More than exponential tail distribution in LAPALICE queues

Philippe Jacquet

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.:(1) 39 63 55 11

Rapports de Recherche

N° 1465

Programme 2
Calcul Symbolique, Programmation
et Génie logiciel

MORE THAN EXPONENTIAL TAIL DISTRIBUTION IN LAPALICE QUEUES

Philippe JACQUET

Juin 1991

More than exponential Tail distribution in LaPalice queues

Philippe Jacquet

June 7, 1991

Abstract. In this note we investigate the tail distribution of some key parameters which arise in the so-called LaPalice queues. In the LaPalice queues, customers are scheduled in such a way that the period between two consecutively scheduled customers is greater than or equal to the service time of the first one. Actual arrival time of each customer in queue is equal to the scheduled time plus a random independent shifting delay. This shifting delay generally destroys the scheduled pre-order and can build substantial queues phenomenon. Our aim is to show that under some general hypotheses the queue size and waiting delays have more than exponential tail distribution. This property is of prime importance for some applications since usual queues generally show exponential tail distribution for those key parameters.

Distributions asymptotiques plus qu'exponentielles dans les files d'attente de La Palice

Résumé. Dans cette note nous étudions la distribution asymptotique de certains paramètres clés qui apparaissent dans les files d'attente de La Palice. Dans une file d'attente de La Palice les clients sont convoqués de telle manière que la période séparant deux clients successifs soit plus grande ou égale au temps de service du premier des deux clients. La date d'arrivée effective d'un client dans la file d'attente est égale à la date de convocation plus un délai d'écart aléatoire et indépendant. En général, ce délai d'écart détruit le bel ordonnancement construit par les convocations et rend finallement possible des phénomènes sensibles de file d'attente. Notre propos est de montrer que sous des hypothèses très générales la taille de la file d'attente ou les délais d'attente en file ont des distributions asymptotiques plus qu'exponentielles. Cette propriété marque une différence significative avec les files d'attente usuelles qui ont en général des distributions asymptotiques exponentielles pour ces paramètres clés.

1 Introduction to LaPalice queues

LaPalice (*Monsieur de La Palice*, Maréchal de France) is a famous French general of François, First of the name, King of France. As it was usual for a soldier, during these times, LaPalice died on a battle field. He was very popular among his soldiers, who celebrated their general in a naive song where the following lyrical sentence held: "[LaPalice] was alive fifteen minutes before his death". This kind of obvious statement (in French *Lapalissade*) is now for ever attached to the popular character of LaPalice (frankly speaking this seems to have been the only memorable fact concerning this LaPalice).

Let us introduce a similar statement that concerns with queueing theory.

Lapalissade 1 *A queue where customers arrive separated by more than their service times has always its waiting room empty.*

As expected, this Lapalissade does not give any interesting information. Let us complicate our model in order to make our statements less innocent. We will call a *LaPalice queue*, every queue where the following holds.

Each customer has an integer index i. S_i is its service time, which we consider independently and identically distributed, with the same distribution as that of a positive random variable S.

Quantity θ_i is the *scheduled* arrival time of customer i. We suppose the fundamental relation: $\theta_{i+1} \geq \theta_i + S_i$. The previous inequality means that customers are scheduled in such a way that they are separated by more than their service times.

The *exact* arrival time is $\theta_i + D_i$ where the D_i's are independently and identically distributed with the same distribution as a random variable D. The D_i's are called the shifting delays.

The server policy can be First-in First-out, or with pre-emptive priorities, or processor sharing *etc.* We abbreviate LaPalice queueing model with single server $L/G/1$. We can also consider queueing models with multiple servers; in that case we refer to $L/G/n$, n being the number of servers. LaPalice queueing model is somewhat related to the class of queues with state dependent interarrival and service times, but to the best author's knowledge, this kind of queueing system seems to be new and the results stated in this paper, about the tail distributions, are new. The reader interested in literature and extensive analysis on queues with state dependent interarrival and service times is referred to [4] and the references cited here.

A. Motivations of LaPalice queueing models in Computer Science

LaPalice queues find numerous applications in Computer Science, especially in computer networks. A LaPalice queue can be seen as a special queue where each customer takes *rendez-vous* with the server but may arrive late to his rendez-vous because of some events which can be modeled as like shifting delays. Let us introduce an application for flow control in high speed switching packet networks.

Generally high speed device show lower storage capacity than lower speed device. Therefore flow control algorithms are very crucial in high speed networks. When propagation delays are much more larger than average packet emission, classic flow control algorithms are of no use, because they can greatly hurt network performance in such a way that you may simply forget about high speed. LaPalice queues are interesting alternative for flow control in high speed networks. Let us assume that for every message of length greater than, say, ten packets, the sender takes rendez-vous with the receiver, and let us suppose that the sender do that in parallel for every of its messages in buffer. In parallel means that the sender does not wait for the answer of the receiver (because

of large propagation delays) before sending the request for another message. Since, for the same sender, messages may have different destination, conflict may occur when messages have to be really transmitted. Added to contention delays on the networks, messages may arrive late at their rendez-vous on the receiver network controller and queues of packets may build up in the receiver high speed device. In the worst case, we can assume that the receiver receives messages from independent source, thus shifting delays may be independent; therefore we are exactly in LaPalice hypotheses. One of our main result is the fact that LaPalice queues very unlikely build up high. For example LaPalice queues generally show more than exponential tail distribution and when all shifting delays are bounded for the above, queue length is also bounded for the above.

The interested reader is referred to [2] for an extensive description and analysis of LaPalice queueing models in high speed networks.

B. Parameters of interest in LaPalice queueing models

The parameters of interest about our LaPalice Queueing naturally are the working load $Q(t)$ at time t and the customer delay W. We assume that Q involves load in waiting room and the remaining load of the customer in service. It is clear that if $D = 0$ in distribution, then the waiting room is always empty and W is identical to S, that is exactly the Lapalissade, above mentioned. Interesting facts occur when $D \neq 0$ in distribution. Note that W and Q do not change when D is translated by a constant quantity.

We restrict our analysis to the tail distributions of random variables W and Q. The knowledge of the tail distribution of Q is an important tool in order to design waiting rooms which are sufficiently large to make overflows very seldom events and small enough to make their cost affordable. One may have similar considerations about the tail distribution of waiting delays. The tail distribution is clearly the key of the benefits we can get from LaPalice application to high speed networks.

For the following, we consider a simple First-in First-out $L/G/1$ system.

Lapalissade 2 *Let Δ and Σ be two non-negative real numbers. If $D \leq \Delta$ and $S \leq \Sigma$ in distribution, then the working loads Q and delays W in the First-in First-out $L/G/1$ are both smaller than or equal to $\Delta + \Sigma$ in distribution.*

This Lapalissade is not obvious although it seems to simply state that total waiting delay is equal to the sum of delay D and service time W. This consideration is in fact a *faux ami* since shifting delays are not included in waiting delays and service time does not involve delay in waiting room. However, the proof to that lapalissade is really very simple, thus it will explain *a posteriori* why we could call it a lapalissade.

