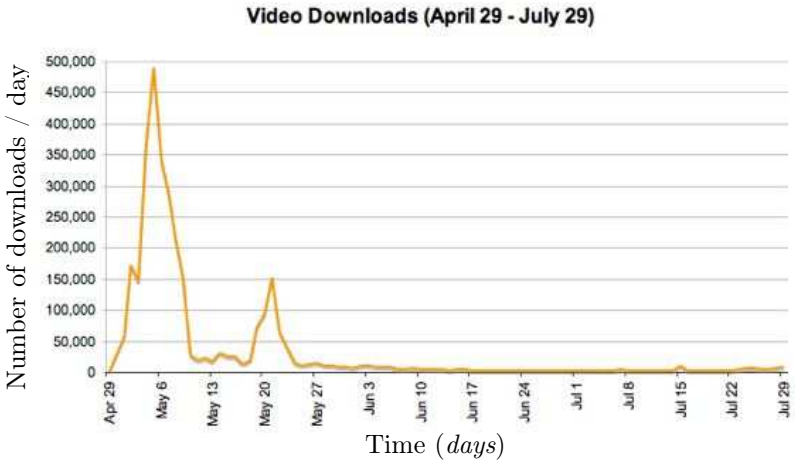


FIGURE 1 – Video server workload : time series displaying a characteristic pattern of flash crowd (buzz effect). Trace obtained from [2].

In this report we follow a constructive approach to propose a stochastic epidemic workload generator for a VoD system based on a Markov model. We show that it succeeds to reproduce the traffic volatility, as exhibited in a real trace, including the buzz occurrence. But the principal interest of our model is that it verifies a Large Deviation Principle (LDP) that gives a probabilistic description of the mean workload of the system over different time scales. It thus adequately allows for statistically characterizing extreme rare events such as the ones produced by buzz transients. Our ultimate objective is to exploit this large deviation information as an input of an probabilistic resource management scheme. However, in order the proposed model to conform with this objective, it needs to be "identifiable" and easily calibrated on real data. The corresponding estimation procedure may not be trivial, since the VoD model is a non-parsimonious model and accounts for complex dynamics. In this report we propose a complete framework for the operators to identify the VoD model parameters based on a server workload trace
After parameter estimation we devise two possible and generic ways to exploit the large deviation information in the context of probabilistic resource provisioning. They can serve as the input of resource management functionalities of the Cloud environment. It is evident that we can not define elasticity without the notion of a time scale ; the Large Deviation Principle (LDP) is capable of automatically integrating the time resolution in automatic description of the system. It is to be noted that Markovian processes do satisfy the LDP, but so do some other models as well. Hence, our proposed probabilistic approach is very generic and can adapt to address any provisioning issues, provided the resource volatility can be resiliently represented by a stochastic process for which the LDP holds true.
In a nutshell our contributions in this report include :

- A Markov based versatile model to generate VoD workload,
- A heuristic identification procedure for the proposed workload model,
- A numerical evaluation of the estimator for each parameter of the model,
- A real case study to assess the adequacy of our model to fit video workload traces,
- An analysis of the Large Deviation property of the proposed Markovian model,
- A discussion on the generic ways to exploit the large deviation information in the context of probabilistic resource provisioning.

Moreover, since we followed a constructive approach, each parameter of the model accounts for a specific component of the system, and so, its estimated value also permits to quantify the importance of the corresponding dynamic effect.
Rest of the paper is organized as follows. In Section 2 we discuss the related works. We describe our model and further analyze it in Section 3. Section 4 outlines the parameter estimation procedure and validates the procedure against synthetic workload traces. In Section 5 we validate both our model and the estimation procedure against the real workload traces. Section 6 presents Large Deviation Principle and numerical interpretations of the Large Deviation Spectrum. Section 7 deals with the probabilistic provisioning scheme, derived from the Large Deviation Spectrum for our use case. Finally we conclude and discuss future works in Section 8.

2 Related Work

Information dissemination in a VoD system has been an active area of research. In [4], it has been already demonstrated that the epidemic algorithms can be used as an effective solution for information dissemination in a VoD like P2P systems. However, in this model an individual process must have a precise idea about the total number of processes in the system. Scalability is also another challenge that the authors addressed in this work. The authors of [5] studied random epidemic strategies like the random peer, latest useful chunk algorithm to achieve optimal information dissemination. But main objective of this work is to demonstrate ways to achieve performance trade-offs using unstructured, epidemic live streaming systems. However, it does not bring any information about the underlying dynamics of the streaming system. Authors of [6] similarly discussed an analytical framework for gossip protocols based on the pairwise information exchange between interacting nodes. However, this model only provides an analysis of networks with lossy channels. Another relevant work to our study is derived in [7] where the authors proposed an approach to predict workload for cloud clients. They considered an auto-scaling approach for resource provisioning and validated the result with real-world cloud client application traces. However, this work depends on similar past occurrences of the current short-term workload history and is not appropriate to deal with sudden and short large variations of workload, as the ones produced by buzz effects. Authors of [8] show a statistical study of streaming traffic. They analyzed VoD session characteristics, amount and types of media delivered, popularity and daily access profile in order to develop a workload generator. However, the model does not involve the dynamics of the process itself, ergo it is not naturally adapted to infer dynamic resource allocation strategies. Authors of [9], [10] and [11] also develop user activity models to describe the usage of system resources. Limitation of these models are that they only give average results. However, dealing with mean workloads might not be sufficient to clearly describe applications because of their potential volatility. In [12] authors proposed a maximum likelihood method for fitting a Markov arrival process (MAP) to the web traffic measurements, collected in commonly available HTTP web server traces. This method achieves reasonable accuracy in predictive models for web workloads but lacks intuitive nature to describe users behavior like a gossip based method. In [13] the authors statistically model traffic volatility in large scale VoD systems using GARCH (generalized autoregressive conditional heteroscedasticity) process. Amazon Cloud-Watch follows this approach and provides a free resource monitoring service to Amazon Web Service customers for a given frequency. Based on such estimates of future demand, each VoD provider can individually reserve a sufficient amount of bandwidth to satisfy *in average* its random future demand within a reasonable confidence. However, according to the authors, this technique only models and forecasts the mean demand, or the expected demand whereas the real demand might vary around this predicted mean. They suggested to provision an additional “risk premium" to the service providers for tolerating the demand fluctuation. In another workload model the authors of [14] [15] proposed a Markov Modulated Poisson Process (MMPP) based approach for buzz modeling and then parameter estimation using the index of dispersion. However, the MMPP model includes only short-term memory in the system and the obtained statistics is not physically interpretable to draw inference about the system dynamics. The model we derive in section 3 of this report has the following advantages :

- It follows a constructive approach, based on a Markov model,
- It is identifiable and succeeds to capture workload volatility,
- It satisfies the large deviation properties, that can be exploited to frame dynamic resource allocation strategies.

3 A VoD system and its modeling

A VoD service delivers video contents to consumers on request. According to Internet usage trends, users are increasingly getting more involved in the VoD and this enthusiasm is likely to grow. A popular VoD provider like Netflix accounts for around 30 percent of the peak downstream traffic in the North America and is the "largest source of Internet traffic overall" [1]. In a VoD system, consumers are video clients who are connected to a *Network Provider*. The source video content is managed and distributed by a *Service Provider* from a central data centre. With the evolution of Cloud Computing and Networking, the service in a VoD system can be made more scalable by dynamically distributing the caching/transcoding servers across the network providers. Video service providers interact with the network service providers and describe the virtual infrastructures required to implement the service (like the number of servers required, their placements and clustering of resources). The resource provider reserves resource for certain time period and may change it dynamically depending on resource requirement. Such a dynamic approach brings benefits of cost saving in the system through dynamic resource provisioning which is important for service providers as VoD workload is highly variable by nature. However, since the virtual resources used by Cloud Networking have a set-up time which is not negligible, analysis and provisioning of such a system can be very critical from the operators perspective (CAPEX versus OPEX trade-off). Figure 2 shows a VoD schematic where the back-end server is connected to the data centre and the transcoding (caching) servers are placed across the network providers. Since VoD has stringent streaming rate requirements, each VoD provider needs to

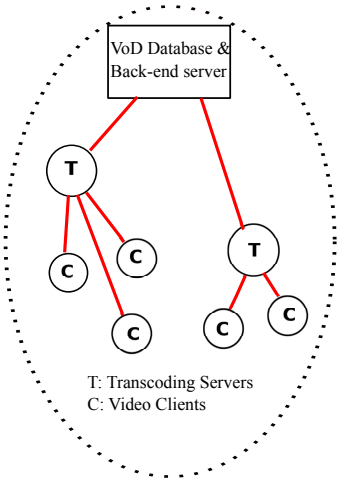


FIGURE 2 – Basic schematics of a VoD system with transcoding/caching servers

reserve a sufficient amount of server outgoing bandwidth to sustain continuous media delivery. When multiple VoD providers (such as Netflix) are on board to use cloud services from cloud providers, there will be a market between VoD providers and cloud providers, and commodities to be traded in such a market consist of bandwidth reservations, so that VoD streaming performance can be guaranteed.

