Queues with M/G/∞ Input: A survey and Some New Results

Dieter Fiems, Koen De Turck

Queues with $M/G/\infty$ input: a survey and some new results

Dieter Fiems and Koen De Turck

SMACS Research Group
Department of Telecommunications and Information Processing, Ghent University
St-Pietersnieuwstraat 41, 9000 Gent, Belgium
Email: {Dieter.Fiems,Koen.DeTurck}@telin.UGent.be

***Abstract*—As input traffic at various nodes in packet switched telecommunication networks typically exhibits considerable correlation, there is a continuing interest in queuing systems with correlated arrivals. The class of $M/G/\infty$ arrival models can account for various types of arrival correlation, has a two-level structure which is appealing in the context of network protocol stacks, and is characterised by only few parameters. Moreover, queues with $M/G/\infty$ input prove to be amenable for a classic queuing analysis. In view of these characteristics, it is not surprising that these models have attracted considerable attention in the past. In this paper, we survey the literature on $M/G/\infty$ type queuing systems, present some new results on these systems, and extend the class of $M/G/\infty$ input models, while retaining the analytic tractability of the corresponding queuing systems.**

I. Introduction

Arrival correlation significantly affects queuing performance and hence there is a continuing interest in tractable queuing models which can accurately capture the correlation in the arrival process. Arrival models of interest include, amongst others, finite state-space Markovian arrival models like the discrete batch Markovian arrival model as well as Markovian arrival models with an infinite but structured state-space. A prime example of the latter class is $M/G/\infty$ input — often referred to as train arrival or session models — which is the subject of this paper.

A continuous-time version of $M/G/\infty$ input was already investigated by Cohen in 1974 [1], while the origins of te discrete-time $M/G/\infty$ input studied here can be traced back to the works of Cox [2]. Formally, let $P = \{P_n, n \in \mathbb{Z}\}$ be a sequence of independent and identically distributed $\mathbb{N}$-valued random variables with common mean λ and ith order moment λ_i. Moreover, let $S = \{S_{n,k}, n \in \mathbb{Z}, k \in \mathbb{N}^*\}$ be a doubly-indexed sequence of independent and identically distributed $\mathbb{N}^*$-valued random variables with mean μ and ith order moment μ_i. The $M/G/\infty$-input process is then defined as follows,

$$A_n = \sum_{k=0}^{\infty} \sum_{\ell=1}^{P_{n-k}} \mathbb{1}_{\{S_{n-k,\ell} > k\}} \,. \tag{1}$$

Here $\mathbb{1}_{\{x\}}$ is the indicator function which evaluates to 1 if x is true and to 0 otherwise. The sequence $A = \{A_n, n \in \mathbb{Z}\}$ is stationary ergodic and corresponds to the queue size process of a discrete-time $M/G/\infty$ with arrival process P and service process S. In literature, P is sometimes assumed to be a sequence of Poisson distributed random variables. However, this restriction will not be imposed unless explicitly stated in the remainder.

The two-level structure of $M/G/\infty$ input is a natural abstraction of an upper and a lower layer in a network protocol stack. At the upper layer, a connection or session is established which continues for a certain duration. The session lengths correspond to the service times of the $M/G/\infty$ queue. During the session, packets are then transmitted by the lower layer in the protocol stack. This natural abstraction is not the only reason $M/G/\infty$ traffic has attracted much attention. Equally important is the fact that $M/G/\infty$ processes are versatile in capturing correlation in the arrival process while they are characterised by only a few parameters: the mean arrival rate λ (assuming Poisson arrivals) and the session length distribution $S(n) = \Pr[S_{0,1} \leq n]$, see [2]–[4]. Finally, we mention the analytical tractability of single server queues with $M/G/\infty$ input as an additional motivation for studying these arrival processes.

