A Storage Model with High Rate and Long Range Dependent On/Off Sources.

A.Simonian, D.Veitch ¹

¹ FT-CNET, 38-40 rue du Général Leclerc, 92131 Issy-les-Moulineaux, France

Abstract

We consider a fluid queueing problem with an infinite reservoir with output rate $C$, fed by an infinite superposition of identical On/Off sources with large instantaneous bit rate $h \geq C$. The On periods have infinite variance, generating Long Range Dependence. It is shown that the distribution of the stationary queue content at embedded time points is heavy tailed with infinite mean. It is explained why this tail behaviour is found rather than Weibullian behaviour which can also arise with heavy On/Off sources. The equilibrium distribution of the maximum queue content during a busy period is shown to be heavy tailed with infinite variance.

Keywords: On/Off Fluid Sources, Heavy Traffic, Long Range Dependence, Power law queues.

1 Introduction

The classical queueing literature is concerned essentially with systems with exponentially decaying covariance (short term dependence) in their arrival processes. This gives rise to distributions of queue content or virtual waiting time with tails which are asymptotically exponential. Recent measurements (Leyland et al. [13] and references therein, Paxson and Floyd [17]) have shown clearly that high speed traffic can contradict this basic assumption, exhibiting Long Range Dependence (LRD) and burstiness over an extremely wide range of timescales. Queueing experiments using real traffic traces (Erramilli et al. [9]) demonstrate that this long range dependence does indeed impact significantly on queueing delays. Recent studies of Norros [15, 16], and Brichet et al. [2], have shown that the presence of LRD in arrival processes does generate non exponential queueing behaviour in certain contexts, with serious implications for queueing delays and losses.

We define LRD in the context of covariant-stationary processes $\{\Lambda_t\}$, representing the arrival rate at time $t$. LRD corresponds to a divergent correlation integral $\int_{0}^{\infty} r(\tau) d\tau$, where $r$ is the correlation function of $\{\Lambda_t\}$. This is essentially equivalent (Cox, [6]) to an ‘unusually’ large asymptotic variance of the corresponding counting (work) process $\mathcal{W}_t = \int_{0}^{t} \Lambda_s ds$, namely $\text{Var}(\mathcal{W}_t) \sim t^{2H}$ for large $t$ and some $H > 1/2$. It is not obvious what the effect of traffic with LRD may be on queueing performance. Intuition in the subject is highly tuned to short
term, Markovian thinking which can be very misleading. Consider a renewal process where the interarrival-time distribution has finite mean but infinite variance. It is not difficult to show [18] that such a process is long range dependent. The standard $GI/M/m$ queueing analysis applies however, where new arrivals find a system containing an exponentially distributed amount of work. Thus LRD alone does not suffice to generate abnormal queueing behaviour. On the other hand, studies by Norros [16] and Brichet et al. [2] show that queueing models with long range dependence in the work process can generate Weibullian tail behaviour, which implies buffer sizes which grow much faster with load than in the classical case.

In this paper we analyze a fluid or storage (see [15]) queueing system with LRD input. Fluid systems have been used before (Brichet et al. [2]), Bensaou et al. [1] and Guibert [11]), to model bursty traffic in ATM networks over time scales where the granularity of the ATM cells, and the quasi deterministic nature of their arrivals, no longer dominate. We consider On/Off sources, that is sources with mutually independent, alternating silence periods with no work arriving, and activity periods where work arriving at a constant rate $h$. We consider a superposition of $N$ identical, independent sources of this type, which flow into an infinite reservoir with a fixed leak rate of $C = 1$. The object of study is the complementary distribution function $Q(.)$ of the stationary queue content (corresponding to the fluid level in the reservoir).

On/Off sources with LRD input were considered by Brichet et al. in [2]. In that paper, as sources were multiplexed together the capacity of the system was raised appropriately to preserve stability. This choice of normalization leads necessarily to sources whose peak rate is very small in comparison with the capacity. This fact is intimately associated with the Weibullian form of the queue (at high load) found for that case. Another, very different choice of normalization is to leave the peak rate and the capacity fixed, but to increase the mean length of the off periods as sources are added. As discussed in [1] (see also [4]), in the limit of a large number of sources, this leads to an aggregate source model where new bursts arrive according a Poisson process. Physically the picture is of large set of high rate sources, each of which emit a single burst of highly variable volume. This is a reasonable model for some of the most important components of future network traffic, for example many independent multi-media ‘World Wide Web’ type connections multiplexed onto a high capacity link. The sizes of documents accessible by the Web are known to be heavy tailed [7].

