Available online at www.sciencedirect.com
ScienceDirect
Performance Evaluation 65 (2008) 152–180
PERFORMANCE EVALUATION
An International Journal
www.elsevier.com/locate/peva

Throughputs in processor sharing models for integrated stream and elastic traffic

Remco Litjens[a], Hans van den Berg[a,b,*], Richard J. Boucherie[c]

[a] Department of Planning, Performance and Quality, TNO ICT, The Netherlands
[b] Department of Computer Science, University of Twente, The Netherlands
[c] Department of Applied Mathematics, University of Twente, The Netherlands

Received 10 December 2003; received in revised form 18 January 2007; accepted 11 May 2007
Available online 17 May 2007

Abstract

We present an analytical study of throughput measures in processor sharing queuing systems with randomly varying service rates, modelling e.g. a communication link in an integrated services network carrying prioritised fixed rate stream traffic and rate-adaptive elastic traffic. A number of distinct throughput measures for the elastic traffic are defined, analysed and compared under various system conditions, both by analytical means and simulation. It is concluded that the call-average throughput, which is most relevant from the user point of view but typically hard to analyse, is very well approximated by the newly proposed so-called expected instantaneous throughput, which is readily obtained from the system's steady state distribution.

Keywords: Throughput; Processor sharing; Random environment

1. Introduction

Processor sharing (PS) queuing models are widely applicable to situations where a common resource is shared by a varying number of concurrent users. In particular, PS models have been fruitfully applied in the field of the performance evaluation of computer systems and telecommunication networks. For instance, the PS service discipline appropriately models the design principle of fair resource sharing by TCP (Transmission Control Protocol) controlled elastic (rate-adaptive) data calls or packet scheduling schemes in, e.g. IP (Internet Protocol), GPRS (General Packet Radio Service), UMTS (Universal Mobile Telecommunications System) networks and WLANs (Wireless Local Area Networks) [1–3,6,22,25].

The 'classical' PS model consists of a single server fairly sharing its fixed capacity among the varying number of present jobs (calls). A relevant extension is the PS queue with randomly varying service capacity, which models e.g. the impact of fluctuating high-priority stream traffic (e.g. speech calls) on low-priority elastic traffic (e.g. video or data calls) sharing a common network link. Important performance measures for PS queues are the sojourn time

* Corresponding author at: Department of Planning, Performance and Quality, TNO ICT, The Netherlands. Tel.: +31 152857031; fax: +31 152857349.

E-mail address: j.l.vandenberg@tno.nl (H. van den Berg).

0166-5316/$ - see front matter
doi:10.1016/j.peva.2007.05.002

and throughput experienced by a job. In the queuing literature, the analyses of PS models are generally focussed towards the (conditional) *sojourn times*, and many analytical results are available. In contrast, although the relevance is apparent from practical applications, *throughput* analyses are rare and few results are known. Therefore, in the present paper, we concentrate on the analysis and comparison of a variety of relevant throughput measures in PS models with fixed or randomly varying service capacity.

Literature Well-known results are the linearity and insensitivity properties, i.e. the expected sojourn time of a tagged job is proportional to its service requirement and independent of the service requirement distribution of the other jobs (see e.g. [20]). The sojourn time distribution for the $M/G/1$ PS queue has been derived by Yashkov [38] and Ott [31]. Cohen [10] considers a generalisation of the $M/G/1$ PS queue, viz. the so-called generalised processor sharing (GPS) model, in which the service rate of the jobs is an arbitrary function of the number of jobs in the system. Note that e.g. the multiple server $M/G/c$ PS queue and the classical Erlang loss model are special cases of the GPS model, which also possesses the linearity and insensitivity properties mentioned above. The reader is referred to [39] and [40] for overviews of the available results on 'classical' PS systems; see also the more recent paper by Zwart and Boxma [41] focusing on sojourn time asymptotics for the $M/G/1$ PS queue with heavy tailed service requirement distributions (e.g. Pareto), and Cheung et al. [7] which provides insensitive bounds for higher moments of the sojourn time.

In the present paper PS systems with *randomly varying* service rates (e.g. due to the presence of higher priority jobs consuming part of the total service capacity) play a particularly important role. Randomly varying service rates severely complicate the analysis, and the nice properties of the steady-state distribution and the expected sojourn time do not hold anymore. Núñez-Queija [28] analyses an $M/M/1$ PS model with an on/off server, and derives closed-form expressions for several sojourn time statistics. In [30], Núñez-Queija et al. consider a GPS model with two priority classes, where each of the high priority jobs takes a fixed amount of the server capacity and the low priority jobs utilise the (fluctuating) remaining service capacity in a PS fashion. For this model, expressions for the (conditional) expected sojourn times of the low priority customers are derived. A generalisation and more extensive treatment of this work can be found in [27,29]. [23] presents and analytically supports the remarkable phenomenon that in the PS model with randomly varying capacity, the expected sojourn times are smaller if the job sizes are more variable, which is a relevant insight in light of the commonly acknowledged property that e.g. WWW pages are heavy tailed [11,21].

Throughput analyses of PS systems are rare in the literature. The only references known to the authors are by Kherani and Kumar [18,19], who use the $M/G/1$ PS queue as a model to evaluate the throughput of TCP-controlled elastic data calls in the Internet, cf. [26,30,34]. While from the user's perspective, the *call-average throughput* is the most relevant average throughput measure, in PS systems the call-average throughput may be hard to determine analytically [18,19]. Therefore, in many papers, other, more tractable throughput measures are selected as a basis for the performance analysis of systems modelled by a PS queue. E.g. in [14,18,19,24] the *time-average throughput*, defined as the expected throughput the 'server' provides to an elastic call at an arbitrary (non-idle) time instant, is applied to approximate the call-average throughput. Many other papers use the *ratio* of the expected transfer volume and the expected sojourn time as an approximation [1,2,4,5,12,32], however they do this mostly without substantiating the validity of this measure.

Contribution The principal objective of the present paper is to investigate and compare, both analytically and numerically, a variety of throughput performance measures in processor sharing models with fixed and varying service capacities. In particular, we introduce the *expected instantaneous throughput*, i.e. the throughput an admitted call experiences immediately upon admission to the system, as a new throughput measure, which can be analysed relatively easily. The experiments demonstrate that the expected instantaneous throughput is the *only* one among the considered throughput measures, which excellently approximates the call-average throughput for each of the investigated PS models and over the entire range of traffic loads.

Aside from a substantial original contribution in the definition, analysis and comparison of throughput measures, known results have been included in order to also establish the survey character of the paper.

Outline The remainder of this paper is organised as follows. Section 2 describes the PS models investigated in this paper in the setting of a communication link shared by different traffic types, and specifies the various throughput measures investigated in this paper. An analytical evaluation of these throughput measures is presented in Section 3. Section 4 presents and discusses the results of an extensive set of numerical experiments carried out to compare the different throughput measures for the different PS models. Additionally, some numerical results are provided for throughput variances in order to assess whether the qualitative conclusions obtained for averages extend to higher moments. The concluding remarks in Section 5 end the main body of this paper. To enhance readability, some lengthy proofs are contained in the Appendix.

2. Models and measures

As mentioned above, we introduce the models in the setting of a communication link shared by different call types. In particular, we consider a communication link with C traffic channels which can be assigned to 'stream' calls, characterised by a fixed channel assignment (e.g. speech telephony), and to 'elastic' calls that can adapt their service requirements and share the traffic channels left over by the stream calls. Concerning the elastic calls, we consider two distinct types: *(i)* elastic calls, whose sojourn time is unaffected by the (dynamically) assigned service rate (e.g. video telephony); and *(ii)* elastic calls, whose sojourn time is affected by the assigned service rate (e.g. data transfer). In the remainder of the paper, the three call types will be referred to by means of the given typical example services. The defining characteristics of the different call types are given below, followed by the specification of the call handling procedures in the two main performance models considered in this paper. An overview of the considered throughput measures ends the section.

2.1. Call characteristics

The three distinct call types mentioned above are characterised in more detail as follows:

Speech calls Speech calls arrive according to a Poisson process with arrival intensity λ_{speech}, and have a generally distributed duration with mean $1/\mu_{\text{speech}}$. A speech call requires a fixed assignment of one traffic channel. The speech traffic load is given by $\rho_{\text{speech}} \equiv \lambda_{\text{speech}}/\mu_{\text{speech}}$, and is expressed in Erlangs.

Video calls Video calls arrive according to a Poisson process with arrival intensity λ_{video}, have a generally distributed duration with mean $1/\mu_{\text{video}}$, and are elastic (*scalable*) in the ideal sense that the assigned number of traffic channels, and thus the video quality can instantaneously, and with perfect granularity, adapt to the varying network load. The number of traffic channels that can be assigned to a video call is constrained by a maximum denoted $\beta_{\text{video}}^{\max}$. On the other hand, acceptable video quality is guaranteed by means of a minimum channel assignment of $\beta_{\text{video}}^{\min} \in \left[0, \beta_{\text{video}}^{\max}\right]$ traffic channels, corresponding to a bit rate of $r_{\text{video}}\beta_{\text{video}}^{\min}$kbits/s, with r_{video} the effective video bit rate per traffic channel. The video traffic load is defined as $\rho_{\text{video}} \equiv \lambda_{\text{video}}/\mu_{\text{video}}$.

Data calls Data calls arrive according to a Poisson process with arrival intensity λ_{data}. A data call is assumed to be the transfer of a file with a generally distributed size, which is expressed in its nominal transfer time assuming a single dedicated traffic channel. The mean call size and effective data bit rate per traffic channel are denoted by $1/\mu_{\text{data}}$ and r_{data} (in kbits/s) respectively, corresponding to an actual mean transfer volume of $r_{\text{data}}/\mu_{\text{data}}$ kbits. Data calls are elastic, in the sense that they are delay tolerant, and can therefore tolerate a dynamic channel assignment, which affects the experienced throughput, and thus the data call's sojourn time. As for the video calls, a maximum assignment denoted $\beta_{\text{data}}^{\max}$ is enforced to incorporate the terminals' technical limitations, while a possible QOS requirement is modelled by means of a minimum channel assignment $\beta_{\text{data}}^{\min}$. The data traffic load is given by $\rho_{\text{data}} \equiv \lambda_{\text{data}}/\mu_{\text{data}}$, while the normalised data traffic load is denoted as $\rho_{\text{data}}^{\star} \equiv \rho_{\text{data}}/C$.

Observe from the specifications above that the key difference between video and data calls is the impact of the channel assignment on the calls' presence in the system. For video calls, the channel assignment influences the perceived audio and image quality experienced on the video terminal, while it does not affect the autonomously sampled video call duration. In case of data calls, the channel assignment affects the rate at which the file is transferred and thus the data call's sojourn time which, aside from the data throughput, is a key performance measure in itself.

2.2. Performance models

As speech calls require a fixed capacity assignment during their lifetimes, the dynamics in their arrivals and departures leave a time-varying residual capacity for the considered elastic call type. In other words, from the point of view of the elastic calls, the system behaves as a processor sharing type of model with varying service capacity. For the case of elastic video calls, the model is denoted by SV, and for elastic data calls by SD. These models are described in more detail below. Let $S(t)$, $V(t)$ and $D(t)$ denote the processes following the number of speech, video and data calls present at time $t \geq 0$, with states denoted s, v and d, respectively.

SV *model* In the SV model, the C traffic channels are dynamically shared by speech and video calls. Aside from the channels that are assigned to ongoing video calls in order to meet their minimum QOS requirements, the remaining service capacity is available with preemptive priority for speech calls. In other words, an arriving speech call is admitted if and only if $s + 1 \leq s_{\max}(v) \equiv \lfloor C - v\beta_{\text{video}}^{\min} \rfloor$, given a presence of s speech and v video calls. Analogously, if $\beta_{\text{video}}^{\min} > 0$, the condition for the admission of a video call is given by $v + 1 \leq v_{\max}(s) \equiv \lfloor (C - s) / \beta_{\text{video}}^{\min} \rfloor$. At any given time, the capacity that is not assigned to speech calls, is fairly shared by the present video calls in a PS fashion, i.e. each video call is assigned an instantaneous channel assignment of $\beta_{\text{video}}(s, v) \equiv \min\left\{(C - s)/v, \beta_{\text{video}}^{\max}\right\}$, which is guaranteed to exceed the minimum QOS requirement due to effects of the call admission control. Observe that the SV model is an example of a multi-rate model (see e.g. [16,33]) incorporating speech and video calls with respective capacity requirements of 1 and $\beta_{\text{video}}^{\min}$ traffic channels.

SD *model* In the SD model, the C traffic channels are dynamically shared by speech and data calls. In line with the above specification of the SV model, the call admission control conditions for the admission of a speech or data call are given by $s + 1 \leq s_{\max}(d) \equiv \lfloor C - d\beta_{\text{data}}^{\min} \rfloor$ and $d + 1 \leq d_{\max}(s) \equiv \lfloor (C - s)/\beta_{\text{data}}^{\min} \rfloor$ (only if $\beta_{\text{data}}^{\min} > 0$), respectively, given a presence of s speech and d data calls. At any given time, the capacity that is not assigned to speech calls, is fairly shared by the present data calls, i.e. each data call is assigned an instantaneous channel assignment of $\beta_{\text{data}}(s, d) \equiv \min\left\{(C - s)/d, \beta_{\text{data}}^{\max}\right\} \geq \beta_{\text{data}}^{\min}$.

The models described above ‘reduce’ to processor sharing models with fixed capacity when the speech call arrival intensity is taken as equal to zero. The resulting models are denoted V and D, and will be treated in the analysis in Section 3 as special cases of the SV and SD models, for which often more (or more explicit) results can be derived.

2.3. Throughput measures

In this subsection, we present the definitions of the different throughput measures to be analyzed and compared in Sections 3 and 4. The definitions apply to both elastic call types, i.e. video and data. Denote by a_k (d_k) the arrival (departure) time of the kth *admitted* elastic call, by $\tau_k \equiv d_k - a_k$ the call's sojourn time and by x_k the associated information volume (in kbits) transferred during its sojourn. Recall that for the video service, the call durations τ_k are autonomously sampled and the transfer volumes x_k are determined by the system dynamics, while for the data service the reverse holds. Let τ and x be the corresponding random variables with expected values $\mathbf{E}\{\tau\}$ and $\mathbf{E}\{x\}$.

The *call-average* throughput $\mathbf{R}^c$ is the most relevant average throughput measure from the user's perspective, defined as the per call throughput averaged over all calls, i.e.,

$$\mathbf{R}^c \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{x_k}{\tau_k} = \mathbf{E}\left\{\frac{x}{\tau}\right\}. \tag{1}$$

The *time-average* throughput $\mathbf{R}^t$ is defined as the expected throughput the server provides to an elastic call at an arbitrary (non-idle) time instant. With $N(t)$ as the number of elastic calls present in the system, and $C(t)$ as the aggregate number of channels assigned to the elastic service at time $t \geq 0$, this throughput measure is expressed as

$$\mathbf{R}^t \equiv \lim_{t\to\infty} \frac{\frac{1}{t}\int_0^t \frac{rC(u)}{N(u)} 1\{N(u) \geq 1\}\,\mathrm{d}u}{\frac{1}{t}\int_0^t 1\{N(u) \geq 1\}\,\mathrm{d}u}, \tag{2}$$

where r denotes the effective information bit rate per traffic channel. Note that $N(t)$ is given by $V(t)$ in the (S)V model or $D(t)$ in the (S)D model, while $C(t)/N(t)$ is given by the channel assignment functions $\beta(\cdot)$. The time-average throughput is used to approximate the call-average throughput in e.g. [14,18,19].

As a new throughput measure, we introduce the expected *instantaneous* throughput, denoted by $\mathbf{R}^i$. It is defined as the expected throughput an admitted call experiences immediately upon admission to the system, i.e.,

$$\mathbf{R}^i \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{rC\,(a_k)}{N(a_k^+)}, \tag{3}$$

where $N(a_k^+)$ denotes the number of ongoing elastic calls immediately after the kth elastic call admission, and thus includes the new call.

The *ratio* $\mathbf{R}^r$ of the expected transfer volume and the expected sojourn time is, like the time-average throughput, also often used as an alternative to the call average throughput, see e.g. [1,2,4,5,12,32]. It is formally defined by

$$\mathbf{R}^r \equiv \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{k=1}^{n} x_k}{\frac{1}{n}\sum_{k=1}^{n} \tau_k} = \frac{\mathbf{E}\{x\}}{\mathbf{E}\{\tau\}}. \tag{4}$$

Note that $\mathbf{R}^r$ can also be written as

$$\mathbf{R}^r = \frac{\lambda\,(1-\mathbf{P})\,\mathbf{E}\{x\}}{\lambda\,(1-\mathbf{P})\,\mathbf{E}\{\tau\}} = \lim_{t\to\infty} \frac{\frac{1}{t}\int_0^t rC\,(u)\,\mathrm{d}u}{\frac{1}{t}\int_0^t N(u)\mathrm{d}u},$$

where λ denotes the elastic call arrival rate and $\mathbf{P}$ the elastic call blocking probability (see also below). This alternate expression for $\mathbf{R}^r$ is given by the ratio of the long-term average aggregate system throughput and the long-term average number of elastic calls in the system. Its equivalence to expression (4) is due to the fact that in equilibrium the aggregate admitted bit rate must be equal to the aggregate processed bit rate (numerator) and Little's law (denominator).