The remainder of this paper is organised as follows. The next section reviews the stochastic characteristics of $M/G/\infty$ input. Basic results on the single-server queue with $M/G/\infty$ input are then presented in Section III. Despite the complex state space of the $M/G/\infty$ process, the expression of the mean queue content of this queuing system is surprisingly simple. This observation leads us to extend the class of $M/G/\infty$ processes while retaining the tractability of the mean queue content. This class of regenerative arrival processes is introduced in Section IV and includes discrete autoregressive processes [5] as well as branching arrival processes [6] in addition to $M/G/\infty$ input.

In terms of applicability, the single server model suffers from the inherent drawback that a single active session always consumes all available bandwidth. To avoid this, either sessions should not continuously produce packets or more servers need to be introduced. In the former case, the arrival process no longer satisfies the regeneration property and no simple expression for the mean queue content can be found while in the latter case, regeneration does not suffice for tractability. Therefore, the focus shifts to approximations for

various performance measures. Light-traffic approximations are considered in Section V. Finally, heavy-traffic limits are the subject of Section VI.

II. Properties of $M/G/\infty$ input

Taking expectation in (1) immediately yields,

$$\mathrm{E}[A_0] = \lambda\mu\,.$$

Analogously, the variance of A_0 can be calculated directly from (1),

$$\mathrm{var}(A_0) = \lambda\mu + (\lambda_2 - \lambda - \lambda^2)\sum_{k=0}^{\infty}(1-S(k))^2\,.$$

More generally, the covariance of the $M/G/\infty$ input is given by,

$$\begin{aligned}\mathrm{cov}(A_0, A_n) &= \lambda\,\mathrm{E}[(S_{0,1}-n)^+] \\ &\quad + (\lambda_2 - \lambda - \lambda^2)\sum_{k=0}^{\infty}(1-S(k))(1-S(k+n))\,.\end{aligned}$$

Assuming P constitutes a sequence of Poisson distributed random variables, these expressions simplify considerably. In this particular case, the distribution of A_0 is known: A_0 is Poisson-distributed with mean $\lambda\mu$. Moreover, the covariance of the $M/G/\infty$ input simplifies to

$$\mathrm{cov}(A_0, A_n) = \lambda\,\mathrm{E}[(S_{0,1}-n)^+] = \lambda\sum_{k=0}^{\infty}\Pr[S_{0,1} > n+k]\,,$$

see e.g. [3], [7].

The covariance function allows for adapting the service time distribution to fit the autocorrelation function of e.g. measurement data. For example, in [3], the service time is chosen to fit the autocorrelation of a number of video traces. $M/G/\infty$ proves to be very versatile in this respect. In fact, if the autocorrelation function $\rho(n) = \mathrm{cov}(A_0, A_n)/\,\mathrm{var}(A_0)$ is decreasing and integer convex with $1 > \rho(1)$ and $\lim_{n\to\infty}\rho(n) = 0$, then the probability mass function of the service times of the $M/G/\infty$ input can be expressed in terms of the autocorrelation function as follows [8],

$$\Pr[S_{0,1} = k] = \frac{\rho(k-1) - 2\rho(k) + \rho(k+1)}{1-\rho(1)}\,.$$

In other words, the service time distribution can be adapted to match the autocorrelation function exactly for a large class of processes.

III. The single server queue

We first focus on the discrete-time single-server queue with $M/G/\infty$ input and single-slot service times by means of classic probability generating functions techniques as investigated in [9]. The queue content at consecutive slot boundaries are related by the following Lindley recursion,

$$U_{n+1} = (U_n - 1)^+ + A_n\,, \tag{2}$$

whereby A_n is defined in (1). Obviously, (U_n, A_n) does not constitute a Markov chain as all active session lengths need to be tracked as well. Therefore, let $H_n^{(k)}$ denote the number of sessions that deliver their kth packet during slot n and let $\mathbf{H}_n$ denote the infinite-length vector with elements $H_n^{(k)}$. Then $(\mathbf{H}_n, U_n)$ constitutes a Markov chain. For $\rho = \mathrm{E}[A_0] = \lambda\mu < 1$, the chain admits a stationary solution $(\hat{\mathbf{H}}_n, \hat{U}_n)$. Let $\mathcal{P}(\mathbf{x}, z)$ denote the joint probability generating function of the marginal distribution of this solution,

