Journal of Applied Mathematics and Stochastic Analysis, **13**:4 (2000), 429-450.

SINGLE SERVER QUEUEING NETWORKS WITH VARYING SERVICE TIMES AND RENEWAL INPUT

PIERRE LE GALL
France Telecom, R&D
4 Parc de la Bérengère
F-92210 Saint-Cloud, France

(Received January, 1999; Revised December, 1999)

Using recent results in tandem queues and queueing networks with renewal input, when *successive* service times of the same customer are *varying* (and when the busy periods are frequently *not broken up* in large networks), the local queueing delay of a *single server* queueing network is evaluated utilizing new concepts of *virtual* and *actual* delays (respectively). It appears that because of an important property, due to the *underlying* tandem queue effect, the usual queueing standards (related to long queues) cannot protect against significant overloads in the buffers due to some possible "*agglutination phenomenon*" (related to short queues). Usual network management methods and traffic simulation methods should be revised, and should monitor the partial traffic streams loads (and not only the server load).

Key words: Queueing Networks, Concentration Tree, Tandem Queue, Local Queueing Delay, Jitter Delay, Agglutination Phenomena, Apparent Queueing Delay, Buffer Overload.

AMS subject classifications: 60K25, 90B22.

1. Introduction

In this paper, we utilize recent results in tandem queues and queueing networks to evaluate the local queueing delay in *single server* queueing networks (excluding ring networks) with *renewal input*, when *successive* service times of the same customer are *varying*, and when busy periods are frequently *not broken up* in large networks (leading to the useful concept of equivalent tandem queue). We assume that customers only gain access to a downstream queue after completion of the upstream service. At each queue, the service discipline is "*first-come-first served*" (FC-FS).

Classical theories (i.e., the product form theory) ignore some queueing phenomena which occur between entry to the network and the considered local point, and influence the local queue. In particular, *the influence of the indistinguishability of the*

customers inside the upstream busy periods affects the distribution towards a given destination. Markovian queueing networks typically use the concept of local transition rates independent of upstream interferences. *For converging traffic streams*, busy periods tend to amalgamate from state-to-stage, leading to a *predominant influence of the longest service times.* In the same way, this indistinguishability in the upstream busy periods leads to a new concept of a local *apparent* queueing delay related to the overall upstream queueing delay. Here, only a part of the delay observed upstream is perceived downstream, leading to some *reduction factor for diverging traffic streams*, and to new concepts of *virtual* and *actual* local queueing delays, respectively.

Curiously and due to this reduction factor, this last phenomenon of indistinguishability tends to assimilate the long local queue to a classical G/G/1 server queue, and the usual queueing standards may only use the G/G/1 server model. This does not mean that the actual queueing model can be represented simply by a G/G/1 server. On the contrary, it may be dangerous to dimension the buffers using only this G/G/1 server model, particularly in case of infinite support for the service time distribution. In the network, servers are occupied by *service time.* But in the buffers, servers are occupied by *sojourn time* (i.e., the sum of the queueing delay and the service time). Short and medium service times may *overload buffers due to the agglutination phenomenon of short service times behind long service times*, leading to the same occupancy for long and short service times. Since busy periods tend to amalgamate, this phenomenon is amplified from stage-to-stage, leading to a *strong influence of the longest service times* with a necessarily limited length.

If the methods of buffer dimensioning have to be revised, the same follows fro *network management methods in which partial traffic streams loads* (i.e., traffic intensities) *should be monitored* (and not only the server loads). This avoids some fast buffer congestion generalizing (in a large area of the network) the inaccessibility to the servers, since the agglutination phenomenon is transmitted downstream immediately.

In Section 2, we define our notation and assumptions. In Section 3, we outline our earlier studies in tandem queues and queueing networks for the case of *single servers* with renewal input. In Section 4, we deduce the evaluation of the local queue distribution in single server queueing networks (with renewal input), with an application to *buffer* ***overload*** and *buffer* ***rejection rate.*** In Section 5, we conduct a numerical study (with *Poisson arrivals*) for a single link packet switched network, in which the links handle three populations of packets (with constant packet lengths). Between these populations, packet lengths are very different, proving the strong impact of the longest packets on buffer overload and buffer rejection rate.

2. Notation and Assumptions

We assume the system is in the *stationary regime.*

2.1 The Local Queue

The total arrival process at a local queue (considered at the entry to the network) is assumed to be *renewal*, with $\boldsymbol{F_0(t)}$ denoting the distribution function of the *interarrival intervals.* Local service times are mutually independent and independent of the

arrival process. The *local queueing delay* is usually assimilated to the queueing delay $\boldsymbol{W_0}$ of a GI/G/1 queue, with the following data for an *arbitrary* customer:

- arrival rate λ;
- *local* service time $\boldsymbol{T}$;
- $\text{Prob}(\boldsymbol{T} < t) = \boldsymbol{F_1(t)}$;
- *local* sojourn time $\boldsymbol{S} = W_0 + T$;
- overall *upstream* service time $\boldsymbol{T'}$;
- load (= traffic intensity) $= \boldsymbol{\rho} = \lambda \cdot T$. (1)

Note that the upstream service times are not necessarily independent of the downstream service times. We set for $\text{Re}(z) \geq 0$:

- $\boldsymbol{\varphi_0(z)} = \int_0^\infty e^{-zt} \cdot dF_0(t)$;
- $Ee^{-zT} = \boldsymbol{\varphi_1(z)} = \int_0^\infty e^{-zt} \cdot dF_1(t)$;
- $Ee^{-zW_0} = \mathcal{W}_0(z)$;
- $Ee^{-zS} = \boldsymbol{\Phi_1(z)} = \mathcal{W}_0(z) \cdot \varphi_1(z)$. (2)

To present the following expressions, we will use Cauchy contour integrals along the imaginary axis in the complex plane $\boldsymbol{u}$. If the contour (followed from the bottom to the top) is to be right of the imaginary axis (the contour being closed at infinity to the right), we write $\int_{+0}$. If the contour is to the left of the imaginary axis, we write $\int_{-0}$. Pollaczek [6] gave the expressions:

$$\mathcal{W}_0(z) = \text{Exp}\left\{\frac{-1}{2\pi i} \int_{-0} \left[\frac{1}{z-u} + \frac{1}{u}\right] \cdot \log[1 - \varphi_0(-u) \cdot \varphi_1(u)] \cdot du\right\},$$
$$\text{Prob}(W_0 = 0) = \boldsymbol{Q_1} = \text{Exp}\left\{\frac{-1}{2\pi i} \int_{-0} \log[1 - \varphi_0(-u) \cdot \varphi_1(u)] \cdot \frac{du}{u}\right\}. \tag{3}$$

For tandem queues, it will be useful to introduce these other expressions, related to the case where $T < t$:

$$\varphi_1(z,t) = \int_0^t e^{-z\alpha} \cdot dF_1(\alpha),$$
$$Q_1(t) = \text{Exp}\left\{\frac{-1}{2\pi i} \cdot \int_{-0} \log[1 - \varphi_0(-u) \cdot \varphi_1(u,t)] \cdot \frac{du}{u}\right\}. \tag{4}$$

Consequently, we have:

$$Q_1 = Q_1(\infty). \tag{5}$$

2.2 The Queueing Network

In Figure 1, representing a *full network*, a number $\boldsymbol{n}$ of identical incoming paths of $\boldsymbol{m}$ successive servers are distributed [at stage $(m+1)$] upon a number $\boldsymbol{n}$ of final servers, defining the final stage. The traffic stream $\boldsymbol{A_i(j)}$ is carried by the $\boldsymbol{i}$th incoming path, and then by the $\boldsymbol{j}$th final server.