As a final measure, the (unitless) call-average *stretch* $\mathbf{S}$ (or the *normalised sojourn time*) is given by

$$\mathbf{S} \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \frac{\tau_k}{\left(\frac{x_k}{rC}\right)} = rC\,\mathbf{E}\left\{\frac{\tau}{x}\right\}, \tag{5}$$

which is relevant for the data service only and is used as a performance measure in e.g. [15,34].

For the special case of unrestricted channel assignments, i.e. $\beta^{\min}_{\text{data}} = \beta^{\min}_{\text{video}} = 0$ and $\beta^{\max} \geq C$, let $\widetilde{\mathbf{R}}^c$, $\widetilde{\mathbf{R}}^t$, $\widetilde{\mathbf{R}}^i$, $\widetilde{\mathbf{R}}^r$ and $\widetilde{\mathbf{S}}$, denote the associated performance measures corresponding to the measures specified above for the more general settings.

3. Performance analysis

In this section, we derive analytical expressions for the performance measures in the four models specified above. In particular, in Sections 3.1 and 3.2 we study the SV and V models, respectively. Sections 3.3 and 3.4 are concerned with the analysis of the SD and D models. For each model, we start by analysing the equilibrium distribution and call blocking probabilities, and then successively consider, for the involved elastic traffic type, the call-average throughput, the time-average throughput, the expected instantaneous throughput, the ratio throughput measure and the call-average stretch (note that this last measure is applicable only to the SD and D models). Finally, an analytical comparison of the throughput measures for the considered model is made.

3.1. Analysis of SV model

Consider the SV model with generally distributed speech and video call durations. The evolution of the system in the SV model can then be described by the continuous-time stochastic process $(S(t), V(t))_{t\geq 0}$, with states denoted (s, v). The process' state space is given by $\mathbb{S} \equiv \left\{(s, v) \in \mathbb{N}_0 \times \mathbb{N}_0 : s + v\beta^{\min}_{\text{video}} \leq C\right\}$. The unique equilibrium probability vector $\boldsymbol{\pi}$ of the stochastic process, given by

$$\pi\,(s,v) = \left(\sum_{(s,v)\in\mathbb{S}} \frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\right)^{-1} \frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}, \quad (s,v)\in\mathbb{S},$$

is *insensitive* to the specific form of the speech and video call distributions, depending on their means only (see e.g. [16,17,33]). For the special case of unrestricted channel assignments to the video service, the state space is equal to $\widetilde{\mathbb{S}} \equiv \{(s,v)\in\mathbb{N}_0\times\mathbb{N}_0 : s\le C\}$, and the equilibrium distribution is given by

$$\widetilde{\pi}\,(s,v) = \exp\left(-\rho_{\text{video}}\right)\left(\sum_{s=0}^{C}\frac{\rho_{\text{speech}}^{s}}{s!}\right)^{-1}\frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}, \quad (s,v)\in\mathbb{S}.$$

Using the well-known PASTA property [37], which states that in equilibrium, under very general conditions, the fraction of Poisson arrivals that find a stochastic process in a particular system state is equal to the fraction of time the process spends in that state, the call blocking probabilities ($\mathbf{P}_{\text{speech}}$, $\mathbf{P}_{\text{video}}$) are readily derived from the equilibrium distribution:

$$\mathbf{P}_{\text{speech}} = \sum_{v=0}^{v_{\max}(0)} \pi\left(s_{\max}\left(v\right),v\right) \quad \text{and} \quad \mathbf{P}_{\text{video}} = \sum_{s=0}^{C}\pi\left(s, v_{\max}\left(s\right)\right).$$

In the case of unrestricted channel assignments to the video service, the speech call blocking probability is simply given by the Erlang loss probability, denoted $\widetilde{\mathbf{P}}_{\text{speech}}$, since speech traffic does not 'see' video traffic in the absence of video QOS guarantees, while the video call blocking probability equals zero.

3.1.1. Call-average throughput

In the analysis of the *call-average* throughput a video call, we first confine ourselves to the case of *exponentially* distributed speech and video call durations. Next, it will be shown (see Theorem 2 and the proof in Appendix B) that this performance measure is *insensitive* to the distributions of the speech and video call durations (apart from their means), i.e. the result derived under the assumption of exponential calls is also valid for generally distributed speech and video call durations.

For each state $(s,v)\in\mathbb{S}_{\text{video}}^{+}\equiv\{(s,v)\in\mathbb{S}: v>0\}$, denote with $\widehat{x}_{s,v}(\tau)$ the conditional expected transfer volume of an admitted video call of duration τ, arriving at a given system state (s,v), where v includes the new video call. The derivation involves a modified version of the Markov chain that is readily specified to describe the evolution of the SV model's stochastic process under the exponentiality assumption. Characterised by the presence of one permanent video call, the modified Markov chain consequently has the reduced state space $\mathbb{S}_{\text{video}}^{+}$. The video call departure rates in the associated infinitesimal generator $\mathcal{Q}_{\text{video}}^{\star}$ reflect the presence of the permanent video call, i.e. $\mathcal{Q}_{\text{video}}^{\star}\left((s,v);(s,v-1)\right) = (v-1)\mu_{\text{video}}$. The equilibrium distribution vector $\boldsymbol{\pi}_{\text{video}}^{\star}\equiv\left(\pi_{\text{video}}^{\star}(s,v),(s,v)\in\mathbb{S}_{\text{video}}^{+}\right)$, lexicographically ordered in (s,v), of the modified Markov chain is, invoking reversibility and truncation of a reversible process [17], readily obtained as

$$\pi_{\text{video}}^{\star}\left(s,v\right) = \frac{\pi\,(s,v-1)}{\sum\limits_{(s',v')\in\mathbb{S}_{\text{video}}^{+}}\pi\,(s',v'-1)}, \quad (s,v)\in\mathbb{S}_{\text{video}}^{+}, \tag{6}$$

i.e. the equilibrium probabilities $\pi_{\text{video}}^{\star}(s,v)$ corresponding to the modified Markov chain with one permanent video call are *equal* to the conditional probabilities that a newly admitted video call brings the system in state (s,v) in the original Markov chain. The equilibrium distribution $\boldsymbol{\pi}_{\text{video}}^{\star}$ can readily be seen to be insensitive to the specific form of the speech and video call duration distributions [16,17,33]. Let $\mathcal{B}_{\text{video}}\equiv\text{diag}(\beta_{\text{video}}(s,v),(s,v)\in\mathbb{S}_{\text{video}}^{+})$ denote the diagonal matrix of video channel assignments, lexicographically ordered in (s,v). We can now formulate Theorem 1 below, which is proven in Appendix A.

Theorem 1. *For exponentially distributed video call durations, the conditional expected video throughput vector* $\widehat{\mathbf{x}}(\tau)/\tau \equiv (\widehat{x}_{s,v}(\tau)/\tau, (s,v) \in \mathbb{S}^+_{\text{video}})$*, lexicographically ordered in* (s,v)*, is given by*

$$\frac{\widehat{\mathbf{x}}(\tau)}{\tau} = r_{\text{video}} \left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}\right) \mathbf{1} + \frac{1}{\tau} \left[\mathcal{I} - \exp\left\{\tau \mathcal{Q}^\star_{\text{video}}\right\}\right] \boldsymbol{\gamma}_{\text{video}},$$

where $\boldsymbol{\gamma}_{\text{video}} \equiv \left(\gamma_{\text{video}}(s,v), (s,v) \in \mathbb{S}^+_{\text{video}}\right)$ *is the unique solution to*

$$\mathcal{Q}^\star_{\text{video}} \boldsymbol{\gamma}_{\text{video}} = r_{\text{video}} \left\{\left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}\right) \mathbf{1} - \mathcal{B}_{\text{video}} \mathbf{1}\right\}, \tag{7}$$

$$\boldsymbol{\pi}^\star \boldsymbol{\gamma}_{\text{video}} = \mathbf{0}. \tag{8}$$

The conditional expected (call-average) video throughput $\mathbf{R}^c_{\text{video}}(s,v,\tau)$ of a video call admitted to the system in state (s,v) with a given holding time τ is given by (recall (1))

$$\mathbf{R}^c_{\text{video}}(s,v,\tau) = \frac{\widehat{x}_{s,v}(\tau)}{\tau}. \tag{9}$$

Deconditioning on the system state upon admission yields the conditional expected (call-average) video throughput of an admitted video call with duration τ, given by

$$\begin{aligned}
\mathbf{R}^c_{\text{video}}(\tau) &= \sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \left(\frac{\pi\,(s, v-1)}{\sum_{(s',v')\in\mathbb{S}^+_{\text{video}}} \pi\,(s', v'-1)}\right) \mathbf{R}^c_{\text{video}}\,(s,v,\tau) \\
&= \boldsymbol{\pi}^\star_{\text{video}} \left\{r_{\text{video}} \left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}\right) \mathbf{1} + \frac{1}{\tau} \left[\mathcal{I} - \exp\left\{\tau \mathcal{Q}^\star_{\text{video}}\right\}\right] \boldsymbol{\gamma}_{\text{video}}\right\} \\
&= r_{\text{video}} \left(\boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}\right) + \frac{1}{\tau} \boldsymbol{\pi}^\star_{\text{video}} \left(\boldsymbol{\gamma}_{\text{video}} - \sum_{k=0}^{\infty} \frac{\left(\tau \mathcal{Q}^\star_{\text{video}}\right)^k}{k!} \boldsymbol{\gamma}_{\text{video}}\right) \\
&= r_{\text{video}} \boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1} = r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \left(\frac{\pi\,(s, v-1)}{\sum_{(s',v')\in\mathbb{S}^+_{\text{video}}} \pi\,(s', v'-1)}\right) \beta_{\text{video}}\,(s,v),
\end{aligned}$$

using (8) and $\boldsymbol{\pi}^\star_{\text{video}} \mathcal{Q}^\star_{\text{video}} = \mathbf{0}$. Observe that $r_{\text{video}} \boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}$ is equal to the time-average video throughput in the SV model with one permanent video call (see also below). Comparing the first and last expressions in the above derivation might confuse the reader into thinking that $\mathbf{R}^c_{\text{video}}\,(s,v,\tau)$ is simply equal to $r_{\text{video}} \beta_{\text{video}}\,(s,v)$, which is however readily seen to be not the case. Observe that $\mathbf{R}^c_{\text{video}}(\tau)$ does not depend on τ, so that the call-average video throughput is given by

$$\mathbf{R}^c_{\text{video}} = \int_{\tau=0}^{\infty} \mathbf{R}^c_{\text{video}}(\tau) \mu_{\text{video}} \exp\left\{-\tau \mu_{\text{video}}\right\} \mathrm{d}\tau = \mathbf{R}^c_{\text{video}}(\tau) = r_{\text{video}} \boldsymbol{\pi}^\star_{\text{video}} \mathcal{B}_{\text{video}} \mathbf{1}. \tag{10}$$

Whereas the above derivations utilised the exponentiality of the speech and video call durations, Theorem 2 implies that the obtained expressions for both $\mathbf{R}^c_{\text{video}}$ and $\mathbf{R}^c_{\text{video}}(\tau)$ (*not* $\mathbf{R}^c_{\text{video}}(s,v,\tau)$) also hold for general distributions of the speech and video call durations.

Theorem 2. *The call-average video throughput* $\mathbf{R}^c_{\text{video}}$*, and the conditional call-average video throughput* $\mathbf{R}^c_{\text{video}}(\tau)$*, are insensitive to the speech and video call duration distributions apart from their means.*

The proof of this theorem is presented in Appendix B.

In the case with unrestricted channel assignments, the (conditional) call-average video throughput can be simplified to

$$
\begin{aligned}
\widetilde{\mathbf{R}}^{c}_{\text{video}} &= \widetilde{\mathbf{R}}^{c}_{\text{video}}(\tau) = r_{\text{video}} \sum_{s=0}^{C} \sum_{v=1}^{\infty} \left(\frac{\frac{\rho_{\text{speech}}^{s}}{s!} \frac{\rho_{\text{video}}^{v-1}}{(v-1)!}}{\sum_{s'=0}^{C} \sum_{v'=1}^{\infty} \frac{\rho_{\text{speech}}^{s'}}{s'!} \frac{\rho_{\text{video}}^{v'-1}}{(v'-1)!}} \right) \frac{C-s}{v} \\
&= r_{\text{video}} \exp\left(-\rho_{\text{video}}\right) \sum_{v=1}^{\infty} \frac{\rho_{\text{video}}^{v-1}}{v!} \sum_{s=0}^{C} \left(\frac{\frac{\rho_{\text{speech}}^{s}}{s!}}{\sum_{s'=0}^{C} \frac{\rho_{\text{speech}}^{s'}}{s'!}} (C-s) \right) \\
&= r_{\text{video}} \exp\left(-\rho_{\text{video}}\right) \left(C - \rho_{\text{speech}} \left(1 - \widetilde{\mathbf{P}}_{\text{speech}}\right)\right) \sum_{v'=1}^{\infty} \frac{\rho_{\text{video}}^{v-1}}{v!} \\
&= r_{\text{video}} \frac{1 - \exp\left(-\rho_{\text{video}}\right)}{\rho_{\text{video}}} \left(C - \rho_{\text{speech}} \left(1 - \widetilde{\mathbf{P}}_{\text{speech}}\right)\right),
\end{aligned}
$$

with $\widetilde{\mathbf{P}}_{\text{speech}}$ being the Erlang loss probability in a system with C channels, and traffic load equal to ρ_{speech} Erlang.

3.1.2. *Time-average throughput*

Using the theory of regenerative processes (e.g. [36,37]), the *time-averaged* video throughput is given by, cf. (2),

$$
\begin{aligned}
\mathbf{R}^{t}_{\text{video}} &\equiv \lim_{t\to\infty} \frac{\frac{1}{t}\int_0^t r_{\text{video}} \beta_{\text{video}}\left(S(u), V(u)\right) 1\left\{V(u) \geq 1\right\} \mathrm{d}u}{\frac{1}{t}\int_0^t 1\left\{V(u) \geq 1\right\} \mathrm{d}u} \\
&= r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \left(\frac{\pi(s,v)}{\sum_{(s',v')\in\mathbb{S}^{+}_{\text{video}}} \pi(s',v')} \right) \beta_{\text{video}}(s,v), \qquad (11)
\end{aligned}
$$

where $\pi(s,v)/\sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \pi(s,v)$ is the equilibrium probability that the system is in state (s,v), conditioned on the presence of at least one video call. The involved Césaro limits are derived using the renewal reward theorem [36,37]. For the special case without channel assignment restrictions, this yields

$$
\begin{aligned}
\widetilde{\mathbf{R}}^{t}_{\text{video}} &= r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \left(\frac{\pi(s,v)}{\sum_{(s',v')\in\mathbb{S}^{+}_{\text{video}}} \pi(s',v')} \right) \beta_{\text{video}}(s,v) \\
&= r_{\text{video}} \sum_{s=0}^{C} \sum_{v=1}^{\infty} \left(\frac{\frac{\rho_{\text{speech}}^{s}}{s!} \frac{\rho_{\text{video}}^{v}}{v!} \left(\frac{C-s}{v}\right)}{\sum_{s'=0}^{C} \sum_{v'=1}^{\infty} \frac{\rho_{\text{speech}}^{s'}}{s'!} \frac{\rho_{\text{video}}^{v'}}{v'!}} \right) = r_{\text{video}} \left(\frac{\sum_{v=1}^{\infty} \frac{\rho_{\text{video}}^{v}}{v v!}}{\sum_{v'=1}^{\infty} \frac{\rho_{\text{video}}^{v'}}{v'!}} \right) \sum_{s=0}^{C} \left(\frac{\rho_{\text{speech}}^{s}/s!}{\sum_{s'=0}^{C} \rho_{\text{speech}}^{s'}/s'!} \right) (C-s) \\
&= \frac{r_{\text{video}}}{\left(\exp\left(\rho_{\text{video}}\right) - 1\right)} \left(\sum_{v=1}^{\infty} \frac{\rho_{\text{video}}^{v}}{v v!} \right) \left(C - \rho_{\text{speech}} \left(1 - \widetilde{\mathbf{P}}_{\text{speech}}\right)\right),
\end{aligned}
$$

where $\widetilde{\mathbf{P}}_{\text{speech}}$ is the Erlang loss probability. Note that the derivation of (11) does not require information on the specific form of the equilibrium distribution. As this equilibrium distribution is insensitive to the call duration distribution (except for its mean), this property is inherited by the time-average video throughput.

