Effective Bandwidths for a Class of Non Markovian Fluid Sources*

Kimon Kontovasilis
National Center for Scientific Research "DEMOKRITOS",
Institute for Informatics & Telecommunications,
GR-15310 AG. PARASKEVI ATTIKHS, PoB 60228, GREECE
kimon@cyclades.nrcps.ariadne-t.gr

Nikolas Mitrou
National Technical University of Athens,
Electrical & Computer Eng. Dept.,
Computer Science Division,
9 IROON POLYTECHNEIOU STR., GR-15773 ZOGRAFOU, GREECE
mitrou@cs.ece.ntua.gr

Abstract

This paper proves the existence of and explicitly determines effective bandwidths for a class of non Markovian fluid source models, featuring multiple data-transmission rates and arbitrary distributions for the times these rates are sustained. The investigated models cover considerably more traffic profiles than the usual Markovian counterparts and have reduced state-space requirements. The effective bandwidth, as a function of the asymptotic loss probability decay rate, is implicitly derivable by the requirement that the spectral radius of an appropriate nonnegative matrix be equal to unity. The effective bandwidth function is shown to be, either strictly increasing, or constant and equal to the mean rate. Sources of the second kind, which are characterized, generalize the notion of 'CBR' traffic. Furthermore, a study for the limiting effective bandwidth, towards a loss-less environment, is undertaken; it is shown that the limiting value may, under some fully identified restrictions on the source behavior, be less than the source's peak rate. Under those restrictions, a source may have reduced bandwidth requirements, even if it features a large peak rate.

1 Introduction

Resource management and the associated control functions to support it present some of the most important problems in the way towards the future broadband communication networks. A large number of the proposed approaches to these problems promotes the effective bandwidth theory as the underlying foundation for constructing the appropriate control mechanisms. This is no accident, since the effective bandwidth theory boils all resource requirements for a given connection down to a conceptually simple scalar descriptor, derivable only from information about the connection's characteristics, the buffering capabilities at the congestion points and the desired quality of service (expressed as a loss probability percentile). No information about the other connections in the system is necessary, a property of great importance which dramatically reduces the dimension of the computational problems. Furthermore, the effective bandwidth is additive: the required bandwidth for the superposition of a number of given independent connections is simply the sum of the effective bandwidths of the individual connections. This property not only simplifies admission control schemes, but also provides a direct link to well known techniques developed for classical circuit switching systems.

Of course the effective bandwidth theory has its shortcomings. Firstly, it's an asymptotic theory, valid to the limit as the buffering capabilities tend to infinity (maintaining a finite constant of proportionality to the logarithm of the QoS value). Secondly, it conservatively ignores the part of the multiplexing gain due to rate cancellations over many connections (if it didn't, the nice additive property wouldn't hold any more). Despite these blemishes, the attractive properties of the theory and the present domain of application, viz. ATM (featuring multiplexers with large buffers and very stringent QoS demands) combine to justify the popularity of the effective bandwidth approach.

The concept of effective bandwidth for high-speed networks was originally proposed by [9, 10, 8] and was applied to iid, Markovian On/Off, and other simple source models. A more general development of the theory, for the Markovian framework [6], and for general stationary sources [2, 11], followed. See also [3] for further references therein and a review of the effective bandwidth theory along the lines of the statistical mechanics viewpoint. Reference [4] provides a recent review of resource management techniques over the effective bandwidth idea.

The present general theory declares that for any source model (either in the fluid, or the discrete cell domain), under the assumption of stationarity and some other mild conditions (to be reviewed later), there exists an effective bandwidth function, derivable on the basis of the moment generating function of the total amount of data generated by the source within time t, asymptotically as $t \to \infty$. This not only provides a 'recipe' for determining the effective bandwidth for arbitrary sources, but also provides, at least in principle, a way to by-pass modeling; instead of being calculated, the asymptotic moment generating function can be

*Work partially funded by the European Union, through the research project AC-235 WATT, in the ACTS programme.

To appear in SIGCOMM'97, September 1997, Cannes, France

measured directly (for work along this line, see [5]). Direct measurement of asymptotic quantities within a reasonable amount of time, however, is difficult and error-prone, at least for sources of non trivial complexity. Consequently, modeling still remains a valuable tool in the arsenal of analysts and designers.

The most general class of models for which effective bandwidth functions have been *explicitly* calculated is the class of multi-rate Markovian models. The purpose of this paper is to provide an explicit calculation of the effective bandwidth function for a more general class of source models. Two main reasons provided the motivation: (a) the process of approximating a given source's behavior by a Markovian model often results in a Markov chain of very large state-space (and associated computational complexity); and (b) many sources exhibit a definitely non-Markovian nature, especially in the presence of traffic shaping or source-based throttle control. The models in the extended class covered in this paper allow an arbitrary (finite) number of data generating rates, mapped to corresponding states. Transitions from state to state still occur according to a (discrete) Markov Chain. However, the sojourn times spent in the various states are allowed *general* distributions. With this arrangement, the state-space is dramatically reduced; as an example, completely general on/off sources are always described with a state-space containing exactly two states. The effective bandwidth, as a function of the asymptotic loss probability decay rate is determined implicitly, by the requirement that the spectral radius of an appropriate nonnegative matrix be equal to unity.

The study of the generalized class of models in this paper reveals new properties, not shared by the usual Markovian fluids. To start with, the effective bandwidth is shown to be, either a strictly increasing function of the asymptotic decay rate (as in the Markovian framework), or constant and equal to the source's mean rate. Sources of the second kind, which may well feature multiple data-generation rates, are completely characterized and generalize the notion of 'CBR' traffic. Furthermore, a study for the limiting effective bandwidth, as the quality requirement becomes unboundedly stringent, is undertaken. It is shown that the limiting effective bandwidth may, under some fully identified restrictions on the source behavior, be less than the source's peak rate (unlike usual Markovian models). In other words, it is shown that a source exhibiting a large peak rate may still have reduced bandwidth requirements, provided that the large peak rate is counter-balanced by other features in the source's behavior. The conditions for a reduced limiting effective bandwidth may be of value either to the designer of a terminal, or to settings were the source's activity is controllable 'on-line'.

The organization of the rest of the paper is as follows. Section 2 establishes some preliminary results, reviews some background material and introduces the main bulk of the notation. Section 3 presents the generalized class of models, proves the existence of, and calculates the effective bandwidth function and investigates properties on monotonicity. The subsequent Section 4 determines the limiting effective rate and comments on the 'large buffer' asymptotic regime under which the effective bandwidth theory is valid. Some numerical results are presented in Section 5 to illustrate key concepts of our work and validate the results. Finally, Section 6 recapitulates the findings of the paper.

2 Preliminaries and background material

2.1 Convex functions

We state the following two lemmas on convex functions, for later usage. The first lemma characterizes the monotonicity of certain functions, through the convexity of other associated functions.

Lemma 2.1 *Let $f(x)$ be a continuous, convex (alt.: strictly convex) function defined on an interval $\mathcal{I} \subseteq \mathbb{R}$ containing 0 and assume that there exists the $\lim_{\substack{x\to 0\\ x\in\mathcal{I}}} (f(x)/x) = l \neq \pm\infty$. Define $g(x) := f(x)/x$, for $x \neq 0$ and set $g(0) = l$. Then, $g(x)$ is a continuous, increasing (alt.: strictly increasing) function on $\mathcal{I}$.*

The second lemma characterizes convex analytic functions.

Lemma 2.2 *A function convex and analytic on an interval $\mathcal{I} \subseteq \mathbb{R}$ is either strictly convex or linear on the whole of $\mathcal{I}$.*

The proofs may be found in the appendix.

2.2 Moment generating functions

Consider a nonnegative random variable (r.v.) X, its PDF denoted as $F(\cdot)$ and let $\gamma(\cdot)$ stand for its moment generating function, viz., $\gamma(\omega) = \mathbb{E}\, e^{\omega X}$. Define its effective domain as $\Omega = \{\, \omega \in \mathbb{R} \mid \gamma(\omega) < \infty \,\}$ and let $\omega^* = \sup \Omega$. Since X is nonnegative, $\omega^* \geq 0$. We will assume throughout that $\omega^* > 0$. This is equivalent to assuming that the tail of $1-F(x)$ has an exponential upper bound; as a consequence, all moments of X are finite. Finally, we make the assumption that $\omega^* \notin \Omega$ (i.e., Ω is open). Then, by a direct application of Fatou's Lemma, $\lim_{\omega\to\omega^*-0} \gamma(\omega) = \infty$.

For all $\omega \in \Omega$, the generator function $\gamma(\omega)$ is a strictly[1] increasing analytic function. It is also a log-convex, hence convex function. The following lemma, to be used in several future occasions, characterizes $\gamma(\omega)$ more completely:

Lemma 2.3 $\log \gamma(\omega)$ *is either strictly convex or linear on the whole of Ω. Linearity applies iff X is a.s. constant.*

Proof Pick two different $\omega_1, \omega_2 \in \Omega$ and choose $0 < h_1, h_2$, such that $h_1 + h_2 = 1$. Define $a_i(x) := e^{h_i \omega_i x}$, and let $p_i = h_i^{-1}$, for $i = 1, 2$. By Hölder's inequality,

$$\begin{aligned} \gamma(h_1\omega_1 + h_2\omega_2) &= \mathbb{E}[a_1(X)a_2(X)] \\ &\leq \left(\mathbb{E}[a_1(X)^{p_1}]\right)^{1/p_1} \left(\mathbb{E}[a_2(X)^{p_2}]\right)^{1/p_2} \\ &= \gamma(\omega_1)^{h_1}\gamma(\omega_2)^{h_2}, \end{aligned}$$

proving log-convexity. Further, by the condition for equality in Hölder's inequality, $\gamma(\omega)$ is strictly log-convex, unless $a_1(X)^{p_1} = C a_2(X)^{p_2}$, a.s., for some constant C. This immediately translates to $X = \log C/(\omega_1 - \omega_2)$ a.s. For the 'if part', when X is a.s. constant direct calculation yields linearity. ∎

We will also need the following fact, immediately arising from the strict concavity of the logarithm:

Lemma 2.4 *For every ω, it holds $\log \gamma(\omega) \geq \omega\, \mathbb{E} X$. Furthermore, for $\omega \neq 0$, equality holds iff X is a.s. constant.*

[1] Unless X is a.s. zero, in which case $\gamma(\omega)$ is constant. Almost surely zero r.v.s do not arise in this paper.

We finally connect the asymptotic behavior of $\gamma(\omega)$ to the extremal values of the r.v. X, that is its essential supremum, $\operatorname{ess\,sup} X \stackrel{\text{def}}{=} \sup\{\, x \mid F(x) < 1 \,\}$ and its essential infimum, $\operatorname{ess\,inf} X \stackrel{\text{def}}{=} \inf\{\, x \mid F(x) > 0 \,\}$. Note that for a nonnegative X, $\operatorname{ess\,inf} X \geq 0$. Define $\omega^{-1} \log \gamma(\omega)$ as an extended real valued function (equal to $+\infty$ for $\omega \notin \Omega$). Because of Lemma 2.3, all assumptions of Lemma 2.1 are fulfilled (with $\lim_{\omega \to 0} \omega^{-1} \log \gamma(\omega) = \gamma'(0)/\gamma(0) = \mathbb{E} X$). Thus, $\omega^{-1} \log \gamma(\omega)$ is an increasing function, hence its limits, as $\omega \to \pm\infty$ exist, in the extended sense. The following lemma states that these limits are equal to the extremal values of X.

Lemma 2.5 $\lim_{\omega \to +\infty} \omega^{-1} \log \gamma(\omega) = \operatorname{ess\,sup} X$. *Similarly,* $\lim_{\omega \to -\infty} \omega^{-1} \log \gamma(\omega) = \operatorname{ess\,inf} X$.