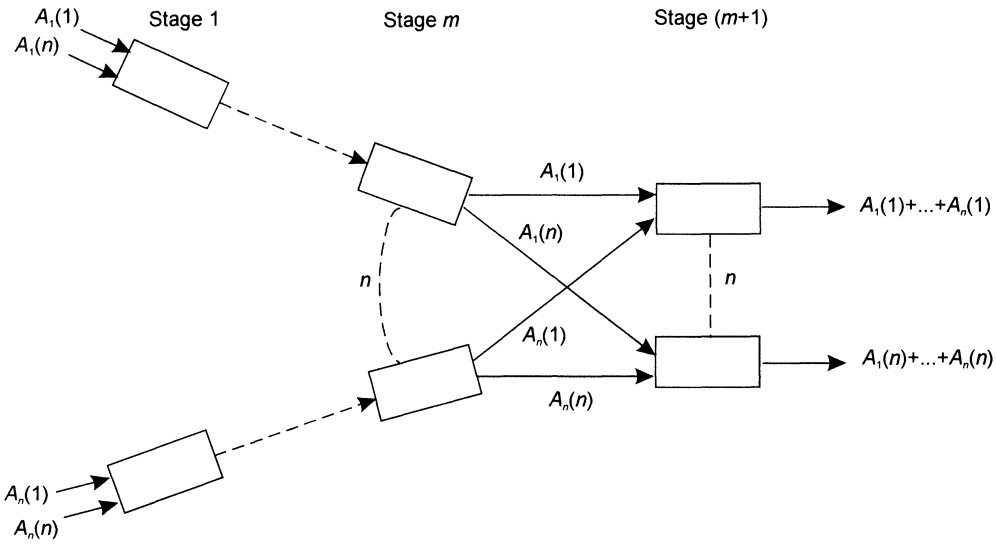

Figure 1: The Full Network

In Le Gall [1], we observe that traffic streams *crossing upstream* have no significant influence on the queueing delay at the final stage (due to our assumption that the busy periods are *not broken up* for m large), an important case in which a local arrival initiating a local busy period made the same upstream. We therefore may *eliminate* these streams. This also holds for *intermediate arrivals* along incoming paths, since their influence is not changed (at the final stage) by assuming that these customers arrive at the entry to the network, and correspond to the service times equal to zero in the first stages.

Our study may be simplified by introducing the *truncated network* shown in Figure 2. At the final stage, we consider only a single server. This server handles the traffic stream $\boldsymbol{A_i}$ coming from the ith incoming path ($i = 1...n$). In this incoming path, the traffic stream interferes with a traffic stream $\boldsymbol{B_i}$, which is *not offered to the final server.* Its customers may be considered as "*premature departures*" *affecting the final local queue*, a phenomenon usually neglected.

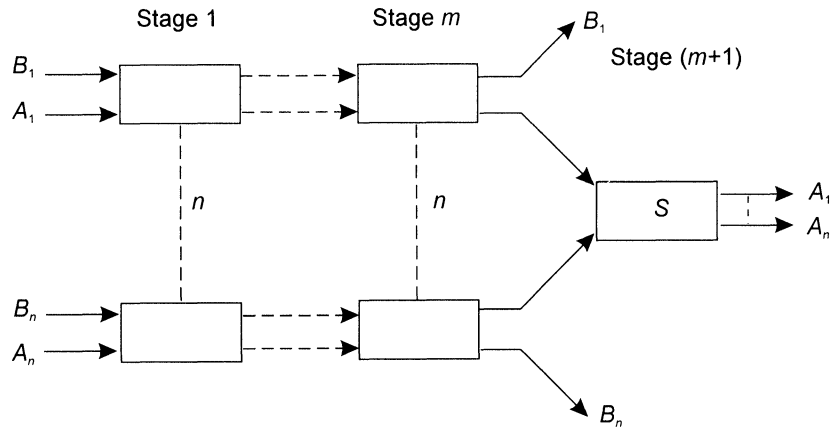

Figure 2: The Truncated Network

In the general case, we have no symmetry. The number m_i of successive servers in the ith incoming path is different in each path, and successive servers may generate different service times (with different distributions). In fact, it will be possible to define a hypothetical network of Figure 2 due to some following properties.

First, we will not consider the special impact of "*premature departures*", related to a *reduction factor* (due to the concept of "apparent" upstream delays), to define an *equivalent tandem queue* with a number m of upstream stages (as defined by a certain relation depending on the number m_i and the overall *upstream* service time T') *on the condition that the busy periods are not broken up.* Moreover, if successive service times of the same customer are not too widely varying compared with the local queueing delays, this equivalent tandem queue may be assimilated to a theoretical packet switched tandem queue in which successive service times are identical. This simplifies the calculations *on the condition that the number* m *of identical successive servers is defined correctly.* In that condition, we define a ***virtual*** *local* queueing delay, *independent* of the considered incoming path. This great simplification cannot exist (for m small) when the busy periods are broken up.

Secondly, we will consider the special impact of "*premature departures*" (due to the concept of "*apparent*" upstream delays) to define the ***actual*** *local* queueing delay. Consequently, a *reduction factor* will appear to lead (in part) towards the GI/G/1 queue. This reduction factor (as a function of B_i and the considered incoming path) may generate the same influence as a hypothetical number of identical incoming paths in Figure 2 truncated network to replace the actual number n. Consequently, Figure 2 will be a reference figure in our paper.

3. Preliminary Theory

3.1 The Equivalent Tandem Queue and the Virtual Local Queueing Delay

Consider the case of a concentration tree (Figure 2) with traffic streams A_i and B_i, with an identical traffic intensity at each stage, without taking into account a certain

reduction factor due to the concept of "apparent" upstream delays (to be considered in Subsection 3.3). Consequently, we define a ***virtual*** *local* queueing delay as considered in Le Gall [1].

Each incoming path is a tandem queue, with the following notations for the ***h***th *customer at stage* $k = 1 \ldots m$:

- local queueing delay w_h^k;
- local service time T_h^k;
- local sojourn time $s_h^k = w_h^k + T_h^k$;
- interarrival interval [between customers $(h-1)$ and h] $\boldsymbol{Y_{h-1}^k}$;
- occasional idle period [during Y_{h-1}^k] $\boldsymbol{e_h^k}$.

In other words, we may write:

$$Y_{h-1}^k = T_h^{k-1} + e_h^{k-1} \tag{6}$$

Moreover, we let for $k = 2 \ldots m$:

$$\begin{aligned} T_h' + \ldots + T_h^k &= T_h^1(k) \\ s_h^2 + \ldots + s_h^k &= S_h^{k-1}. \end{aligned} \tag{7}$$

In Le Gall [2], we recall that this concentration tree (with the same traffic intensity at each stage) gives the same local queueing delay distribution (at the final stage) as does an *equivalent tandem queue*, concerning an *arbitrary customer* (coming from an arbitrary incoming path) with the same *local* service time $\boldsymbol{T}$, and the same *upstream* overall service time $\boldsymbol{T'}$ [notation (1)]. To simplify the calculations, *when the busy periods are not broken up* (during the upstream busy periods), we defined an *equivalent tandem queue with* $\boldsymbol{(m_0+1)}$ *successive* ***identical*** *single servers* (as in packet switching), where $\boldsymbol{m_0}$ is given by the relation:

$$\mathrm{Var}(\boldsymbol{m_0} \cdot T) = \mathrm{Var}\boldsymbol{T'}, \tag{8}$$

if $\boldsymbol{T}$ and $\boldsymbol{T'}$ are not constant. We deduce:

$$\boldsymbol{m_0^2} = \frac{\mathrm{Var}T'}{\mathrm{Var}T}. \tag{9}$$

When $\boldsymbol{T}$ and $\boldsymbol{T'}$ are *highly varying*, we have: $\mathrm{Var}(m_0 \cdot T) \cong m_0^2 \cdot ET^2$, $\mathrm{Var}T' \cong \boldsymbol{ET'^2}$, and consequently,

$$m_0^2 = \frac{\boldsymbol{ET'^2}}{ET^2}. \tag{10}$$

In queueing formulae, when $\boldsymbol{m_0}$ is between two successive integers, we will use an interpolation between delays related to these integers, or directly the possible fractional $\boldsymbol{m_0}$ in formulae. In Le Gall [2], relation (3), and in Le Gall [3], relation (7), we gave the following condition for busy periods not broken up:

Hypothesis 1: (Busy periods not broken up) *We assume that the following relation is satisfied:*

$$T_h^{k-1} \leq s_{h-1}^k, \quad k = 2 \ldots (m+1). \tag{11}$$

This is the case when:

(*a*) All successive service times of the same customer are *identical* (because in this case: $T_h^{k-1} = T_h^k < s_h^k = s_{h-1}^k$); and

(*b*) In *heavy load*, congestion becomes high enough to increase s_{h-1}^k.

This hypothesis is not very restrictive. To simplify in the sequel, *we shall replace* $\boldsymbol{m_0}$ *by* $\boldsymbol{m}$. In that case, we recall in Le Gall [2] that we get th following relation at stage $(m+1)$:

$$(T_h^1 + w_h^2) + \ldots + (T_h^m + w_h^{m+1}) = \mathrm{Max}[T'_h(m), S_{h-1}^m - e_h^1]. \tag{12}$$

The left-hand side may be written:

$$S_h^m + (T_h^1 - T_h^{m+1}).$$

To simplify calculations, we introduce a second hypothesis (see Le Gall [2]), which is usually satisfied:

Hypothesis 2: (Limitation to successive service time variations) *If* ε *is an arbitrary small positive number, we suppose that the following condition*

$$\lim_{m \to \infty} \mathrm{Prob}\left(\left| \frac{T_h^1 - T_h^{m+1}}{T_h^2 + \ldots + T_h^{m+1}} \right| < \varepsilon \right) \to 1, \tag{13}$$

is satisfied for any $\boldsymbol{h} > 0$.

In that case, recurrence relation (12) is equivalent (in probability for $\boldsymbol{m}$ large enough) to the stochastic recurrence [notation (7)]:

$$S_h^m = \mathrm{Max}\left[T'_h(m), S_{h-1}^m - e_h^1\right]. \tag{14}$$

Note that this equivalence could be also valid (for $\boldsymbol{m}$ large) in the case of mutual independence for successive service times, if T_h^k $(k = 2, \ldots, m+1)$ is replaced by s_h^k in the denominator of (13), at heave load (s_h^k being high).

Finally, Hypotheses 1 and 2 allow handling of the (almost) general case by the simple packet switched case. In Le Gall [5], we gave the *virtual* local sojourn time distribution (at the final stage). We let, from notations (1) and (4)

$$\boldsymbol{v_0(t)} = \frac{Q_1}{Q_1(t)} \cdot F_1(t), \quad \boldsymbol{v(t,u)} = \mathrm{Exp}\left\{ -u \cdot \int_t^\infty \frac{1 - F_1(v)}{Q_1(v)} \cdot dv \right\},$$

$$\boldsymbol{u_m(t,u)} = v_0(t) \cdot [v(t,u)]^m, \tag{15}$$

$$\boldsymbol{d_m(t,u)} = \int_0^t v\left(\frac{t+w}{m}, u\right) \cdot d_w u_{m-1}\left(\frac{w}{m-1}, u\right).$$

Finally, a distribution function of the ***virtual*** *local sojourn time* $\boldsymbol{U(m)}$, at final stage $(m+1)$, and for an arbitrary customer, is from notations (2) and (3):

(*a*) *Case of renewal input*:

$$\boldsymbol{U(t,m)} = \frac{1}{2\pi i} \cdot \int_{+0} \frac{\varphi_0(-u) \cdot \Phi_1(u)}{Q_1} \cdot d_m(t,u) \cdot \frac{du}{u} \tag{16}$$

(*b*) *Case of Poisson input*:

$$\boldsymbol{U(t,m)} = d_m(t,\lambda) \cong v_0(t) \cdot \left[v\left(\frac{t}{m}, \lambda\right) \right]^m \tag{17}$$

In Le Gall [5], we also gave an approximated expression for relation (17), for the situation when $\boldsymbol{m}$ increases. Consider *the longest service time* $\boldsymbol{T_N}$, and the service times almost so long (total arrival rate: $\boldsymbol{\lambda_N}$; total load: $\boldsymbol{\rho_N}$). Since (λ_N/λ) is low, the approximation for $\boldsymbol{U(t,m)}$, independent of any segmentation of service time with medium lengths, is:

for $0 < t \leq T_N$:

$$\boldsymbol{U_1(t,m)} = \left(1 - \frac{\lambda_N}{\lambda}\right) \cdot \frac{1-\rho}{1-(\rho-\rho_N)} \cdot \exp\left\{ \frac{-m\rho_N}{1-(\rho-\rho_N)} \cdot \left(1 - \frac{t}{m \cdot T_N}\right) \right\}; \tag{18}$$

for $t > T_N$:

$$\boldsymbol{U_1(t,m)} = 1.$$

The impact of the longest service times comes from the "*agglutination phenomenon*". During any busy period, relation (14) gives, when $e_h^1 = 0$: $S_h^m = S_{h-1}^m = \ldots = S_{h_0}^m$, where $\boldsymbol{h_0}$ corresponds to the customer initiating the busy period. The *sojourn time is the same for any customer of the same busy period.* A busy period initiated by a long service time, leading to long sojourn times inside this busy period, tends to amalgamate with subsequent busy periods. From stage to stage, the phenomenon is amplified, leading to a strong impact of the longest service times. This agglutination phenomenon leads to a local sojourn time independent of the considered incoming path. This also follows for the ***virtual*** local queueing delay.

Note: In the case of infinite support for the service time distribution function $\boldsymbol{F_1(t)}$, we have seen in Le Gall [1], Subsection 3.3, that a stationary condition exists when $[1 - F_1(t)]$ decreases asymptotically as a negative exponential distribution. This is not the case for a *Pareto* distribution (corresponding to "*Fractal*" processes), which decreases asymptotically as $(at)^{-\alpha}$, $(\alpha > 2)$. Consequently, it will be the same in a queueing network, in which a "*Pareto*" distribution cannot be handled (on the contrary of a single GI/G/1 server). Practically, we will restrict this paper to the case of finite support, in which $\boldsymbol{T_N}$ is the longest service time.

3.2 The Jitter Effect

The equivalent tandem queue keeps the same order of arrivals at the entry to the network, and at final stage $(m+1)$. The actual arrival process at this final stage is different, since the incoming traffic streams are mutually independent. This difference generates a *local jitter effect.* independent of the above queueing delay, which was evaluated approximately in Le Gall [2], Subsection 3.1, for large $\boldsymbol{n}$ (> 5). The

distribution function of this local jitter delay is:

for $0 < t \leq T'' - T$:

$$\boldsymbol{J(t)} = \frac{1-\rho''}{1-\rho'' \cdot \frac{t+T}{T''}}, \text{ with } T'' = \frac{T}{1-\rho''}, \rho'' = (1-\tfrac{1}{n}) \cdot \rho.$$

for $t > T'' - T$:

$$\boldsymbol{J(t)} = 1. \tag{19}$$

The accuracy is better when the load of long service times is lower than 0.3ρ.