C. History vectors and sub-markovian gaps distribution

Let service times S be generally distributed with mean μ. If the distribution is such that S cannot be less than a given threshold $\sigma > 0$, we say that service time is bounded for the below. If S cannot exceed another threshold Σ, we say that service time is bounded for the above. We assume that D has a general distribution. When S and D are both bounded for the above we are in the case of the previous lapalissade.

In order to completely describe our $L/G/1$ system, it remains to quantify the *gaps* between service times, namely the quantities $\delta_{i+1} = \theta_{i+1} - \theta_i - S_i$. We can assume, for simplicity, that the δ_i are independent and identically distributed as an exponential random variable with parameter λ; but in fact much more general conditions can be assumed for this stochastic quantity. For the

following we simply assume that the gaps are part of a renewal process. Let $1/\lambda$ be the average gap duration.

We can describe our system by two time axis that we call *history vectors*. One time axis depicts points θ_i and $\theta_i + S_i$. We color in black intervals $[\theta_i, \theta_i + S_i]$ (for service times) and in white the interval $[\theta_i + S_i, \theta_{i+1}]$ (for gaps). The second time axis depicts arrival times $\theta_i + D_i$. See figure 1 for illustration. We call the past history vectors of t these two time axis truncated of their extension beyond t, which actually belongs to the future of t.

Let us introduce the *sub-markovian* gaps distribution concept. Let us consider a given time t. Let $r(t) > t$ the function of t defined by the fact that the gaps or the fractions of gaps contained by the interval $[t, t + r(t)]$ have cumulated length exactly equal to 1 time unit. The gaps distribution is sub-markovian when there exists a non-negative real number $\rho > 0$ with the following property: for any time t and for any past history vectors, the average value of $r(t)$ given the past history vectors is always less than ρ.

Note that when the gaps are simply identically and independently distributed as a random variable δ, and when S has at least an exponential tail distribution the gap distribution is sub-markovian. But much more general scheduling processes (defining the δ_i's and the S_i's) also lead to sub-markovian gaps distribution and, therefore, are concerned by our results. For example, when the scheduling process is identical to the output process of a $M/G/1$,or $G/G/1$, *etc*, systems.

When S and D are not bounded for the above, we will prove the following theorem.

Theorem 1 *Let us consider a $L/G/1$ queue with sub-markovian gaps distribution and service time bounded for the below. If there exist non-negative real numbers A, α, B and β such that $\alpha > 1$, $\alpha > \beta$ and quantities* $\limsup \log \Pr\{S \geq x\}/Ax^{\alpha}$ *and* $\limsup \log \Pr\{D \geq x\}/Bx^{\beta}$ *are both less than -1, then quantities* $\limsup \log \Pr\{Q \geq x\}/Cx^{\gamma}$ *and* $\limsup \log \Pr\{W \geq x\}/Cx^{\gamma}$ *are both less than -1 when $x \to \infty$, with $\gamma = \beta + 1 - \beta/\alpha$ and $C = \frac{(A\alpha)^{1/\alpha}}{\gamma}(B\frac{(\alpha-1)}{\alpha})^{(\alpha-1)/\alpha}$.*

Note that $\alpha > \gamma > \beta$ and $\gamma > 1$. This theorem shows that generally when random variable S has more than exponential tail, therefore random variables Q and W also have more than exponential tail. One may conclude that $L/G/1$ systems have better delays than $M/G/1$ systems, since $M/G/1$ systems always lead to simple exponential tails for their delays and working loads. A recent paper [3] of Sadowsky and Szpankowski shows that heterogeneous $GI/GI/c$ queues also have exponential tail distributions. $L/G/1$ queues show also smaller queue size than $M/G/\infty$ queues. $M/G/\infty$ queue size is directly Poisson (therefore tail distribution is $\exp\{-O(x \log x)\}$). It is interesting to notice that if parameter β tends to zero, then $\gamma \to 1$, *i.e.* we tend to exponential tail distribution. Clearly when the shifting delay distribution is stretched to infinity, which may illustrate the fact that $\beta \to 0$, one can consider that our $L/G/1$ queue tends to a limiting $M/G/1$. For example, let us consider that the shifting delay is uniform on a given interval $[a, b]$; when $b \to \infty$, with a fixed, correlations between arrival times and past scheduled times tends to disappear and the inter-arrival times to be distributed as a Poisson memoryless process.

One may consider our assumption of S as bounded for the below as a restrictive statement, but we conjecture that our theorem remains true without this assumption, but we have not been able to prove this. However the assumption fits the major application of LaPalice queues. For example for flow control in high speed network, it is clear that LaPalice queues only concern with large messages (a rendez-vous costs at least two packets).

Clearly the statements of our theorem are not very simple and their proofs are really technical, far from obvious, at least from the author's point of view. This is why we decided to state them in

a theorems, not in a lapalissades (obviousness has its limit).

The following is build around three main sections. The first one is the proof of the lapalissade, where we include very simple lemmas, all are not totally necessary for the lapalissade but this will shed some light on the following section: the proof of the theorem. We conclude our paper by some considerations about specific cases and further applications.

2 Proof of the Lapalissade

Quantity $Q(t)$ is expected to experience large variations with respect to variable t (Heaviside like behaviour). Therefore we introduce two parallel parameters $B(t)$ and $L(t)$ which will have smoother variations.

Let us make things precise. We will say that a customer i is already scheduled at time t if $\theta_i < t$. We will say that a customer i has already arrived at time t if $\theta_i + D_i < t$. A customer which is scheduled but has not yet arrived (such that $\theta_i < t \leq \theta_i + D_i$) is said to be in *purgatory*. A customer which has arrived (such that $\theta_i + D_i < t$) and whose service time is not yet expired is said to be in *queue*.

We call $B(t)$ the *load in purgatory* at time t. By load in purgatory at time t we mean the cumulated service times of the customers which are in purgatory at time t, with the convention that if one of these customers is scheduled at time $t - \delta$ and its service time is greater than δ, then only fraction δ is counted in $B(t)$. This statement allows us smoothing the variations of $B(t)$.

We call $L(t)$ the *queued load* at time t. The queued load at time t is the cumulated remaining service times of the customers which are in queue at time t. By remaining service time we mean the the service time minus its expired fraction at time t. We adopt similar convention than with $B(t)$: if one of these customers has been scheduled at $t - \delta$ and its service time is greater than δ, then only δ is is counted in $L(t)$ (minus its expired fraction). The distinction with the working load in queue, $Q(t)$ is relevant only when there is such a customer that we say to be in *current arrival*. To be precise, quantity $Q(t)$ involves the whole service time of the customer in current arrival. We always have $L(t) \leq Q(t)$.

The classic *busy period* denotes any time interval $[a, b]$ such that $Q(t) > 0$ for all $a < t < b$ and $Q(t) = 0$ for all $t \in [a - \varepsilon, a[\cup[b, b + \varepsilon[$ for some $\varepsilon > 0$. The server is permanently busy during a busy period. We call initial global load of the busy period, the quantity $L(a) + B(a)$. In opposition, the partial initial load is only $L(a)$ (will be of no use in this paper).

