INVITED PAPER *Special Section on Networking Technologies for Cloud Services*

Dynamic Resource Management in Clouds: A Probabilistic Approach

Paulo GONÇALVES[†a)], **Shubhabrata ROY**[†], **Thomas BEGIN**[†], ***and*** **Patrick LOISEAU**[††], ***Nonmembers***

SUMMARY 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 work 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. We show that the resulting model verifies a Large Deviation Principle that statistically characterizes extreme rare events, such as the ones produced by "buzz/flash crowd 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, resource management, epidemic model, workload generator, large deviation principle, service level agreements, video on demand, buzz/flash crowd*

1. Introduction

Users of a Cloud Computing platform can have several numbers of choices regarding server selection (some are compute intensive, some provide better I/O performance, some are superior in networking). Cloud provider such as Amazon offers many different server instances that differ in many aspects with respect to CPU speed, network bandwidth and memory capacity. Each of these instances provides a certain amount of dedicated resource and charges per instance-hour consumed [1]. A Service Provider finds it to be extremely difficult to optimize the best combination of servers to be deployed in a *Cloud* for his business on a certain application. This problem differs from the concept of traditional distributed computing (like Grid), since the numbers of servers are virtually unlimited but bandwidth is limited. The choice of deployment of resources can be dynamically tuned using cloud virtualization, that abstracts the IT resources to allow communication and control on-line. Cost of resources varies significantly depending on server types and Cloud Service Providers.

In most applications, the amount of IT resource that is actually used, is a highly variable quantity that follows the instantaneous activity, and in particular the volume of exchanged traffic when network infrastructures are concerned. Depending on the type of application, the generated workload can be a highly varying process that turns difficult to find an acceptable trade-off between an expensive overprovisioning able to anticipate peak loads and a sub performing resource allocation that does not mobilize enough resources. To bypass this challenge, dynamic bandwidth allocation is an original approach that we chose to investigate in the context of network virtualization. We aim to demonstrate the proof of concept for the case of a Video on Demand (VoD) system by adaptively tuning the provisioned bandwidth to the current application workload. In this paper we have resorted to probabilistic provisioning of resource management; however in some situations it can be used to anticipate resource requirements that can serve as inputs for dynamic resource allocation.

Our work attempts to capture some properties that describe user behaviors or workload generating mechanism of the system and fits them to a mathematical model satisfying particular properties. We leverage these properties to derive a probabilistic assumption on the mean workload of the system at different time resolutions. Embedding the notion of time scale is very important since time scale is by essence intrinsic to dynamicity. In this study we build our system using epidemic models where Markovian models are widely used and happen to satisfy to the specific property mentioned above.

Epidemic information dissemination has been an active area of research in distributed systems, such as Peer-to-Peer (P2P) or VoD systems. In [2], it has been already demonstrated that the epidemic algorithms can be used as an effective solution for information dissemination in the P2P systems as deployed on Internet or ad-hoc networks. The authors of [3] studied random epidemic strategies like the random peer, latest useful chunk algorithm to achieve optimal information dissemination. However the most relevant work to our study is derived in [4] where the authors proposed an approach to predict workload for cloud clients. They used auto-scaling algorithm for resource provisioning and validated the result with real-world Cloud client application traces. Our approach encompasses both constructive Markovian model to reproduce epidemic information dissemination and workload provisioning aspects. However, we insist on the fact that its originality stems from the analysis of the Large Deviation property of the proposed Marko-

Manuscript received December 9, 2011.
Manuscript revised March 22, 2012.

[†] The authors are with LIP, UMR 5668 Inria - ENS Lyon - UCB Lyon 1 - CNRS, France.
[††] The author is with EURECOM, Sophia Antipolis, France.
a) E-mail: paulo.goncalves@ens-lyon.fr

DOI: 10.1587/transcom.E95.B.2522

vian model. The resulting characterization can be viewed as a multi-resolution extension of the classical steady-state distribution for the observable mean value of the random process over different aggregated time scales.

After constructing the Markovian mathematical model, we propose two possible and generic ways to exploit these 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.