As a buyer in such a market, each VoD provider can periodically make reservations for bandwidth capacity to satisfy its random future demand. A simple way to achieve this is to estimate expectation and variance of its future demand using historical demand information, which can easily be obtained from cloud monitoring services. As an example, Amazon Cloud-Watch provides a free resource monitoring service to Amazon Web Service customers for a

given frequency. Based on such estimates of future demand, each VoD provider can individually reserve a sufficient amount of bandwidth to satisfy *in average* its random future demand within a reasonable confidence. However, this information is not helpful in case of a “buzz" or a “flash crowd" when a video becomes popular very quickly leading to a *flood* of user requests on the VoD servers. In situations like the one described in Figure 1, variance estimation or more generally steady state distribution can not explain burstiness of such event as time resolution is excluded from the description. The LDP, by virtue of its multi-resolution extension of the classical steady-state distribution, can describe the dynamics of rare events like this, which we believe can be of some interest for the VoD service providers.

3.1 Markov Model to describe the VoD user behavior

Epidemic models commonly subdivide a population into several compartments : susceptible (noted S) to designate the persons who can get infected, and contagious (noted C) for the persons who have contracted the disease. This contagious class can further be categorized into two parts : the infected subclass (I) corresponding to the persons who are currently suffering from the disease and can spread it, and the recovered class (R) for those who got cured and do not spread the disease anymore [16]. There can be more categories that fall outside the scope of our current work. In these models $(N_S(t))_{t\geq 0}$, $(N_I(t))_{t\geq 0}$ and $(N_R(t))_{t\geq 0}$ are stochastic processes representing the time evolution of susceptible, infected and recovered populations respectively. Similarly, information dissemination in a social network can be viewed as an epidemic spreading (through gossip), where the “buzz" is a special event where interest for some particular information increases drastically within a very short period of time. Following the lines of related works, we claim that the above mentioned epidemic models can appropriately be adapted to represent the way information spreads among the users in a VoD system. In the case of a VoD system, infected I refers to the people who are currently watching the video and can spread the information about it. In our setting, I directly represents the current workload which is the current aggregated video requests from the users. Here, we consider the workload as the total number of current viewers, but it can also refer to total bandwidth requested at the moment. The class R refers to the past viewers. In contrast to the classical epidemic case, we introduce a memory effect in our model, assuming that the R compartment can still propagate the gossip during a certain random latency period. Then, we define the probability within a small time interval $\mathrm{d}t$, for a susceptible individual to turn into an active viewer, as follows :

$$\mathbb{P}_{S\to C} = (l + (N_I(t) + N_R(t))\,\beta)\mathrm{d}t + o(\mathrm{d}t) \tag{1}$$

where $\beta > 0$ is the rate of information dissemination per unit time and $l > 0$ fixes the rate of spontaneous viewers. The instantaneous rate of newly active viewers in the system at time t is thus :

$$\lambda(t) = l + (N_I(t) + N_R(t))\beta. \tag{2}$$

Equation (2) corresponds to the arrival rate $\lambda(t)$ of a non-homogeneous (state dependant) Poisson process. This rate varies linearly with $N_I(t)$ and $N_R(t)$.

To complete our model we assume that the watch time of a video is exponentially distributed with rate γ. As already mentioned, it also deems reasonable to consider that a past viewer will not keep propagating the gossip about a video indefinitely, but remains active only for a latency random period that we also assume exponentially distributed with rate μ (in general $\mu \ll \gamma$). Another important consideration of the model is the maximum allowable viewers ($I_{\max}$) at any instant of time. This assumption conforms to the fact that the resources in the system are physically limited. For the sake of numerical tractability and without loss of generality, we also

assume the number of past (but spreading rumour) viewers at a given instant to be bounded by a maximum value ($R_{\max}$). With these assumptions, and posing $(N_I(t) = i, N_R(t) = r)$ the current state of the Markov processes, the probability that the process reaches a different state $(i' < I_{\max}, r' < R_{\max})$ at time $t + \mathrm{d}t$ ($\mathrm{d}t$ being small) reads :

$$\begin{aligned} &\mathbb{P}(i', r'|i, r) \qquad (3)\\ &= (l + (i + r)\beta)\mathrm{d}t + o(\mathrm{d}t) \quad \text{for } (i' = i + 1, r' = r),\ ^{1}\\ &= (\gamma i)\mathrm{d}t + o(\mathrm{d}t) \quad \text{for } (r' = r + 1, i' = i - 1),\\ &= (\mu r)\mathrm{d}t + o(\mathrm{d}t) \quad \text{for } (r' = r - 1, i' = i),\\ &= o(\mathrm{d}t) \quad \text{otherwise.} \end{aligned}$$

This process defining the evolution of the current viewer and past viewer populations is a finite and irreducible Markov chain. It is to be noted that $l > 0$ precludes the process to reach an absorbing state. This chain is ergodic and admits a stationary regime.
Above mentioned descriptions define the mechanism of information dissemination in the community in normal situations. A buzz event differs from this situation by a sudden increase of the dissemination rate β. In order to adapt the model to buzz we resort to Hidden Markov Model (HMM) to be able to reproduce the change in β. Without loss of generality we consider only two states. One with dissemination rate $\beta = \beta_1$ corresponds to the buzz-free case described above, and another hidden state corresponding to the buzz situation, where the value of β increases significantly and takes on a value $\beta_2 \gg \beta_1$. Transitions between these two hidden and memoryless Markov states occur with rates a_1 and a_2 respectively (see Figure 3). These rates characterize the buzz in terms of frequency, magnitude and duration.

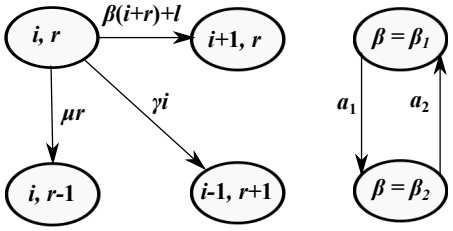


FIGURE 3 – Markov chain diagram representing the evolution of the Current viewers (i) and Past Viewers (r) populations with a Hidden Markov Model.

A closed-form expression for the steady state distribution of the workload (i) of this model is not trivial to derive. However, we could easily express the analytic mean workload of the system solving the flow balance equation, i.e. equaling the incoming and outgoing flow rates in steady regime. For ease, we start with $\beta = \beta_1 = \beta_2$ and generalize the result to $\beta_1 \neq \beta_2$ thereafter. We get :

$$\mathbb{E}(i) = \frac{\mu l}{\mu\gamma - \mu\beta - \gamma\beta}, \qquad (4)$$

which, to be a positive and finite quantity, yields the stability criterion in buzz-free regime :

$$\beta^{-1} > \mu^{-1} + \gamma^{-1}. \qquad (5)$$

1. In a closed system, where the total number of viewers (susceptible, current and past) is constant, say N, the transition probability for $(i' = i+1, r' = r)$ needs to be modified, since it would then depend on the number of susceptible viewers as well,, i.e $(N - i - r)$. The transition probability in this case would be $(l + (i + r)\frac{\beta}{N})(N - i - r)\mathrm{d}t + o(\mathrm{d}t)$. Therefore, Eq. 4 and 5 need to be modified accordingly.

We now extend these results to the case where the model may exhibit a buzz activity. As β alternates between the hidden states $\beta = \beta_1$ and $\beta = \beta_2$, with respective state probabilities $a_2/(a_1 + a_2)$ and $a_1/(a_1 + a_2)$, one can simply replace β in Eq. (4) and (5) with the equivalent average value :

$$\bar{\beta} = \frac{\beta_1 a_2}{a_1 + a_2} + \frac{\beta_2 a_1}{a_1 + a_2}. \tag{6}$$

In order to illustrate the flexibility of our workload model and to validate Eq. (4), we generate three synthetic traces corresponding to the different sets of parameters verifying the stability condition of relation (5) and reported in Table 1. Particular realizations of these processes generated over 2^{21} points are displayed in Figure 4. While the synthetic traces corresponding to

TABLE 1 – Parameters value used in the workload model to generate the three traces plotted in Fig. 4. The last two rows correspond to the theoretical mean workload of Eq. (4) and to the sample mean value estimated from the traces.

	case (a)	case (b)	case (c)
β_1	4.762×10^{-4}	3.225×10^{-5}	2.439×10^{-5}
β_2	0.0032	0.0032	0.0032
γ	0.0111	0.0020	0.0011
μ	5×10^{-4}	3.289×10^{-5}	2.5×10^{-5}
l	10^{-4}	10^{-4}	10^{-4}
a_1	10^{-7}	10^{-7}	10^{-7}
a_2	0.0667	0.0667	0.0667
$\mathbb{E}(i)$	1.92	15.68	44.72
Emp. mean $\langle i \rangle$	1.74	16.72	45.23

cases (b) and (c) reproduce distinct and easily identifiable buzz regimes, the parameter set of case (a) leads to a workload variation distinct from the typical shape of Figure 1. Nonetheless, for all 3 configurations, the empirical means estimated from the 2^{21} samples of the traces are in good agreement with the expected values of Eq. (4).
Finally, let us notice that even though we consider exponentially distributed random variables in our model, any other distributions could be used, which, according to the same balance principle, would lead to a mean workload and to a stability condition of the same kind as (4) and (5). However, the estimation procedure we derive in the next section strongly relies on the exponential assumption and it would need to be thoroughly reworked to adapt to different hypotheses.

