Faculty of Mathematical Sciences

University of Twente
The Netherlands

P.O. Box 217
7500 AE Enschede
The Netherlands
Phone: +31-53-4893400
Fax: +31-53-4893114
Email: memo@math.utwente.nl
www.math.utwente.nl/publications

Memorandum No. 1618

Elastic calls in an integrated services network: the greater the call size variability the better the QoS

R. Litjens[1] and R.J. Boucherie

March 2002

ISSN 0169-2690

[1]Expertise Group QoS Control, KPN Research, The Netherlands

Elastic Calls in an Integrated Services Network: the Greater the Call Size Variability the Better the QoS

Remco Litjens
Expertise Group QoS Control, KPN Research, The Netherlands
Tel: +31 70 446 3419, Fax: +31 70 446 3477, E-mail: R.Litjens@kpn.com

Richard J. Boucherie*
Faculty of Mathematical Sciences, University of Twente, The Netherlands
Tel: +31 53 489 3432, Fax: +31 53 489 3069, E-mail: R.J.Boucherie@math.utwente.nl.

Abstract

We study a telecommunications network integrating prioritized stream calls and delay tolerant elastic calls that are served with the remaining (varying) service capacity according to a processor sharing discipline. The remarkable observation is presented and analytically supported that the expected elastic call holding time is decreasing in the variability of the elastic call size distribution. As a consequence, network planning guidelines or admission control schemes that are developed based on deterministic or lightly variable elastic call sizes are likely to be conservative and inefficient, given the commonly acknowledged property of e.g. WWW documents to be heavy tailed. Application areas of the model and results include fixed IP or ATM networks and mobile cellular GSM/GPRS and UMTS networks.

Keywords: Integrated services networks, performance analysis, Quality-of-Service, varying service capacity, heavy tail distributions, processor sharing models.
AMS Subject classifications. Primary: 90B18, 90B22. Secondary: 60K25.

1 Introduction

As the offered telecommunications services developed from voice telephony only to a much wider variety of services including electronic mail, WWW browsing and video-conferencing, separate networks were typically installed alongside the traditional public switched telephone network, for reasons of operational simplicity. In recent years, the operators' need for flexibility and efficiency has shifted the focus towards integrated services networks, i.e. single networks supporting a broad range of services on a selected set of prespecified bearers. As an essential input for proper network planning and traffic management, analytical modelling and performance evaluation have evolved in parallel from simple circuit-oriented models, such as the classic Erlang loss system,

*The research of Richard J. Boucherie is partly supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs, The Netherlands.

towards more complicated models, integrating multiple circuit-oriented services or, very recently, integrating circuit- and packet-oriented services.

The present study falls into the latter category: we consider a telecommunications system integrating a circuit-oriented *stream* service, requesting a fixed capacity assignment (e.g. voice telephony), with a packet-oriented *elastic* service that is delay-tolerant (e.g. WWW browsing). As the stream calls are assumed to have strict and preemptive priority over elastic calls, a *varying amount of capacity* is available to serve the elastic calls. At any time, the available capacity is equally distributed over the present elastic calls according to a *processor sharing* (PS) service discipline. The performance measures of interest are the Grade-Of-Service (GOS: e.g. stream call blocking probability) and the Quality-Of-Service (QOS: e.g. elastic call holding times). The generic model can be applied to evaluate link sharing policies on a transmission link in a fixed IP or ATM network (e.g. [15]), in a single cell in a mobile cellular network integrating GSM for voice (stream) calls and GPRS for data (elastic) calls (e.g. [8]), or in a multi-service UMTS network.

Analytical studies based on Markovian models (e.g. [8, 15]) assume exponentially distributed elastic call sizes for reasons of mathematical tractability. Note that the exponential distribution is rather lightly tailed with a coefficient of variation equal to one. In light of the commonly acknowledged property of e.g. World Wide Web (WWW) pages to be *heavy tailed* (see e.g. [7, 16]), we have performed a series of dynamic simulations in order to investigate the impact of the elastic call size distribution on the experienced QOS, observing that *a greater elastic call size variability improves the* QOS. The core explanation for this phenomenon is that when increasing the variability of the elastic call size distribution, the number of small elastic calls that enters the system increases while large elastic calls become larger yet increasingly rare. The fraction of time the system is in a relatively favorable state of light load is then proportional to the elastic call size variability, so that the expected holding time of an elastic call decreases as the elastic call size variability becomes larger.

In order to illustrate the significance of the presented phenomenon, we note that in a (fixed capacity) *first-in first-out* (FIFO) queue the effect is known to be reversed: the QOS *degrades* under a greater call size variability (depending only on its first and second moments), while a similar trend is argued and demonstrated to hold for a FIFO queue with varying capacity. In contrast, the elastic call holding times in a fixed capacity PS queue are known to be *insensitive* to the call size variability. The contribution of our paper is to demonstrate and analytically support the phenomenon that in the practically most relevant PS queue with varying capacity, the QOS *improves* under a greater call size variability. The presented numerical results indicate however that it is not trivial which elastic call size variability measure captures the essence of its impact on the QOS. Within a family of probability distributions, both the second moment and the tail heaviness are observed to suffice, while these variability measures turn out to be less useful in a comparison across distinct families of distributions.

The outline of the paper is as follows. Section 2 reviews the relevant literature. The considered integrated services model is described in Section 3, featuring a PS service discipline for admitted elastic calls. Section 4 presents the observation that the expected elastic call holding time is decreasing in the variability of the elastic call size distribution. Subsequently, Section 5 provides

theoretical support and intuition based on an analysis of extreme cases. As is demonstrated in Section 6, the QOS improvement with greater call size variability is specifically due to the PS nature of the service process. In particular, in Section 6 we explicitly investigate the trade-off between the FIFO and PS service disciplines and study the impact of the elastic call size variability on the QOS in an extended model that queues rather than rejects elastic calls that cannot enter service immediately upon arrival. Section 7 ends this paper with some concluding remarks.

2 Literature

In the literature some relevant and/or related results are available regarding the principal aspects of the presented study. For standard $M/G/1/PS$ *processor sharing* queueing systems with *fixed* service capacity, it is a well-known fairness result that the conditional expected holding time of a call with a given service requirement x, is equal to $x/(1-\rho)$, where $\rho \equiv \lambda \mathbf{E}\{x\}$ denotes the offered load and λ is the Poisson call arrival rate (see e.g. [18]). As the conditional expected holding time depends on the service requirement distribution only through its first moment, also the expected holding time is insensitive to the shape of the service requirement distribution.

Performance studies of systems with *varying service capacity* have recently been published. Focussing on the performance of elastic calls, such models fall within the class of queues in a random environment (see e.g. [12, Chapter 6]). In [15] an analytical comparison of segregated and integrated policies is presented for link sharing in an integrated services IP or ATM network. In [8] the model proposed in [15] is applied to an integrated GSM/GPRS network, and extended to include the option of queueing elastic calls that cannot start transfer immediately upon admission. In both papers analytical expressions are derived for GOS and QOS performance measures, including channel utilization, stream and elastic call blocking probabilities and (conditional) expected elastic call holding times. In [2] a single cell in a GSM network is studied serving voice (stream), video and data (both elastic) calls, evaluating a proposed capacity sharing policy in terms of channel utilization and blocking probabilities. A similar model is studied in [9], further extending the analysis of [8] to include the (conditional) QOS analysis of elastic video calls. A principal characteristic that is common to these papers is the Markovian model that forms the basis for the analyses, in particular assuming exponentially distributed elastic call sizes.

Regarding the analysis of queueing models with *heavy tailed* service requirement distributions, it was shown in [5] that in the $GI/G/1/FIFO$ queue the tail of the holding time distribution is 'one degree' heavier than that of the service requirement distribution. In [20] it is proved that both tails are equally heavy in an $M/G/1/PS$ queue. The generally considered desirable property of this tail equivalence is extended in [13] (Theorem 5.3.1) to an on/off server model (varying capacity) under certain conditions.

3 Model

Consider a telecommunications system where C traffic channels are shared between *stream* and *elastic* calls, with preemptive priority for stream calls (see Figure 1). Both call types arrive

according to two mutually independent Poisson processes, with arrival intensities λ_s and λ_e calls per second, respectively.

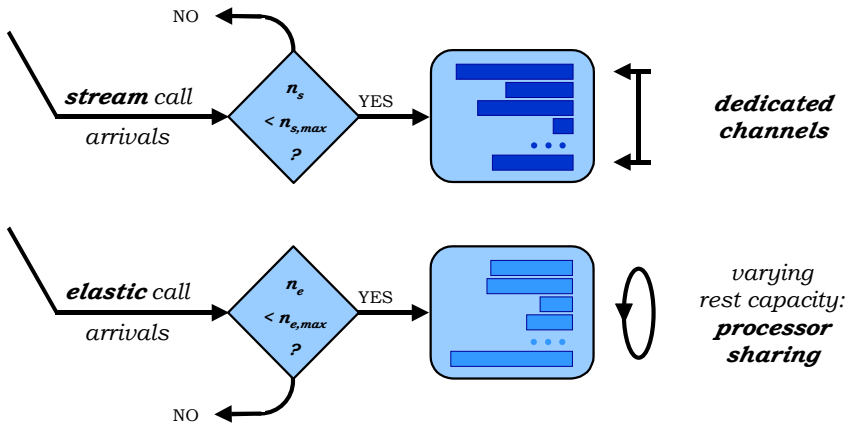


Figure 1: *The model.*

The holding time of a *stream call* is assumed to have Probability Density Function (PDF) φ_s with mean μ_s^{-1}, and the stream call traffic load is denoted $\rho_s \equiv \lambda_s/\mu_s$ (in Erlang). An admitted stream call is served with a fixed assignment of one traffic channel. The call handling scheme serves stream calls with preemptive priority, so that an arriving stream call is blocked if and only if there are already $n_{s,\max} \equiv C$ stream calls present, in which case the call is cleared from the system. Denote with $N_s(t)$ the number of stream calls in the system at time t, with state space denoted $\mathbb{S} \equiv \{0, 1, \cdots, C\}$. Define $\mathbb{S}_+ \equiv \mathbb{S} \backslash \{C\}$ for later convenience.

The *nominal* holding time (or: *size*) of an *elastic call* is defined as the call holding time under a fixed assignment of a single traffic channel, and is assumed to have PDF φ_e with mean μ_e^{-1}. The elastic call traffic load is denoted $\rho_e \equiv \lambda_e/\mu_e$ (in Erlang). The defining characteristic of an elastic call is that it is delay tolerant and can handle a variable channel assignment which may lie anywhere between 0 and C. As a result, the actual holding time and the nominal holding time of an elastic call may differ. Admission of elastic calls is governed by a predetermined maximum number of elastic calls that is allowed in the system, denoted $n_{e,\max}$, and blocked elastic calls are cleared from the system. At any time t, the channel sharing policy distributes the $C - N_s(t)$ available channels fairly over the present elastic calls according to the typically used PS service discipline, so that each elastic call is assigned $(C - N_s(t))/N_e(t)$ channels, with $N_e(t)$ the number of elastic calls present at time t. Not only is PS an attractive service discipline due to its inherent fairness property, but it also implicitly models the (idealized) effects of TCP (Transmission Control Protocol) flow control in the sense of (instantaneous) adjustment of the transfer rate in accordance with the traffic congestion level (see also [11, 15, 17]). Section 6 considers an adjustment of the current model where elastic calls that cannot enter the PS queue immediately upon arrival are queued rather than blocked.