PROOF We prove only the first limit, the arguments for the other being similar. Assume first that $\operatorname{ess\,sup} X = a < +\infty$. Due to the definition of $\operatorname{ess\,sup} X$, for every $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$, such that $F(a - \epsilon) = 1 - \delta(\epsilon)$. Thus,

$$e^{\omega a} \geq \gamma(\omega) \geq \int_{a-\epsilon}^{+\infty} e^{\omega x}\, dF(x) \geq e^{(a-\epsilon)\omega} \delta(\epsilon), \ \forall \omega, \epsilon > 0,$$

the first inequality being due to the fact that $\operatorname{ess\,sup} X = a$. By taking logarithms, dividing by ω and letting first $\omega \to +\infty$ and then $\epsilon \to 0$, we obtain $\lim_{\omega \to +\infty} \omega^{-1} \log \gamma(\omega) = a = \operatorname{ess\,sup} X$.

Now assume that $\operatorname{ess\,sup} X = +\infty$, but suppose that $\lim_{\omega \to +\infty} \omega^{-1} \log \gamma(\omega) = a < +\infty$. We obtain,

$$e^{\omega a} \geq \gamma(\omega) \geq \int_{\xi+}^{+\infty} e^{\omega x}\, dF(x) \geq e^{\xi \omega} \big(1 - F(\xi)\big), \ \forall \omega \geq 0, \forall \xi,$$

the first inequality being due to the fact that $\omega^{-1} \log \gamma(\omega)$ is an increasing function and its limiting value as $\omega \to +\infty$ is a. By rearranging, we obtain $1 - F(\xi) \leq e^{(a-\xi)\omega}$, for all $\omega \geq 0$, which implies that $F(\xi) = 1, \forall \xi > a$, i.e., $\operatorname{ess\,sup} X \leq a < +\infty$, a contradiction. Thus, we have shown the intended equality for infinite values too. ■

2.3 Nonnegative matrices

Throughout the paper, we use the following notation on matrix (or vector) relations: given two matrices A and B, we write $A \geq B$, when the inequality holds element-wise, for all elements. We say that $A > B$ when $A \geq B$ and $A \neq B$, i.e., the inequality is strict for at least a pair of elements. Finally, we write $A \stackrel{+}{>} B$ when the inequality is strict for all elements of A and B.

Consider a nonnegative matrix $A \geq 0$. The standard Perron-Frobenius theory (see e.g., [1]) states that the spectral radius $\rho(A)$ is an eigenvalue, corresponding to a nonnegative eigenvector. If, furthermore, A is irreducible, then $\rho(A)$ is a simple eigenvalue and the corresponding eigenvector $\boldsymbol{x}$ is strictly positive ($\boldsymbol{x} \stackrel{+}{>} \mathbf{0}$). Additionally, every other nonnegative eigenvector $\boldsymbol{y}$ is a multiple of $\boldsymbol{x}$ [1, p. 27], that is, it necessarily corresponds to $\rho(A)$.

The spectral radius of nonnegative matrices is an increasing function of any of their elements, i.e., if $0 \leq A \leq B$, then $\rho(A) \leq \rho(B)$. Irreducibility strengthens this result as follows [1, p. 27]: if $0 \leq A < B$ and $A+B$ is irreducible, then $\rho(A) < \rho(B)$. Consequently, the spectral radius of an irreducible nonnegative matrix is a strictly increasing function of any of its elements.

The following result provides a lower bound for the spectral radius of nonnegative matrices:

Lemma 2.6 (see [7], Theorem 3.1) *Let A be a nonnegative matrix with $\rho(A) > 0$, such that it admits strictly positive left and right Perron eigenvectors, $\boldsymbol{u}$ and $\boldsymbol{v}$, normalized so that $\sum_j u_j v_j = 1$. Then, for every positive definite diagonal matrix D it holds*

$$\log \rho(DA) \geq \sum_j u_j v_j \log d_j + \log \rho(A).$$

If $A^T A$ is irreducible, equality holds iff $D = aI$.

Note that ref. [7] states the theorem for irreducible A and the condition for strict inequality to require $A \stackrel{+}{>} 0$ (relaxed in comments to the condition that A is completely indecomposable). The proof however immediately reveals that the result holds as stated in Lemma 2.6. The upper bound is 'best-possible' in the sense that there exist $A \geq 0$ for which the bound is attained for every positive definite D. Indeed, this hold trues if A is a $k \times k$ cyclic matrix with period k, viz.,

$$A = \begin{pmatrix} 0 & \times & 0 & \cdots & 0 \\ 0 & 0 & \times & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \times & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{1}$$

where the nonzero elements are marked as $\times$. Note that although the condition for strict inequality in the lemma is satisfied if A is completely indecomposable, this is not so if A is merely irreducible and aperiodic. For example, if the element in the last row and second column within (1) becomes positive, then A becomes primitive, but $A^T A$ is completely decomposable and there exists D, not of all elements equal, that achieves equality in the bound. (Thus the Remark 3.2, in page 471 of [7] is wrong.) Lastly, we mention that the condition for strict inequality is sufficient but not necessary. In general, even if $A^T A$ is reducible, the elements of D must satisfy a certain relation (depending on the structure of A) to obtain equality.

2.4 Effective bandwidths of general stationary sources

We now summarize some elements of the effective bandwidth theory necessary in our later developments, drawing from reference [2]. Consider a fluid[2] information source, generating data at rate $r(t)$, $t \geq 0$. The amount of information generated within the interval $[t_1, t_2]$ is $V(t_1, t_2) = \int_{t_1}^{t_2} r(t)\, dt$. Define

$$V^*(\theta, t) \stackrel{\text{def}}{=} \sup_{s \geq 0} \log \mathbb{E}\, e^{\theta V(s, t+s)} \tag{2}$$

and consider the following 'regularity' conditions, collectively referred to as the '*conditions R*':

R1 *The process $\{\, r(t),\ t \geq 0 \,\}$ is stationary and ergodic;*

R2 $a(\theta) \stackrel{\text{def}}{=} \theta^{-1} \lim_{t \to \infty} \big(V^*(\theta, t)/t\big)$ *exists for all $\theta > 0$;*

R3 $\theta a(\theta)$ *is strictly convex and differentiable for all $\theta > 0$;*

[2]Entirely analogous results hold for sources generating information in discrete quanta, and in fact [2] states its results within the discrete setting.

Subject to conditions R1–2, we call $a(\theta)$ the effective bandwidth function of the process. The superposition of a number of independent processes (each one of them being subject to conditions R1–2) has effective bandwidth equal to the sum of the effective bandwidths of the constituent processes. Note that, by (2) and Lemma 2.3, $\theta a(\theta)$, if it exists, is convex, hence $a(\theta)$ is, by Lemma 2.1, increasing. Furthermore, if R3 also holds, by the same reasoning we see that $a(\theta)$ is strictly increasing.

A traffic source with eff. bandwidth function $a(\theta)$ enjoys the following property: if the source loads an infinite queue with output capacity equal to c and if, for some $\theta > 0$, $a(\theta) < c$, then the queue length $q(t)$, at all times, has an exponential upper bound with parameter θ, i.e., there exists a constant $d(\theta)$, such that, for all $t \geq 0$ and all $x \geq 0$, it holds $\Pr\{\, q(t) \geq x \,\} \leq d(\theta)e^{-\theta x}$. In other words, the condition $a(\theta) < c$ guarantees that the queue length decays exponentially at all times, with a rate at least as good as θ. Conversely, if $c < a(\theta)$ then the queue content cannot be bounded exponentially with respect to θ.

The two statements taken together imply that, to achieve an exponential decay with rate at least as θ, the output capacity must be chosen at least equal to $a(\theta)$. In the general case, if this is done the system achieves an exponential decay with rate equal to θ^*, where $\theta^* = \sup\{\, \theta \mid a(\theta) < c \,\}$. If the additional condition R3 holds, then $a(\theta)$ is strictly increasing and, instead of an upper bound, we have an asymptotic 'match', i.e., there exists a stationary distribution of the queue length $q(\infty)$ and $\lim_{x\to\infty}(-\log \Pr\{\, q(\infty) \geq x \,\}/x) = \theta$, with $\theta = a^{-1}(c)$.

3 The model and its effective bandwidth

The class of models we consider in this paper comprises a number K of states, each corresponding to a data-generation rate r_j, $j = 1, \ldots, K$. Not all rates are equal (otherwise, the model describes a plain CBR source). We write $\hat{r}$ for the peak rate $\max_j r_j$ and denote $R = \mathrm{diag}\{r_j\}$. Each time the model enters a state, it spends there a random amount of time, then moves to another state. The movement from state to state occurs according to a (discrete) irreducible Markov Chain, described by the transition probability matrix P. Let $\boldsymbol{\pi}$ denote the invariant probability vector of this Markov chain, viz. $\boldsymbol{\pi} P = \boldsymbol{\pi}$ and $\sum_j \pi_j = 1$. The sojourn time for a particular visit at some state is independent of all the states visited in the past or future and the corresponding sojourn times. Finally, different visits to a particular state correspond to iid random variables.

Let T_j denote a 'typical' sojourn time at state j. In the notation of subsection 2.2, let $\gamma_j(\omega)$, Ω_j and ω_j^* denote respectively the moment generating function for T_j, its effective domain and the supremum of the latter. Arbitrary distributions are allowed for the variables T_j, the only restrictions being those mentioned in subsection 2.2, viz. for all j: $\omega_j^* > 0$ (i.e., no 'long range dependence'), and $\omega_j^* \notin \Omega_j$ (open effective domains). We use τ_j, τ_j^+ and τ_j^- to denote respectively the expected value, essential supremum and essential infimum of T_j.

Without loss of generality, we assume that there is no state k with T_k being a.s. zero, otherwise this state is trivial and can be removed from the system (the adjusted Markov chain being the stochastic complement of P, with respect to the non-trivial states). Thus, for all j:, $\tau_j, \tau_j^+ > 0$. Furthermore, note that if T_j is to reflect the *whole* sojourn in state j before moving to another state $k \neq j$, then P must feature $p_{jj} = 0$, $\forall j$. The system may always easily be transformed to this 'normal' form, by appropriate adaptation of P and the moment generators. The results in the paper do *not* require the model to be in normal form.

Standard Markov Renewal theory predicts that the data generation rate for this class of models is a stationary and ergodic process (thus condition R1 of subsection 2.4 holds). The corresponding mean rate is

$$\bar{r} = \frac{\sum_{j=1}^{K} \pi_j \tau_j r_j}{\sum_{j=1}^{K} \pi_j \tau_j}. \tag{3}$$

We introduce two variables, θ and u, and consider the values $\gamma_j(r_j\theta - u)$, for all the combinations that satisfy $r_j\theta - u \in \Omega_j$. We write $\Gamma(\theta, u)$ for $\mathrm{diag}\{\gamma_j(r_j\theta - u)\}$ and $\Gamma'(\theta, u)$ for $\mathrm{diag}\{\gamma_j'(r_j\theta - u)\}$. For each permissible pair of values the matrix $A(\theta, u) \stackrel{\text{def}}{=} \Gamma(\theta, u)P$ is a nonnegative irreducible matrix. Let

$$\phi(\theta, u) \stackrel{\text{def}}{=} \rho\big(A(\theta, u)\big). \tag{4}$$

By reference to Lemma 2.6, we immediately obtain the lower bound

$$\log \phi(\theta, u) \geq \sum_j \pi_j \log \gamma_j(r_j\theta - u). \tag{5}$$

The following theorem provides deeper insight into the structure of the construction just described:

Theorem 3.1 *For every $\theta \geq 0$, there exists a unique $u(\theta)$, such that $\phi\big(\theta, u(\theta)\big) = 1$. This value satisfies $\bar{r}\theta \leq u(\theta) \leq \hat{r}\theta$. The function $u(\theta)$ is analytic and strictly increasing. In particular, $u'(0) = \bar{r}$.*

PROOF By construction, $A(\theta, u)$ is continuous and differentiable in both its arguments and by irreducibility, its spectral radius $\phi(\theta, u)$ is a simple eigenvalue. Through standard eigenvalue perturbation theory (see e.g., [12]) we see that $\phi(\theta, u)$ is continuous and differentiable in both its arguments. Furthermore, $\partial A(\theta, u)/\partial u = -\Gamma'(\theta, u)P < 0$, which, together with the irreducibility of $A(\theta, u)$ implies strict monotonicity for $\phi(\theta, u)$, specifically, $\partial\phi(\theta, u)/\partial u < 0$. Therefore, for some fixed θ, existence of a single u satisfying $\phi(\theta, u) = 1$, will follow if we determine two values of u corresponding to values of $\phi(\theta, \cdot)$ respectively greater and less than unity. To this end, $\Gamma(\theta, \hat{r}\theta) < I$, because not all rates are equal. Thus, by the discussion in subsection 2.3, $\phi(\theta, \hat{r}\theta) < \rho(P) = 1$. For the other bound, let $u^* = \max_j\{r_j\theta - \omega_j^*\}$ (u^* may be equal to $-\infty$). In case $u^* \geq \bar{r}\theta$, we get $\lim_{u\to u^*+0} \phi(\theta, u) = +\infty$, because at least one term to the right of (5) will tend to infinity and all the possibly negative terms will remain bounded. By continuity, there exists $u > \bar{r}\theta$ such that $\phi(\theta, u) > 1$. In case $u^* < \bar{r}\theta$, the combined application of Lemma 2.4, (3) and (5) yields $\log \phi(\theta, \bar{r}\theta) \geq \sum_j \pi_j \log \gamma_j(r_j\theta - \bar{r}\theta) \geq \sum_j \pi_j \tau_j (r_j - \bar{r})\theta = 0$. Since $\partial\phi(\theta, u)/\partial u < 0$ for every θ, u, an application of the implicit function theorem completes the proof for the existence and the bounds.