3.1.3. *Expected instantaneous throughput*

Again applying the theory of regenerative processes, the expected *instantaneous* video throughput as defined in (3) is obtained as

$$
\begin{aligned}
\mathbf{R}^{i}_{\text{video}} &\equiv \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} r_{\text{video}} \beta_{\text{video}} \left(S(a_k), V(a_k^{+}) \right) \\
&= r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \left(\frac{\pi(s, v-1)}{\sum_{(s',v')\in\mathbb{S}^{+}_{\text{video}}} \pi(s', v'-1)} \right) \beta_{\text{video}}(s, v) \\
&= r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} \pi^{\star}_{\text{video}}(s, v)\, \beta_{\text{video}}(s, v). \qquad (12)
\end{aligned}
$$

As for the time-averaged throughput, the expected instantaneous video throughput measure inherits its insensitivity with respect to the specific form of the video call duration distribution from the insensitivity of $\pi^{\star}_{\text{video}}$. Observe that the expected instantaneous video throughput is equal to the call-average video throughput, and hence so is the special case with unrestricted channel assignments.

3.1.4. Ratio throughput measure

The *ratio* of the expected video call transfer volume and the expected video call duration is given by

$$
\mathbf{R}^{r}_{\text{video}} = \frac{\mathbf{E}\left\{\tau \mathbf{R}^{c}_{\text{video}}(\tau)\right\}}{\mu^{-1}_{\text{video}}} = \mathbf{R}^{c}_{\text{video}}
$$

(cf. (4)), where the numerator is indeed equal to the expected transfer volume of a video call, using the fact that $\mathbf{R}^{c}_{\text{video}}(\tau) = \mathbf{R}^{c}_{\text{video}}$ does not depend on τ. It is readily seen that also for the special case of unrestricted channel assignments, the ratio throughput measure is equal to the corresponding call-average video throughput.

3.1.5. Comparison of throughput measures

From the results derived above, it appears that the *call-average* video throughput, the expected *instantaneous* video throughput and the *ratio* of the expected video call transfer volume and the expected video call duration are identical, i.e.

$$
\mathbf{R}^{c}_{\text{video}} = \mathbf{R}^{i}_{\text{video}} = \mathbf{R}^{r}_{\text{video}},
$$

and hence what remains is to compare these measures with the *time-average* throughput. Based on the explicit expressions (10) and (11), it can be shown for the case of $\beta^{\min}_{\text{video}} \in \{0, 1, \ldots, C\}$ that the time-average throughput exceeds the call-average throughput:

Theorem 3. *In the* SV *model with $\beta^{\min}_{\text{video}} \in \{0, 1, \ldots, C\}$, the call-average video throughput is less than or equal to the time-average video throughput: $\mathbf{R}^{c}_{\text{video}} \leq \mathbf{R}^{t}_{\text{video}}$.*

The proof of this theorem is given in Appendix C

As an interesting corollary, we obtain that the time-average video throughput is monotonous in the offered video traffic load. This is noted to be non-trivial: while for $\rho_{\text{speech}} = 0$ (V model), this monotonicity can readily be concluded via stochastic monotonicity, for $\rho_{\text{speech}} > 0$ speech calls may take the place of video calls, thus destroying stochastic monotonicity.

Corollary 1. *The time-average video throughput is non-increasing in the video traffic load for $\beta^{\min}_{\text{video}} \in \{0, 1, \ldots, C\}$, i.e.,*

$$
\frac{\partial \mathbf{R}^{t}_{\text{video}}}{\partial \rho_{\text{video}}} \leq 0.
$$

The proof of this corollary is given in Appendix D.

3.2. *Analysis of* V *model*

Since all relevant video throughput measures have been derived in closed-form for the SV model, including those for the case of unrestricted channel assignments, an explicit consideration of the V model would be superfluous, as it is merely a special case of the SV model with $\rho_{\text{speech}} = 0$. Also the ordering of the different throughput measures is as under the SV model.

3.3. *Analysis of* SD *model*

Consider the SD model with *exponentially* distributed speech call durations and data call sizes. The evolution of the system in the SD model can then be described by an irreducible two-dimensional continuous-time Markov chain $(S(t), D(t))_{t\geq 0}$, with states denoted (s, d). The state space of the Markov chain is given by $\mathbb{S} \equiv \{(s, d) \in \mathbb{N}_0 \times \mathbb{N}_0 : s + d\beta_{\text{data}}^{\min} \leq C\}$, while its infinitesimal generator $\mathcal{Q}$ is readily specified in terms of the speech and data call arrival and departure rates (see e.g. [22]). The irreducibility of the finite state space Markov chain $(S(t), D(t))_{t\geq 0}$ ensures the existence of a unique probability vector $\boldsymbol{\pi}$ that satisfies the system of global balance equations $\boldsymbol{\pi}\mathcal{Q} = \mathbf{0},$ with $\mathbf{0}$, the vector with all entries zero. The equilibrium distribution is *not* insensitive to the specific form of the speech call duration and data call size distributions. For the Markovian case, the equilibrium distribution can be determined numerically, e.g. by a successive overrelaxation procedure [36].

Using PASTA, the speech and data call blocking probabilities are given by

$$\mathbf{P}_{\text{speech}} = \sum_{d=0}^{d_{\max}(0)} \pi(s_{\max}(d), d) \quad \text{and} \quad \mathbf{P}_{\text{data}} = \sum_{s=0}^{C} \pi(s, d_{\max}(s)).$$

In the special case of unrestricted channel assignments to the data service, the speech call blocking probability becomes equal to the Erlang loss probability, as speech traffic does not 'see' data traffic in the absence of data QOS guarantees, while the data call blocking probability becomes zero.

3.3.1. *Call-average throughput*

Compared to other data throughput measures, obtaining explicit expressions for the *call-average* data throughput $\mathbf{R}_{\text{data}}^{c}$ is more involved. We first concentrate on the distribution of the data call sojourn times, conditional on the data call size. For each state $(s, d) \in \mathbb{S}_{\text{data}}^{+} \equiv \{(s, d) \in \mathbb{S} : d > 0\}$ define $\tau_{s,d}(x)$ as the random time it takes to transfer a file of size x, arriving at a given system state (s, d), where d includes the new data call. Define the Laplace–Stieltjes transform of the distribution of $\tau_{s,d}(x)$ by

$$T_{s,d}(\zeta, x) \equiv \mathbf{E}\left\{\exp\left\{-\zeta\tau_{s,d}(x)\right\}\right\}, \quad \text{Re}(\zeta) \geq 0,\ (s, d) \in \mathbb{S}_{\text{data}}^{+}$$

and let $\mathbf{T}(\zeta, x) = \left(T_{s,d}(\zeta, x), (s, d) \in \mathbb{S}_{\text{data}}^{+}\right)$ be lexicographically ordered in $(s, d) \in \mathbb{S}_{\text{data}}^{+}$.

In an analogous manner to that used to determine the conditional expected transfer volumes of video calls in the SV model, the derivation of an explicit expression for $\mathbf{T}(\zeta, x)$ involves a modified version of the original Markov chain, governed by infinitesimal generator $\mathcal{Q}_{\text{data}}^{\star}$, characterised by the presence of one permanent data call, and with state space $\mathbb{S}_{\text{data}}^{+}$. The data call departure rates in the modified chain reflect the presence of the permanent data call, and are equal to $\mathcal{Q}_{\text{data}}^{\star}((s, d) ; (s, d-1)) = \beta_{\text{data}}(s, d)(d-1)\mu_{\text{data}}$. Denote with $\boldsymbol{\pi}_{\text{data}}^{\star}$ the unique equilibrium distribution of the modified Markov chain, and let $\mathcal{B}_{\text{data}} \equiv \text{diag}(\beta_{\text{data}}(s, d), (s, d) \in \mathbb{S}_{\text{data}}^{+})$ be the diagonal matrix of data channel assignments, lexicographically ordered in (s, d). Partition $\mathbb{S}_{\text{data}}^{+}$ into $\mathbb{S}_{\text{data},0}^{+} \equiv \left\{(s, d) \in \mathbb{S}_{\text{data}}^{+} : \beta_{\text{data}}(s, d) = 0\right\}$ and its complement $\mathbb{S}_{\text{data},+}^{+} \equiv \mathbb{S}_{\text{data}}^{+} \backslash \mathbb{S}_{\text{data},0}^{+}$, and reorder the rows and columns in $\mathcal{Q}_{\text{data}}^{\star}$, $\mathcal{B}_{\text{data}}$, $\boldsymbol{\pi}_{\text{data}}^{\star}$ and $\mathbf{T}(\zeta, x)$ in accordance with the introduced state space partitioning, in order to allow the partitioning

$$\mathcal{Q}_{\text{data}}^{\star} = \begin{bmatrix} \mathcal{Q}_{++}^{\star} & \mathcal{Q}_{+0}^{\star} \\ \mathcal{Q}_{0+}^{\star} & \mathcal{Q}_{00}^{\star} \end{bmatrix}, \qquad \mathcal{B}_{\text{data}} = \begin{bmatrix} \mathcal{B}_{+} & \mathcal{O} \\ \mathcal{O} & \mathcal{O} \end{bmatrix},$$

and

$$\boldsymbol{\pi}_{\text{data}}^{\star} = \left(\boldsymbol{\pi}_{\text{data},0}^{\star}, \boldsymbol{\pi}_{\text{data},+}^{\star}\right), \qquad \mathbf{T}(\zeta, x) = \left(\mathbf{T}_{0}(\zeta, x), \mathbf{T}_{+}(\zeta, x)\right),$$

where we omit the 'data' subscript in the submatrices of $\mathcal{Q}^{\star}_{\text{data}}$ and $\mathcal{B}_{\text{data}}$ for enhanced readability. We note that in case $\beta^{\min}_{\text{data}} > 0$, this implies that $\mathbb{S}^{+}_{\text{data},0} = \emptyset$, leading to a slightly simplified analysis (see [27, Section 4.2]).

As shown in [27, Section 4.4], for $x \geq 0$ and $\text{Re}(\zeta) \geq 0$, a closed-form expression for $\mathbf{T}(\zeta, x)$ is given by

$$\mathbf{T}_0(\zeta, x) = -\left(\mathcal{Q}^{\star}_{00} - \zeta\mathcal{I}\right)^{-1} \mathcal{Q}^{\star}_{0+}\mathbf{T}_+(\zeta, x),$$

and

$$\mathbf{T}_+(\zeta, x) = \exp\left\{x\mathcal{B}^{-1}_+ \left(\mathcal{Q}^{\star}_{++} - \mathcal{Q}^{\star}_{+0}\left(\mathcal{Q}^{\star}_{00} - \zeta\mathcal{I}\right)^{-1}\mathcal{Q}^{\star}_{0+} - \zeta\mathcal{I}\right)\right\}\mathbf{1}.$$

The conditional expected throughput $\mathbf{R}^{c}_{\text{data}}(s, d, x)$ of a data call admitted to the system in state (s, d), and with a given size x is given by

$$\begin{aligned}
\mathbf{R}^{c}_{\text{data}}(s, d, x) &= r_{\text{data}}\mathbf{E}\left\{\frac{x}{\tau_{s,d}(x)}\right\} \\
&= r_{\text{data}}\int_{\tau=0}^{\infty}\frac{x}{\tau}\mathrm{d}\Phi_{s,d,x}(\tau) \\
&= r_{\text{data}}x\int_{\tau=0}^{\infty}\left(\int_{\zeta=0}^{\infty}\exp\{-\zeta\tau\}\,\mathrm{d}\zeta\right)\mathrm{d}\Phi_{s,d,x}(\tau) \\
&= r_{\text{data}}x\int_{\zeta=0}^{\infty}\left(\int_{\tau=0}^{\infty}\exp\{-\zeta\tau\}\,\mathrm{d}\Phi_{s,d,x}(\tau)\right)\mathrm{d}\zeta \\
&= r_{\text{data}}x\int_{\zeta=0}^{\infty}T_{s,d}(\zeta, x)\mathrm{d}\zeta,
\end{aligned}$$

where $\Phi_{s,d,x}(\tau)$ denotes the cumulative distribution function of $\tau_{s,d}(x)$, given data call size x and system state (s, d) upon the considered data call's admission. Deconditioning on the system state (s, d) upon admission yields

$$\mathbf{R}^{c}_{\text{data}}(x) = \sum_{(s,d)\in\mathbb{S}^{+}_{\text{data}}}\left(\frac{\pi(s, d-1)}{\sum_{(s',d')\in\mathbb{S}^{+}_{\text{data}}}\pi(s', d'-1)}\right)\mathbf{R}^{c}_{\text{data}}(s, d, x),$$

while subsequently deconditioning on the exponentially distributed data call size x gives the call-average data throughput as:

$$\mathbf{R}^{c}_{\text{data}} = \mu_{\text{data}}\sum_{(s,d)\in\mathbb{S}^{+}_{\text{data}}}\left(\frac{\pi(s, d-1)}{\sum_{(s',d')\in\mathbb{S}^{+}_{\text{data}}}\pi(s', d'-1)}\right)\int_{x=0}^{\infty}\exp\left(-\mu_{\text{data}}x\right)\mathbf{R}^{c}_{\text{data}}(s, d, x)\mathrm{d}x.$$

Since the equilibrium distribution can only be numerically obtained, the above expression does not simplify for the special case of unrestricted channel assignments.

3.3.2. *Time-average throughput*

The *time-average* data throughput can be derived is a similar way as the time-average video throughput in the SV model, cf. (11). In particular, we obtain:

$$\mathbf{R}^{t}_{\text{data}} = r_{\text{data}}\sum_{(s,d)\in\mathbb{S}^{+}_{\text{data}}}\left(\frac{\pi(s, d)}{\sum_{(s',d')\in\mathbb{S}^{+}_{\text{data}}}\pi(s', d')}\right)\beta_{\text{data}}(s, d).$$

This expression does not simplify for the special case of unrestricted channel assignments.

3.3.3. Expected instantaneous throughput

Similar to the derivation of the corresponding measure (12) for the SV model, the expected *instantaneous* data throughput is given by

$$\mathbf{R}^{i}_{\text{data}} = r_{\text{data}} \sum_{(s,d)\in\mathbb{S}^{+}_{\text{data}}} \left(\frac{\pi(s,d-1)}{\sum\limits_{(s',d')\in\mathbb{S}^{+}_{\text{data}}} \pi(s',d'-1)} \right) \beta_{\text{data}}(s,d).$$

This result does not simplify for the special case of unrestricted channel assignments.

3.3.4. Ratio throughput measure

The *ratio* of the expected data call size and the expected data call sojourn time is equal to

$$\mathbf{R}^{r}_{\text{data}} = \left(\frac{r_{\text{data}}}{\mu_{\text{data}}}\right) \Bigg/ \left(\frac{\sum\limits_{(s,d)\in\mathbb{S}} d\pi(s,d)}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})} \right) = r_{\text{data}} \frac{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}{\sum\limits_{(s,d)\in\mathbb{S}} d\pi(s,d)},$$

where Little's formula (see e.g. [36]) is applied to express the expected data call sojourn time in terms of the equilibrium distribution. The resulting formula does not simplify for the special case of unrestricted channel assignments.

3.3.5. Call-average stretch

The call-average *stretch*, or normalised sojourn time, is given by the expected ratio of the actual and the minimum sojourn time, where the latter is given by the service requirement (data call size). Using

$$\mathbf{E}\left\{ \frac{\tau_{s,d}(x)}{x} \middle| x, s, d \right\} = -\frac{1}{x} \frac{\partial}{\partial \zeta} T_{s,d}(\zeta, x) \Big|_{\zeta=0},$$

with the Laplace–Stieltjes transform $T_{s,d}(\zeta, x)$ as defined above, the expected (call-average) data stretch is given by

$$\begin{aligned} \mathbf{S}_{\text{data}} &= C\mathbf{E}\left\{ \frac{\tau_{s,d}(x)}{x} \right\} \\ &= -C\mu_{\text{data}} \sum_{(s,d)\in\mathbb{S}^{+}_{\text{data}}} \left(\frac{\pi(s,d-1)}{\sum\limits_{(s',d')\in\mathbb{S}^{+}_{\text{data}}} \pi(s,d-1)} \right) \times \left\{ \int_{x=0}^{\infty} \frac{1}{x} \exp(-\mu_{\text{data}} x) \left(\frac{\partial}{\partial \zeta} T_{s,d}(\zeta, x) \Big|_{\zeta=0} \right) \mathrm{d}x \right\}, \end{aligned}$$

conforming to the definition given by (5), and noting that in the above analysis the data call size x is expressed in units of r_{data} kbits (see also Section 2.1). This result does not simplify for the special case of unrestricted channel assignments.