For stream calls the blocking probability $\mathbf{P}_s$ is the only relevant *performance measure,* which is readily determined using the classical $M/G/C/C$ Erlang loss model. For elastic calls, however, we are primarily interested in the experienced QOS in terms of the expected call holding time $\mathbf{T}_e$ and the conditional expected call holding time $\mathbf{T}_e(x)$ of an elastic call of given size x. More specifically,

as the document sizes of e.g. WWW calls are commonly acknowledged to be heavy tailed, we seek to determine the impact of the call size variability on these QOS measures. Although for finite $n_{e,\max}$ the elastic call blocking probability $\mathbf{P}_e$ is another essential performance measure that is considered, the principal result is most transparently conveyed for a model with $n_{e,\max} = \infty$, in order to prevent any ambiguity in the overall performance evaluation. This issue is readdressed in the next section.

Knowing that in a PS system with *fixed* capacity, the (conditional) expected holding time of an elastic call is independent of the call size distribution, we are also interested in the impact of the degree of variability in the service capacity on the QOS of elastic calls. To this end, the elastic service performance in the integrated services model is compared with that in a fixed capacity $M/G/1/n_{e,\max}/PS$ queueing model. Since this model is defined for a single channel, we must scale the elastic service requirements appropriately, resulting in $\mu_e^\star \equiv \mu_e C^\star$ and $\rho_e^\star \equiv \lambda_e/\mu_e^\star = \rho_e/C^\star$, with fixed channel capacity $C^\star \equiv C - \rho_s(1 - \mathbf{P}_s)$, i.e. the average number of available channels in the integrated services model. The conditional expected holding time of an elastic call in the fixed capacity model is given by (see [6])

$$\mathbf{T}_e(x) = \frac{x \sum_{n_e=0}^{n_{e,\max}-1} (\rho_e^\star)^{n_e} (n_e + 1)}{C^\star \sum_{n_e=0}^{n_{e,\max}-1} (\rho_e^\star)^{n_e}}, \tag{1}$$

while the expected elastic call holding time $\mathbf{T}_e$ trivially follows given that μ_e^{-1} is the expected elastic call size. Here $\mathbf{T}_e(x)$ and $\mathbf{T}_e$ are insensitive to the elastic call size distribution. For the specific case of $n_{e,\max} = \infty$, i.e. the standard $M/G/1/PS$ queueing model, we obtain $\mathbf{T}_e(x) = x/(C^\star - \rho_e)$ and $\mathbf{T}_e = 1/(\mu_e(C^\star - \rho_e))$ (see [18]), requiring $\rho_e < C^\star$ for stability. As will be demonstrated, the $M/G/1/n_{e,\max}/PS$ model also serves as a limit case and lower bound for the (conditional) expected elastic call holding time in the integrated services model under extremely high elastic traffic load, regardless of the degree of variability in the service capacity.

4 Observations

An extensive simulation study has been carried out in order to investigate the impact of the elastic call size distribution tail and the degree of variability in the stream call arrival and termination process, on the identified QOS measures.

Regarding the choice of the *stream* call holding time distribution (φ_s) a variety of options has been considered, including the exponential distribution and a bimodal mixture of lognormal distributions (see [4] for empirical results on the distribution of voice calls in (cellular) communication networks). While the same qualitative results are demonstrated by all options for the considered φ_s, the presented graphs depict simulation results based on the exponentiality assumption, as it enables us to generate part of the results analytically (see [8, 14, 15]), and to provide analytical support in the next section. The included numerical results are however representative for the general trends.

Regarding the choice of the *elastic* call size distribution (φ_e), the simulation study comprises of a variety of options as well, including the Weibull and Pareto distributions, which have been selected to demonstrate the principal results. The specifications of the considered distributions are summarized in Table 1 below, where $\Gamma(\cdot)$ denotes the gamma function. Note that the degenerate PDF with $\varphi_e(x) = \delta\left(x - \mu_e^{-1}\right)$ is a special case of both the Weibull ($\alpha \to \infty$, $\beta = \mu_e^{-1}$) and Pareto ($\alpha \to \infty$, $c = \mu_e^{-1}$) PDFs, while the exponential PDF with $\varphi_e(x) = \mu_e e^{-x\mu_e}$ is a special case of the Weibull PDF ($\alpha = 1$ and $\beta = \mu_e^{-1}$).

Weibull(α, β)**,** $\alpha, \beta > 0$			**Pareto**(α, c)**,** $\alpha, c > 0$		
$\varphi_e(x)$	$=$	$\alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^\alpha}$, $x > 0$	$\varphi_e(x)$	$=$	$\alpha c^\alpha x^{-(\alpha+1)}$, $x > c$
μ_e^{-1}	$=$	$\frac{\beta}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$	μ_e^{-1}	$=$	$\begin{cases} \frac{\alpha c}{\alpha-1} & \text{if } \alpha > 1 \\ \infty & \text{if } \alpha \leq 1 \end{cases}$
η_e	$=$	$\sqrt{2\alpha\frac{\Gamma(2\alpha^{-1})}{\Gamma^2(\alpha^{-1})} - 1}$	η_e	$=$	$\begin{cases} \frac{1}{\sqrt{\alpha(\alpha-2)}} & \text{if } \alpha > 2 \\ \infty & \text{if } \alpha \leq 2 \end{cases}$

Table 1: *Specifications of the considered elastic call size distributions.*

As a characterization of the variability of a given elastic call size distribution, both the coefficient of variation η_e and the heaviness of the distribution tail, captured by shape parameter α, are considered. If it exists, η_e is an inversely proportional function of α, which is in correspondence with the property that for $x_0 > \beta$, $\Pr(x > x_0) = \exp\left(-(x_0/\beta)^\alpha\right)$ (Weibull) and for $x_0 > c$, $\Pr(x > x_0) = (c/x_0)^\alpha$ (Pareto) are decreasing in α, i.e. the tail becomes heavier as α is smaller. Note that the Pareto tail is generally heavier than the Weibull tail, regardless of the parameter choices, in the sense that

$$\lim_{x\longrightarrow\infty} \frac{\int_x^\infty \varphi_e^{weibull}(y)dy}{\int_x^\infty \varphi_e^{pareto}(y)dy} = \lim_{x\longrightarrow\infty} \frac{\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha_w}\right)}{\left(\frac{c}{x}\right)^{\alpha_p}} = 0,$$

for all α_w, β, α_p, $c > 0$, where the α-parameters of both distributions have been given an identifying subscript to express that they are generally different. Remark 1 at the end of the section comments on the applicability of these variability measures on the QOS comparison under different elastic call size distributions.

The Weibull distribution is particularly useful for our purposes as it enables a straightforward simulation study for a variety of η_e. In contrast, the Pareto distribution induces rather tedious simulations in order to obtain sufficient statistical accuracy. Still, the Pareto distribution is included in our study as it is probably the best-known distribution that satisfies all existing definitions of being truly *heavy tailed*. In this light we refer to [19], where both the class of subexponential distributions (e.g. Weibull (for $\alpha \in (0, 1)$), lognormal and Pareto) and the class of regularly varying distributions (generalizations of the Pareto distribution) are identified as heavy tailed distributions. We stress that for our purposes it is not essential to consider heavy tailed distributions per se, but rather to be able *to vary the balance between a small number of extremely large calls and a large number of small calls*, which is captured in the heaviness of the distribution tail, and study its impact on the experienced QOS.

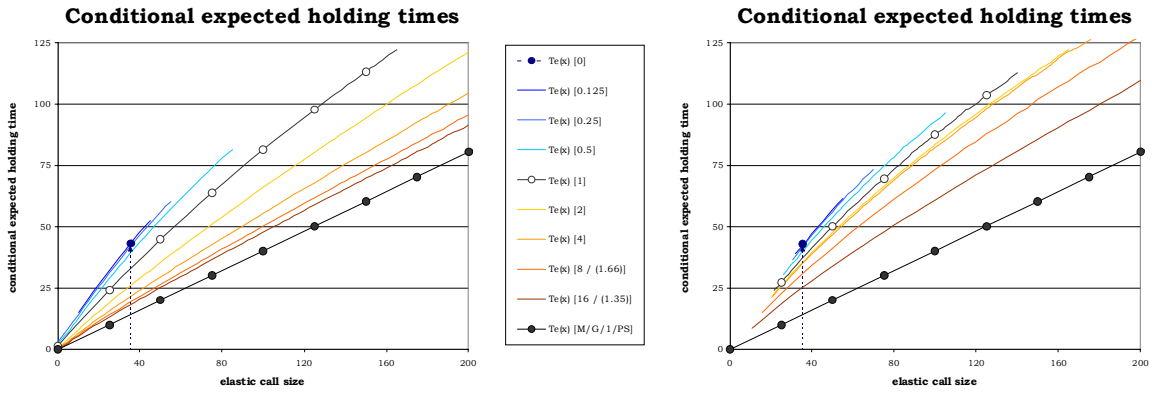
The distribution parameter settings are based on an integrated GSM/GPRS network. Typical values for such a case are $C = 22$, corresponding to 3 GSM frequencies, $\mu_s^{-1} = 50$ seconds and $\rho_s = 13.651$ Erlang, resulting in a stream (voice) call blocking probability of 1%, which is a typical target value for GSM operators. Regarding the elastic (data) service, the transmission rate on a single channel is 9.05 kbps under GPRS channel coding scheme CS-1, so that the average nominal holding time to transfer an e-mail with an average size of 320 kbits, say, is $\mu_e^{-1} = 320/9.05 \approx 35.359$ seconds. The coefficient of variation η_e of the elastic calls is taken from the set $\left\{0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16\right\}$ for the Weibull PDFs, while for the Pareto case we considered $\eta_e \in \left\{0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4\right\}$ to allow cross-PDF comparisons, as well as two cases with infinite η_e ($\alpha_1 = 1.66$, $\alpha_2 = 1.35$, and the c's set to establish the intended average elastic call size) to present even heavier tails. The first three experiments assume $n_{e,\max} = \infty$, to prevent possible distortion of the results due to elastic call blocking, while the fourth experiment assumes a finite $n_{e,\max}$ in order to evaluate the system under an extreme elastic traffic load and to obtain insight into the distortion effect of elastic call blocking on the QOS.

In the figures included below to present the results of the four numerical experiments, each left chart is based on Weibull PDFs while the charts on the right represent Pareto PDFs. The 'o'-marked curves in the Weibull plots correspond to an exponential elastic call size distribution ($\eta_e = 1$), while the '•'-marked curves (all plots) correspond to the $M/G/1/PS$ queueing model with load $\rho_e^\star$, all of which can be calculated analytically. All other curves are obtained from simulations for which 95% confidence intervals have been determined with no worse than 2.5% relative precision. Within each experiment, the marker for $\eta_e = 0$ as well as the $M/G/1/PS$ curves are identical in both the Weibull and Pareto plots. In order to help distinguish the different curves, we note that the included legend follows the order of the actual curves (top-to-bottom).

Experiment 1 Figure 2 shows the conditional expected holding time $\mathbf{T}_e(x)$ of an elastic call as a function of its size x, for elastic traffic load $\rho_e = 6$. For all $\eta_e > 0$, $\mathbf{T}_e(x)$ is displayed for elastic call sizes up to the 99%-percentile of the corresponding distribution, truncated at $x_{\max} = 200$ in both plots in order to enable easy comparison (still capturing 97% (99%) of the most variable Weibull (Pareto) distribution). Whereas the Weibull distribution allows elastic call sizes down to 0 kbits, we note that the Pareto curves start at the corresponding c-values (minimum sample value), which is decreasing in η_e.