As developed in the following, such a storage model has points in common with a $M/G/1$ model where heavy bursts correspond in some sense to service times with infinite second moments. In the case of the $M/G/1$ queue [2], such ‘heavy’ service times generate a power law tail for $Q(x)$ as $x$ grows to infinity and, in particular, imply infinite mean waiting times. We are interested in seeing whether such a result is also true of the On/Off storage model. It is certainly reasonable to expect queues with tails that are at least Weibullian, since now bursts of very long duration have in addition a large rate. Power law tails have been found numerically by Erramilli et al. [8] for a queueing system modelled by certain chaotic mappings
of the plane, where the deterministically generated traffic is essentially On/Off-like with weak correlations between neighbouring silence and active periods.

General expressions for $Q(x)$ or for bounds on $Q(x)$ are given in [2] for both $h < 1$ and $h \geq 1$. In Cohen [4], an exact expression for the Laplace transform of the queue content $V$ at certain time instants is given for the analytically convenient case $h = 1$. The inclusion of long range dependence however makes these expressions extremely difficult to evaluate. We succeed in doing so, however by expanding the expression derived in Cohen [4] mentioned above in the case of an infinite number of sources. We also extend this expression to include the case $h > 1$. We then find that the distribution function $V(\cdot)$ of $V$ is indeed heavy tailed with infinite mean for $h \geq 1$. We also calculate the maximum queue content during a busy cycle and find that it has a tail behaviour similar to that of the burst duration, and hence has infinite variance. Note that $h > 1$ is physically relevant, for example in the case of a router or bridge connecting two networks with unequal access rates. Traffic passing in the fast-slow direction corresponds to a rate $h > 1$.

We do not consider the case of $N$ finite here. Similar results are expected to hold, but the calculations are more cumbersome. The case $N = 1$ is particularly simple however and we can make the following comments. As shown in the following sections, the key problem in our approach is to determine the tail properties of the random variable $G$ which is in general a complicated function of $B$ (see (4.5)). For $N = 1$ however we have simply $G = (h - 1)B$. Note that for $h = 1$ the queue is always empty. This simplicity in the case of a single source allows generalisations beyond On/Off sources to be considered. Using $G/G/1$ theory, Choudhury and Whitt [3] consider the case of a single source with several possible flow rates, some of which are below $C = 1$, others which exceed it. The durations of these flow rate periods are given by mutually independent heavy-tailed random variables. A power law tail with infinite mean is again found for the queue content distribution.

The case $h < 1$ is not studied here, as the fact that the queue content can decrease even whilst sources are active invalidates the $M/G/1$-related approach. It is not clear at this point whether this sharp change in the structure of the problem for $h < 1$ results in a corresponding discontinuity in the form of the tail of the queue content distribution and/or other properties of the queueing system.

2 Superposition of On/Off sources

Consider a superposition of $N$ identical On/Off sources, each with mutually independent, alternating silence periods $A$ and activity periods $B$. We denote by $A(\cdot)$ and $B(\cdot)$ the probability distribution functions of the silence and activity periods, respectively. The peak rate of a source when active is $h$ and the queue output rate is denoted by $C = 1$. Given $E[A] = 1/\nu$ and $E[B] = 1/\mu$, the total load is $\rho = Nh\nu/(\nu + \mu)$ and we require that $\rho < 1$ so that the
queue has a stationary regime.

Consider the particular limiting form of the above superposition where \( \nu = \lambda / N \), \( \lambda \) fixed, and let \( N \) tend to infinity. The load \( \rho \) tends to \( \rho = h \lambda / \mu \) and the burst arrival times tend to arrival epochs of a Poisson process with rate \( \lambda \). Each of the original sources now contributes just a single burst. The number of active sources at an arbitrary instant is given by a Poisson variable with parameter \( \lambda / \mu \). Thus the probability that no source is active remains appreciable, in sharp contrast to the alternative normalization \( h \mapsto h / N \) of Bensaou et al. [1], where it tends to zero. This is essential in the case \( h \geq 1 \), where the storage level of a non-empty reservoir can decrease only when all sources are silent.