In Figure 1, we show the behaviour of these parameters in a concrete example. All horizontal lines are parallel time axis. The two first line are history vectors. The first line shows the θ_i and the S_i for 5 customers, $i = 1, \ldots, 5$. The black area after each picked θ_i corresponds to the duration of the service times S_i. The second line shows the t_i's, note that the relative order is now t_1, t_3, t_2, t_5, t_4. We depict quantity $B(t)$ of load in purgatory, below this line in inversed perspective. The third line corresponds to quantities $L(t)$ (solid) and $Q(t)$ (dotted), note that for all t: $Q(t) \geq L(t)$. When $L(t) = Q(t)$, the line is solid. The fourth line show quantity $L(t) + B(t)$.

Lemma 1 *Function $B(t)$ is lower semicontinuous, function $L(t)$ is upper semicontinuous, function $B(t) + L(t)$ is continuous.*

We refer to figure 1 to illustrate this proposition.

Remark: Functions $B(t)$ and $L(t)$ are both left semicontinuous.

Lemma 2 *Let t an arbitrary time. Let B^* be the right limit of $B(\theta)$ when $\theta \to t$. For all $\theta < t$ we have $B(\theta) \geq B^* - (t - \theta)$.*

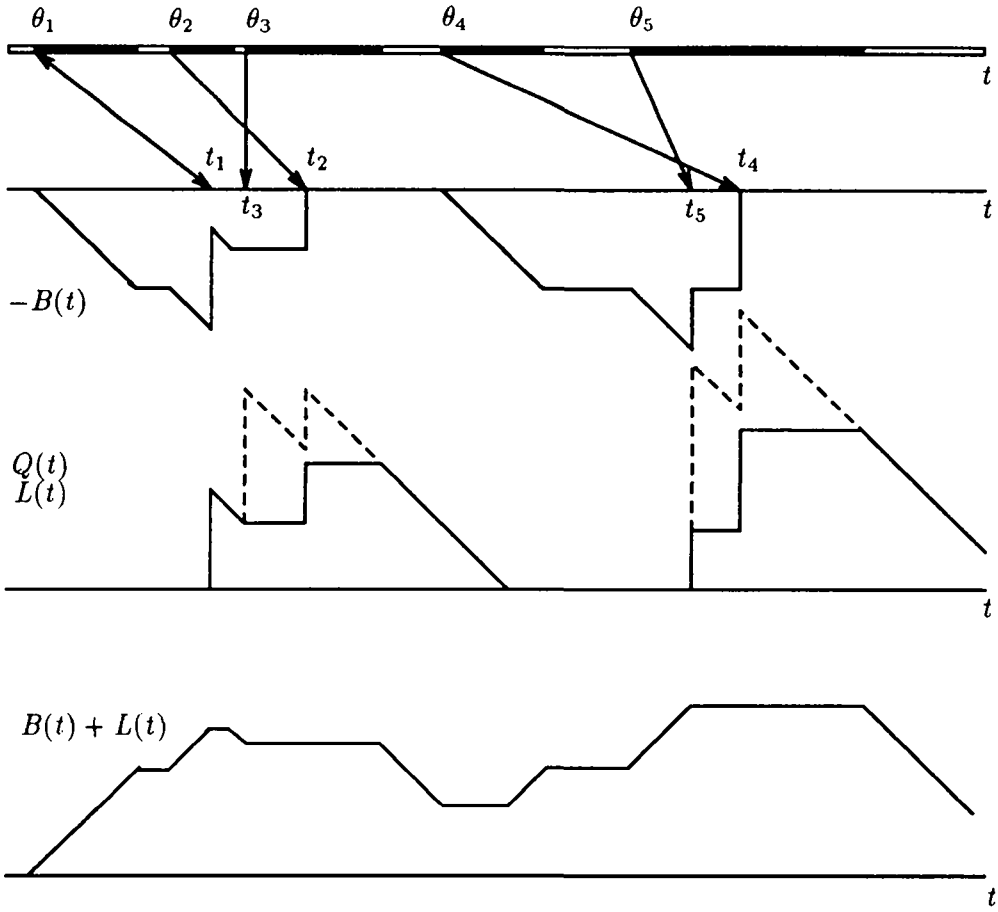

Figure 1: Illustration of history vectors and the behaviour of $B(t)$, $L(t)$ and $B(t) + L(t)$ (solid) and $Q(t)$ (dotted)

Proof. $B(t)$ cannot increase with rate higher than 1. ∎

Lemma 3 *Quantity $L(t) + B(t)$ decreases during a busy period.*

Proof. Quantity $L(t) + B(t)$ is the service time of customer already scheduled but not exhausted. This load enters the system with continuous rate equal to 1 if there is one customer such that $\theta_i < t < \theta_i + S_i$, and equal to zero in the alternative case. Since the load exits the system with continuous service rate 1 during a busy period, the lemma is proved. ∎

Lemma 4 *Let $\Delta > 0$. If random variable D is always smaller than or equal to a given constant Δ then $B(t)$ is always smaller than or equal to Δ.*

Proof. It is clear that at every time t the oldest customer in purgatory cannot be older than Δ time units. Since the intervals $[\theta_i, \min\{t, \theta_i + S_i\}]$ are pair disjoint and must fit the interval $[t-\Delta, t]$ when i is the index of a customer in purgatory, since $B(t)$ is exactly the sum of the $\min\{t+i+S_i, t\} - t_i$'s, the lemma is proved. ∎

Lemma 5 *If random variable D is always smaller than or equal to a given constant Δ, then every busy period has its initial global load less than or equal to Δ.*

Proof. Let $[a, b]$ be a busy period, since $L(t) + B(t)$ is continuous we have $\lim_{t \to a} L(t) + B(t) = L(a) + B(a)$. But for $t \in [a - \varepsilon, a[$ for some $\varepsilon > 0$ we have $L(t) = 0$, therefore $L(t) + B(t) = B(t)$ and consequently $L(t) + B(t) \leq \Delta$, using previous lemma. Thus $L(a) + B(a) \leq \Delta$. ∎

Lemma 6 *If random variable D is always smaller than or equal to a given constant Δ, then we always have $L(t) \leq \Delta$ in distribution.*

Proof. Obvious with previous lemma and lemma 3. ∎

Lemma 7 *If random variables D and S are always smaller than or equal to respectively given constant Δ and Σ, then for all t $Q(t) \leq \Delta + \Sigma$.*

Proof. We already know that $Q(t) = L(t)$ when there is no current arrival at t. When the youngest customer in queue is in current arrival at time t and x is the remaining part of its service time ($x = S_i + \theta_i - t$, we simply have $Q(t) = L(t) + x$. Therefore we always have $Q(t) \leq L(t) + \Sigma$ for all t. The previous lemma allows us to conclude. ∎

Lemma 8 *If random variables D and S are always smaller than or equal to respectively given constant Δ and Σ, then for all customers $W \leq \Delta + \Sigma$.*

Proof. Since we consider a First-in First-out policy, the total waiting delay (including service time) of a customer that is queued at time t is exactly $Q(t)$. ∎

3 Proof of the theorem

In the following we consider a First-in First-out $L/G/1$ with sub-markovian gaps distribution. Let μ be the average service time. By hypothesis there is a real number $\rho < \infty$ such that the following proposition is valid for arbitrarily chosen past history vectors: the average time necessary to see enough gaps that fill a unit time interval is always less than ρ when we condition our observation by giving a fixed past history vectors.