Observe from the figures for both distribution types that *(i)* the greater the coefficient of variation of x, the smaller the conditional expected holding times; equivalently, for the Pareto cases with infinite η_e: the smaller α, the heavier the tail and the smaller the conditional expected holding times; *(ii)* the curves all have a concave shape; *(iii)* conditional expected holding times are lowest in a system with fixed capacity. Furthermore, *(iv)* we observe that for a given $\eta_e > 0$ the conditional expected elastic call holding times appear to be lower for the Weibull distribution than for the corresponding Pareto distribution (see also Remark 1 at the end of this section).

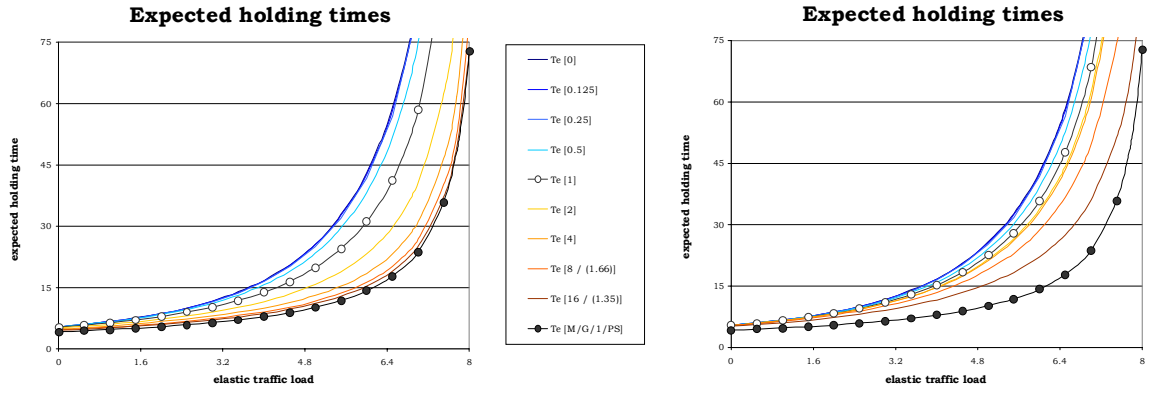
Experiment 2 Figure 3 shows the expected transfer time $\mathbf{T}_e$ of an elastic call as a function of the elastic traffic load ρ_e which is varied between 0 and 8 Erlang. The most important observation



(Weibull) (Pareto)

Figure 2: *Numerical results of experiment 1.*

that can be made from the figure is that *(i)* the greater the coefficient of variation of x, the smaller the expected holding time or equivalently, for the Pareto cases with infinite η_e, the smaller α, the heavier the distribution tail and the smaller the expected holding time. Observe further that *(ii)* for $\rho_e \longrightarrow C^\star \approx 8.486$ Erlang, the expected holding times increase exponentially, whereas for even greater values of ρ_e the system becomes unstable. Note again that *(iii)* the elastic calls achieve optimal QOS in a system with fixed capacity, and that *(iv)* for a given η_e the QOS appears to be better for the Weibull distribution than for the corresponding Pareto distribution (see also Remark 1).

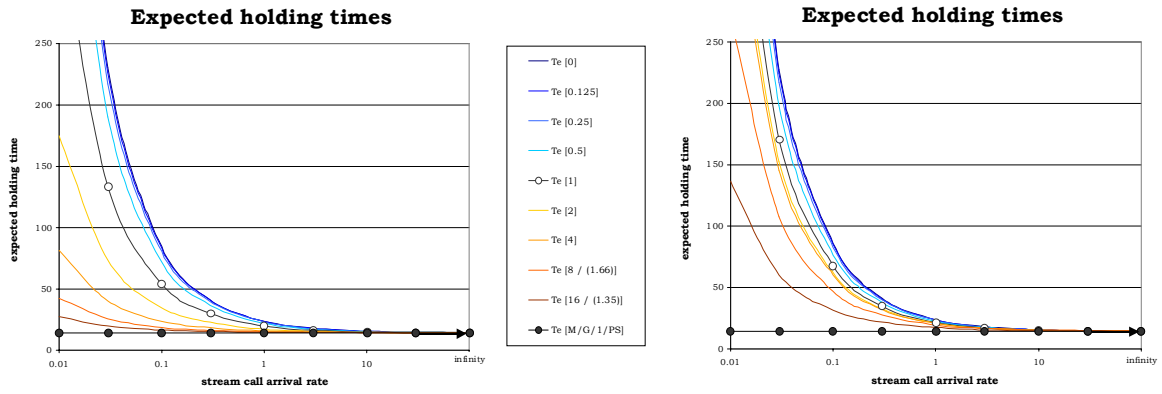


(Weibull) (Pareto)

Figure 3: *Numerical results of experiment 2.*

Experiment 3 In order to investigate the impact of the degree of variability in the service capacity on the QOS of elastic calls, the system has been simulated for the parameter settings of Figure 2, but varying λ_s and μ_s while keeping stream traffic load $\rho_s = 13.651$ constant. The elastic traffic load is set to $\rho_e = 6$. The numerical results in Figure 4 allow a few conclusions. Most importantly, *(i)* the greater the rate of change of the service capacity, the smaller the difference in

performance between the different elastic call size distributions. Intuitively this is not surprising as during the lifetime of an elastic call very rapid variations of the available capacity average out to a growing extent as λ_s and μ_s increase. In fact, the limiting scenario $(\lambda_s, \mu_s \longrightarrow \infty)$ corresponds to a system with fixed capacity $C^\star \approx 8.486$. The expected elastic holding time is then readily determined using standard results for an $M/G/1/PS$ queueing system: $\mathbf{T}_e = 1/\left(\mu_e(C^\star - \rho_e)\right) \approx 14.226$. Furthermore, *(ii)* the figure provides additional support for the claim that the QOS experienced by elastic calls is better as the call sizes are more variable, as well as *(iii)* for the fact that the elastic QOS is best under fixed rather than varying capacity. Lastly, in accordance with the numerical results of experiments 1 and 2, *(iv)* for a given η_e, Weibull PDFs appear to yield better QOS than Pareto PDFs (see also Remark 1).

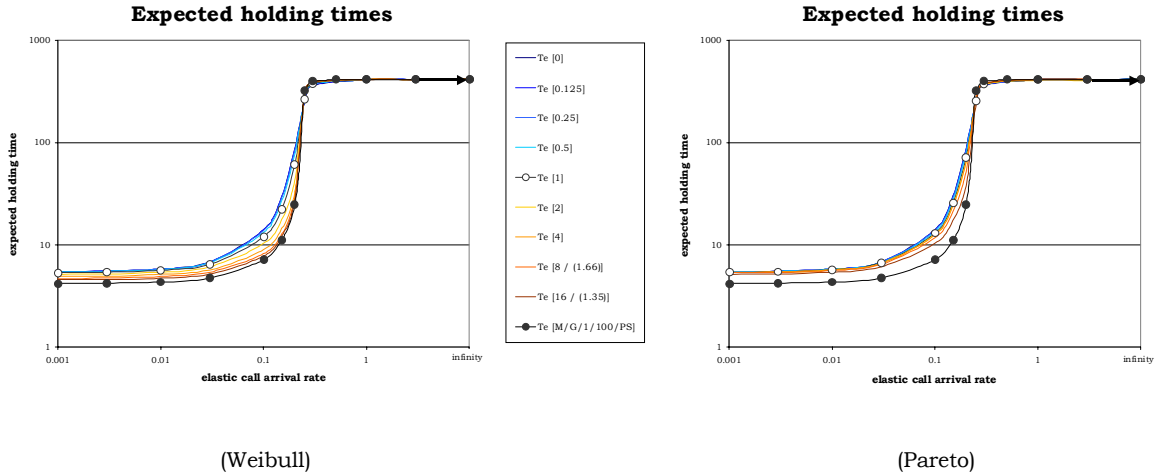


(Weibull) (Pareto)

Figure 4: *Numerical results of experiment 3.*

Experiment 4 As stated earlier, to prevent possible distortion of the results due to elastic call blocking, no elastic call admission control (i.e. $n_{e,\max} = \infty$) was applied in order to obtain the above simulation results. Indeed, a final simulation experiment that has been carried out indicates that in case $n_{e,\max} < \infty$, elastic call blocking probability is lower if the elastic call sizes are more variable, implying a higher *carried* elastic traffic load, which in turn *may* yield higher elastic call holding times through increased competition for resources. The net effect of the elastic call size distribution is unclear in such a case, which is the very reason for avoiding elastic call blocking in the previous experiments. The numerically obtained expected holding times obtained for $n_{e,\max} = 100$ are plotted in Figure 5.

Three relevant observations can be made from the figure. The figure illustrates *(i)* the convergence of the expected elastic call holding times in a system with varying capacity to those in the corresponding $M/G/1/n_{e,\max}/\ PS$ queueing model. Furthermore, *(ii)* note that the expected holding times start to increase exponentially as $\rho_e \longrightarrow C^\star \approx 8.486$ (cf. Figure 3), corresponding with $\lambda_e \longrightarrow 0.240$, until the number of elastic calls present in the system becomes practically deterministic at $n_{e,\max}$ and the expected holding time flattens out at its maximum value. As an illustration of the above argument, *(iii)* note that for a range of λ_e values around 0.240 the elastic call QOS appears to be better (with 95% statistical significance) under a more lightly tailed elastic



(Weibull) (Pareto)

Figure 5: *Numerical results of experiment 4.*

call size distribution, which is due to a relatively strong heterogeneity in the carried elastic traffic load. In fact, for this range of λ_e the $M/G/1/n_{e,\max}/PS$ system with fixed capacity performs worst in QOS and best in GOS (elastic call blocking probability).

Although it was not so clear in Figure 3 (experiment 2) due to the linearity of the vertical axis, the left (Weibull) chart in Figure 5 more clearly indicates that as $\lambda_e \longrightarrow 0^+$, the expected elastic call holding times for the different η_e converge to distinct values that are ordered in accordance with the our principal result. The reason for pointing out this extreme case is that it is one of the cases considered in the next section to provide analytical support and intuition.

Remark 1 *The above experiments have demonstrated that in a processor sharing system with varying service capacity, the elastic call* QOS *improves as the distribution tail becomes heavier, i.e. as smaller calls become more frequent, while the rare large calls become larger. This effect has been observed within a given family of elastic call size distributions. Comparison of the numerical results for the Weibull and Pareto distributions also reveal that a more general statement regarding the impact of the elastic call size distribution on the* QOS *is rather difficult to give, as it is not trivial to identify the variability measure that fully captures the essence of its impact on the* QOS*. The impact is not purely determined by the coefficient of variation (∼ variance, second moment), since in that case the Weibull and Pareto curves would have to coincide for a given η_e. Also, the impact is not purely determined by the heaviness of the distribution tail, as the Pareto tail has been shown to generally outweigh the Weibull tail, while it is not the case in the presented numerical experiments that all Weibull curves lie strictly above all Pareto curves. Hence the* QOS *impact is determined by the characteristics of the elastic call size distribution in a broader sense. We refer to Remark 4 in Section 5.1.2 for a further elaboration on the cross-*PDF *comparisons.*

5 Analytical support and intuition

This section provides theoretical support for the observations presented in Section 4, by means of an analytical treatment of two extreme cases of the model defined in Section 3, as well as

a discussion of the performance in between these extreme cases. Section 5.1 treats the limit case of $\lambda_e \longrightarrow 0^+$, proving that a greater elastic call size variability indeed leads to better QOS. Subsequently, Section 5.2 shows that at the other extreme, i.e. $\lambda_e \longrightarrow \infty$ (and assuming $n_{e,\max} < \infty$ to ensure stability), the elastic QOS becomes not only *insensitive* to the elastic call size distribution, but also equal to that achieved in an $M/G/1/n_{e,\max}/PS$ system with *fixed* capacity. Finally, in Section 5.3 a brief intuitive discussion is provided regarding the intermediate case of $\lambda_e \in (0, \infty)$.