In the superposition process, a ‘silence’ is a period when no source is active. We denote its duration by the random variable \( S \). Following Cohen [4], we call a ‘continuous inflow period’ an interval between the end of one silence and the beginning of the next. Denote by \( D \) the random variable describing the duration of a continuous inflow period, and \( W \) the normalized work arriving in this period, that is work/h. Thus the superposed process consists of alternating silences and continuous inflow periods which are in general dependent.

Now, introduce long range dependence into the source model. A natural way to do this is to assume an asymptotically heavy tail for the distribution of the activity period for each source, that is for large \( t \)

\[
1 - B(t) \sim g(t)t^{-\beta}, \quad 1 < \beta \leq 2
\]  

(2.1)

where \( g \) is a function with slow variation (see Feller [10, p.268]). The range of \( \beta \) chosen ensures a finite expectation but infinite variance. To eliminate ‘oscillatory’ behaviour in the tail, we assume that the derivative \( g' \) satisfies \( tg'(t) = o(g(t)) \) for large \( t \). This implies that the density \( b(.) \) associated with distribution function \( B(.) \) exists and verifies \( b(t) \sim \beta g(t)t^{-\beta - 1} \), which is decreasing for large \( t \). It can then be shown [2] that under these conditions the Laplace transform \( B^* : s > 0 \mapsto E(e^{-sB}) \) of \( B \) takes the form

\[
B^*(s) = 1 - \frac{s}{\mu} + \frac{\Gamma(2 - \beta)}{\beta - 1} g \left( \frac{1}{s} \right) s^\beta + o \left( g \left( \frac{1}{s} \right) s^\beta \right)
\]

(2.2)

where \( \Gamma \) is Euler’s function. We now make a specific choice for \( B(.) \) which will simplify our calculations greatly, although we expect that the results will be essentially the same for any variable satisfying (2.1). We use the classical Pareto law with density

\[
b(t) = \begin{cases} 
0, & t < T \\
\beta T^\beta / t^{1+\beta}, & t \geq T
\end{cases}
\]

(2.3)

where \( T \) is a positive parameter. Here \( g \) is the constant \( T^\beta \) and the mean burst duration \( E[B] \) equals \( 1/\mu = \beta T / (\beta - 1) \).
3 Connection to an $M/G/1$ queue.

It has been noted before [1], [4], that for $h \geq 1$ the output processes of the On/Off storage system and an associated $G/G/1$ system with service distribution $B(.)$ are identical. Thus the busy period distributions for the two systems are the same. It is not clear, however, how useful the busy period is for the determination of delay and loss for individual connections. To infer a connection between the respective queue sizes is more useful for this purpose but requires more care.

To use the known results of the $M/G/1$ queue, we need the silences to be exponentially distributed. This could be achieved by letting the Off period $A$ have an exponential distribution, or by considering the limiting model described above. In either case, it is not difficult to see that the distribution $S$ is exponential with parameter $\lambda$, and silences and inflow periods are mutually independent. The essential difference now between the storage and $M/G/1$ systems is that in the latter, the work brought in a burst arrives instantaneously, whereas in the former bursts are spread out and take time to empty fully into the reservoir. During silences, however, this distinction vanishes as there are no partially arrived bursts in the On/Off input. In particular, at the set of epochs defined by the end of silences, the work level in the reservoir is precisely the equilibrium waiting time in the $M/G/1$ system where the service time is given by $G = hW - D$, and the Poisson arrival process has parameter $\lambda$. In fact, if $V_n$ is the storage level at the beginning of the $n$th continuous inflow period then the level at the $(n+1)$th is given by

\[ V_{n+1} = [V_n + hW_n - (D_n + S_n)]^+ \]

\[ = [V_n + (hW_n - D_n) + S_n]^+ \]  
(3.1)

from which the above correspondence follows.