Now note that that all $\gamma_j(\cdot)$ are analytic functions of their argument, hence $\Gamma(\cdot, \cdot)$ is analytic in both u and θ. Furthermore, $A(\theta, u) = \Gamma(\theta, u)P$ is an irreducible matrix, hence, by (4), $\phi(\theta, u)$ is a simple eigenvalue and, by standard perturbation theory [12], an analytic function in perturbations of $\Gamma(\cdot, \cdot)$. The analyticity of $\phi(\theta, u)$ in both its arguments and, by the implicit function theorem, of $u(\theta)$ follows by combining the two previous statements.

For the monotonicity, observe that $u'(\theta) = \frac{-\partial\phi(\theta,u(\theta))/\partial\theta}{\partial\phi(\theta,u(\theta))/\partial u}$ and we have already seen that $\partial\phi/\partial u < 0$. Let $\boldsymbol{x}(\theta)$ and $\boldsymbol{y}(\theta)$ denote respectively the left and right Perron eigenvectors of $A\big(\theta, u(\theta)\big)$; then by standard perturbation analysis [12], $\partial\phi(\theta, u(\theta))/\partial\theta = \frac{\boldsymbol{x}(\theta)R\Gamma'(\theta,u(\theta))P\boldsymbol{y}(\theta)}{\boldsymbol{x}(\theta)\boldsymbol{y}(\theta)} > 0$, since $\boldsymbol{x}(\theta)^T, \boldsymbol{y}(\theta) \overset{+}{>} \mathbf{0}$. Therefore, $u'(\theta) > 0$. Finally, note that $u(0) = 0$ and, therefore, $A(0,0) = P$, $\boldsymbol{x}(0) = \boldsymbol{\pi}$ and $\boldsymbol{y}(0) = \mathbf{1}$; application of the previous formula yields $\partial\phi(0,0)/\partial\theta = \sum_j \pi_j\tau_j r_j$. A similar perturbation formula establishes that $\partial\phi(0,0)/\partial u = -\sum_j \pi_j\tau_j$, completing the proof. ■

The importance of $u(\theta)$ stems from the next theorem, which links the function to the effective bandwidth theory; specifically, in the notation of subsection 2.4, we have:

Theorem 3.2 *For every $\theta \geq 0$, the $\lim_{t\to\infty}(V^*(\theta,t)/t)$ exists and is equal to $u(\theta)$.*

PROOF For $\theta = 0$ trivially, $\lim_{t\to\infty}(V^*(0,t)/t) = 0 = u(0)$. Therefore assume $\theta > 0$ and define $q(s, s+t) \overset{\text{def}}{=} V(s, s+t) - t(u(\theta)/\theta) = \int_s^{s+t}[r(t) - u(\theta)/\theta]\,dt$; thus

$$\mathbb{E}\, e^{\theta q(s,s+t)} = e^{-u(\theta)t}\, \mathbb{E}\, e^{\theta V(s,s+t)}. \tag{6}$$

Let n stand for the number of state-transitions having occurred within $[s, s+t]$, at the instances $t_0 \leq s \leq t_1 \leq \ldots \leq t_n \leq s+t \leq t_{n+1}$ and denote the destination states as $\iota_0, \ldots, \iota_{n+1}$. Finally, let $S_j \overset{\text{def}}{=} t_{j+1} - t_j$; the S_j are mutually independent r.v.s with the same distributions as T_j. In this notation,

$$q(s, s+t) \leq q^+(s, s+t) = \begin{cases} [(r_{\iota_0} - u(\theta)/\theta)S_0]^+ & n = 0, \\ [(r_{\iota_0} - u(\theta)/\theta)S_0]^+ + \\ \displaystyle\sum_{j=1}^{n-1}(r_{\iota_j} - u(\theta)/\theta)S_j + \\ [(r_{\iota_n} - u(\theta)/\theta)S_n]^+ & n > 0, \end{cases}$$

where $[x]^+ = \max(x, 0)$. (The exact value of $q(s, s+t)$ would have been obtained, if S_0 and S_n had been replaced by $t_1 - s$ and $s + t - t_n$, respectively, and the $[\cdot]^+$ operators had been dropped.) Similarly, a lower bound $q^-(s, s+t)$ may be obtained by replacing $[x]^+$, by $[x]^- \overset{\text{def}}{=} -[-x]^+$. We will now show that there exist constants $b^+(\theta)$ and $b^-(\theta)$, such that

$$\begin{aligned} 0 &< b^-(\theta) \leq \mathbb{E}\, e^{\theta q^-(s,s+t)} \leq \mathbb{E}\, e^{\theta q(s,s+t)} \leq \cdots \\ \cdots &\leq \mathbb{E}\, e^{\theta q^+(s,s+t)} \leq b^+(\theta) < \infty, \quad \forall s, t,\ \forall\theta. \end{aligned} \tag{7}$$

We restrict ourselves to the upper bound, the arguments for the lower bound being completely analogous. Consider $\mathbb{E}\big(e^{\theta q^+(s,s+t)} \bigm| n\big)$; a standard backward-equations argument, slightly modified to accommodate the boundary conditions, yields $\mathbb{E}\big(e^{\theta q^+(s,s+t)} \bigm| n\big) = \boldsymbol{\pi}(s)\boldsymbol{w}_n(\theta)$, where $\boldsymbol{\pi}(s)$ denotes the vector of state-occupancy probabilities at time s and where $\boldsymbol{w}_n(\theta) = (w_{n,1}(\theta), \ldots, w_{n,K}(\theta))^T$ is given by

$$\boldsymbol{w}_n(\theta) = \begin{cases} \Gamma^+\big(\theta, u(\theta)\big)\mathbf{1} & n = 0, \\ \Gamma^+\big(\theta, u(\theta)\big)A(\theta, u(\theta))^{n-1}\Gamma^+\big(\theta, u(\theta)\big)\mathbf{1} & n > 0, \end{cases}$$

where $\Gamma^+(\theta, u) = \operatorname{diag}\big\{\gamma_j\big([r_j\theta - u]^+\big)\big\}$. By its structure, $\boldsymbol{w}_n(\theta) \overset{+}{>} \mathbf{0}$, for all n, θ. Recall that $A(\theta, u(\theta))$ is a nonnegative irreducible matrix, and $\rho\big(A(\theta, u(\theta))\big) = \phi(\theta, u(\theta)) = 1$. If, in addition, $A(\theta, u(\theta))$ is primitive, then it is convergent and $\lim_n A(\theta, u(\theta))^n = Q$, a strictly positive matrix of rank one. Hence, $\lim_n w_{n,j}(\theta) = \big(\Gamma^+\big(\theta, u(\theta)\big)Q\Gamma^+\big(\theta, u(\theta)\big)\mathbf{1}\big)_j < \infty$. The other possibility is that $A(\theta, u(\theta))$ is cyclic of period (say) m, in which case $\lim_n A(\theta, u(\theta))^{mn} = Q^*$, the direct sum of m strictly positive matrices of rank one. Then, the sequences $w_{n,j}(\theta)$ are not convergent, but $\limsup_n w_{n,j}(\theta) = \max_{0\leq l\leq m-1}\big(\Gamma^+\big(\theta, u(\theta)\big)A(\theta, u(\theta))^l Q^*\Gamma^+\big(\theta, u(\theta)\big)\mathbf{1}\big)_j < \infty$. We have thus established that the sequences $w_{n,j}(\theta)$ are upper bounded in both cases; it follows that there exists a vector $\boldsymbol{w}^*(\theta) \overset{+}{>} \mathbf{0}$, such that $\boldsymbol{w}_n(\theta) \leq \boldsymbol{w}^*(\theta)$, for all n. Consequently, for all s, $\mathbb{E}\, e^{\theta q^+(s,s+t)} = \mathbb{E}\,\mathbb{E}\big(e^{\theta q^+(s,s+t)} \bigm| n\big) \leq \boldsymbol{\pi}(s)\boldsymbol{w}^*(\theta) \leq \max_j w_j^*(\theta) \equiv b^+(\theta)$, and the upper bound is proved.

With the bounds of (7) at hand, equation (6) directly yields $\log b^-(\theta) + u(\theta)t \leq V^*(\theta, t) = \sup_s \log \mathbb{E}\, e^{\theta V(s,s+t)} \leq \log b^+(\theta) + u(\theta)t$, and the result follows. ■

Theorem 3.2 proves the existence of an effective bandwidth for the class of models under consideration and, at the same time, determines the effective bandwidth function. Indeed, the theorem directly verifies that condition R2 holds (R1 had been verified earlier), establishing the existence of the effective bandwidth $a(\theta)$. Comparison with condition R2 yields

$$a(\theta) = \frac{u(\theta)}{\theta} \qquad \forall\theta > 0. \tag{8}$$

Theorem 3.1 assures that $a(\theta)$ is an analytic function. Furthermore, the bounds of the same theorem directly translate to the 'typical' bounds

$$\bar{r} \leq a(\theta) \leq \hat{r} \qquad \forall\theta > 0. \tag{9}$$

Note that the lower bound is tight, since, by L' Hospital's rule and Theorem 3.1, $\lim_{\theta\to 0} a(\theta) = u'(0) = \bar{r}$.

We now examine the monotonicity of the effective bandwidth. By the discussion in subsection 2.4 we know that $a(\theta)$ is increasing in general. However, for the class of models at hand, we can be more precise. To this end, we introduce the following conditions, collectively referred to as 'conditions S':

S1 *For every state j, such that $r_j \neq \bar{r}$, T_j is a.s. constant;*

S2 *The matrix P^TP is completely decomposable.*

Given conditions S, we can state:

Theorem 3.3 *The function $u(\theta)$ (respectively: $a(\theta)$) is either strictly convex (resp.: strictly increasing) or linear and equal to $\bar{r}\theta$ (resp.: constant and equal to $\bar{r}$) for all $\theta \in [0, \infty)$. Linearity (resp.: $a(\theta) = \bar{r}$) applies only if conditions S hold.*

PROOF The function $u(\theta)$ is analytic (by Theorem 3.1) and convex (by Theorem 3.2 and the discussion in subsection 2.4). Therefore, Lemma 2.2 applies and $u(\theta)$ is either strictly convex or linear in its whole domain. If it is linear, necessarily $u(\theta) = \bar{r}\theta$, in order to satisfy $u(0) = 0$ and $u'(0) = \bar{r}$. Assume linearity; by employing (5), Lemma 2.4 and (3), it follows $0 = \log\phi(\theta, \bar{r}\theta) \geq \sum_j \pi_j \log\gamma_j(r_j\theta - \bar{r}\theta) \geq \sum_j \pi_j\tau_j(r_j - \bar{r})\theta = 0$, for all θ, and all inequalities must be equalities. Lemma 2.4 suggests that the second inequality can be equality only if condition S1 holds. Similarly, Lemma 2.6 requires condition S2 for the other inequality. The assertions for $a(\theta)$ follow immediately by (8) and Lemma 2.1. ■

Theorem 3.3 reveals that for the models under consideration two possibilities may occur. In the first case, the effective bandwidth function $a(\theta)$ is strictly increasing, condition R3 holds and, according to the discussion in subsection 2.4, a queue with output capacity c, when loaded by the source under discussion, reaches steady-state, whereupon the CPDF of the buffer occupancy has an exponential tail with parameter $a^{-1}(c)$. Equivalently, for admission control purposes, if the buffering capacity is $B \gg 0$ and the target QoS is $\epsilon \ll 1$, bandwidth equal to $a(-\log\epsilon/B)$ must be allocated to the source. The usual Markovian fluid models always fall in this category, since they violate condition S1.