5.1 Limit case: $\lambda_e \longrightarrow 0^+$

Consider the case of exponentially distributed stream call holding times. The presented analysis for the limit case $\lambda_e \longrightarrow 0^+$ is broken up into two stages. First we determine a closed-form expression for the conditional expected holding time $\mathbf{T}_e(x)$ of an elastic call, indicating that it is concave in x. Subsequently, we show that for a given concave $\mathbf{T}_e(x)$ the expected holding time is decreasing in the coefficient of variation of the elastic call size.

5.1.1 Determine $\mathbf{T}_e(x)$

Denote with $\tau_{n_s}(x)$ the random holding time of an admitted elastic call of size x, that finds n_s active stream calls in the system upon arrival, and let $\widehat{\tau}_{n_s}(x) \equiv \mathbf{E}\{\tau_{n_s}(x)\}$ be its expectation. Define the vector $\widehat{\boldsymbol{\tau}}_+(x) \equiv (\widehat{\tau}_{n_s}(x),\ n_s \in \mathbb{S}_+)$. For the limit case of $\lambda_e \longrightarrow 0^+$, the conditional expected holding time $\mathbf{T}_e(x)$ of an elastic call of size x is then equal to

$$\mathbf{T}_e(x) = \sum_{n_s \in \mathbb{S}} \pi(n_s)\, \widehat{\tau}_{n_s}(x), \tag{2}$$

where $\pi(n_s)$ is the $M/M/C/C$ equilibrium probability that there are n_s stream calls in the system. We stress that the probability that an *admitted* elastic call finds n_s stream calls upon arrival is *not* in general equal to $\pi(n_s)$, since the thinned arrival process of *admitted* elastic calls needs not be Poisson, as will be demonstrated in the next section. In the case of $\lambda_e \longrightarrow 0^+$, however, the elastic call blocking probability converges to zero and thus the arrival process of *admitted* elastic calls becomes Poisson, so that PASTA can be applied as in (2).

In order to obtain explicit expressions for $\widehat{\tau}_{n_s}(x)$, $n_s \in \mathbb{S}$, we note that for $\lambda_e \longrightarrow 0^+$ the number of elastic calls in the system never exceeds 1, so that we may set $n_{e,\max} = 1$, without affecting the results.

Theorem 2 *Let $\mathcal{Q}$ be the infinitesimal generator of the $M/M/C-1/C-1$ model with stream call arrival rate λ_s and average stream call holding time μ_s^{-1}, and let $\boldsymbol{\pi}$ be the stationary distribution vector of the $M/M/C/C$ model with the same rates. Let $\boldsymbol{\pi}_+ \equiv (\pi(n_s),\ n_s \in \mathbb{S}_+)$. Denote with $\mathcal{B}_0 \equiv diag\,(C - n_s,\ n_s \in \mathbb{S}_+)$ the diagonal matrix containing the number of channels available for an elastic call in states $n_s \in \mathbb{S}_+$. Given $\mathbf{u} = (1, 1, \cdots, 1 + \lambda_s/(\mu_s C)) \in \mathbb{R}^C$, let $\boldsymbol{\gamma}$ be the unique solution to the system of linear equations*

$$\mathcal{Q}\,\boldsymbol{\gamma} = -\mathbf{u} + \frac{1}{\boldsymbol{\pi}_+\,\mathcal{B}_0\,\mathbf{1}}\mathcal{B}_0\,\mathbf{1}, \tag{3}$$

$$\boldsymbol{\pi}_+\,\mathcal{B}_0\,\boldsymbol{\gamma} = 0. \tag{4}$$

where $\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1} = C - \rho_s (1 - \mathbf{P}_s)$ *is the average number of channels available for elastic calls with* $\mathbf{P}_s = \pi(C)$*. Then the conditional expected holding time* $\widehat{\tau}_{n_s}(x)$ *of an elastic call of size* x *entering the system in the presence of* n_s *stream calls,* $n_s \in \mathbb{S}$*, is given by the closed-form expressions*

$$\widehat{\boldsymbol{\tau}}_+(x) = \frac{x}{\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1}} \mathbf{1} + \left[\mathcal{I} - \exp\left\{x \mathcal{B}_0^{-1} \mathcal{Q}\right\}\right] \boldsymbol{\gamma}, \tag{5}$$

$$\widehat{\tau}_C(x) = 1/(\mu_s C) + \widehat{\tau}_{C-1}(x). \tag{6}$$

Proof. Although the result is a special case of the more general analysis presented in [14], this dedicated proof is included both for reasons of clarity and to make our claim self-contained. We begin by noting that (6) indeed holds, since an elastic call that finds C stream calls present upon arrival is idle for an expected time $1/(\mu_s C)$ until one of the stream calls terminates, after which the unaffected elastic call finds itself in the presence of $C - 1$ stream calls.

Equation (5) is proven using marginal analysis. Let $n_s \in \mathbb{S}_+$ and consider a time interval of length $\Delta > 0$, with Δ sufficiently small such that the elastic call cannot terminate within this time, i.e. $\Delta < x/C$. Conditioning on all the possible events occurring in this interval, we get for $\widehat{\tau}_{n_s}(x)$:

$$\begin{aligned}\widehat{\tau}_{n_s}(x) = \Delta + \lambda_s \Delta \widehat{\tau}_{n_s+1}(x - O(\Delta)) + n_s \mu_s \Delta \widehat{\tau}_{n_s-1}(x - O(\Delta)) \\ + (1 - \lambda_s \Delta - n_s \mu_s \Delta) \widehat{\tau}_{n_s}(x - (C - n_s)\Delta) + o(\Delta),\end{aligned}$$

where the Landau symbols $O(\Delta)$ and $o(\Delta)$ are standard notation for some unspecified functions $F(\Delta)$ and $f(\Delta)$, having the property that $\lim_{\Delta \to 0} F(\Delta) = 0$ and $\lim_{\Delta \to 0} f(\Delta)/\Delta = 0$, respectively. Rearranging terms and letting $\Delta \downarrow 0$, we obtain the system of differential equations

$$(C - n_s) \frac{\partial \widehat{\tau}_{n_s}(x)}{\partial x} = 1 + \lambda_s \widehat{\tau}_{n_s+1}(x) + n_s \mu_s \widehat{\tau}_{n_s-1}(x) - (\lambda_s + n_s \mu_s) \widehat{\tau}_{n_s}(x), \quad n_s \in \mathbb{S}_+,$$

which, using (6), may equivalently be written in matrix notation:

$$\mathcal{B}_0 \frac{\partial}{\partial x} \widehat{\boldsymbol{\tau}}_+(x) = \mathbf{u} + \mathcal{Q} \widehat{\boldsymbol{\tau}}_+(x).$$

The system is complemented with the initial condition $\widehat{\boldsymbol{\tau}}_+(0) = \mathbf{0}$ reflecting the fact that the transfer time $\tau_{n_s}(0)$, $n_s \in \mathbb{S}_+$, of an 'empty' elastic call is zero, almost surely.

The existence and uniqueness of a solution $\widehat{\boldsymbol{\tau}}_+(x)$ for every initial vector, immediately follows by checking the conditions in e.g. Braun [3] (Theorem 3, page 412). The existence of a vector $\boldsymbol{\gamma}$ that satisfies (3) and its uniqueness up to a translation along the vector $\mathbf{1}$, are guaranteed by results in Markov decision theory (see e.g. Tijms [18]). Normalizing $\boldsymbol{\gamma}$ as in (4) yields a unique solution for $\boldsymbol{\gamma}$. The proposition is then proven by substituting the claimed unique solution into the system of differential equations and verifying that it indeed holds, using (3). To conclude the proof, observe that (5) satisfies the initial condition $\widehat{\boldsymbol{\tau}}_+(0) = \mathbf{0}$. ■

The resulting expressions for $\widehat{\tau}_{n_s}(x)$ may be convex or concave in x, or neither convex nor concave, depending on the choices of C and $n_s \in \mathbb{S}$. We have derived explicit expressions for $\widehat{\tau}_{n_s}(x)$ for $C \in \{1, 2, 3, 4\}$ and $n_s \in \mathbb{S}$, and have observed that $\widehat{\tau}_{n_s}(x)$ is typically convex for small

n_s, neither convex nor concave for medium n_s, and concave for large n_s. For $C = 1$ and $C = 2$ the expressions are sufficiently compact to be included below. For $C = 1$ we find

$$\begin{cases} \widehat{\tau}_0(x) = (\rho_s + 1)\ x \\ \\ \widehat{\tau}_1(x) = \widehat{\tau}_0(x) + \mu_s^{-1}, \end{cases}$$

while for $C = 2$ we find

$$\begin{cases} \widehat{\tau}_0(x) = \frac{\rho_s^2+2\rho_s+2}{2(\rho_s+2)}\, x + \frac{\rho_s(\rho_s+1)}{\mu_s(\rho_s+2)^2}\left(\exp\left\{-\frac{1}{2}x\mu_s\left(\rho_s+2\right)\right\} - 1\right), \\ \\ \widehat{\tau}_1(x) = \frac{\rho_s^2+2\rho_s+2}{2(\rho_s+2)}\, x - \frac{2(\rho_s+1)}{\mu_s(\rho_s+2)^2}\left(\exp\left\{-\frac{1}{2}x\mu_s\left(\rho_s+2\right)\right\} - 1\right), \\ \\ \widehat{\tau}_2(x) = \widehat{\tau}_1(x) + (2\mu_s)^{-1}. \end{cases}$$

Note that for $C = 1$ both $\widehat{\tau}_0(x)$ and $\widehat{\tau}_1(x)$ are linear in x, whereas for $C = 2$, $\widehat{\tau}_0(x)$ is strictly convex while $\widehat{\tau}_1(x)$ and $\widehat{\tau}_2(x)$ are strictly concave. An expression for $\mathbf{T}_e(x)$ is then readily derived using (2) and the $M/M/C/C$ equilibrium distribution $\pi(n_s) = \mathbf{G}^{-1}\rho_s^{n_s}/n_s!$, $n_s \in \mathbb{S}$, with $\mathbf{G} \equiv \sum_{n_s\in\mathbb{S}} \rho_s^{n_s}/n_s!$ the appropriate normalization constant. For $C = 1$, the conditional expected holding time is given by

$$\mathbf{T}_e(x) = \frac{\rho_s}{\mu_s\left(\rho_s+1\right)} + (\rho_s + 1)\ x,$$

which is linear in x, while for $C = 2$ we obtain

$$\begin{aligned} \mathbf{T}_e(x) \quad = \quad & \frac{\rho_s^2}{2\mu_s\left(\rho_s^2+2\rho_s+2\right)} + \frac{\rho_s^2+2\rho_s+2}{2\left(\rho_s+2\right)}\, x + \\ & - \left(\frac{2\rho_s\left(\rho_s+1\right)^2}{\mu_s\left(\rho_s+2\right)^2\left(\rho_s^2+2\rho_s+2\right)}\right) \times \left(\exp\left\{-\frac{1}{2}x\mu_s\left(\rho_s+2\right)\right\} - 1\right), \end{aligned}$$

which is strictly concave in x, as it also is for $C \in \{3, 4\}$.