4 Estimation procedure

In this section, we address the identifiability of our model and design a calibration algorithm to fit workload data. We start constructing empirical estimators for each parameter of the model and we numerically evaluate their performance on synthetic traces.

4.1 Parameters estimation

Considering a standard epidemic process X with propagation rate θ, the maximum likelihood estimate $\widehat{\theta}_{\text{MLE}}$ is derived in [16], [17] and reads :

$$\widehat{\theta}_{\text{MLE}} = n \cdot \left(\int_0^T X(t)\, \mathrm{d}t \right)^{-1}, \tag{7}$$

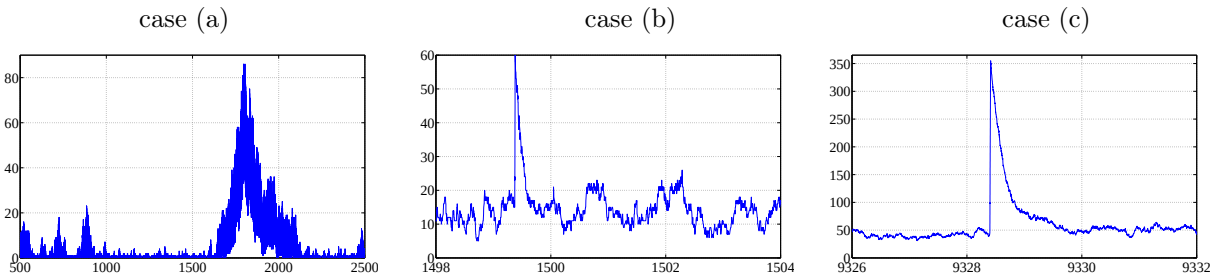


FIGURE 4 – Illustration of our model ability at generating different dynamics of workload $I(t)$. See Table 1 for the parameter values corresponding to each of these three cases. The $X-$axis corresponds to time (in hours unit) while the $Y-$axis indicates the number of active viewers.

where n is the number of contaminations (i.e. number of increments of X) occurring within the time interval T.

Very often, maximum likelihood approach yields optimal results (in terms of estimate variance and or bias) but it is not always possible to get a closed-form expression for the estimated parameters. This can either be due to the likelihood function that is impossible to derive analytically, or to missing data that preclude straightforward application of the maximum likelihood principle. Nonetheless, solutions, such as the Expectation-Maximization (EM) or the Monte Carlo Markov Chain (MCMC) algorithms exist, which in some cases can approximate maximum likelihood estimators.

Returning to our model depicted in Figure 3, each parameter needs to be empirically estimated, assuming that the instantaneous workload time series is the only available observation.

Watching parameter γ. As γ is the departure rate of users that leave the infected state after they finished watching a video, it can directly be inferred from the number n of decrements of the observable process $I(t)$. Therefore, the MLE of Eq. (7) straightforwardly applies and leads to :

$$\widehat{\gamma}_{\text{MLE}} = n \cdot \left(\int_0^T I(t)\, \mathrm{d}t \right)^{-1}. \tag{8}$$

Memory parameter μ. This rate at which past viewers leave the recovery compartment and stop propagating the virus (gossip), relates to the decrement density of the non-observed process $R(t)$. It is thus impossible to simply apply the MLE of Eq. (7) unless we first construct a substitute $\widehat{R}(t)$ to the missing data from the observable data set $I(t)$. Let us recall that in our model, all current viewers turn and remain contagious for a mean period of time $\gamma^{-1} + \mu^{-1}$. Then, in first approximation, we can consider that $R(t)$ derives from the finite memory cumulative process :

$$\widehat{R}(t) = \int_{t-(\gamma^{-1}+\mu^{-1})}^{t} I(u)\, \mathrm{d}u, \tag{9}$$

which itself, depends on the parameter to be estimated μ. We propose an estimation procedure based on the inherent exponential property of the model. From the Poisson assumption, the inter-arrival time $\mathbf{w}$ between the consecutive arrivals of two new viewers is an exponentially distributed random variable such that $\mathbb{E}\left(\mathbf{w} | \, I(t) + R(t) = x\right) = (\beta\, x + l)^{-1}$. Its means that, for x fixed, the normalized random variable $\widetilde{\mathbf{w}} = \mathbf{w}/\mathbb{E}(\mathbf{w}|x)$ is exponentially distributed with unitary parameter and becomes independent of x. Ideally then, for each value of $R(t) + I(t) = x$, all the

sub-series $\mathrm{w}_x = \{w_n \ : \ R(t_n)+I(t_n) = x\}$, after normalization by their own empirical mean, yield independent and identically distributed realizations of a unitary exponential random variable. In practice though, as $R(t)$ is not observable, only if $\widehat{R}(t)$ is accurately estimated, should this unitary exponential i.i.d. assumption hold true. From there, we propose the following algorithm : for different values of μ spanning a *reasonable* interval, we use $\widehat{R}_\mu(t)$ estimated from Eq. (9) to build the normalized series $\widetilde{\mathrm{w}}_\mu$. A statistical test applied to each $\widetilde{\mathrm{w}}_\mu$ allows for assessing the exponential i.i.d. hypothesis and then to select the value of μ that yield the best score.
More concretely, we apply to $\widetilde{\mathrm{w}}_\mu = (\widetilde{w}_n)_{n=1,\dots,N}$ the statistical exponentially test derived in [18] : Form the *normalized spacings* $\mathrm{v}_\mu = \left(v_{(n)} = (N-n+1)(\widetilde{w}_{(n)} - \widetilde{w}_{(n-1)})\right)_{n=1,\dots,N}$ where $\left(\widetilde{w}_{(n)}\right)_{n=1,\dots,N}$ stands for $\widetilde{\mathrm{w}}_\mu$ rearranged in ascending order. Let F and G denote the cumulative distribution functions of $\widetilde{\mathrm{w}}_\mu$ and v_μ respectively, and compute the classical Kolmogorov-Smirnov distance :

$$T_\mu = \sqrt{\frac{1}{N}} \sup_{1\le k\le N} |F(k) - G(k)|. \tag{10}$$

As F and G are identical for an exponentially i.i.d. random series, we then expect T_μ to reach its minimum for the value of μ that gives the best estimate $\widehat{R}_\mu(t)$ of $R(t)$:

$$\begin{cases} \widehat{\mu} = \mathrm{argmin}_\mu T_\mu \\ \widehat{R} = \widehat{R}_{\widehat{\mu}}. \end{cases} \tag{11}$$

Plots of Figure 5 show the evolution of the Kolmogorov-Smirnov distance corresponding to the traces displayed in Figure 4. In the 3 cases, T_μ clearly attains its minimum bound for $\widehat{\mu}$ close to to the actual value. The corresponding estimated processes $\widehat{R}(t)$ derived from Eq. (11) match fairly well the real evolution of the (**R**) class in our model (see Figure 6).

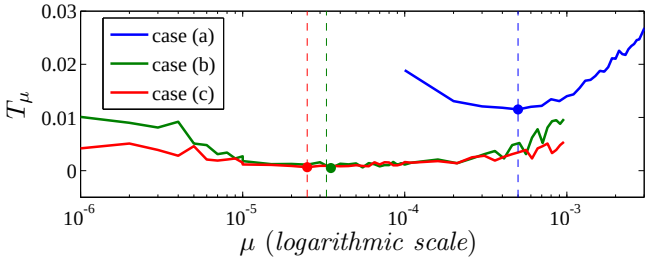


FIGURE 5 – Evolution of the exponential test statistics (10) applied to the traces of Figure 4. Dotted vertical lines locate the actual value of μ for each case ; dot markers on each curve indicate the estimated value $\widehat{\mu}$ corresponding to the minimum point of the statistical test T_μ.