3.3.6. Comparison of measures

The expressions for the various throughput measures derived above for the SD model do not allow an analytical comparison. A numerical comparison is presented in Section 4.

3.4. Analysis of D *model*

The D model is a special case of the SD model with $\rho_{\text{speech}} = 0$. Moreover, the D model is equivalent to the $M/G/1/d_{\max}$ GPS queuing model with state-dependent aggregate service rates given by $dr_{\text{data}}\beta_{\text{data}}(d) = dr_{\text{data}} \min\left\{C/d, \beta^{\max}_{\text{data}}\right\}$, see [10]. For this model, the equilibrium distribution is known to be *insensitive* to the specific form of the data call size distribution, and is given by

$$\pi(d) = \frac{\left(\rho^{\star}_{\text{data}}\right)^{d} \phi(d)}{\sum\limits_{d'=0}^{d_{\max}} \left(\rho^{\star}_{\text{data}}\right)^{d'} \phi(d')}, \quad \text{with } \phi(d) \equiv \left(\prod_{d'=1}^{d} \frac{d'\beta_{\text{data}}(d')}{C} \right)^{-1}, \ d = 0, \ldots, d_{\max},$$

where $\rho^{\star}_{\text{data}} \equiv \rho_{\text{data}}/C$ denotes the normalised data traffic load and $\phi(0) \equiv 1$ by convention. For the special case of unrestricted channel assignments, $d_{\max} = \infty$ and the D model reduces to the standard $M/G/1$ PS queuing model, which has a geometric equilibrium distribution:

$$\widetilde{\pi}(d) = \left(1-\rho^{\star}_{\text{data}}\right)\left(\rho^{\star}_{\text{data}}\right)^{d}, \quad d \geq 0,$$

requiring $\rho^{\star}_{\text{data}} < 1$ for stability.

Using PASTA, the data call blocking probability is equal to

$$\mathbf{P}_{\text{data}} = \pi(d_{\max}),$$

while it is equal to zero in the case of unrestricted channel assignments.

3.4.1. Call-average throughput

In this section, we assume exponentially distributed data call sizes. We first derive a closed-form expression for $\mathbf{T}(\zeta, x) \equiv (T_d(\zeta, x), d = 1, \ldots, d_{\max})$ with $T_d(\zeta, x)$ the Laplace–Stieltjes transform of the distribution of $\tau_d(x)$, i.e. the random sojourn time of a data call of size x admitted to the system in the presence of $d-1$ other data calls. Recall that x is expressed in the nominal sojourn time (in seconds). By analogy with the similar analysis presented for the SD model, $\mathcal{B}_{\text{data}}$ is the diagonal matrix of channel assignments, and $\mathcal{Q}^{\star}_{\text{data}}$ is the infinitesimal generator corresponding the D model's modified Markov chain with one permanent data call. In this data-only model, $\beta_{\text{data}}(d) > 0$ for all $d \geq 1$, so that no partitioning of $\mathbf{T}(\zeta, x)$ is required. As a specific instance of the result presented in [27, Section 4.2], for $x \geq 0$ and $\text{Re}(\zeta) \geq 0$, $\mathbf{T}(\zeta, x)$ is given by the closed-form expression

$$\mathbf{T}(\zeta, x) = \exp\left\{x\mathcal{B}^{-1}_{\text{data}}\left(\mathcal{Q}^{\star}_{\text{data}} - \zeta\mathcal{I}\right)\right\}\mathbf{1}.$$

By analogy with the analysis for the SD model, expressions for the conditional expected throughput measures $\mathbf{R}^{c}_{\text{data}}(d, x)$ and $\mathbf{R}^{c}_{\text{data}}(x)$ are readily derived. We limit ourselves here to stating the (unconditional) call-average data throughput:

$$\mathbf{R}^{c}_{\text{data}} = \mu_{\text{data}} \sum_{d=1}^{d_{\max}} \left(\frac{\pi(d-1)}{\sum\limits_{d'=1}^{d_{\max}} \pi(d'-1)}\right) \int_{x=0}^{\infty} \exp\left(-\mu_{\text{data}} x\right)\left(r_{\text{data}} x \int_{\zeta=0}^{\infty} T_d(\zeta, x) \mathrm{d}\zeta\right) \mathrm{d}x.$$

For the case of unrestricted channel assignments, $\widetilde{\mathbf{R}}^{c}_{\text{data}}\left(\underset{\sim}{x}\right)$ can be obtained using the following closed-form expression for the deconditioned Laplace–Stieltjes transform $\underset{\sim}{T}(\zeta, x)$ as derived in [9]:

$$\begin{aligned}
\widetilde{T}(\zeta, x) \equiv \mathbf{E}\left\{\exp\{-\zeta\tau(x)\}\right\} &= \sum_{d=1}^{\infty} \left(\frac{\pi(d-1)}{\sum\limits_{d'=1}^{\infty} \pi(d'-1)}\right) \widetilde{T}_d(\zeta, x) \\
&= \frac{\left(1-\rho^{\star}_{\text{data}}\right)\left(1-\rho^{\star}_{\text{data}} r^2\right) \exp\left\{-\left(\lambda_{\text{data}}(1-r)+\zeta\right)x\right\}}{\left(1-\rho^{\star}_{\text{data}} r\right)^2 - \rho^{\star}_{\text{data}}(1-r)^2 \exp\left\{-\mu x\left(1-\rho^{\star}_{\text{data}} r^2\right)/r\right\}},
\end{aligned}$$

with $\text{Re}(\zeta) \geq 0$ and r given by

$$r = \frac{(\lambda_{\text{data}} + \mu_{\text{data}} + \zeta) - \sqrt{(\lambda_{\text{data}} + \mu_{\text{data}} + \zeta)^2 - 4\lambda_{\text{data}}\mu_{\text{data}}}}{2\lambda_{\text{data}}},$$

so that the conditional expected (call-average) data throughput is given by

$$\widetilde{\mathbf{R}}^{c}_{\text{data}}(x) = \sum_{d=1}^{\infty} \pi(d-1)\left(r_{\text{data}} x \int_{\zeta=0}^{\infty} \widetilde{T}_d(\zeta, x) \mathrm{d}\zeta\right)$$

$$= r_{\text{data}} x \int_{\zeta=0}^{\infty} \left(\sum_{d=1}^{\infty} \pi(d-1) \widetilde{T}_d(\zeta, x) \right) \mathrm{d}\zeta$$
$$= r_{\text{data}} x \int_{\zeta=0}^{\infty} \widetilde{T}(\zeta, x) \mathrm{d}\zeta$$
$$= r_{\text{data}} x \int_{\zeta=0}^{\infty} \frac{(1-\rho)\left(1-\rho r^2\right) \exp\{-(\lambda(1-r)+\zeta)x\}}{(1-\rho r)^2 - \rho(1-r)^2 \exp\left\{-\mu x\left(1-\rho r^2\right)/r\right\}} \mathrm{d}\zeta.$$

3.4.2. Time-average throughput

Evaluating definition (2) for the D model, the *time-average* data throughput is given by

$$\mathbf{R}^t_{\text{data}} = r_{\text{data}} \sum_{d=1}^{d_{\max}} \left(\frac{\pi(d)}{\sum\limits_{d'=1}^{d_{\max}} \pi(d')} \right) \beta_{\text{data}}(d).$$

In the case of unrestricted channel assignments, this simplifies to

$$\widetilde{\mathbf{R}}^t_{\text{data}} = r_{\text{data}} \sum_{d=1}^{\infty} \left(\frac{\left(1-\rho^\star_{\text{data}}\right)\left(\rho^\star_{\text{data}}\right)^d}{\sum\limits_{d'=1}^{\infty} \left(1-\rho^\star_{\text{data}}\right)\left(\rho^\star_{\text{data}}\right)^{d'}} \right) \frac{C}{d}$$
$$= r_{\text{data}} C \left(\frac{1-\rho^\star_{\text{data}}}{\rho^\star_{\text{data}}} \right) \sum_{d=1}^{\infty} \left(\frac{\left(\rho^\star_{\text{data}}\right)^d}{d} \right)$$
$$= r_{\text{data}} C \left(\frac{1-\rho^\star_{\text{data}}}{\rho^\star_{\text{data}}} \right) \ln\left(\frac{1}{1-\rho^\star_{\text{data}}} \right),$$

requiring $\rho^\star_{\text{data}} < 1$ for stability. Note that due to the insensitivity of the equilibrium distribution, these expressions for the time-average throughput are also insensitive to the specific form of the data call size distribution.

3.4.3. Expected instantaneous throughput

It is readily seen that the D model definition (3) for the expected *instantaneous* data throughput yields

$$\mathbf{R}^i_{\text{data}} = r_{\text{data}} \sum_{d=1}^{d_{\max}} \left(\frac{\pi(d-1)}{\sum\limits_{d'=1}^{d_{\max}} \pi(d'-1)} \right) \beta_{\text{data}}(d). \tag{13}$$

Only in the special case of unrestricted channel assignments, the expression for the expected instantaneous data throughput is equal to that for the time-average data throughput:

$$\widetilde{\mathbf{R}}^i_{\text{data}} = r_{\text{data}} \sum_{d=1}^{\infty} \left(\frac{\left(1-\rho^\star_{\text{data}}\right)\left(\rho^\star_{\text{data}}\right)^{d-1}}{\sum\limits_{d'=1}^{\infty} \left(1-\rho^\star_{\text{data}}\right)\left(\rho^\star_{\text{data}}\right)^{d'-1}} \right) \frac{C}{d}$$
$$= r_{\text{data}} C \left(\frac{1-\rho^\star_{\text{data}}}{\rho^\star_{\text{data}}} \right) \sum_{d=1}^{\infty} \frac{1}{d} \left(\rho^\star_{\text{data}}\right)^d$$
$$= r_{\text{data}} C \left(\frac{1-\rho^\star_{\text{data}}}{\rho^\star_{\text{data}}} \right) \ln\left(\frac{1}{1-\rho^\star_{\text{data}}} \right),$$

requiring $\rho^\star_{\text{data}} < 1$ for stability. The equality of $\widetilde{\mathbf{R}}^t_{\text{data}}$ and $\widetilde{\mathbf{R}}^i_{\text{data}}$ can be understood from the geometric equilibrium distribution for the number of present data calls (which is stressed to only hold in the case of unrestricted channel assignments) with the associated memorylessness property, and the PASTA property for Poisson arrivals. Once again, the above expressions for the expected instantaneous throughputs inherit the insensitivity property of the equilibrium distribution.

3.4.4. Ratio throughput measure

Following definition (4), the *ratio* of the expected data call size and the expected data call sojourn time is equal to

$$\mathbf{R}^r_{\text{data}} = \left(\frac{r_{\text{data}}}{\mu_{\text{data}}}\right) \Bigg/ \left(\frac{\sum_{d=0}^{d_{\max}} d\pi(d)}{\lambda_{\text{data}}(1-\mathbf{P}_{\text{data}})}\right) = r_{\text{data}} \frac{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}{\sum_{d=0}^{d_{\max}} d\pi(d)},$$

again applying Little's formula. In the case of unrestricted channel assignments we have

$$\widetilde{\mathbf{R}}^r_{\text{data}} = r_{\text{data}} \frac{\rho_{\text{data}}}{\sum_{d=0}^{\infty} d\left(1-\rho^\star_{\text{data}}\right)\left(\rho^\star_{\text{data}}\right)^d} = r_{\text{data}} \frac{C}{\left(1-\rho^\star_{\text{data}}\right)\sum_{d=0}^{\infty} d\left(\rho^\star_{\text{data}}\right)^{d-1}} = r_{\text{data}} C\left(1-\rho^\star_{\text{data}}\right),$$

requiring $\rho^\star_{\text{data}} \leq 1$. Both expressions are insensitive to the data call size distribution, aside from its mean.

3.4.5. Call-average stretch

The call-average *stretch* is given by

$$\mathbf{S}_{\text{data}} = \mathbf{E}\left\{\mathbf{S}_{\text{data}}(x)\right\} = C\,\mathbf{E}\left\{\frac{\mathbf{T}_{\text{data}}(x)}{x}\right\} = C\mathbf{E}\left\{\frac{1}{x}\left(x\frac{\sum_{d=0}^{d_{\max}} d\pi(d)}{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}\right)\right\} = \frac{\sum_{d=0}^{d_{\max}} d\pi(d)}{\rho^\star_{\text{data}}(1-\mathbf{P}_{\text{data}})},$$

using the known linearity in x of the conditional expected sojourn time $\mathbf{T}_{\text{data}}(x)$ of a data call of size x [10,36]. The call-average stretch for the case of unrestricted channel assignments is readily derived to be equal to

$$\widetilde{\mathbf{S}}_{\text{data}} = \frac{1}{1-\rho^\star_{\text{data}}},$$

requiring $\rho^\star_{\text{data}} < 1$ for stability. Note that the effect of the channel rate r_{data} is captured only in the definition of the data traffic load $\rho^\star_{\text{data}}$.

3.4.6. Comparison of measures

We now present a number of results on relations between the different throughput measures derived above. Our first result relates the call average throughput and the ratio throughput measure.

Theorem 4. *For the* D *model,*

$$\mathbf{R}^c_{\text{data}} \geq \mathbf{R}^r_{\text{data}}. \tag{14}$$

Proof. The result is a straightforward extension of the equivalent result given in [18] for the case of unrestricted channel assignments. Applying Jensen's inequality (see e.g. [35]) with convex mapping $\psi(x) \equiv 1/x$:

$$\begin{aligned}\mathbf{R}^c_{\text{data}} &= r_{\text{data}}\mathbf{E}\left\{\psi\left(\frac{\mathbf{T}_{\text{data}}(x)}{x}\right)\right\} \\ &\geq r_{\text{data}}\psi\left(\mathbf{E}\left\{\frac{\mathbf{T}_{\text{data}}(x)}{x}\right\}\right) = r_{\text{data}}\left(\mathbf{E}\left\{\frac{1}{x}\left(x\frac{\sum_{d=0}^{d_{\max}} d\pi(d)}{\rho_{\text{data}}(1-\mathbf{P}_{\text{data}})}\right)\right\}\right)^{-1}\end{aligned}$$

$$= r_{\text{data}} \frac{\rho_{\text{data}}(1 - \mathbf{P}_{\text{data}})}{\sum_{d=0}^{d_{\max}} d\pi(d)} = \mathbf{R}^{r}_{\text{data}}. \quad \blacksquare$$

We further adopt the following result for the case of unrestricted channel assignments and deterministic data call sizes.

Theorem 5 (*Kherani and Kumar [18]*). *In case of deterministic data call sizes, the following inequality holds:*

$$\widetilde{\mathbf{R}}^{t}_{\text{data}} > \widetilde{\mathbf{R}}^{c}_{\text{data}}. \tag{15}$$

Lastly, the explicitly derived expressions above revealed that, *only* for the case of unrestricted channel assignments, the time-average throughput is equal to the expected instantaneous throughput:

$$\widetilde{\mathbf{R}}^{t}_{\text{data}} = \widetilde{\mathbf{R}}^{i}_{\text{data}},$$

while in general it holds that

$$\mathbf{R}^{r}_{\text{data}}\mathbf{S}_{\text{data}} = \widetilde{\mathbf{R}}^{r}_{\text{data}}\widetilde{\mathbf{S}}_{\text{data}} = r_{\text{data}}C.$$

4. Numerical experiments

In this section, we present the results from a set of numerical experiments, carried out in order to provide further insight in the throughput performance of elastic (video or data) calls in a system with a fixed or varying service capacity. As an example setting used for the presented experiments, we have selected the wireless environment of a GSM/GPRS cell, although we argue that the revealed qualitative trends are unaffected by the actual parameter settings and thus apply to other contexts equally well. The applied system and traffic parameter settings are summarised in Section 4.1 below. Subsequently, Section 4.2 presents a numerical evaluation of the conditional expected throughput in the V and D models as a function of the (exponentially distributed) elastic call size, the number of competing elastic calls found upon admission and the CAC threshold. In Section 4.3, an extensive numerical comparison is presented of the various (unconditional) throughput measures in the (S)V and (S)D models, considering different elastic call size distributions where relevant. As the results will demonstrate, the expected instantaneous throughput is the only (average) throughput measure that closely approximates the call-average throughput for all considered scenarios. Finally, Section 4.4 presents some results on the coefficient of variation associated with the distinct throughput measures. Although the principal focus of the paper is on througput *averages*, these results are included to assess whether the qualitative conclusions obtained for averages extend to higher moments.

