Adv. Appl. Prob. **49,** 629–651 (2017)
*doi:*10.1017/apr.2017.15

ALLOCATION SCHEMES OF RESOURCES WITH DOWNGRADING

CHRISTINE FRICKER,* *INRIA*
FABRICE GUILLEMIN,** *Orange Labs*
PHILIPPE ROBERT * *** AND
GUILHERME THOMPSON,* **** *INRIA*

Abstract

We consider a server with large capacity delivering video files encoded in various resolutions. We assume that the system is under saturation in the sense that the total demand exceeds the server capacity C. In such a case, requests may be rejected. For the policies considered in this paper, instead of rejecting a video request, it is downgraded. When the occupancy of the server is above some value $C_0 < C$, the server delivers the video at a minimal bit rate. The quantity C_0 is the bit rate adaptation threshold. For these policies, request blocking is thus replaced with bit rate adaptation. Under the assumptions of Poisson request arrivals and exponential service times, we show that, by rescaling the system, a process associated with the occupancy of the server converges to some limiting process whose invariant distribution is computed explicitly. This allows us to derive an asymptotic expression of the key performance measure of such a policy, namely the equilibrium probability that a request is transmitted at requested bitrate. Numerical applications of these results are presented.

Keywords: Resource allocation; scaling method; loss system

2010 Mathematics Subject Classification: Primary 60K25; 60K30
Secondary 90B18

1. Introduction

Over the past few years video streaming applications have become the dominant applications in the internet and generate the prevalent part of traffic in today's internet protocol (IP) networks; see, for instance, [10] for an illustration of the application breakdown in a commercial IP backbone network. Video files are currently downloaded by customers from large data centers, such as Google's data centers for YouTube® files. In the future, it is very likely that video files will be delivered by smaller data centers located closer to end users, for instance cache servers disseminated in a national network. It is worth noting that, as shown in [11], caching is a very efficient solution for YouTube traffic. While this solution can improve performances by reducing delays, the limited capacity of those servers in terms of bandwidth and computing can cause overload.

One possibility to reduce overload is to use bit rate adaptation. Video files can indeed be encoded at various bit rates (for example, standard and high definition video). If a node cannot

Received 5 April 2016; revision received 24 February 2017.
* Postal address: INRIA, 2 Rue Simone IFF, CS 42112, 75589 Paris Cedex 12, France.
** Postal address: CNC/NCA Orange Labs, 2 Avenue Pierre Marzin, 22300 Lannion, France.
*** Email address: philippe.robert@inria.fr
**** Research supported by the Brazilian Government/CAPES under grant BEX 13748-13-0.

serve a file at a high bit rate then the video can be transmitted at a lower rate. It is remarkable that video bit rate adaptation has become very popular in the past few years with the specification of dynamic adaptive streaming over HTTP (DASH), also known as MPEG-DASH standard, where it is possible to downgrade the quality of a given transmission; see [1], [7], [17], [18], and [21]. Adaptive streaming is also frequently used in mobile networks where bandwidth is highly varying. In this paper we investigate the effect of bit rate adaptation in a node under saturation.

1.1. Downgrading policy

We assume that customers request video files encoded at various rates, say, A_j for $j = 1, \ldots, J$, with $1 = A_1 < A_2 < \cdots < A_J$. Jobs of class $j \in \{1, \ldots, J\}$ require bit rate A_j. The total capacity of the communication link is C. If $\ell = (\ell_j)$ is the state of the network at some moment, with ℓ_j being the number of class j jobs, the quantity $\langle A, \ell\rangle = A_1\ell_1 + \cdots + A_J\ell_J$ has to be less than C. The quantity $\langle A, \ell\rangle$ is defined as the *occupancy* of the link. The algorithm has a parameter $C_0 < C$ and works as follows. If there is an arrival of a job of class $1 \le j_0 \le J$,

- if $\langle A, \ell\rangle < C_0$ then the job is accepted;
- if $C_0 \le \langle A, \ell\rangle < C$ then the job is accepted but as a class 1 job, i.e. it has an allocated bit rate of $A_1 = 1$ and service rate μ_1;
- if $\langle A, \ell\rangle = C$ then the job is rejected.

For $1 \le j \le J$, jobs of class j arrive according to a Poisson process with rate λ_j and have an exponentially distributed transmission time with rate μ_j. Additionally, it is assumed that

$$\mu_1 \le \min(\mu_j, 2 \le j \le J).$$

1.2. A scaling approach

To study this allocation scheme, a scaling approach is used. It is assumed that the server capacity is very large, namely, scaled up by a factor N. The bit rate adaptation threshold and the request arrival rates are scaled up accordingly, i.e.

$$\lambda_j \mapsto \lambda_j N, \quad 1 \le j \le J, \qquad C_0 \mapsto c_0 N \quad \text{and} \quad C \mapsto cN. \tag{1}$$

1.2.1. *Performances of the algorithm.* Our main result shows that, for the downgrading policy and if c_0 is chosen conveniently, then

- the equilibrium probability of rejecting a job converges to 0 as N goes to ∞;
- the equilibrium probability of accepting a job without downgrading it converges to

$$\pi^- := \left(c_0\mu_1 - \sum_{j=1}^{J}\lambda_j\right) \Big/ \left(\mu_1 \sum_{j=1}^{J}\frac{\lambda_j}{\mu_j}A_j - \sum_{j=1}^{J}\lambda_j\right) \quad \text{as } N \to \infty.$$

See Theorem 2 and Corollary 1.

The above formula gives an explicit expression of the success rate of this allocation mechanism. The quantity $1 - \pi^-$, the probability of downgrading requests, can be seen as the 'price' of the algorithm to avoid rejecting jobs.

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

The scaling (1) was introduced by Kelly to study loss networks [14]. The transient behavior of these networks under this scaling has been analyzed by Hunt and Kurtz [12]. This last reference provides essentially a framework to establish convenient convergence theorems involving stochastic averaging principles. This line of research was developed in the 1990s to study uncontrolled loss networks where a request is rejected as soon as its demand cannot be accepted.

When the demand can be adapted to the state of the network, for controlled loss networks, several (scarce) examples have also been analyzed during that period of time; see, for example, [4], [5], [22], and [23]. Our model can be seen as a 'controlled' loss network instead of a pure loss network. Controlled loss networks may have mechanisms such as trunk reservation or may allocate requests according to some complicated schemes depending on the state of the network. In our case, the capacity requirements of requests are modified when the network is in a 'congested' state.

Contrary to classical uncontrolled loss networks, as it will be seen, the Markov process associated to the evolution of the vector of the number of jobs for each class is not reversible. Additionally, the invariant distribution of this process does not seem to have a closed-form expression. Kelly's approach [13] is based on an optimization problem, it cannot be used in our case to obtain an asymptotic expression of some characteristics at equilibrium. For this reason, the equilibrium behavior of these policies is investigated in a two step process.

(i) *Transient Analysis*. We investigate the asymptotic behavior of some characteristics of the process on a finite-time interval when the scaling parameter N goes to ∞.

(ii) *Equilibrium*. The stability properties of the limiting process are analyzed, we prove that the equilibrium of the system for a fixed N converges to the equilibrium of the limiting process.

For our model, the transient analysis involves the *explicit* representation of the invariant distribution of a specific class of Markov processes. It is obtained using complex analysis arguments. As it will be seen, this representation plays an important role in the analysis of the asymptotic behavior at equilibrium.

It should be noted that related models have recently been introduced to investigate resource allocation in a cloud computing environment where the nodes receive requests of several types of resources; see, for example, [8], [19], and [20].We believe that this domain will receive a renewed attention in the coming years. It could be said that there is renewed interest in the study of loss networks. Part of the motivation of this paper to shed some light on the methods that can be used to study these systems.

1.3. Outline of the paper

We consider a system in overload. Because of bit rate adaptation, requests may be downgraded but not systematically rejected as in a pure loss system. As it will be seen, the stability properties of this algorithm are linked to the behavior of a Markov process associated to the occupation of the link. Under exponential assumptions for interarrival and service times, this process turns out to be, after rescaling by a large parameter N, a bilateral random walk instead of a reflected random walk as in the case of loss networks. Using complex analysis methods, an explicit expression of the invariant distribution of this random walk is obtained. With this result, the asymptotic expression of the probability that, at equilibrium, a job is transmitted at its requested rate (and, therefore, does not experience a bit rate adaptation) is derived.

This paper is organized as follows. In Section 2 we present the model used to study the network under some saturation condition. Convergence results when the scaling factor N tends

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

to ∞ are proved in Section 3. The invariant distribution of a limiting process associated to the occupation of the link is computed in Section 4 by means of complex analysis techniques. Applications are discussed in Section 5.

2. Model description

We consider a service system where J classes of requests arrive at a server with bandwidth/capacity C. Requests of class j, $1 \leq j \leq J$, arrive according to a Poisson process $\mathcal{N}_{\lambda_j}$ with rate λ_j. A class j request has a bandwidth requirement of A_j units for a duration of time which is exponentially distributed with parameter μ_j. For the systems investigated in this paper, there is no buffering, requests have to be processed at their arrival otherwise they are rejected. Without any flexibility on the resource allocation, this is a classical loss network with one link; see, for example, [14].

In this paper we investigate allocation schemes which consist of reducing the bandwidth allocation of arriving requests to a minimal value when the link has a high level of congestion. In other words, the service is downgraded for new requests arriving during a saturation phase. If the system is correctly designed, it will reduce significantly the fraction of rejected transmissions and, hopefully, few jobs will in fact experience downgrading.

