Journal of Applied Mathematics and Stochastic Analysis, **14**:4 (2001), 381-398.

SINGLE SERVER TANDEM QUEUES AND QUEUEING NETWORKS WITH NON-CORRELATED SUCCESSIVE SERVICE TIMES

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

(Received January, 2000; Revised July, 2001)

To evaluate the local **actual** queueing delay in general single server queueing networks with non-correlated successive service times for the same customer, we start from a recent work using the *tandem queue effect*, when two successive local arrivals are not separated by "*premature departures*". In that case, two assumptions were made: busy periods not broken up, and there are limited variations for successive service times. These assumptions are given up after having crossed two stages. The local arrivals become *indistinguishable* for the sojourn time inside a given busy period. It is then proved that the local sojourn time of this tandem queue effect may be considered as the sum of two components: the first (independent of the local interarrival time) corresponding to the case where upstream, successive service times are supposed to be identical to the local service time, and the second (negligible after having crossed 2 or 3 stages) depending on local interarrival times increasing because of broken up busy periods. The consequence is the possible occurrence of the *agglutination phenomenon* of indistinguishable customers *in the buffers* (when there are limited "premature departures"), due to a stronger impact of long service times upon the local actual queueing delay, which is not consistent with the traditional concept of local traffic source only generating distinguishable customers.

Key words: Queueing Networks, Tandem Queues, Tandem Queue Effect, Non-Correlated Successive Service Times, Local Queueing Delay, Agglutination Phenomenon, Buffer Overload, Local Traffic Source Concept, Indistinguishability.

AMS subject classifications: 60K25, 90B22.

1. Introduction

In a recent work (see Le Gall [6]) utilizing renewal input, intermediate queues and local "first come-first served" discipline, we evaluated the local queue in single server

queueing networks using the *tandem queue effect* as presented in Le Gall [4]. When two successive local arrivals come from the same upstream traffic stream, the local queueing delay is strongly dependent on the upstream service time, since the local interarrival time is equal to the upstream service time in case of congestion. In that case, the sum of the upstream service time and of the local queueing delay is equal to the local sojourn time of the preceding customer (for successive service times correlated or not for the same customer). But we used two restrictive assumptions: busy periods not broken up and there are limited variations for successive service times. In that case, interferences with other traffic streams crossing upstream could be neglected. Now, when successive service times of the same customer are not correlated, we intend to give up these assumptions when customers have already crossed two stages. It will be proved (for this tandem queue effect) that, when added to the service time of the customer initiating the busy period, the local sojourn time may be considered as the sum of two supplementary components:

The first component corresponds to the case of equivalent upstream service times being identical to the local service time, which leads to our earlier results (with the agglutination phenomenon of *indistinguishable* customers in the buffers), but with a lower number of equivalent upstream stages.

The second component (which is specific for a given customer and may be neglect after having crossed 2 or 3 stages) corresponds to an actual queueing delay generated by a GI/G/1 server, where interarrival times are increasing (from stage-to-stage) due to broken up busy periods, and where service times are rapidly decreasing, corresponding to the supplementary part compared with these new interarrival times.

Network behavior will appear similar to our earlier work, with a stronger impact of long service times, but with a great difference: we now suppose that *successive service times* (for the same customer) *are not correlated.* Consequently, from stage-to-stage, busy periods will amalgamate more slowly. For example, in the M/M/1 case and compared with Jackson's queueing network theory, the increase in mean local queueing delay may be detected after having crossed approximately 20 stages. But in the case of different populations of packet traffics (with highly varying packet lengths), this increase may be faster, still leading (in case of a predominant traffic stream) to a strong *agglutination phenomenon of indistinguishable* customers *in the buffers* (due to the identical sojourn times inside a given busy period), which may become congested even for a slight load (when half of this load comes from the same upstream traffic stream). Markovian theories are not appropriate to evaluate congestion in buffers.

In this paper, we assume that customers only gain access to a downstream queue after completion of upstream service. We will begin evaluating the tandem queue effect. In Section 2, we define our notation and assumptions. In Section 3, we outline our earlier studies in single server tandem queues. In Section 4, we consider tandem queues with non-correlated successive service times. In Section 5, we consider single server queueing networks and buffers in the case of non-correlated successive service times.

2. Notations and Assumptions for Single Server Tandem Queues

2.1 The Tandem Queue

The tandem queue is made of $(m+1)$ successive stages of single servers, with the following notations for the h^{th} customer at stage k $(=1,\ldots,m+1)$:

- 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]: Y_{h-1}^k;
- occasional idle period [during Y_{h-1}^k]: e_h^{k-1}.

In other words, we may write

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

Moreover, we let for $k = 2,\ldots,m+1$:

$$\begin{cases} T_h^1 + \ldots + T_h^k = T_h'(k), \\ s_h^i + \ldots + s_h^k = S_h(i,k). \end{cases} \tag{2}$$

For the h^{th} customer, $T_h'(k)$ is the *overall service time* from stage 1 to stage k, and $S_h(i,k)$ is the overall sojourn time from stage i to stage k.