The proof of the theorem first establishes appropriate upper bounds of the tail distribution of quantity $B(t)$ and then we extend this result to the tail distributions of the other quantities such that $L(t)$, $Q(t)$ and $W(t)$, *via* classic relations between these parameters (*e.g.*, Little's formula).

A. Loads in purgatory

In this subsection we are interested in the tail distribution of the quantity $B(t)$ when random variables D and S are not bounded for the above.

Lemma 9 *Let $\ell(x)$ be the average length of a busy period starting at time t with $L(t) + B(t) = x$, we have*

$$\ell(x) \leq (x + 1)\rho \, .$$

Proof. In order to give an upper bound of $\ell(x)$ we build a modified system at time t which bounds for the above the original system. The modified system parallels the original one before time t. At time t we immediately queue in the modified system all customers which were in purgatory at time t in the original system. Furthermore, customers which are scheduled after time t in the original system do not stay in purgatory in the modified system and are queued exactly at their scheduled

time (as if their shifting delays were zero). Let $L^*(t)$ be the queued load of the modified parallel system, it is clear that the busy period of the modified system is always longer than the busy period of the original system. It is not difficult to see that the busy period of the modified system is exactly the time interval $[t, t']$ which contains gaps or fractions of gaps of cumulated length equal to x time units. Since the gaps distribution is sub-markovian the mean quantity $t' - t$ is less than $(x+1)\rho$, the lemma is proved. ∎

In the following we denote by $P_S(x)$ the probability that $S \geq x$ (thus $P_S(x) = \exp\{-f(x)\}$) and $P_D(y)$ the probability that $D \geq y$. We denote by $dP_S(x)$ the probability measure of random variable S at value x (formally the probability that $S = x$): $P_S(x) = \int_x^\infty dP_S(y)$.

Let us consider the case where there are n customers in purgatory at time t. We rank customers in the decreasing order of their scheduled times (from the youngest to the oldest). We suppose that their respective service times are $x_1, \ldots, x_n$. We consider the case where customer number 1 was scheduled more than its service time before t. Therefore $B(t) = x_1 + \cdots + x_n$. Let $dp_0(x_1, \ldots, x_n)$ be the probability measure on R_+^n corresponding to the statistic distribution of the n-tuples $(x_1, \ldots, x_n)$.

In the case where customer 1 was scheduled less than its service time before t, we define x_1 as the fraction of this service time already counted in purgatory. Therefore $B(t) = x_1 + \cdots + x_n$ still holds. Let $dp_1(x_1, \ldots, x_n)$ be the probability measure on R_+^n of such event.

Lemma 10 *We have the following inequalities*

$$dp_0(x_1, \ldots, x_n) \leq \prod_{i=1}^{i=n} P_D(x_1 + \cdots + x_i) dP_S(x_i) \ . \tag{1}$$

and

$$dp_1(x_1, \ldots, x_n) \leq P_D(x_1)\mu^{-1} P_S(x_1) dx_1 \prod_{i=2}^{i=n} dP_S(x_i) P_D(x_1 + \cdots + x_i) \ . \tag{2}$$

Proof. Inequality (1) comes from the fact that customer i must independently have a service time equal to x_i (thus a factor $dP_S(x_i)$) and must have waited more than $x_1 + \cdots + x_i$ in purgatory (thus a factor $P_D(x_1 + \cdots + x_n)$). Inequality (2) is easy to derive by replacing in (1) factor $dP_S(x_1)$ by $P_S(x_1)dx_1/\mu$, since μ is the mean value of service time S. Quantity $P_S(x_1)dx_1/\mu$ is the probability customer 1 have exactly x_1 of its service time in purgatory when it has been scheduled less than its service time before t. ∎

Lemma 11 *For all $b > a \geq 0$, we have*

$$\Pr\{B(t) \in [a,b]\} = \sum_{n=1}^{n=\lceil b/\sigma \rceil + 1} \int_{x_1 + \cdots + x_n \in [a,b]} \{dp_0(x_1, \cdots, x_n) + dp_1(x_1, \cdots, x_n)\} \ , \tag{3}$$

where $\lfloor x \rfloor = \lceil x \rceil - 1$ denotes the integer part of x.

Proof. When $B(t) \leq b$ the number of customers in purgatory cannot exceed $\lceil b/\sigma \rceil + 1$, since service time cannot be less than σ (the $+1$ comes from the case when we consider that customer 1 has only a fraction of its service time counted in the load in purgatory). ∎

Corollary 1 *We have*

$$\Pr\{B(t) \in [a,b]\} \leq 2(\frac{b}{\sigma}+2)\Pr\{D > a\} \ .$$

Proof. It is clear that

$$\int_{x_1+\cdots+x_n\in[a,b]} dp_0(x_1,\ldots,x_n) \leq P_D(a) \ ,$$

since the integral $\int_{R_+^n} \prod dP_S(x_i)$ is expected to be equal to 1. We have the same inequality with $dp_1(x_1,\ldots,x_n)$. ∎

Let $f(x) = \min\{Ax^\alpha, -\log P_S(x)\}$ and $g(x) = \min\{Bx^\beta, -\log P_D(x)\}$, according to our hypotheses about asymptotics of $-\log P_S(x)$ and $-\log P_D(x)$, it is clear that $f(x) \sim Ax^\alpha$ and $g(x) \sim Bx^\beta$. We have the following non-obvious lemma.

Lemma 12 *There exists $H > 0$ such that for all $b > a > \sigma$ the following inequality holds*

$$\begin{aligned}\Pr\{B(t) \in [a,b]\} \leq H \sum_{n=1}^{n=\lceil b/\sigma\rceil+1} \\ \frac{n!}{\sigma^n}\int_{x_1+\cdots+x_n\in[a-\sigma,b]} \exp\{-\sum_{i=1}^{i=n} f(x_i) + g(x_1+\cdots+x_i)\}dx_1\cdots dx_n \ ,\end{aligned} \tag{4}$$

We defer the proof of this technical lemma to the appendix.