This jitter effect is only significant for heavily loaded networks.

3.3 The Impact of Premature Departures and the Actual Local Queueing Delay

We now consider the special impact of "*premature departures*" at the final stage $(m+1)$ to define the ***actual*** local queueing delay. Without considering this special impact, and excluding the jitter effect, the *virtual* local sojourn time at stage $(m+1)$, as perceived at the entry to the network, is:

$$\boldsymbol{U(m)} = (W_0 + T) + \boldsymbol{S(m)}, \tag{20}$$

where $\boldsymbol{S(m)}$ is due to the $\boldsymbol{m}$ supplementary (upstream) stages. In Le Gall [2, 3], we noted that the overall delays observed upstream are defined ***without distinguishing*** *customers in the upstream busy periods.* In the case of "*premature departures*", the final queue only perceives ***apparent*** delays. In the ith incoming path (at the upstream stage), we introduce:

- total load $\boldsymbol{a_i}$ (excluding the load of cross-traffic streams);
- part of this total load offered to the final queue $\boldsymbol{a'_i}$.

For an *arbitrary* customer (coming from the ith and from any incoming path, respectively), we introduce the following ratios (defining the numbers $\boldsymbol{n'_i}$ and $\boldsymbol{n_1}$, respectively), replacing the number $\boldsymbol{n}$ in Figure 2:

$$\frac{1}{n'_i} = \frac{a'_i}{a_i}; \quad \frac{1}{n_1} = \sum_i \frac{a'_i}{a} \cdot \left(\frac{a'_i}{a_i}\right), \text{ with } a = \sum_i a'_i. \tag{21}$$

Finally, ***from the*** ith ***incoming path***, the final queue perceives the ***apparent*** delay $h_i \cdot \boldsymbol{S(m)}$, where h_i is a random number $= 1$, with probability $\left(\frac{1}{n'_i}\right)$, and $= 0$ with probability $\left(1 - \frac{1}{n'_i}\right)$. Stochastic expression (20) becomes, for the ***actual*** local sojourn time S_i including the jitter delay J, with $D(m) = U(m) + J$:

$$\boldsymbol{S_i} = h_i \cdot \boldsymbol{D(m)} + (1 - h_i) \cdot (\boldsymbol{W_0} + \boldsymbol{T}),$$

$$\boxed{Ee^{-zS_i} = \left(\frac{1}{n'_i}\right) \cdot Ee^{-zD(m)} + \left(1 - \frac{1}{n'_i}\right) \cdot Ee^{-z(W_0+T)}}. \tag{22}$$

Finally, in Le Gall [2, 3], we gave the following expression defining the ***actual*** *local*

sojourn time S at final stage $(m+1)$, for $Re(z) \geq 0$, and *from any incoming path*:

$$\boxed{Ee^{-zS} = \left(\frac{1}{n_1}\right) \cdot Ee^{-zD(m)} + \left(1 - \frac{1}{n_1}\right) \cdot Ee^{-z(W_0 + T)}}, \tag{23}$$

with

$$D(m) = U(m) + J, \tag{24}$$

and

W_0 = queueing delay of the GI/G/1 queue; J = Jitter delay,

$U(m)$ = ***virtual*** local sojourn time of the equivalent tandem queue (identical for each customer of the same busy period).

Conclusions:

(a) In case of a ***heavily*** *loaded environment* $\left(a_i' \cong 0 \rightarrow \left(\frac{1}{n_1}\right) \cong 0\right)$, (i.e., in case of ***traffic streams diverging***, and consequently generating a lot of "premature departures"), we have for the *actual* local queueing delay w: $Ee^{-zw} \cong Ee^{-zW_0}$. This is a classical result.

(b) In case of a ***slightly*** *loaded environment* $\left(a_i' \cong a_i \rightarrow \left(\frac{1}{n_1}\right) \cong 1\right)$, (i.e., in case of no "premature departures"), the *tandem queue effect* appears with the ***agglutination phenomenon***, increasing *buffer overloads*, but the value of $U(m)$ is different since the upstream traffic intensities become lower.

(c) *In the general case* (in stationary regime), and due to the concept of *apparent* queueing delay used above, we can state the ***basic*** following *properties:*

(α) *An* ***arbitrary*** *local customer can see the* ***tandem queue effect*** *(with the* ***agglutination phenomenon*** *and the buffer overload) with a probability* $(1/n_1)$, *and the classical* ***GI/G/1 queue*** *in the other cases.* This probability corresponds to the case of successive local customers coming from the same incoming path.

(β) Despite the breaking up of the traffic streams at each stage, the aggregation of small parts of upstream busy periods gives rise to a *new local busy period corresponding to the maximum upstream sojourn times* (see our comment on the agglutination phenomenon at the end of Subsection 3.1). Consequently, *the agglutination phenomenon is amplified from stage to stage.* In fact, *inside a local busy period*, any customer appears to be ***indistinguishable***, and it follows that the concentration tree may be assimilated to a single equivalent tandem queue. This explains why *the tandem queue effect increases from stage to stage*, for a given probability $(1/n_1)$ to perceive this effect.

Note: This tandem queue effect is generated by the *equivalent tandem queue*, which has to satisfy *Hypothesis* 1 (busy periods not broke up) *and* 2 (limitation to successive service time variations). Consequently, the upstream loads considered are equal to the final local load, since successive *equivalent* servers are *identical* to the final server, in case of a normally loaded environment.

(d) When we may approximately use Hypothesis 2, in the case of *mutual*

independent successive service times, the tandem queue effect exists for m large, since the upstream service time (in case of congestion) is also the downstream interarrival time (for two successive arrivals from the same incoming path). This property is not consistent with Jackson's theory (due to the indistinguishability of customers inside any local busy period).

(*e*) Expression (22) proves the need to monitor the partial traffic streams loads, and not just the server load (i.e., traffic intensity). This is the same for traffic simulation methods. A direct observation of the local queueing delay cannot detect the possible correlation of the local interarrival time with the upstream service time (i.e, *tandem queue effect*). Finally, a classical GI/G/1 queue may be observed instead of the tandem queue effect. To avoid this difficulty, it is necessary to directly observe the *difference* between the two successive (upstream and local) *overall sojourn times* for a ***given traffic stream.***

After having outlined our earlier papers, we now can define and evaluate the local queueing delay in a single server queueing network with varying service times, as well as for a renewal input (at the entry to the network) related to the local queueing process.

4. The Local Queue in the Network

We will evaluate the local queue distribution, the buffer overload, and the rejection rate in the local buffer.

4.1 The Local Queue Distribution

The ***actual*** *local sojourn time* S is defined by expression (23), referring to the classical queueing delay W_0 of the local GI/G/1 queue, and to the ***virtual*** local sojourn time $D(m)$, sum of the *jitter delay*J and the *virtual* local sojourn time $U(m)$ of the *equivalent tandem queue* [J and $U(m)$ being mutually independent]. The distribution of $U(m)$ is defined by expressions (15) and (16). When m increases, we may simplify the expression $d_m(t,u)$, by introducing the following approximations:

$$v\left(\frac{t+w}{m},u\right) \cong v\left(\frac{t}{m},u\right); \quad u_{m-1}\left(\frac{w}{m-1},u\right) \cong u_{m-1}\left(\frac{w}{m},u\right).$$

For $v_0(t)$ relating to the case $m = 1$ only, we may write, finally from expression (16):

$$d_m(t,u) \cong v_0(t) \cdot \left[v\left(\frac{t}{m},u\right)\right]^m. \tag{25}$$

In the case of *Poisson arrivals*, we have only one pole [for $Re(u) > 0$] in the integrand of expression (16): $u = \lambda$. Expression (25) above leads to the approximations in expressions (17) and (18).