2.2 Assumptions

We assume the system is in the *stationary regime.* The arrival process (at the entry) is *renewal.* At each stage, there is an *intermediate queue* using a "*first come-first served*" discipline. There are *no intermediate arrivals.*

The *arrival rate* is:

$$\lambda = (1/EY_{h-1}^k), \quad (k = 1,\ldots,m+1). \tag{3}$$

$F_0(t)$ is the distribution function of the interarrival intervals. $F_k(t)$, $(k = 1,\ldots, m+1)$ is the distribution function of the service time at stage k, independent of the considered customer. All successive service times (for the same customer) are mutually independent. They are also independent of the arrival process and of the service times related to other customers. At stage k, the load (= traffic intensity) is:

$$\rho_k = \lambda E(T_h^k) < 1. \tag{4}$$

3. Preliminary Results in Single Server Tandem Queues

3.1 Case of the General Distribution for Service Times

In Le Gall [1], we gave the fundamental stochastic recurrence relation for the sojourn time in any single server tandem queue (Formula 3.5):

$$\left\{\begin{array}{l} Exp(-\sum_{i=1}^{m+1} z_i s_n^i) = \prod_{k=1}^{m+1} \frac{1}{2\pi i} \int_{+0} (\frac{1}{z_k - u_k} + \frac{1}{u_k - u_{k+1}}) du_k \\ \times Exp(u_1 Y_{n-1}^1) \cdot \left[Exp(-z_1 T_n^1 - \sum_{k=2}^{m+1} (z_k T_n^k - u_k T_n^{k-1}) \right] \\ \times \left[Exp(-\sum_{k=1}^{m+1} u_k s_{n-1}^k) \right], \\ \text{with: } 0 < R(u_{k+1}) < R(u_k) < R(z_k),\ k = 1,\ldots,m+1;\ u_{m+2} = 0. \end{array}\right. \tag{5}$$

The symbol $\prod$ denotes a repeated integral in the $(u_1,\ldots,u_{m+1})$ successive planes. We use (Cauchy) contour integrals along the imaginary axis in the complex planes $\boldsymbol{u_k}$. If the contour (followed from the bottom to the top) is to the 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 $\int_{-0}$. If it is not necessary to define the side, we may simply write $\int_0$.

If, in the successive planes $\boldsymbol{u_k}$ $(k > 1)$, the following condition

$$s_{n-1}^k \geq T_n^{k-1},\quad k = 2,\ldots,m+1, \tag{6}$$

is satisfied, the kernel in (5) is holomorphic for $R(u_k) > 0$ $(k > 1)$. Now set $z_k = z$ $(k = 1,\ldots,m+1)$. The poles $u_1 = z_1$ and $u_{k+1} = u_k$ $(k = 1,\ldots,m)$ remain. We can thus apply the residue theorem at the preceding poles u_k $(k = 2,\ldots,m+1)$, and we deduce the following stochastic relation, with notation (2):

$$S_n(1,m+1) = Max[T'_n(m+1), S_{n-1}(1,m+1) + T_n^{m+1} - Y_{n-1}^1]. \tag{7}$$

At stage $\boldsymbol{i}$ $(i = 1,\ldots,m)$, we may write in the same way, with notation (1):

$$S_n(i,m+1) = Max[T_n^i + \ldots + T_n^{m+1}, S_{n-1}(i,m+1) + T_n^{m+1} - Y_{n-1}^i], \tag{8}$$

or using expression (1):

$$S_n(i,m+1) = Max[T_n^i + \ldots + T_n^{m+1}, S_{n-1}(i,m+1) + (T_n^{m+1} - T_n^{i-1}) - e_n^{i-1}]. \tag{9}$$

In Le Gall [6], we used an assumption to simplify (hypothesis [2]) to delete the term $(T_n^{m+1} - T_n^{i-1})$ in the right-hand side, corresponding to a limitation on successive service time variations. In the present paper, we intend to give up this simplification and condition (6) after having crossed several stages.

The key point in stochastic relations (7) and (9) is to note that a customer initiating a busy period at stage $(m+1)$, and corresponding to the first member in brackets, does not wait upstream: he also initiates the upstream busy periods. This point has not yet been detected in classical theories, in particular, for the tandem queue M/M/1 $\rightarrow$M/1. Note, on the contrary, that a customer initiating a busy

period at stage i may experience some queues downstream.

3.2 Case of Identical Successive Service Times

3.2.1 Equation and equivalence

In Le Gall [1], we solved the case of identical successive service times for the same customer. Stochastic relation (9), with notation (2), becomes (for $i = 2$):

$$\begin{cases} S_n(2, m+1) = Max[T'_n(m), S_{n-1}(2, m+1) - e_n^1], \\ \quad \text{with: } T'_n(m) = T_n^1 + \ldots + T_n^m = m \cdot T_n^{m+1}. \end{cases} \tag{10}$$

When we used this relation (10), even in the case of non-identical successive service times, as an approximation of relation (9) for $i = 2$, we stressed in Le Gall [2] that the local sojourn time distribution is practically defined by the two first moments, and finally by $VarT'_n(m)$. Consequently, in Le Gall [3], we introduced the parameter $\boldsymbol{m_0}$ such as:

$$\boldsymbol{m_0^2} \cdot Var(T_n^{m+1}) = Var(m_0 \cdot T_n^{m+1}) = VarT'_n(m), \tag{11}$$

when excluding the case of T_n^{m+1} and $T'_n(m)$ constant. When $\boldsymbol{m_0}$ is between two successive integers, we will use an interpolation between the delays related to these integers, or directly the possible fractional $\boldsymbol{m_0}$ in formulae. Thus, the local sojourn time is practically the same as for a *single server packet tandem queue of* $(m_0 + 1)$ *stages*, and corresponding to identical successive service times: $T_n^1 = \ldots = T_n^{m_0+1} = T_n^{m+1}$, when $\boldsymbol{m_0}$ is an integer. When $\boldsymbol{m_0}$ is not an integer, the distribution function may be used with this new value $\boldsymbol{m_0}$.