2.1. Downgrading policy $\mathcal{D}(C_0)$

We introduce $C_0 < C$, the parameter C_0 will indicate the level of congestion of the link. It is assumed that the vector of integers $A = (A_j)$ is such that $A_1 = 1 < A_2 < \cdots < A_J$. The condition $A_1 = 1$ is used to simplify the presentation of the results and to avoid problems of irreducibility in particular, but this is not essential.

If the network is in state $\ell = (\ell_j)$ and if the occupancy $\langle A, \ell \rangle$ is less than C_0, then any arriving request is accepted. If the occupancy is between C_0 and $C - 1$, it is accepted but with a minimal allocation, as a class 1 job. Finally, it is rejected if the link is fully occupied, i.e. $\langle A, \ell \rangle = C$. It is assumed that $\mu_1 \leq \mu_j$ for $1 \leq j \leq J$, i.e. class 1 jobs are served with the smallest service rate.

Mathematically, the stochastic model is close to a loss network with the restriction that a job may change its requirements depending on the state of the network. This is a controlled loss network; see [23]. It does not appear that, such as in uncontrolled loss networks, the associated Markov process giving the evolution of the vector ℓ has reversibility properties, or that its invariant distribution has a product-form expression. Related schemes with product form are trunk reservation policies for which requests of a subset of classes are systematically rejected when the level of congestion of the link is above some threshold; see, for example, [4] and [22]. Concerning controlled loss networks, mathematical results are more scarce. We can mention networks where jobs requiring congested links are redirected to less loaded links. Several mathematical approximations have been proposed to study these models; see the surveys [14] and [23]. In our model, in the language of loss networks, the control is on the change of capacity requirements instead of a change of link.

2.2. Scaling regime

The invariant distribution being, in general, not known, a scaling approach is used. The network is investigated under Kelly's regime, i.e. under heavy traffic regime with a scaling factor N. This regime was introduced in [13] in order to study the equilibrium of uncontrolled networks. The arrival rates are scaled by N: λ_j is replaced by $\lambda_j N$ as well as the capacity C

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

by C^N and the threshold C_0 by C_0^N, which are such that

$$C^N = cN + o(N) \quad \text{and} \quad C_0^N = c_0 N + o(N), \quad \text{for } 0 < c_0 < c.$$

Definition 1. For $1 \le j \le J$ and $t \ge 0$, $L_j^N(t)$ denotes the number of class j jobs at time t in this system and $L^N(t) = (L_j^N(t),\ 1 \le j \le J)$.

It will be assumed that the system is overloaded when the jobs have their initial bandwidth requirements

(R) $\langle A, \rho\rangle > c$ and $\Lambda/\mu_1 < c$,

with $\Lambda = \lambda_1 + \cdots + \lambda_J$ and $\rho_j = \lambda_j/\mu_j$, $1 \le j \le J$. The first condition gives that, without any change on the bandwidth requirement of jobs, the system will reject jobs. The second condition implies that the network could accommodate all jobs without losses (with high probability) if all of them would require the reduced bit rate $A_1 = 1$ and service rate μ_1.

It should be noted that, from the point of view of the design of algorithms, the constant c_0 has to be defined. If we take $c_0 \in (\Lambda/\mu_1, c)$ then the following hold:

(R_1) $\langle \rho, A\rangle > c_0$,

(R_2) $\Lambda/\mu_1 < c_0$.

If $\langle A, \rho\rangle < c$, it is not difficult to see that the system is equivalent to a classical underloaded loss network with one link and multiple classes of jobs. There is, of course, no need to use downgrading policies since the system can accommodate incoming requests without any loss when N is large; see, for example, [14] or [15, Chapter 6, Section 7].

3. Scaling results

In this section we prove convergence results when the scaling parameter N goes to ∞. These results are obtained by studying the asymptotic behavior of the occupation of the link around C_0^N,

$$m^N(t) = \langle A, L^N(t)\rangle - C_0^N. \tag{2}$$

In the context of loss networks, the analogue of such a quantity is the number of empty places. The following proposition shows that, for the downgrading policy, the boundary C^N does not play a role after some time if condition (R_2) holds.

Proposition 1. *Under condition (R_2) and if the initial state is such that*

$$\lim_{N\to+\infty}\left(\frac{L_j^N(0)}{N}\right) = \ell(0) = (\ell_{j,0}) \in \mathcal{S} := \{x \in \mathbb{R}_+^J : \langle A, x\rangle < c\},$$

then, for $\varepsilon > 0$, there exists $t_\varepsilon \ge 0$ such that, for $T > t_\varepsilon$,

$$\lim_{N\to+\infty} \mathbb{P}\Big(\sup_{t_\varepsilon \le t \le T} \langle A, L^N(t)\rangle < (c_0 + \varepsilon)N\Big) = 1.$$

Proof. Define

$$(\tilde{L}_j^N(t)) := (D_1^N(t) + X^N(t), D_2^N(t), \ldots, D_J^N(t)),$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

where $(X^N(t))$ is the process of the number of jobs of an independent M/M/∞ queue with $X^N(0) = 0$, service rate μ_1, and arrival rate $\Lambda = \lambda_1 + \cdots + \lambda_J$, and, for $1 \le j \le J$,

$$D_j^N(t) = \sum_{k=1}^{L_j^N(0)} \mathbf{1}\{E_{\mu_j,k} > t\},$$

where $(E_{\mu_j,k})$ is a sequence of independent and identically distributed (i.i.d.) exponentially distributed random variables with rate μ_j and where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The quantity $D_j^N(t)$ is the number of initial class j jobs still present at time t. Using Theorem 6.13 of [15], we obtain the convergence in distribution

$$\lim_{N\to+\infty}\left(\frac{X^N(t)}{N}\right) = \left(\frac{\Lambda}{\mu_1}(1-\mathrm{e}^{-\mu_1 t})\right),$$

and, consequently,

$$\lim_{N\to+\infty}\left(\frac{1}{N}\langle A, \tilde{L}^N(t)\rangle\right) = \left(\frac{\Lambda}{\mu_1}(1-\mathrm{e}^{-\mu_1 t}) + \sum_{j=1}^{J} A_j \ell_{j,0}\mathrm{e}^{-\mu_j t}\right). \tag{3}$$

Since $\mu_1 \le \mu_j$ for $1 \le j \le J$,

$$\begin{aligned}\frac{\Lambda}{\mu_1}(1-\mathrm{e}^{-\mu_1 t}) + \sum_{j=1}^{J} A_j \ell_{j,0}\mathrm{e}^{-\mu_j t} &\le \frac{\Lambda}{\mu_1}(1-\mathrm{e}^{-\mu_1 t}) + \mathrm{e}^{-\mu_1 t}\langle A, \ell(0)\rangle \\ &\le \max(c_0, \langle A, \ell(0)\rangle)\end{aligned}$$

by condition (R_2). Note that the asymptotic occupancy, when N is large, remains below the initial occupancy.

If $0 < \varepsilon N < C^N - C_0^N$ and $L^N(0) \in \mathbb{N}^J$ such that $C_0^N + \varepsilon N < \langle A, L^N(0)\rangle < C^N$, let

$$\tau_N = \inf\left\{t > 0 \colon \langle A, L^N(t)\rangle \le C_0^N + \tfrac{1}{2}\varepsilon N\right\},$$

then, on the event $\{\tau_N > T\}$, the downgrading policy yields that the identity in distribution

$$((L_j^N(t)), 0 \le t \le T) \overset{\mathrm{D}}{=} ((\tilde{L}_j^N(t)), 0 \le t \le T) \tag{4}$$

holds. From condition (R_2) we have the existence of t_ε such that

$$\frac{\Lambda}{\mu_1}(1-\mathrm{e}^{-\mu_1 t_\varepsilon}) + \sum_{j=1}^{J} A_j \ell_{j,0}\mathrm{e}^{-\mu_j t_\varepsilon} = c_0 + \frac{\varepsilon}{2}.$$

From convergence (3) we see that the sequence (τ_N) converges in distribution to t_ε.

Note that, if $S \in (t_\varepsilon, T)$, as long as the process $(\langle A, L^N(t)\rangle)$ stays above C_0^N on $I = [t_\varepsilon, S)$, a relation similar to (4) holds. Using again convergence (3), we see that, as N goes to ∞, the process $(\langle A, L^N(t)\rangle/N)$ remains below $c_0 + \varepsilon$ with probability close to 1 on I. The proposition is proved. □

We are now ready to investigate the asymptotic behavior of the process $(m^N(t))$ defined by relation (2). The variable indicates if the network is operating in saturation at time t, $m^N(t) \ge 0$, or not, $m^N(t) < 0$. In pure loss networks, when N is large, up to a change of time scale, the analogue of this process, the process of the number of empty places converges to a reflected

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

random walk in $\mathbb{N}$. In our case, as it will be seen, the corresponding process is in fact a random walk on $\mathbb{Z}$.

Definition 2. For $\ell = (\ell_j) \in \mathcal{S}$, let $(m_\ell(t))$ be the Markov process on $\mathbb{Z}$ whose Q-matrix Q_ℓ is defined by, for $x \in \mathbb{Z}$ and $1 \le j \le J$,

$$\begin{gathered} Q_\ell(x, x - A_j) = \mu_j \ell_j, \qquad Q_\ell(x, x + A_j) = \lambda_j \quad \text{if } x < 0, \\ Q_\ell(x, x + 1) = \Lambda \quad \text{if } x \ge 0, \end{gathered} \tag{5}$$

with $\Lambda := \lambda_1 + \lambda_2 + \cdots + \lambda_J$.

In the following proposition we summarize the stability properties of the Markov process $(m_\ell(t))$.