The other possibility, for which conditions S are necessary prerequisites, is that the effective bandwidth is constant and equal to the source's mean rate. For these sources, any value of $c > \bar{r}$ yields $\sup\{\,\theta \mid a(\theta) < c\,\} = \infty$ and opens the possibility that such sources may enjoy loss-less performance, when being fed to a queue with a finite amount of buffering, even when only bandwidth equal to their mean rate has been allocated to them. We will return to this point shortly.

Note that conditions S are necessary but *not* sufficient. As the remarks following Lemma 2.6 have already pointed out, there exist cases where both S1–2 are fulfilled, but $a(\theta)$ is still strictly increasing. In general, the rates r_j must satisfy some relation (dependent on the structure of P), in addition to conditions S, for $a(\theta)$ to be constant. In section 4 we will provide a replacement for S2 which, together with S1, constitute necessary and sufficient conditions for the degenerate case.

We now focus on a subclass of the considered models that contains exactly the cases where the Markov chain P is of the fixed form (1) (where, necessarily, all nonzero elements are equal to one). We will call models in this subclass *completely cyclic*. These models include all general On/Off traffic streams and in fact generalize them, with respect to the number of possible data-generation rates. As has already been mentioned in subsection 2.3, for completely cyclic models relation (5) applies with equality, regardless of the distribution of the various T_j. (It can be additionally verified that $\pi_j = 1/K$, $\forall j$.) Direct calculation will convince that condition S1 is necessary and sufficient for a constant effective bandwidth. We coin the term *deterministic completely cyclic* sources for those completely cyclic sources which satisfy condition S1. Although such sources may still display random aspects (if they possess states with rate equal to $\bar{r}$ that are sustained for random time intervals) this is immaterial for all intents and purposes of interest, as it will be clarified momentarily.

Let us now return to the issue of loss-less performance, for the case of deterministic completely cyclic sources. Suppose such a source loads a queue with output capacity equal to the source's mean rate, which in this case equals $\bar{r} = (\sum_j \tau_j r_j)/(\sum_j \tau_j)$; we are about to study how the queue content builds up with time. As it was just declared, the possible presence of states with rate equal to $\bar{r}$ is immaterial; indeed, during sojourns at such states, the queue content remains invariant. Similarly, inspection of (3) verifies that addition or removal of states with rate equal to the mean rate does not modify the value of the latter. In consequence, such states may readily be removed and we are left with a completely deterministic system. Furthermore, since the source periodically cycles through all states, marking a state as 'first' is also immaterial. Due to the periodicity, we just have to study a single cycle. Within such a cycle, the queue content evolves in a piece-wise linear fashion, increasing if, for the current state j, $r_j > \bar{r}$, decreasing otherwise. Consequently, the local extrema of the queue content occur at the sojourn boundaries. At the end of the cycle, the queue content is equal to the value it had at the cycle's beginning. In result, starting with an empty buffer, the maximum buffer content during the cycle is

$$B_{\max} = \max_{1\le j\le K} \sum_{l=1}^{j} (r_l - \bar{r})\tau_l. \tag{10}$$

If this amount of buffering is supplied to the source, indeed loss-less performance is achieved. Note that although the relative sequence in which the states are visited is immaterial for the determination of the mean rate value (which, in this case, is the required bandwidth to be allocated), the sequence matters for determining the maximal buffer occupancy. To complete the thread of thought, see that in a setting where multiple deterministic completely cyclic sources are multiplexed, a conservative upper bound for the total buffer requirements is the sum of the individual quantities $B_{\max}$, as given by (10). It is conservative since it does not take into account rate cancellations arising from multiplexing.

The basic ideas underlying this trivial model will be elaborated in section 4 to cover the general class of models under scope.

4 The limiting effective rate

In the previous section we encountered sources, namely those of the deterministic completely cyclic kind, for which $a(\theta) = \bar{r}$, for all θ. In consequence, for such sources, $a(\infty) \stackrel{\text{def}}{=} \lim_{\theta\to\infty} a(\theta) = \bar{r} < \hat{r}$. It follows that for the class of models discussed in this paper, it is not necessary for the effective bandwidth to approach the peak rate as the performance constraint becomes unboundedly stringent. This section will investigate the issue in its generality.

We start by revisiting the simple completely cyclic case (this time imposing no requirement on the distributions of the sojourn times) and marking the important elements that allow for $a(\infty) < \hat{r}$. Indeed, if we load a queue with output capacity c with a completely cyclic source, the buffer content having been contributed by the source and not yet served by the end of the cycle is $\sum_j (r_j - c)T_j$. If c is chosen equal to the limiting effective bandwidth of the source, namely $a(\infty)$, then this buffer content must be zero, under the most stringent circumstances, i.e., letting $T_j = \tau_j^+$ for all those j featuring $r_j > c$, while letting $T_j = \tau_j^-$ when $r_j < c$. This reasoning leads to the conclusion that $a(\infty) \ge r_j$ for all j featuring $\tau_j^+ = \infty$. Furthermore, if $a(\infty)$ is to be maintained less than $\hat{r}$, then there must exist states j with $r_j < \hat{r}$, such that $\tau_j^- > 0$, otherwise, the positive data content contributed while the peak-rate state was visited may not be counterbalanced.

The previous arguments, although derived heuristically on the basis of the simplest models under scope, are in fact in the heart of the general answer. Indeed, we can immediately formalize and generalize the first of these heuristic arguments. To this end, let $\mathcal{S}_{\inf} \stackrel{\text{def}}{=} \{\,j \mid \tau_j^+ = \infty\,\}$. We can now state:

Lemma 4.1 *If $\mathcal{S}_{\inf} \neq \emptyset$ then $a(\infty) \ge \max_{j\in\mathcal{S}_{\inf}} r_j \stackrel{\text{def}}{=} r^*$.*

PROOF Assume there exists $k \in \mathcal{S}_{\inf}$, such that $r_k > a(\infty)$. Combining (5) with (8) yields $\sum_j \pi_j \log\gamma_j(\theta r_j - \theta a(\theta)) \le$

$\log \phi(\theta, \theta a(\theta)) = 0$, for all $\theta \geq 0$. Let $\mathcal{L} = \{ j \mid a(\infty) \leq r_j \}$ and denote its complement by $\tilde{\mathcal{L}}$. Since $a(\theta)$ is an increasing function, $a(\theta) \leq r_l$, for all $l \in \mathcal{L}$ and the corresponding terms in the previous inequality are nonnegative; by excluding all of them except the one that corresponds to k, we are led to $\sum_{j \in \tilde{\mathcal{L}}} \pi_j \log \gamma_j\big(\theta(r_j - a(\theta))\big) + \pi_k \log \gamma\big(\theta(r_k - a(\theta))\big) \leq 0$. Dividing by θ and taking the limit, we obtain (due to Lemma 2.5 and the fact that all τ_j^- are finite), $+\infty = \sum_{j \in \tilde{\mathcal{L}}} \pi_j (r_j - a(\infty))\tau_j^- + \pi_k (r_k - a(\infty))\tau_k^+ \leq 0$, a contradiction. ■

To proceed further we need to take a closer look on the combinatorial structure of the underlying Markov chain. Let $\boldsymbol{\iota} = \big(\iota_1, \ldots, \iota_{\ell(\boldsymbol{\iota})}\big)$ be a sample path of states visited in sequence by the Markov chain and denote its length by $\ell(\boldsymbol{\iota})$. Given that the Markov chain starts at state ι_1, the path $\boldsymbol{\iota}$ is realized with probability $\varpi_{\boldsymbol{\iota}} = \prod_{j=1}^{\ell(\boldsymbol{\iota})-1} P_{\iota_j, \iota_{j+1}}$. Of course, the path exists iff $\varpi_{\boldsymbol{\iota}} > 0$. A path $\boldsymbol{\iota}$ is called a circuit if $\boldsymbol{\iota}' = \big(\iota_1, \ldots, \iota_{\ell(\boldsymbol{\iota})}, \iota_1\big)$ is a realizable path, i.e., $\varpi_{\boldsymbol{\iota}'} > 0$. (Although it is customary to define a circuit to include the 'closing' state both at the beginning and the end, in this setting it is appropriate to define the circuit to end just before it closes.) A circuit is called simple if all its states are visited once. A given circuit can always be analyzed into simple circuits.

Let $\mathcal{C}$ denote the set of simple circuits defined by the combinatorial structure of P. (Note that, although the set of circuits is infinite, $\mathcal{C}$ is always a finite set.) The functions on simple circuits that we will encounter are invariant to state permutations; accordingly, for each $\boldsymbol{\iota} \in \mathcal{C}$ we define $\boldsymbol{v}(\boldsymbol{\iota}) = \big(v_1, \ldots, v_{\ell(\boldsymbol{\iota})}\big)$ as a permutation of $\boldsymbol{\iota}$, such that $r_{v_1(\boldsymbol{\iota})} \leq \ldots \leq r_{v_{\ell(\boldsymbol{\iota})}(\boldsymbol{\iota})}$; $\boldsymbol{v}(\boldsymbol{\iota})$ need not be an actual path. In this notation, $r_{v_{\ell(\boldsymbol{\iota})}(\boldsymbol{\iota})} = \max_{1 \leq j \leq \ell(\boldsymbol{\iota})} r_{\iota_j}$. We further let $\mathcal{V} \stackrel{\text{def}}{=} \{ \boldsymbol{v}(\boldsymbol{\iota}) \mid \boldsymbol{\iota} \in \mathcal{C} \}$ (note that they may exist more than one $\boldsymbol{\iota} \in \mathcal{C}$ leading to the same $\boldsymbol{v}(\boldsymbol{\iota})$) and $\mathcal{V}_f \stackrel{\text{def}}{=} \{ \boldsymbol{v} \in \mathcal{V} \mid v_{\ell(\boldsymbol{v})} \notin \mathcal{S}_{\inf} \} \subseteq \mathcal{V}$. Finally we define

$$\chi_{\boldsymbol{\iota}}(\theta) \stackrel{\text{def}}{=} \theta^{-1} \sum_{j=1}^{\ell(\boldsymbol{\iota})} \log \gamma_{\iota_j}(\theta r_{\iota_j} - \theta a(\theta)), \quad \forall \theta > 0, \quad \forall \boldsymbol{\iota} \in \mathcal{C}. \tag{11}$$

By construction, $\chi_{\boldsymbol{\iota}}(\theta) = \chi_{\boldsymbol{v}(\boldsymbol{\iota})}(\theta)$. The following lemma describes properties to be used later on:

Lemma 4.2 *For every $\boldsymbol{\iota} \in \mathcal{C}$, $\limsup_{\theta \to \infty} \chi_{\boldsymbol{\iota}}(\theta) \leq 0$. Furthermore, there exists $\boldsymbol{\iota}^* \in \mathcal{C}$, such that $a(\infty) \leq r_{v_{\ell(\boldsymbol{\iota}^*)}(\boldsymbol{\iota}^*)}$ and $\limsup_{\theta \to \infty} \chi_{\boldsymbol{\iota}^*}(\theta) = 0$.*

The proof is in the appendix.