Lemma 13 *Let $h(x)$ the function defined by*

$$h(x) = \min_{n,x_1+\cdots+x_n=x}\{\sum_{i=1}^{i=n} f(x_i) + g(x_1+\cdots+x_i)\} \ ,$$

where all the x_i's are positive. Let $x \geq 2\sigma$, we have

$$\Pr\{B(t) \in [x-\sigma, x]\} \leq (\frac{x+1}{\sigma})^{x+1}\int_{x-2\sigma}^{x} \exp\{-h(y)\}dy \ .$$

Proof. It is a well known result (undergraduate course in mathematics) that

$$\int_{x_1+\cdots+x_n\leq x} dx_1\cdots dx_n = \frac{x^n}{n!} \ .$$

Thus, by rewriting (4),

$$\begin{aligned}\Pr\{B(t) \in [x-\sigma, x]\} \leq H \sum_{n=1}^{n=\lceil x/\sigma\rceil+1} (\frac{n}{x})(\frac{x}{\sigma})^n \int_{x_1+\cdots+x_n\geq x} \\ \exp\{-\sum_{i=1}^{i=n} f(x_i) + g(x_1+\cdots+x_i)\}dx_1\cdots dx_n \ ,\end{aligned}$$

we easily obtain the expected inequality. ∎

Finally, to establish our main result we apply following theorem from [1].

Theorem 2 (stationary inf-convolution) *We have* $h(x) \sim Cx^\gamma$ *when* $x \to \infty$.

Thus, the following lemma holds.

Lemma 14 *We have*

$$\limsup_{x\to\infty} \frac{\log \Pr\{L(t) \geq x\}}{Cx^\gamma} \leq -1 \ .$$

Proof. Let $H(x) = \log[(x+1)^{x+1} \int_{x-2\sigma}^{x} \exp\{-h(y)\}dy$, it is readily that $H(x)$ is also equivalent to $-Cx^\gamma$ when $x \to \infty$ since $\log(x+1)^{x+1} \sim x \log x$ which is negligible before Cx^γ. It is also clear that $\log[\sum_{i=0}^{\infty} \exp\{-H(x+i\sigma)] \sim -Cx^\gamma$. Since $\Pr\{B(t) \geq x\} \leq \sum_{i=1}^{\infty} \Pr\{B(t) \in [x+(i-1)\sigma, x+i\sigma]\}$ the corollary is proved. ∎

B. Analysis of the load in queue

Let $\nu(x)$ be the frequence of busy period occurrences starting with initial global load greater than or equal to x.

Lemma 15 *We have*

$$\limsup_{x\to\infty} \frac{\log \nu(x)}{Cx^\gamma} \leq -1 \ .$$

Proof. According to lemma 2 it is clear that if t is the beginning of a busy period such that $L(t)+B(t) \geq x$, then for all $\theta \in [t-\sigma, t]$: $B(\theta) \geq x-\sigma$. Thus, since beginnings of busy periods are necessarily separated by more than σ, the intervals $[t_i - \sigma, t_i[$, where the t_i's are beginnings of busy periods starting with $L(t)+B(t) \geq x$, are necessarily pair disjoint. On this interval $B(t) \geq x-\sigma$, therefore we necessarily have $\sigma\nu(x) \leq \Pr\{B(t) \geq x-\sigma\}$. ∎

Lemma 16 *We have* $\Pr\{L(t) \geq x]\} \leq \sum_{n=0}^{\infty} \nu(x+n\sigma)\ell(x+(n+1)\sigma)$.

Proof. Let $L^*(t)$ be a random variable which majorizes the actual $L(t)$ for every t. We define $L^*(\theta) = 0$ when θ is outside busy periods of the original system (therefore $L(\theta) = 0$ also), and $L^*(\theta) = L(t) + B(t)$ when θ is inside a busy period of the real system and t is the beginning of the busy period ($L^*(\theta)$ is equal to the global initial load of the busy period). We know that $L(\theta) \leq L(t) + B(t)$ because of lemma 3. The frequency of busy period occurrences with initial global load in the interval $[x, x+\sigma]$ is less than $\nu(x)$, the average length of such periods is less than $\ell(x+\sigma)$, therefore the following holds

$$\Pr\{L^*(t) \in [x, x+\sigma]\} \leq \ell(x+\sigma)\nu(x) \tag{5}$$

And summing on $[x+n\sigma, x+(n+1)\sigma]$, and then using $L^*(t) \geq L(t)$ for all t, the lemma is proved. ∎

Lemma 17 *We have*

$$\limsup_{x\to\infty} \frac{\log \Pr\{L(t) \geq x\}}{Cx^\gamma} \leq -1 \ .$$

Proof. Since $\ell(x) \leq (x+1)\rho$ and using lemma 15 about the asymptotic behaviour of $\log \nu(x)$ we obtain the asymptotic behaviour of $\log \nu(x)\ell(x+\sigma)$ and, equivalently the asymptotic behaviour of $\log \sum_n \nu(x+n)\ell(x+(n+1)\sigma)$. ∎

C. The working load in queue $Q(t)$

We are interested in deriving the tail distribution of $Q(t)$, not $L(t)$. But $L(t)$ and $Q(t)$ are not far for away. It is clear that $Q(t) = L(t)$ when t is not inside a current arrival in queue (*i.e.* there is a customer i in queue such that $t < \theta_i + S_i$). In the case where t is selected inside of current arrival of length S (corresponding to a service time S), we have $Q(t) \leq L(t) + S$.

Lemma 18 *The average number of current arrivals that occur inside a busy period starting with $L(t) + B(t) = x$ is less than $\ell(x)/\sigma$.*

Proof. The length of a current arrival cannot be less than σ. ∎

Lemma 19 *The average number of current arrivals of length y that occurs during a busy period with global initial load equal to x is less than $\ell(x)dP_S(y)/\sigma$.*

Proof. Quantity $dP_S(y)$ refers to the distribution measure of random variable S. Each of the current arrivals, taken separately, has probability $dP_S(y)$ corresponding to a service time equal to y, independently of the past events of the system, in particular the fact that it occurs inside a busy period with initial global load x. ∎

For the following we denote by $d\nu(x)$ the frequency measure of busy period occurrences with initial global load equal to x. Therefore $\int_x^\infty d\nu(y) = \nu(x)$.

Lemma 20 *We have the following inequality*

$$\Pr\{Q(t) \geq x\} \leq \Pr\{L(t) \geq x\} + \\ + \frac{1}{\sigma} \int_{x_1+x_2 \geq x} \ell(x_1) x_2 d\nu(x_1) dP_S(x_2) + \int_{x_1+x_2 \geq x} x_2 d\nu(x_1) dP_S(x_2) \ . \tag{6}$$

Proof. There are three terms in the right hand side. The first one is clear, it gives an obvious majorization in the case where t is selected outside any current arrival in queue. In this case $Q(t) = L(t)$. The second term is obtained by corollary to previous lemmas, in the sense that it gives a majorization in the case where t is chosen when a customer is in current arrival during a busy period. If the busy period started with initial global load x_1 and the current arrival is of length x_2, we have $Q(t) \leq x_1 + x_2$. To recall a current arrival at time t is when a customer i is queued with quantity $\theta_i + S_i$ greater than current time t.

The third term contributes when the busy period itself starts during a current arrival. In this case we do not have independence between initial global load of the busy period and the length of the current arrival. In order to recover independence we have to remove from the initial load the fraction due to the arriving customer. But our formula is still valid since we are dealing with upper bounds. ∎

The following lemma gives a simple result which is mentioned in [1].