Proposition 2. *If $\ell = (\ell_j) \in \mathcal{S}$ then the Markov process $(m_\ell(t))$ is ergodic if $\ell \in \Delta_0$ with*

$$\Delta_0 := \left\{ x \in \mathcal{S} \colon \langle A, x \rangle = c_0, \sum_{j=1}^{J} (\lambda_j - \mu_j x_j) A_j > 0 \text{ and } \Lambda < \sum_{j=1}^{J} \mu_j x_j A_j \right\}, \tag{6}$$

where π_ℓ denotes the corresponding invariant distribution.

Proof. The Markov process $(m_\ell(t))$ on $\mathbb{Z}$ behaves like a random walk on each of the two half-lines $\mathbb{N}$ and $\mathbb{Z}_-^*$. Definition (6) implies that if $\ell \in \Delta_0$ then the drift of the random walk is positive when in $\mathbb{Z}_-^*$ and negative when in $\mathbb{N}$. This property implies the ergodicity of the Markov process by using, for example, the Lyapunov function $F(x) = |x|$; see, for example, [15, Corollary 8.7]. □

We now extend the expression π_ℓ for the values $\ell \in \mathcal{S} \setminus \Delta_0$. This will be helpful in order to describe the asymptotic dynamic of the system. See Theorem 1 for further details.

Definition 3. We denote $\pi_\ell = \delta_{-\infty}$, the Dirac measure at $-\infty$ when $\ell \in \Delta_-$, with

$$\Delta_- := \left\{ x \in \mathcal{S} \colon \langle A, x \rangle = c_0, \sum_{j=1}^{J} (\lambda_j - \mu_j x_j) A_j \le 0 \right\} \cup \{ x \in \mathcal{S} \colon \langle A, x \rangle < c_0 \},$$

and $\pi_\ell = \delta_{+\infty}$ if $\ell \in \Delta_+$, with

$$\Delta_+ := \left\{ x \in \mathcal{S} \colon \langle A, x \rangle = c_0, \sum_{j=1}^{J} \mu_j x_j A_j \le \Lambda \right\} \cup \{ x \in \mathcal{S} \colon \langle A, x \rangle > c_0 \}.$$

3.1. Stochastic evolution equations

For $\xi > 0$, denote by $\mathcal{N}_\xi(\mathrm{d}t)$ a Poisson process on $\mathbb{R}_+$ with rate ξ and $(\mathcal{N}_{\xi,i}(\mathrm{d}t))$ an i.i.d. sequence of such processes. All Poisson processes are assumed to be independent. Classically, the process $(L^N(t))$ can be seen as the unique solution to the following stochastic differential

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

equations:

$$\mathrm{d}L_1^N(t) = -\sum_{k=1}^{L_1^N(t-)} \mathcal{N}_{\mu_1,k}(\mathrm{d}t) + \mathbf{1}\{m^N(t-) < C^N - C_0^N\}\mathcal{N}_{\lambda_1 N}(\mathrm{d}t) + \sum_{j=2}^{J} \mathbf{1}\{0 \le m^N(t-) < C^N - C_0^N\}\mathcal{N}_{\lambda_j N}(\mathrm{d}t), \tag{7a}$$

$$\mathrm{d}L_j^N(t) = -\sum_{k=1}^{L_j^N(t-)} \mathcal{N}_{\mu_j,k}(\mathrm{d}t) + \mathbf{1}\{m^N(t-) < 0\}\mathcal{N}_{\lambda_j N}(\mathrm{d}t), \qquad 2 \le j \le J, \tag{7b}$$

with initial condition $(L_j^N(0)) \in \mathbb{N}^J$ such that $\langle A, L^N(0)\rangle \le C^N$.

Theorem 1. (Limiting dynamical system.) *Under condition (R_2), if the initial conditions are such that $m^N(0) = m \in \mathbb{Z}$ and*

$$\lim_{N\to+\infty}\left(\frac{L_j^N(0)}{N}\right) = (\ell_j(0)) \in \mathcal{S},$$

then there exists a continuous process $(\ell(t)) = (\ell_j(t))$ such that the convergence in distribution

$$\lim_{N\to+\infty}\left(\left(\frac{L_j^N(t)}{N}\right), \int_0^t f(m^N(u))\,\mathrm{d}u\right) = \left((\ell_j(t)), \int_0^t \int_{\mathbb{Z}} f(x)\pi_{\ell(u)}(\mathrm{d}x)\,\mathrm{d}u\right) \tag{8}$$

holds for any function f with finite support on $\mathbb{Z}$. Furthermore, there exists $t_0 > 0$ such that $(\ell(t),\ t \ge t_0)$ satisfies the differential equations

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\ell_1(t) &= -\mu_1\ell_1(t) + \lambda_1 + \pi_{\ell(t)}(\mathbb{N})\left(\sum_{k=2}^{J}\lambda_k\right), \\ \frac{\mathrm{d}}{\mathrm{d}t}\ell_j(t) &= -\mu_j\ell_j(t) + \lambda_j\pi_{\ell(t)}(\mathbb{Z}_-^*), \qquad 2 \le j \le J, \end{aligned} \tag{9}$$

where π_ℓ, for $\ell \in \mathcal{S}$, is the distribution of Proposition 2 and Definition 3.

It should be noted that, since the convergence holds for the convergence in distribution of processes, the limit $(\ell(t))$ is *a priori* a *random* process.

Proof. Using the same method as Hunt and Kurtz [12], we obtain the analogue of their Theorem 3. Fix $\varepsilon > 0$ such that $c_0 + \varepsilon < c$, from Proposition 1, we obtain the existence of t_0 such that

$$\lim_{N\to+\infty} \mathbb{P}\Big(\sup_{t_0\le t\le T} \langle A, L^N(t)\rangle < (c_0+\varepsilon)N\Big) = 1,$$

which implies that the boundary condition $m^N(t) < C^N - C_0^N$ in the evolution equations (7a) and (7b) can be removed. Consequently, only the boundary condition of $(m^N(t))$ at 0 plays a role, which yields relation (9) as in Hunt and Kurtz [12]. Note that, contrary to the general situation described in this reference, we have indeed a convergence in distribution because, for any $\ell \in \mathcal{S}$, $(m_\ell(t))$ has exactly one invariant distribution (which may be a Dirac mass at ∞) by Proposition 2; see [12, Conjecture 5]. □

In the following proposition we provide a characterization of the equilibrium point of the dynamical system $(\ell(t))$.

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

Proposition 3. (Fixed point.) *Under conditions (R_1) and (R_2), there exists a unique equilibrium point $\ell^* \in \Delta_0$ of the process $(\ell_j(t))$ defined by (8) given by*

$$\ell_1^* = c_0 - \pi^-(\rho_2 A_2 + \cdots + \rho_J A_J), \qquad \ell_j^* = \rho_j \pi^-, \quad 2 \le j \le J, \tag{10}$$

where

$$\pi^- := \frac{c_0 - \Lambda/\mu_1}{\langle A, \rho\rangle - \Lambda/\mu_1}, \tag{11}$$

with $\Lambda = \lambda_1 + \cdots + \lambda_J$. The process $(m_{\ell^*}(t))$ is ergodic in this case.

Proof. Assume that there exists an equilibrium point $\ell^* = (\ell_j^*)$ of $(\ell_j(t))$ defined by (8), it is also an equilibrium point of the dynamical system defined by (9), then

$$\mu_1 \ell_1^* = \lambda_1 + (\lambda_2 + \cdots + \lambda_J)(1 - \pi^-), \qquad \mu_j \ell_j^* = \lambda_j \pi^-, \quad 2 \le j \le J, \tag{12}$$

with $\pi^- = \pi_{\ell^*}(\mathbb{Z}_-^*)$. We obtain

$$\sum_{j=1}^J \lambda_j = \sum_{j=1}^J \mu_j \ell_j^* < \sum_{j=1}^J \mu_j \ell_j^* A_j = \pi^- \sum_{j=1}^J \lambda_j A_j + (1 - \pi^-) \sum_{j=1}^J \lambda_j < \sum_{j=1}^J \lambda_j A_j. \tag{13}$$

We now show that the vector ℓ^* is on the boundary, i.e.

$$\sum_{j=1}^J A_j \ell_j^* = c_0. \tag{14}$$

If we assume that

$$\lim_{N \to +\infty} \left(\frac{L_j^N(0)}{N} \right) = (\ell_j^*),$$

from Theorem 1 and the definition of $(m^N(t))$, we know that, for the convergence of processes, the following relation holds:

$$\lim_{N \to +\infty} \left(\frac{m^N(t)}{N} \right) = (\kappa_0), \quad \text{with } \kappa_0 := \sum_{j=1}^J A_j \ell_j^* - c_0.$$

For $N_0 \in \mathbb{N}$, $\varepsilon > 0$, and $N \ge N_0$,

$$\int_0^1 \mathbf{1}\{|m^N(u)| \ge \varepsilon N\} \, \mathrm{d}u \le \int_0^1 \mathbf{1}\{|m^N(u)| \ge \varepsilon N_0\} \, \mathrm{d}u.$$

Using again Theorem 1 and the fact that ℓ^* is an equilibrium point of the dynamical system, we have, for the convergence in distribution,

$$\lim_{N \to +\infty} \int_0^1 \mathbf{1}\{|m^N(u)| \le \varepsilon N_0\} \, \mathrm{d}u = \pi_{\ell^*}([-\varepsilon N_0, \varepsilon N_0]).$$

The left-hand side of the above expression can be arbitrarily close to 1 when N_0 is large. By convergence of the sequence $(m^N(t)/N)$ to (κ_0), it follows that, for the convergence in distribution, the relation

$$\lim_{N \to +\infty} \int_0^1 \mathbf{1}\{|m^N(u)|/N \ge \varepsilon\} \, \mathrm{d}u = 0$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

holds for $\varepsilon > 0$, which implies that $\kappa_0 = 0$. Thus, relation (14) holds. Finally, from (12) and (14), we obtain (10). We conclude, therefore, that $\ell^* \in \Delta_0$, the associated process $(m_{\ell^*}(t))$, is necessarily ergodic by Proposition 2 and (13).