The rest of the paper is organized as follows. In Sect. 2 we discuss the VoD system as our use case, followed by a Markovian description of the model in the Sect. 3. Section 4 presents Large Deviation Principle. We discuss the numerical interpretations in Sect. 5. Section 6 deals with the probabilistic provisioning scheme, derived from the Large Deviation Spectrum for our use case followed by the conclusion in Sect. 7.

2. Use Case: Video on Demand (VoD)

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” [5]. 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 1 shows a VoD schematic

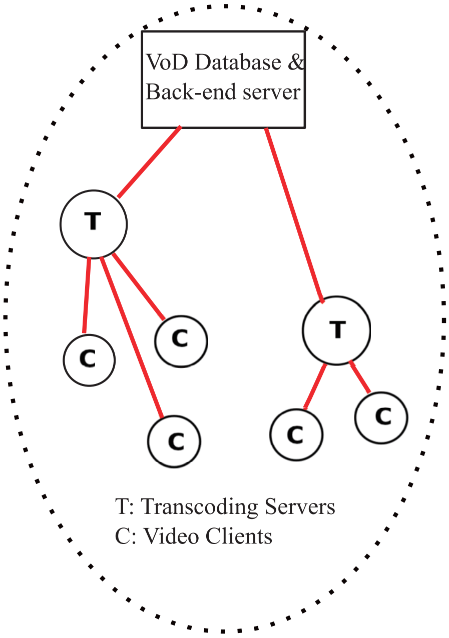

Fig. 1 Basic schematics of a VoD system with transcoding/caching servers.

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 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. Following is one example of “buzz” where interest over a video“Star Wars Kid” [6] grew very quickly within a very short timespan. According to [7] it was viewed more than 900 million times within a short interval of time making it one of the top viral videos. Figure 2 plots the original server logs for the Star Wars Kid debacle [6].

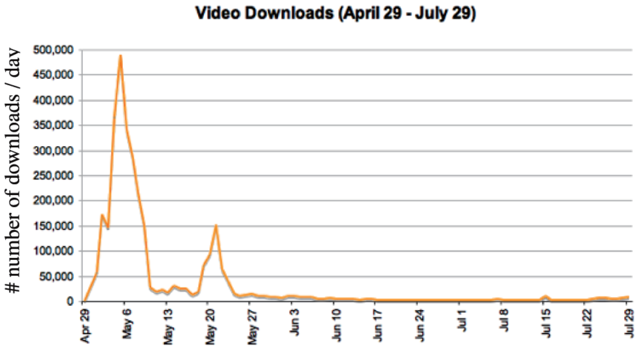

Fig. 2 Video server workload: time series displaying a characteristic pattern of flash crowd (buzz effect). Trace obtained from URL: http://waxy.org/2008/05/star_wars_kid_the_data_dump/

In situations like the one described in Fig. 2, 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. Markov Model to Describe the Behavior of the Users

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 [8]. 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 dt, 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 of 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 $t + \mathrm{d}t$ (dt being small) reads:

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

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

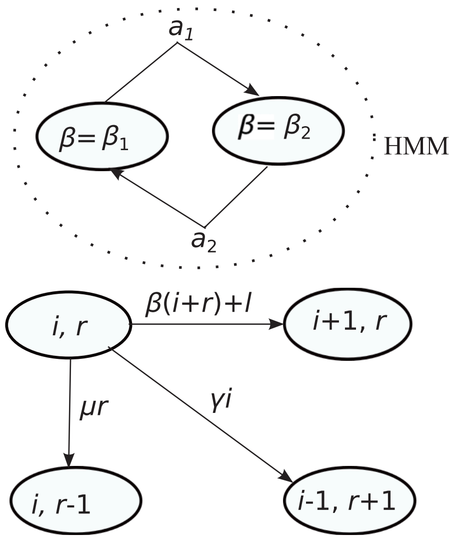

Fig. 3 Markov chain diagram representing the evolution of the Current viewers (i) and Past Viewers (r) populations with a Hidden Markov Model.

two hidden and memoryless Markov states occur with rates a_1 and a_2 respectively (see Fig. 3). These rates characterize the buzz in terms of frequency, magnitude and duration.