4.1. Parameter settings

The system and traffic parameter settings applied for the numerical experiments are summarised in Table 1. As stated above, the parameter settings are based on the example context of a single GSM/GPRS cell. The number of traffic channels C in the integrated services SV/SD models is based on a cell with 22 traffic channels (corresponding to 3 GSM frequencies *minus* 2 control channels). The capacity selected for the single service V/D models is equal to the average number of idle traffic channels in the SV/SD models, i.e. $22 - \rho_{\text{speech}} \left(1 - \mathbf{P}_{\text{speech}}\right)$, where ρ_{speech} is chosen such the corresponding speech call blocking probability is 1%. The speech call durations are exponentially distributed. An average call duration of 50 s is assumed for both the speech and video service. The average data file transfer is set at 320 kbits, which normalises to the given expected duration of μ^{-1}_{data} s. The video (data) bit rate per traffic channel is set to 13.4(9.05) kbits/s, based on an assumed GPRS coding scheme CS-2 (CS-1). The video and data traffic loads are varied between 0 and the applicable value of C. Potential practical upper bounds on the channel assignment are disregarded. In the conditional throughput analyses for the V/D models, the minimum QOS requirements are varied within the range $[0, C]$, so that corresponding CAC thresholds between 1 and ∞ are considered, while no such restrictions are imposed for the unconditional throughput analyses.

Table 1
Summary of the parameter settings assumed for the numerical experiments, based on the chosen context of a single cell in a GSM/GPRS network

	SV model	V model	SD model	D model
C	22	8.486	22	8.486
μ_{speech}^{-1}	50 s	–	50 s	–
ρ_{speech}	13.651 Erlang	–	13.651 Erlang	–
μ_{video}^{-1}	50 s	50 s	–	–
ρ_{video}	$\in (0, C)$	$\in (0, C)$	–	–
r_{video}	13.4 kbits/s	13.4 kbits/s	–	–
$\beta_{\text{video}}^{\min}$	0 channels	$\in [0, C]$ channels	–	–
μ_{data}^{-1}	–	–	35.359 s	35.359 s
ρ_{data}	–	–	$\in (0, C)$	$\in (0, C)$
r_{data}	–	–	9.05 kbits/s	9.05 kbits/s
$\beta_{\text{data}}^{\min}$	–	–	0 channels	$\in [0, C]$ channels
$\beta_{\text{GPRS}}^{\max}$	C	C	C	C

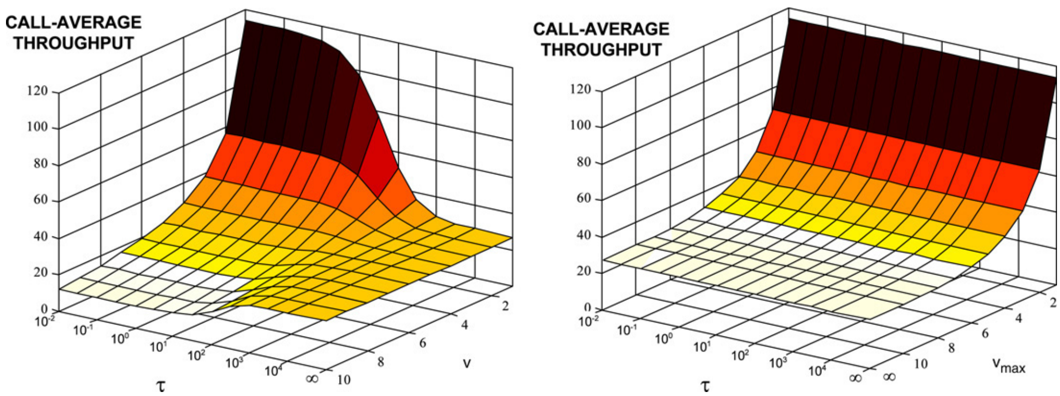

Fig. 1. Conditional expected throughput performance in the V model. The left chart shows the call-average throughput of a tagged video call as a function of its duration τ and the number of video calls v found upon admission, and the right chart shows it as a function of the duration τ and the CAC threshold $v_{\max}$.

4.2. Conditional throughput performance

We now present the results of the numerical conditional throughput analyses that have been carried out for the single service V and D models, respectively.

V model Fig. 1 shows the conditional call-average video throughputs (in kbits/s) for the case of exponentially distributed video call durations and $\rho_{\text{video}} = \frac{1}{2}C = 11$. A logarithmic scale is used for the video call duration τ (expressed in seconds). The results in the left chart assume a CAC threshold of $v_{\max} = 10$, which is achieved by setting $\beta_{\text{video}}^{\min} \in (0.7715, 0.8486]$, and leads to a video call blocking probability of $\mathbf{P}_{\text{video}} = 0.0075$. The depicted curve for $\mathbf{R}_{\text{video}}^{c}(v, \tau)$ is obtained using a special case of the result presented in (9), i.e. without speech traffic. As $\tau \downarrow 0$, the call-average throughputs conditional on the system state v upon admission approach $r_{\text{video}}\beta_{\text{video}}(v) = 113.7023/v$. As τ increases, the impact of the system state upon admission vanishes, and for each v the call-average throughput converges towards the time-average video throughput in a system with one permanent video call, which was seen to be equal to $\mathbf{R}_{\text{video}}^{c}$, i.e. the call-average video throughput in the original model without a permanent video call, h.l. equal to 26.6132. Observe that for low (high) v, convergence is from above (below), in accordance with intuition.

The right chart shows $\mathbf{R}_{\text{video}}^{c}(\tau)$ for $\beta_{\text{video}}^{\min} \in [0, C]$ and hence $v_{\max} \in \{1, 2 \ldots, \infty\}$. The corresponding video call blocking probabilities are as follows:

$v_{\max}$	1	2	3	4	5	10	∞
$\mathbf{P}_{\text{video}}$	0.8093	0.6319	0.4719	0.3336	0.2206	0.0075	0.0000

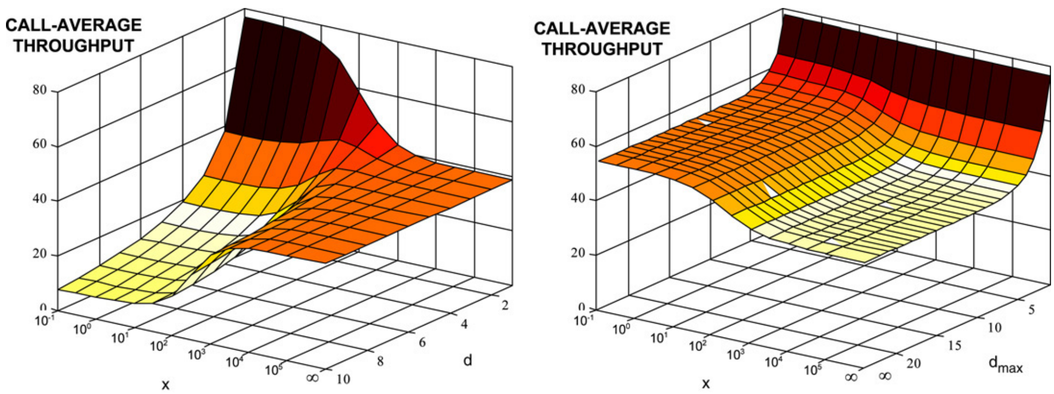

Fig. 2. Conditional expected throughput performance in the D model. The left chart shows the call-average throughput of a tagged data call as a function of its size x and the number of data calls d found upon admission, and the right chart shows it as a function of the size x and the CAC threshold $d_{\max}$.

Observe that, in accordance with the exact result demonstrated in Section 3, $\mathbf{R}^c_{\text{video}}(\tau)$ is independent of the video call duration τ, which reflects the equivalence of the expected instantaneous throughput and call-average throughput measures. For $v_{\max} = 1$, the call-average video throughput is trivially equal to the aggregate service rate $r_{\text{video}}C = 113.7023$, while for $v_{\max} \to \infty$ the conditional video throughput decays exponentially to

$$r_{\text{video}}C\left(\frac{1-\exp(-\rho_{\text{video}})}{\rho_{\text{video}}}\right) = 26.4149.$$

Note that the case for $v_{\max} = 10$ is identical to the converged values in the left chart (for $\tau \to \infty$).

D *model* Fig. 2 shows the conditional call-average data throughputs in the D model, for the case of exponentially distributed data call sizes and $\rho_{\text{data}} = \frac{1}{2}C = 11$ ($\rho^\star_{\text{data}} = 0.5$). Equivalent to the above experiment for the V model, the results for $\mathbf{R}^c_{\text{data}}(d, x)$ (with x expressed in nominal transfer seconds, as explained in Section 2) in the left chart assume a CAC threshold of $d_{\max} = 10$, which is achieved by setting $\beta^{\min}_{\text{data}} \in (0.7715, 0.8486]$. At the considered data traffic load, the selected CAC threshold causes virtually no data call blocking. The profile of the left chart is very similar to that of the left chart in Fig. 1: $\lim_{x\downarrow 0} \mathbf{R}^c_{\text{data}}(d, x)$ is given by the instantaneous throughput $r_{\text{data}}\beta_{\text{data}}(d) = 76.7915/d$, while $\lim_{x\to\infty} \mathbf{R}^c_{\text{data}}(d, x)$ is independent of d and given by the time-average data throughput in a data-only system with one permanent call, readily derived to be

$$r_{\text{data}}C\frac{\left(1-\rho^\star_{\text{data}}\right)\left(1-\left(\rho^\star_{\text{data}}\right)^{d_{\max}}\right)}{\left(1-\left(\rho^\star_{\text{data}}\right)^{d_{\max}+1}\right)-\left(d_{\max}+1\right)\left(\rho^\star_{\text{data}}\right)^{d_{\max}}\left(1-\rho^\star_{\text{data}}\right)} = 38.5843. \quad (16)$$

In contrast with the V model, in the D model the time-average throughput in the modified Markov chain with one permanent data call is *not* equal to the call-average throughput in the original Markov chain.

The right chart shows $\mathbf{R}^c_{\text{data}}(x)$ for various CAC thresholds $d_{\max} \in \{1, 2, \ldots, \infty\}$, with the corresponding data call blocking probabilities given by

$d_{\max}$	1	2	3	4	5	10	∞
$\mathbf{P}_{\text{data}}$	0.3333	0.1429	0.0667	0.0323	0.0159	0.0005	0.0000

In the trivial case of $d_{\max} = 1$, the call-average data throughput is equal to the aggregate service rate $r_{\text{data}}C = 76.7915$, independent of the data call size x. As $d_{\max}$ increases, not only does $\mathbf{R}^c_{\text{data}}(x)$ decrease due to an increased carried data traffic load and hence a greater competition for resources, it is also no longer independent of x. For a given CAC threshold of $d_{\max}$, $\mathbf{R}^c_{\text{data}}(x)$ decreases from the corresponding expected instantaneous data throughput $\mathbf{R}^i_{\text{data}}$ (cf. expression (13)) to the expected time-average data throughput in the associated modified Markov chain with one permanent data call (cf. expression (16)). Observe that the expected instantaneous throughput is an upper bound for the call-average throughput. Unlike in the V model, in the D model small calls experience a higher throughput than

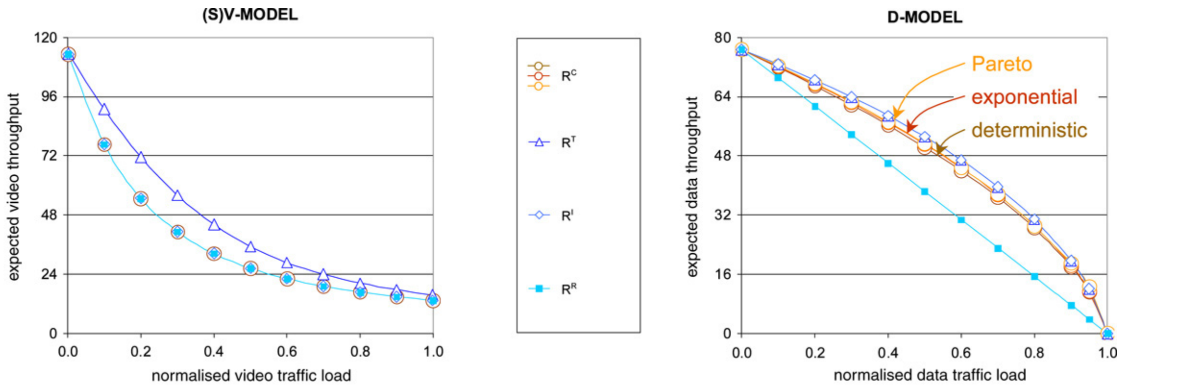

Fig. 3. Comparison of different throughput measures in the SV, V and D models. The (insensitive) throughput measures in the left chart are identical for the SV and V models, given an appropriately normalised video traffic load. The right chart depicts, for the D model, the insensitive $\mathbf{R}^t_{\text{data}}$, $\mathbf{R}^i_{\text{data}}$ and $\mathbf{R}^r_{\text{data}}$ measures, along with the sensitive $\mathbf{R}^c_{\text{data}}$ measure for three distinct data call size distributions.

large calls. It is stressed, however, that the expected sojourn time is proportional in the data call size, so that the expected stretch is insensitive to the data call size. The potential confusion is due to the fact that the reciprocal of the expectation of a random variable is generally unequal to the expectation of the reciprocal of that random variable. Observe that the expected instantaneous throughput is an upper bound for the call-average throughput.

4.3. Unconditional throughput performance

We now concentrate on the unconditional throughput as a function of the elastic traffic load, with a principal focus on the proximity of the various throughput measures in the different PS models.

(S)V *model* Consider the SV and V models. Fig. 3 depicts the various (unconditional) throughput performance measures as functions of the normalised elastic traffic load. In all considered cases, no channel assignment restrictions have been imposed on the elastic services. The left chart covers both the SV and the V models, for which all throughput measures are identical for any given normalised video traffic load $\rho^\star_{\text{video}} \equiv \rho_{\text{video}}/C$, with C appropriately chosen in each model (see Table 1). The chart reveals both the demonstrated equality of $\mathbf{R}^c_{\text{video}}$, $\mathbf{R}^i_{\text{video}}$ and $\mathbf{R}^r_{\text{video}}$, and the proven ordering of $\mathbf{R}^t_{\text{video}} \geq \mathbf{R}^c_{\text{video}}$. It can be observed from the numerical results that $\mathbf{R}^t_{\text{video}}$ may exceed $\mathbf{R}^c_{\text{video}}$ by more than 36%.

D *model* The right chart of Fig. 3 concentrates on the D model. Since (only) the call-average throughput measure $\mathbf{R}^c_{\text{data}}$ is sensitive to the data call size distribution and no explicit expression could be derived, three distinct curves have been obtained via dynamic simulations for deterministic (zero variance), exponential and Pareto (with shape parameter $\alpha = 1.35$: infinite variance) data call size distributions. Sufficient numerical accuracy is ensured in the simulation experiment, indicated by a relative precision of the constructed 95% confidence intervals that is no worse than 5%. Observe that the call-average throughput is higher for more variable data call sizes, as also observed in [18], although the discrepancies are extremely small. This is probably due to the fact that a more variable data call size distribution features a relatively large number of small data calls, which appear to experience higher throughputs than large data calls (cf. the right chart of Fig. 2).

As shown in Section 3, the insensitive time-average and expected instantaneous throughput measures are identical, and appear to offer a very good, only slightly overestimating (cf. (15)), approximation for the call-average throughput. Finally, $\mathbf{R}^r_{\text{data}}$ significantly underestimates the call-average throughput (cf. (14)), for high data traffic loads even by a factor exceeding 2.