Propagation parameters β and l. According to our model, the arrival rate $\lambda(t)$ of new viewers is given by Eq. (2). It linearly depends on the current number of active and past viewers. So, from the observation $I(t)$ and the reconstructed process $\widehat{R}(t)$ of Eq. (11), we could formally apply the maximum likelihood Eq. (7) to estimate β. In practice however, we have to bear in mind that : *(i)* the arrival process of rate $\lambda(t)$ comprises a spontaneous viewers ingress that is governed by parameter l and which is independent of the current state of the system ; *(ii)* depending on the current hidden state of the model (buzz-free *versus* buzz state), it is alternately $\beta = \beta_1$ and $\beta = \beta_2$ that fix the propagation rate in Eq. (2). We designed an estimation procedure based on a weighted linear regression, that simultaneously addresses these two issues. We decompose our

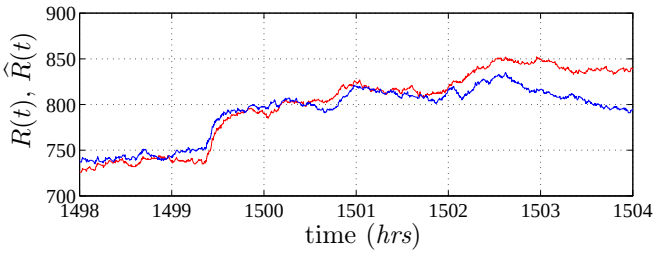


FIGURE 6 – Evolution of the number of active past viewers. Comparison of the actual (non observable) process $R(t)$ (blue curve) with the estimated process $\widehat{R}(t)$ (red curve) derived from expression (9).

rationale in two steps : First, let us consider the buzz-free state only and $\beta = \beta_1$. As discussed in the estimation of μ the inter-arrival time $\mathbf{w}$ between the consecutive arrivals of two new viewers is an exponentially distributed random variable such that $\mathbb{E}(\mathbf{w}|\, I(t) + R(t) = x) = (\beta\, x + l)^{-1}$. Concretely then, for different values of the sum $I(t) + \widehat{R}(t)$, we calculate the conditional empirical mean :

$$\Omega(x) = \frac{1}{|\mathcal{I}(x)|} \sum_{t_n \in \mathcal{I}(x)} w_n \;:\; \mathcal{I}(x) = \{t_n : I(t_n) + \widehat{R}(t_n) = x\}. \tag{12}$$

The linear regression of $(\Omega(x))^{-1}$ against x yields at one go, both parameters estimation $\widehat{\beta}$ (slope) and $\widehat{l}$ (intercept).
Let us now return to the general form of our model with alternation of buzz and buzz-free periods. In the buzz-free case, $\beta = \beta_1$ corresponds to a normal workload activity, meaning that the sum $I(t) + \widehat{R}(t)$ takes on rather moderate values. Conversely, when the system undergoes a buzz, $\beta = \beta_2$ and the population $I(t) + \widehat{R}(t)$ suddenly increases to reach significantly larger values. Yet, in both cases, the quantity Ω^{-1} defined in Eq. (12) remains linear with x but with two different regimes (slopes) depending on the amplitude of $I(t) + \widehat{R}(t) = x$. As a result, it is possible to reduce the bias that β_2 causes on the estimation of β_1, using a weighted linear regression of Ω^{-1} *vs* x where the weights $p(x)$ are proportional to the cardinal of the indicator sets $\mathcal{I}(x)$. Indeed, $|\mathcal{I}(x)|$ should be smaller for larger values of x because buzz episodes are expected to be less frequent than nominal activity periods. Figure 7 confirms the claim : the plots (x, Ω^{-1}) show a manifest linear trend with higher variability at x's large, meaning a fewer terms entered the sum of Eq. (12).

Formally, we can apply the exact same procedure to estimate β_2, but considering opposite weights to favor the large values of x's. However, due to the large fluctuations of $(\Omega(x))^{-1}$ in the corresponding region, the slope $\widehat{\beta_2}$ is subject to a very poor estimation variance. Instead, we propose to apply the ML estimator described in Eq. (7) on the restriction of $I(t)$ to the buzz periods only. Strictly speaking, we should consider $\widehat{R}(t)$ as well, but since a buzz event normally occurs on very small interval of time, we assume that $\widehat{R}(t)$ (resp. $R(t)$) remains constant in the meanwhile (flash crowd viewers will enter in R compartment only after the visualization time). In practice, to automatically identify the buzz periods, we threshold $I(t)$ and consider only the persistent increasing parts that remain above the threshold.

Transition rates a_1 and a_2. As we already said, at time t, the inter-arrival time $\mathbf{w}$ separating to new incomers is a random variable drawn from an exponential law of parameter $\lambda = \beta(i + r) + l$, where $I(t) + R(t) = i + r$ and β is either equal to β_1 or to β_2. We denote $f_1(\mathbf{w})$ and $f_2(\mathbf{w})$ the corresponding densities built upon the reconstructed process $\widehat{R}(t)$ and the estimated parameters

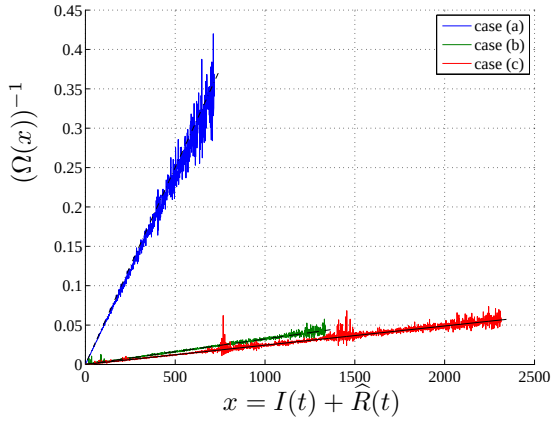


FIGURE 7 – Weighted linear regression of Ω^{-1} vs x corresponding to the three traces of Figure 4. Superimposed are the linear trends fitted on the respective data.

$(\widehat{\beta_1}, \widehat{l})$ and $(\widehat{\beta_2}, \widehat{l})$ respectively. For a given inter-arrival time $\mathbf{w} = w_n$ observed at time t_n, we form the likelihood ratio $f_2(w_n)/f_1(w_n)$ to determine whether the system is in buzz or in buzz-free state. Moreover, in order to avoid non-significant state transitions we resort to a restoration method inspired by the Viterbi algorithm [19]. Once we have identified the hidden states of the process, we estimate the transitions rates $\widehat{a_1}$ and $\widehat{a_2}$ from the average times spent in each state.

4.2 Numerical Validation

To evaluate the statistical performance of our estimation procedure, we resort to numerical experiments to empirically get the first and the second order moments of each parameter estimator. Owing to the versatility of our model, we must ensure that the proposed calibration algorithm performs well for a variety of workload dynamics. To this end, we systematically reproduce the experiments considering the 3 sets of parameters reported in Table 1. For each configuration, we generate 10 independent realizations of processes similar to the ones depicted in Figure 4, and use these to derive descriptive statistics.

The box-and-whisker plots of Figure 8 indicate for each estimated parameter (centered and normalized by the corresponding actual value) the sample median (red line), the inter-quartile range (blue box height) along with the extreme samples (whiskers) obtained from time series of length 2^{21} points each. As expected (owing to the maximum likelihood procedure), estimation of γ shows to be the most accurate, both in terms of bias and variance. But more surprisingly though, although the estimation $\widehat{\beta_1}$ derives from a heuristic procedure that itself depends on the raw approximation $\widehat{R}(t)$ of Eq. (9), the resulting performance is remarkably good : bias is always negligible (less than 5% in the worst case (c)) and the variance always confines to 10% of the actual value of β_1. Notice also that the estimation of β_1 goes from a slight underestimation in case (a) to a slight overestimation in case (c), as the buzz effect, i.e. the value of β_2, grows from traces (a) to (b). Compared to $\widehat{\beta_1}$, the estimation of β_2 behaves more poorly and proves to be the most difficult parameter to estimate. But we have to keep in mind that this latter is only based on buzz periods which represent only a small fraction of the entire time series. Regarding the parameter μ, its estimation remains within a 20% inter-quartile range but cases (a) and (c) show a systematic bias (median hits the lower quartile bound). Let us then recall that the procedure, described by Eq. (11) to determine $\widehat{\mu}$ selects within some discretized interval, the value of μ that yields the best

T_μ score. It is then very likely that the true value does not coincide with any sampled point of the interval and therefore, the procedure picks the closest one that systematically lies beneath or above. Finally, estimation of the transition parameters a_1 and a_2 between the two hidden states relies on all other parameters estimation, cumulating so all relative inaccuracies. Nonetheless and despite a systematic underestimating trend, precision remains within a very acceptable confidence interval. Convergence rate of the empirical estimators is another important feature that binds