4. Large Deviation Principle

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 [9], gives an 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:

$$\lim_{\epsilon\to 0}\lim_{\tau\to\infty}\frac{1}{\tau}\log\mathbb{P}\left\{\int_0^\tau \Phi(X_s)\mathrm{d}s \in [\alpha-\epsilon, \alpha+\epsilon]\right\} = f(\alpha). \quad (4)$$

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:

$$f(\alpha) = \sup_{q\in\mathbb{R}}\{q\alpha - \Lambda(q)\}\,, \forall\alpha \in \mathbb{R}. \quad (5)$$

As described in Eq. (4), $\alpha_\tau = \langle i\rangle_\tau$ 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:

$$\mathbb{P}\{\langle i\rangle_\tau \approx \alpha\} \sim e^{\tau\cdot f(\alpha)}. \quad (6)$$

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" [10]. 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 $\alpha \in \mathbb{R}$:

$$f_\tau(\alpha,\epsilon_\tau) = \frac{1}{\tau}\log\frac{\#\left\{j : \int_{I_{j,\tau}}\Phi(X_s)\mathrm{d}s \in [\alpha-\epsilon_\tau, \alpha+\epsilon_\tau]\right\}}{k_\tau} \quad (7)$$

$$\text{and } \lim_{\tau\to\infty} f_\tau(\alpha,\epsilon_\tau) = f(\alpha).$$

In practice, for the empirical estimation of the large deviations spectrum, we use a similar estimator as the one derived in [11] and also used in [12]. 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 (7) and estimate the scale-dependant empirical *log-pdf* $f_\tau(\alpha,\epsilon_\tau)$, with the adaptive choices derived in [11]:

$$\alpha_\tau = \Lambda'_\tau(q) \quad \text{and} \quad \epsilon_\tau = \sqrt{\frac{-\Lambda''_\tau(q)}{\tau}}. \quad (8)$$

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 within a period τ.

5. Numerical Interpretations

We simulate the proposed workload model and generate two time series corresponding to the buzz and to the buzz free situations. We developed our simulator in *C* programming environment, by creating several parallel child processes (client) that communicate with a parent process (server) to disseminate information. The child process is in any of the susceptible, active viewers or past viewers states at a particular instant of time. When it is in the past viewers state it randomly chooses another process (using process id) and communicates with the parent to infect him. The parent process maintains a table with the status (which state a process is in) of each process. If the chosen process is not already in active viewers or in past viewers states it gets infected. We