4.2 The Approximated Expression of the Distribution

Expressions (1) define the data, related to an *arbitrary local* customer, defining the local GI/G/1 queue. As for expression (18), we suppose that T_N corresponds to the *longest service times* (in case of finite support), and we define the total *arrival rate*

λ_N (and the total load: ρ_N) for the local arbitrary customers corresponding to a service time closed to T_N. We want to evaluate expression (25) for t *closed to* T_N. We introduce $\varphi_1^*(z)$ and Q_1^*, *when the customers corresponding to the set* (λ_N, T_N) *are excluded.* From expression (4), we have:

$$F_1(t) = \frac{\sum_{j=1}^{N-1} \lambda_j}{\sum_{j=1}^{N} \lambda_j} = 1 - \frac{\lambda_N}{\lambda},$$

$$\varphi_1(z,t) = \frac{\sum_{j=1}^{N-1} \lambda_j}{\sum_{j=1}^{N} \lambda_j} \cdot \varphi_1^*(z) = \left(1 - \frac{\lambda_N}{\lambda}\right) \cdot \varphi_1^*(z). \tag{26}$$

Usually, T_N is much longer than $E(T)$. Consequently, (λ_N/λ) may be neglected, and we may write:

$$\varphi_1(z,t) \cong \varphi_1^*(z), \quad Q_1(t) \cong Q_1^*. \tag{27}$$

Finally, expression (25) becomes:

$$d_m(t,u) \cong \left(1 - \frac{\lambda_N}{\lambda}\right) \cdot \frac{Q_1}{Q_1^*} \cdot \exp\left\{ -\frac{u}{\lambda} \cdot \frac{m \cdot \rho_N}{Q_1^*} \cdot \left[1 - \frac{t}{m \cdot T_N}\right]\right\}. \tag{28}$$

For $u = \lambda$, in the case of Poisson arrivals, expression (28) leads to approximation (18) and, in Le Gall [5], we have seen that this approximation (18) is practically valid for any time t. Consequently, *expressions* (16) *and* (28) *are useful approximations to evaluate the distribution of the* ***virtual local sojourn time*** $U(m)$ for any time t.

4.3 The Buffer Load

First, we want to evaluate the mean $\overline{U(m)}$. We let:

$$A(u) = \frac{\rho_N}{Q_1^*} \cdot \frac{u}{\lambda}. \tag{29}$$

In expression (16), if we replace $d_m(t,u)$ by one, $U(t,m)$ becomes equal to one. Consequently,

$$1 - U(t,m) = \frac{1}{2\pi i} \cdot \int_{+0} \frac{\varphi_0(-u) \cdot \Phi_1(u)}{Q_1} \cdot [1 - d_m(t,u)] \cdot \frac{du}{u}. \tag{30}$$

We let:

$$\overline{U(m)} = \int_0^{T_N} [1 - U(t,m)] \cdot dt, \tag{31}$$

$$D(m,u) = \int_0^{T_N} [1 - d_m(t,u)] \cdot dt.$$

From expressions (16), (28) and (29), we deduce:

$$\overline{U(m)} = \frac{1}{2\pi i} \cdot \int_{+0} \frac{\varphi(-u) \cdot \Phi_1(u)}{Q_1} \cdot \boldsymbol{D_m(u)} \cdot \frac{du}{u},$$

with (32)

$$\boldsymbol{D_m(u)} = \boldsymbol{T_N} \cdot \left\{ 1 - \left(1 - \frac{\lambda_N}{\lambda} \right) \cdot \frac{Q_1}{Q_1^*} \cdot \frac{\mathrm{Exp}[-(m-1) \cdot A(u)] - \mathrm{Exp}[-m \cdot A(u)]}{A(u)} \right\}.$$

Finally, taking definitions (1) and relations (23) into account, we deduce the *mean* ***actual*** *local sojourn time* [at stage $(m+1)$] of an *arbitrary* customer, corresponding to his *occupancy in the local buffer*:

$$\boldsymbol{s_1(m+1)} = \frac{1}{n_1} \cdot \left[\overline{U(m)} + \overline{J} \right] + \left[1 - \frac{1}{n_1} \right] \cdot \left[\overline{W_0} + \overline{T} \right], \tag{33}$$

with $\boldsymbol{\overline{U(m)}}$ being given by expression (32). In the same way, we could evaluate the second moment.

4.4 The Buffer Rejection Rate

If $\boldsymbol{K}$ is the *buffer capacity*, a customer is rejected on his local arrival if the number of customers $\boldsymbol{j}$ waiting is such as: $j \geq K-1$. If $\boldsymbol{j}$ denotes the number of customers in the local system (waiting or with service in progress), the rejection condition becomes: $j \geq K$. We suppose that a rejected customer repeats his arrival at the entry to the network. It follows that traffic handled locally is not decreasing, with the queue length distribution giving a good approximation of the rejection rate, when its value is low.

In conclusion (*c*) of Subsection 3.3, we noted that an arbitrary local customer can see the tandem queue effect (*with its rejection rate*) with a probability $(\boldsymbol{1/n_1})$, including the jitter effect, and the classical GI/G/1 queue (*with its rejection rate*) in the other cases. We introduce the following notation for an *arbitrary* local customer (in stationary regime):

- $\boldsymbol{R_0(K)}$ rejection rate due to the GI/G/1 queue;
- $\boldsymbol{R_1(K)}$ rejection rate due to the tandem queue;
- $\boldsymbol{R(K)}$ the total rejection rate.

Note that the rejection rate due to the jitter effect may be neglected. Due to our comment above, and using relation (23), we may write:

$$\boldsymbol{R(K)} = \left(\frac{1}{n_1} \right) \cdot R_1(K) + \left(1 - \frac{1}{n_1} \right) \cdot R_0(K). \tag{34}$$

(*a*) ***Evaluation of*** $\boldsymbol{R_0(K)}$. For the classical, local GI/G/1 queue, we recall our notation (2). $\boldsymbol{F_0(t)}$ is the distribution function of the *interarrival intervals* (at the entry to the network). $\boldsymbol{W_0(t)}$ is the distribution function of the local queueing delay. Expression (35) in Le Gall [4] gives:

$$\boldsymbol{R_0(K)} = \int_0^{\infty} [1 - W_0(t)] \cdot d[F_0(t)]^{(K-1)}, \tag{35}$$

where $\boldsymbol{[F_0(t)]^{(K)}}$ denotes the $\boldsymbol{K}$-fold convolution of $F_0(t)$.

(*b*) ***Evaluation of*** $\boldsymbol{R_1(K)}$. Following our comment after expression (18), related to the *agglutination phenomenon* of short service times behind a long service time initiating a local busy period, the interarrival times $\boldsymbol{T^*}$, during this busy period, correspond to the service times (excluding the longest service times $\boldsymbol{T_N}$). Due to notations (26) and (27), we denote by $\boldsymbol{F_1^*(t)}$ the service time distribution function, excluding the set (λ_N, T_N). From notation (6), the *rejection condition*, at stage $(m+1)$, becomes:

$$\begin{gathered} s_h^{m+1} - (Y_h^{m+1} + \ldots + Y_{h+K-1}^{m+1}) > 0, \\ (\text{i.e., } s_h^{m+1} - \boldsymbol{K} \cdot \boldsymbol{T^*} > 0), \end{gathered} \tag{36}$$

which leads to expression:

$$\begin{aligned} \boldsymbol{R_1(K)} &= \int_0^{T_N} [1 - U(t,m)] \cdot d[F_1^*(t)]^K, \text{ if } K \leq \frac{T_N}{T_1}; \\ &= 0, \qquad\qquad \text{if } K > \frac{T_N}{T_1}, \end{aligned} \tag{37}$$

where $[F_1^*(t)]^K$ denotes the K-fold convolution of $\boldsymbol{F_1^*(t) \cdot T_1}$ corresponds to the shortest service times. $\boldsymbol{U(t,m)}$ is defined by expressions (16) and (15), or more simply by approximation (25). In case of *Poisson input*, $\boldsymbol{U(t,m)}$ is defined by simple expression (17).