$$\mathcal{P}(\mathbf{x}, z) = \mathrm{E}\left[\prod_{k=1}^{\infty} x_k^{\hat{H}_0^{(k)}} z^{\hat{U}_0}\right] = \mathrm{E}\left[\mathbf{x}^{\hat{\mathbf{H}}_0} z^{\hat{U}_0}\right]\,, \tag{3}$$

with $\mathbf{x} = [x_1, x_2, x_3, \ldots]$. This joint probability generating function then satisfies the following functional equation,

$$\mathcal{P}(\mathbf{x}, z) = \frac{P(x_1 z)}{z}\left(\mathcal{P}(\mathbf{C}(\mathbf{x}z), z) + (z-1)(1-\rho)\right)\,, \tag{4}$$

with $\mathbf{C}(\mathbf{x}) = [C_1(x_2), C_2(x_3), \ldots]$ a vector with elements,

$$C_k(x) = 1 - (1-z)\frac{1-S(k)}{1-S(k-1)}\,.$$

Iterating (4) yields an explicit expression for $\mathcal{P}$, although it contains infinite sums and products. The moments of the queue content can then be obtained by the moment generating property of probability generating functions. In particular, the mean queue content equals,

$$\mathrm{E}[\hat{U}_0] = \rho + \frac{\mu^2(\lambda_2 - \lambda) + (\mu_2 - \mu)\lambda}{1-\rho}\,. \tag{5}$$

In addition to the results above, single-server queues with $M/G/\infty$ input have been studied under the assumption of specific session lengths, e.g. deterministic session lengths [10], [11] or geometrically distributed session lengths [12], [13]. Extending $M/G/\infty$ input, the queuing model in [14] and [15] allows for arrival correlation of the sessions. A two-state Markov chain is introduced to capture the correlation of session arrivals (an interrupted batch arrival process IBP); this is a queue with $IBP/G/\infty$ input. In all references above, a first-in-first-out scheduling discipline is assumed. This is not the case in [16] where priority disciplines are investigated under the (simplifying) assumption of geometrically distributed session lengths.

Note that, despite the size of the state space of the $M/G/\infty$ process, equation (5) above is intriguingly simple and only contains the first and second order moments of the arrivals and the session lengths. This observation has sparked the investigation of the regenerative arrival processes introduced next.

IV. Regenerative arrivals

The analysis of the queuing model of the preceding section makes use of the following property. If the queue is empty, then there cannot be active sessions. Indeed, if this were the case, there would be packets present in the queue. In mathematical terms, this means that the arrival process regenerates whenever the queue is empty. Regeneration turns out to be the key property which makes the $M/G/\infty$ single-server queue analytically tractable. In fact, regeneration is the only property

required to make the mean queue content analysis tractable as shown in this section.

That is, we now investigate the single-server queue with single-slot service times for the following arrival process. $A = \{A_n, n \in \mathbb{Z}\}$ is a stationary ergodic sequence of non-negative integer random variables which adheres the following assumptions.

[A1] The arrival process regenerates when there are no arrivals. That is,

$$\begin{aligned}&\Pr[A_{k+1} = i_{k+1}, \ldots, A_{k+\ell} = i_{k+\ell} | A_k = 0, \mathcal{F}_{-\infty}^k] \\ &\quad = \Pr[A_{k+1} = i_{k+1}, \ldots, A_{k+\ell} = i_{k+\ell} | A_k = 0] \,, \quad (6)\end{aligned}$$

for $\ell \in \mathbb{N}^* = \{1, 2, \ldots\}$ and $i_j \in \mathbb{N}$ for all $j \in \mathbb{Z}$ and where $\mathcal{F}_{-\infty}^k$ is the natural filtration of the arrival process $\{A_k\}$.