We are now ready to embark on the investigation for the value of $a(\infty)$. The basic idea is to examine all the simple circuits $\boldsymbol{\iota} \in \mathcal{C}$ in isolation, as if each of them belonged to a completely cyclic model. Since the quantities of interest are invariant to permutation of states, we base our investigation on $\mathcal{V}$ instead. For a $\boldsymbol{v} \notin \mathcal{V}_f$ things are not very interesting, since by Lemma 4.1 we know that $a(\infty) \geq r_{v_{\ell(\boldsymbol{v})}}$. Thus, we further restrict ourselves to $\mathcal{V}_f$. For every $\boldsymbol{v} \in \mathcal{V}_f$ we define

$$l(\boldsymbol{v}) \stackrel{\text{def}}{=} \min\{ j \mid 1 \leq j \leq \ell(\boldsymbol{v}),\ v_k \notin \mathcal{S}_{\inf}, \forall k \geq j \}. \tag{12}$$

Since $\boldsymbol{v} \in \mathcal{V}_f$, $l(\boldsymbol{v})$ is well defined. We now introduce the quantities

$$\bar{r}_j(\boldsymbol{v}) \stackrel{\text{def}}{=} \frac{\displaystyle\sum_{k<j} r_{v_k} \tau_{v_k}^- + \sum_{k \geq j} r_{v_k} \tau_{v_k}^+}{\displaystyle\sum_{k<j} \tau_{v_k}^- + \sum_{k \geq j} \tau_{v_k}^+}, \ \forall \boldsymbol{v} \in \mathcal{V}_f,\ l(\boldsymbol{v}) \leq j \leq \ell(\boldsymbol{v}) \tag{13}$$

and

$$\begin{aligned} B_j(\boldsymbol{v}) \stackrel{\text{def}}{=} & \sum_{k<j} (r_{v_k} - r_{v_j}) \tau_{v_k}^- \\ & + \sum_{k \geq j} (r_{v_k} - r_{v_j}) \tau_{v_k}^+, \ \forall \boldsymbol{v} \in \mathcal{V}_f,\ l(\boldsymbol{v}) \leq j \leq \ell(\boldsymbol{v}). \end{aligned} \tag{14}$$

These quantities have the following 'physical' meaning: each $\boldsymbol{\iota} \in \mathcal{C}$ that corresponds to $\boldsymbol{v}$ is assumed associated with an isolated completely cyclic model; then the sojourn times are replaced by a.s. constant counterparts, some equal to the essential supremum of the original r.v., the rest to the essential infimum. In fact, by (13), $\bar{r}_j(\boldsymbol{v})$ is then equal to the mean rate of the induced deterministic completely cyclic model. Similarly, $B_j(\boldsymbol{v})$ is equal to the buffer content (contributed during a cycle and not yet served) of a virtual queue possessing an output capacity equal to r_{v_j} and being loaded by the induced deterministic completely cyclic model. The usage of the extrema (τ_j^+ and τ_j^-) are in accordance with the heuristic discussion at the beginning of this section. Incidentally, the definition of $l(\boldsymbol{v})$ above was chosen so that (13) and (14) always yield finite values. The following lemma summarizes the basic properties of $\bar{r}_j(\boldsymbol{v})$ and $B_j(\boldsymbol{v})$:

Lemma 4.3 *For every $\boldsymbol{v} \in \mathcal{V}_f$:*

a. *For all j such that $l(\boldsymbol{v}) \leq j \leq \ell(\boldsymbol{v})$ it holds:*

$$\bar{r}_j(\boldsymbol{v}) \begin{array}{c} > \\ = \\ < \end{array} r_{v_j} \iff B_j(\boldsymbol{v}) \begin{array}{c} > \\ = \\ < \end{array} 0.$$

b. *$B_j(\boldsymbol{v})$ is decreasing in j and there exists $j \leq \ell(\boldsymbol{v})$ such that $B_j(\boldsymbol{v}) \leq 0$.*

c. *Let $j^*(\boldsymbol{v}) = \min\{ j \mid B_j(\boldsymbol{v}) \leq 0 \}$. Then, $\bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) = \max_{l(\boldsymbol{v}) \leq j \leq \ell(\boldsymbol{v})} \bar{r}_j(\boldsymbol{v}) \stackrel{\text{def}}{=} \bar{r}^*(\boldsymbol{v})$ and it holds $r_{v_1} \leq \cdots \leq r_{v_{j^*(\boldsymbol{v})-1}} < \bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) \leq r_{v_{j^*(\boldsymbol{v})}} \leq \cdots \leq r_{v_{\ell(\boldsymbol{v})}}$.*

PROOF Let $W_j(\boldsymbol{v})$ denote the denominator in (13). Then, item (a) follows immediately from the identity $B_j(\boldsymbol{v}) = (\bar{r}_j(\boldsymbol{v}) - r_{v_j}) W_j(\boldsymbol{v})$. To prove (b), observe that $B_{j+1}(\boldsymbol{v}) = B_j(\boldsymbol{v}) + (r_{v_j} - r_{v_{j+1}}) W_{j+1}(\boldsymbol{v}) \leq B_j(\boldsymbol{v})$, because $r_{v_j} \leq r_{v_{j+1}}$ for all j; thus $B_j(\boldsymbol{v})$ is a decreasing sequence. Furthermore, $B_{\ell(\boldsymbol{v})}(\boldsymbol{v}) = \sum_{k<\ell(\boldsymbol{v})} \tau_{v_k}^- (r_{v_k} - r_{v_{\ell(\boldsymbol{v})}}) \leq 0$, proving the existence of a nonpositive $B_j(\boldsymbol{v})$. Finally, for item (c), a few algebraic manipulations will show that $\bar{r}_j(\boldsymbol{v}) - \bar{r}_{j+1}(\boldsymbol{v}) = -(\tau_{v_j}^+ - \tau_{v_j}^-) B_j(\boldsymbol{v}) / (W_j(\boldsymbol{v}) W_{j+1}(\boldsymbol{v}))$. By the definition of $j^*(\boldsymbol{v})$ we see that $\bar{r}_j(\boldsymbol{v})$ is increasing for $j < j^*(\boldsymbol{v})$ and decreasing for $j > j^*(\boldsymbol{v})$, proving in effect that $\bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) = \bar{r}^*(\boldsymbol{v})$. The ordering follows from $r_{v_{j^*(\boldsymbol{v})-1}} < \bar{r}_{j^*(\boldsymbol{v})-1}(\boldsymbol{v}) \leq \bar{r}^*(\boldsymbol{v}) = \bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) \leq r_{v_{j^*(\boldsymbol{v})}}$, the first and last inequalities being due to the definition of $j^*(\boldsymbol{v})$ and item (a). ■

Part of Lemma 4.3 may be interpreted as follows: for each $\boldsymbol{v} \in \mathcal{V}_f$, $r_{v_{j^*(\boldsymbol{v})}}$ is the smallest actual rate (i.e., corresponding directly to a state among those states within $\boldsymbol{v}$) that, if used as the output capacity of the virtual queue will allow loss-less performance. Furthermore, $\bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) = \bar{r}^*(\boldsymbol{v})$ is the 'optimal' rate for $\boldsymbol{v}$, in the sense that it removes the excess slack, so that at the end of each cycle the buffer residue is zero. Lemma 4.3, besides establishing this fact, also provides the algorithmic way to determine $\bar{r}^*(\boldsymbol{v})$. One starts with $j = l(\boldsymbol{v})$ and progresses onwards, until $B_j(\boldsymbol{v})$ becomes nonpositive. The current $\bar{r}_j(\boldsymbol{v})$ is the required value.

Since the quantity $\bar{r}^*(\boldsymbol{v})$ characterizes, in some sense, optimally all $\boldsymbol{\iota} \in \mathcal{C}$ featuring $\boldsymbol{v}(\boldsymbol{\iota}) = \boldsymbol{v}$, intuition suggests that an appropriate maximization, over the 'most stringent path' will provide the sought value of $a(\infty)$ in general. This is in fact true and is put precisely in the following:

Theorem 4.1 *If $\mathcal{V}_f \neq \emptyset$, let $\bar{r}^* \stackrel{\text{def}}{=} \max_{\boldsymbol{v}\in\mathcal{V}_f} \bar{r}^*(\boldsymbol{v})$, otherwise let $\bar{r}^* = 0$. Similarly, if $\mathcal{S}_{\text{inf}} \neq \emptyset$ let r^* be as in Lemma 4.1, otherwise set $r^* = 0$. Lastly, define $r_{\max}$ as $r_{\max} \stackrel{\text{def}}{=} \max\{\bar{r}^*, r^*\}$. It then holds $a(\infty) = r_{\max}$.*

PROOF Suppose $a(\infty) < r_{\max}$. Due to Lemma 4.1 this may only be possible if $r^* \leq a(\infty) < \bar{r}^*$, implying also that $\mathcal{V}_f \neq \emptyset$. Pick a $\boldsymbol{v} \in \mathcal{V}_f$ that achieves $\bar{r}^*$, a $\boldsymbol{\iota} \in \mathcal{C}$, such that $\boldsymbol{v}(\boldsymbol{\iota}) = \boldsymbol{v}$ and let $j^*(\boldsymbol{v})$ be as in part (c) of Lemma 4.3. By this part we get $r_{v_{j^*(\boldsymbol{v})}} \geq \bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) = \bar{r}^* > a(\infty)$. Furthermore, there exists a $k < j^*(\boldsymbol{v})$ such that $a(\infty) \geq r_{v_k}$. For, if not, $a(\infty) < r_{v_j}$ for all j and for the chosen circuit $\boldsymbol{\iota}$ we obtain, by Lemma 2.5, that $\limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}}(\theta) = \sum_j (r_{v_j} - a(\infty)) \lim_{\omega\to+\infty} \omega^{-1} \log \gamma_{v_j}(\omega) = \sum_j (r_{v_j} - a(\infty))\tau^+_{v_j} > 0$, contradicting Lemma 4.2. We have thus established that

$$\exists k \leq j^*(\boldsymbol{v}): \qquad r_{v_{k-1}} \leq a(\infty) < r_{v_k}. \tag{15}$$

If equality occurs on the left side, we let $\mathcal{E} = \{ j \mid r_{v_j} = r_{v_{k-1}} \}$, otherwise we set $\mathcal{E} = \emptyset$. Observe that for all $j \in \mathcal{E}$ we have $r_{v_j} \geq a(\theta)$ for all θ, because $a(\theta)$ is increasing. Consequently, $\chi_{\boldsymbol{\iota}}(\theta) \geq \theta^{-1} \sum_{j\notin\mathcal{E}} \log \gamma_{v_j}(\theta r_{v_j} - \theta a(\theta))$ and to the limit (using Lemmas 2.5 and 4.2)

$$\begin{aligned}
0 &\geq \limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}}(\theta) \\
&\geq \sum_{\substack{j\leq k-1 \\ j\notin\mathcal{E}}} (r_{v_j} - a(\infty))\tau^-_j + \sum_{j\geq k} (r_{v_j} - a(\infty))\tau^+_j \\
&= \sum_{j\leq k-1} (r_{v_j} - a(\infty))\tau^-_j + \sum_{j\geq k} (r_{v_j} - a(\infty))\tau^+_j,
\end{aligned}$$

which, upon rearrangement yields

$$a(\infty) \geq \bar{r}_k(\boldsymbol{v}) \tag{16}$$

Now, if $k = j^*(\boldsymbol{v})$ then (16) leads to $a(\infty) \geq \bar{r}^*(\boldsymbol{v}) = \bar{r}^*$, contradicting our assumption. Therefore, necessarily $k < j^*(\boldsymbol{v})$; in this case, however, Lemma 4.3 yields $B_k(\boldsymbol{v}) > 0$, which (by part (a) of the same lemma) leads to $\bar{r}_k(\boldsymbol{v}) > r_{v_k}$. This last relation, combined with (16) contradicts (15). It follows that, necessarily $a(\infty) \geq r_{\max}$.