To prove that the ℓ^* defined by (10) and (11) is indeed an equilibrium point of the dynamical system defined by (9), we need to show that the right-hand side of (11) is indeed equal to $\pi_{\ell^*}(\mathbb{Z}_-^*)$. This is proved in Proposition 5 of Section 4. □

3.2. Convergence of invariant distributions

In this section we use our main result to establish the convergence of the invariant distribution of the process $(m^N(t))$ as N gets large. This will yield, in particular, the convergence with respect to N of the probability of not downgrading a request at equilibrium.

Lemma 1. *If the process $(\tilde{L}_j^N(t))$ is the process $(L_j^N(t))$ at equilibrium then, for any $\varepsilon > 0$ and $T > 0$,*

$$\lim_{N\to+\infty} \mathbb{P}\left(\sup_{0\le t\le T} \sup_{2\le j\le J} \frac{\tilde{L}_j^N(t)}{N} \le \rho_j + \varepsilon \right) = 1.$$

Proof. Let $(L_j^N(t))$ be the process with initial state empty, then we can easily construct a coupling such that

$$L_j^N(t) \le \tilde{Q}_j^N(t), \qquad t \ge 0,\ 2 \le j \le J,$$

holds almost surely, where $(Q_j^N(t))$ is the M/M/∞ queue associated to class j requests. We deduce that

$$\tilde{L}_j^N(0) \le_{\mathrm{st}} \tilde{Q}_j^N(0),$$

where $\tilde{Q}_j^N(0)$ is a Poisson random variable with parameter $\rho_j N$ and '$\le_{\mathrm{st}}$' is the stochastic ordering of random variables. We can, therefore, construct another coupling such that

$$\tilde{L}_j^N(t) \le \tilde{Q}_j^N(t), \quad t \ge 0,\ 2 \le j \le J,$$

where $(\tilde{Q}_j^N(t))$ is a stationary version of the M/M/∞ queue associated to class j requests. The lemma is then a consequence of the following convergence in distribution of processes:

$$\lim_{N\to+\infty} \left(\frac{\tilde{Q}_j^N(t)}{N} \right) = (\rho_j) \quad \text{for } 2 \le j \le J;$$

see, for example, Theorem 6.13 of [15, p. 159]. □

Definition 4. Let $(y(t))$ be the dynamical system on $\mathcal{S}$ satisfying

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} y_1(t) &= -\mu_1 y_1(t) + \lambda_1 + \left(\sum_{k=2}^{J} \lambda_k \right) \frac{1}{\Lambda_A} \sum_{k=1}^{J} A_k(\lambda_k - \mu_k y_k(t)), \\ \frac{\mathrm{d}}{\mathrm{d}t} y_j(t) &= -\mu_j y_j(t) + \lambda_j \frac{1}{\Lambda_A} \sum_{k=1}^{J} (A_k \mu_k y_k(t) - \lambda_k), \qquad 2 \le j \le J, \end{aligned} \tag{15}$$

with

$$\Lambda_A = \sum_{k=1}^{J} \lambda_k (A_k - 1).$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

Lemma 2. *If $y(0) \in \Delta_0$ and if there exists an instant $T > 0$ such that $y(t) \in \Delta_0$ for $t \in [0, T]$, then $(y(t))$ and $(\ell(t))$ coincide on the time interval $[0, T]$, where $(\ell(t))$ is the solution of (9) with $\ell(0) = y(0)$.*

Proof. The proposition is a simple consequence of the representation (9) of the differential equations defining the dynamical system $(\ell(t))$ and of the explicit expression of the quantity $\pi_\ell(\mathbb{Z}_-^*)$ given by (22) when $\ell \in \Delta_0$; see (6). □

In the next proposition we investigate the stability properties of $(y(t))$.

Proposition 4. *Let H_0 be the hyperplane*

$$H_0 := \{z \in \mathcal{S} : \langle A, z\rangle = c_0\}.$$

If $y(0) \in H_0$ then $y(t) \in H_0$ for all $t \geq 0$ and $(y(t))$ is converging exponentially fast to ℓ^* defined in Proposition 3.

Proof. It is easily checked that

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle A, y(t)\rangle = 0,$$

so that, if $y(0) \in H_0$, then the function $t \mapsto \langle A, y(t)\rangle$ is constant and equal to c_0, hence, $y(t) \in H_0$ for all $t \geq 0$.

For $2 \leq j \leq J$,

$$\frac{\mathrm{d}}{\mathrm{d}t} y_j(t) = \lambda_j b_0 - \mu_j y_j(t) + \lambda_j \sum_{k=2}^{J} b_k y_k(t),$$

with

$$b_0 = \frac{\mu_1 c_0 - \Lambda}{\Lambda_A} \quad \text{and} \quad b_j = \frac{A_j(\mu_j - \mu_1)}{\Lambda_A}.$$

In matrix form, if $z(t) = (y_2(t), \ldots, y_J(t))$, it can be expressed as

$$\frac{\mathrm{d}}{\mathrm{d}t} z(t) = e_b + Bz(t), \tag{16}$$

with $e_b = b_0(\lambda_2, \ldots, \lambda_J) \in \mathbb{R}^{J-1}$ and $B = (B_{jk}, 2 \leq j, k \leq J)$ with

$$B_{jk} = \lambda_j b_k - \mu_j \mathbf{1}\{k = j\}.$$

If $v = (v_2, \ldots, v_J)$ is an eigenvector for the eigenvalue x of B, then

$$(x + \mu_j)v_j = \lambda_j \sum_{k=2}^{J} b_k v_k, \qquad 2 \leq j \leq J,$$

hence, x is an eigenvalue if and only if it is a solution of

$$F(x) := \sum_{j=2}^{J} \frac{b_j \lambda_j}{x + \mu_j} = 1.$$

If L is the number of distinct values of μ_j, $2 \leq j \leq J$, such that $\mu_j \neq \mu_1$, then the above equation shows that an eigenvalue is a zero of a polynomial of degree at most L. Using condition (R),

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

it is easy to check that the relation $F(0) < 1$ holds. In particular, 0 is not an eigenvalue and, consequently, B is invertible. Due to the poles of F at the $-\mu_j$, $2 \le j \le J$, and the relations $F(0) < 1$ and $\mu_j \ge \mu_1$ for $2 \le j \le J$, we have already L negative solutions of the equation $F(x) = 1$. All eigenvalues of B are thus negative, consequently, $\exp(tB)$ converges to 0; see, for example, [2, Corollary 2, Chapter 25].)

Equation (16) can be solved as

$$z(t) = \mathrm{e}^{tB}(z(0) + B^{-1}e_b) - B^{-1}e_b.$$

Therefore, the function $(z(t))$ has a limit at ∞ given by $-B^{-1}e_b$, which is clearly $(\ell_j^*, 2 \le j \le J)$. The proposition is proved. □

We are now ready to prove the main result of this section.

Theorem 2. *If ℓ^* is the quantity defined in Proposition 3 then the equilibrium distribution of $(m^N(t))$ converges to π_{ℓ^*} when N goes to ∞.*

Proof. Recall that $m^N(t) = \langle A, L^N(t)\rangle - C_0^N$ and let Π^N be the invariant distribution of $(L^N(t))$. It is assumed that the distribution of $L^N(0)$ is Π^N for the rest of the proof. In particular, $(m^N(t))$ is a stationary process.

We first prove that $(L^N(0)/N)$ converges in distribution to ℓ^*. The boundary condition $\langle A, L^N(0)\rangle \le C^N$ yields that the sequence of random variables $(L^N(0)/N)$ is tight. If $(L^{N_k}(0)/N_k)$ is a convergent subsequence to some random variable ℓ^∞, by Theorem 1, it follows that, for the convergence in distribution,

$$\lim_{k\to+\infty}\left(\left(\frac{L^{N_k}(t)}{N_k}\right)\right) = (\ell(t))$$

holds, where $(\ell(t))$ is a solution of (9) with initial point at $\ell(0) = \ell^\infty$. Note that $(\ell(t))$ is a stationary process, its distribution is invariant under any time shift.

By Lemma 1, it follows that the relation $\ell_j(t) \le \rho_j$, for $2 \le j \le J$, holds almost surely on any finite-time interval and, by Proposition 1, $\langle A, \ell(t)\rangle \le c_0$ also holds almost surely on finite-time intervals.

Assume that $\langle A, \ell(0)\rangle < c_0$ holds. The ordinary differential equations defining the limiting dynamical system are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell_j(t) = -\mu_j\ell_j(t) + \lambda_j, \qquad 1 \le j \le J,$$

as long as the condition $\langle A, \ell(t)\rangle < c_0$ holds, hence, on the corresponding time interval, we have

$$\ell_j(t) = \rho_j + (\ell_j(0) - \rho_j)\mathrm{e}^{-\mu_j t}, \qquad 1 \le j \le J,$$

so that

$$\langle A, \ell(t)\rangle = \langle A, \rho\rangle + \sum_{j=1}^{J} A_j(\ell_j(0) - \rho_j)\mathrm{e}^{-\mu_j t}.$$

Since $\langle A, \rho\rangle > c_0$, there exists some $t_1 > 0$ such that $\langle A, \ell(t_1)\rangle = c_0$.