3.2.2 Local sojourn time distribution

In Le Gall [5], w evaluated the local sojourn time distribution *at stage* $(m+1)$, in the case of a *stationary regime* and successive service times *identical* to the local service time T_n^{m+1}, with $F_{m+1}(t)$ being the distribution function. For $R(z) \geq 0$, we set:

$$\begin{cases} \varphi_0(z) = \int_0^\infty e^{-zt} dF_0(t), \\ \varphi_{m+1}(z,t) = \int_0^t e^{-z\alpha} dF_{m+1}(\alpha), \\ \Phi_{m+1}(z) = Exp\left(-\frac{1}{2\pi i}\int_{-0} [\frac{1}{z-u} + \frac{1}{u}] log[1 - \varphi_0(-u)\varphi_{m+1}(u,\infty)]du\right)\varphi_{m+1}(z,\infty), \\ Q_{m+1}(t) = Exp\left(-\frac{1}{2\pi i}\int_{-0} log[1 - \varphi_0(-u)\varphi_{m+1}(u,t)]\frac{du}{u}\right), \text{ and} \\ Q_{m+1} = Q_{m+1}(\infty). \end{cases} \tag{12}$$

And we introduce the following expressions, using (12):

$$\begin{cases} v_0(t) = \dfrac{Q_{m+1}}{Q_{m+1}(t)} F_{m+1}(t), \\ v(t,u) = Exp(-u\int_t^{\infty} \dfrac{1 - F_{m+1}(v)}{Q_{m+1}(v)} dv), \\ u_m(t,u) = v_0(t)[v(t,u)]^m, \text{ and} \\ d_m(t,u) = \int_0^t v(\frac{t+w}{m}, u) d_w u_{m-1}(\frac{w}{m-1}, u). \end{cases} \tag{13}$$

Finally, the distribution function of the *local sojourn time* $\boldsymbol{U(m+1)}$, *at stage* $(m+1)$ and for an *arbitrary* customer is, with expressions (12) and (13):

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

$$U(t, m+1) = \frac{1}{2\pi i} \int_{+0} \frac{\varphi_0(-u)\Phi_{m+1}(u)}{Q_{m+1}} d_m(t,u) \frac{du}{u}. \tag{14}$$

b) **Case of Poisson input:**

$$U(t, m+1) = d_m(t, \lambda) \cong v_0(t)[v(\tfrac{t}{m}, \lambda)]^m. \tag{15}$$

In the case of finite support for $F_{m+1}(t)$, we stressed the influence of the longest service time (i.e., length $= \boldsymbol{T_N}$), corresponding to a load (= traffic intensity):

$$\rho_N = \lambda_N T_N < 1. \tag{16}$$

This influence is due to the "*agglutination phenomenon*" in the *buffers*. Due to this long service time upstream, the queue disappears downstream and customer $\boldsymbol{n_0}$ does not wait downstream; rather, he initiates a busy period. In case of congestion upstream, we may write $\boldsymbol{e_n^1} = 0$ in relation (10), and we deduce during this busy period:

$$S_n(2, m+1) = S_{n-1}(2, m+1) = \ldots = S_{n_0}(2, m+1) \cong \boldsymbol{T_N}.$$

Finally, *all successive local sojourn times* of the local busy period *are equal to the long service time* $\boldsymbol{T_N}$, with the customers becoming *indistinguishable* in the buffer. Based on the increase of the busy period duration (from stage-to-stage), it follows that *the busy periods tend to* ***amalgamate*** *with subsequent busy periods.* The phenomenon is amplified when $\boldsymbol{m}$ increases, leading to a strong impact of the longest service times. We deduced the following practical approximation for expression (15), i.e., for the ***local sojourn time distribution*** of an ***arbitrary*** customer, at ***stage*** $\boldsymbol{(m+1)}$, and in case of *Poisson* input:

for $0 < t \leq T_N$:

$$U_1(t, m+1) = (1 - \frac{\lambda_N}{\lambda}) \frac{1 - \rho_{m+1}}{1 - (\rho_{m+1} - \rho_N)} Exp(\frac{-m\rho_N}{1 - (\rho_{m+1} - \rho_N)}[1 - \frac{t}{mT_N}]);$$

for $t > T_N$:

$$U_1(t, m+1) = 1, \tag{17}$$

where ρ_{m+1} is the total load (= traffic intensity) at stage $(m+1)$.