[A2] Let $f_n = \Pr[A_n = 0, A_{n-1} > 0, \ldots, A_1 > 0 | A_0 = 0]$ denote the regeneration distribution. We assume the existence of its second moment,

$$\sum_{n=1}^{\infty} n^2 f_n < \infty \,.$$

The following is our main result on queuing systems with regenerative arrivals, see [17] for its proof. For stationary ergodic $\{A_k\}$ satisfying properties [A1] and [A2] and with $\mathrm{E}[A_0] < 1$, the mean steady-state queue content $\mathrm{E}[\hat{U}_0]$ is given by,

$$\begin{aligned}\mathrm{E}[\hat{U}_0] = \frac{\mathrm{E}[A_0] - \mathrm{E}[(A_0)^2]}{2(1 - \mathrm{E}[A_0])} + \sum_{m=0}^{\infty} (\mathrm{E}[A_{m+1} | A_0 = 0] - \mathrm{E}[A_0]) \\ + \sum_{m=0}^{\infty} \frac{\mathrm{E}[A_0 A_m] - \mathrm{E}[A_0]^2}{1 - \mathrm{E}[A_0]} \,. \quad (7)\end{aligned}$$

The mean queue content is finite provided the infinite sums in (7) converge.

Seemingly somewhat artificial, many arrival processes adhere to the regeneration property. It holds for the above-cited $M/G/\infty$ input, autoregressive and branching-type arrival processes [6]. However, the class of these zero-regenerative arrival processes is considerably larger as illustrated by the following two examples.

A. Train arrivals with autoregressive session arrivals

A first example is the $DAR(1)/G/\infty$ input model. In contrast to $M/G/\infty$ input, new sessions arrive in accordance with a discrete autoregressive arrival process. This arrival model is characterised by the sequence S as defined in Section I and by (1) a sequence $B = \{B_k, k \in \mathbb{Z}\}$ of iid Bernoulli distributed random variables with $\mathrm{E}[B_0] = p$ and (2) an iid sequence $\{N_k\}$ of $\mathbb{N}$-valued random variables. N_k denotes the number of new trains in slot k if $B_k = 0$. The number of arrivals in slot k is still given by (1) but now S_k is not an iid sequence. Instead, S_k is expressed in terms of B_k and N_k as follows,

$$S_{k+1} = B_k S_k + (1 - B_k) N_k \,. \quad (8)$$

By calculating $\mathrm{E}[A_0 A_n]$ and $\mathrm{E}[A_n | A_0 = 0]$ for this arrival process, the following expression for the mean queue content is found,

$$\begin{aligned}\mathrm{E}[\hat{U}_0] = \frac{\mathrm{E}[N]^2 \mu \mu_2 - \mathrm{E}[N] \mu_2}{(1 - \mathrm{E}[N] \mu} \\ + \frac{\mathrm{E}[N] \mu (1 - 2p) - \mathrm{E}[N]^2 \mu^2 (1 - 2p)}{(1 - p)(1 - \mathrm{E}[N] \mu)} \\ + \frac{(\mathrm{E}[N^2] - \mathrm{E}[N]^2) \mu^2 (1 + p)}{2(1 - p)(1 - \mathrm{E}[N] \mu)} \,. \quad (9)\end{aligned}$$

For $p = 0$, the arrival model simplifies to the train arrival model where the number of new trains is a sequence of iid random variables. Plugging $p = 0$ in (9) yields (5). In addition, assuming single slot train-lengths — this means $\mu = \mu_2 = 1$ — we obtain the single-server queuing system with discrete autoregressive arrivals of [5]. The expression of the mean queue content then simplifies to,

$$\mathrm{E}[\hat{U}_0] = \frac{\mathrm{E}[N_0](1 - 3p) + \mathrm{E}[(N_0)^2](1 + p) - 2(1 - p)\,\mathrm{E}[N_0]^2}{2(1 - \mathrm{E}[N_0])(1 - p)} \,.$$