Hence, by the stationarity in distribution of $(\ell(t))$, we can shift time at t_0 and assume that $\langle A, \ell(0)\rangle = c_0$. On this event

$$\sum_{j=1}^{J}\mu_j\ell_j(0)A_j \ge \mu_1\sum_{j=1}^{J}\ell_j(0)A_j = \mu_1 c_0 > \Lambda = \sum_{j=1}^{J}\lambda_j. \tag{17}$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

Similarly, since $\ell_j(0) \le \rho_j$ for all $2 \le j \le J$,

$$\begin{aligned}\sum_{j=1}^{J} A_j(\lambda_j - \mu_j \ell_j(0)) &= \lambda_1 - \mu_1 c_0 + \mu_1 \sum_{j=2}^{J} A_j \ell_j(0) + \sum_{j=2}^{J} A_j(\lambda_j - \mu_j \ell_j(0)) \qquad (18)\\ &= -\mu_1 c_0 + \sum_{j=1}^{J} A_j(\lambda_j + (\mu_1 - \mu_j)\ell_j(0))\\ &\ge -\mu_1 c_0 + \sum_{j=1}^{J} A_j(\lambda_j + (\mu_1 - \mu_j)\rho_j)\\ &= -\mu_1 c_0 + \sum_{j=1}^{J} A_j \lambda_j \frac{\mu_1}{\mu_j}\\ &= \mu_1(\langle A, \rho\rangle - c_0)\\ &> 0,\end{aligned}$$

and the last quantity is independent of $\ell(0)$. From (17) and (18) we see that $\ell(0) \in \Delta_0$ and, by (9) and (15), they also hold for t in a small neighborhood I of 0 independent of $\ell(0)$ so that $\ell(t) \in \Delta_0$ for $t \in I$. Consequently, the dynamical system $(\ell(t))$ never leaves Δ_0. From Lemma 2 we see that the two dynamical systems $(\ell(t))$ and $(y(t))$ (with $y(0) = \ell(0)$) coincide. Hence, on the one hand, $(\ell(t))$ is a stationary process and, on the other hand, it is a dynamical system converging to ℓ^*, we deduce that it is constant and equal to ℓ^*. We have thus proved that the sequence $(L^N(0)/N)$ converges in distribution to ℓ^*.

Using again Theorem 1, it follows that, for the convergence in distribution,

$$\lim_{N\to+\infty} \int_0^1 f(m^N(u))\,\mathrm{d}u = \int_{\mathbb{Z}} f(x)\pi_{\ell^*}(\mathrm{d}x)$$

holds for any function f with finite support on $\mathbb{Z}$. Using the stationarity of $(m^N(t))$ and Lebesgue's Theorem, we obtain

$$\lim_{N\to+\infty} \mathbb{E}(f(m^N(0))) = \int_{\mathbb{Z}} f(x)\pi_{\ell^*}(\mathrm{d}x).$$

The theorem is proved. □

Since a job arriving at time t is not downgraded if $m^N(t) < 0$, we have the following corollary.

Corollary 1. *As N goes to ∞, the probability that, at equilibrium, a job is not downgraded in this allocation scheme but is converging to π^-, as defined in (10), is*

$$\pi^- = \frac{c_0 - \Lambda/\mu_1}{\langle A, \rho\rangle - \Lambda/\mu_1}.$$

4. Invariant distribution

We assume in this section that $\ell \in \Delta_0$, as defined in Proposition 2, so that $(m_\ell(t))$ is an ergodic Markov process. The goal of this section is to derive an explicit expression of the

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

invariant distribution π_ℓ on $\mathbb{Z}$ of $(m_\ell(t))$. At the same time, Proposition 5 below gives the required argument to complete the proof of Proposition 3 on the characterization of the fixed point of the dynamical system.

4.1. Functional equation

In the following we denote by Y_ℓ a random variable with distribution $\pi_\ell = (\pi_\ell(n),\ n \in \mathbb{Z})$. For $r > 0$, we will use the notation

$$D(r) = \{z \in \mathbb{C},\ |z| < r\}, \qquad D^c(r) = \{z \in \mathbb{C},\ |z| > r\}, \qquad \gamma(r) = \{z \in \mathbb{C},\ |z| = r\}.$$

For the sake of simplicity, we will use $D = D(1)$ and $D^c = D^c(1)$.

Lemma 3. *With the notation*

$$\varphi_+(z) = \mathbb{E}(z^{Y_\ell}\mathbf{1}\{Y_\ell \geq 0\}), \qquad \varphi_-(z) = \mathbb{E}(z^{Y_\ell}\mathbf{1}\{Y_\ell < 0\}),$$

the random variable Y_ℓ is such that

$$P_1(z)\varphi_+(z) = P_2(z)\varphi_-(z), \tag{19}$$

where P_1 and P_2 are polynomials defined by

$$\begin{aligned} P_1(z) &= \sum_{j=1}^{J}[(\lambda_j + \mu_j\ell_j)z^{A_J} - \lambda_j z^{A_J+1} - \mu_j\ell_j z^{A_J-A_j}], \\ P_2(z) &= \sum_{j=1}^{J}[\lambda_j z^{A_J+A_j} + \mu_j\ell_j z^{A_J-A_j} - (\lambda_j + \mu_j\ell_j)z^{A_J}]. \end{aligned} \tag{20}$$

Proof. For $z \in \gamma(1)$, define $f_z\colon \mathbb{Z} \mapsto \mathbb{C}$ such that $f_z(x) = z^x$ for $x \in \mathbb{Z}$. Equilibrium equations for $(m_\ell(t))$ yield the identity

$$\sum_{x,y\in\mathbb{Z},\, x\neq y} \pi_\ell(x)Q_\ell(x,y)(f_z(y) - f_z(x)) = 0,$$

where Q_ℓ is the Q-matrix of $(m_\ell(t))$ given by (5). After some simple reordering, we have

$$\begin{aligned} &\mathbb{E}(z^{Y_\ell}\mathbf{1}\{Y_l \geq 0\}) \sum_{j=1}^{J}(\lambda_j(1-z) + \mu_j\ell_j(1 - z^{-A_j})) \\ &\quad = -\mathbb{E}(z^{Y_l}\mathbf{1}\{Y_l < 0\}) \sum_{j=1}^{J}(\lambda_j(1 - z^{A_j}) + \mu_j\ell_j(1 - z^{-A_j})). \end{aligned} \tag{21}$$

Using the definition of $\varphi_+(z)$ and $\varphi_-(z)$, (21) can be rewritten as (19). □

Proposition 5. *If $\ell \in \Delta_0$ then*

$$\pi_\ell(\mathbb{Z}_-^*) = \frac{\sum_{j=1}^{J}(A_j\mu_j\ell_j - \lambda_j)}{\sum_{j=1}^{J}\lambda_j(A_j - 1)}. \tag{22}$$

In particular, if $\ell^* \in \mathcal{S}$ is given by (10) then

$$\pi_{\ell^*}(\mathbb{Z}_-^*) = \frac{c_0 - \Lambda/\mu_1}{\langle A, \rho\rangle - \Lambda/\mu_1}.$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

Note that the right-hand side of this equation is precisely π^- of (11), which is the result necessary to complete the proof of Proposition 3.

Proof. With the same notation as before, from (19),

$$\frac{\varphi_-(z)}{\varphi_+(z)} = \frac{P_1(z)}{P_2(z)}$$

holds for $z \in \mathbb{C}$, with $z \in \gamma(1)$. By definition of $\varphi_-(z)$ and $\varphi_+(z)$,

$$\lim_{z\to 1} \varphi_-(z) = \pi_\ell(\mathbb{Z}_-^*) \quad \text{and} \quad \lim_{z\to 1} \varphi_+(z) = 1 - \pi_\ell(\mathbb{Z}_-^*).$$

Since 1 is a zero of P_1 and P_2, this yields

$$\frac{\pi_\ell(\mathbb{Z}_-^*)}{1-\pi_\ell(\mathbb{Z}_-^*)} = \frac{P_1'(1)}{P_2'(1)} = \frac{\sum_{j=1}^J (A_j\mu_j\ell_j - \lambda_j)}{\sum_{j=1}^J A_j(\lambda_j - \mu_j\ell_j)}.$$

Using the expression of (ℓ_j^*), with some algebra, we obtain

$$\pi_{\ell^*}(\mathbb{Z}_-^*) = \left(c_0 - \sum_{j=1}^J \frac{\lambda_j}{\mu_1}\right) \Big/ \left(\sum_{j=1}^J \rho_j A_j - \sum_{j=1}^J \frac{\lambda_j}{\mu_1}\right) = \pi^-.$$

The proposition is proved. □

Relation (19) is valid on the unit circle, however, the function φ_+ (respectively, φ_-) is defined on D (respectively, D^c). This can then be expressed as a Wiener–Hopf factorization problem analogous to the one used in the analysis of reflected random walks on $\mathbb{N}$. This is used in the analysis of the GI/GI/1 queue; see, for example, [3, Chapter VIII] or [15, Chapter 3]. In a functional context, this is a special case of the Riemann problem; see [9]. In our case, this is a random walk in $\mathbb{Z}$, with a drift depending on the half-space where it is located. The first (respectively, second) condition in the definition of the set Δ_0 in definition (6) implies that the drift of the random walk in $\mathbb{Z}_-^*$ (respectively, in $\mathbb{N}$) is positive (respectively, negative).

The first step in the analysis of (19) is to determine the locations of the zeros of P_1 and P_2. This is the purpose of the following lemma.