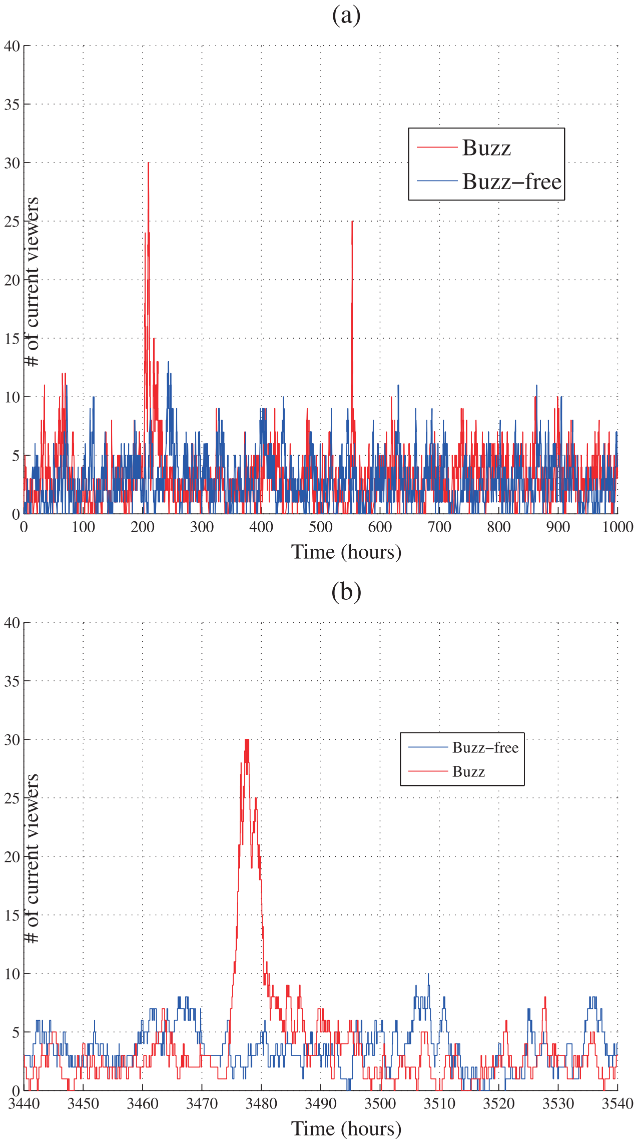

Fig. 4 Plot (a): Workload $N_I(t)$ generated according to the model depicted in Fig. 3 (For the buzz case: $\beta_1 = 0.1$, $\beta_2 = 0.8$, $\gamma = 0.7$, $\mu = 0.3$, $l = 1.0$, $a_1 = 0.006$ and $a_2 = 0.6$. For the buzz-free case: $\beta_1 = \beta_2 = \beta = 0.1$, $\gamma = 0.7$, $\mu = 0.3$, $l = 1.0$). In both cases, $I_{\max} = 30, R_{\max} = 60$. Plot (b): Zoomed in view of a buzz event.

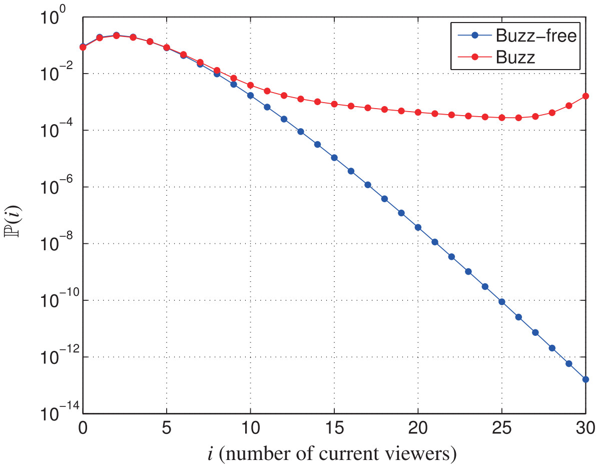

Fig. 5 Steady-state probabilities for the number of current viewers with buzz and buzz-free scenarios (Y-axis in log-scale).

have chosen UDP socket-pairs in order to facilitate communication between the processes. For fair and consistent comparisons, we carefully tuned the values of the model parameters so as to obtain the same mean workload for both resulting traces. In Fig. 4(a) the bursty transients represent the buzz effect. It reflects sudden and sharp increases of workload due to intense dissemination of popular videos. The zoomed in view displayed in Fig. 4(b) shows the characteristic pattern of a buzzy transient, that is to say a sharp increase ($\beta_1 \to \beta_2$) and a slower decrease (owing to $\beta_2 \to \beta_1$ and to the memory effect of the model).

This clear evidence of our model's ability to capturing the buzz effect is moreover confirmed by the numerical steady-state distributions $\mathbb{P}(i)$ displayed in Fig. 5. As compared to the buzz-free case, the buzz distribution presents a thicker tail indicating that the instantaneous workload i takes on larger values with higher probability. To include the notion of time scale in the results one needs to consider along with the steady-state distribution the time coherence of the underlying process, viz. its covariance structure. However, except for the trivial case of uncorrelated processes, deriving the statistics of the local average process at any resolution is a hard problem in general.

Intrinsically, Large Deviation Principle naturally embeds this time scale notion into the statistical description of the aggregated observable at different time resolutions. As expected, the theoretical LD spectra displayed in Fig. 6(a) reach their maximum for the same mean number of users. This apex is the almost sure value as described in Sect. 4. 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.