SD *model* For the SD model, all the throughput measures are more or less sensitive to the data call size distribution, so that for reasons of clarity the numerical results are presented in the two separate charts of Fig. 4 (for each marker in the legend, the left (right) throughput measure is depicted in the left (right) chart). In all cases, observe again that a more variable data call size distribution appears to lead to higher expected throughputs, which is in agreement with the sojourn time results of [23]. In this data model with varying service capacity, both the time-average throughput ($\mathbf{R}^t_{\text{data}}$) and the ratio of the expected data call size and the expected sojourn time ($\mathbf{R}^r_{\text{data}}$) are significantly lower than

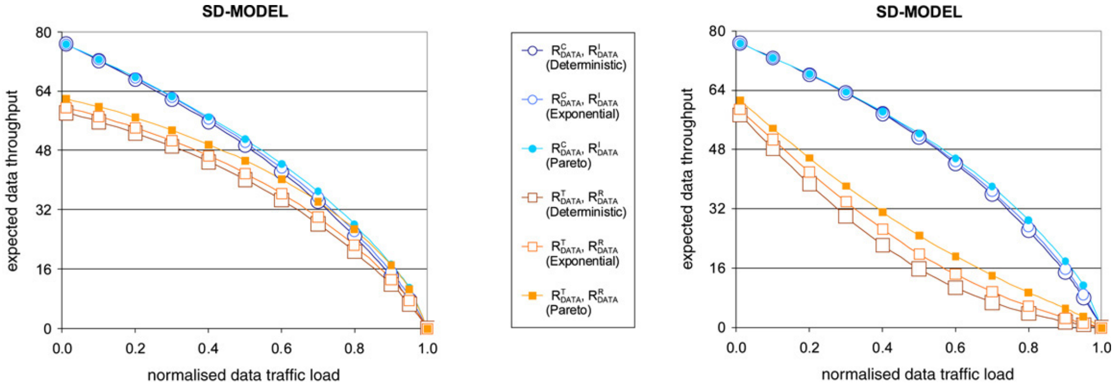

Fig. 4. Comparison of different throughput measures in the SD model. All throughput measures are sensitive to the data call size distributions. The performance induced by three distinct distributions is shown.

the call-average throughput ($\mathbf{R}^c_{\text{data}}$), in particular for lower data traffic loads. In contrast, the expected instantaneous throughput ($\mathbf{R}^i_{\text{data}}$) remains a very good and fairly insensitive approximation for $\mathbf{R}^c_{\text{data}}$, across the entire range of data traffic loads. The *slight over*estimation of the call-average throughput seems to be not significant enough to lead to perilously loose Call Admission Control schemes or planning guidelines.

Comparing the throughput results for the D and SD models, observe that the call-average data throughput appears to be fairly insensitive to the variability of the available capacity, as also observed in [12] (recall that for the SV and V models, the call-average video throughputs were identical). Only for heavy data traffic loads, is the call-average data throughput non-negligibly higher for the fixed capacity D model.

In order to get a better grasp on the large discrepancy between e.g. the time- and call-average data throughputs in the SD model, the left chart of Fig. 5 shows the time-average data throughput versus the normalised data traffic loads for various degrees of acceleration of the speech call arrival and departure process. Keeping ρ_{speech} fixed at 13.651 Erlang, we multiply both λ_{speech} and μ_{speech} by the acceleration factor $\vartheta \in \{1, 10, 100, \infty\}$. The case of $\vartheta = 1$ refers to the original model, and the associated curve is identical to the one for $\mathbf{R}^t_{\text{data}}$ in Fig. 4 (left chart). At the other extreme, in the case of $\vartheta \to \infty$, the speech calls arrive and depart so quickly that from the perspective of the data traffic, the available capacity is deterministic at $C - \rho_{\text{speech}} \left(1 - \mathbf{P}_{\text{speech}}\right)$, and hence the accelerated model corresponds with the D model. As a consequence, the associated curve is identical to the one for $\mathbf{R}^t_{\text{data}}$ in Fig. 3 (right chart). Observe that as the capacity fluctuation process is accelerated, i.e. when ϑ is increased from 1 to ∞, the time-average throughput curves gradually approach the one corresponding to the extreme case of the D model, and the time-average throughput thus approximates the call-average throughput more and more closely. Additional numerical experiments (not included) indicate that among the different throughput measures, the ratio throughput measure is most sensitive to the degree of speech call dynamics in the SD model. While the call-average and expected instantaneous throughputs are largely insensitive to ϑ, and the time-average throughput converges to a significantly lower, yet positive value as $\vartheta \downarrow 0$, the ratio throughput measure becomes negligible for very small ϑ.

The right chart of Fig. 5 shows the expected stretch of a data call for both the SD and D models. As noted in Section 3, the expected stretch in the D model is insensitive to the data call size distribution. For the SD model, such insensitivity does not hold, as is demonstrated by the three expected stretch curves for deterministic, exponential and Pareto (with shape parameter $\alpha = 1.35$) data call size distributions. In correspondence with the throughput performance, the expected stretch appears to be smaller (better) for more highly variable data call sizes. A noteworthy observation from the numerical experiments that is not included in the figure, is that the expected stretch turns out to be infinitely large for the considered subexponential Weibull data call size distributions, i.e. with coefficient of variation greater than 1, for any data traffic load. In contrast, for highly variable Pareto distributions such as the one included in the figure, the expected stretch was nicely finite within the stable regime of data traffic loads. The probable reason for this phenomenon is that a subexponential Weibull distribution features many very small data calls, which may suffer from excessively large relative sojourn times in the case of a varying service capacity that is even equal to zero at times. Pareto distributions are inherently truncated at the lower end, however, so that extremely small data

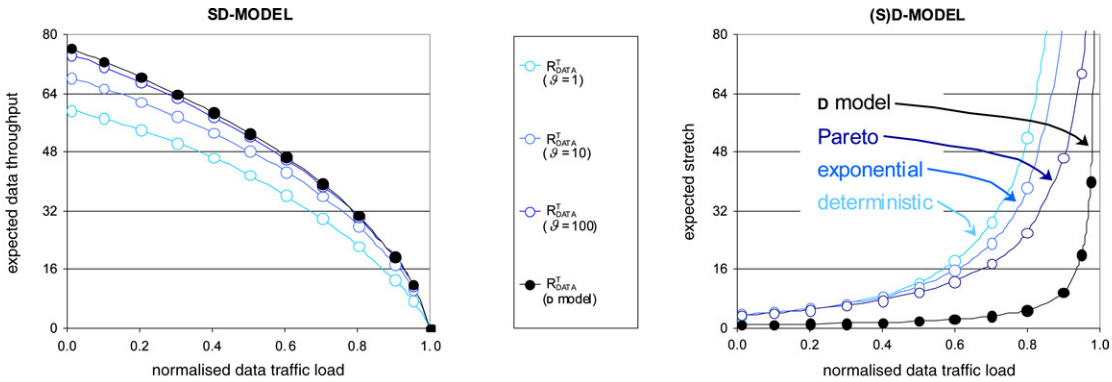

Fig. 5. The impact of acceleration of the speech call arrival and departure process on $\mathbf{R}^t_{\text{data}}$ in the SD model (left chart). The expected stretch performance for different data call size distributions (SD model), as well as the insensitive values for the D model.

calls simply do not occur. In any case, the expected stretch thus appears to be less useful as a measure for throughput performance.

4.4. *Coefficient of variation of the throughput*

Although the focus of the paper is on the analysis and comparison of distinct call- and time-centric *average* throughput measures, in this subsection we present some numerical results for the Coefficients of Variation (CoV) associated with the considered throughput measures. Note that the CoV is defined as the ratio of the standard deviation of the throughput and its average. More specifically, we consider the CoV of the instantaneous, call-[1] and time-centric[2] throughputs, noting that no CoV measure exists that can be associated with the ratio throughput measure. For the (S)V and (S)D models, and three distinct CoV measures, Fig. 6 depicts the numerical results, which are analytically obtained when possible, or via dynamic simulations otherwise. For instance, for the SV model, closed-form expressions for the CoV of the instantaneous and time-centric throughput are readily derived from the equilibrium distribution. As was shown to be the case for the associated averages, these CoVs are different and insensitive to the call size distribution. Other CoV measures that can be derived analytically are the CoV of the instantaneous and time-centric throughputs in the D model, which appear to be identical and also insensitive to the call size distribution, which was also derived to hold for the associated averages.

Considering the numerical results depicted in the figure, a number of observations can be made. Note first the distinct trends of the curves associated with the (S)V and (S)D models, respectively, which reflect the net effect of a generally decreasing trend of both the standard deviations and the averages of all throughput measures in all models. Apparently, for the CoV in the (S)V model, the decreasing trend of the standard deviation is dominant for moderate to heavy traffic loads, while in the (S) D model, the *exponential* decline of the average throughput as the data traffic load grows (see also Fig. 4) outweighs the milder decline of the standard deviation. Another general observation that can be made is that the CoV of the elastic calls' throughput is generally larger in the models with speech traffic, as the varying presence of speech traffic provides an additional source of throughput variation, besides the variation that is due to the competition among elastic calls themselves.

Comparing the CoV curves for the distinct throughput measures, we observe that for the D model, the CoV appears to be rather insensitive to both the applied throughput measure and the call size distribution. Furthermore, for the SD model, the CoV of the instantaneous throughput appears to be very close to that of the call-centric throughput, while the CoV of the time-centric throughput is generally slightly higher. For (S)V model, neither the CoV of the time-centric, nor the one of the instantaneous throughput appear to be very good approximations for the CoV of the call-centric throughput, except for very low video traffic loads. For moderate-to-high video traffic loads, the CoV of the instantaneous throughput, however, still appears to offer the closest approximation among the readily attainable

[1] Cf. 'call-average'.
[2] Cf. 'time-average'.

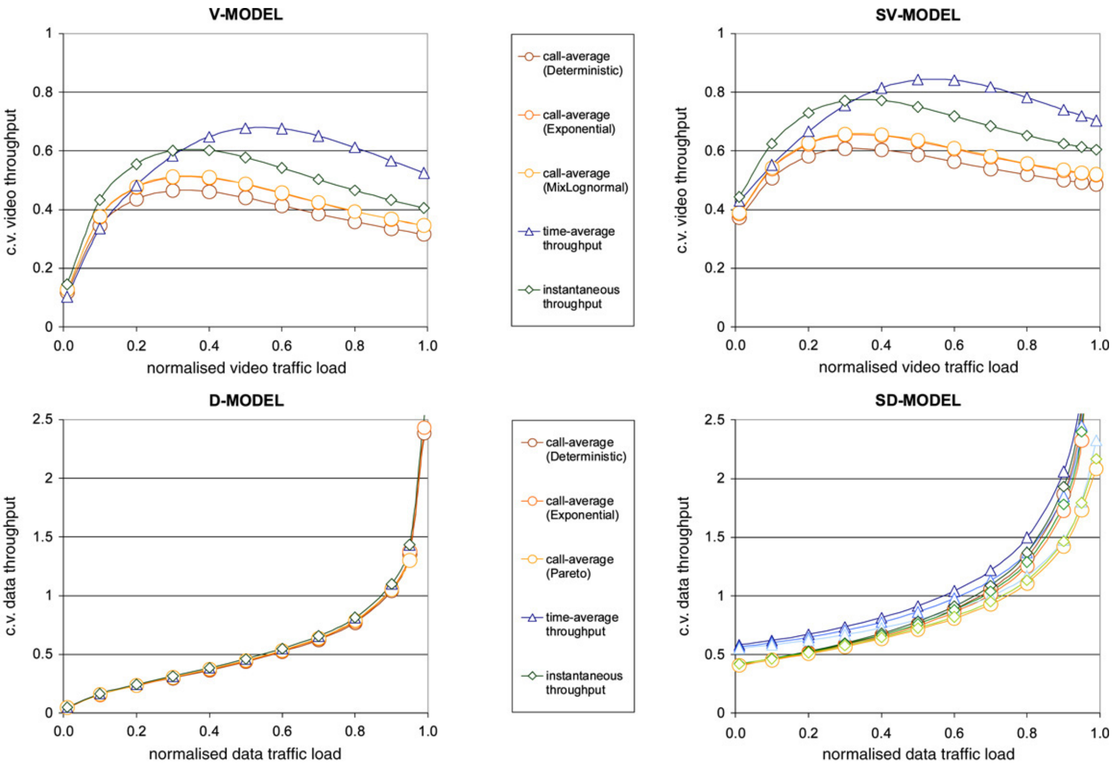

Fig. 6. Comparison of the Coefficient of Variation associated with the different considered throughput measures in the V, SV, D and SD models. The CoV of the instantaneous and time-centric throughput measures in the (S)V and D models are insensitive to the data call size distributions. For the CoV measures that are not insensitive, the performance induced by three distinct distributions is shown. For those curves, the different applied shades indicate the assumed call size/duration distributions, as only explicitly presented in the legend for the call-average throughput.

measures. Also, observe that the approximation offers an upper bound and thus, when used for network dimensioning purposes, is expected to yield conservative guidelines.

5. Concluding remarks

In this paper we have specified, derived and compared, both analytically and numerically, a set of throughput measures in PS queuing systems modelling a communication link carrying elastic video or data calls. The available service capacity was either fixed or randomly varying, corresponding to an integrated services network link, where elastic calls utilise the capacity left idle by prioritised speech traffic. The call-average throughput is arguably the most appropriate indicator of the experienced average Quality Of Service, which, however, for models involving elastic calls of the data type, is hard to determine analytically. Among the alternative throughput measures, the newly proposed and readily analytically derived expected instantaneous throughput is the only measure which excellently approximates (or is even equal to) the call-average throughput in all considered system models and across the entire range of considered elastic traffic loads. In particular, for the practically most relevant model integrating speech and data traffic, other typically applied throughput measures such as the time-average throughput or the ratio of the expected call size and the expected sojourn time, significantly underestimate the call-average throughput. An intuitive reason for the generally (near-)perfect fit of the expected instantaneous throughput is that apparently, the throughput an elastic call experiences immediately upon arrival is an excellent predictor of what the call is likely to experience throughout its lifetime. Moreover, among the considered alternative throughput measures, the expected instantaneous throughput is the *only* measure that is truly *call*-centric. Considering higher moments, the instantaneous throughput again generally provides the most adequate predictor for the coefficient of variation of the call-centric throughput, although these approximations are not always as accurate as in the case of throughput averages.

The analytical evaluation further revealed that the expected call-average throughput of elastic video calls in the considered PS models is *insensitive* to both the variability of the available capacity and the call duration distribution, while the numerical experiments indicated that this insensitivity property also holds for the data service to a considerable degree. As seen in [23], the latter insensitivity does not hold if the data performance is measured by the (conditional) expected sojourn time.

Acknowledgments

The authors would like to thank the anonymous referees for their constructive comments. This research has been carried out partly within the project BEYOND 3G supported by the Dutch Ministry of Economic Affairs under the program Technologische Samenwerking ICT-Doorbraakprojecten, project number TSIT1025. The research is partly supported by the Technology Foundation STW, Applied Science Division of NWO and the technology programme of the Dutch Ministry of Economic Affairs, The Netherlands.

Appendix A. Proof of Theorem 1

Proof. Define the Laplace–Stieltjes transform of the distribution of $x_{s,v}(\tau)$ by

$$X_{s,v}(\zeta,\tau) \equiv \mathbf{E}\left\{\exp\left\{-\zeta x_{s,v}(\tau)\right\}\right\}, \quad \mathrm{Re}(\zeta) \geq 0,\ (s,v) \in \mathbb{S}^{+}_{\text{video}},$$

and let $\mathbf{X}(\zeta,\tau)$ be the vector with the $X_{s,v}(\zeta,\tau)$, ordered lexicographically in $(s,v) \in \mathbb{S}^{+}_{\text{video}}$. Using marginal analysis, we will first derive a differential equation and initial condition for $\mathbf{X}(\zeta,\tau)$, for $\tau \geq 0$ and $\mathrm{Re}(\zeta) \geq 0$. Consider a time interval of length $\Delta > 0$,, with Δ sufficiently small such that the tagged video call cannot terminate within this time, i.e. $\Delta < \tau$. Condition on all the possible events occurring in this interval, starting in state $(s,v) \in \mathbb{S}^{+}_{\text{video}}$. For notational convenience and readability, the boundary constraints are not explicitly considered. Equations for the boundary can be derived by analogy with the results below.

$$\begin{aligned}
X_{s,v}(\zeta,\tau) &\equiv \mathbf{E}\left\{\exp\left\{-\zeta x_{s,v}(\tau)\right\}\right\} \\
&= \lambda_{\text{speech}}\Delta \exp[-\zeta(r_{\text{video}}\beta_{\text{video}}(s,v)(\Delta - O(\Delta)) \\
&\quad + r_{\text{video}}\beta_{\text{video}}(s+1,v)\,O(\Delta))]X_{s+1,v}(\zeta,\tau-\Delta) \\
&\quad + s\mu_{\text{speech}}\Delta \exp[-\zeta(r_{\text{video}}\beta_{\text{video}}(s,v)(\Delta - O(\Delta)) \\
&\quad + r_{\text{video}}\beta_{\text{video}}(s-1,v)\,O(\Delta))]X_{s-1,v}(\zeta,\tau-\Delta) \\
&\quad + \lambda_{\text{video}}\Delta \exp[-\zeta(r_{\text{video}}\beta_{\text{video}}(s,v)(\Delta - O(\Delta)) \\
&\quad + r_{\text{video}}\beta_{\text{video}}(s,v+1)\,O(\Delta))]X_{s,v+1}(\zeta,\tau-\Delta) \\
&\quad + (v-1)\mu_{\text{video}}\Delta \exp[-\zeta(r_{\text{video}}\beta_{\text{video}}(s,v)(\Delta - O(\Delta)) \\
&\quad + r_{\text{video}}\beta_{\text{video}}(s,v-1)\,O(\Delta))]X_{s,v-1}(\zeta,\tau-\Delta) \\
&\quad + (-\lambda_{\text{speech}}\Delta - s\mu_{\text{speech}}\Delta - \lambda_{\text{video}}\Delta - (v-1)\mu_{\text{video}}\Delta) \\
&\quad \times \exp[-\zeta r_{\text{video}}\beta_{\text{video}}(s,v)\,\Delta]\,X_{s,v}(\zeta,\tau-\Delta) \\
&\quad + \left(1 - \zeta r_{\text{video}}\beta_{\text{video}}(s,v)\,\Delta + \sum_{j=2}^{\infty}\frac{(-\zeta r_{\text{video}}\beta_{\text{video}}(s,v)\,\Delta)^j}{j!}\right)X_{s,v}(\zeta,\tau-\Delta) + o(\Delta).
\end{aligned}$$