Lemma 21 *Let $\zeta(x)$ and $\phi(x)$ two positive increasing functions such that $\zeta(x) \sim Ax^\alpha$ and $\phi(x) \sim Cx^\gamma$ when $x \to \infty$. We define $\Phi(x) = \min_{y \in [0,x]}\{\phi(y) + \zeta(x-y)\}$. We have $\Phi(x) \sim Cx^\gamma$, when $x \to \infty$.*

This leads to the following result.

Lemma 22 *We have*

$$\limsup (Cx^{\gamma})^{-1} \log \int_{x_1+x_2 \geq x} \ell(x_1) x_2 d\nu(x_1) dP_S(x_2) \leq -1 \ .$$

Proof. Since measures $d\nu(x)$ and $dP_S(x)$ are zero as soon $x \leq \sigma$, we can apply techniques we used in the proof of lemma 12 (see appendix) and obtain for all $x \geq \sigma$ the inequality

$$\int_{x_1+x_2 \geq x} \ell(x_1) x_2 d\nu(x_1) dP_S(x_2) \leq \frac{2}{\sigma^2} \int_{x_1+x_2 \geq x-\sigma} \ell(x_1) x_2 \nu(x_1) P_S(x_2) dx_1 dx_2 \ .$$

Let $\phi(x) = -\log(\ell(x)\nu(x))$ and $\zeta(x) = -\log(xP_S(x))$. Let $\Phi(x) = \min_{y \in [0,x]}\{\phi(y) + \zeta(x-y)\}$. It is clear that $\zeta(x) \sim Ax^{\alpha}$ and $\liminf \phi(x)(Cx^{\gamma})^{-1} \geq 1$ therefore, using lemma 21, we get $\liminf \Phi(x)(Cx^{\gamma})^{-1} \geq 1$. Since

$$\int_{x_1+x_2 \geq x} \ell(x_1) x_2 d\nu(x_1) dP_S(x_2) \leq \int_{x-\sigma}^{\infty} \exp\{-\Phi(y)\} dy \ ,$$

the lemma is proved. ∎

Lemma 23 *We have*

$$\limsup (Cx^{\gamma})^{-1} \log \Pr\{Q(t) \geq x\} \leq -1 \ .$$

Proof. It suffices to prove the convergence when replacing $\Pr\{Q(t) \geq x\}$ by each of the three right hand side terms of inequation (6). The result is clear about the first term because of lemma 17. Same consideration about the second term with lemma 22. About the third term, similar treatment can be done as in lemma 22. ∎

D. Analysis of the waiting time in queue

The random variable W denotes the total waiting time spent in queue by any random customer. Our final purpose is to derive the asymptotic behaviour of the tail distribution of W.

Let us first introduce intermediate random variable ω. The random variable $\omega(t)$ is attached to a random customer and a random time. $\omega(t)$ is equal to the remaining delay that the customer has to spend in queue when regarding at time t.

Figure 2 illustrates this definition. Horizontal scale is time t vertical scale is load and waiting times in the same time unit. We show quantity $Q(t)$ (solid line) and the different $\omega(t)$'s (dotted lines) of the different customers in queue at time t. Note that the highest $\omega(t)$ is identical to $Q(t)$ and therefore is confused with the solid line. We mark the level x in order to illustrate lemma 25

Lemma 24 *The average number of customers such that $\omega \geq x$ when considering the queue at a random time is exactly*

$$\frac{\lambda}{1+\lambda\mu} \int_x^{\infty} \Pr\{W \geq y\} dy \ . \tag{7}$$

Proof. This lemma is a simple application of Little's formula, since $\lambda/(1+\lambda\mu)$ is the average number of customers queued during one time unit. ∎

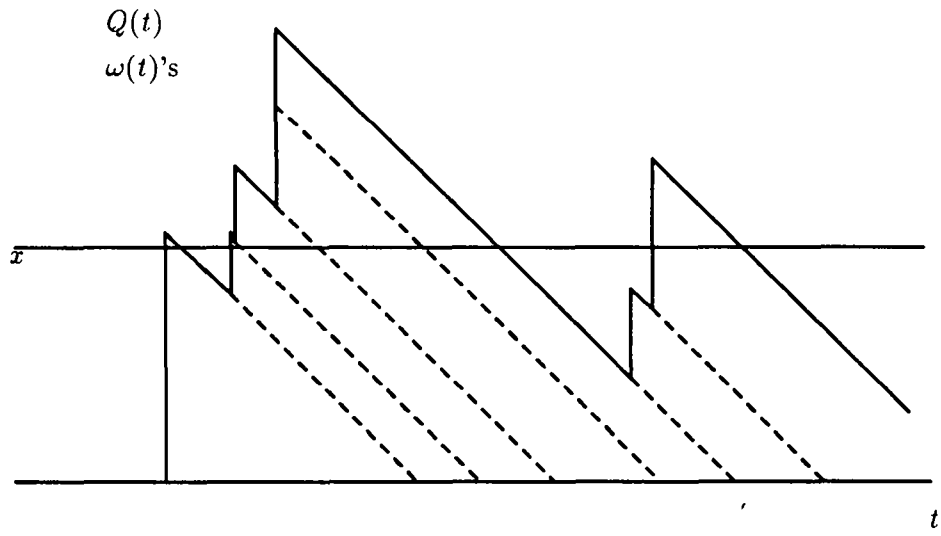

Figure 2: Illustration of the behaviour of $Q(t)$ (solid) and the $\omega(t)$'s (dotted)

Lemma 25 *We have the following inequality*

$$\frac{\lambda}{1+\lambda\mu}\int_x^\infty \Pr\{W \geq y\}dy \leq \sum_{n=0}^{\infty} \Pr\{Q(t) \geq x + n\sigma\} \tag{8}$$

Proof. Since service times are at least σ the ω's of any pair of queued customers considered at the same given time t show discrepancy greater than σ. Since the queue is First-in, First-out, the ω's are all less than or equal to $Q(t)$. Therefore the number of customers such that $\omega \geq x$ at time t is less than $\lceil Q(t)/\sigma \rceil$. See figure 2 for illustration ∎

Proof of theorem 1. A rough derivation of inequality (8) readily gives $\sigma \Pr\{W \geq x+\sigma\} \leq (\frac{1}{\lambda}+\mu)\phi(x)$ with $\phi(x) = \sum_{n=0}^{\infty} \Pr\{Q(t) \geq x+n\sigma\}$. Since we readily have $\limsup \log \phi(x)/Cx^\gamma \leq -1$ the theorem is proven. ∎

4 Further considerations

One may ask about theorem 1 when $\beta > \alpha$. But in this case the theorem holds with $\gamma = \beta$ and $C = B$. The reason for this is that the asymptotic behaviour of function $h(x)$ defined in lemma 13 is clearly $h(x) \sim Bx^\beta$, since we obviously have $g(x) < h(x) < g(x) + f(x)$.

When reading [1], we notice that the case where random variable S is upper bounded, can be treated as in theorem 1. Let Σ be the upper bound of random variable S. In this case we say that function $f(x)$ is a degenerate function with edge Σ, *i.e.* $f(x) = \infty$ when $x > \Sigma$. This case is treated as an eccentric case in [1] and we get a companion theorem of theorem 2.