4.5 Conclusion

Due to expressions (32) and (37), introducing a limitation to (T_N/T_1) for the impacts of tandem queue load and length, respectively, an important property follows from relation (23), from combining the tandem queue and GI/G/1 queue models:

- "*For a queue length longer and a buffer capacity larger than* (T_N/T_1), *the G/G/1 queue model alone exists.*"

If this property, typical for single server queueing networks *with varying service times* (without breaking up the busy periods), is not perceived, it may be a real danger to use only the GI/G/1 queue model for the design, dimensioning, and management of the network, due to significant buffer overloads not being perceived. Buffer congestion leads to servers' inaccessibility, and the agglutination phenomenon is transmitted downstream (and upstream by reattempts) immediately, generating the blocking of a large area in the network.

The usual concept of local traffic source should be revised in some network queueing theories (e.g., product form theory), due to the existence of the underlying tandem queue effect. In Markovian queueing networks, particularly, the concept of a local transition coefficient is not consistent with the tandem queue model above. And finally, some standards could be very useful for introducing ***a limitation to the ratio*** $\boldsymbol{(T_N/T_1)}$, a typical constraint when service times are varying and not perceived up to now.

5. Case of a Packet Switch Network

5.1 The Packet Traffic Streams

The truncated network of Figure 2 is our reference figure, with traffic streams $\boldsymbol{A_i}$ and $\boldsymbol{B_i}$. Even though traffic streams are not identical, to define the *equivalent tandem queue* (of Subsection 3.1) with the same local queue at the final stage (for an *arbitrary* packet), we may consider identical traffic streams in each incoming path.

The total traffic stream (at final stage) handles a mixture of $\boldsymbol{N}$ packet populations, each labeled $\boldsymbol{j}$ ($j = 1 \ldots N$). Each population corresponds to a partial *Poisson* traffic stream $\boldsymbol{j}$ with constant (i.e., deterministic) packet length T_j ($T_1 < T_2 < \ldots < T_N$), a partial arrival rate $\boldsymbol{\lambda_j}$, and a partial load (*in stationary regime*) $\boldsymbol{\rho_j} = \lambda_j \cdot T_j$. For the total traffic stream, the total arrival rate (for an arbitrary packet) is: $\boldsymbol{\lambda} = \sum_{j=1}^{N} \lambda_j$, and the total load is: $\boldsymbol{\rho} = \sum_{j=1}^{N} \rho_j$. With notation (2), we compute for the transform of the sending (i.e., service) time distribution function of any packet, and for $Re(z) \geq 0$:

$$\varphi_1(z) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} \cdot e^{-zT_j}. \tag{38}$$

5.2 The Distribution of the Local Queue

Relation (23) defines the distribution of the *actual* local sojourn time $\boldsymbol{S}$ with two components:

- the case of the M/G/1 queue (well known);
- the case of the equivalent tandem queue.

The distribution function $\boldsymbol{U(t,m)}$ of the ***virtual*** *local sojourn time* $\boldsymbol{U(m)}$, at the final stage of the equivalent tandem queue, as defined by expressions (15) and (17), has been given in Le Gall [5], formula (36):

for $t < T_1$:

$$\boldsymbol{U_0(t,m)} = 0;$$

for $T_k < t < T_{k+1}$:

$$\boldsymbol{U_k(t,m)} = v_0(t) \cdot \left[v\left(\frac{t}{m}, \lambda\right) \right]^m;$$

with (39)

$$v_0(t) = \frac{\lambda_1 + \ldots + \lambda_k}{\lambda} \cdot \frac{1-\rho}{1-(\rho_1 + \ldots + \rho_k)};$$

$$v(t,\lambda) = \mathrm{Exp}\left\{ -\left(\frac{\lambda_{k+1} + \ldots + \lambda_N}{1-(\rho_1 + \ldots + \rho_k)} \cdot (T_{k+1} - t) + \ldots \right.\right.$$

$$\left.\left. + \frac{\lambda_N}{1-(\rho_1 + \ldots + \rho_{N-1})} \cdot (T_N - T_{N-1}) \right) \right\};$$

for $t > T_N$:

$$\boldsymbol{U_N(t,m)} = 1.$$

In fact, when m increases, a good approximation $U_1(t,m)$ is given by expression (18), depending only on the set (λ_N, T_N).

5.3 The Buffer Load

The mean *actual* local sojourn time [at stage $(m+1)$] of an arbitrary packet (*from any incoming path*), corresponding to its *occupancy in the local buffer*, may be written from definitions (1) and relation (23):

$$\overline{S} = s(m+1) = \frac{1}{n_1} \cdot \left[\overline{U(m)} + \overline{J}\right] + \left[1 - \frac{1}{n_1}\right] \cdot \left[\overline{W_0} + \overline{T}\right]. \tag{40}$$

In expression (33), $\overline{U(m)}$ is given by approximated expression (32), but here $\overline{U(m)}$ has to be deduced from expressions (39).

(*a*) ***Exact expression*** $s(m+1)$**:**

(α) *Calculation of* $\overline{U(m)}$

From expressions (39), we may write:

$$\overline{U(m)} = \sum_{k=0}^{N-1} \int_{T_k}^{T_{K+1}} [1 - U_k(t,m)] \cdot dt. \tag{41}$$

(β) *Calculation of* $\overline{J}$

In Figure 2, we have n incoming paths. Expression (19) gives:

$$\overline{J} = \int_0^{T'' - T} [1 - J(t)] \cdot dt. \tag{42}$$

(γ) *Calculation of* $(\overline{W_0} + \overline{T})$

From expression (38), we deduce that for the service time distribution of an arbitrary packet:

$$\overline{T} = E(T) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} \cdot T_j, \quad m_2 = E(T^2) = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda} \cdot (T_j)^2. \tag{43}$$

The classical Pollaczek formula gives:

$$\overline{W_0} = \frac{1}{2} \cdot \frac{\rho}{1-\rho} \cdot \frac{m_2}{\overline{T}}. \tag{44}$$

Depending on the actual incoming paths, expression (21) defines n_1 and, finally, we may evaluate the buffer occupancy $s(m+1)$, at stage $(m+1)$, as defined by expression (40).

(*b*) ***Approximated expression*** $s_1(m+1)$**:** A general approximated expression $s_1(m+1)$ of the buffer occupancy (for an arbitrary packet) is given by (33), using (32) for $u = \lambda$ in case of Poisson arrivals, with $Q_1^* = (1 - (\rho - \rho_N)$:

$$\overline{U(m)} = D_m(\lambda) = T_N \cdot \left\{ 1 - \left(1 - \frac{\lambda_N}{\lambda}\right) \cdot \frac{1-\rho}{1-(\rho-\rho_N)} \cdot \frac{\mathrm{Exp}[-(m-1)A] - \mathrm{Exp}[-mA]}{A} \right\},$$

with

$$A = \frac{\rho_N}{1-(\rho-\rho_N)}. \tag{45}$$

Note: It follows that the *virtual* sojourn time distribution is not practically changed when we *segment* packets of medium lengths.