Rearranging terms and letting $\Delta \downarrow 0$ gives the system of differential equations

$$\begin{aligned}
\frac{\partial X_{s,v}(\zeta,\tau)}{\partial\tau} &= \lambda_{\text{speech}}X_{s+1,v}(\zeta,\tau) + s\mu_{\text{speech}}X_{s-1,v}(\zeta,\tau) \\
&\quad + \lambda_{\text{video}}X_{s,v+1}(\zeta,\tau) + (v-1)\mu_{\text{video}}X_{s,v-1}(\zeta,\tau) \\
&\quad + \left(-\lambda_{\text{speech}} - s\mu_{\text{speech}} - \lambda_{\text{video}} - (v-1)\mu_{\text{video}}\right)X_{s,v}(\zeta,\tau) \\
&\quad - \zeta r_{\text{video}}\beta_{\text{video}}(s,v)\,X_{s,v}(\zeta,\tau),
\end{aligned}$$

using the continuity of $X_{s,v}(\zeta,\tau)$ in τ. This system of differential equations may equivalently be written in the matrix notation as follows:

$$\frac{\partial}{\partial\tau}\mathbf{X}(\zeta,\tau) = \left(\mathcal{Q}^{\star}_{\text{video}} - \zeta r_{\text{video}}\mathcal{B}_{\text{video}}\right)\mathbf{X}(\zeta,\tau). \tag{17}$$

The initial condition

$$\mathbf{X}(\zeta, 0) = \mathbf{1}, \tag{18}$$

simply reflects the fact that the transfer volume $x_{s,v}(0)$ of a video call with a duration of zero seconds equals zero bits:

$$\mathbf{X}(\zeta, 0) = \left(\mathbf{E}\left\{\exp\left\{-\zeta x_{s,v}(0)\right\}\right\}\right)_{(s,v)\in\mathbb{S}^{+}_{\text{video}}} = \mathbf{1}.$$

The *existence* and *uniqueness* of a solution $\mathbf{X}(\zeta, \tau)$ to the system of differential equations (17) with initial condition (18) follows from e.g. [8, Chapter 1, Section 8]. The unique solution is readily observed and verified to be given by

$$\mathbf{X}(\zeta, \tau) = \exp\left\{\tau\left(\mathcal{Q}^{\star}_{\text{video}} - \zeta r_{\text{video}}\mathcal{B}_{\text{video}}\right)\right\}\mathbf{1}. \tag{19}$$

Using this closed-form expression for the Laplace–Stieltjes transform of the distribution of $x_{s,v}(\tau)$, an explicit expression of the conditional expected transfer volume and, consequently, the conditional expected video throughput can now be derived. Recall that $\boldsymbol{\pi}^{\star}_{\text{video}} \equiv \left(\pi^{\star}_{\text{video}}(s,v), (s,v)\in\mathbb{S}^{+}_{\text{video}}\right)$ is the equilibrium probability distribution vector corresponding to the Markov chain with one permanent video call, i.e. $\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{Q}^{\star}_{\text{video}} = \mathbf{0}$, while $\boldsymbol{\gamma}_{\text{video}} \equiv \left(\gamma_{\text{video}}(s,v), (s,v)\in\mathbb{S}^{+}_{\text{video}}\right)$ is the unique solution to the system of linear equations given by (7) and (8). The existence of a vector $\boldsymbol{\gamma}_{\text{video}}$ that satisfies (7), and its uniqueness up to a translation along the vector $\mathbf{1}$, are guaranteed by results in Markov reward chain theory. Interpreting $\boldsymbol{\gamma}_{\text{video}}$ as the vector of relative rewards in a Markov reward chain governed by the infinitesimal generator $\mathcal{Q}^{\star}_{\text{video}}$ and with immediate reward vector $\frac{1}{\eta}r_{\text{video}}\left(\mathcal{B}_{\text{video}}\mathbf{1} - \left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right)$ with η is the maximum rate of change in the Markov chain, and understanding that the long-term average rewards are zero, $\frac{1}{\eta}\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\left(\mathcal{B}_{\text{video}}\mathbf{1} - \left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right) = 0$, [36, Theorem 3.1, page 167] can be applied after a uniformisation of the continuous-time Markov chain. A translation of $\boldsymbol{\gamma}_{\text{video}}$ along the vector $\mathbf{1}$ indeed does not alter the solution, since for any $\alpha \in \mathbb{R}$, $\mathcal{Q}^{\star}_{\text{video}}\left(\boldsymbol{\gamma}_{\text{video}} + \alpha\mathbf{1}\right) = \mathcal{Q}^{\star}_{\text{video}}\boldsymbol{\gamma}_{\text{video}}$, or equivalently, $\left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\left(\boldsymbol{\gamma}_{\text{video}} + \alpha\mathbf{1}\right) = \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}}$, which readily follows from using the Taylor expansion of $\exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}$. The single degree of freedom that exists in choosing $\boldsymbol{\gamma}_{\text{video}}$ in expression (7), is used to normalise $\boldsymbol{\gamma}_{\text{video}}$ as in (8).

The vector of conditional expected transfer volumes $\widehat{\mathbf{x}}(\tau)$ is then obtained by taking the derivative of $\mathbf{X}(\zeta, \tau)$ with respect to ζ, and subsequently setting $\zeta = 0$.

$$\begin{aligned}
\widehat{\mathbf{x}}(\tau) &= -\frac{\partial}{\partial\zeta}\mathbf{X}(\zeta,\tau)\bigg|_{\zeta=0} = -\frac{\partial}{\partial\zeta}\sum_{k=0}^{\infty}\frac{\left(\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right) + \left(-\zeta\tau r_{\text{video}}\mathcal{B}_{\text{video}}\right)\right)^{k}}{k!}\mathbf{1}\bigg|_{\zeta=0} \\
&= -\left(\sum_{k=1}^{\infty}\sum_{i=0}^{k-1}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-i-1}\left(-\tau r_{\text{video}}\mathcal{B}_{\text{video}}\right)\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{i}}{k!}\right)\mathbf{1} = \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\tau r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\left[\tau r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1} - \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1}\right] \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} - \left(\sum_{k=1}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k-1}}{k!}\right)\tau\mathcal{Q}^{\star}_{\text{video}}\boldsymbol{\gamma}_{\text{video}} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left(\mathcal{I} - \sum_{k=0}^{\infty}\frac{\left(\tau\mathcal{Q}^{\star}_{\text{video}}\right)^{k}}{k!}\right)\boldsymbol{\gamma}_{\text{video}} \\
&= \tau\left(\boldsymbol{\pi}^{\star}_{\text{video}}r_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}}
\end{aligned}$$

where after the third equality sign, only those matrix cross-products appear that remain after differentiating the terms in the preceding expression, and setting ζ to 0. The subsequent equality sign uses $\mathcal{Q}^{\star}_{\text{video}}\mathbf{1} = 0$, so that all terms with $i > 0$ disappear. A similar argument is used to obtain the fifth equality. Eq. (7) is used for the sixth equality.

Finally, the conditional expected throughput vector is given by

$$\frac{\widehat{\mathbf{x}}(\tau)}{\tau} = r_{\text{video}}\left(\boldsymbol{\pi}^{\star}_{\text{video}}\mathcal{B}_{\text{video}}\mathbf{1}\right)\mathbf{1} + \frac{1}{\tau}\left[\mathcal{I} - \exp\left\{\tau\mathcal{Q}^{\star}_{\text{video}}\right\}\right]\boldsymbol{\gamma}_{\text{video}}. \quad \blacksquare$$

Appendix B. Proof of Theorem 2

Proof. The stationary joint distribution $\pi\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right)$ of the number of speech (S) and video calls (V) present in the system, and the associated residual call durations $\boldsymbol{\Theta}_{\text{speech}} \equiv (\Theta_{\text{speech}}(1), \ldots, \Theta_{\text{speech}}(S))$ and $\boldsymbol{\Theta}_{\text{video}} \equiv (\Theta_{\text{video}}(1), \ldots, \Theta_{\text{video}}(V))$, is given by (see e.g. [13]):

$$\begin{aligned}\pi\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right) &= \Pr\{S = s, V = v, \boldsymbol{\Theta}_{\text{speech}} \in \left[\boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{speech}} + d\boldsymbol{\vartheta}_{\text{speech}}\right], \\ &\qquad \boldsymbol{\Theta}_{\text{video}} \in [\boldsymbol{\vartheta}_{\text{video}}, \boldsymbol{\vartheta}_{\text{video}} + d\boldsymbol{\vartheta}_{\text{video}}]\} \\ &= G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C\right)\left\{\frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\prod_{s'=1}^{s}\left(\frac{\overline{\Phi}_{\text{speech}}\left(\vartheta_{\text{speech}}\left(s'\right)\right)}{\mu_{\text{speech}}^{-1}}d\vartheta_{\text{speech}}\left(s'\right)\right)\right. \\ &\qquad \left.\times \prod_{v'=1}^{v}\left(\frac{\overline{\Phi}_{\text{video}}\left(\vartheta_{\text{video}}\left(v'\right)\right)}{\mu_{\text{video}}^{-1}}d\vartheta_{\text{video}}\left(v'\right)\right)\right\},\end{aligned}$$

for $(s, v) \in \mathbb{S} = \mathbb{S}(C) \equiv \left\{(s, v) \in \mathbb{N}_0 \times \mathbb{N}_0 : s + v\beta_{\text{video}}^{\min} \le C\right\}$, $\boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}} \ge \mathbf{0}$, where the vectors $d\boldsymbol{\vartheta}_{\text{speech}}$ and $d\boldsymbol{\vartheta}_{\text{video}}$ consist of the infinitesimally small elements

$$G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C\right) \equiv \left(\sum_{(s,v)\in\mathbb{S}(C)} \frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\right)^{-1},$$

and where $\overline{\Phi}_{\text{speech}}$ and $\overline{\Phi}_{\text{video}}$ denote the complementary cumulative distributions of the speech and video call durations, respectively.

As the arrival process is a Poisson process, the joint distribution $\pi_{\text{video}}^{\bullet}\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right)$ of $(S, V, \boldsymbol{\Theta}_{\text{speech}}, \boldsymbol{\Theta}_{\text{video}})$ upon admission of a tagged video call can readily be obtained as the conditional distribution seen by an arriving call, given that it is admitted. Invoking PASTA to obtain the distribution seen by an arriving call, we obtain

$$\begin{aligned}\pi_{\text{video}}^{\bullet}\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right) &= \Pr\{S = s, V = v, \boldsymbol{\Theta}_{\text{speech}} \in \left[\boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{speech}} + d\boldsymbol{\vartheta}_{\text{speech}}\right], \\ &\qquad \boldsymbol{\Theta}_{\text{video}} \in [\boldsymbol{\vartheta}_{\text{video}}, \boldsymbol{\vartheta}_{\text{video}} + d\boldsymbol{\vartheta}_{\text{video}}] \mid s + v\beta_{\text{video}}^{\min} \le C - \beta_{\text{video}}^{\min}\} \\ &= G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C - \beta_{\text{video}}^{\min}\right) \\ &\quad \times \left\{\frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\prod_{s'=1}^{s}\frac{\overline{\Phi}_{\text{speech}}\left(\vartheta_{\text{speech}}\left(s'\right)\right)}{\mu_{\text{speech}}^{-1}}\prod_{v'=1}^{v}\frac{\overline{\Phi}_{\text{video}}\left(\vartheta_{\text{video}}\left(v'\right)\right)}{\mu_{\text{video}}^{-1}}\right\},\end{aligned}$$

for $(s, v) \in \mathbb{S}\left(C - \beta_{\text{video}}^{\min}\right)$, where v *excludes* the newly admitted tagged video call.

Observe that $\pi_{\text{video}}^{\bullet}\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right)$ is equal to the stationary joint distribution of the number of speech and video calls and their residual call durations in a corresponding system with capacity $C - \beta_{\text{video}}^{\min}$ instead of C, or equivalently, in the original system, but with one *permanent* video call (where v *excludes* this call). Hence the system state remains stochastically identical throughout the duration of the tagged video call. The associated (partially deconditioned) system state distribution $\pi_{\text{video}}^{\bullet}(s, v)$ is given by

$$\begin{aligned}\pi_{\text{video}}^{\bullet}(s, v) &= \int_{\vartheta_{\text{speech}}(1)=0}^{\infty}\int_{\vartheta_{\text{speech}}(s)=0}^{\infty}\cdots\int_{\vartheta_{\text{video}}(1)=0}^{\infty}\int_{\vartheta_{\text{video}}(v)=0}^{\infty}\pi_{\text{video}}^{\bullet}\left(s, v, \boldsymbol{\vartheta}_{\text{speech}}, \boldsymbol{\vartheta}_{\text{video}}\right) \\ &= G\left(\rho_{\text{speech}}, \rho_{\text{video}}, C - \beta_{\text{video}}^{\min}\right)\left\{\frac{\rho_{\text{speech}}^{s}}{s!}\frac{\rho_{\text{video}}^{v}}{v!}\right\},\end{aligned} \tag{20}$$

for $(s, v) \in \mathbb{S}\left(C - \beta_{\text{video}}^{\min}\right)$. Since the throughput of the tagged video call is completely determined by the distribution of the number of speech and *other* video calls present during its lifetime, as given in (20), it is now immediately clear that the conditional call-average throughput $\mathbf{R}_{\text{video}}^{c}(\tau)$ of the tagged video call is *independent* of its duration

τ, i.e. $\mathbf{R}^c_{\text{video}}(\tau) = \mathbf{R}^c_{\text{video}}$, for all $\tau \geq 0$. In particular, it is equal to the expected instantaneous video throughput experienced upon admission, which inherits its insensitivity from the insensitivity of $\pi^{\star}_{\text{video}}$ (see also Section 3.1.3). Note that the stationary probability $\pi^{\bullet}_{\text{video}}(s, v)$ is given in expression (20), where v *excludes* the tagged (permanent) video call, is readily verified to be equivalent to the conditional probability $\pi^{*}_{\text{video}}(s, v+1)$ given in expression (6), where v *includes* the newly admitted video call. ■

Appendix C. Proof of Theorem 3

Proof. For the extreme cases of infinitesimally small or infinitely large video traffic loads, it is readily argued that the call- and time-average video throughput measures are identical. Under an extremely *light* video traffic load ($\rho_{\text{video}} \downarrow 0$), a (rarely) occuring system state $(s, v) \in \mathbf{S}^+_{\text{video}}$ must have $v = 1$, almost surely, for both the original stochastic process, and the modified process with one permanent video call. As a consequence, the time-average video throughputs of both processes are identical, and hence so are the call- and time-average video throughputs of the original process. We thus have that

$$\lim_{\rho_{\text{video}} \downarrow 0} \mathbf{R}^t_{\text{video}} = \lim_{\rho_{\text{video}} \downarrow 0} \mathbf{R}^c_{\text{video}}$$

as can readily be verified from (10) and (11).