To complete the analysis, we require only a knowledge of the variable $G = hW - D$, more specifically its Laplace Transform, before employing the Pollaczek-Khintchine formula [12] for the Laplace transform of the stationary distribution of the variables $V_n$.

In examining the queueing system at the end of silence periods we consider a conditional distribution $V(.)$ and not the distribution $Q(.)$ at an arbitrary instant. We believe however that $V(.)$ is strongly characteristic of the stationary behaviour of the system. In addition, once $V(.)$ is known the equilibrium distribution of a dynamic quantity of interest, the maximum queue size during a busy cycle, can be calculated easily using known results.

4 Behaviour of ‘service time’ $G$.

In view of (3.1), the ‘service time’ $G = hW - D$ is of primary interest. To gain insight into the distribution of this variable, we first make the following observations. By taking the simplest
possible inflow period, that consisting of just a single burst, we note that \( P\{W > t\} \geq P\{D > t\} \geq P\{B > t\} \) for all \( t \). We can immediately deduce that if variables \( W \) and \( D \) have heavy tailed distributions with respective indices \( \gamma_w, \gamma_d \), then \( 1 < \gamma_w \leq \gamma_d \leq \beta \leq 2 \) (recall that the first moments exist). It is interesting to note that \( W - D = 0 \) for a single-burst inflow period, so \( G = 0 \) with positive probability if \( h = 1 \). We now give a preliminary result on the existence of the two moments of \( G \).

**Proposition 4.1** In the case when \( E(B) < \infty \) and \( E(B^2) = \infty \) corresponding to an exponent \( \beta \in ]1, 2] \), we also have \( E(G) < \infty \) and \( E(G^2) = \infty \).

**Proof:** recall the following results from Cohen [4] for the On/Off storage system with \( h = 1 \). With \( a = \lambda/\mu \), we have

\[
\lambda E[\mathcal{W}] = ae^a, \quad (4.1)
\]

and

\[
\lambda E[\mathcal{D}] = e^a - 1
\]

and

\[
E[\mathcal{W}^2] = e^a(1 + 2a)E[B^2] + a^2E[D^2],
\]

\[
E[\mathcal{W}D] = e^aE[B^2] + aE[D^2] \quad (4.2)
\]

\[
e^aE[B^2] \leq E[D^2] \leq \frac{e^a(e^a - 1)}{a}E[B^2].
\]

Hence for \( G = hW - D \) with \( h \geq 1 \), we have

\[
\lambda E[hW - D] = 1 - (1 - ha)e^a, \quad (4.3)
\]

\[
E[(hW - D)^2] = (1 - ha)^2E[D^2] - e^a(1 - 2ah^2 -(h - 1)^2)E[B^2],
\]

\[
\geq e^a(h(a + 1) - 1)^2E[B^2] \geq a^2E[B^2]. \quad (4.4)
\]

From (4.4) with \( E[B^2] = \infty \), we conclude that \( E[(hW - D)^2] \) is infinite.

This result is inconsistent with an exponential or Weibullian tail for the distribution \( G(\cdot) \) of \( G \), but consistent with the form of (2.1) for some exponent \( \gamma \). By definition, we clearly have \( \gamma_w \leq \gamma \leq 2 \). For \( h > 1 \), we also have \( \gamma_w \leq \gamma \leq \gamma_d \) since \( P\{hW - D > t\} \geq P\{(h - 1)D > t\} = P\{D > t/(h - 1)\} \). To actually find \( \gamma \) and to determine other properties of \( G(\cdot) \), we use the following expression for its Laplace Transform \( G^* \).

**Proposition 4.2** For \( \Re(s) > 0 \) and \( \rho = h\lambda E[B] < 1 \), we have

\[
\frac{s - h\lambda(1 - B^*(s))}{s - h\lambda(1 - G^*(s/h))} = 1 - \lambda \int_0^{+\infty} R(s, t)e^{-\lambda t + \lambda M(s, t)}dt \quad (4.5)
\]

with

\[
R(s, t) = E[e^{-s(B - t/h)}(B \geq t)], \quad M(s, t) = E[t - B]e^{-sB}(B \leq t). \quad (4.6)
\]
The derivation of this result is given in the Appendix beginning with results from Cohen [4]. As shown in the following section, a classical Tauberian theorem applied to the Laplace transform \(G^*\) enables the specification of the tail behaviour of \(G(\cdot)\) along with that of the queue length distribution \(V(\cdot)\). For this purpose it turns out that the expansion in \(s\) of (4.5) for small \(s\) need only be taken up to first order. This expansion is calculated in the following technical lemma.