4. Single Server Tandem Queues with Non-Correlated Successive Service Times

4.1 Equations and Results

When successive service times (for the same customer) are non-correlated, it will appear as a supplementary local queueing delay. In that case, it is useful to introduce a relation due to Pollaczek [7], Formula (15) for $\nu = 1, \ldots, n$ and $R(z) \geq 0$:

$$\left\{ \begin{array}{c} Exp(-zMax^+ x_\nu) = (\prod_{\nu=1}^{n} \frac{1}{2\pi i} \int_{+0} \frac{du_\nu}{u_\nu}) Exp(-\sum_{\nu=1}^{n} x_\nu u_\nu) \frac{z}{z - \sum_{\nu=1}^{n} u_\nu}, \\ \text{with: } Max^+(x_\nu) = Max(0, x_1, \ldots, x_n), \nu = 1, \ldots, n, \\ 0 < R(u_\nu), R(\sum_{\nu=1}^{n} u_\nu) < R(z). \end{array} \right. \tag{18}$$

In our case, we have $R(x_1) > 0$. We may apply the residue theorem at pole: $u_1 = z - \sum_{\nu=2}^{n} u_\nu$. For $z = u$ and $\nu = 1, \ldots, n$ we deduce:

$$\left\{ \begin{array}{c} Exp(-uMax(x_1, \ldots, x_n)) = (\prod_{\nu=1}^{n} \frac{1}{2\pi i} \int_{+0} \frac{du_\nu}{u_\nu}) u \; Exp(-\sum_{\nu=1}^{n} x_\nu u_\nu) \\ \text{with: } \sum_{\nu=1}^{n} u_\nu = u. \end{array} \right. \tag{19}$$

We let:

$$\theta_n^{m+1} = Max(T_n^1, \ldots, T_n^{m+1}). \tag{20}$$

We deduce from expression (19):

$$\left\{ \begin{array}{c} Exp(-u_1 \theta_n^{m+1}) = (\prod_{k=2}^{m+1} \frac{1}{2\pi i} \int_{+0} Exp[-(u_{k-1} - u_k) T_n^{k-1}] \cdot \frac{du_k}{u_{k-1} - u_k}) \\ \cdot (\frac{u_1}{u_{m+1}}) Exp[-(u_{m+1} \cdot T_n^{m+1})], \\ \text{with: } 0 < R(u_{m+1}) < \ldots < R(u_1). \end{array} \right. \tag{21}$$

Let us consider the basic relation (5), on writing:

$$\left\{\begin{array}{c} Exp[-z_1 T_n^1 - \sum_{k=2}^{m+1}(z_k T_n^k - u_k T_n^{k-1})] \rightarrow \\ Exp[-\sum_{k=1}^{m+1}(z_k - u_k)T_n^k] \cdot Exp[-\sum_{k=1}^{m+1}(u_k - u_{k+1})T_n^k], \\ \text{with: } u_{m+2} = 0. \end{array}\right. \tag{22}$$

In relation (5), the kernel becomes the product of two quantities:

$$\left\{\begin{array}{c} A = Exp(u_1 Y_{n-1}^1) \cdot Exp[-\sum_{k=1}^{m+1}(z_k - u_k)T_n^k] \cdot Exp[-\sum_{k=1}^{m+1} u_k s_{n-1}^k], \\ B = u_1 Exp[-\sum_{k=1}^{m+1}(u_k - u_{k+1})T_n^k]. \end{array}\right. \tag{23}$$

As in expression (5), we set: $z_k = z(k = 1,\ldots,m+1)$. For $0 < R(u_1) < \delta$, where δ is a positive number sufficiently low, the kernel in (5) is holomorphic. Consequently, in the planes $\boldsymbol{u_k}$ $(k = 2,\ldots,m+1)$, we only find the poles $u_{k+1} = u_k$. We may apply the residue theorem and put $u_k = u_1$ $(k = 2,\ldots,m+1)$ in expression A above. Expression B is, in fact, the kernel in expression (21). Finally, the basic relation (5) becomes, on using (20):

$$\left\{\begin{array}{c} Exp[-zS_n(1,m+1)] = \frac{1}{2\pi i}\int_{+0} Exp[-(z-u_1)T'_n(m+1)] \\ \cdot Exp[-u_1(\theta_n^{m+1} - Y_{n-1}^1)] \\ \times Exp[=u_1 S_{n-1}(1,m+1)] \cdot \frac{z du_1}{(z-u_1)u_1}, \\ \text{with: } 0 < R(u_1) < R(z). \end{array}\right. \tag{24}$$

This expression corresponds to the following stochastic relationship, and is *valid even when the successive service times* (of the same customer) *are correlated*:

$$S_n(1,m+1) = Max[T'_n(m+1), S_{n-1}(1,m+1) + \theta_n^{m+1} - Y_{n-1}^1]. \tag{25}$$

Note that variables $T'_n(m+1)$ and θ_n^{m+1} are correlated. Compared with relation (7), the first member between brackets has not changed. It is the same for customer $\boldsymbol{n_0}$ initiating a busy period at stage $(m+1)$. On the contrary, during a busy period (see the second member between brackets), the term T_n^{m+1} is replaced by θ_n^{m+1}, which is not correlated with $T'_{n_0}(m+1)$. Let us use the relations:

$$\left\{\begin{array}{c} S_n(1,m+1) = w_n^1 + T_n^1 + S_n(2,m+1), \\ S_{n-1}(1,m+1) - Y_{n-1}^1 = (w_n^1 - e_n^1) + S_{n-1}(2,m+1). \end{array}\right. \tag{26}$$