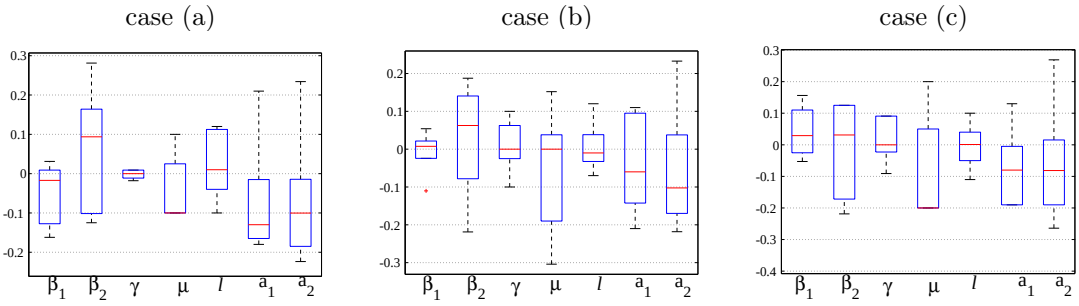


FIGURE 8 – Box-and-Whisker plots for relative estimation errors of the model parameters for the three different sets of prescribed parameters reported in Table 1. For each case (a)-(c), statistics are computed over 10 independent realizations of time series of length 2^{21} points.

the estimate precision to the amount of available data. Using the same data set, the bar plots of Figure 9 depicts the evolution of the mean square error $\mathrm{MSE}(\widehat{\theta}) = \mathbb{E}\{(\widehat{\theta} - \theta)^2\}$ – where generic θ stands for any parameter of the model – with the length N of the observable time series $I(t)$. As our purpose is to stress the rate of convergence of these quantities towards zero, to ease the comparison, we normalize the MSE of each parameter by its particular value at maximum data length (i.e 2^{21} points here). Then, the estimator rate of convergence α_θ corresponds to the decaying slope of the MSE with respect to N in a *log-log* plot, i.e. $\mathrm{MSE}(\widehat{\theta}) \sim O(N^{-\alpha_\theta})$. For the different parameters of our model we obtain convergence rates that lie between $\alpha_{\beta_1} = 0.9$ and $\alpha_{a_2} = 0.2$, leading each time to sub-optimal convergence ($\alpha_\theta < 1$). It is worth noticing that, despite its relatively *ad hoc* construction, the estimator of β_1 has an almost optimal convergence rate, which proves the rationality of our approach.

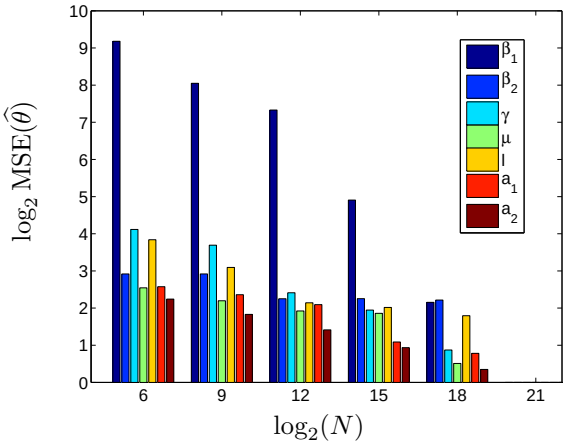


FIGURE 9 – Evolution of the Mean Square Error *versus* the data length N in a *log-log* plot. For the sake of conciseness, we only show here the results corresponding to the case (b) of Table 1.

5 Validation of the estimation procedure against a real workload trace

We now apply the calibration procedure detailed in the previous section, to fit our model on real data and to assess the data-model adequacy in the context of highly volatile workloads. As a paradigm for variable demand applications, we use a VoD trace, released by the Greek Research and Technology Network (GRNET) VoD servers [20]. Since the trace shows modest activity with a workload level that is not sufficient to stress-test our system, we scale up the data, shrinking all inter-arrival times by a factor of 10. The resulting workload time series displayed in Figure 10, clearly shows two distinct periods of steady activity before and after the time index $t = 200$. We consider the sub-series on both sides of this cutoff time, as two individual workload processes, referred to as trace I and trace II respectively, and we calibrate our model on each of them separately.

Results of the parameters estimation are reported in Table 2, and we verified that in both cases the stability condition of Eq. (5) was satisfied. In the same vein, we also compared the empirical mean workload of each trace with its corresponding theoretical value given by the formula (4). We obtain for trace I a relative difference of 12% ($\mathbb{E}(i) = 5.59$ compared to $\langle i \rangle = 4.99$), and of 12.5% for trace II ($\mathbb{E}(i) = 0.621$ compared to $\langle i \rangle = 0.71$). Naturally, the correspondence here is not as striking as it is with the synthetic traces of Section 3. But we must bear in mind that first, *ab initio* nothing guarantees that the underlying system matches our model dynamics and, second, traces I and II can possibly encompass short scale non-stationary periods (e.g. day *versus* night activity) which are not accounted for in our model. Notwithstanding this, the match we observe is quite satisfactory and we now focus on higher order statistics to further characterize the data model adequacy. As we do not have a closed-form for the steady state distribution of our Markov process model, nor we have an analytic expression for its autocorrelation function, we use the two sets of estimated parameters of Table 2 to synthesize two time series that we compare to the real workload traces I and II. We refer to those synthetic traces as to the fitted traces I and II. The plots in Figure 11 show the empirical steady state densities and the sample autocorrelation functions of both the real and the fitted traces. The superimposition of the different curves is a clear evidence of our model ability at catching the statistical distribution of the number

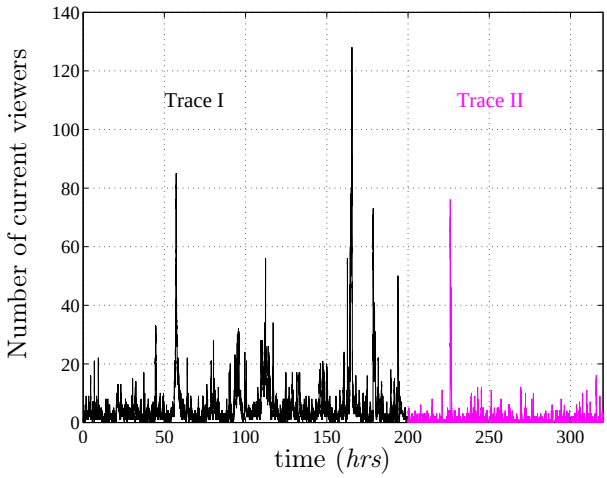


FIGURE 10 – Real workload time series corresponding to a VoD server demand from [20]. Initial trace was scaled up by a factor of 10 to increase the mean workload. Trace is chopped into two separate processes (Trace I and II) corresponding to different activity levels.

of current viewers along time. But also, and perhaps more importantly, it demonstrates that the dynamical mechanism underlying our constructive model is able to perfectly reproduce the temporal structure of the real traces, by imposing the correct statistical dependencies between distant samples $I(t)$ and $I(t+\tau)$ of the process.

In addition to serve as a mean to evaluate the goodness-of-fit of our model, the estimated parameters bring on their own, a valuable insight about the system itself. For instance let us compare the propagation rates $\widehat{\beta_1}$ and $\widehat{l}$ estimated from traces I and II, successively. In the first case, $\widehat{\beta_1} < l$, meaning that arrival of new viewers is dominated by spontaneous incomers and is not so much due to information propagation through gossip. Conversely, $\widehat{\beta_1} > \widehat{l}$ for the second workload regime, indicating that the spontaneous attraction of the server has severely dropped whereas the peer-to-peer diffusion component significantly increased but not sufficiently to sustain the mean workload activity. At the same time, the index $\widehat{\mu}$ tripled, meaning that the mean memory period for propagation shrank by a factor of 3. This parameter could then be used as an indicator of the content interest delivered by the server, and of its lifetime in users mind.

TABLE 2 – Estimated Parameters from traces I and II separately.