Now suppose $a(\infty) > r_{\max}$. Pick the special circuit $\boldsymbol{\iota}^*$ of Lemma 4.2 and let $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{\iota}^*)$. Then, necessarily $\boldsymbol{v} \in \mathcal{V}_f$. For, if not, $v_{\ell(\boldsymbol{v})} \in \mathcal{S}_{\text{inf}}$ and, by Lemma 4.1, $a(\infty) \geq r_{v_{\ell(\boldsymbol{v})}}$. However, by Lemma 4.2 we also have $a(\infty) \leq r_{v_{\ell(\boldsymbol{v})}}$; in combination, $a(\infty) = r_{v_{\ell(\boldsymbol{v})}} \leq r^* \leq r_{\max}$, a contradiction to our hypothesis. Additionally, there exists an index k such that $r_{v_k} \leq a(\infty)$, for otherwise $r_{v_j} > a(\infty)$ for all j and we are led to $a(\infty) < \bar{r}^*(\boldsymbol{v}) \leq \bar{r}^* \leq r_{\max}$, again a contradiction to our hypothesis. Furthermore, by Lemma 4.1, $r_{v_{\ell(\boldsymbol{v})}} \geq a(\infty)$, thus $a(\infty)$ is also bounded from above. We conclude that there exists k such that $r_{v_{k-1}} \leq a(\infty) \leq r_{v_k}$, and we obtain

$$\begin{aligned}
0 &= \limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}^*}(\theta) \\
&\leq \sum_m \limsup_{\theta\to\infty} \left\{ (r_{v_m} - a(\theta)) \frac{\log \gamma_{v_m}(\theta r_{v_m} - \theta a(\theta))}{\theta(r_{v_m} - a(\theta))} \right\} \\
&\leq \sum_{m\leq k-1} (r_{v_m} - a(\infty))\tau^-_{v_m} + \sum_{m\geq k} (r_{v_m} - a(\infty))\tau^+_{v_m}.
\end{aligned}$$

(Note that for all m such that $r_{v_m} \neq a(\infty)$, lim sup is plain limit. In case there is m with $r_{v_m} = a(\infty)$, the results does not change, since, by our hypothesis $a(\infty) > r_{\max} \geq r^*$, i.e., $v_m \notin \mathcal{S}_{\text{inf}}$. Accordingly, both the respective left and right terms tend to 0.) The obtained inequality, upon rearrangement, yields $a(\infty) \leq \bar{r}_k(\boldsymbol{v}) \leq \bar{r}^*(\boldsymbol{v}) \leq \bar{r}^* \leq r_{\max}$, contradicting our hypothesis. Thus necessarily $a(\infty) = r_{\max}$. ∎

Theorem 4.1, together with Lemma 4.3, provide a general 'recipe' for computing the limiting effective bandwidth of the models considered in this paper. Its assertion, however, can be also used to determine the cases where the limiting value $a(\infty)$ remains less than the peak rate. To pursue this matter, let $\mathcal{P}$ denote the set of states with rate equal to the peak rate, viz., $\mathcal{P} = \{ k \mid r_k = \hat{r} \}$. The answer is then given by the following:

Theorem 4.2 *$a(\infty) < \hat{r}$ if and only if the following two conditions hold:*

i. *$\mathcal{P} \cap \mathcal{S}_{\text{inf}} = \emptyset$;*

ii. *Every $\boldsymbol{\iota} \in \mathcal{C}$ containing a state $k \in \mathcal{P}$ also contains at least one state $n \notin \mathcal{P}$, such that $\tau^-_n > 0$.*

PROOF Only if: assume $\mathcal{P} \cap \mathcal{S}_{\text{inf}} \neq \emptyset$ and let k belong to the intersection; we obtain $\hat{r} = r_k \leq r^* \leq r_{\max} = a(\infty)$. Now suppose condition (i) holds but (ii) doesn't. Pick a $\boldsymbol{\iota}$ that violates (ii) and let $\boldsymbol{v} = \boldsymbol{v}(\boldsymbol{\iota})$. Since (i) holds, $\boldsymbol{v} \in \mathcal{V}_f$. Then, using also Lemma 4.3, $a(\infty) = r_{\max} \geq \bar{r}^* \geq \bar{r}_{j^*(\boldsymbol{v})}(\boldsymbol{v}) \geq \bar{r}_{\ell(\boldsymbol{v})} = (\sum_{j<\ell(\boldsymbol{v})} \tau^-_{v_j} r_{v_j} + \tau^+_{\ell(\boldsymbol{v})} r_{\ell(\boldsymbol{v})})/(\sum_{j<\ell(\boldsymbol{v})} \tau^-_{v_j} + \tau^+_{\ell(\boldsymbol{v})}) = \hat{r}$, because all $\tau^-_{v_j} = 0$, except possibly when $r_{v_j} = \hat{r}$.

If: since $\mathcal{P} \cap \mathcal{S}_{\text{inf}} = \emptyset$, we obtain $r^* < \hat{r}$. Now pick a $\boldsymbol{v} \in \mathcal{V}_f$. If $r_{v_j} < \hat{r}$ for all j, then, by construction, $\bar{r}^*(\boldsymbol{v}) < \hat{r}$. Otherwise, there exists a k, such that $v_k \notin \mathcal{P}$ and $\tau^-_{v_k} > 0$. In this case all $\bar{r}_j(\boldsymbol{v})$ are strictly convex combinations of the rates of the participating states, hence $r_{v_j}(\boldsymbol{v}) < \hat{r}$ for all j. It follows again that $\bar{r}^*(\boldsymbol{v}) < \hat{r}$. Maximizing over all $\boldsymbol{v} \in \mathcal{V}_f$ brings $\bar{r}^* < \hat{r}$. Combining with $r^* < \hat{r}$, the result follows. ∎

As it can be verified by comparison to the beginning of this section, the general conditions for $a(\infty) < \hat{r}$ are essentially those heuristically derived there for the simple case of completely cyclic sources. The only difference is that one has to apply the 'tests' to all simple circuits defined by the underlying Markov chain. As a simple application, see that for on/off sources, the conditions for $a(\infty) < \hat{r}$ reduce exactly to the requirement that $\tau^+_{\text{on}} < \infty$ and $\tau^-_{\text{off}} > 0$. By following Theorem 4.1, we obtain $a(\infty) = \hat{r}\tau^+_{\text{on}}/(\tau^+_{\text{on}} + \tau^-_{\text{off}})$.

We close by noting that Theorem 4.1 may be used to provide a necessary and sufficient condition for detecting degenerate models that feature $a(\infty) = \bar{r}$ (recall that the combinations of conditions S1–2 in section 3 was necessary, but not sufficient). The result is established by recognizing that condition S1 simplifies the expression in (13) and that we must require $\bar{r}^*(\boldsymbol{v}) \leq \bar{r}$, for all $\boldsymbol{v} \in \mathcal{V}_f$. The precise result is as follows:

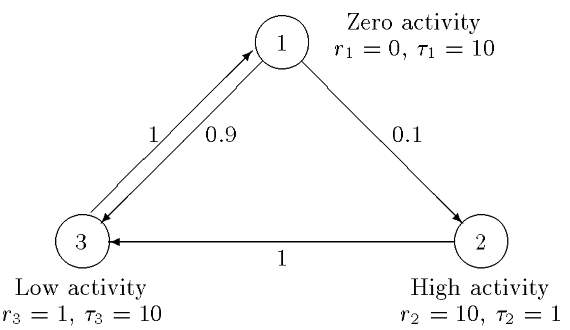

Figure 1: State transitions, traffic rates and mean sojourn times in the example model

Theorem 4.3 *The effective bandwidth function $a(\theta)$ is constant and equal to $\bar{r}$ if and only if in addition to condition S1 there holds*

$$\frac{\sum_{j=1}^{\ell(\boldsymbol{v})} \tau_{v_j} r_{v_j}}{\sum_{j=1}^{\ell(\boldsymbol{v})} \tau_{v_j}} \le \bar{r}, \quad \forall \boldsymbol{v} \in \mathcal{V}_f.$$

PROOF Given condition S1, all states feature constant sojourn times, except possibly for some states featuring a rate equal to $\bar{r}$. Let $\mathcal{E}$ denote the set of these exceptional states. We have $\mathcal{S}_{\inf} \subseteq \mathcal{E}$, thus $r^* = \bar{r}$, if $\mathcal{S}_{\inf} \ne \emptyset$, and $r^* = 0$ otherwise. Furthermore, $\mathcal{V}_f \ne \emptyset$, since it includes $\boldsymbol{v}(\iota)$ for all simple circuits ι that contain a state in $\mathcal{P}$, because the sojourn times at all such states are constant and thus the respective essential suprema finite. Since $a(\theta) \ge \bar{r}$ for all θ, Theorem 4.1 and the arguments just stated combine to $a(\infty) = r_{\max} = \bar{r}^* \ge \bar{r} \ge r^*$. Therefore, the statement $a(\infty) = \bar{r}$ is equivalent to $\bar{r}^* \le \bar{r}$, this in turn being equivalent to $\bar{r}_j(\boldsymbol{v}) \le \bar{r}$, for all $\boldsymbol{v} \in \mathcal{V}_f$ and all $j : l(\boldsymbol{v}) \le j \le \ell(\boldsymbol{v})$. Since all states k with constant sojourn times feature $\tau_k^- = \tau_k^+ = \tau_k$, (13) yields

$$\bar{r}_j(\boldsymbol{v}) = \frac{\displaystyle\sum_{k:v_k \notin \mathcal{E}} r_{v_k}\tau_{v_k} + \bar{r}\left(\sum_{k<j:v_k\in\mathcal{E}} \tau_{v_k}^- + \sum_{k\ge j:v_k\in\mathcal{E}} \tau_{v_k}^+\right)}{\displaystyle\sum_{k:v_k \notin \mathcal{E}} \tau_{v_k} + \left(\sum_{k<j:v_k\in\mathcal{E}} \tau_{v_k}^- + \sum_{k\ge j:v_k\in\mathcal{E}} \tau_{v_k}^+\right)}.$$

Now consider general fractions $f(x) = (A + sx)/(B + x)$, where all parameters are nonnegative. It may be trivially verified that, for any $x \ge 0$, there holds $f(x) \le s$ if and only if $f(0) = A/B \le s$. Applying this to the fraction equal to $\bar{r}_j(\boldsymbol{v})$ above, it follows that, instead of checking directly for $\bar{r}_j(\boldsymbol{v}) \le \bar{r}$ one may equivalently replace the parenthesized quantity with $\sum_{k:v_k\in\mathcal{E}} \tau_{v_k}$ and check this modified fraction against $\bar{r}$. However, the modified fraction is immediately seen to be equal to the one in the statement of the theorem. ■

5 A numerical example

We now present a numerical example that highlights some of the concepts and results in the paper. Consider a source-model featuring three states, as depicted in Figure 1. The traffic pattern corresponding to this model alternates between silent and active periods. Most of the times the active period features a low traffic generation rate of one bandwidth unit[3] and lasts on the average for 10 time units. Occasionally, however, the activity commences with a brief excitation, lasting on the average for 1 t.u., whereupon the source generates data at a rate of 10 b.u.; the excitation is followed by the 'usual' activity, which itself leads to the silent state.

Name	Parameters	Generator $\gamma(\omega)$	Mean τ	Sup. τ^+	Inf. τ^-
Exponen.	$s > 0$	$\frac{1}{1-\omega s}$	s	∞	0
Shifted exponen.	$s_1, s_2 > 0$	$\frac{e^{s_1\omega}}{1-s_2\omega}$	$s_1 + s_2$	∞	s_1
Uniform	$b > a \ge 0$	$\frac{e^{b\omega}-e^{a\omega}}{\omega(b-a)}$	$\frac{a+b}{2}$	b	a

Table 1: Classes of PDFs used in the example & associated properties

The transition matrix for the Markov Chain corresponding to this model obviously is

$$P = \begin{pmatrix} 0 & 0.1 & 0.9 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix},$$

and the corresponding stationary probability vector may be trivially verified equal to $\boldsymbol{\pi} = \big(10/21, 1/21, 10/21\big)$. Usage of (3) directly yields a mean rate equal to $\bar{r} = 0.5473$ b.u. The peak rate corresponds to the state with label 2 ($\hat{r} = r_2 = 10$ b.u.). Note the high burstiness of the source (equal to $10/0.573 \approx 17.5$). The mean rate value is also the limiting effective bandwidth value, as $\theta \to 0$, regardless of any other details in the stochastic behavior of the various sojourn times.