Plots (b)-(c) of Fig. 6 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

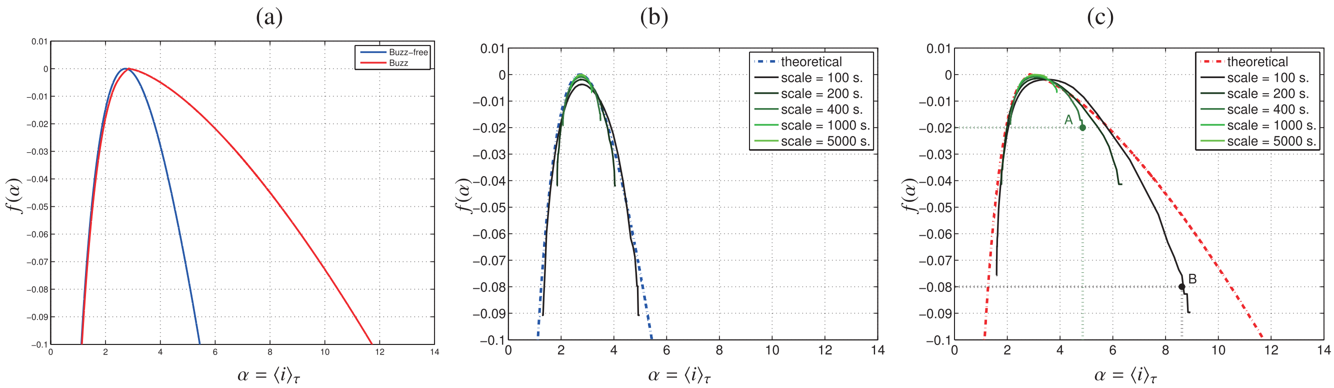

Fig. 6 Large Deviations spectra corresponding to the traces of Fig. 4. (a) Theoretical spectra for the buzz free (blue) and for the buzz (red) scenarii. (b) & (c) Empirical estimations of $f(\alpha)$ at different scales from the buzz free and the buzz traces, respectively.

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).

6. Probabilistic Provisioning

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.

6.1 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 (6), the minimum reconfiguration time scale τ^* for dynamic resource allocation, that verifies the sought compromise, is simply the solution of the following inequality:

$$\tau^* = \max\{\tau : \mathbb{P}\{\langle i \rangle_\tau \geq \alpha^*\} = \int_{\alpha^*}^{\infty} e^{\tau f_\tau(\alpha)}\, d\alpha \geq \sigma^*\}, \quad (9)$$

with $f_\tau(\alpha)$ as defined in expression (7).

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 (9). 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 (9).

6.2 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 Sect. 4, this margin C_0 is determined by the resolution of the following inequality:

$$\begin{aligned} C_0 \ &: \int_{\alpha_{\text{a.s.}}+C_0}^{\infty} \int_{\tau_{min}}^{\tau_{max}} e^{\tau \cdot f(\alpha)} \,\mathrm{d}\tau \,\mathrm{d}\alpha \ \leq \ p_{\text{loss}} \\ &: \int_{\alpha_{\text{a.s.}}+C_0}^{\infty} \frac{e^{\tau_{max}\cdot f(\alpha)} - e^{\tau_{min}\cdot f(\alpha)}}{f(\alpha)} \,\mathrm{d}\alpha \ \leq \ p_{\text{loss}} \end{aligned} \qquad (10)$$

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,

$$\tau_{\min} = \frac{Q}{\alpha - (\alpha_{a.s.} + C_0)}, \qquad (11)$$

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 $\tau_{\max}$, 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 $\tau_{\max}$ can then simply be set to infinity. With these particular integration bounds, Eq. (10) simplifies to

$$C_0 = C - \alpha_{a.s.} : \int_{C}^{\infty} \frac{-1}{f(\alpha)} \, e^{\frac{Q}{\alpha - C} f(\alpha)} \,\mathrm{d}\alpha \ \leq \ p_{\text{loss}}, \qquad (12)$$

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 (as schematized in Fig. 7). 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:

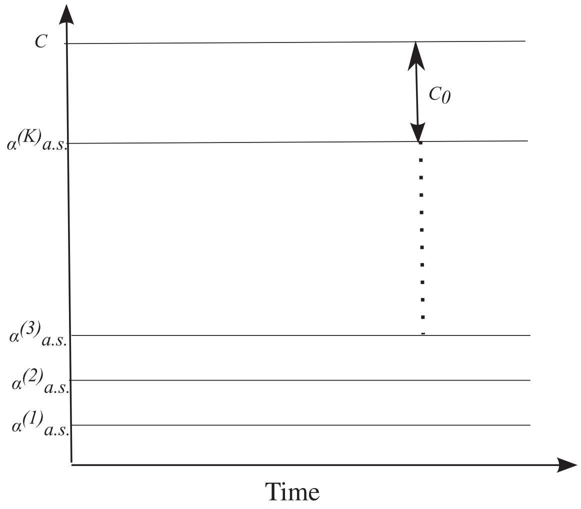

Fig. 7 Dimensioning K, the number of hosted servers sharing a fixed capacity link C. The safety margin C_0 is determined according to the probabilistic loss rate negotiated in the *Service Level Agreement* between the infrastructure provider and the VoD service provider.

$$K = \arg\max_{K} \, (C - K \cdot \alpha_{\text{a.s.}}) \leq C_0, \qquad (13)$$

where the safety margin C_0 is defined as in expression (12).