Theorem 3 *If $f(x)$ is a degenerate function with edge Σ, and $g(x) \sim Bx^\beta$ when $x \to \infty$, then $h(x) \sim \frac{B}{(\beta+1)\Sigma} x^{\beta+1}$.*

This allows us to state a companion theorem of theorem 1.

Theorem 4 *Let us consider a $L/G/1$ queue with sub-markovian gaps distribution and service times bounded for the below. If service time is also bounded by the above by a constant Σ and quantity $\limsup(Bx^{\beta})^{-1}\log\Pr\{D \geq x\}$ is less than -1 when $x \to \infty$, then the following quantities $\limsup(Cx^{\gamma})^{-1}\log\Pr\{Q \geq x\}$ and $\limsup(Cx^{\gamma})^{-1}\log\Pr\{W \geq x\}$ are both less than -1 when $x \to \infty$, with $C = \frac{B}{(\beta+1)\Sigma}$.*

About theorems 1 and 4, we have the following conjecture.

Conjecture 1 *In general the statements of theorem 1 and 4 hold when we replace all "$\limsup$" by "$\lim$", and all "less than -1" by "equal to -1".*

The reader is referred to appendix for partial proof of this conjecture. The conjecture is proven in the case where gaps are stochastically smaller than an independently and identically distributed random variable. This very condition fits of course the case where the gaps are themselves independently and identically distributed, but also the case where the schedule process is equivalent to the output process of a $M/G/1$ queue (or the output process of most $G/G/1$ queues).

The fact that $L/G/1$ queueing models lead to more than exponential tail is interesting since it drastically reduces overflow occurrence in limited buffer queueings. Numerous applications allow us to consider service time bounded for the above and exponential tail for the shifting delay: $-\log\Pr\{D \geq x\} \sim Bx$, when $x \to \infty$. Therefore the overflow probability of a buffer of capacity x is $\Pr\{Q(t) \geq x\}$ and we know that $-\log\Pr\{Q(t) \geq x\}$ is at least equivalent to $Bx^2/2\Sigma$, which is equivalent to a gaussian tail. If D had itself a gaussian tail distribution, namely $-\log\Pr\{D \geq x\} \sim Bx^2/2$, we obtain a cubic exponential tail for $Q(t)$: $-\log\Pr\{Q(t) \geq x\}$ is at least equivalent to $Bx^3/6\Sigma$. This perspective looks very promising in comparison with usual $M/G/1$'s which lead at least to exponential tail for $Q(t)$: in general we have $-\log\Pr\{Q(t) \geq x\} \sim C(\lambda)x$ where $C(\lambda)$ is a function of the interarrival time parameter.

It is clear that the appropriate tool to transform usual $M/G/1$ into $L/G/1$ is the schedule through *rendez-vous*. This can find numerous extension in computer science problems where the cost of a simple rendez-vous may be lower than large capacity buffer. We discussed in the introduction the application in flow control in high speed networks. In high speed metropolitan network it is difficult to use classic flow control algorithms since large propagation delays obliterate performance. On the other hand, high speed devices show lower integration facilities that limit buffer capacities at low level. LaPalice queueing is an interesting alternative to classic flow control algorithm in such systems. This proposition is extensively analyzed in [2].

References

[1] P. JACQUET "A theorem about the asymptotic behaviour of some inf-convolution equations," 1991, submitted.

[2] P. JACQUET AND P. MÜHLETHALER "Flow control algorithms using LaPalice queueings in high speed networks," 1991, submitted.

[3] J. SADOWSKY AND W. SZPANKOWSKI "The Probability of Large Queue Lengths and Waiting Times in a Heterogeneous $GI/GI/c$ Queue," 1991 submitted.

[4] W. WHITT "Queues with service times and interarrivals times depending linearly and randomly upon waiting times," Queueing Systems, 6, pp. 335-352, 1990.

APPENDIX

Proof of lemma 12 Let us concentrate on $\int dp_0(x_1, \cdots, x_n)$. Let Ω be a measurable subset of R_+^n, we have

$$\int dp_0(x_1, \cdots, x_n) \leq \int \prod_{i=1}^{i=n} dP_S(x_i) P_D(x_1 + \cdots + x_n) \ . \tag{9}$$

For reasons of convenience, let $\mathbf{x}$, $(dP_S(\mathbf{x}))^n$ and $(P_S(\mathbf{x})d\mathbf{x})^n$ be respectively short hand notations for n-uple $(x_1, \cdots, x_n)$, the expanded measures $\prod dP_S(x_i)$ and $\prod P_S(x_i)dx_i$. Let $\phi(\mathbf{x})$ be an arbitrary measurable positive function. Some easy algebra gives

$$\int_\Omega \phi(\mathbf{x})(P_S(\mathbf{x})d\mathbf{x})^n = \int_\Omega \phi(\mathbf{x}) V(\Omega, \mathbf{x})(dP_S(\mathbf{x}))^n \ , \tag{10}$$

where $V(\Omega, \mathbf{x})$ is the volume of the intersection set between Ω and the set of all n-uples $(y_1, \ldots, y_n)$ such that $\forall i$: $y_i \leq x_i$.

Let $\phi(\mathbf{x}) = \prod P_D(x_1 + \cdots + x_i)$. We note $|\mathbf{x}| = x_1 + \cdots + x_n$. Since $dP_S(x)$ is zero when $x < \sigma$, we obtain

$$\int_{|\mathbf{x}| \in [a-\sigma, b]} \phi(\mathbf{x})(dP_S(\mathbf{x}))^n \geq V(\sigma) \int_{|\mathbf{x}| \in [a,b]} \phi(\mathbf{x})(P_S(\mathbf{x})d\mathbf{x})^n \ ,$$

with $V(\sigma) = \int_{|\mathbf{x}| \leq \sigma} dx_1 \cdots dx_n$. A classical result gives

$$\int_{x_1 + \cdots x_n \leq x} dx_1 \cdots dx_n = \frac{x^n}{n!} \ ,$$

which achieves the proof of the lemma, when treating similarly the case with $dp_1(x_1, \cdots, x_n)$. ∎

Partial proof of the conjecture. Let us prove this conjecture in the very case where gaps are independently distributed. We denote $J = (\frac{\gamma C}{\alpha A})^{1/(\alpha-1)}$ and $\delta = \frac{\gamma - 1}{\alpha - 1}$. We denote $R(x)$ the polynomial $x - Jx^\delta$. We will use the following lemma.

Lemma 26 *Let Δ be an arbitrary non-negative real number. Let $G(x)$ the function defined by*

(i) $G(x) = Ax^\alpha + Bx^\beta$, *if* $R(x) \leq \Delta$;

(ii) $G(x) = G(R(x)) + A(x - R(x))^\alpha + Bx^\beta$ *else.*

We have $\limsup G(x)(Cx^\gamma)^{-1} \leq 1$ *when* $x \to \infty$.