**Proposition 4.3** Expression (4.5) can be expanded as

\[
\frac{s - h\lambda(1 - B^*(s))}{s - h\lambda(1 - G^*(s/h))} = e^{-a} + O(s)
\]  

(4.7)

for small positive \(s\).

The difficulty in the estimation of (4.5) lies in the product terms of the form \(t \times s\) in the integrand. Expansions for small \(s\) cannot be taken directly since for large \(t\) the product \(st\) is not negligible. Note that expressions derived from the Benes approach in [1] lead to similar terms.

**Proof:** first note that \(M(s,t) = 0\) and \(R(s,t) = e^{st/h}B^*(s)\) for \(t \leq T\) as the Pareto density vanishes over \([0,T]\). Hence

\[
\int_0^{+\infty} R(s,t)e^{-\lambda t + \lambda M(s,t)} dt = \int_0^T e^{st/h}B^*(s)e^{-\lambda t} dt + \int_T^{+\infty} R(s,t)e^{-\lambda t + \lambda M(s,t)} dt
\]

\[
= B^*(s) \frac{e^{(s/h - \lambda)T} - 1}{s/h - \lambda} + \int_T^{+\infty} R(s,t)e^{-\lambda t + \lambda M(s,t)} dt
\]  

(4.8)

Now consider the terms of the integrand in (4.8), beginning with \(R(s,t)\). From (4.6) for \(t > T\) and using expression (2.3) for the density \(b(\cdot)\) of \(B\), we have

\[
R(s,t) = \int_t^{+\infty} e^{-s(u - t/h)} dB(u)
\]

\[
= \int_t^{+\infty} e^{-s(u - t/h)} \frac{\beta T^\beta}{u^{\beta+1}} du = \beta T^\beta s^\beta e^{st/h} \Gamma(-\beta, st)
\]  

(4.9)

where \(\Gamma(-\beta, \cdot)\) denotes the complementary incomplete Gamma function with non integer parameter \(-\beta\). Turning to \(M(s,t)\), we first expand \(e^{-stB}\), reverse the order of summation and integration and integrate term by term, giving

\[
M(s,t) = \int_0^t (t - u)e^{-su} dB(u) = \int_0^t (t - u)e^{-su}\beta T^\beta \frac{du}{u^{\beta+1}}
\]

\[
= \beta T^\beta \int_0^t \frac{t - u}{u^{\beta+1}} \sum_{k=0}^{+\infty} (-1)^k s^k u^k \frac{du}{k!}
\]

\[
= \beta \sum_{k=0}^{+\infty} (-1)^k s^k \left[ \frac{T^{k+1}}{k + 1 - \beta} + t \frac{T^k}{\beta - k} + \frac{T^k}{(\beta - k)(\beta - k - 1)} \right].
\]
It is important to note that this expansion is regular in s. Recalling that $1/\mu = \beta T/(\beta - 1)$, the important low order terms $t$, $-1/\mu$ and $-ts/\mu$ can readily be identified and the others regrouped to obtain finally

$$
- \lambda t + \lambda M(s, t) = -[a + \eta_1(s)] - st[a + \eta_2(s)] + s^{\beta-1} \varphi(st)
$$

(4.10)

where $a = \lambda/\mu$ and

$$
\begin{align*}
\eta_1(s) &= \lambda \beta \sum_{k=1}^{+\infty} (-1)^{k-1} \frac{s^k}{k!} \frac{T^{k+1}}{k + 1 - \beta}, \\
\eta_2(s) &= \lambda \beta \sum_{k=1}^{+\infty} (-1)^{k-1} \frac{s^k}{(k + 1)!} \frac{T^{k+1}}{k + 1 - \beta}, \\
\varphi(z) &= \lambda \beta T^\beta \sum_{k=0}^{+\infty} \frac{(-1)^k}{k!} \frac{1}{(\beta - k)(\beta - k - 1)} z^{k+1-\beta}.
\end{align*}
$$