During a busy period, we may write from the second term in expression (25):

$$S_n(2,m) = S_{n-1}(2,m) + [\theta_n^m - T_n^1] - e_n^1,$$

or:

$$Y_{n-1}^{m+1} = \theta_n^m = Y_{n-1}^2 + [S_n(2m) - S_{n-1}(2m)]. \tag{27}$$

When we consider the case of upstream service times identical to T_n^{m+1}, the term $[\theta_n^m - T_n^1]$ does not exist, leading to a first component for the sojourn time at stage $(m+1)$. Moreover, inside the busy periods, the non-correlation between successive upstream service times leads to a local interarrival time θ_n^m *at stage* $(m+1)$, as given by expression (27). Besides, the term above (between brackets) is now existing and leads to a local service time $\tau_n^{m+1} = \theta_n^{m+1} - \theta_n^m$. At this stage, from expression (25), this quantity generates a supplementary local queueing delay corresponding to a GI/G/1 server defined by the set $(\theta_n^m, \tau_n^{m+1})$ during busy periods. To evaluate the local *actual* queueing delay we may summarize for a *single server tandem queue*:

Proposition 1: (The two components of the local sojourn time) *At stage* $(m+1)$ *in stationary regime* $(m>1)$, *the* local sojourn time s_n^{m+1} *is the sum of two components:*

First component $\boldsymbol{U_n^{m+1}}$ *corresponds to the case of successive upstream service times supposed to be identical to the local service time* T_n^{m+1}, *with the number of stages* $(\boldsymbol{m_0+1})$ *being defined by expression* (11), *and the local sojourn time distribution being given by expressions* (14), (15) *or* (17);

Second component $\boldsymbol{V_n^{m+1}}$ *corresponds to the queueing delay generated by a GI/G/1 server, defined by service time:*

$$\tau_n^{m+1} = \theta_n^{m+1} - \theta_n^m = [T_n^{m+1} - \theta_n^m]^+, \tag{28}$$

and by the local interarrival time θ_n^m *during busy periods, where* θ_n^m *is defined by expression* (20) *and where* T_n^{m+1} *and* θ_n^m *are not correlated when no correlations exist between successive service times.*

In fact, the first component corresponds to any customer inside the entire busy period and is independent of the local interarrival time θ_n^m. On the contrary, the second component is specific for a given customer. The case of stage 2 is considered in Annex 1. For $m>1$ in case of Poisson input, the 2nd component $\boldsymbol{V_n^{m+1}}$ is evaluated in Annex 2a when successive service times are not correlated, and in Annex 2b when successive service times are correlated. Consequently, *Proposition* 1 *is valid even in the case of correlations.*

Notes: a) We note that the *concept of local traffic source cannot exist.* At stage 2, the local interarrival time is given by relation (1). When the downstream queue is empty (during the upstream busy period), the downstream busy period may be broken up when $T_n^2 < T_n^1$. In fact, for the following customer, no change appears if we suppose $T_n^2 = T_n^1$. Consequently, for evaluation of the local, *actual* queueing delay (at stage 3), we may consider that the busy period (at stage 2) is not broken up (during the busy period at stage 1). If $T_n^2 > T_n^1$, $[T_n^2 - T_n^1]^+$ may be considered as a service time generating a supplementary GI/G/1 server.

Finally, at stage 3 we may introduce two kinds of interarrival times: an *actual* idle period equal to $(\theta_n^2 + e_n^2)$, and a *virtual* idle period (when the busy period is

broken up) equal to θ_n^2. More generally at stage $(m+1)$, the interarrival time (during busy periods) may be considered as equal to θ_n^m, not influencing the first component U_n^{m+1} and avoiding the impact of broken up busy periods if we combine these periods with the service time τ_n^{m+1}. This comment (concerning the evaluation of the *actual* queueing delay), valid even when successive service times for the same customer are correlated, is consistent with relation (25), but not with classical tandem queue theories, since we have to consider these busy periods as not broken up (at stage > 2).

Proposition 1 avoids the difficulty of distinguishing the two cases of idle periods in order to evaluate the *actual* queueing delay. We may keep the assumption of busy periods not broken up as in our recent work (see Le Gall [6]). The agglutination phenomenon appears, and the phenomenon of amalgam (or coalescence) of busy periods amplifies from stage-to-stage. Note that, at stage 2, the server may be considered as a classical GI/G/1 server, since the concept of virtual idle period cannot exist at stage 1 (see in Annex 1).

b) When $T_n^k = T^k$ is deterministic, (25) leads to a very known result: the overall waiting time corresponds to the GI/G/1 queue defined by the set $(\theta_n^{m+1}, Y_{n-1}^1)$, when $\lambda.E(\theta_n^{m+1}) < 1$.

c) When T_n^k is not deterministic, we find $\lambda.E(\theta_n^m) > 1$ after having crossed 2 or 3 stages. Since the moments of θ_n^m increase when m increases, the second component will rapidly decrease, when the distribution function of T_n^{m+1} is independent of m. Finally, after having crossed several stages (e.g., 3 stages) and in order to evaluate the *actual* queueing delay, the local queue may be considered as *generated by* a tandem queue (not influenced by θ_n^m) with *identical successive service times* for the same customer. Practically, we may apply relation (9) (for e.g., $i = 4$) in which:

$$(T_n^{m+1} - T_n^{i-1}) \rightarrow (\theta_n^{m+1} - \theta_n^{i-1}) \cong 0.$$

As already mentioned in Section 3.2.2, the customers become *indistinguishable* since all the sojourn times are identical insider the local busy period.