	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\gamma}$	$\widehat{\mu}$	$\widehat{l}$	$\widehat{a_1}$	$\widehat{a_2}$
I	0.0013	0.0084	0.0039	0.0028	0.0032	$3.13.10^{-4}$	0.022
II	0.0049	0.0183	0.0118	0.0095	0.0005	$1.32.10^{-5}$	0.041

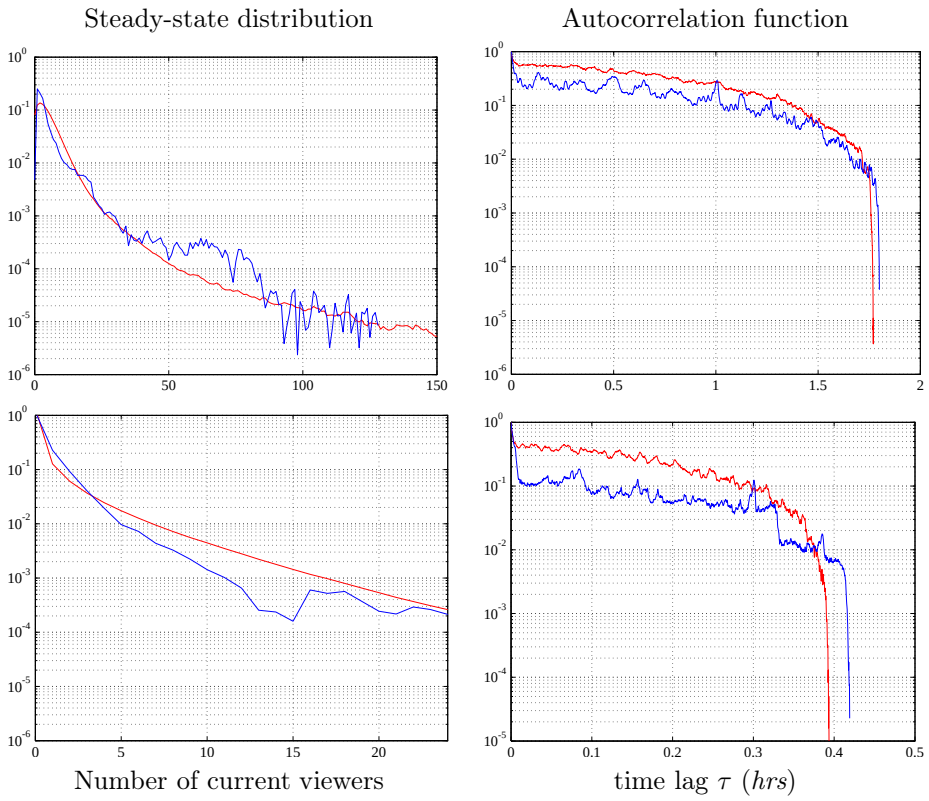


FIGURE 11 – Comparison of the empirical steady-state distribution and of the autocorrelation function of the real (blue curves) and the fitted (red curves) traces. Top two plots correspond to trace I and bottom plots correspond two trace II.

6 Large Deviation Principle and its interpretation

Consider a continuous-time Markov process $(X_t)_{t\geq 0}$, taking values in a finite state space S, of rate matrix $A = (A_{ij})_{i\in S, j\in S}$. In our case X is a vectorial process $X(t) = (N_I(t), N_R(t)), \forall t \geq 0$, and $S = \{0, \cdots, I_{\max}\} \times \{0, \cdots, R_{\max}\}$. If the rate matrix A is irreducible, then the process X admits a unique steady-state distribution π satisfying $\pi A = 0$. Moreover, by Birkhoff ergodic theorem, it is known that for any mapping $\Phi : S \to \mathbb{R}$, the sample mean of $\Phi(X)$ at scale τ, i.e. $1/\tau \cdot \int_0^\tau \Phi(X_s)\mathrm{d}s$ converges almost-surely towards the mean of $\Phi(X)$ under the steady-state distribution, as τ tends to infinity. The function Φ is often called the *observable*. In our case, as we are interested in the variations of the current number of users $N_I(t)$, Φ will simply be the function that selects the first component : $\Phi(N_I(t), N_R(t)) = N_I(t)$. The large deviations principle (LDP), which holds for irreducible Markov processes on a finite state space [21], gives a efficient way to estimate the probability for the sample mean calculated over a large period of time τ to be around a value $\alpha \in \mathbb{R}$ that deviates from the almost-sure mean :

$$\lim_{\epsilon\to 0} \lim_{\tau\to\infty} \frac{1}{\tau} \log \mathbb{P}\left\{ \int_0^\tau \Phi(X_s)\mathrm{d}s \in [\alpha - \epsilon, \alpha + \epsilon] \right\} = f(\alpha). \tag{13}$$

The mapping $\alpha \mapsto f(\alpha)$ is called the large deviations spectrum (or the rate function). For a given function Φ, it is possible to compute the theoretical large deviations spectrum from the

rate matrix A as follows. One first computes, for each values of $q \in \mathbb{R}$, the quantity $\Lambda(q)$ defined as the principal eigenvalue (*i.e.,* the largest) of the matrix with elements $A_{ij} + q\delta_{ij}\Phi(j)$ ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise). Then the large deviations spectrum can be computed as the Legendre transform of Λ :

$$f(\alpha) = \sup_{q \in \mathbb{R}} \{q\alpha - \Lambda(q)\}, \forall \alpha \in \mathbb{R}. \tag{14}$$

As described in Equation(13), $\alpha_\tau = \langle i \rangle_\tau$ corresponds in our study case, to the mean number of users i observable over a period of time of length τ and $f(\alpha)$ relates to the probability of its occurrence as follows :

$$\mathbb{P}\{\langle i \rangle_\tau \approx \alpha\} \sim e^{\tau \cdot f(\alpha)}. \tag{15}$$

Interestingly also, if the process is strictly stationary (*i.e.* the initial distribution is invariant) the same large deviation spectrum $f(\cdot)$ can be estimated from a single trace, provided that it is "long enough" [22]. We proceed as follows : At a scale τ, the trace is chopped into k_τ intervals $\{I_{j,\tau} = [(j-1)\tau, j\tau[,\ j = 1, \ldots, k_\tau\}$ of length τ and we have (almost-surely), for all $\alpha \in \mathbb{R}$:

$$\begin{gathered} f_\tau(\alpha, \epsilon_\tau) = \frac{1}{\tau} \log \frac{\#\left\{j : \int_{I_{j,\tau}} \Phi(X_s)\mathrm{d}s \in [\alpha - \epsilon_\tau, \alpha + \epsilon_\tau]\right\}}{k_\tau} \\ \text{and } \lim_{\tau \to \infty} f_\tau(\alpha, \epsilon_\tau) = f(\alpha). \end{gathered} \tag{16}$$

In practice, for the empirical estimation of the large deviations spectrum, we use a similar estimator as the one derived in [23] and also used in [24]. At scale τ, we compute for each $q \in \mathbb{R}$ the values of $\Lambda'_\tau(q)$ and $\Lambda''_\tau(q)$, where $\Lambda_\tau(q) = \tau^{-1} \log\left(k_\tau^{-1} \sum_{j=1}^{k_t} \exp\left(q \int_{I_{j,\tau}} \Phi(X_s)\mathrm{d}s\right)\right)$. Then, for each value of τ, we count the number of intervals $I_{j,\tau}$ verifying the condition in expression (16) and estimate the scale-dependant empirical *log-pdf* $f_\tau(\alpha, \epsilon_\tau)$, with the adaptive choices derived in [23] :

$$\alpha_\tau = \Lambda'_\tau(q) \quad \text{and} \quad \epsilon_\tau = \sqrt{\frac{-\Lambda''_\tau(q)}{\tau}}. \tag{17}$$

Let us now illustrate the LDP in the context of the specific VoD use case, where X would correspond to (i, r), the bi-variate Markov process. $\Phi(X)$ is i, the observable and $\int_0^\tau \Phi(X_S)\, ds = \langle i \rangle_\tau$ corresponds to the average number of users with a period τ.

6.1 Numerical Interpretations

For ease of computation we estimate the Large Deviation Spectrum for cases where $I_{\max} = 30, R_{\max} = 60$. We also choose the parameters accordingly (so that it does not saturate with the maximum value) for buzz and buzz-free scenarios. For the first case $\beta_1 = 0.1$, $\beta_2 = 0.8$, $\gamma = 0.7$, $\mu = 0.3$, $l = 1.0$, $a_1 = 0.006$ and $a_2 = 0.6$. For the buzz-free case : $\beta_1 = \beta_2 = \beta = 0.1$, $\gamma = 0.7$, $\mu = 0.3$, $l = 1.0$. Intrinsically, Large Deviation Principle naturally embeds the time scale notion into the statistical description of the aggregated observable at different time resolutions. As expected, the theoretical LD spectra displayed in Figure 12(a) reach their maximum for the same mean number of users. This apex is the almost sure value as described in Section **??**. As the name suggests almost sure workload ($\alpha_{a.s}$) corresponds to the mean value that we almost surely observe on the trace. More interestingly though, the LD spectrum corresponding to the buzz case, spans over a much larger interval of observable mean workloads than that of the buzz-free case. This remarkable support widening of the theoretical spectrum shows that LDP can accurately quantify the occurrence of extreme, yet rare events.