B. Stationary ergodic arrivals during trains

As a second example we consider a train arrival model in which trains produce packets in accordance with a stationary ergodic process. As before, each train generates at least one packet such that regeneration is ensured. However, trains may produce more packets. Let $H_{k,n,\ell}$ denote the number of packets produced in the lth slot of the nth train arriving in slot k. The processes $\{H_{k,n,\ell}, \ell = 0, 1, \ldots\}$ constitute a doubly-indexed (by k and n) sequence of stationary ergodic positive random processes. Let A_k denote the number of arrivals in slot k, we then have,

$$A_k = \sum_{m=0}^{\infty} \sum_{n=1}^{P_{k-m}} \mathbb{1}_{\{S_{k-m,n} > m\}} H_{k-m,n,m} \,. \quad (10)$$

Let $\beta = \mathrm{E}[H_{0,1,1}]$, $\beta_2 = \mathrm{E}[(H_{0,1,1})^2]$ and $\kappa_n = \mathrm{E}[H_{0,1,0} H_{0,1,n}]$ characterise the arrival process during sessions and let $\rho = \lambda \mu \beta$ denote the arrival load. The mean queue content then equals,

$$\begin{aligned}\mathrm{E}[\hat{U}_0] = -\frac{1}{2} \lambda \beta \mu_2 + \frac{\rho + \lambda \sum_{m=0}^{\infty} (1 - S(m)) \sum_{n=0}^{m} \kappa_n}{1 - \rho} \\ + \frac{(\lambda_2 - \lambda) \beta^2 \mu^2 - 3 \lambda^2 \mu^2 \beta^2 - \lambda \mu \beta_2}{2(1 - \rho)} \,. \quad (11)\end{aligned}$$

Under the additional assumption that the number of packets produced in the consecutive slots is an independent sequence, the former expression further simplifies to,

$$\begin{aligned}\mathrm{E}[\hat{U}_0] = -\frac{1}{2} \lambda \beta \mu_2 + \frac{2\rho + \lambda \mu_2 \beta^2 - \rho \beta}{2(1 - \rho)} \\ + \frac{(\lambda_2 - \lambda) \beta^2 \mu^2 - 3 \lambda^2 \mu^2 \beta^2 - \lambda \mu \beta_2}{2(1 - \rho)} \,, \quad (12)\end{aligned}$$

in correspondence with results published in [13].

V. Light traffic

We now focus on approximations of performance measures in case the regeneration property does not hold. In this section, we consider approximations when the queue is subject to light traffic. Performance of heavily loaded queues is the subject of the following section.

Recalling the construction of $M/G/\infty$ input, there are two natural ways to reduce the arrival load. The reduction is either achieved by reducing the arrival rate of the sessions or of the amount of traffic produced in a session. The latter reduction can be accomplished by so-called π-thinning: every packet arrival in the original system also arrives in the system with reduced load with some probability q and is discarded otherwise.

Tsoukatos and Makowski [18] study approximations when the arrival rate of sessions is reduced. For the multi-server case, approximations for the probability that the queue is empty are obtained. Regarding the moments, a structural property is obtained for both single-server and multi-server systems is obtained, but an explicit expansion is obtained only for the single-server queue. However, the latter results have limited practical value as in this case an explicit expression is available.

Reducing the load by π-thinning is investigated in [19] under the additional assumption that the session lengths are geometrically distributed. Under this assumption and without thinning, this model was independently investigated in [20] and [21]. Let q denote the packet arrival probability and let p denote the probability that a session continues, $p = 1 - \mu^{-1}$. The following Taylor series expansions in q for the first two moments of the stationary queue content were obtained in [19],

$$\begin{aligned}\mathrm{E}[\hat{U}_0] = \frac{\lambda}{1-p}q + \left(\frac{\lambda_2}{2(1-p^2)} + \frac{2p\lambda^2 - (1-p)\lambda}{2(1+p)(1-p)^2}\right)q^2 \\ + \left(\frac{p}{2(1-p)(p^2+p+1)}\lambda_3 \right. \\ + \frac{(5p^3+3p^2+1)\lambda - p(1+2p)(1-p)^2}{2(1-p)^2(p^2+p+1)(1+p)}\lambda_2 \\ + \frac{4\,p^4 - p^3 + p^2 - 1}{2(1+p)(p^2+p+1)(1-p)^2}\lambda^2 \\ \left. + \frac{2p(1+2\,p^3)\lambda^2 - 2\,p^3(1-p)^2}{2(1+p)(p^2+p+1)(1-p)^3}\lambda\right)q^3 + O(q^4)\,, \quad (13)\end{aligned}$$