To obtain more information on the effective bandwidth function, explicit assignment of PDFs must be made to all sojourn times. The rest of this section presents detailed graphs of the effective bandwidth functions that result from various such assignments. The graphs were obtained by solving numerically for u the equation $\phi(\theta, u) = 1$ (see eq. (4)), for all desired values of θ and then applying (8). All PDFs in the test-cases below were chosen among three different classes of distribution functions, namely: exponential, shifted exponential, and uniform. Various properties of these classes are summarized in Table 1.

Note that the shifted exponential and uniform classes depart drastically from exponentiality; thus, such PDFs would require a dramatically increased state space for an adequate approximation by an ordinary Markovian model[4]. Furthermore, note that a shifted exponential distribution represents the convolution between two r.v.s: one constant and one exponentially distributed. Thus, in all test cases below where such PDFs occur, the respective states could be replaced by a tandem of two 'virtual' states (each with the same traffic generation rate), featuring constant and exponential sojourn times, respectively. The fact that no such 'virtual'

[3] The various traffic rates and mean holding times in figure 1 are expressed in unspecified bandwidth and time units, respectively. Accordingly, the data unit is implicitly defined as one bandwidth unit times one time unit. The 'free variable' θ in the effective bandwidth function is expressed in inverse such data units.

[4] More specifically, each state featuring a non-exponential PDF would have to be 'broken' into a potentially big number of 'virtual' exponential states, appropriately interconnected. This would provide a rational generator approximating the exact generator of the non-exponential distribution.

states have to be introduced is a consequence of the generality of PDFs allowed for the sojourn times of the various states. This generality permits the description of a given model through a Markov Chain of minimal order.

All individual test-cases, to be presented momentarily, are accompanied with a derivation of the corresponding limiting effective bandwidth value, as $\theta \to \infty$, through application of the results in Section 4. For this purpose, we now identify the simple circuits occurring in the example-model. Direct inspection of Figure 1 (or of the form of the matrix P for that matter) verifies that two such circuits (to be pedantic: two classes of circuits, equivalent upon rotation) arise, namely,

$$\begin{aligned} \boldsymbol{\iota}_1 &= (1,3) \quad &\text{with} \quad & \boldsymbol{v}_1 \equiv \boldsymbol{v}(\boldsymbol{\iota}_1) = (1,3), \\ \boldsymbol{\iota}_2 &= (1,2,3) \quad &\text{with} \quad & \boldsymbol{v}_2 \equiv \boldsymbol{v}(\boldsymbol{\iota}_2) = (1,3,2). \end{aligned} \tag{17}$$

The function $\boldsymbol{v}(\cdot)$ maps the circuits into permuted vectors of states, featuring increasing rates, as explained in Section 4.

For the first test-case we let the sojourn times spent at all states be exponentially distributed (with parameters matching the given, fixed, mean values τ_j). This arrangement furnishes a usual Markovian fluid. Since the essential supremum of the sojourn time at state 2 is infinite, Theorem 4.2 directly yields $a(\infty) = \hat{r} = 10$ (the first condition is violated). Note that the same result would have been obtained even if T_2 possessed a different PDF, as long as the property $\tau_2^+ = \infty$ was still maintained. This is because the limiting effective bandwidth depends only on the values of the essential suprema and infima of the sojourn times. As another example, $a(\infty)$ would still be equal to 10, independently of *any* assumptions on the PDF of T_2, provided that both τ_1^- and τ_3^- were zero (e.g., by having T_1 be uniformly distributed in $[0,20]$ and T_3 be exponentially distributed), because then the second condition in Theorem 4.2 would be violated. (States 2 and 3 are the states with rate less than the peak rate, which are visited in the only simple circuit, namely $\boldsymbol{\iota}_2$ in (17), that contains a state featuring the peak rate). The full graph of $a(\theta)$ when all T_j are exponential is displayed as the top-most curve in Figure 2. Note how the curve approaches the limiting value 10, as θ increases.

The rest of the numerical examples are chosen so as to demonstrate cases where $a(\infty) < \hat{r}$. To this end, in all subsequent instances the PDF for T_2 remains uniform, in the interval $[0,2]$ (so that $\tau_2 = 1$); thus, $\tau_2^+ = 2 < \infty$ (see Table 1). Furthermore, at least one of τ_1^- and τ_3^- is always kept strictly positive.

Firstly, consider a configuration where $\tau_1^- = 5$, $\tau_3^- = 0$ and $\tau_3^+ = \infty$. Referring to the source model in Figure 1, this choice models the fact that silence is always forced to last at least half its average duration, while the duration of the low activity period is unrestricted. (The choice for the PDF of T_2 in the previous paragraph has also imposed an upper bound for the duration of the high activity state.) Two variants conforming to this configuration will be presented. In both variants, T_3 has an exponential distribution. However, in the first variant T_1 has a shifted exponential distribution with parameters $s_1 = s_2 = 5$, resulting in $\tau_1^+ = +\infty$, while in the second variant T_1 is distributed uniformly in the interval $[5,15]$ and now $\tau^+ = 15$.

To determine $a(\infty)$ we must examine all $\boldsymbol{v} \in \mathcal{V}_f$, as explained in Section 4. Since $3 \in \mathcal{S}_{\text{inf}}$, $\mathcal{V}_f$ contains only a single vector, namely $\boldsymbol{v}_2$ (see (17)). Thus $\bar{r}^*(\boldsymbol{v}_2) = \bar{r}^* = a(\infty)$. Furthermore, an application of (12) yields $l(\boldsymbol{v}_2) = 3 = \ell(\boldsymbol{v}_2)$, for both variants, again because $3 \in \mathcal{S}_{\text{inf}}$. (Recall that $l(\boldsymbol{v})$ indexes *elements of* $\boldsymbol{v}$, not states. In the case at

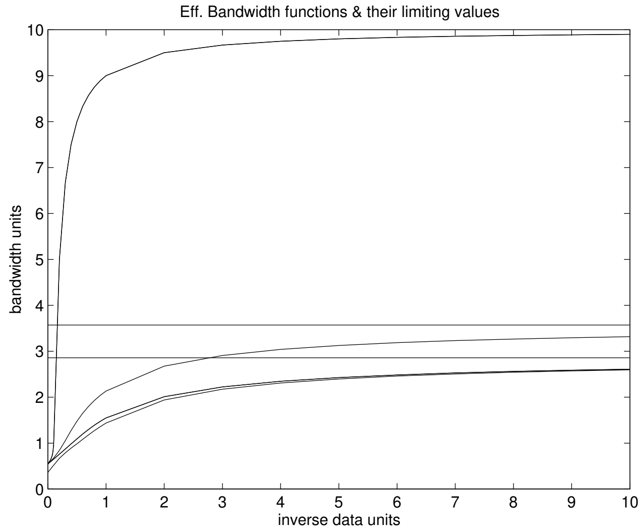

Figure 2: Effective bandwidth functions for various choices of the sojourn times

hand, $v_{2,l(\boldsymbol{v}_2)} = v_{2,3} = 2$.) Thus, we immediately obtain $\bar{r}_3(\boldsymbol{v}_2) = \bar{r}^*(\boldsymbol{v}_2)$ (see Lemma 4.3). Direct computation using (13) yields

$$\begin{aligned} \bar{r}^* = \bar{r}_3(\boldsymbol{v}_2) &= \frac{\tau_{v_{2,1}}^- r_{v_{2,1}} + \tau_{v_{2,2}}^- r_{v_{2,2}} + \tau_{v_{2,3}}^+ r_{v_{2,3}}}{\tau_{v_{2,1}}^- + \tau_{v_{2,2}}^- + \tau_{v_{2,3}}^+} \\ &= \frac{\tau_1^- r_1 + \tau_3^- r_3 + \tau_2^+ r_2}{\tau_1^- + \tau_3^- + \tau_2^+} \\ &= \frac{5 \times 0 + 0 \times 1 + 2 \times 10}{5 + 0 + 2} \approx 2.8571. \end{aligned}$$

Since $r^* = r_3 = 1$, we obtain (see Theorem 4.1) $a(\infty) = \bar{r}^* \approx 2.8571$. Observe how smaller this value is, compared with the peak rate $\hat{r} = 10$. Furthermore, note that only the infimum of the silence periods matters; the limiting eff. bandwidth value isn't affected when very large silences are introduced. (This is the major 'qualitative' difference between the two variants.) Thus, a potential shaping scheme should focus into enforcing only lower bounds for silences. The full graphs for $a(\theta)$, for both variants, appear as the bottom-most group of two curves in Figure 2, along with a straight line marking the limiting value. Note that the two variants share not only the same limiting value, but also a quite similar evolution towards this limiting value.

The final example is somewhat of a reciprocal to the previous case. The silence periods are allowed to be arbitrarily short ($\tau_1^- = 0$) but the low activity period is restricted so as to last at least half its expected duration ($\tau_3^- = 5$). Specifically, now T_1 is exponentially distributed and T_3 has a shifted exponential distribution (thus, $\tau_3^+ = \infty$). The high activity period is still uniformly distributed, as previously. The derivation of the limiting effective bandwidth matches exactly the procedures in the previous case, except for the difference in the values of τ_1^- and τ_3^-. Applying the new values yields $a(\infty) = 25/7 \approx 3.5714$, still about one third of the peak rate. The outcome is that regulation to enforce adequate silence periods may be replaced with the enforcement of minimal periods of low activity (while always restricting the maximal duration during high activity, of course), albeit with a milder effect. The full graph of $a(\theta)$ for this example appears as the middle curve in Figure 2, along with a straight line marking the corresponding limiting value.

Comparing retrospectively the last two examples with the plain Markovian counterpart presented first, it becomes apparent that the presence of non-exponential sojourn times, may lead to markedly different properties of the effective bandwidth function. In these settings, plain Markovian models may severely over-emphasize (or, in other cases, not exhibited here, under-emphasize) the effective bandwidth requirements. This is clearly demonstrated in Figure 2.

Finally, another comment is in order, in connection with settings where the sojourn times of some states are bounded away from zero and/or infinity, as is typical in the presence of traffic shaping or throttle control. In such cases, the limiting effective bandwidth value $a(\infty)$ may be bounded away from the source's peak rate. The results of Section 4 enable the computation of this limiting value, by means of elementary calculations, involving only the extremal values of the sojourn times, which are directly available from the parameters of the shaping or other control function. Thus, calculation of $a(\infty)$ provides the means for assessing the order of magnitude of the effectiveness of a given control scheme, with a minimum of assumptions on the sources' details.

6 Conclusions

In this paper we studied a class of fluid source models with multiple data generation rates and non-Markovian behavior, in the sense that the random times with which the data rates are sustained are allowed to be arbitrarily distributed. This class of models generalizes the usual Markovian models and allows a wider coverage of traffic profiles, with reduced state-space requirements. The paper established that the models under investigation possess a well defined effective bandwidth function. This function was calculated, as the solution of an implicit function problem arising from the requirement that the spectral radius of an appropriate non-negative matrix be equal to unity.

The models investigated in the paper presented a more variable behavior than the usual Markovian fluids. In particular, their effective bandwidth function was shown to either be a strictly increasing function ('regular' behavior) or constant and equal to the source's mean rate. Sources of the second kind, which were completely characterized, extend the notion of 'CBR' traffic (although they present multiple data generation rates) and they were shown to be able to achieve loss-less performance, given a finite amount of buffering and allocated bandwidth equal to their mean rate.