(4.11)

Performing the variable change $z = st$, substituting (4.9) and (4.10) into the integral in (4.8) yields

$$
\int_T^{+\infty} R(s, t)e^{-\lambda t + \lambda M(s, t)} dt = \beta T^\beta s^{\beta-1} e^{-a - \eta_1(s)} \int_{st}^{+\infty} e^{z/h} \Gamma(-\beta, z) e^{-z(a + \eta_2(s))} e^{s^{\beta-1} \varphi(z)} dz.
$$

(4.12)

Denote the right hand side of (4.12) by $\Theta(s)$ and consider its expansion about $s = 0$. Since $\eta_1(s)$ and $\eta_2(s)$ are clearly $O(s)$ for small $s$, for clarity of exposition we first assume that they are identically zero. Expanding the exponential $e^{s^{\beta-1} \varphi(z)}$, we have

$$
\Theta(s) = \beta T^\beta e^{-a} \sum_{j=0}^{+\infty} \frac{s^{(j+1)(\beta-1)}}{j!} K_j(s)
$$

(4.13)

where

$$
K_j(s) = \int_{st}^{+\infty} e^{-az} \varphi(z)^j e^{z/h} \Gamma(-\beta, z) dz.
$$

From definition (4.11), the derivative $\varphi'$ satisfies $\varphi'(z) = \lambda \beta T^\beta \Gamma(-\beta, z)$ which tends to $\beta \lambda T^\beta \Gamma(-\beta)$ when $z \to +\infty$ (see [14, pp.338-9]) and thus $\varphi(z) = O(z)$ for large positive $z$. It can now be seen that the $K_j(s)$ are finite since $e^{z/h} \Gamma(-\beta, z) \leq e^z \Gamma(-\beta, z) = O(z^{\beta-1})$ for large $z$. Now, from the definition we have $\varphi(z) \sim Az^{1-\beta}$ for small $z$, where we set $A = \lambda T^\beta/(\beta - 1)$. Using $\Gamma(-\beta, z) = z^{-\beta}/\beta - z^{-(\beta+1)}/(\beta - 1) + O(1)$, for each $j \geq 0$ we obtain $e^{-az} \varphi(z)^j e^{z/h} \Gamma(-\beta, z) \sim A^j z^{-j(j+1)/}\beta$ with $j - (j + 1)\beta < -1$ since $\beta > 1$. Each $K_j(s)$ thus
tends to $+\infty$ as $s \downarrow 0$ and is asymptotic to

$$K_j(s) \sim \frac{A_j}{\beta} \int_{sT}^{\infty} z^{j-(\beta-1)}dz \sim \frac{A_j}{(j+1)\beta(\beta-1)}(sT)^{j-(\beta-1)}. \quad (4.14)$$

Substituting (4.14) into (4.13), we thus conclude that $\Theta(s) \to \Theta(0)$ as $s \downarrow 0$ with

$$\Theta(0) = \beta T^\beta e^{-a} \sum_{j=0}^{+\infty} \frac{A_j}{j! (j+1)^{\beta}(\beta-1)} T^{j-(\beta-1)},$$

which simply sums to

$$\Theta(0) = \frac{e^{-a}}{\lambda} \left[ e^{\lambda T^\beta} - 1 \right] = \frac{1}{\lambda} \left[ e^{-\lambda T} - e^{-a} \right]. \quad (4.15)$$

This term substituted into (4.8) with $s = 0$ yields $e^{-a}$ as claimed.

Similarly, at further order in $z$ it can be shown that

$$e^{-az} \varphi(z) = \frac{A_j}{\beta} z^{j-(\beta-1)} \Gamma(-\beta, z) = \frac{A_j}{\beta} \cdot z^{j-(\beta-1)} \left[ \Gamma(-\beta) + z \left( a + 1 - h + \frac{1}{\beta-1} + \frac{j \beta}{\beta-2} \right) + o(z) \right].$$

After integrating the second term of this expansion over $[sT, z_0]$ for any fixed $z_0$ and multiplying by $s^{(\beta-1)}/(\beta-1)$, we obtain a term of order $O(sT)$ for all $j$ in (4.13). It is easily verified that this remains true when we reinstate $\eta(s) = O(s)$, $\eta_2(s) = O(s)$ and $e^{-a-\eta(s)} = e^{-a}(1 + O(s))$ in (4.12), so we conclude that $\Theta(s) = \Theta(0) + O(s)$. Recalling that $\Theta(s)$ is the integral in (4.8), the proof is complete.