$$\begin{aligned}\mathrm{E}[\hat{U}_0^2] = \frac{\lambda}{1-p}q + \left(\frac{3\lambda_2}{2(1-p^2)} + \frac{6p\lambda^2 - 3(1-p)\lambda}{2(1+p)(1-p)^2}\right)q^2 \\ + \left(\frac{(15p^3+13p^2+2\,p+3)\lambda - 6p^4 + 6p}{2(1-p)^2(1+p)(1+p+p^2)}\lambda_2 \right. \\ - \frac{(9p+2)}{3(1-p)(p^2+p+1)}\lambda_3 \\ + \frac{2p(6p^3+2p^2+3)\lambda - 12p^5 + 15p^4 - 2p^3 + p^2 + p - 3}{2(1+p)(p^2+p+1)(1-p)^3}\lambda^2 \\ \left. - \frac{9p^3 - 2p - 2}{3(p^2+p+1)(1-p^2)}\lambda\right)q^3 + O(q^4)\,. \quad (14)\end{aligned}$$

The methodology in [19] allows for obtaining higher order expansions and expressions for higher order moments as well. However, the expressions for these expansions grow quickly in size and are therefore only practical for numerical evaluation.

VI. Other limiting results

The parsimony and versatility of the $M/G/\infty$ input model is also apparent when considering various limiting regimes. Along with fractional Brownian motion and On/Off sources, $M/G/\infty$ models have been among the most popular input processes, especially in the context of evaluating the impact of long-range dependence on buffer performance.

We summarize results from the fields of heavy-traffic and large-deviations theory [7], [18], [22], [23], [25], [26]. As a fully rigourous exposition of the results contained in these papers, would easily exceed the total length of this survey, we sacrifice rigour for ideas and intuition. The queuing model considered in this section is the early-arrival and multi-server variant of (2):

$$U_{n+1} = (U_n + A_n - c)^+.$$

Across the different methodologies, a dichotomy is observed between the cases that the input process is short-range dependent (srd) resp. long-range dependent (lrd). Recall that a process A is srd (resp. lrd) if the covariance function $v(k) \doteq \mathrm{cov}(A_0, A_k)$ is summable resp. non-summable. In [7] it is shown that

$$\sum_{k=0}^{\infty} |v(k)| = \frac{\lambda}{2}(\mu_2 + \mu), \quad (15)$$

hence the input process is srd (lrd) if and only if μ_2 is finite (infinite).

Also recall the definition of regularly varying distributions: a mass function $f(.)$ is regularly varying of index h, if and only if for all $y > 0$:

$$\frac{f(yt)}{f(t)} \to y^h, \quad t \to \infty.$$

If $h = 0$ then we call $f(.)$ slowly varying; and if $h \in [0, 1)$ then $f(.)$ is subexponentially varying.

The heavy-traffic limit considered in [18] is as follows: consider a series of queuing systems indexed by $r \in \mathbb{N}$. The rth model is fed by an input process with session arrival rate λ_r, with $\lambda_r \uparrow c/\mu$ as $r \to \infty$. Note that the distribution of the $M/G/\infty$ service times, as well as the release rate c remains the same for every system in the sequence. The heavy traffic limit is then the distribution Q such that $q_\infty^r / w_r \Rightarrow Q$, where q_∞^r is the steady-state buffer content distribution for the rth system, w_r is an appropriate scaling function and $\Rightarrow$ denotes weak convergence.

If the input process is srd, then we find a convergence to Brownian motion, with formulas than go back as far as Kingman's. If on the other hand, $S_{0,1}$ has a Pareto distribution with parameter α, $1 < \alpha < 2$, then the solution is expressed in terms of a Mittag-Leffler special function, and the tail is found to be Pareto with parameter $\alpha - 1$. We note that the

limiting process is Gaussian in the srd case, but non-Gaussian in the Pareto case.