Finally, when we consider any busy period as not broken up, an important consequence is: a customer, initiating a busy period at stage $(m+1)$, also initiates the upstream busy periods. *A strong correlation appears between the successive local sojourn times* for the same customer, in spite of the non-correlation between the successive service times. And this strong correlation *allows us to eliminate the possible interferences with the other upstream cross-traffic streams.*

d) As a consequence, when we observe a local queue directly, we cannot detect this correlation. We observe a *virtual* local queueing delay not experienced by the customer, due to the impact of the virtual idle period. To get a correct observation it is *necessary to observe*, for the same customer, the *two successive* (upstream and local) ***overall*** *sojourn times*, directly, in order to avoid observing the virtual idle periods. This is a consequence of the *non-existence of the concept of* ***local*** *traffic source*: this concept does not take into account the difference in correlations between the occupancy state and the idle period.

4.2 Case of the M/M/1→M/1 Tandem Queue and of Other Queues

The results above may be surprising since the tandem queue M/M/1→M/1 has been considered as a succession of mutually independent M/M/1 queues from a long time,

leading to an overall sojourn time considered as the sum of independent local sojourn times. In fact, the negative exponential service time distribution frequently does not generate long service times. Let us consider an example with $E(T_n^{m+1}) = 1$.

We set, in stationary regime and for $R(z) \geq 0$:

$$\begin{cases} \varphi_{m+1}(z) = \varphi_{m+1}(z, \infty) = \int_0^\infty e^{-z\alpha} dF_{m+1}(\alpha) = \dfrac{1}{1+z}, \\ \psi_2(z) = \dfrac{1}{1+z}\dfrac{1}{2}\left(1 + \dfrac{1}{1+z}\right), \\ \psi_m(z) = E(e^{-z\theta_n^m}), \text{ with } m \geq 2. \end{cases} \tag{29}$$

After stage 3, we may neglect the influence of the 2nd component $\boldsymbol{V_n^{m+1}}$ in *Proposition* 1, as defined in Annex 2. For the distribution function of U_n^k, we start from the numerical value of the sojourn time deduced from (11) and (15). In our example, we consider the case of an arrival rate $\lambda(=\rho) = 0.7$. We get, at stage $(m+1)$, for $E(s_n^{m+1})$:

Stage 1: $E(s_n^1) = 3.3$,

Stage 2: $E(s_n^2) = 3.3$, (see in Annex 1, *Proposition* 2.*a*);

Stage 5: $E(s_n^5) = 3.4$,

Stage 10: $E(s_n^{10}) = 3.7$,

Stage 20: $E(s_n^{20}) = 4.0$,

Stage 100: $E(s_n^{100}) = 4.8$.

With classical Jackson's queueing theory, in which departure process at stage k is equal to arrival process at stage $(k+1)$, we get $E(s_n^k) = 3.3$ at any stage for the *virtual* local sojourn time (as observed directly at a given stage by an external observer). For the mean *actual* local sojourn time (as experienced by the customer), the discrepancy may only be observed after having crossed 10 or 20 stages, despite the high increase (from stage 3) of the local interarrival time θ_n^m. But, when the long service times occur more frequently, the discrepancy with Markovian (or product form) theories appears more rapidly. For instance, in Le Gall [6], we mentioned that the service time *Pareto distribution* cannot be handled: when t increases indefinitely, the complementary distribution function decreases asymptotically as $(\text{at})^{-\alpha} (\alpha > 2)$, only, instead of a negative exponential decrease. And now, when successive service times of the same customer are not correlated, the result is unchanged, the service time *Pareto* distribution cannot yet be handled.

In the case of finite support for the service time distribution $\boldsymbol{T_1}$ and $\boldsymbol{T_N}$ denote the shortest and the longest service times, respectively. To avoid significant congestion in the buffers due to the "*agglutination phenomenon*", Le Gall [6] mentioned the need for the *buffer capacity* $\boldsymbol{K}$ (in number of customers) to satisfy the condition:

$$K > \frac{T_N}{T_1}. \tag{30}$$

Since Markovian theories cannot detect the "*agglutination phenomenon*", they are not appropriate to dimension the buffers.

5. Single Server Queueing Networks with Non-Correlated Successive Service Times

With the same assumptions as above in the case of non-correlated successive service times for the same customer, when he has already crossed two stages, we may apply our recent results in Le Gall [6] to consider single server queueing networks. A significant impact of upstream traffic streams appears when they are distributed, or not (at the adjacent upstream stage) towards different downstream directions, with this distribution generating indistinguishable "*premature departures*". The basic property used (see Le Gall [4]), considered the possibility of correlation or non-correlation between the local interarrival time and the upstream service time. When two successive local arrivals are separated by "premature departures" in the same upstream traffic stream, this correlation cannot exist. The local queue appears as a single GI/G/1 queue; the total arrival process is considered at the entry to the network. Note that it is the result traditionally used, but is justified by the concept of a local traffic source. This argument is wrong since these local traffic sources do not exist and could lead to significant errors in evaluating the influence of the upstream part of the network.