Lemma 4. (Location of the zeros of P_1 and P_2.) *Let ℓ be in Δ_0.*

(i) *Polynomial P_2 has exactly two positive real roots* 1 *and* $z_2 \in]0, 1[$. *There are $A_J - 1$ roots in $D(z_2)$ and $A_J - 1$ roots whose modulus are strictly greater than* 1.

(ii) *Polynomial P_1 has exactly two positive real roots* 1 *and* $z_1 > 1$. *The $A_J - 1$ remaining roots have a modulus strictly smaller than* 1.

Proof. First note that P_2 is a polynomial with the same form as the f defined in [4, Equation (13)] (with $e_j = A_j$, $\kappa_j = \lambda_j$, and $\hat{e} = A_J$). The roots of Q are exactly the roots of f. Lemma 4(i) follows from [4, Lemma 2.2].

The proof of assertion (ii) uses an adaptation of the argument for the proof of [4, Lemma 2.2]. Define the function $f(z) = z^{-A_J} P_1(z)$. Recall that P_1 is a polynomial with degree $A_J + 1$. There are exactly two real positive roots for P_1. Indeed, $f(1) = 0$ and it is easily checked that f is strictly concave with

$$f'(1) = \sum_{j=1}^J (-\lambda_j + A_j\mu_j\ell_j) > 0,$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

since $\ell \in \Delta_0$, by the second condition in definition (6). Hence, P_1 has a real zero z_1 greater than 1.

Let $r \in (1, z_1)$ be fixed, note that $P_1(r) > 0$. Define

$$f_1(z) = Kz^{A_J} \quad \text{with } K = \sum_{j=1}^{J}(\lambda_j + A_j\mu_j\ell_j), \qquad f_2(z) = \sum_{j=1}^{J}(\lambda_j z^{A_J+1} + \mu_j\ell_j z^{A_J-A_j}),$$

so that $P_1 = f_1 - f_2$.

Fix some $z \in \gamma(r)$. By expressing these functions in terms of real and imaginary parts,

$$z^{A_J} = \alpha_1 + \mathrm{i}\beta_1 \quad \text{and} \quad f_2(z) = \alpha_2 + \mathrm{i}\beta_2,$$

we obtain

$$\begin{aligned} |f_1(z) - f_2(z) - bz^{A_J}|^2 &= |K(\alpha_1 + \mathrm{i}\beta_1) - b(\alpha_1 + \mathrm{i}\beta_1) - (\alpha_2 + \mathrm{i}\beta_2)|^2 \\ &= (K\alpha_1 - \alpha_2)^2 + (K\beta_1 - \beta_2)^2 + H \\ &= |f_1(z) - f_2(z)|^2 + H, \end{aligned} \tag{23}$$

with

$$\begin{aligned} H &= (b\alpha_1)^2 - 2b\alpha_1(K\alpha_1 - \alpha_2) + (b\beta_1)^2 - 2b\beta_1(K\beta_1 - \beta_2) \\ &= b(b - 2K)(\alpha_1^2 + \beta_1^2) + 2b(\alpha_1\alpha_2 + \beta_1\beta_2). \end{aligned}$$

Using the Cauchy–Schwarz inequality, we obtain

$$\alpha_1\alpha_2 + \beta_1\beta_2 \le \frac{1}{K}|f_2(z)||f_1(z)| \le \frac{1}{K}f_2(r)f_1(r),$$

since $|f_i(z)| \le f_i(|z|)$ for $i = 1, 2$. Thus,

$$\begin{aligned} \frac{H}{b} &= (b - 2K)(\alpha_1^2 + \beta_1^2) + 2(\alpha_1\alpha_2 + \beta_1\beta_2) \\ &\le (b - 2K)\frac{f_1(r)^2}{K^2} + 2f_2(r)\frac{f_1(r)}{K} \\ &= \frac{f_1(r)}{K^2}((b - 2K)f_1(r) + 2Kf_2(r)) \\ &= \frac{f_1(r)}{K^2}(bf_1(r) - 2KP_1(r)). \end{aligned}$$

Since $P_1(r) > 0$, b can be chosen so that $bf_1(r) < 2KP_1(r)$. From the above relation and (23), it follows that, for $z \in \gamma(r)$,

$$|f_1(z) - f_2(z) - bz^{A_J}| < |f_1(z) - f_2(z)|$$

holds. By Rouché's theorem, it follows that, for any $r \in (1, z_1)$, P_1 has exactly A_J roots in $D(r)$. We conclude that P_1 has exactly A_J roots in $\bar{D}$. It is easily checked that if $z \in \gamma(1)$ and $z \notin \mathbb{R}$ then the real part of $P_1(z)$ is positive, hence, z cannot be a root of the polynomial P_1. Consequently, P_1 has exactly $A_J - 1$ roots in D. The lemma is proved. □

Definition 5. For $U \in \{P_1, P_2\}$, denote by $\mathcal{Z}_U$ the set of the zeros of U different from 1.

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

Define

$$\Phi(z) = \begin{cases} -\varphi_+(z)\lambda_J{}^{-1}(z - z_1) \displaystyle\prod_{q \in \mathcal{Z}_{P_2} \cap D^c} (z - q)^{-1}, & z \in D, \\ \varphi_-(z)\Lambda^{-1} \displaystyle\prod_{q \in \mathcal{Z}_{P_2} \cap D} (z - q) \prod_{p \in \mathcal{Z}_{P_1} \cap D} (z - p)^{-1}, & z \in D^c, \end{cases}$$

with $\Lambda = \lambda_1 + \cdots + \lambda_J$ and retaining the same notation as before. By definition, function Φ is holomorphic in D and D^c and, from (19), is continuous on $\gamma(1)$. The analytic continuation theorem (for example, [16, Theorem 16.8]) yields that Φ is holomorphic on $\mathbb{C}$. For $z \in D^c$,

$$|\varphi_-(z)| \le \mathbb{E}(\mathbf{1}\{Y_\ell < 0\}|z|^{Y_\ell}) \le \frac{1}{|z|},$$

since the cardinality of $\mathcal{Z}_{P_1} \cap D$ (respectively, $\mathcal{Z}_{P_2} \cap D$) is $A_J - 1$ (respectively, A_J), the holomorphic function Φ is, therefore, bounded on $\mathbb{C}$. By Liouville's theorem, Φ is constant and equal to $\kappa \in \mathbb{C}$. Therefore,

$$\begin{aligned} \varphi_+(z) &= -\kappa\lambda_J(z - z_1)^{-1} \prod_{q \in \mathcal{Z}_{P_2} \cap D^c} (z - q), && z \in D, \\ \varphi_-(z) &= \kappa\Lambda \prod_{q \in \mathcal{Z}_{P_2} \cap D} (z - q)^{-1} \prod_{p \in \mathcal{Z}_{P_1} \cap D} (z - p), && z \in D^c. \end{aligned} \tag{24}$$

Recall that $\varphi(z) = \varphi_+(z) + \varphi_-(z) = \mathbb{E}(z^{Y_\ell})$ is a generating function, in particular, $\varphi(1) = 1$. Substituting the previous expressions for φ_+ and φ_- into $\varphi_+(1) + \varphi_-(1) = 1$, we obtain

$$1 = -\kappa \prod_{q \in \mathcal{Z}_{P_2} \cap D} (1 - q)^{-1} \frac{1}{1 - z_1}(P_1'(1) + P_2'(1)),$$

hence, using (20),

$$\kappa = \frac{z_1 - 1}{\Lambda_A} \prod_{q \in \mathcal{Z}_{P_2} \cap D} (1 - q),$$

where Λ_A is introduced in Definition 4. Note that κ is positive. We can now state the main result of this section.

Proposition 6. (Invariant measure.) *If $\ell \in \Delta_0$ is defined by (6) then the invariant measure π_ℓ can be expressed, for $n \in \mathbb{Z}$, as*

$$\pi_\ell(n) = \begin{cases} -\kappa \displaystyle\sum_{q \in \mathcal{Z}_{P_2} \cap D} \frac{P_1(q)q^{-n-1}}{(q - z_1)(q - 1)R_D'(q)}, & n < 0, \\ \kappa\left(\alpha_n + \dfrac{P_2(z_1)z_1^{-n-1}}{(z_1 - 1)R_D(z_1)}\right), & 0 \le n < A_J - 1, \\ \kappa \dfrac{P_2(z_1)z_1^{-n-1}}{(z_1 - 1)R_D(z_1)}, & n \ge A_J - 1, \end{cases}$$

where z_1 is defined in Lemma 4, and P_1 and P_2 by (20),

$$R_D(z) = \prod_{q \in \mathcal{Z}_{P_2} \cap D} (z - q), \qquad \kappa = \frac{(z_1 - 1)R_D(1)}{\Lambda_A} \quad \textit{for } 0 \le n < A_J - 1,$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

and α_n is the coefficient of degree n of the polynomial

$$-\frac{1}{z-z_1}\left(\frac{P_2(z)}{(z-1)R_D(z)}-\frac{P_2(z_1)}{(z_1-1)R_D(z_1)}\right).$$

Proof. Note that, for $z \in \mathbb{C}$,

$$\prod_{p\in\mathcal{Z}_{P_1}\cap D}(z-p) = -\frac{1}{\Lambda}\frac{P_1(z)}{(z-z_1)(z-1)}.$$

For $z \in D^c$,

$$\varphi_-(z) = \kappa\Lambda \prod_{q\in\mathcal{Z}_{P_2}\cap D}(z-q)^{-1}\prod_{p\in\mathcal{Z}_{P_1}\cap D}(z-p).$$