Mandjes considers a different limit in [25]. In particular, he studies the Gaussian process that retains the first two moments of the original input process, that is, the average number of arrivals during stationarity $\mathrm{E}[A_0]$ and the covariance function. It is then shown that the Pareto case yields a logarithmic many-sources asymptotic the form $\frac{1}{n}\log p_n(b,c) \to O(b^{\alpha-1})$, provided that the buffer level is also large. This suggests Weibullian instead of Pareto tails for the buffer content, which is another evidence of the fact that the manner in which a limit of a series of stochastic processes is attained is very important.

Direct large-deviations analyses (without first transforming the input process into a Gaussian form) have been undertaken as well. For example, Parulekar and Makowski [7] obtain large-buffer asymptotics through large-deviation theory, whereas Duffield [28] and Mandjes [26] obtain many-sources asymptotics. It is observed that srd inputs invariably lead to exponential tails, whereas with lrd crucial differences are caused by the shape of the session duration distribution. If session durations are subexponentially varying, then the input rate during overflow will significantly exceed the drain rate c, and the time to overflow is proportional to the buffer level. In contrast, with slowly-varying inputs, the buffer will fill very slowly, at a rate only slightly higher than c. Also note that for srd $M/G/\infty$ inputs, the machinery developed for logarithmic asymptotics for general single-server queues with light-tailed arrival processes suffices (as developed for example in [27]).

Another strand of research [22]–[24] constitutes the analysis of fluid queues fed by $M/G/\infty$ input. Each active session causes the queue level to rise at a fixed rate. A number of observations can be made from these papers; the first being a confirmation of the fact that light-tailed session lengths lead to exponential tails while heavy-tailed service times lead to subexponential tails. In [24], upper and lower bounds are obtained, leading to exact asymptotics for Pareto-distributed session lengths if a certain peak-rate condition is fulfilled.

In [22], the more general but interesting situation of several heterogeneous $M/G/\infty$-processes is considered, each with different session rates and session durations, some of which are heavy-tailed. The typical configuration of long sessions that causes overflow (the path to overflow) is identified through an integer program, which paves the way for the exact asymptotics of the workload behavior.

REFERENCES

[1] J. W. Cohen, "Superimposed renewal processes and storage with gradual input," *Stochastic Processes and their Applications*, vol. 2, no. 1, pp. 31–57, 1974.

[2] D. Cox, *Statistics: an appraisal*, chapter "Long-range dependence: a review." Iowa State University Press, 1984.

[3] M. Krunz and A. Makowski, "Modeling video traffic using $M|G|\infty$ input processes: a compromise between Markovian and LRD models," *IEEE journal on Selected Areas in Communications*, vol. 16, no. 5, p. 1998, 1998.

[4] B. D'Auria and S. Resnick, "Data network models of burstiness," *Advances in Applied Probability*, vol. 38, no. 2, pp. 373–404, 2006.

[5] F. Kamoun, "The discrete-time queue with autoregressive inputs revisited," *Queueing Systems*, vol. 54, pp. 185–192, 2006.

[6] D. Fiems, J. Walraevens, and H. Bruneel, "Queues with Galton-Watson-type arrivals," in *Proceedings of the Belarussian Winter Workshop on Queueing Theory*, (Minsk, Belarussia), 2009.

[7] M. Parulekar and A. Makowski, "Tail probabilities for $M/G/\infty$ input processes (I): Preliminary asymptotics," *Queueing Systems*, vol. 27, no. 3–4, pp. 271–296, 1997.

[8] W. Poon and K. Lo, "A refined version of $M/G/\infty$ processes for modelling VBR video traffic," *Computer Communications*, vol. 24, no. 11, pp. 1105–1114, 2001.

[9] S. Wittevrongel and H. Bruneel, "Correlation effects in ATM queues due to data format conversions," *Performance Evaluation*, vol. 32, no. 1, pp. 35–56, 1998.