The expressions for the elastic call holding time $\mathbf{T}_e(x)$ have a straightforward interpretation that is most transparent for the case of $C = 1$. It is readily verified that $\mathbf{T}_e(x)$ is equal to the nominal holding time (x) *plus* the expected residual waiting time upon arrival ($\pi(1)\,\mu_s^{-1}$) *plus* the expected number of transfer interruptions ($\lambda_s\, x$) *times* the duration of such an interruption (μ_s^{-1}), applying Wald's equation (see e.g. [18]). To see that the expected number of transfer interruptions is given by $\lambda_s\, x$, note that it is equal to the expected number of Poisson arrivals (at rate λ_s) during the nominal holding time x.

5.1.2 A greater elastic call size variability improves the QOS

In this subsection we demonstrate that a greater elastic call size variability improves the QOS if the conditional expected elastic call holding time $\mathbf{T}_e\left(x\right)$ is concave in x, starting with a simple yet very insightful intuitive argument. Although in general $\mathbf{T}_e\left(x\right)$ may depend on the distribution of x, it is insensitive in the considered limit case of $\lambda_e \longrightarrow 0^+$, as is shown above. This enables

us to compare the effect of the elastic call size distribution φ_e on the expected holding time $\mathbf{T}_e$ for a single given $\mathbf{T}_e(x)$.

Starting out with a degenerate distribution in $x = \mu_e^{-1}$ with a coefficient of variation equal to zero and $\mathbf{T}_e = \mathbf{T}_e\left(\mu_e^{-1}\right)$, we increase the coefficient of variation by shifting an equal amount of probability mass equally far up and down the x scale, such that the mean elastic call size remains μ_e^{-1} (see Figure 6). Due to the concavity of $\mathbf{T}_e(x)$ the QOS gain that is made by moving some probability mass towards the lower end of the x scale, outweighs the QOS cost that is incurred by moving the same probability mass equally far towards the upper end of the x scale, so that indeed the net effect is a QOS improvement.

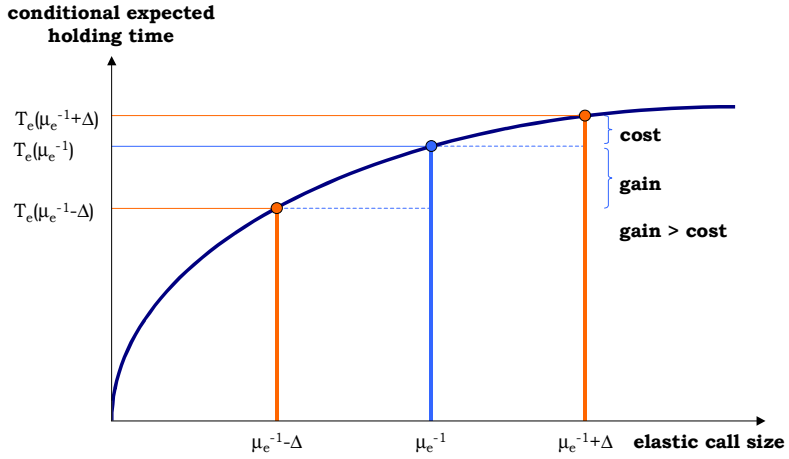


Figure 6: *Analytical support: heavier tails improve* QOS *under a concave* $\mathbf{T}_e(x)$.

After this rather intuitive argument, two more rigorous approaches follow to support the claim that if the expected holding time $\mathbf{T}_e(x)$ is concave in x the expected holding time $\mathbf{T}_e$ is lower if the PDF φ_e has a larger coefficient of variation (is more variable), given the same mean. Firstly, for the case that the elastic call size x has a discrete PDF on two values, Theorem 3 below proves analytically that concavity of $\mathbf{T}_e(x)$ implies that $\mathbf{T}_e$ is decreasing in η_e.

Theorem 3 *Assume that the elastic call size x has a discrete* PDF *that can take on two values. If the conditional expected holding time $\mathbf{T}_e(x)$ of an elastic call of size x is concave in x, then the expected holding time $\mathbf{T}_e$ is decreasing in the coefficient of variation of x, for a given mean $\mu_e^{-1} \equiv \mathbf{E}\{x\}$.*

Proof. Denote the two possible elastic call sizes with $x_0 \equiv \mu_e^{-1} - \vartheta_0$ and $x_1 \equiv \mu_e^{-1} + \vartheta_1$, for $\vartheta_0 \in \left[0, \mu_e^{-1}\right]$ and $\vartheta_1 \in [0, \infty)$. It is readily verified that $\xi_0 \equiv \Pr\{x = x_0\} = \vartheta_1/(\vartheta_0 + \vartheta_1)$ and $\xi_1 \equiv \Pr\{x = x_1\} = \vartheta_0/(\vartheta_0 + \vartheta_1)$ must hold, in order to establish that $\mathbf{E}\{x\} = \mu_e^{-1}$. The coefficient of variation η_e is equal to $\mu_e\sqrt{\vartheta_0\vartheta_1}$, which is strictly increasing in both ϑ_0 and ϑ_1. For a given μ_e, the expected holding time of an elastic call can be written as a function of ϑ_0 and $\eta_e^\star \equiv \vartheta_0\vartheta_1$:

$$\theta(\vartheta_0, \eta_e^\star) = \frac{\eta_e^\star/\vartheta_0}{\vartheta_0 + \eta_e^\star/\vartheta_0}\mathbf{T}_e\left(\mu_e^{-1} - \vartheta_0\right) + \frac{\vartheta_0}{\vartheta_0 + \eta_e^\star/\vartheta_0}\mathbf{T}_e\left(\mu_e^{-1} + \eta_e^\star/\vartheta_0\right),$$

where $\theta(\cdot,\cdot)$ denotes the expected holding time $\mathbf{T}_e$ which is a function of ϑ_0 and $\eta_e^\star$. In order to prove the theorem, it suffices to show that $\theta(\vartheta_0, \eta_e^\star)$ is decreasing in $\eta_e^\star$, since the coefficient of

variation η_e is a simple increasing function of $\eta_e^\star$. The first derivative of $\theta\left(\vartheta_0, \eta_e^\star\right)$ with respect to $\eta_e^\star$ is given by

$$\frac{\partial}{\partial \eta_e^\star}\theta\left(\vartheta_0, \eta_e^\star\right) = -\frac{\mathbf{T}_e\left(\mu_e^{-1} + \eta_e^\star/\vartheta_0\right) - \mathbf{T}_e\left(\mu_e^{-1} - \vartheta_0\right)}{\left(\vartheta_0 + \eta_e^\star/\vartheta_0\right)^2} + \frac{\mathbf{T}_e'\left(\mu_e^{-1} + \eta_e^\star/\vartheta_0\right)}{\vartheta_0 + \eta_e^\star/\vartheta_0}.$$

Using the property of concave functions that the tangent in any point lies above the function itself, the result immediately follows:

$$\left(\vartheta_0 + \eta_e^\star/\vartheta_0\right)\mathbf{T}_e'\left(\mu_e^{-1} + \eta_e^\star/\vartheta_0\right) \leq \mathbf{T}_e\left(\mu_e^{-1} + \eta_e^\star/\vartheta_0\right) - \mathbf{T}_e\left(\mu_e^{-1} - \vartheta_0\right) \Longleftrightarrow \frac{\partial}{\partial \eta_e^\star}\theta\left(\vartheta_0, \eta_e^\star\right) \leq 0.$$

■

Secondly, for the Weibull and Pareto PDFs introduced in Section 4 (including the deterministic case), Table 2 contains the expected elastic call holding times,

$$\mathbf{T}_e \equiv \int_{x=0}^{\infty} \mathbf{T}_e\left(x\right)\, \varphi_e(x)\, dx,$$

given $C \in \{1, 2, 3, 4\}$ and the conditional elastic call holding times $\mathbf{T}_e\left(x\right)$ from Section 5.1.1. Recall that the two Pareto cases with infinite coefficient of variation are defined by $\alpha_1 = 1.66$, $\alpha_2 = 1.35$, and the c's such that the expected value is as intended. The stream call process is defined by an average call holding time of $\mu_s^{-1} = 50$ seconds, while for each C the stream call arrival rate λ_s is set such that the resulting stream traffic load $\rho_s \equiv \lambda_s/\mu_s$ induces a 1% stream call blocking probability.

Weibull		η_e									$M/G/1/PS$	
C	ρ_s	0	1/8	1/4	1/2	1	2	4	8	16	$(\star)$	$(\diamond)$
1	0.010	36.216	36.216	36.216	36.216	36.216	36.216	36.216	36.216	36.216	36.216	35.716
2	0.153	20.375	20.371	20.359	20.314	20.186	19.977	19.782	19.644	19.552	19.374	19.124
3	0.455	15.297	15.293	15.280	15.232	15.093	14.847	14.598	14.412	14.287	14.038	13.871
4	0.869	12.731	12.727	12.715	12.669	12.533	12.283	12.016	11.812	11.672	11.388	11.263
Pareto		η_e									$M/G/1/PS$	
C	ρ_s	0	1/8	1/4	1/2	1	2	4	∞_1	∞_2	$(\star)$	$(\diamond)$
1	0.010	36.216	36.216	36.216	36.216	36.216	36.216	36.216	36.216	36.216	36.216	35.716
2	0.153	20.375	20.372	20.362	20.338	20.301	20.273	20.262	20.187	20.047	19.374	19.124
3	0.455	15.297	15.293	15.284	15.258	15.216	15.184	15.171	15.081	14.906	14.038	13.871
4	0.869	12.731	12.727	12.718	12.693	12.652	12.620	12.606	12.513	12.329	11.388	11.263

Table 2: *Expected elastic call holding times.*

Since for $C = 1$, $\mathbf{T}_e\left(x\right)$ is linear in x, the shape of the elastic call size distribution has no impact on the expected holding times. For $C \in \{2, 3, 4\}$, i.e. the cases where $\mathbf{T}_e\left(x\right)$ is strictly concave in x, the values in the table demonstrate that indeed the expected elastic call holding time is decreasing in η_e. For the Pareto cases with $\eta_e = \infty$, the QOS improves with lower α (heavier tail). The final column of Table 2 (marked $(\diamond)$) contains the expected holding times $\mathbf{T}_e = 1/\mu_e^\star$ of an elastic call if it were served in a fixed capacity $M/G/1/PS$ system (with $\lambda_e \longrightarrow 0^+$), while in the adjacent column (marked $(\star)$) these values are raised by $\pi(C)/(\mu_s C)$, the expected delay before an admitted elastic call may start its transfer in the system with varying capacity. The significance of the latter values is that they are limit values for the system with varying capacity, as the variability of x grows. Another observation that can be made from the table, is that the

QOS improves as C increases, which is trivial as more capacity remains to be assigned to each admitted elastic call. Finally, for a given $\eta_e < \infty$, the QOS induced by the Weibull elastic call size distribution is better than that induced by the Pareto distribution (see also Remark 1 at the end of Section 4 and Remark 4 below).