Since $|\mathcal{Z}_{P_1}\cap D| = A_J - 1 < A_J = |\mathcal{Z}_{P_2}\cap D|$ by Lemma 4, φ_- has the following partial fraction decomposition:

$$\begin{aligned}\varphi_-(z) &= -\kappa\sum_{q\in\mathcal{Z}_{P_2}\cap D}\frac{P_1(q)}{(q-z_1)(q-1)R_D'(q)}\frac{1}{z-q}\\ &= \sum_{i=0}^{\infty}-\kappa\sum_{q\in\mathcal{Z}_{P_2}\cap D}\frac{P_1(q)q^i}{(q-z_1)(q-1)R_D'(q)}\frac{1}{z^{i+1}}.\end{aligned}$$

Denote

$$R_{D^c}(z) = \prod_{q\in\mathcal{Z}_{P_2}\cap D^c}(z-q) = \frac{P_2(z)}{\lambda_J(z-1)R_D(z)},$$

then

$$\varphi_+(z) = -\kappa\lambda_J\frac{R_{D^c}(z)}{z-z_1} = \kappa\left(-\lambda_J\frac{R_{D^c}(z)-R_{D^c}(z_1)}{z-z_1}+\frac{P_2(z_1)}{(1-z_1)R_D(z_1)}\frac{1}{z-z_1}\right).$$

We conclude by using the expression of κ obtained before. □

4.2. Some moments of (π_{ℓ^*})

Using the probability generating function $\varphi(z)$ of π_{ℓ^*} from (24), we can derive an explicit expression of the mean, the variance, and the skewness of such a distribution. The skewness of a random variable X is a measure of the asymmetry of the distribution of X,

$$\mathrm{skew}(X) := \mathbb{E}([X-\mathbb{E}(X)]^3);$$

see, for example, [6].

Proposition 7. *If Y_{ℓ^*} is a random variable with distribution π_{ℓ^*} then*

$$\begin{aligned}\mathbb{E}(Y_{\ell^*}) &= A_J + \frac{\theta_2}{2\theta_1} - S(1),\\ \mathrm{var}(Y_{\ell^*}) &= \frac{\theta_2+2\theta_3}{6\theta_1} - \left(\frac{\theta_2}{2\theta_1}\right)^2 - (S(1)+S'(1)),\\ \mathrm{skew}(Y_{\ell^*}) &= \frac{\theta_2^3}{4\theta_1^3} + \theta_2\frac{\theta_2-2\theta_3}{4\theta_1^2} + \frac{\theta_4-\theta_3}{4\theta_1} - (S(1)+3S'(1)+S''(1)),\end{aligned}$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

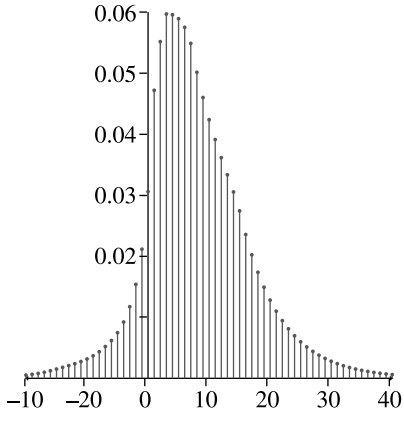

FIGURE 1: The histogram of Y_{ℓ^*} with the parameters $J = 5$, $A = (1, 2, 4, 8, 16)$, $\lambda = (0.25, 0.2, 0.15, 0.1, 0.05)$, $\mu = (1, 1, 1, 1, 1)$, and $c_0 = 0.97$.

where, for $i \geq 1$,

$$\theta_i = \sum_{j=2}^{J} \lambda_j A_j^{i-1}(A_j - 1),$$

and

$$S(z) = \frac{1}{z - z_1} + \sum_{q \in \mathcal{Z}_{P_2} \cap D} \frac{1}{z - q},$$

with $R_D(z)$ *defined in Proposition 6.*

The proof is straightforward, modulo some tedious calculations of the successive derivatives of $\varphi(z)$ evaluated at 1. In Figure 1, we see that the distribution of Y_{ℓ^*} is significantly asymmetrical; for this example, $\mathbb{E}(Y_{\ell^*}) = 8.048\,19$, $\text{var}(Y_{\ell^*}) = 77.2284$, and $\text{skew}(Y_{\ell^*}) = 0.967\,069$.

5. Applications

5.1. Comparison with a pure loss system

In this case, a request which cannot be accommodated is rejected right away. Recall that, with probability 1, our algorithm does not reject any request. The purpose of this section is to discuss the price of such a policy. Intuitively, at equilibrium the probability W_{L} of accepting a job at requested capacity in a pure loss system is greater that the corresponding quantity W_{D} for the downgrading algorithm; see Proposition 8 below. A further question is to assess the impact of such a policy, i.e. the order of magnitude of the difference $W_{\text{L}} - W_{\text{D}}$.

Under the same assumptions about the arrivals and under condition (R), with $\Lambda = \lambda_1 + \cdots + \lambda_J$, then, as N gets large, the equilibrium probability that a request of class $1 \leq j \leq J$ is accepted in the pure loss system is converging to β^{A_j}, where $\beta \in (0, 1)$ is the unique solution of the equation

$$\sum_{j=1}^{J} A_j \rho_j \beta^{A_j} = c; \tag{25}$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

see [13]. Consequently, the asymptotic load of accepted requests is given by

$$W_{\mathrm{L}} := \frac{1}{\Lambda} \sum_{j=1}^{J} \rho_j \beta^{A_j}.$$

Under the downgrading policy, the equilibrium probability that a job is accepted without degradation is given by π^-, the asymptotic load of requests accepted without degradation is

$$W_{\mathrm{D}} := \frac{1}{\Lambda} \sum_{j=1}^{J} \rho_j \frac{c_0 - \Lambda/\mu_1}{\langle \rho, A \rangle - \Lambda/\mu_1} \quad \text{for } c_0 \in (\Lambda/\mu_1, c).$$

Note that, when the service rates are constant and equal to 1 then W_{L} (respectively, W_{D}) is the asymptotic throughput of accepted requests (respectively, of nondegraded requests).

The following proposition establishes the intuitive property that a pure loss system has better performances in terms of acceptance.

Proposition 8. *For $c_0 \in (\Lambda/\mu_1, c)$, the relation $W_{\mathrm{D}} \le W_{\mathrm{L}}$ holds.*

Proof. The representation of these quantities yields that the relation needed for the proof is equivalent to

$$\sum_{j=1}^{J} \rho_j \beta^{A_j} \left(\sum_{j=1}^{J} \rho_j A_j - \sum_{j=1}^{J} \frac{\lambda_j}{\mu_1} \right) - \sum_{j=1}^{J} \rho_j \left(c_0 - \sum_{j=1}^{J} \frac{\lambda_j}{\mu_1} \right) \ge 0.$$

Using the fact that $c_0 < c$ and (25), it is enough to show that the quantity

$$\Delta := \sum_{j=1}^{J} \rho_j \beta^{A_j} \left(\sum_{i=1}^{J} \rho_j A_i - \sum_{i=1}^{J} \frac{\lambda_i}{\mu_1} \right) - \sum_{j=1}^{J} \rho_j \left(\sum_{i=1}^{J} A_i \rho_j \beta^{A_i} - \sum_{i=1}^{J} \frac{\lambda_i}{\mu_1} \right)$$

is positive. But this is clear since

$$\begin{aligned} \Delta &= \sum_{1 \le i,j \le J} \rho_i \rho_j (A_j (\beta^{A_i} - \beta^{A_j})) + \sum_{1 \le i,j \le J} \rho_j \frac{\lambda_i}{\mu_1} (1 - \beta^{A_j}) \\ &= \sum_{1 \le i < j \le J} \rho_i \rho_j ((A_j - A_i)(\beta^{A_i} - \beta^{A_j})) + \sum_{1 \le i,j \le J} \rho_j \frac{\lambda_i}{\mu_1} (1 - \beta^{A_j}) \end{aligned}$$

and the terms of both series of the right-hand side of this relation are nonnegative due to the fact that $0 < \beta < 1$. □

Numerical experiments have been performed to estimate the difference $W_{\mathrm{L}} - W_{\mathrm{D}}$; see Figure 2. The general conclusion is that, at moderate load under condition (R), the downgrading algorithm performs quite well with only a small fraction of downgraded jobs. As it can be seen this is no longer true for high load, where, as expected, most of the requests are downgraded but no requests are lost.

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

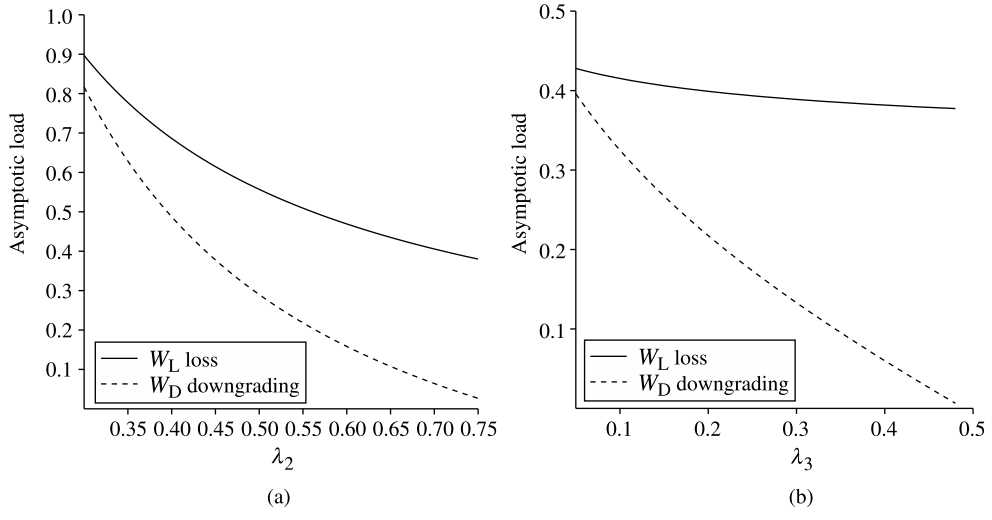

FIGURE 2: Asymptotic load of nondowngraded/accepted requests with $A_1 = 1$, $c = 1$, $c_0 = 0.99$, and all service rates equal to 1. (a) $J = 2$, $A_2 = 3$, $\lambda_1 = 0.2$, and (R) with condition $\lambda_2 \in (0.2633, 0.79)$. (b) $J = 3$, $A_2 = 2$, $\lambda_1 = \lambda_2 = 0.2$, and (R) with condition $\lambda_3 \in (0.03, 0.49)$.