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.

7. Conclusion

The objective of this work is to harness probabilistic methods for resource provisioning in the Clouds. We illustrate our purpose with a Video on Demand scenario, a characteristic service whose demand relies on information spreading. Adopting a constructive approach to capture the users' behavior, we proposed a simple, concise and versatile model for generating the workload variations in such context. 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.

References

[1] Amazon, "Amazon web service server instance choices," http://aws.amazon.com/ec2/instance-types/

[2] P. 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.

[3] 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.

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

[5] Sandvine, "Sandvines spring 2011 global internet phenomena report reveals new internet trends," May 2011.

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

[7] BBC, "Star wars kid is top viral video, month = november," 2006.

[8] A. Barrat, M. Barthelemy, and A. Vespignani, Dynamical Processes on Complex Networks, 1st ed., Cambridge University Press, Nov. 2008.

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

[10] 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.

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

[12] P. Loiseau, P. Gonçalves, J. Barral, and P. Vicat-Blanc Primet, "Modeling TCP throughput: An elaborated large-deviations-based model and its empirical validation," Proc. IFIP Performance, Nov. 2010.

Paulo Gonçalves graduated from the Signal Processing Department of ICPI Lyon (now CPE Lyon), France in 1993. He received the Masters (DEA) and Ph.D. degrees in signal processing from the Institut National Polytechnique de Grenoble, France, in 1990 and 1993 respectively. While working toward his Ph.D. degree, he was with Ecole Normale Suprieure de Lyon (ENS-Lyon). Since 1996, he is associate researcher at Institut National de Recherche en Informatique et Automatique (INRIA). He is currently head of the INRIA team "RESO" at the Laboratoire de l'Informatique du Paralllisme (LIP) of ENS-Lyon. P. Gonalves research interests are in multiscale analysis (signals, images and systems) and in wavelet-based statistical inference. His principal application is in metrology and deals with grid traffic statistical characterization and modelling for protocol quality assessment and control.

Shubhabrata Roy did his Bachelors in Electrical Engineering from the Jadavpur University, India and Masters in Communication and Networks at SSSUP in CNR, Pisa. He is currently pursuing his Ph.D. at Ecole Normale Suprieure de Lyon under the supervision of Paulo Gonçalves and Thomas Begin. His research interests include Network Virtualization, Cloud Computing and Stochastic Processes.

Thomas Begin is an Assistant Professor at Universit Claude Bernard Lyon 1. He joined this university in September 2009 and is a member of the INRIA RESO Team at the LIP Laboratory. He received his Ph.D. degree in Computer Science from the University Pierre et Marie Curie in 2008, after earning a M.Sc. in Computer Networks from University Pierre et Marie Curie in 2005 and a M.Sc. in Electronics Engineering from ISEP in 2003. In Spring 2009, he was invited to University of California Santa Cruz as a visiting researcher. His research interests include performance evaluation, queueing theory and wireless networks.

Patrick Loiseau received a M.Sc. degree in physics (2006) and a Ph.D. degree in computer science (2009) from Ecole Normale Suprieure de Lyon (France). He received a M.Sc. degree in mathematics (2010) from UPMC (U. Paris 6 and Ecole Polytechnique). He was a post-doctoral fellow at INRIA Paris-Rocquencourt (2010) and at UC Santa Cruz (2011). He is currently Assistant Professor in the networking and security department at EURECOM (France). Patrick Loiseau's main research interests are in the areas of probability, statistics and game theory with applications to networks modeling. He is specifically interested in network traffic modeling, performance evaluation, inference of traffic characteristics (sampling), resource pricing and modeling of network security interactions. He has also worked on large deviations with applications to heart-rate modeling.