Alternatively, an infinitely *heavy* video traffic load ($\rho_{\text{video}} \to \infty$, assuming $\beta^{\min}_{\text{video}} > 0$ for stability) leads to a (complete or near) crowding out of speech calls, and implies the everlasting presence of $v_{\max}(0) = \lfloor C_{\text{total}}/\beta^{\min}_{\text{video}} \rfloor \geq 1$ video calls, and hence again the performance of the original and the modified process are the same. In particular, all video throughput measures are identical, so that

$$\lim_{\rho_{\text{video}} \to \infty} \mathbf{R}^t_{\text{video}} = \lim_{\rho_{\text{video}} \to \infty} \mathbf{R}^c_{\text{video}}.$$

Now assume that $0 \leq \rho_{\text{video}} < \infty$. Then, from (10) and (11), we have

$$\begin{aligned}
\mathbf{R}^c_{\text{video}} \leq \quad & \mathbf{R}^t_{\text{video}} \iff r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s, v-1)}{\sum_{(s',v')\in\mathbb{S}^+_{\text{video}}} \pi(s', v'-1)} \right) \beta_{\text{video}}(s, v) \\
\leq \quad & r_{\text{video}} \sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \left(\frac{\pi(s, v)}{\sum_{(s',v')\in\mathbb{S}^+_{\text{video}}} \pi(s', v')} \right) \beta_{\text{video}}(s, v) \\
\iff & \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^{v-1}_{\text{video}}}{s!\,(v-1)!} \beta_{\text{video}}(s, v) \right) \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}}\, \rho^v_{\text{video}}}{s!v!} \right) \\
& + - \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^v_{\text{video}}}{s!v!} \beta_{\text{video}}(s, v) \right) \left(\sum_{(s,v)\in\mathbb{S}^+_{\text{video}}} \frac{\rho^s_{\text{speech}} \rho^{v-1}_{\text{video}}}{s!\,(v-1)!} \right) \leq 0 \\
\iff & \left(\sum_{v=0}^{v_{\max}-1} \rho^v_{\text{video}} \sum_{s=0}^{C-\beta^{\min}(v+1)} \frac{\rho^s_{\text{speech}}}{s!v!} \beta_{\text{video}}(s, v+1) \right) \left(\sum_{w=0}^{v_{\max}-1} \rho^w_{\text{video}} \sum_{s=0}^{C-\beta^{\min}(w+1)} \frac{\rho^s_{\text{speech}}}{s!\,(w+1)!} \right) \\
& - \left(\sum_{v=0}^{v_{\max}-1} \rho^v_{\text{video}} \sum_{s=0}^{C-\beta^{\min}(v+1)} \frac{\rho^s_{\text{speech}}}{s!\,(v+1)!} \beta_{\text{video}}(s, v+1) \right) \\
& \times \left(\sum_{w=0}^{v_{\max}-1} \rho^w_{\text{video}} \sum_{s=0}^{C-\beta^{\min}(w+1)} \frac{\rho^s_{\text{speech}}}{s!w!} \right) \leq 0.
\end{aligned}$$

Recognising that the LHS is a polynomial in ρ of degree $2\,(v_{\max}-1) = 2\left(\lfloor C/\beta_{\text{video}}^{\min}\rfloor - 1\right)$, the above condition can be written in the following form:

$$\sum_{k=0}^{2(v_{\max}-1)} \rho_{\text{video}}^{k} \sum_{v+w=k} \zeta_{v,w} \leq 0, \tag{21}$$

where the coefficients $\zeta_{v,w}$, $v, w, = 0, \ldots, v_{\max}-1$, are given by

$$\begin{aligned}
\zeta_{v,w} &= \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!v!}\beta_{\text{video}}\,(s, v+1)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!\,(w+1)!}\right) \\
&\quad - \left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!\,(v+1)!}\beta_{\text{video}}\,(s, v+1)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!w!}\right) \\
&= \frac{1}{v!w!}\left(\frac{1}{w+1} - \frac{1}{v+1}\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\beta_{\text{video}}\,(s, v+1)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\right).
\end{aligned}$$

Note that $\zeta_{v,v} = 0$, $v = 0, \ldots, v_{\max}-1$, so that the coefficients for ρ^{0} and $\rho^{2(v_{\max}-1)}$ vanish.

Observe that since $\rho_{\text{video}} \geq 0$, a sufficient condition for (21) is that all coefficients $\sum_{v+w=k} \zeta_{v,w} \leq 0$, $k = 1, \ldots, 2v_{\max}-1$. To this end, we will show that $\zeta_{v,w} + \zeta_{w,v} \leq 0$, where we take $v < w$ without loss of generality, i.e.,

$$\begin{aligned}
\zeta_{v,w} + \zeta_{w,v} \leq\;& 0 \\
\iff & \left(\frac{1}{w+1} - \frac{1}{v+1}\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\beta_{\text{video}}\,(s, v+1)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\right) \\
& + \left(\frac{1}{v+1} - \frac{1}{w+1}\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\beta_{\text{video}}\,(s, w+1)\right)\left(\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\right) \leq 0
\end{aligned}$$

or, equivalently,

$$\frac{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\beta_{\text{video}}\,(s, v+1)}{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(v+1)} \frac{\rho_{\text{speech}}^{s}}{s!}} \geq \frac{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}\beta_{\text{video}}\,(s, w+1)}{\sum_{s=0}^{C-\beta_{\text{video}}^{\min}(w+1)} \frac{\rho_{\text{speech}}^{s}}{s!}},$$

i.e.,

$$\mathbf{E}\left\{\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(v+1)}, v+1\right)\right\} \geq \mathbf{E}\left\{\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(w+1)}, w+1\right)\right\},$$

where S_x is a random variable distributed as the queue length in a standard Erlang loss model with capacity x and traffic load ρ_{speech}. Observe that we have effectively reduced the inequality $\mathbf{R}_{\text{video}}^{c} \leq \mathbf{R}_{\text{video}}^{t}$ for the SV model to a set of inequalities for a speech-only model, i.e. for the standard Erlang loss model.

To complete the proof, we will show that $\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(v+1)}, v+1\right)$ is almost surely non-increasing in v, for $v = 0, \ldots, v_{\max}-1$. Substituting $y = C - \beta_{\text{video}}^{\min}\,(v+1)$, we have that

$$\beta_{\text{video}}\left(S_{C-\beta_{\text{video}}^{\min}(v+1)}, v+1\right) = \beta_{\text{video}}\left(S_y, \frac{C-y}{\beta_{\text{video}}^{\min}}\right),$$

which we will demonstrate to be almost surely non-decreasing in y, by comparing the above expression for $y, y+\beta_{\text{video}}^{\min} \in \left[0, C-\beta_{\text{video}}^{\min}\right]$, where the lower (upper) bound corresponds with $v = v_{\max}-1$ ($v = 0$). First observe that the sample paths of the Erlang loss model with capacity y and $y + \beta_{\text{video}}^{\min}$ can readily be compared. Clearly, for an identical input of interarrival times and call lengths, it must be that the sample path of the system with capacity $y + \beta_{\text{video}}^{\min}$ is never below that of the system with capacity y. In fact, starting with an empty system, the sample paths coincide until a call is blocked in the system with capacity y. Then, during the period that the system with capacity y is full, it may be that one or more additional calls are admitted to the system with capacity $y + \beta_{\text{video}}^{\min}$. Note that at most $\beta_{\text{video}}^{\min}$ additional calls can be accepted. The sojourn times of the additional calls are independent of the sojourn times of the other calls in the system with capacity $y + \beta_{\text{video}}^{\min}$, which are also present in the system with capacity y. Hence, with probability 1,

$$S_y \leq S_{y+\beta_{\text{video}}^{\min}} \leq S_y + \beta_{\text{video}}^{\min} \text{ and } S_y \leq y.$$

Combining these results with the fact that $y + \beta_{\text{video}}^{\min} \leq C$ and, in general, for $a, b \in \mathbb{R}$, it holds that if $a \geq b > \epsilon$, then $\left(\frac{a-\epsilon}{b-\epsilon}\right) \geq \frac{a}{b}$, implies that

$$\frac{C - S_{y+\beta_{\text{video}}^{\min}}}{C - \left(y + \beta_{\text{video}}^{\min}\right)} \geq \frac{C - \left(S_y + \beta_{\text{video}}^{\min}\right)}{C - \left(y + \beta_{\text{video}}^{\min}\right)} \geq \frac{C - S_y}{C - y},$$

with probability 1. Recall that

$$\beta_{\text{video}}\left(S_y, \frac{C - y}{\beta_{\text{video}}^{\min}}\right) = \min\left\{\beta^{\max}, \beta_{\text{video}}^{\min} \frac{C - S_y}{C - y}\right\},$$

so that

$$\beta_{\text{video}}\left(S_{y+\beta_{\text{video}}^{\min}}, \frac{C - \left(y + \beta_{\text{video}}^{\min}\right)}{\beta_{\text{video}}^{\min}}\right) \geq \beta_{\text{video}}\left(S_y, \frac{C - y}{\beta_{\text{video}}^{\min}}\right),$$

with probability 1, which completes the proof. ■

Appendix D. Proof of Corollary 1

Proof. The proof follows from manipulating the inequality proven in Theorem 3, using expressions (10) and (11), and relating it to the derivative of the time-average video throughput expression (11) with respect to ρ_{video} :

$$\begin{aligned}
\mathbf{R}_{\text{video}}^{c} \leq \mathbf{R}_{\text{video}}^{t} \iff & \left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s, v-1)\,\beta_{\text{video}}(s,v)\right)\left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v)\right) \\
& - \left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v)\,\beta_{\text{video}}(s,v)\right)\left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v-1)\right) \leq 0 \\
\iff & \frac{\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v-1)\,\beta_{\text{video}}(s,v)}{\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v)} \\
& - \frac{\left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v)\,\beta_{\text{video}}(s,v)\right)\left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v-1)\right)}{\left(\sum_{(s,v)\in\mathbb{S}_{\text{video}}^{+}} \pi(s,v)\right)^{2}} \leq 0 \\
\iff & \frac{\partial \mathbf{R}_{\text{video}}^{t}}{\partial \rho_{\text{video}}} \leq 0. \quad ■
\end{aligned}$$

References

[1] S. Ben Fredj, T. Bonald, A. Proutiere, G. Régnié, J.W. Roberts, Statistical bandwidth sharing: A study of congestion at flow level, in: Proceedings of SIGCOMM '01, San Diego, USA, 2001.
[2] N. Benameur, S. Ben Fredj, F. Delcoigne, S. Oueslati-Boulahia, J.W. Roberts, Integrated admission control for streaming and elastic traffic, in: Proceedings of the 2nd International Workshop on Quality of Future Internet Services, Coimbra, Portugal, 2001.
[3] J.L. van den Berg, R. Litjens, J.F. Laverman, HSDPA flow level performance: The impact of key system and traffic aspects, in: Proc. of MSWIM '04, Venice, Italy, 2004.
[4] T. Bonald, L. Massoulié, Impact of fairness on internet performance, in: Proceedings of SIGMETRICS '01, Cambridge, USA, 2001.
[5] T. Bonald, J.W. Roberts, Performance of bandwidth sharing mechanisms for service differentiation in the internet, in: Proceedings of the ITC Specialist Seminar on IP Traffic Measurement, Modelling and Management, Monterey, USA, 2000, pp. 22.1–22.10.
[6] S.C. Borst, User-level performance of channel-aware scheduling algorithms in wireless data networks, in: Proceedings of INFOCOM '03, San Francisco, USA, 2003.
[7] S.K. Cheung, J.L. van den Berg, R.J. Boucherie, Insensitive bounds for the moments of the sojourn time distribution in the $M/G/1$ processor sharing queue, Queuing Systems: Theory and Applications 53 (1–2) (2006) 7–18.
[8] E.A. Coddington, N. Levinson, Theory of Ordinary Differential Equations, McGraw-Hill, New York, USA, 1955.
[9] E.G. Coffman, R.R. Muntz, H. Trotter, Waiting time distributions for processor-sharing systems, Journal of the Association for Computing Machinery 17 (1970) 123–130.
[10] J.W. Cohen, The multiple phase service network with generalized processor sharing, Acta Informatica 12 (1979) 245–284.
[11] M. Crovella, A. Bestavros, Self-similarity in World Wide Web traffic: Evidence and possible causes, Transactions on Networking 5 (6) (1997) 835–846.
[12] F. Delcoigne, A. Proutière, G. Régnié, Modelling integration of streaming and data traffic, in: Proceedings of the ITC Specialist Seminar on Internet Traffic Engineering and Traffic Management, Würzburg, Germany, 2002.
[13] N.M. van Dijk, Queuing Networks and Product Forms: A System's Approach, John Wiley & Sons, New York, USA, 1993.
[14] C. Douligeris, Multiobjective flow control in telecommunication networks, in: Proceedings of INFOCOM '92, Florence, Italy, 1992.
[15] N. Joshi, S.R. Kadaba, S. Patel, G.S. Sundaram, Downlink scheduling in CDMA data networks, in: Proceedings of MOBICOM '00, Boston, USA, 2000.
[16] J.S. Kaufman, Blocking in a shared resource environment, IEEE Transactions on Communications 29 (10) (1981) 1474–1481.
[17] F.P. Kelly, Reversibility and Stochastic Networks, John Wiley & Sons, New York, USA, 1979.
[18] A.A. Kherani, A. Kumar, Performance analysis of TCP with nonpersistent sessions, in: Proceedings of the Workshop on Modelling of Flow and Congestion Control, Paris, France, 2000.
[19] A.A. Kherani, A. Kumar, Stochastic models for throughput analysis of randomly arriving elastic flows in the internet, in: Proceedings of INFOCOM '02, New York, USA, 2002.
[20] L. Kleinrock, Queuing Systems, vol. II, John Wiley & Sons, New York, USA, 1976.
[21] W.E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, On the self-similar nature of ethernet traffic, in: Proceedings of SIGCOMM '93, San Francisco, USA, 1993, pp. 183–193.
[22] R. Litjens, R.J. Boucherie, Radio resource sharing in a GSM/GPRS network, in: Proceedings of the ITC Specialist Seminar on Mobile Systems and Mobility, Lillehammer, Norway, 2000, pp. 261–274.
[23] R. Litjens, R.J. Boucherie, Elastic calls in an integrated services network: The greater the call size variability the better the Quality-Of-Service, Performance Evaluation 52 (4) (2003) 193–220.
[24] R. Litjens, J.L. van den Berg, M.J. Fleuren, Spatial traffic heterogeneity in HSDPA networks and its impact on network planning, in: Proceedings of ITC 19, Beijing, China, 2005.
[25] R. Litjens, F. Roijers, J.L. van den Berg, R.J. Boucherie, M. Fleuren, Performance analysis of WLANs: An integrated packet/flow level approach, in: Proceedings of ITC18, Berlin, Germany, September 2003.
[26] L. Massoulié, J.W. Roberts, Arguments in favour of admission control for TCP flows, in: Proceedings of ITC 16, Edinburgh, Scotland, 1999.
[27] R. Núñez Queija, Processor-sharing models for integrated-services networks, Ph.D. Thesis, Technische Universiteit Eindhoven, 1999.
[28] R. Núñez Queija, Sojourn times in a processor sharing queue with service interruptions, Queuing Systems 34 (2000) 351–386.
[29] R. Núñez Queija, Sojourn times in non-homogeneous QBD processes with processor sharing, Stochastic Models 17 (2001) 61–92.
[30] R. Núñez Queija, J.L. van den Berg, M.R.H. Mandjes, Performance evaluation of strategies for integration of elastic and stream traffic, in: Proceedings of ITC 16, Edinburgh, Scotland, 1999, pp. 1039–1050.
[31] T. Ott, The sojourn time distribution in the $M/G/1$ queue with processor sharing, Journal of Applied Probability 21 (1984) 360–378.
[32] S. Rácz, B.P. Gerö, G. Fodor, Flow level performance analysis of a multi-service system supporting elastic and adaptive services, Performance Ealuation 49 (1–4) (2002) 451–469.
[33] J.W. Roberts, A service system with heterogeneous user requirements—Application to multiservice telecommunications systems, in: G. Pujolle (Ed.), Performance of Data Communications Systems and their Applications, North Holland, New York, USA, 1981.
[34] J.W. Roberts, L. Massoulié, Bandwidth sharing and admission control for elastic traffic, in: Proceedings of the ITC Specialist Seminar on Teletraffic Issues Related to Multimedia and Nomadic Communications, Yokohama, Japan, 1998.
[35] S.M. Ross, A First Course in Probability, Macmillan Publishing Company, New York, USA, 1984.
[36] H.C. Tijms, Stochastic Modelling and Analysis: A Computational Approach, John Wiley & Sons, Chichester, England, 1986.
[37] R.W. Wolff, Stochastic Modeling and the Theory of Queues, Prentice-Hall, Englewood Cliffs, USA, 1989.
[38] S.F. Yashkov, A derivation of response time distribution for a $M/G/1$ processor sharing queue, Problems in Control and Information Theory 12 (1983) 133–148.
[39] S.F. Yashkov, Processor sharing queues: some progress in analysis, Queuing Systems 2 (1987) 1–17.
[40] S.F. Yashkov, Mathematical problems in the theory of processor sharing queuing systems, Journal of Soviet Mathematics 58 (1992) 101–147.
[41] A.P. Zwart, O.J. Boxma, Sojourn time asymptotics in the $M/G/1$ processor sharing queue, Queuing Systems 35 (2000) 141–166.