5.2. Application to video transmission

We consider now a link with a large bandwidth, 10.0Gbps, in charge of video streaming. Requests that cannot be immediately served are lost. Video transmission is offered in two standard qualities, namely, low quality (LQ) and high quality (HQ). From Añorga *et al.* [1], the bandwidth requirement for YouTube's videos at 240p is 1485Kbps, and for 720p it is 2737.27Kbps.

Using the values above, after renormalization, we take $A_1 = 1$, $N = C^N = 7061$ and $A_2 = 2$, $c = 1$. Jobs arrive at rate λ_2 in this system asking for HQ transmission, but clients accept to watch the video in LQ. In particular, $\lambda_1 = 0$. Service times are assumed to be the same for both qualities and taken as the unity, $\mu_1 = \mu_2 = 1$. Condition (R) is satisfied when

$$0.5 < \lambda_2 < 1.$$

We define $C_0 = \alpha C$, with $0 < \alpha < 1$. The quantity α_ε is defined as the largest value of α such that the loss probability of a job is less than $\varepsilon > 0$. With the notation of Section 4, we write

$$\alpha_\varepsilon = \sup\{\alpha \in (0, 1) : \mathbb{P}(Y_{\ell^*} + C_0 > C) < \varepsilon\}.$$

Note that this is an approximation, since the variable Y_{ℓ^*} corresponds to the case when the scaling parameter N goes to ∞.

Using the explicit expression of the distribution of Y_{ℓ^*} of Proposition 6, in Figure 3 we plot the threshold α_ε that ensures a loss rate less than ε as a function of ε, for several values of λ_2. In the numerical example, taking $C_0 = 0.98C$ is sufficient to obtain a loss probability less than 10^{-7}.

Now let π_ε^- be the value of π^- defined by Corollary 1 for $C_0 = \alpha_\varepsilon C$. Recall that π_ε^-, given by (11), is the asymptotic equilibrium probability that a job is not downgraded, i.e.

$$\pi_\varepsilon^- = \frac{\alpha_\varepsilon}{\lambda_2} - 1.$$

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

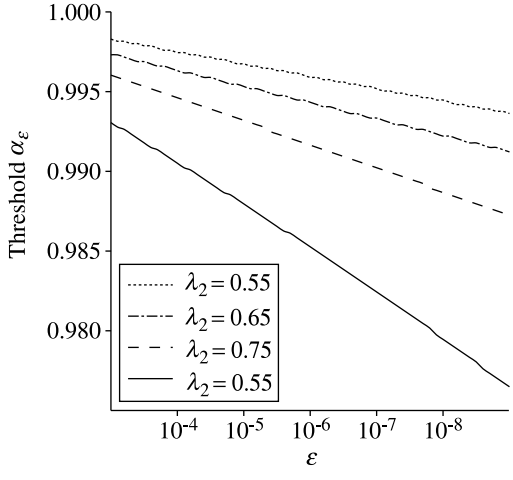

FIGURE 3: Maximal threshold for a loss probability equal to ε.

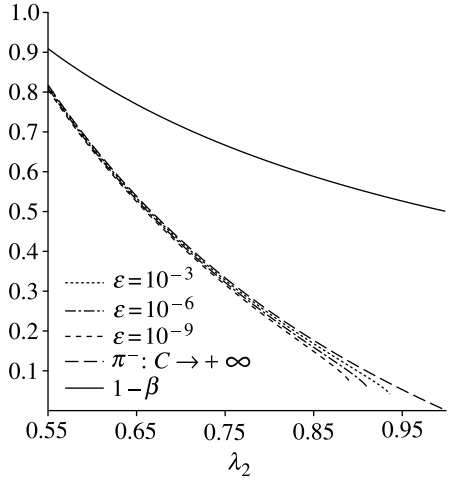

FIGURE 4: Fraction of nondowngraded jobs at equilibrium for the downgrading policy compared to the fraction of lost jobs in a pure loss system.

For comparison, β is defined as the corresponding acceptance probability when no control is used in the system. We show, in Figure 4, the relation between these quantities and the workload λ_2, for fixed loss rates of 10^{-3}, 10^{-6}, and 10^{-9}. We have $\beta = 1 - 1/(2\lambda_2)$; see [15, Proposition 6.19]. The difference $\beta - \pi^-$ can be seen as the fraction of jobs which are downgraded for our policy but lost in the uncontrolled policy. Intuitively, it can be seen as the price of not rejecting any job. Note also that the curves plotting π_ε^- for $\varepsilon = 10^{-3}, 10^{-6}, 10^{-9}$ are close and that β is larger than π^-. We remark nevertheless that, for high loads, the system cannot hold these demands because our policy is no longer effective.

https://doi.org/10.1017/apr.2017.15 Published online by Cambridge University Press

Acknowledgement

The authors are very grateful to an anonymous referee for pointing out a gap in the proof of Theorem 2 in the first version of this work.

References

[1] Añorga, J. (2015). *et al.* YouTube's DASH implementation analysis. In *Recent Advances in Communications* (Proc. 19th Internat. Conf. on Communications; Recent Adv. Electrical Eng. Ser. **50**), pp. 61–66.
[2] Arnol'd, V. I. (1992). *Ordinary Differential Equations*. Springer, Berlin.
[3] Asmussen, S. (2003). *Applied Probability and Queues* (Appl. Math. **51**), 2nd edn. Springer, New York.
[4] Bean, N. G., Gibbens, R. J. and Zachary, S. (1995). Asymptotic analysis of single resource loss systems in heavy traffic, with applications to integrated networks. *Adv. Appl. Prob.* **27,** 273–292.
[5] Bean, N. G., Gibbens, R. J. and Zachary, S. (1997). Dynamic and equilibrium behavior of controlled loss networks. *Ann. Appl. Prob.* **7,** 873–885.
[6] Doane, D. P. and Seward, L. E. (2011). Measuring skewness: a forgotten statistic? *J. Statist. Education* **19,** 1–18.
[7] Fricker, C., Guillemin, F., Robert, P. and Thompson, G. (2016). Analysis of downgrading for resource allocation. *ACM SIGMETRICS Performance Evaluation Rev.* **44,** 24–26.
[8] Fricker, C., Guillemin, F., Robert, P. and Thompson, G. (2016). Analysis of an offloading scheme for data centers in the framework of Fog computing. *ACM Trans. Model. Performance Evaluation Comput. Systems* **1,** 12pp.
[9] Gakhov, F. D. (1990). *Boundary Value Problems*. Dover, New York.
[10] Guillemin, F., Houdoin, T. and Moteau, S. (2013). Volatility of YouTube content in Orange networks and consequences. In *Proc. IEEE Internat. Conf. on Communications*, IEEE, pp. 2381–2385.
[11] Guillemin, F., Kauffmann, B., Moteau, S. and Simonian, A. (2013). Experimental analysis of caching efficiency for YouTube traffic in an ISP network. In *Proc. 25th Internat. Teletraffic Congress*, IEEE, 9pp.
[12] Hunt, P. J. and Kurtz, T. G. (1994). Large loss networks. *Stoch. Process. Appl.* **53,** 363–378.
[13] Kelly, F. P. (1986). Blocking probabilities in large circuit-switched networks. *Adv. Appl. Prob.* **18,** 473–505.
[14] Kelly, F. P. (1991). Loss networks. *Ann. Appl. Prob.* **1,** 319–378.
[15] Robert, P. (2003). *Stochastic Networks and Queues* (Appl. Math. **52**). Springer, Berlin.
[16] Rudin, W. (1987). *Real and Complex Analysis*, 3rd edn. McGraw-Hill, New York.
[17] Schwarz, H., Marpe, D. and Wiegand, T. (2007). Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Trans. Circuits Systems Video Tech.* **17,** 1103–1120.
[18] Sieber, C. *et al.* (2013). Implementation and user-centric comparison of a novel adaptation logic for DASH with SVC. In *2013 IFIP/IEEE Internat. Symp. on Integrated Network Management*, IEEE, pp. 1318–1323.
[19] Stolyar, A. L. (2013). An infinite server system with general packing constraints. *Operat. Res.* **61,** 1200–1217.
[20] Stolyar, A. L. (2015). Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems* **80,** 341–361.
[21] Vadlakonda, S. *et al.* (2010). System and method for dynamically upgrading / downgrading a conference session. Patent US7694002 B2.
[22] Zachary, S. and Ziedins, I. (2002). A refinement of the Hunt–Kurtz theory of large loss networks, with an application to virtual partitioning. *Ann. Appl. Prob.* **12,** 1–22.
[23] Zachary, S. and Ziedins, I. (2011). Loss networks. In *Queueing Networks* (Internat. Ser. Operat. Res. Manag. Sci. **154**), eds R. J. Boucherie and N. M. van Dijk, Springer, New York, pp. 701–728.