5 Applications to the Queueing Problem

The Pollaczek-Khintchine formula for the Laplace transform of the waiting time $\mathcal{V}^*$ of the M/G/1 queue with service time $\mathcal{G}$ and arrival rate $\lambda$ reads

$$\mathcal{V}^*(s) = \frac{s(1 - \lambda \mathbb{E}[\mathcal{G}])}{s - \lambda(1 - \mathcal{G}^*(s))}$$

for $s > 0$. The denominator of this expression can be obtained from (4.7) after a change of variable $s \mapsto sh$, and using (4.3) for the expectation of $\mathcal{G}$ we have

$$\mathcal{V}^*(s) = s(1 - \rho)e^a \cdot \frac{e^{-a} + O(s)}{s - \lambda(1 - \mathcal{B}^*(sh))} \approx \frac{s(1 - \rho)}{s - \lambda(1 - \mathcal{B}^*(sh))}$$

9
for small $s$. Remarkably, this is exactly the same as the equilibrium waiting time for $M/G/1$ with service time $hB$. In addition, this shows that $G(.)$ is asymptotically very simply related to $B(.)$ and, in particular, that the exponent $\gamma$ of $G(.)$ equals $\beta$. Now, using the expansion (2.2) of $B^*(s)$ for small $s$, we obtain

$$V^*(s) = 1 - \frac{\lambda(2-\beta)(Th)^\beta}{(1-\rho)(\beta-1)}s^{\beta-1} + o(s^{\beta-1})$$ (5.1)

for small $s$. Applying a Tauberian theorem [10, p.445] to $(1 - V^*(s))/s$ while using the asymptotics (5.1) yields the power law tail

$$1 - V(x) \sim \frac{\lambda(Th)^\beta}{(1-\rho)(\beta-1)} \frac{1}{x^{\beta-1}}$$ (5.2)

for large $x$, which implies a distribution with infinite mean since $1 < \beta < 2$.

Using this result for the tail of the embedded queue, we can determine the asymptotic form of the maximum queue content during a busy cycle using the theorem from [4, p.46]. The theorem states that the maximum content $Q_{\text{max}}$ of the buffer during a busy cycle has the same distribution as that of the maximum virtual waiting time $V_{\text{max}}$ during a busy cycle of an $M/G/1$ queue with service time distribution $W - D$ and arrival rate $\lambda$. It is not difficult to see from the proof in [4] that this result holds true in the case of $h > 1$ if we replace $W - D$ by $G$. From [5, p.618] we have the following result for $V_{\text{max}}$ applied to our imbedded $M/G/1$ queue,

$$P\{V_{\text{max}} \geq x\} = \frac{1}{\lambda} \frac{d}{dx} \log(1 - V(x))$$ (5.3)

which from (5.2) for large $x$ takes the form

$$P\{V_{\text{max}} \geq x\} \sim \frac{(Th)^\beta}{1 - \rho} \frac{1}{x^{\beta}}.$$ (5.4)

In other words $Q_{\text{max}}$, like $B$, is a random variable with an index-$\beta$ power law tail, implying a finite mean (given by [5] $\frac{1}{\lambda} \log((1 - \lambda E[G])^{-1}) = -\frac{1}{\lambda}(\log(1 - \rho) + a)$, but infinite variance.