Remark 4 *As a sequel to Remark 1, we utilize the exact numerical results of Table 2 to provide additional insight in cross-*PDF *comparisons of the impact of the call size variability on the experienced* QOS*. As observed in various places in the paper, the Weibull elastic call size distribution appears to yield better* QOS *than the Pareto distribution, for the same mean and coefficient of variation. Consider e.g. the case of $\eta_e = 1$ in Table 2. In light of the intuitive argument that the* QOS *is influenced by the balance between a small number of extremely large calls and a large number of small calls, refer to Figure 7 for the corresponding* CDF*s (solid curves).*

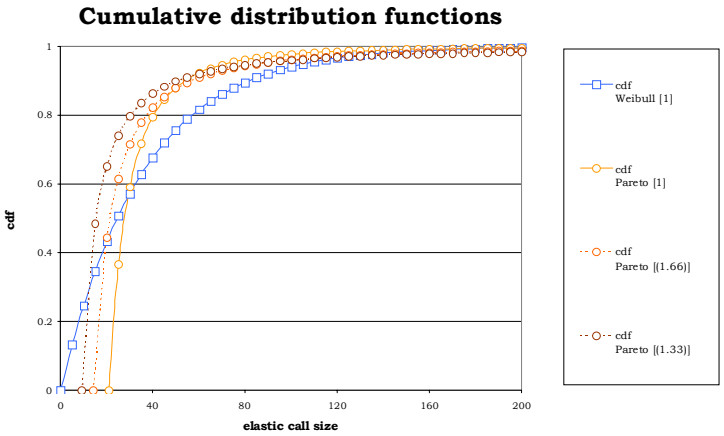


Figure 7: *Cross-*PDF *comparisons.*

In contrast to the Weibull CDF*, which exhibits a significant probability mass for small elastic call sizes, the Pareto* CDF *is positive only for call sizes greater than $c \approx 20.713$, which is rather large compared to the mean. This is in line with the intuitive argument given above. If we were to increase the call size variability (only) of the Pareto* CDF*, e.g. consider the cases with $\eta_e = \infty$, we learn that the* QOS *is equivalent (Table 2: '∞_1') or better ('∞_2') compared to the Weibullian case, in correspondence with the shift of probability mass towards smaller call sizes that is indicated in Figure 7 (dashed curves: the minimum call size c decrease to 14.058 and 9.167, respectively).*

Although fundamental insight has been provided, the quest for a well-defined variability measure that captures the essential characteristics of a PDF *to allow cross-*PDF *comparisons, remains as an open issue for further research.*

5.1.3 Accelerated stream call process

The results presented in Theorem 2 can be used to prove that as the stream call arrival and termination process is accelerated, i.e. $\lambda_s, \mu_s \longrightarrow \infty$ while ρ_s remains fixed, the QOS converges to that of an $M/G/1/PS$ queueing system, and thus becomes insensitive to the elastic call size distribution. Define $\lambda_s^\vartheta \equiv \vartheta\lambda_s$ and $\mu_s^\vartheta \equiv \vartheta\mu_s$, with $\vartheta \in \mathbb{R}_+$, so that $\mathcal{Q}^\vartheta = \vartheta\mathcal{Q}$ is the infinitesimal generator of the Markov chain that describes the accelerated ($\vartheta > 1$) stream call process. The

modification affects only $\mathcal{Q}$, so that the vector $\boldsymbol{\gamma}_\vartheta$ that solves the system (3), (4) is equal to $\boldsymbol{\gamma}/\vartheta$, with $\boldsymbol{\gamma}$ the solution in the basic model ($\vartheta = 1$). As a consequence, we find

$$\widehat{\boldsymbol{\tau}}_+(x) = \frac{x}{\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1}}\mathbf{1} + \left[\mathcal{I} - \exp\left\{\vartheta x \mathcal{B}_0^{-1}\mathcal{Q}\right\}\right] \frac{\boldsymbol{\gamma}}{\vartheta} \longrightarrow \frac{x}{\boldsymbol{\pi}_+ \mathcal{B}_0 \mathbf{1}}\mathbf{1} \quad (\vartheta \longrightarrow \infty). \tag{7}$$

To see this, note that $\mathcal{B}_0^{-1}\mathcal{Q}$ is the generator of an irreducible finite state space Markov chain, with equilibrium distribution vector $\boldsymbol{\pi}\mathcal{B}_0/(\boldsymbol{\pi}\mathcal{B}_0\,\mathbf{1})$, so that $\lim_{\vartheta\to\infty} \exp\left\{\vartheta x \mathcal{B}_0^{-1}\mathcal{Q}\right\} = \mathbf{1}\,\boldsymbol{\pi}\mathcal{B}_0/(\boldsymbol{\pi}\mathcal{B}_0\,\mathbf{1})$. Furthermore, also $\widehat{\tau}_C(x) \longrightarrow x/(\boldsymbol{\pi}_+ \mathcal{B}_0\,\mathbf{1})$, as $\vartheta \longrightarrow \infty$, which immediately follows from (6), (7) and the fact that $1/(\vartheta\mu_s C) \longrightarrow 0$, as $\vartheta \longrightarrow \infty$. Hence

$$\mathbf{T}_e(x) = \sum_{n_s \in \mathbb{S}} \pi(n_s)\,\widehat{\tau}_{n_s}(x) \longrightarrow \frac{x}{\boldsymbol{\pi}_+ \mathcal{B}_0\,\mathbf{1}} \quad (\vartheta \longrightarrow \infty),$$

which is precisely the conditional expected holding time in an $M/G/1/PS$ model with fixed capacity $\boldsymbol{\pi}_+ \mathcal{B}_0\,\mathbf{1} = C^\star = C - \rho_s\,(1 - \mathbf{P}_s)$ and an infinitesimally low (elastic) call arrival rate. It is obvious that then also the expected holding times must be equal.

5.2 Limit case: $\lambda_e \longrightarrow \infty$

Consider the case of exponentially distributed stream call holding times and assume $n_{e,\max} < \infty$. In this subsection we prove that in the limit of $\lambda_e \longrightarrow \infty$, ensuring a continuous presence of $n_{e,\max}$ elastic calls, the conditional expected holding time $\mathbf{T}_e(x)$ of an elastic call is linear in x, insensitive to the elastic call size distribution, and moreover, $\mathbf{T}_e(x)$ is equal to the conditional expected holding time in the corresponding $M/G/1/n_{e,\max}/PS$ queueing model with fixed capacity.

5.2.1 Determine $\mathbf{T}_e(x)$

With $\tau_{n_s}(x)$ and $\widehat{\tau}_{n_s}(x)$ as introduced in Section 5.1.1, define the vector $\widehat{\boldsymbol{\tau}}_+(x) \equiv (\widehat{\tau}_{n_s}(x),\ n_s \in \mathbb{S}_+)$. The conditional expected holding time $\mathbf{T}_e(x)$ of an elastic call of size x is then equal to

$$\mathbf{T}_e(x) = \sum_{n_s \in \mathbb{S}} \pi^\star(n_s)\,\widehat{\tau}_{n_s}(x),$$

where $\pi^\star(n_s)$ is the probability that an *admitted* elastic call finds n_s stream calls upon arrival, which is *not* equal to the equilibrium probability $\pi(n_s)$ that n_s stream calls are present in the system. For instance, $\pi(C) \neq \pi^\star(C) = 0$ since an elastic call must terminate in a system state $n_s \in \mathbb{S}_+$ while in the case of $\lambda_e \to \infty$ its freed service position is immediately taken by a fresh elastic call.

In order to determine the probabilities $\pi^\star(n_s)$, $n_s \in \mathbb{S}_+$, time is rescaled as illustrated by Figure 8 for the case of $C = 2$. The random time between two successive stream call arrival or termination events during which n_s stream calls are present in the system is weighted by the corresponding number of channels assigned to an elastic call, $n_{e,\max}^{-1}\,(C - n_s)$, so that on the new time scale a fixed capacity of one channel is continuously available for each of the elastic calls.

On the new time scale the coinciding elastic call arrival and termination instants form a renewal process with random interrenewal times equal to the elastic call sizes. The probability $\pi^\star(n_s)$ that a fresh elastic call finds n_s stream calls present upon admission is then given by the

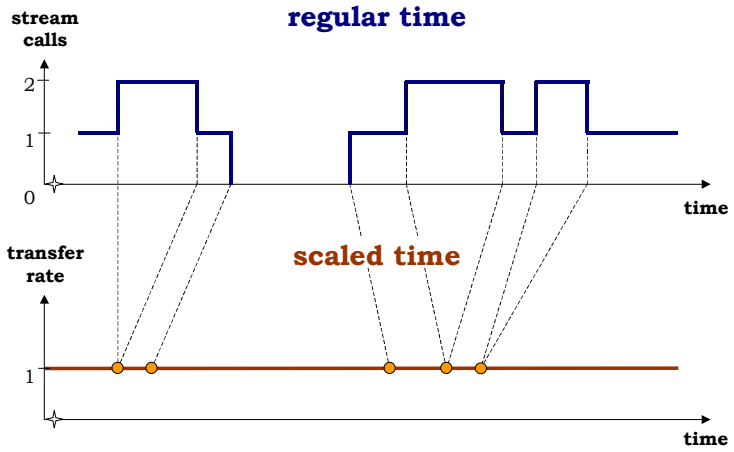


Figure 8: *Analytical support: rescaling of time.*

stationary distribution of the system state at such a renewal event, i.e. the fraction of time that the system serves n_s stream calls:

$$\pi^\star(n_s) \equiv \frac{n_{e,\max}^{-1}\,(C-n_s)\;\pi(n_s)}{\sum\limits_{n_s'\in\mathbb{S}_+} n_{e,\max}^{-1}\,(C-n_s')\;\pi(n_s')},\; n_s\in S_+ \quad\Longleftrightarrow\quad \boldsymbol{\pi}_+^\star \equiv (\pi^\star(n_s),\; n_s\in\mathbf{S}_+) = \frac{\boldsymbol{\pi}_+\,\mathcal{B}_\infty}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}},$$

with $\pi(n_s)$ the time fraction that n_s stream calls are present in an $M/M/C/C$ Erlang loss system with stream call traffic load ρ_s on a regular time scale. The diagonal matrix $\mathcal{B}_\infty \equiv diag\left(n_{e,\max}^{-1}\,(C-n_s),\;\; n_s\in\mathbb{S}_+\right) = n_{e,\max}^{-1}\mathcal{B}_0$ contains the number of channels available for an elastic call in states $n_s\in\mathbb{S}_+$.

For the considered case of $\lambda_e \longrightarrow \infty$, explicit expressions for $\widehat{\tau}_{n_s}(x)$, $n_s\in\mathbb{S}$, can be obtained by analogy with Theorem 2, using $\mathcal{B}_\infty$ rather than $\mathcal{B}_0$, as the number of elastic calls in the system is now continuously equal to $n_{e,\max}$. Note that the solution $\boldsymbol{\gamma}$ to the system of equations (3) and (4) is the same, regardless of whether $\mathcal{B}_0$ or $\mathcal{B}_\infty$ is used. In light of the earlier remark that an infinite elastic call arrival rate implies that no call is ever admitted in the presence of C stream calls and hence each elastic call can start service immediately upon admission, we note that nonetheless an expression is obtained for $\widehat{\tau}_C(x)$. Naturally, since $\pi^\star(C)=0$, $\widehat{\tau}_C(x)$ does not contribute to $\mathbf{T}_e(x)$.