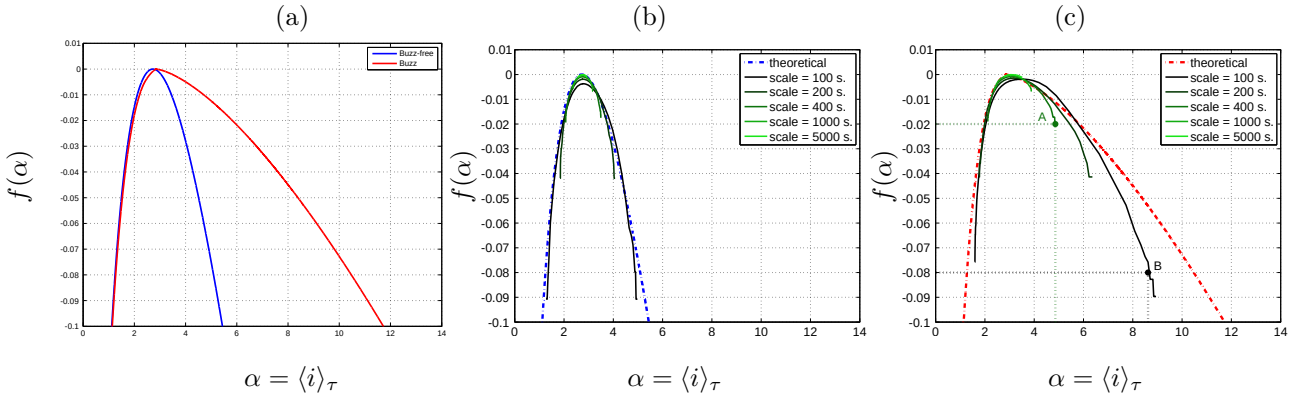


FIGURE 12 – Large Deviations spectra corresponding to the traces of Figure 4. (a) Theoretical spectra for the buzz free (blue) and for the buzz (red) scenarii. (b) & (c) Empirical estimations of $f(\alpha)$ at different scales from the buzz free and the buzz traces, respectively.

Plots (b)-(c) of Figure 12 compare theoretical and empirical large deviation spectra obtained for the two traces. For each given scale (τ) the empirical estimation procedure yields one LD estimate. These empirical estimates at different scales superimpose for a given range of α. This is reminiscent of the scale invariant property underlying the large deviation principle. If we focus on the supports of the different estimated spectra, the larger the time scale τ is, the smaller becomes the interval of observable value of α. This is coherent with the fact that for a finite trace-length the probability to observe a number of current viewers, that in average, deviates from the nominal value ($\alpha_{a.s}$) during a period of time (τ) decreases exponentially fast with τ. To fix the ideas, the estimates of plot (c), indicate that for a time scale $\tau = 400\,sec.$, the maximum observable mean number of users is around 5 with probability $2^{400\cdot(-0.02)} \approx 35.10^{-5}$ (point A), while it increases up to 9 with the same probability ($2^{100\cdot(-0.08)}$) for $\tau = 100\,sec.$ (point B).

7 Resource management policies

Retuning to our VoD use case, we now sketch two possible schemes for exploiting the Large Deviation description of the system to dynamically provision the allocated resources :

- *Identification of the reactive time scale for reconfiguration* : Find a relevant time scale that realizes a good trade-off between the expectable level of overflow associated to this scale and a sustainable OPEX cost induced by the resources reconfiguration needed to cope with the corresponding flash crowd.
- *Link capacity dimensioning* : Considering a maximum admissible loss probability, find the safety margin that it is necessary to provision on the link capacity, to guarantee the corresponding QoS.

7.1 Identification of the reactive time scale for reconfiguration

We consider the case of a VoD service provider who wants to determine the reactivity scale at which it needs to reconfigure its resource allocation. This quantity should clearly derive from a good compromise between the level of congestion (or losses) it is ready to undergo, i.e. a tolerable performance degradation, and the price it is willing to pay for a frequent reconfiguration of its infrastructure. Let us then assume that the VoD provider has fixed admissible bounds for these two competing factors, having determined the following quantities :

- $\alpha^* > \alpha_{\text{a.s.}}$: the deviation threshold beyond which it becomes worth (or mandatory) considering to reconfigure the resource allocation. This choice is uniquely determined by a CAPEX performance concern.
- σ^* : an acceptable probability of occurrence of these overflows. This choice is essentially guided by the corresponding OPEX cost.

Let us moreover suppose, that the LD spectrum $f(\alpha)$ of the workload process was previously estimated, either by identifying the parameters of the Markov model used to describe the application, or empirically from collected traces. Then, recalling the probabilistic interpretation we surmised in relation (15), the minimum reconfiguration time scale τ^* for dynamic resource allocation, that verifies the sought compromise, is simply the solution of the following inequality :

$$\tau^* = \max\{\tau : \mathbb{P}\{\langle i \rangle_\tau \geq \alpha^*\} = \int_{\alpha^*}^{\infty} e^{\tau f_\tau(\alpha)}\, d\alpha \geq \sigma^*\}, \tag{18}$$

with $f_\tau(\alpha)$ as defined in expression (16).

From a more general perspective though, we can see this problem as an underdetermined system involving 3 unknowns (α^*,τ^* and σ^*) and only one relation (18). Therefore, and depending on the sought objectives, we can imagine to fix any other two of these variables and to determine the resulting third so that it abides with the same inequality as in expression (18).

7.2 Link capacity dimensioning

We now consider an architecture dimensioning problem from the infrastructure provider perspective. Let us assume that the infrastructure and the service providers have come to a Service Level Agreement (SLA), which among other things, fixes a tolerable level of losses due to link congestion. We start considering the case of a single VoD server and address the following question : What is the minimum link capacity C that has to be provisioned such that we meet the negotiated QoS in terms of loss probability ? Like in the previous case, we assume that the estimated LD spectrum $f(\alpha)$ characterizing the application has been priorly identified. A rudimentary SLA would be to guarantee a loss free transmission for the *normal* traffic load only : this loose

QoS would simply amount to fix C to the almost sure workload $\alpha_{\text{a.s.}}$. Naturally then, any load overflow beyond this value will result in goodput limitation (or losses, if there is no buffer to smooth out exceeding loads). For a more demanding QoS, we are led to determine the necessary safety margin $C_0 > 0$ one has to provision above $\alpha_{\text{a.s.}}$ to absorb the exact amount of overruns corresponding to the loss probability p_{loss} that was negotiated in the SLA. From the interpretation of the large deviation spectrum provided in Section **??**, this margin C_0 is determined by the resolution of the following inequality :

$$\begin{aligned} C_0 \quad &: \quad \int_{\alpha_{\text{a.s.}}+C_0}^{\infty} \int_{\tau_{min}}^{\tau_{max}} e^{\tau \cdot f(\alpha)} \, \mathrm{d}\tau \, \mathrm{d}\alpha \;\leq\; p_{\text{loss}} \\ &: \quad \int_{\alpha_{\text{a.s.}}+C_0}^{\infty} \frac{e^{\tau_{max} \cdot f(\alpha)} - e^{\tau_{min} \cdot f(\alpha)}}{f(\alpha)} \, \mathrm{d}\alpha \;\leq\; p_{\text{loss}} \end{aligned} \tag{19}$$

In this expression, τ_{min} is typically determined by the size Q of the buffers that is usually provisioned to dampen the traffic volatility. In that case,

$$\tau_{\text{min}} = \frac{Q}{\alpha - (\alpha_{a.s.} + C_0)}, \tag{20}$$

corresponds to the maximum burst duration that can be buffered without causing any loss at rate $\alpha > C = \alpha_{a.s.} + C_0$. As for τ_{max}, it relates to the maximum period of reservation dedicated to the application. Most often though, the characteristic time scale of the application exceeds the dynamic scale of flash crowds by several orders of magnitude, and τ_{max} can then simply be set to infinity. With these particular integration bounds, Equation (19) simplifies to

$$C_0 = C - \alpha_{a.s.} : \int_{C}^{\infty} \frac{-1}{f(\alpha)} \, e^{\frac{Q}{\alpha - C} f(\alpha)} \, \mathrm{d}\alpha \;\leq\; p_{\text{loss}}, \tag{21}$$

a decreasing function of C, which can be solved using a simple bisection technique.
As long as the server workload remains below C, this resource dimensioning guarantees that no loss occurs. All overrun above this value will produce losses, but we ensure that the frequency (probability) and duration of these overruns are such that the loss rate remains conformed to the SLA. The proposed approach clearly contrasts with resource over-provisioning that does not seek at optimizing the CAPEX to comply with the loss probability tolerated in the SLA.