[10] Y. Xiong and H. Bruneel, "Buffer contents and delay for statistical multiplexers with fixed-length packet-train arrivals," *Performance Evaluation*, vol. 17, no. 1, pp. 31-42, 1993.

[11] F. Kamoun, "Performance analysis of a discrete-time queuing system with a correlated train arrival process," *Performance Evaluation*, vol. 63, no. 4–5, pp. 315–340, 2006.

[12] H. Bruneel and I. Bruyland, "Performance study of statistical multiplexing in case of slow message generation," in: Proceedings of the IEEE International Conference on Communications, ICC '89, pages 951–955, Boston, June 1989.

[13] B. Feyaerts, S. De Vuyst, S. Wittevrongel, and H. Bruneel, "Session delay in file server output buffers with general session lengths," in *IEEE International Conference on Communications (ICC 2010)*, (Cape Town, South Africa), May 2010.

[14] S. De Vuyst, S. Wittevrongel, and H. Bruneel, "Statistical multiplexing of correlated variable-length packet trains: an analytic performance study," *Journal of the Operational Research Society*, vol. 52, no. 3, pp. 318–327, 2001.

[15] S. De Vuyst, S. Wittevrongel, and H. Bruneel, "Mean value and tail distribution of the message delay in statistical multiplexers with correlated train arrivals." *Performance Evaluation*, vol. 48, no. 1–4, pp. 103–129.

[16] J. Walraevens, S. Wittevrongel, and H. Bruneel, "A discrete-time priority queue with train arrivals," *Stochastic Models*, vol. 23, no. 3, pp. 489–512, 2007.

[17] D. Fiems and K. De Turck, "Mean queue content of discrete-time queues with zero-regenerative arrivals," technical report, Ghent University, 2011.

[18] K. Tsoukatos and A. Makowski, "Power-law vs exponential queueing in a network traffic model," *Performance Evalution*, vol. 65, no. 1, pp. 32–50, 2008.

[19] K. De Turck, D. Fiems, S. Wittevrongel, and H. Bruneel, "A Taylor series expansions approach to queues with train arrivals," in *Proceedings of Valuetools 2011*, (Cachan, France), May 2011.

[20] H. Bruneel and I. Bruyland, "Performance study of statistical multiplexing in case of slow message generation," in *Proceedings of the IEEE International Conference on Communications, ICC '89*, (Boston), pp. 951–955, June 1989.

[21] A. Brandt, M. Brandt, and H. Sulanke, "A single server model for packet-wise transmission of messages," *Queueing Systems*, vol. 6, pp. 287–310, 1990.

[22] Sem Borst and Bert Zwart, "Fluid Queues with Heavy-Tailed $M/G/\infty$ input," Mathematics of Operations Research, vol. 30, no. 4, pp. 852–879, 2005.

[23] Sidney Resnick and Gennady Samorodnitsky, "Steady-state distribution of the buffer content for $M/G/\infty$ input fluid queues," Bernoulli vol. 7, no. 2 (2001), pp. 191-210.

[24] Nikolay Likhanov, "Bounds on the buffer occupancy probability with self-similar input traffic," in: Self-Similar Network. Traffic and Performance Evaluation, Wiley, New York, pp. 193–214, 2000.

[25] M. Mandjes, "Large deviations for Gaussian queues: modelling communication networks," Wiley and Sons, New York, 2007.

[26] M. Mandjes, "A note on queues with $M/G/\infty$ input," *Operations Research Letters*, vol. 28, no. 5, pp. 233–242, 2001.

[27] N.G. Duffield and Neil O'Connell, "Large deviations and overflow probabilities for the general single server queue, with applications," Mathematical Proceedings of the Cambridge Philosophical Society, vol. 118, pp. 363–374, 1995.

[28] N. Duffield, N.G., "Queueing at large resources driven by long-tailed $M/G/\infty$-modulated processes," *Queueing Systems*, vol. 28 no. 1, pp. 245–266, 1998.