It is readily proven that for any C, $\mathbf{T}_e(x)$ is linear in x, with $\mathbf{T}_e(0)=0$:

$$\begin{aligned}\mathbf{T}_e(x) &= \frac{\boldsymbol{\pi}_+\,\mathcal{B}_\infty}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}}\left\{\frac{x}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}}\mathbf{1} + \left[\mathcal{I}-\exp\left\{x\mathcal{B}_\infty^{-1}\mathcal{Q}\right\}\right]\boldsymbol{\gamma}\right\}\\ &= \frac{x}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}} + \frac{\boldsymbol{\pi}_+\,\mathcal{B}_\infty}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}}\left\{\sum_{i=1}^{\infty}\frac{\left(x\mathcal{B}_\infty^{-1}\mathcal{Q}\right)^i}{i!}\right\}\boldsymbol{\gamma} = \frac{x}{\boldsymbol{\pi}_+\,\mathcal{B}_\infty\,\mathbf{1}},\end{aligned}$$

using $\boldsymbol{\pi}_+\mathcal{Q}=0$, due to the reversibility of the $M/M/C/C$ queue (note that $\boldsymbol{\pi}_+$ contains equilibrium probabilities of the $M/M/C/C$ queue, while $\mathcal{Q}$ is the infinitesimal generator of the $M/M/C-1/C-1$ queue). For $C=1$, the conditional expected holding time is given by

$$\mathbf{T}_e(x) = n_{e,\max}\;(\rho_s+1)\;x,$$

which is simply equal to the linear function $\widehat{\tau}_0(x)$, while for $C = 2$ we obtain

$$\mathbf{T}_e(x) = n_{e,\max} \frac{\rho_s^2 + 2\rho_s + 2}{2(\rho_s + 2)} x,$$

which is a weighted average of $\widehat{\tau}_0(x)$ and $\widehat{\tau}_1(x)$, whose respective convexity and concavity balance out, as is also the case for $C \in \{3, 4\}$.

5.2.2 The QOS is insensitive to the elastic call size variability

Since $\mathbf{T}_e(x)$ is linear in x for all C and finite $n_{e,\max}$, the shape of the elastic call size distribution has no impact on the expected holding times. Moreover, $\mathbf{T}_e$ is equal to $n_{e,\max}/(\mu_e C^\star)$, the average elastic call size divided by the average number of available channels. Note that this is precisely the expected holding time in an $M/G/1/n_{e,\max}/PS$ system:

$$\mathbf{T}_e = \lim_{\lambda_e \longrightarrow \infty} \frac{1}{\mu_e^\star} \left(\sum_{n_e=0}^{n_{e,\max}-1} (\rho_e^\star)^{n_e} (n_e + 1) \right) \Bigg/ \left(\sum_{n_e=0}^{n_{e,\max}-1} (\rho_e^\star)^{n_e} \right) = \frac{n_{e,\max}}{\mu_e C^\star},$$

using l'Hôpital's rule. This result is not so surprising as for $\lambda_e \longrightarrow \infty$ the distribution of the system state observed upon arrival by an admitted elastic call is independent of the elastic call size distribution, so that the fact that the system continuously contains $n_{e,\max}$ elastic calls implies that the expected amount of capacity an admitted elastic call enjoys is equal to the expected amount of capacity that is available, divided by the (deterministic) number $n_{e,\max}$ of elastic calls sharing this capacity.

5.2.3 Accelerated stream call process

As a final note, in contrast with the results of Section 5.1.3, the (conditional) expected holding times are insensitive to an *acceleration* of the stream call arrival and termination process, since it influences $\mathbf{T}_e(x)$ only through ρ_s and not through λ_s and μ_s individually.

5.3 Intermediate case: $\lambda_e \in (0, \infty)$

So far we have investigated the effect of the elastic call size distribution on the QOS measures for the extreme cases of $\lambda_e \longrightarrow 0^+$ and $\lambda_e \longrightarrow \infty$. The principal reason why these limit cases are analytically tractable is that the level of competition of a tagged elastic call is independent of the elastic call size distribution, as the number of elastic calls competing for the same resources is either 0 for $\lambda_e \longrightarrow 0^+$, or $n_{e,\max} - 1$ for $\lambda_e \longrightarrow \infty$. Moreover, in these cases the distribution of the number of present stream calls that an admitted elastic calls sees upon arrival can be explicitly determined and is insensitive to the elastic call size distribution. For $\lambda_e \in (0, \infty)$, however, the number of competing elastic calls that a tagged elastic call endures is not only a random variable, but it is also sensitive to the elastic call size distribution. As a consequence, the distribution of the system state upon departure of an elastic call is also sensitive to the elastic call size distribution, and hence the distribution of the system state upon admission of an elastic call is as well.

As an illustrative argument, recall from Section 5.1.1 the expressions for the conditional expected call holding time $\widehat{\tau}_{n_s}(x)$, $n_s \in \mathbb{S}$, of an elastic call of size x, admitted to the system in the

presence of n_s stream calls (with $n_{e,\max} = 1$). It was noted that for $C = 2$, $\widehat{\tau}_0(x)$ is strictly convex while $\widehat{\tau}_1(x)$ and $\widehat{\tau}_2(x)$ are strictly concave. The weighted average $\mathbf{T}_e(x) \equiv \sum_{n_s \in \mathbb{S}} p(n_s)\, \widehat{\tau}_{n_s}(x)$ of these expressions is concave if sufficient weight lies on the concave $\widehat{\tau}_1(x)$ and $\widehat{\tau}_2(x)$, i.e. if $p(0)$ is sufficiently small. We conjecture that $p(0)$ increases monotonously from $\pi(0)$ to $\pi^\star(0)$, as λ_e runs from 0 to ∞, which is backed by some additional simulations (not included). Besides the fact that this monotonicity implies that indeed $\mathbf{T}_e(x)$ is concave for all λ_e, so that a greater elastic call size variability enhances the QOS, simulation results further revealed that for each $\lambda_e \in (0, \infty)$, the probability $p(0)$ is inversely proportional to the coefficient of variation η_e of the elastic call size. In light of the discussion in Section 4 regarding the potentially distorting effect of a finite $n_{e,\max}$ on elastic call blocking and the experienced QOS, we remark that this phenomenon was not observed for $n_{e,\max} = 1$, so that the above argument is valid.

6 An extended model

In the previous sections we have demonstrated and analytically supported the phenomenon that QOS improves under more variable elastic call sizes in a PS model with varying capacity. We would now like to shed some light on an interesting extension of the basic model considered above (see Figure 9). As opposed to the basic model of Figure 1, in the current model elastic calls that cannot find a free position in the PS queue (*transfer queue*) are not blocked but rather queued in an infinite FIFO *access queue*. This model was previously studied in [10] for exponentially distributed elastic call sizes, while [1] considers the elastic call performance in isolation, i.e. without the capacity fluctuations due to the prioritized stream call arrival and departure process.

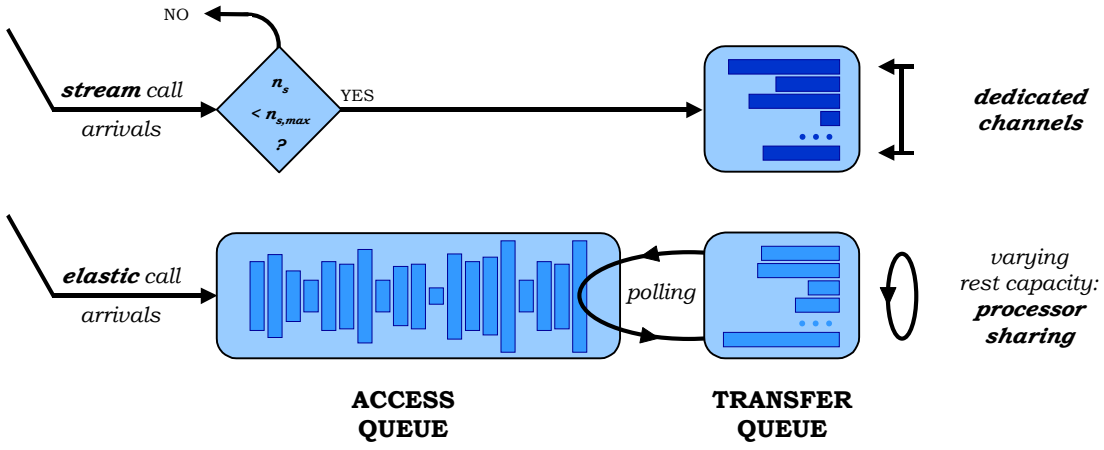


Figure 9: *The extended model.*

The number of service positions in the transfer queue is denoted by $n_{e,\max}^{\text{transfer}}$, and we are primarily interested in the impact of $n_{e,\max}^{\text{transfer}}$ on the relative performance of the different elastic call size distribution tails. Moreover, we compare this impact in both the integrated services model with varying PS capacity and the model with fixed PS capacity $C^\star \equiv C - \rho_s(1 - \mathbf{P}_s)$, i.e. the *average* number of available channels in the integrated services model.

With regards to the *fixed* capacity model, we note that the extreme cases of $n_{e,\max}^{\text{transfer}} \in \{1, \infty\}$ represent the pure FIFO and PS models, respectively, and allow exact expressions for the expected

elastic call holding times. For the case $n_{e,\max}^{\text{transfer}} = 1$ (FIFO), the expected elastic call holding time is given by the well-known Pollaczek-Khintchine formula (see e.g. [18]):

$$\mathbf{T}_e = \frac{\left(\eta_e^2 + 1\right)}{2} \cdot \frac{\rho_e / C^\star}{\mu_e \left(C^\star - \rho_e\right)} + \frac{1}{\mu_e C^\star},$$

implying that heavier tails (higher η_e) induce *worse* QOS. At the other extreme of $n_{e,\max}^{\text{transfer}} = \infty$ (PS), the expected elastic call holding time is equal to

$$\mathbf{T}_e = \frac{1}{\mu_e \left(C^\star - \rho_e\right)},$$

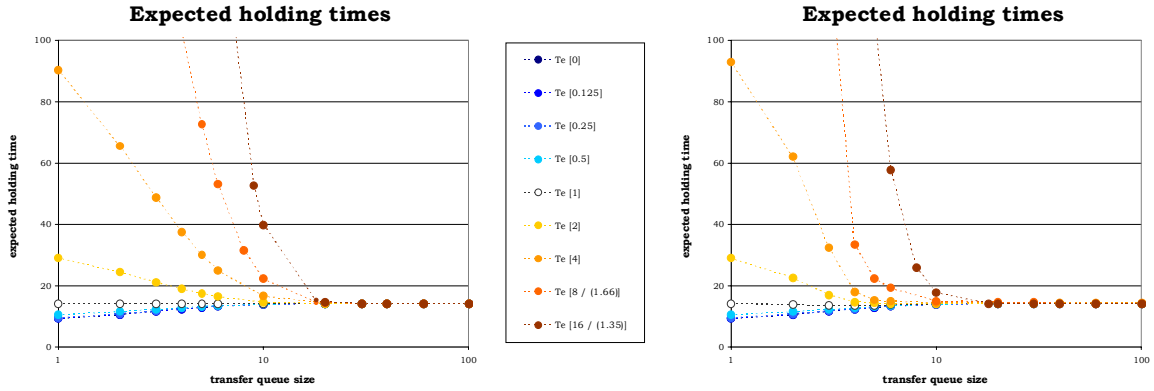
as already stated in Section 3, which expresses insensitivity of $\mathbf{T}_e$ with respect to η_e. In [1] the following approximation of the expected elastic call holding time is presented for a fixed capacity system that applies for all $n_{e,\max}^{\text{transfer}}$:

$$\mathbf{T}_e \cong \frac{\left(\eta_e^2 + 1\right)}{2} \cdot \frac{\left(\rho_e / C^\star\right)^{n_{e,\max}^{\text{transfer}}}}{\mu_e \left(C^\star - \rho_e\right)} + \frac{1 - \left(\rho_e / C^\star\right)^{n_{e,\max}^{\text{transfer}}}}{\mu_e \left(C^\star - \rho_e\right)},$$

where the first term approximates the expected access time and the second term approximates the expected transfer time. It is readily verified that the approximation provides exact results for $n_{e,\max}^{\text{transfer}} \in \{1, \infty\}$, representing pure FIFO and PS models, respectively, as well as for the case of exponentially distributed elastic call sizes, where $\mathbf{T}_e$ is independent of $n_{e,\max}^{\text{transfer}}$. Both the exact extreme cases and the approximation suggest that lower η_e yield better QOS for small $n_{e,\max}^{\text{transfer}}$ while its impact vanishes as $n_{e,\max}^{\text{transfer}}$ increases.