The same provisioning scheme can straightforwardly be generalized to the case of several applications sharing a common set of resources. To fix the idea, let us consider an infrastructure provider that wants to host K VoD servers over the same shared link. A corollary question is then to determine how many servers K can the fixed link capacity C support, while guaranteeing a prescribed level of losses. If the servers are independent, the probability for two of them to undergo a flash crowd simultaneously is negligible. For ease and without loss of generality, we moreover suppose that they are identically distributed and modeled by the same LD spectrum $f^{(k)}(\alpha) = f(\alpha)$ with the same nominal workload $\alpha_{\text{a.s.}}^{(k)} = \alpha_{\text{a.s.}}$, $k = 1, \ldots K$. Then, following the same reasoning as in the previous case of a single server, the maximum number K of servers reads :

$$K = \arg\max_{K} \left(C - K \cdot \alpha_{\text{a.s.}} \right) \leq C_0, \tag{22}$$

where the safety margin C_0 is defined as in expression (21).

Then, depending on the agreed *Service Level Agreements*, the infrastructure provider can easily offer different levels of probability losses (QoS) to its VoD clients, and adapt the number of hosted servers, accordingly.

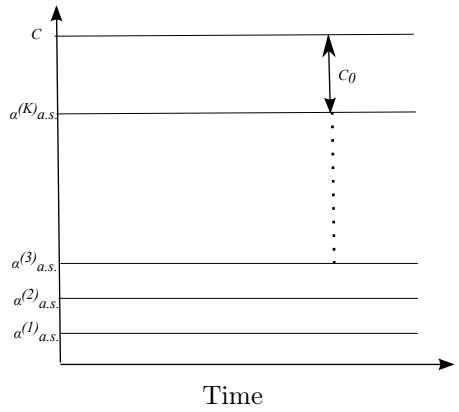


FIGURE 13 – Dimensioning K, the number of hosted servers sharing a fixed capacity link C. The safety margin C_0 is determined according to the probabilistic loss rate negotiated in the *Service Level Agreement* between the infrastructure provider and the VoD service provider.

8 Conclusion

Many applications deployed on a cloud infrastructure, such as a Video on Demand service, are well known for undergoing highly volatile demand, making their workload hard to qualitatively and quantitatively characterize. Adopting a constructive approach to capture the VoD users' behavior, in this report we proposed a simple, concise and versatile model for generating the workload variations in such context. We also devised an heuristic identification procedure that aims at estimating the parameters values of the model from a single collected trace. First, we numerically evaluated the accuracy of this procedure using several synthetic traces. Our experiments show that the procedure introduces little bias and typically recovers the actual parameters value with a relative error of about 10%. Second, we apply this same procedure against two real video server workload traces. Obtained results demonstrate that, once the model has been calibrated, it succeeds to reproduce the statistical behavior of the real trace (in terms of both the steady-state probabilities and the autocorrelations for the workload time series). Moreover, owing to the constructive nature of our model, the estimated values of the parameters provide valuable insight on the application that it would be difficult, or even impossible, to infer from the raw traces. The captured information may answer questions of practical interest to cloud oriented service providers, like : is the application workload mostly driven by spontaneous behaviors, or is it rather subject to a gossip phenomenon ?

Furthermore, a key-point of this model is that it permits to reproduce the workload time series with a Markovian process, which is known to verify a Large Deviation Principle (LDP). This particularly interesting property yields a large deviation spectrum whose interpretation enriches the information conveyed by the standard steady state distribution : For a given observation (workload trace), LDP allows to infer (theoretically and empirically) the probability that the time average workload, calculated at an arbitrary aggregation scale, deviates from its nominal value (i.e. almost sure value).

We leveraged this multiresolution probabilistic description to conceptualize two different management schemes for dynamic resource provisioning. As explained, the rationale is to use large deviation information to help network and service providers together to agree on the best CAPEX-OPEX trade-off. Two major stakes of this negotiation are : *(i)* to determine the largest reconfiguration time scale adapted to the workload elasticity and *(ii)* to dimension VoD server so as

to guarantee with upmost probability the Quality of Service imposed by the negotiated Service Level Agreement.
More generally though, the same LDP based concepts can benefit any other “Service on Demand" scenarii to be deployed on dynamic cloud environments.

Références

[1] Sandvine, "Sandvine's spring 2011 global internet phenomena report reveals new internet trends," May 2011, `http://www.sandvine.com/news/pr_detail.asp?ID=312/`.

[2] B. Andy, "Star kids the data dump," `http://waxy.org/2008/05/star_wars_kid_the_data_dump/`.

[3] BBC, "Star wars kid is top viral video," November 2006, `http://news.bbc.co.uk/2/hi/entertainment/6187554.stm/`.

[4] P.T Eugster, R. Guerraoui, A. Kermarrec, and L. Massoulie, "Epidemic information dissemination in distributed systems," *IEEE Computer Society*, vol. 37, no. 5, pp. 60–67, May 2004.

[5] T. Bonald, L. Massoulié, F. Mathieu, D. Perino, and A. Twigg, "Epidemic live streaming : Optimal performance trade-offs," *ACM SIGMETRICS Performance Evaluation Review - SIGMETRICS '08*, vol. 36, no. 1, pp. 325–336, June 2008.

[6] R. Bakhshi, D. Gavidia, W. Fokkink, and M. van Steen, "A modeling framework for gossip-based information spread," in *Proc. 8th International Conference on Quantitative Evaluation of SysTems*, September 2011.

[7] E. Caron, F. Desprez, and A. Muresan, "Pattern matching based forecast of non-periodic repetitive behavior for cloud clients," *Journal of Grid Computing*, vol. 9, no. 1, pp. 49–64, March 2011.

[8] R. Garcia, X. Paneda, V. Garcia, D. Melendi, and M. Vilas, "Statistical characterization of a real video on demand service : User behaviour and streaming-media workload analysis," *Simulation Modelling Practice and Theory, doi :10.1016/j.simpat.2007.02.004*, vol. 15, pp. 672–689, 2007.

[9] V.O.K. Li, W. Lao, X. Qiu, and E. W. M. Wong, "Performance model of interactive video-on-demand systems," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATION*, vol. 14, pp. 1099–1108, 1996.

[10] D. Melendi, R. Garcia, X. G. Paneda, and V. Garcia, "Multivariate distributions for workload generation in video on demand systems," *IEEE COMMUNICATIONS LETTERS*, vol. 13, pp. 348–350, 2009.

[11] S. Kanrar, "Analysis and implementation of the large scale video-on-demand system," *International Journal of Applied Information Systems*, vol. 2, no. 2, February 2012.

[12] S. Pacheco-Sanchez, G. Casale, B. Scotney, S. McClean, G. Parr, and S. Dawson, "Markovian workload characterization for qos prediction in the cloud," in *Proc. IEEE Cloud*, June 2011.

[13] D. Niu, B. Li, and S. Zhao, "Understanding demand volatility in large vod systems," in *Proc. NOSSDAV*, June 2011.

[14] D. Perez-Palacin, J. Merseguer, and R. Mirandola, "Analysis of bursty workload-aware self-adaptive systems," in *Proc ICPE*, April 2012.

[15] R. Gusella, "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS*, vol. 9, no. 2, February 1991.

[16] A. Barrat, M. Barthelemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, 1st edition, November 2008.

[17] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*, vol. 151, Springer, Lecture Notes in Statistics, 2000.

[18] S.R. Jammalamadaka and E. Taufer, "Testing exponentiality by comparing the empirical distribution function of the normalized spacings with that of the original data," *J. Nonparametric Statistics*, vol. 15, pp. 719 – 729, 2003.

[19] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proc. the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

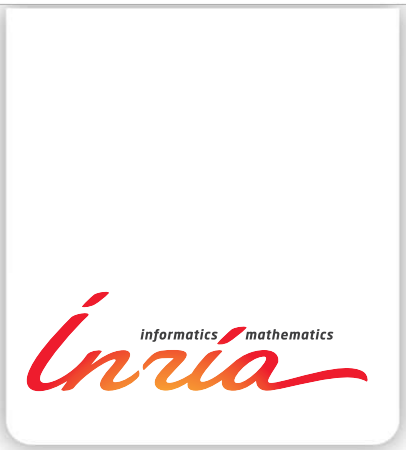
[20] GRNET, "Video traces obtained from grnet," 2011, `http://vod.grnet.gr/`.

[21] S.R.S. Varadhan, "Large deviations," *The Annals of Probability*, vol. 36, no. 2, pp. 397–419, 2008.

[22] J. Barral and P. Loiseau, "Large deviations for the local fluctuations of random walks," *Stochastic Processes and their Applications*, vol. 121, no. 10, pp. 2272–2302, 2011.

[23] J. Barral and P. Gonçalves, "On the estimation of the large deviations spectrum," *Journal of Statistical Physics*, vol. 144, no. 6, pp. 1256–1283, 2011.

[24] P. Loiseau, P. Gonçalves, J. Barral, and P. Vicat-Blanc Primet, "Modeling TCP throughput : an elaborated large-deviations-based model and its empirical validation," in *Proceedings of IFIP Performance*, Nov 2010.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399