6 Conclusion

We have investigated the queueing behaviour of a storage system with a superposition of On/Off fluid sources as input with heavy On periods of infinite variance. A limiting superposition was considered where the number of sources $N$ tends to infinity, which implies a choice of normalisation to maintain the system load constant with increasing $N$. By normalising by increasing the mean silence duration with $h \geq 1$ fixed, power law tail behaviour was found for the distribution of queue content for the limiting system where each source contributes a single
burst. This is in accordance with speculation based on the $M/G/1$ system with service times of infinite variance, which is the analogous system for instantaneous arrivals. See also [8] for suggestive numerical results in a related context. In sharp contrast, normalising by reducing the peak inflow rate $h$ with increasing $N$ for each source whilst leaving the distributions for the On and Off periods unchanged leads to Weibullian tail behaviour [2]. Although the queue content distribution at a set of embedded time points was calculated rather than the stationary distribution, the fact that a power law form was found with an exponent corresponding to an infinite mean, a very strong result, leads us to expect that our results will be characteristic of the stationary results. The form of the tail of the distribution for the maximum queue content during a busy cycle was also calculated, and found to be power law with finite mean but infinite variance. Finally, the trivial $N = 1$ finite source case was treated, and results similar to the infinite source case were found to hold.

For future work it would be valuable to gain insight into the case where $h < 1$ and fixed, for example the case of a finite number of sources each with $h < 1$. This is because it is not clear whether Weibullian, power law, or some other behaviour, is to be expected for such a case. This scenario would arise in practice if sources were allowed to use a substantial proportion of available bandwidth on a multiplexed link.
Appendix.

We derive expression (4.5) of Proposition 4.2, which is a generalisation to $h \geq 1$ of Theorem C-2.10 from Cohen [4, pp.38]. Note that equation and theorem references from [4] will be preceded by a “C-”. Our notation differs substantially from Cohen’s. The correspondence Cohen $\mapsto$ here is \( \{s, \rho, \rho_0, \beta, \beta(.), \Lambda \} \mapsto \{z, s_0, B, B^*(.), \lambda \} \).

Beginning with formula C-A.4.3 from the proof of Theorem C-2.10, we apply C-2.2.9, C-2.2.12, and definition (4.6) for \( M(s,t) \) to obtain the expression

\[
(a.1) \quad \frac{-z - \lambda(1 - B^*(s))}{-z - \lambda(1 - \mathbb{E}[e^{-s(W-D)]})} = 1 - \lambda \int_0^{+\infty} \mathbb{E}[e^{-s(B^*(B \geq t))}] e^{-\lambda t + \lambda M(s,t)} dt.
\]

As shown in the paragraph following C-A.4.3, (a.1) is valid for \( \Re(z) \geq -\Re(s) \), and both sides are analytic functions of \( s \) for \( \Re(s) > s_0 \geq 0 \), where \( s_0 \) is the zero of the function \( s - \lambda(1 - \mathbb{E}[e^{-s(W-D)]}) \) with the largest real part. Lemma C-2.7 states that \( s_0 \) is unique, non-negative real, and equal to zero if \( \lambda \mathbb{E}[W-D] < 1 \). This latter condition is implied by \( h\lambda \mathbb{E}[W-D]/h = \lambda \mathbb{E}[G] < 1 \), which itself is satisfied due to its equivalence to the queue stability requirement \( \rho = h\lambda \mathbb{E}[B] < 1 \), as can easily be seen from (4.3).

Let \( z = -s/h \), a substitution which satisfies \( \Re(z) \geq -\Re(s) \) since \( h \geq 1 \). The above expression (a.1) becomes

\[
(a.2) \quad \frac{s - h\lambda(1 - B^*(s))}{s - h\lambda(1 - \mathbb{E}[e^{-s(h(W-D)/h)]})} = 1 - \lambda \int_0^{+\infty} \mathbb{E}[e^{-s(B^*(B \geq t))/h})] e^{-\lambda t + \lambda M(s,t)/h} dt
\]

for \( \Re(s) > 0 \).

From here (4.5) is established by noting that the Laplace Transform of \( (hW-D)/h = G/h \) is just \( s \mapsto G^*(s/h) \) where \( G^* \) is the Laplace transform of \( G \). The numerator and denominator of the left hand side of (a.2) have essential singularities at \( s = 0 \) since the respective random variables \( B \) and \( G/h \) have infinite variance, however the limit of their quotient as \( s \) tends to zero from the right exists. The right hand side of (a.2) exists at \( s = 0 \). We do not know however whether the left or right hand sides of (a.2) are analytic at \( s = 0 \). Cohen shows that they are if \( B^*(s) \) itself is analytic at \( s = 0 \).

References