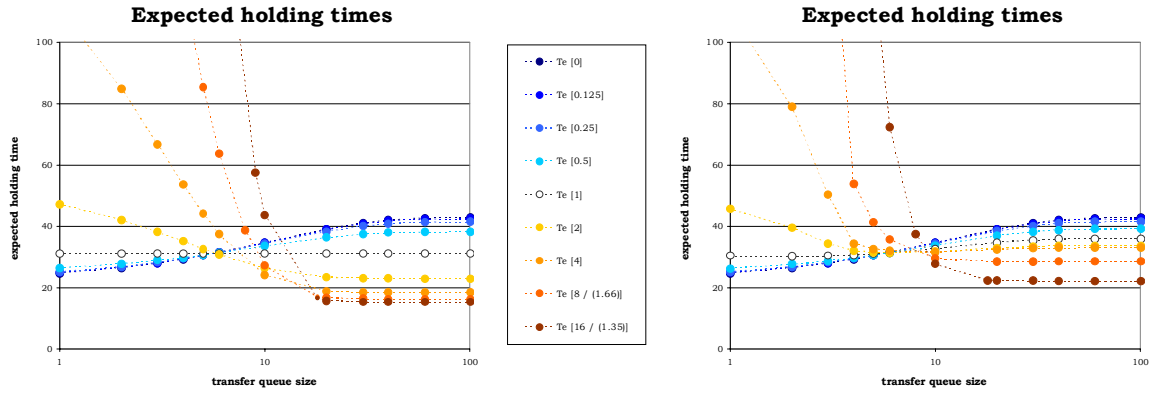
Figure 10 shows the numerical results that back this expectation. The plotted values are obtained by exact calculations where possible and simulations elsewhere. It is noted that in the Pareto cases with $\eta_e = \infty$ the expected elastic call holding times are infinite for small $n_{e,\max}^{\text{transfer}}$ and finite for large $n_{e,\max}^{\text{transfer}}$, and simulation experiments as used to generate Figure 10 can only loosely indicate the minimum transfer queue size that guarantees a finite expected holding time. As a side result, the above approximation appears to be rather good for the Weibullian elastic call sizes but very poor for the Pareto case, especially for moderate values of $n_{e,\max}^{\text{transfer}}$, where it occasionally even underestimates $\mathbf{T}_e$ by a factor greater than five.

For the more interesting case of a *varying* server capacity we expect that for small values of $n_{e,\max}^{\text{transfer}}$ the FIFO queue dominates and a heavier tail degrades the QOS. We argue that the varying capacity does not affect the qualitative phenomenon that relatively small elastic calls suffer greatly from relatively large elastic calls ahead of them in the queue, which is typical for FIFO queues. On the other hand, as $n_{e,\max}^{\text{transfer}}$ increases, the significance of the FIFO access queue diminishes and the system performance is more and more determined by the PS transfer queue. Based on the observations and analysis in Sections 4 and 5 we know that the reverse impact of η_e on $\mathbf{T}_e$ then applies. The numerical results in Figure 11 demonstrate the expected reversal of the ordering of the $\mathbf{T}_e$ curves as $n_{e,\max}^{\text{transfer}}$ is raised from 1 to ∞. We claim that in a well-dimensioned network, the access time is relatively small compared to the transfer time, corresponding with a relatively large number of service positions $n_{e,\max}^{\text{transfer}}$ (with respect to the typical occupation of the access queue), so that the 'PS effect' dominates: the greater the elastic call size variability, the better the QOS.



(Weibull) (Pareto)

Figure 10: *Numerical results of experiment 5f.*



(Weibull) (Pareto)

Figure 11: *Numerical results of experiment 5v.*

The above remark regarding the finiteness of the expected elastic call holding time of the Pareto cases with $\eta_e = \infty$ also applies here. Observe further that there is no generally uniform $n_{e,\max}^{\text{transfer}}$ where the ordering is reversed, although the curves corresponding to $\eta_e \leq 2$ do appear to jointly cross one another at about $n_{e,\max}^{\text{transfer}} = 6$ (both distributions).

Finally, comparing the figures of experiments 5f and 5v, we observe that for all depicted cases the QOS under varying capacity is worse than under fixed capacity, while for small (large) $n_{e,\max}^{\text{transfer}}$ the absolute difference increases (decreases) in the elastic call size variability. In particular, in the extreme case of $n_{e,\max}^{\text{transfer}} = 1$ (FIFO) the absolute QOS differences between the varying and fixed capacity scenarios worsens as the elastic call sizes become more variable, while in the extreme case of $n_{e,\max}^{\text{transfer}} = \infty$ (PS) the reverse effect is observed.

7 Concluding remarks

This paper reports an investigation into the impact of the elastic call size distribution on the experienced QOS in a Processor Sharing (PS) system with $n_{e,\max}$ service positions, whose capacity varies in time due to the arrival and termination of prioritized stream calls. The remarkable observation that the (conditional) expected holding time of a call is inversely proportional to the degree of variability of the elastic call size, is demonstrated by means of a series of simulation charts. The QOS disparities between the different elastic call size distributions that were considered are most significant if the elastic traffic load ρ_e is high (given $n_{e,\max} = \infty$) and if the time-scale at which the stream calls arrive and terminate is relatively large (large λ_s, μ_s). The disparities fade away as $\rho_e \longrightarrow \infty$ (given $n_{e,\max} < \infty$) and as $\lambda_s, \mu_s \longrightarrow \infty$ (keeping $\rho_s \equiv \lambda_s/\mu_s$ fixed), with the (conditional) expected elastic call holding time converging to that experienced in an $M/G/1/PS$ queue whose fixed capacity is equal to the average remaining capacity in the integrated services model. Valuable insight into the validity of the presented observation is provided by means of an analytical treatment of extreme cases.

In light of the fact that the QOS in a FIFO system degrades under a greater elastic call size variability, while it is insensitive (fixed capacity) or even improves (varying capacity) in a PS system, we have explicitly studied the trade-off between the FIFO and PS service disciplines in a extended system model. This model extends the basic model by queueing rather than rejecting elastic calls that find all service positions occupied upon arrival, in an infinite FIFO access queue. We have observed that in the extended model the impact of the elastic call size variability strongly depends on the number of service positions $n_{e,\max}^{\text{transfer}}$ in the transfer queue. In particular, for small $n_{e,\max}^{\text{transfer}}$ the FIFO access queue dominates and the QOS degrades under a greater elastic call size variability, while for large $n_{e,\max}^{\text{transfer}}$ the PS transfer queue dominates and the impact of the variability is reversed. We argue that in a well-dimensioned network the 'PS effect' typically dominates: the greater the elastic call size variability, the better the QOS.

The relevance of the principal result lies in the performance analysis of integrated services telecommunications networks, serving prioritized stream calls (e.g. voice telephony) and delay-tolerant elastic calls (e.g. WWW traffic) such as (fixed) ATM networks or (mobile) GSM/GPRS or UMTS networks. In light of the commonly acknowledged property of e.g. WWW traffic to be heavy tailed, the result indicates that assuming deterministic or lightly variable elastic call sizes, as is typically done for reasons of tractability in mathematical analysis or simulations, may lead to an overestimation of the experienced QOS. As a consequence, admission control schemes or network planning guidelines that are derived from such a model, are likely to be conservative and inefficient.

Acknowledgments

The authors would like to thank Rudesindo Núñez Queija of the Center for Mathematics and Computer Science (The Netherlands), and Hans van den Berg of KPN Research, Leidschendam (The Netherlands), for helpful discussions and comments.

References

[1] B. Avi-Itzhak and S. Halfin, "Expected response times in a non-symmetric time sharing queue with a limited number of service positions," *Proceedings of ITC 12*, pp. 5.4B.2.1-7, 1988.

[2] L. Begain, "Scalable multimedia services in GSM-based networks: an analytical approach," *Proceedings of the ITC specialist seminar on mobile systems and mobility*, Lillehammer, Norway, pp. 73-83, 2000.

[3] M. Braun, *"Differential equations and their applications"*, Springer-Verlag. New York, USA, 1983.

[4] E. Chlebus, "Empirical validation of call holding time distribution in cellular communication systems," *Proceedings of ITC 15*, Washington DC, USA, pp. 1179-88, 1997.

[5] J.W. Cohen, "Some results on regular variation in queueing and fluctuation theory," *Journal of applied probability*, vol. 10, pp. 343-53, 1973.

[6] J.W. Cohen, "The multiple phase service network with generalized processor sharing," *Acta Informatica*, vol.12, pp. 245-84, 1979.

[7] M. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM transactions on networking*, vol. 5, no. 6, pp. 835-46, 1997.

[8] R. Litjens and R.J. Boucherie, "Radio resource sharing in a GSM/GPRS network," *Proceedings of the ITC specialist seminar on mobile systems and mobility*, Lillehammer, Norway, pp. 261-74, 2000.

[9] R. Litjens and R.J. Boucherie, "Quality-of-service differentiation in an integrated services GSM/GPRS network," *submitted for publication.*

[10] R. Litjens and R.J. Boucherie, " Performance analysis of fair channel sharing policies in an integrated cellular voice/data network," to appear in *Telecommunications Systems*. Preprint: `http://www.math.utwente.nl/~boucheri`.

[11] L. Massoulié and J.W. Roberts, "Arguments in favour of admission control for TCP flows," *Proceedings of ITC 16*, Edinburgh, Scotland, 1999.

[12] M.F. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Baltimore, MD: Johns Hopkins University Press, 1981.

[13] R. Núñez Queija, *"Processor-sharing models for integrated-services networks,"* Ph.D. thesis, Technische Universiteit Eindhoven, 2000.

[14] R. Núñez Queija, "Sojourn times in non-homogeneous QBD processes with processor sharing," *Stochastic Models,* vol. 17, pp. 61-92, 2001.

[15] R. Núñez Queija, J.L. van den Berg, and M.R.H. Mandjes, "Performance evaluation of strategies for integration of elastic and stream traffic," *Proceedings of ITC 16*, Edinburgh, Scotland, pp. 1039-50, 1999.

[16] A. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modelling," *IEEE/ACM transactions on networking*, vol. 3, no. 3, pp. 226-44, 1995.

[17] J.W. Roberts and L. Massoulié, "Bandwidth sharing and admission control for elastic traffic," *Proceedings of the ITC specialist seminar on Teletraffic issues related to multimedia and nomadic communications*, Yokohama, Japan, 1998.

[18] H.C. Tijms, *Stochastic modelling and analysis: a computational approach.* Chichester, England: John Wiley & Sons, 1986.

[19] A.P. Zwart, *"Queueing systems with heavy tails,"* Ph.D. thesis, Technische Universiteit Eindhoven, 2001.

[20] A.P. Zwart and O.J. Boxma, "Sojourn time asymptotics in the $M/G/1$ processor sharing queue," *Queueing systems,* vol. 35, pp. 141-166, 2000.