A Numerical Example: It may be interesting to consider the case of Figure 2, with a load ρ in each incoming path i ($i = 1...n$), and in the considered terminal link j. But we introduce some *dissymmetry* in the distribution of this load by introducing the following distribution $\rho_i(j)$, related to the load of the partial traffic stream $A_i(j)$ with j being given:

$$\rho_j(j) = \rho_0, \qquad \rho_i(j) = \frac{\rho-\rho_0}{n-1} \quad (i \neq j). \tag{46}$$

We let:

$$\rho_0 = h \cdot \frac{\rho}{n} \quad (h = 1...n). \tag{47}$$

In relations (33) and (40), n_1 is defined by expression (21), which gives:

$$\frac{1}{n_1} = \frac{\rho_0}{\rho}\cdot\left(\frac{\rho_0}{\rho}\right) + \frac{\rho-\rho_0}{\rho}\cdot\left(\frac{\rho-\rho_0}{n-1}\cdot\frac{1}{\rho}\right) = \left(\frac{\rho_0}{\rho}\right)^2 + \frac{1}{n-1}\cdot\left(\frac{\rho-\rho_0}{\rho}\right)^2. \tag{48}$$

In fact, the buffer load increases with ρ_0, the load of the partial traffic stream $A_j(j)$, which is its contribution to the total load ρ of the terminal link j [at stage $(m+1)$]. This contribution may be much higher than the other contributions $\rho_i(j)$. The case $h = 1$ is the pure symmetrical case: $\rho_i(j) = \rho_j(j) = \frac{\rho}{n}$. On the contrary, the case $h = 10$ $(= n)$ corresponds to $A_j(j)$ becoming a pure tandem queue. For the intermediate case $h = 5$, $\rho_0 = \frac{\rho}{2}$, half of the total load (at the final stage) comes from a single incoming path ($N^0 j$); the tandem queue effect begins to appear.

In Table 1, we consider the case of an **"IP" (Internet Protocol) traffic**: $n = 10$, $m = 6$, $\rho = 0.6$, and a total packet traffic stream with three partial traffic streams:

$$T_1 = 1, \quad T_2 = 5, \quad T_3 = 30; \text{ and } \rho_1 = \rho_2 = \rho_3 = 0.2.$$

			exact		approximated	
h	n_1	ρ_0	$s(m+1)$	C	$s_1(m+1)$	C_1
1	10	0.06	13.08	1.14	13.04	1.14
5	3.6	0.30	16.01	1.40	15.90	1.39
10	1	0.60	27.91	2.44	27.52	2.41
M/G/1 queue: $\overline{W_0}+\overline{T}$			11.43			

Table 1: Mean Actual Local Sojourn Time (= Occupancy in the Local Buffer) $s(m+1)$, Formula (40), and Overload Coefficient $C = \frac{s(m+1)}{\overline{W_0}+\overline{T}}$, in Figure 2 as a Function of the Contribution of $A_j(j)$, load ρ_0, to the Total Terminal Link's Load ρ (N^0 j).

Approximated Values: $s_1(m+1)$, formulae (33) and (45), and $C_1 = \frac{s_1(m+1)}{\overline{W_0}+\overline{T}}$.

Data: $n = 10,\ m = 6,\ \rho = 0.6,\ \rho_j(j) = \rho_0 = h \cdot \frac{\rho}{n}$.

Traffic stream (3 components): $T_1 = 1, T_2 = 5, T_3 = 30;\ \rho_1 = \rho_2 = \rho_3 = 0.2$.

We give the *overload coefficients*:

$$C = \frac{s(m+1)}{\overline{W_0}+\overline{T}};\quad C_1 = \frac{s_1(m+1)}{\overline{W_0}+\overline{T}}. \tag{49}$$

For the usual case ($h = 1$), the buffer load is just 14% higher than the buffer load given by the M/G/1 queue model. For $h = 5$, the buffer load is already 40% more higher. Moreover, following expression (22), the buffer occupancy related to the traffic stream $A_j(j)$ is even 80% higher than the buffer load given by the M/G/1 queue model. For $h = 10$, the buffer load becomes 144% higher (i.e., case of the tandem queue model). Finally, *the tandem queue effect becomes significant when* ***half*** *of the local load comes from a single incoming path only* (case $h = 5$). An example of this is the case of a ***virtual circuit***, or of a ***traffic stream concentration*** *through a supplementary upstream node.*

The approximation $s_1(m+1)$ is a good approximation, proving that the set (λ_N, T_N) of *the longest service times* is *sufficient to generate the agglutination phenomenon.*

5.4 The Buffer Rejection Rate

To evaluate the *buffer rejection rate* $R(K)$ in a stationary regime, we use expression (34), which causes us to evaluate the rejection rate $R_0(K)$ due to the M/G/1 queue model, as well as the rejection rate $R_1(K)$ due to the tandem queue model, where K is the *buffer capacity.*

(*a*) ***Expression of*** $R_0(K)$: $R_0(K)$ is given by expression (35) for the GI/G/1 queue model. In Le Gall [4], formula (28), the expression has been given for the M/G/1 queue model:

$$R_0(K) \cong \int_{T_N}^{\infty} [1 - W_1(t)] \cdot H(t, K-2) \cdot \lambda dt, \tag{50}$$

with

$$H(t,j) = e^{-\lambda t} \cdot \frac{(\lambda t)^j}{j!}. \tag{51}$$

$W_1(t)$ corresponds to the asymptotic expression of the queueing delay distribution $W_0(t)$ for the M/G/1 queue model. In Le Gall [4], expression (32) gives:

$$W_1(t) \cong 1 - \frac{1-\rho}{\rho_1 \cdot e^{\beta_0 T_1} + \ldots + \rho_N \cdot e^{\beta_0 T_N} - 1} \cdot e^{-\beta_0 t}, \tag{52}$$

where $q = \beta_0$ (> 0) is the real (positive) root, closest to the origin, of the equation:

$$q + \lambda - \lambda_1 \cdot e^{qT_1} - \ldots - \lambda_N \cdot e^{qT_N} = 0. \tag{53}$$

Expressions (50) and (52) may be used, as a useful approximation, for $\rho < 0.8$ and $R_0(K) \leq 10^{-2}$.

(*b*) ***Expression of $R_1(K)$:*** $\boldsymbol{R_1(K)}$ is given by expression (37), where $\boldsymbol{U(t,m)}$ is given by the approximated expression (18). To evaluate $[F_1^*(t)]^K$, we consider the Laplace transform, excluding the set (λ_N, T_N):

$$[\varphi_1^*(z)]^K = \sum_{j=0}^{K} K! \frac{\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^j}{j!} \cdot e^{-jzT_1} \cdot \frac{\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{K-j}}{(K-j)!} \cdot e^{-(K-j)zT_2},$$

with the condition (for a single busy period): $jT_1 + (K-j)T_2 < T_3$. Expression (37) gives:

$$\boldsymbol{R_1(K)} = \sum_{j=0}^{K} K! \frac{\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^j}{j!} \cdot \frac{\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{K-j}}{(K-j)!} \cdot [1 - U(jT_1 + (K-j)T_2, m)], \tag{54}$$

with $\boldsymbol{j}$ such as: $jT_1 + (K-j)T_2 < T_3$.

Practically, this last condition leads only to the case $\boldsymbol{j = K}$; ***the rejections are generated by the shortest packets.***

(*c*) ***A Numerical Example:*** We consider the same example as in Subsection 5.3.*c* and in Table 1. The expression of $\boldsymbol{n_1}$ is defined by expression (48). Table 2 gives the numerical results for $\boldsymbol{R_0(K)} = 10^{-2}$, 10^{-3} and 10^{-4}, corresponding to the buffer capacity $K = 19$, 28 and 40, respectively.