When these two successive arrivals are coming from the same upstream traffic stream without being separated by "premature departures", we are in the case of a tandem queue, which was considered in the preceding sections. Due to the fact that a local arrival, having already crossed two or several stages, and initiating a busy period, has also initiated the upstream busy periods in this tandem queue (on excluding other cross-traffic streams), the assumption of non-influence of the other traffic streams crossing upstream is justified. We have seen that, after having crossed three stages, we again find the equivalence with the case of identical successive service times. The concept of an *equivalent tandem queue* may be used with equivalence relation (11) in the case of successive local arrivals coming from different incoming paths (and consequently, not separated by premature departures). Finally, we deduce the second proposition:

Proposition 2: (Network with non-correlated successive service times) *In the case of a stationary regime, and after having crossed three stages in a single server queueing network with* non-correlated *successive service times for the same customer, the local* actual *queueing delay may be considered generated as in the case of hypothetical successive upstream service times supposed to be* identical *to the local service time, with the equivalent number of stages being defined by relation* (11).

Notes: a) This equivalence allows use of the solution given in Le Gall [6], in the case of identical successive service times for the same customer. In the local queue when not separated by "premature departures", the customers (and the incoming paths) become *indistinguishable* since all sojourn times become identical inside the local busy period. The traditional concept of local traffic source (generating distinguishable arrival epochs and queueing delays) disappears, sweeping away traditional theories.

b) However, in the relation of equivalence (11), $VarT'_n(m)$ is proportional to $\boldsymbol{m^2}$ in the case of identical successive service times. But, in the case of non-correlated successive service times, $VarT'_n(m)$ is proportional only to $\boldsymbol{m}$. So the parameter $\boldsymbol{m_0}$, obtained in the case of identical successive service times, is equal to $\sqrt{\boldsymbol{m_0}}$ only in the case of non-correlated successive service times. Consequently, from stage-to-stage, busy periods amalgamate slowly, and the mean local queueing delay may be slow to

increase.

c) This is not the case when service time durations vary highly, leading to a significant congestion in buffers generated by the "*agglutination phenomenon*" of indistinguishable customers (due to the identical sojourn times inside a given busy period), even when the local load (= traffic intensity) is slight, where half of the local load corresponds to two successive local arrivals not separated by "premature departures". It is necessary to satisfy condition (30) for buffer dimensioning, even when successive service times are non-correlated.

d) *Proposition* 2 relates to the evaluation of the local *actual* queueing delay (dependent on upstream stages). For the evaluation of the local *virtual* queueing delay (at a given stage), it is sufficient to apply classical theories.

6. Conclusion

Due to the relation of equivalence (11) and to *Propositions* 1 and 2, it was simple to refer to our recent work given in Le Gall [6] and prove that classical theories are not appropriate for the evaluation of the local *actual* queueing delay (as experienced by the customer) and for buffer dimensioning, even when successive service times (for the same customer) are non-correlated.

This discrepancy corresponds to the case of two successive local arrivals not separated by "*premature departures*", leading to the combination of *indistinguishable* customers (due to identical sojourn times inside the upstream busy period), which is not consistent with traditional assumptions. Consequently, two successive sojourn times (of the same customer) are correlated as in packet traffic (i.e., with successive identical service times). After having crossed three stages, the *tandem queue effect* appears with the non-influence of the local interarrival time and with the *agglutination phenomenon* in buffers, which is not detected in Markovian theories. Due to the amalgam (or coalescence) of busy periods from stage-to-stage, the customer is waiting more after having crossed several stages, which cannot be detected by an external observer considering a specific single stage, without distinguishing virtual and actual idle periods.

Moreover, due to this tandem queue effect, the impact of the longest service times is stronger than in Markovian theories, particularly in large networks in case of overload in a given incoming path (even for a slight total local load). But, to observe this phenomenon, it is necessary to apply the method as recommended in Section 4.1, Note (d), because classical and local observation methods are appropriate to observe the broken up busy periods and the local *virtual* queueing delay, only. In other words, it is necessary to follow the customer instead of observing directly the local queue concerned, because a broken up busy period cannot be perceived by the customer. Finally, the customer can perceive busy periods much longer and he can find buffer occupancies much higher.