Proof. Let x be a non-negative real number larger than Δ. It is clear that the sequence $R^n(x)$ decreases while $R^n(x) \geq 0$. Thus, if $R^n(x) > 0$ for all n, then the sequence $R^n(x)$ converges to 0, since 0 is the only fixed point of $R(x)$ on the non-negative real axis. Therefore there exists n which is the smallest integer such that $R^n(x) > \Delta$ (and $R^{n+1}(x) \leq \Delta$); thus the definition of $G(x)$ is consistent. For $i = 1, \ldots, n$ we denote $x_i = R^{n-i}(x) - R^{n-i+1}(x)$ for $i \geq 2$ and $x_1 = R^n(x)$. Note that $x_1 + \cdots + x_n = x$ and $x_i \geq \Delta - R(\Delta)$ for all i. We have

$$G(x) = \sum_{i=1}^{i=n} A(x_i)^\alpha + B(x_1 + \cdots + x_i)^\beta \ . \tag{11}$$

We want to prove that $G(x) \sim Cx^\gamma$ when $x \to \infty$. The subset of non-negative real numbers such that $R(x) \leq \Delta$ is a compact set $K(\Delta)$. Let x be a non-negative real number outside $K(\Delta)$, we have

$$G(x) - Cx^\gamma = G(R(x)) + A(x - R(x))^\alpha + Bx^\beta - Cx^\gamma \ . \tag{12}$$

Elementary algebra gives the following identities: $\alpha\delta = \beta$, $\beta - \gamma + 1 = \delta$ and $AJ^\alpha + B = \gamma JC$. Therefore

$$G(x) - Cx^\gamma = G(R(x)) - Cx^\gamma(1 - \gamma J x^{\delta-1}) \ . \tag{13}$$

Using the trick $(1-\mu)^a \geq 1 - \mu a$ when $\mu \geq 0$, we get

$$G(x) - Cx^\gamma \leq G(R(x)) - C(R(x))^\gamma \ , \tag{14}$$

and finally

$$G(x) - Cx^\gamma \leq G(R^n(x)) - C(R^n(x))^\gamma \ . \tag{15}$$

Since $G(R^n(x)) = A(R^n(x))^\alpha + B(R(x))^\beta$ and $R^n(x) \in K(\Delta)$, the right hand side of equation (15) is uniformly bounded, which terminates the proof of the lemma. ∎

Let us continue the proof of our conjecture. we start by giving a minoration of $\Pr\{B(t) \geq x\}$ when x tends to infinity.

The average gap duration is $1/\lambda$. Thus, by Chebychev the probability that the gap be less than $2/\lambda$ is greater than $1/2$. Let $v(x)$ a function of x such that $P_S(v(x)) \leq P_S(x)/2$. It is clear that $v(x) \sim x$ when $x \to \infty$.

Let $\varepsilon > 0$ be an arbitrarily small non-negative real number. Let $\Delta > 0$ such that for all $x > \Delta - R(\Delta)$: $2\log 2 - \log P_S(x) \leq (1+\varepsilon)Ax^\alpha$, $v(x) + 2/\lambda \leq x\sqrt[2\beta]{1+\varepsilon}$ and $-\log P_D(x)\lambda \leq \sqrt{1+\varepsilon}Bx^\beta$. Using this Δ we can define function $G(x)$ according to previous lemma. Let us consider x large enough such we can define integer n and n-tuple $(x_1, \ldots, x_n)$ as in the proof of lemma 26.

Let us consider the following event: the n last scheduled customers are yet in purgatory, ranked from the youngest to the oldest one, for all $i \in [1, n)$, the service time of the ith customer is in the interval $[x_i, v(x_i)]$, and all the gaps between them are less than or equal to $2/\lambda$. Let $p(x)$ be the probability of such event. It is clear that $\Pr\{B(t) \geq x\} \geq p(x)$ (To recall $x_1 + \cdots + x_n = x$). On the other hand

$$p(x) \geq (1/2)^n \prod_{i=1}^{i=n} [P_S(x_i) - P_S(v(x_i)] P_D(\sum_{j=1}^{j=i} v(x_j) + 2/\lambda) \ . \tag{16}$$

Using the definition of $v(x)$ we obtain

$$p(x) \geq \prod_{i=1}^{i=n} 1/4 P_S(x_i) P_D(\sum_{j=1}^{j=i} v(x_j) + 2/\lambda) \ . \tag{17}$$

Taking the logarithms, and the definition of Δ we have $-\log p(x)$ smaller than $(1+\varepsilon)G(x)$. Therefore, by lemma 26, and tuning ε arbitrarily small we get $-\log \Pr\{B(t) \geq x\}$ to be at most equivalent to Cx^γ when $x \to \infty$.

Our proof should have ended there if our parameter of interest was only $B(t)$. But we are interested into the limiting distribution of working load $Q(t)$ and waiting time W. To end our proof we use similar analysis than for $B(t)$.

Let us tune Δ such that for all $x \geq \Delta - R(\Delta)$: $3\log 2 - \log P_S(x)(1+\varepsilon)Ax^{\alpha}$, the other conditions remain unchanged. With this new Δ we define a new function $G(x)$.
Let $d(x)$ be the mean quantity $D - x$ given that $D \geq x$. we have

$$d(x) = 1/P_D(x) \int_x^{\infty} P_D(y)dy \ .$$

We readily have $d(x) \sim B^{-1}x^{1-\beta}$. There exists a such that for all $x \geq 0$: $\forall y \leq x : \ d(y) \leq a + d(x)$.

For large enough x, let us consider the following event:

(i) at time t, the n last scheduled customers are yet in purgatory, their service times are respectively in $[x_i, v(x_i)]$ and the gaps between them are all less than $2/\lambda$;

(ii) at time $t + 2(a + d(x))$ these n customers are queued.

Let $q(x)$ the probability of such event. It is clear that $q(x)$ minorizes $\Pr\{Q(t) \geq x - 2(a + d(x))\}$ and $\Pr\{W \geq x - 2(a + d(x))\}$ since at time $t + 2(a + d(x))$, the workload $Q(t)$ is necessarily greater than $x - 2(a + d(x))$ and at least one of the customer have waiting time greater than $x - 2(a + d(x))$ because of FIFO strategy.

For each of the n last customers in purgatory at time t to be queued before time $t + 2(a + d(x))$ is greater than $1/2$, by Chebichev. Thus we have the minoration

$$q(x) \geq \prod_{i=1}^{i=n} 1/8 P_S(x_i) P_D(\sum_{j=1}^{j=i} v(x_j) + 2/\lambda) \ . \tag{18}$$

Using the new definition of Δ we get $-\log q(x)$ smaller than $(1+\varepsilon)G(x)$. That allows us to conclude that both $-\log \Pr\{Q(t) \geq x\}$ and $-\log \Pr\{W \geq x\}$ are at most equivalent to Cx^{γ} when $x \to \infty$ (since $x - 2(a + d(x)) \sim x$). ∎

Remark: In the case where service times are all uniformly bounded for the above by a constant Σ, the conjecture is easier to prove by simply taking $x_i = \Sigma$ and following similar analysis.

ISSN 0249-6399