			K	19	28	40
			$R_0(K)$	10^{-2}	10^{-3}	10^{-4}
h	n_1	ρ_0	$C(K)$			
1	10	0.06		3.9	26	0.90
5	3.6	0.30		8.9	70	0.72
10	1	0.60		30	250	0

Table 2: Local Rejection Rate $\boldsymbol{R(K)} = C(K) \cdot R_0(K)$, for a Buffer Capacity $\boldsymbol{K}$, Formulae (34) and (54). (Examples of Table 1)
[Rejection rate in the M/G/1 queue model: $\boldsymbol{R_0(K)}$, formula (50)]

For $K = 19$ and 28, a strong influence of the agglutination phenomenon appears: $\boldsymbol{R(K)} = C(K) \cdot R_0(K)$, with a high value for $C(K)$. For $K > 30$, this phenomenon has disappeared: $C(K) < 1$.

Due to the impact of the shortest packets, *we may avoid the* ***rejections due to the agglutination phenomenon*** *if the buffer capacity* $\boldsymbol{K}$ *is such as:*

$$K > \frac{T_N}{T_1}, \tag{55}$$

where T_N and T_1 are the longest and shortest packet lengths, respectively. And this conclusion to dimension the buffer is the same as for the M/G/1 queue alone: See Le Gall [4], Subsection 4.3. Moreover, we may note that this condition (55) does not change if we segment the medium packet lengths into the shortest packet lengths. In that case, we can say that condition (55) states that *the buffer capacity should be higher than the agglutination size* (after segmentation).

6. Conclusion

(*a*) We have studied the single server queueing networks (excluding ring networks), *when successive service times are different* (or not) ***without breaking up the busy periods*** from stage-to-stage, leading to a great approximate simplification: the existence of an *equivalent tandem queue* effect. At the end of Subsection 3.1, and in case of infinite support for the service time distribution function $F_1(t)$, we stressed the need that $[1 - F_1(t)]$ decreases exponentially at infinity. In particular, in case of a "**Pareto**" distribution (for "**Fractal**" processes) decreasing as $(at)^{-\alpha}$, $(\alpha > 2)$, this kind of traffic cannot be carried in queueing networks (on the contrary of a single GI/G/1 server) because of the tandem queue effect: in this case link overload generates more successive local arrivals from the same incoming path. Practically, we restricted this paper to the realistic case of finite support.

(*b*) Due to the possible correlation between two successive local arrivals from the same incoming path, a curious "*double faced*" traffic model appears, as for the Janus divinity: An agglutination phenomenon (of short service times T_1 behind a long service time T_N) results for queue lengths and buffer capacities lower than (T_N/T_1), and the classical GI/G/1 queue for queue lengths and buffer capacities higher than (T_N/T_1). This property (amplifying from stage-to-stage) has not yet been detected with classical network queueing theories, (e.g., Markovian queues, product form theory), due to an incorrect concept of local traffic source (i.e., eliminating the occurrence of some agglutination and the concept of *upstream apparent delay*). These theories assumed the local combination of distinguishable customers (with distinguishable queueing delay), instead of parts of upstream busy periods with *indistinguishable* customers (and indistinguishable queueing delay depending on the maximum sojourn time initiating the new downstream busy period). These theories cannot be consistent with the concept of the *equivalent tandem queue.*

(*c*) Some significant consequences appear, since any link overload comes from a given incoming path which generates the tandem queue effect (i.e., correlation between local interarrival time and upstream service time). The usual queueing standards (related to long queues) cannot protect against subsequent, significant overloads in the buffers due to some possible "*agglutination phenomenon*" (related to short queues). Usual network management methods should be revised, and should monitor the partial traffic streams loads (and not only the server load). Moreover, when we connect two networks (or a network and a user), as in circuit switching (with buffer associated), it follows that $n_1 = 1$ in expression (23), and the inlet (downstream) works with a sojourn time (= local transfer delay) practically equal to T_N instead of $(W_0 + T)$, leading to a *supplementary mean transfer delay* $[T_N - (W_0 + T)]$. This is the same in the case of a *traffic stream concentration* through a supplementary upstream node. In that case, leading to some packet re-

jections in connectionless IP-based telecommunication networks, some moving pictures on the screen may be stopped a long time before continuing.

(*d*) This longest packet length, T_N, may be bounded in a "virtual circuit mode" (depending on the considered link). On the contrary, in a "*connectionless mode*", the usual routing (or flow control) methods cannot be efficient to bound T_N for a given link with its buffer. Finally, buffer dimensioning [see condition (55)] depends on the limitation needed by the network itself. Any change on this general limitation may disturb all the buffers of the network.

This disturbance may become considerable when the network tries to handle bursty traffics with *non-transparent bursts* (i.e., with no possibility that other packets cross the burst inside the intervals between packets of the considered burst). If $\boldsymbol{L}$ is the *maximum burst duration*, $\boldsymbol{T_N}$ now has to be replaced by $\boldsymbol{L}$ in condition (55), leading to a ratio (L/T_1) much greater than (T_N/T_1) and, consequently, to an agglutination phenomenon much higher in the buffers (and not detected by flow control). Due to congestion in buffers, a large area of the network may rapidly become inaccessible. The classical traffic modeling, with local Poisson arrivals and no correlation with upstream service times, could not lead to this important consequence (even when successive service times are mutually independent, due to Hypothesis 2).

(*e*) The traffic modeling for this case of correlated, and *different successive* service times is more general than the case of *constant* service times. The occurrence of agglutinations generates the concept of *indistinguishability* as in nuclear physics, which may affect the concept of a local transition coefficient in Markovian queues, and the concept of product form. Moreover, this concept of upstream *indistinguishability* introduces some strong interferences with "premature departures", leading to the new concepts of *virtual* and *actual* local queueing delays, respectively.

(*f*) Finally, we stress our conclusion (*e*) in Subsection 3.3. To be sure that the tandem queue effect may be detected by traffic simulation methods, it is necessary to directly observe the two successive (upstream and local) overall sojourn times for a local arrival in a ***given*** *traffic stream*, instead of directly observing the local queueing delay for an arbitrary arrival. In this latter case, a simple, local Poisson arrival process without any correlation with the upstream stage could appear, leading to the evaluation of a simple, apparent M/G/1 local server. The whole observation of the network could be altered, and in case of rejected packets repeated, it could be impossible to understand some high rejection rates observed, even in the case of local and total load at a medium level (e.g., 0.5 erlang), when condition (55) is not satisfied.

References

[1] Le Gall, P., Bursty traffic in packet switched networks, *Proc. ITC-14* (Antibes, France, June 1994), The Fund. Role of Teletraffic in the Evolution of Telecom. Networks, Elsevier Science B.V., **1a** (1994), 535-549.

[2] Le Gall, P., The theory of networks of single server queues and the tandem queue model, *J. of Appl. Math and Stoch. Anal.* **10**:4 (1997), 363-381.

[3] Le Gall, P., Multiserver queueing networks and the tandem queue model, *J. of Appl. Math. and Stoch. Anal.* **11**:3 (1998), 377-390.

[4] Le Gall, P., The queue-length in GI/G/s queues, *Math. Prob. in Eng.* **6**:1 (2000), 1-11.

[5] Le Gall, P., The stationary local sojourn time in single server tandem queues

with renewal input, *J. of Appl. Math. and Stoch. Anal.* **12**:4 (1999), 417-428.

[6] Pollaczek, F., Problèmes stochastiques posés par le phénomène de formation d'une queue d'attente à un guichet et par des phénomènes apparentés, *Mémorial des Sciences Mathématiques*, Gauthier-Villars, Paris **CXXXVI** (1957). (= **GI/G/1** queue; in French).