The qualitative investigation behind the mechanics of the curious 'CBR' property was deepened, and it was formally found that the limiting effective bandwidth value, towards the regime of unboundedly stringent QoS, need not be equal to the source's peak rate, but may remain smaller. An easily applicable method for determining the limiting effective bandwidth in general was formally established and necessary and sufficient conditions for this value to be smaller than the peak rate were given. Since the limiting value is an upper bound to the effective bandwidth requirements for all levels of QoS, comparison of the limiting value to the peak rate provides a clear indication of the order of magnitude of the bandwidth savings that may be achieved when the source's behavior is appropriately controlled. Consequently, the results on the limiting effective bandwidth may be of value either to the designer of a terminal, or to settings were the source's activity is controllable 'on-line', even when there is no interest for loss-less performance,

A Appendix

A.1 Proof of Lemma 2.1

The function $g(x)$ is continuous by construction. Furthermore, since $\lim_{x\to 0} g(x)$ exists, it follows $\lim_{x\to 0} f(x) = 0$, and by the continuity of f, $f(0) = 0$. To prove the monotonicity, let '$\overset{s}{\le}$' stand for '$<$', if $f(x)$ is strictly convex, and for '$\le$' otherwise. Assume for the moment that $0 < \sup \mathcal{I}$. Pick $x_1, x_2 \in \mathcal{I}$, such that $0 < x_1 < x_2$; then $h := x_1/x_2 < 1$. By convexity,

$$f(x_1) = f\big((1-h)0+hx_2\big) \overset{s}{\le} (1-h)f(0)+hf(x_2) = \frac{x_1}{x_2} f(x_2),$$

leading directly to $g(x_1) \overset{s}{\le} g(x_2)$. A continuity argument yields $g(0) \overset{s}{\le} g(x)$, for all $x > 0$. The same arguments (with a couple of changes in the direction of appropriate inequalities) may be repeated for negative values of x, if applicable.

A.2 Proof of Lemma 2.2

If f is strictly convex on $\mathcal{I}$ there is nothing to prove. Therefore, suppose that there exist $x_0, x_1 \in \mathcal{I}$, with $x_0 \ne x_1$ and an $h \in (0,1)$, such that

$$f((1-h)x_0 + hx_1) = (1-h)f(x_0) + hf(x_1). \tag{18}$$

Let $(f(x_1) - f(x_0))/(x_1 - x_0) = r$ and define $g(x) \overset{\text{def}}{=} f(x) - f(x_0) - r(x - x_0)$. All we need is to prove that $g(x) = 0$, for all $x \in \mathcal{I}$. In fact, it suffices to prove it for all x in the interior of $\mathcal{I}$, since if this holds, then by continuity, g will be also nullified at any endpoint of $\mathcal{I}$.

The function g is convex and analytic on $\mathcal{I}$, since f is. Denote $\bar{x} = (1-h)x_0 + hx_1$. Due to (18) and to the definitions of g and r, we have $g(x_0) = g(\bar{x}) = g(x_1) = 0$. Thus, by Rolle's theorem there exist $x_- \in (x_0, \bar{x})$ and $x_+ \in (\bar{x}, x_1)$, such that $g'(x_-) = g'(x_+) = 0$. Since g' is increasing (due to the convexity of g), $g'(x) = 0$, $\forall x \in [x_-, x_+]$. Consequently, $g(x) = g(\bar{x}) = 0$, for all $x \in [x_-, x_+]$.

Let $\mathcal{S} = \big\{ s \in \mathcal{I} \bigm| g(x) = 0 \ \forall x \in [s, x_+] \big\}$. Obviously, $[x_-, x_+] \subseteq \mathcal{S} \subset \mathcal{I}$. Denote $s^* = \inf \mathcal{S}$ and suppose $s^* > \inf \mathcal{I}$. Since g is analytic, $g^{(n)}(s^*) = \lim_{x\to s^*+0} g^{(n)}(x) = 0$, for all $n \ge 0$. Furthermore, for some suitably small $\epsilon > 0$ and any $h \in [0,1]$,

$$g(s^* - h\epsilon) = \sum_{n=0}^{\infty} \frac{(-h\epsilon)^n}{n!} g^{(n)}(s^*) = 0,$$

showing that $s^* - \epsilon \in \mathcal{S}$, which contradicts the definition of s^*. We conclude that $\inf \mathcal{S} = \inf \mathcal{I}$. With a similar argument, one may show that $\sup\big\{ s \in \mathcal{I} \bigm| g(x) = 0 \ \forall x \in [x_-, s] \big\} = \sup \mathcal{I}$, completing the proof.

A.3 Proof of Lemma 4.2

Pick a $\iota \in \mathcal{C}$, let $\ell(\iota) = n$ and denote $\boldsymbol{\iota}' = (\iota_1, \ldots, \iota_n, \iota_1)$. Further, let $\mathcal{I}(i, j; l)$ be the set of all sample paths from state i to state j, in l steps. Obviously, $\boldsymbol{\iota}' \in \mathcal{I}(\iota_1, \iota_1; n)$. In this notation,

$$\begin{aligned}
1 &= \phi(\theta, u(\theta))^n = \rho\left(A(\theta, u(\theta))^n\right) \\
&\geq \rho\left(\operatorname{diag}\left\{\left(A(\theta, u(\theta))^n\right)_{j,j}\right\}\right) \geq \left(A(\theta, u(\theta))^n\right)_{\iota_1,\iota_1} \\
&= \sum_{\boldsymbol{k} \in \mathcal{I}(\iota_1,\iota_1;n)} \varpi_{\boldsymbol{k}} \prod_{l=1}^{n} \gamma_{k_l}(\theta r_{k_l} - \theta a(\theta)) \\
&\geq \varpi_{\boldsymbol{\iota}'} \prod_{l=1}^{n} \gamma_{\iota_l}(\theta r_{\iota_l} - \theta a(\theta))
\end{aligned}$$

and, by taking logarithms we are led to $\theta\chi_{\boldsymbol{\iota}}(\theta) \leq -\log \varpi_{\boldsymbol{\iota}'}$. Dividing this last relation by θ and letting $\theta \to \infty$, we conclude that $\limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}}(\theta) \leq 0$.

For the second part we first prove that the assertion: *"there exists a constant $h < 1$ such that: $\forall\theta \geq \theta_0$, and $\forall\boldsymbol{\iota} \in \mathcal{C}$, $\prod_{k=1}^{\ell(\boldsymbol{\iota})} \gamma_{\iota_k}(\theta r_{\iota_k} - \theta a(\theta)) \leq h$"* is false. Suppose it is true; let $n^* = \max_{\boldsymbol{\iota}\in\mathcal{C}} \ell(\boldsymbol{\iota})$ and define

$$M(\theta) = \max_{i,j} \max_{0\leq l\leq n^*-1} \max_{\boldsymbol{k}\in\mathcal{I}(i,j;l)} \prod_{\nu=1}^{l} \gamma_{k_\nu}(\theta r_{k_\nu} - \theta a(\theta)).$$

where for $l = 0$ the inner maximum is conventionally assumed equal to 1. Then, for any $n = mn^* + l$ with $0 \leq l \leq n^* - 1$ we obtain

$$\begin{aligned}
\left(A(\theta, u(\theta))^n\right)_{i,j} &= \sum_{\boldsymbol{\iota}\in\mathcal{I}(i,j;n)} \varpi_{\boldsymbol{\iota}} \prod_{k=1}^{n} \gamma_{\iota_k}(\theta r_{\iota_k} - \theta a(\theta)) \\
&\leq h^m M(\theta) \sum_{\boldsymbol{\iota}\in\mathcal{I}(i,j;n)} \varpi_{\boldsymbol{\iota}} \\
&= h^m M(\theta)\left(P^n\right)_{i,j}, \quad \forall i, j, \quad \forall\theta \geq \theta_0,
\end{aligned}$$

because any path of length mn^*+l contains at least m simple circuits and any remaining part has length less than n^*. By compiling the previous inequalities to matrix form and taking spectral radii, we obtain $1 \leq M(\theta)h^m$. By letting $m \to \infty$ we reach a contradiction, because $h < 1$. We conclude that the assertion is false.

Given this result, there must exist a $\boldsymbol{\iota} \in \mathcal{C}$, such that $\limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}}(\theta) = 0$. For if not, all limit suprema are negative and $-b \equiv \max_{\boldsymbol{\iota}\in\mathcal{C}} \limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}}(\theta) < 0$. Then, for an appropriate θ_0, for all $\theta \geq \theta_0$ we get $\chi_{\boldsymbol{\iota}}(\theta) < -b/2$, for all $\boldsymbol{\iota} \in \mathcal{C}$, and the assertion holds true with $h = e^{-\theta_0 b/2}$. Thus, by letting $\mathcal{C}_0 \stackrel{\text{def}}{=} \{\, \boldsymbol{\iota} \in \mathcal{C} \mid \limsup_{\theta\to\infty} \chi_{\boldsymbol{\iota}}(\theta) = 0 \,\}$, we conclude that $\mathcal{C}_0 \neq \emptyset$.

By the previous reasoning, only $\boldsymbol{\iota} \in \mathcal{C}_0$ may violate the assertion and there must be at least one of them violating it, since the assertion is false. Denote such a circuit $\boldsymbol{\iota}^*$. Then, necessarily, $a(\infty) \leq r_{v_{\ell(\boldsymbol{\iota}^*)}(\boldsymbol{\iota}^*)}$. For, if not, $a(\infty)$ is greater than all $r_{\iota_k^*}$ and, since both $a(\theta)$ and all the generators are increasing functions, we obtain that for a suitable θ_0 and a suitable $\epsilon > 0$ there holds $\prod_{k=1}^{\ell(\boldsymbol{\iota}^*)} \gamma_{\iota_k^*}(\theta r_{\iota_k^*} - \theta a(\theta)) \leq \prod_{k=1}^{\ell(\boldsymbol{\iota}^*)} \gamma_{\iota_k^*}(\theta_0 r_{\iota_k^*} - \theta_0(r_{v_{\ell(\boldsymbol{\iota}^*)}(\boldsymbol{\iota}^*)} + \epsilon)) \equiv h < 1$, for all $\theta \geq \theta_0$, and the assertion holds for $\boldsymbol{\iota}^*$. Thus, by contradiction, we obtain $a(\infty) \leq r_{v_{\ell(\boldsymbol{\iota}^*)}(\boldsymbol{\iota}^*)}$.

References

[1] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.

[2] C.-S. Chang. "Stability, queue length, and delay of deterministic and stochastic queueing networks". *IEEE Trans. Automat. Contr.*, 39(5):913–931, May 1994.

[3] C.-S. Chang and J. A. Thomas. "Effective bandwidth in high-speed digital networks". *IEEE JSAC*, 13(6):1091–1100, 1995.

[4] G. de Veciana, G. Kesidis, and J. Walrand. "Resource management in wide-area ATM networks using effective bandwidths". *IEEE JSAC*, 13(6):1081–1090, 1995.

[5] N. G. Duffield, J. T. Lewis, N. O'Connell, R. Russell, and F. Toomey. "Entropy of ATM traffic streams: A tool for estimating QoS parameters". *IEEE JSAC*, 13(6):981–990, 1995.

[6] A. I. Elwalid and D. Mitra. "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks". *IEEE/ACM Trans. Networking*, 1(3):329–343, June 1993.

[7] S. Friedland and S. Karlin. "Some inequalities for the spectral radius of non-negative matrices and applications". *Duke Mathematical Journal*, 42(3):459–490, 1975.

[8] R. J. Gibbens and P. J. Hunt. "Effective bandwidths for the multi-type UAS channel". *Queueing Sys.*, 9:17–28, 1991.

[9] R. Guérin, H. Ahmadi, and M. Naghshineh. "Equivalent capacity and its application to bandwidth allocation in high-speed networks". *IEEE JSAC*, 9(7):968–981, September 1991.

[10] F. P. Kelly. "Effective bandwidths at multi-class queues". *Queueing Sys.*, 9:5–15, 1991.

[11] G. Kesidis, J. Walrand, and C.-S. Chang. "Effective bandwidths for multiclass Markov fluids and other ATM sources". *IEEE/ACM Trans. Networking*, 1(4):424–428, August 1993.

[12] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.