Finally, when there are not many premature departures, the concept of local *actual* queueing delay is not consistent with the traditional concept of local traffic source, generating distinguishable customers influenced by the local interarrival time, as usually considered in large queueing networks, following our comments in Section 4.1, Note (a). Due to relation (25) with expression (20) to increase the interarrival time of distinguishable local arrivals during congestion, their influence decreases and gives place to the impact of the sojourn time of the distinguishable customer initiating the

actual busy period and depending on the actual idle period, only. In particular, we may observe a curious phenomenon in *concentration nodes*, where each output buffer is serving a single (and different) direction, working with the tandem queue effect (*since there are no premature departures*). In that case, the downstream input buffer, receiving different links from similar concentration nodes and combining indistinguishable customers, also generates an *equivalent tandem queue* with indistinguishable customers. Due to the agglutination phenomenon, this input buffer may be permanently congested *even at slight load* when the local service times are highly varying, and when condition (30) is not satisfied. This phenomenon needs to standardize service time variations and define more appropriate structures in the network, which is not yet clearly understood by engineers accustomed to traditional Markovian theories, particularly when applied to distinguishable customers. On the contrary, when condition (30) is satisfied, leading to larger buffer capacities (in number of customers), the classical theories may be used again. A double faced traffic modeling appears, as for Janus divinity: a tandem queue effect for small buffers and indistinguishable customers, and a traditional process for large buffers and distinguishable customers.

This double faced traffic modeling cannot be detected by simplified traffic simulation methods. Instead of separately observing each customer from stage-to-stage, it may be faster to globally manage (at a given stage) all the local arrivals on writing the similarity between the departure process of preceding customers and the process formed at the beginning of service of next customers. Unhappily, it is not true in the case of a customer served a long time upstream and not waiting downstream: the *virtual* idle period and the increase of the interarrival time (see expression (27)) cannot be detected, evading the impact of the longer upstream service times. Consequently, this kind of simplified simulation leads to the principle of independence of stages, removing the impact of upstream link overloads and of long service times, i.e., liquidating the *tandem queue effect* that appears due to some incoming link overloads.

Annex 1. Two-Stage Tandem Queues with Poisson Input

Proposition 1 cannot be applied to the second stage in single server tandem queues, because (at stage 2) the local interarrival time satisfies relation (1), independent of expression (20). To simplify, we assume a *Poisson input* at stage 1 (with *stationary regime*) and evaluate the *actual* queueing delay at stage 2. In that case, the concept of a *virtual* idle period (see Section 4.1, Note (a)) does not exist at stage 1. The load (i.e., traffic intensity) at stage i is denoted ρ_i.

a) The Arrival Process at Stage 2: From relation (1) and for the **nth** customer, the interarrival time at stage 2 is: $T_n^1 + e_n^1$. In the case of busy server 1, this time is T_n^1 only, even when the busy period is broken up at stage 2. In the case of an idle period (at stage 1), this time is $(T_n^1 + e_n^1)$, where the density of arrivals in e_n^1 is λ. We let:

$$Q_0 = Prob(w_n^1 = 0) = 1 - \rho_1. \tag{1a}$$

In this case of Poisson input (at stage 1), we note that the sequences T_n^1 and e_n^1 are independent and that each sequence is identically distributed. Consequently, the

arrival process (at stage 2) is regenerative and, using notations of Section 2.1, we may write for an arbitrary arrival:

$$Prob(Y^2_{n-1} \leq x) = (1 - Q_0) \cdot F_1(x) + Q_0 \cdot F_1(x) * (1 - e^{-\lambda x}). \tag{2a}$$

From the notations of Section 2.2 and from (12), the Laplace-Stieltjes transform for this distribution is:

$$\gamma_2(z) = (1 - Q_0) \cdot \varphi_1(z) + Q_0 \cdot \varphi_1(z) \cdot \frac{\lambda}{\lambda + z},$$

or

$$\gamma_2(z) = \varphi_1(z) \cdot \frac{\lambda + (1 - Q_0)z}{\lambda + z}, \quad \text{with } 0 < Q_0 < 1. \tag{3a}$$

We deduce:

Proposition 1a: *In the case of* Poisson input *at stage* 1 *and a* stationary regime, *the* actual *queueing delay* (*of an arbitrary customer*) *at stage* 2 *is governed by the GI/G/1 server* $[\gamma_2(z), \varphi_2(z)]$, *with* $\gamma_2(z)$ *being defined by* (3*a*) *and* $\varphi_2(z)$ *by Section* 2.2.

Notes: We assume that service times are not deterministic and an arbitrary customer at stage 2 means that he can wait or not at stage 1.

b) The Distribution of w_n^2: As for expressions (12), we refer to Pollaczek [8] to define the Laplace-Stieltjes transform $\Phi_0(z)$ of the distribution of the *actual* queueing delay, w_n^2. From *Proposition* 1a for a stationary regime, we may write:

$$\Phi_0(z) = Ee^{-zw_n^2} = Exp\left(-\frac{1}{2\pi i} \int_{-0} \left[\frac{1}{z-u} + \frac{1}{u}\right] \cdot log[1 - \gamma_2(-u) \cdot \varphi_2(u)] \cdot du \right). \tag{4a}$$

From Pollaczek [8], we deduce the probability of a (virtual or actual) idle period at stage 2:

$$\boldsymbol{Q_1} = Exp\left(-\frac{1}{2\pi i} \int_{-0} log[1 - \gamma_2(-u) \cdot \varphi_2(u)] \cdot \frac{du}{u} \right) < 1. \tag{5a}$$

We get the rth cumulant of the distribution of the *actual* queueing delay w_n^2:

$$C_r = \frac{(-1)^r}{2\pi i} \cdot \int_{-0} log[1 - \gamma_2(u) \cdot \varphi_2(u)] \cdot \frac{du}{u^{r+1}}. \tag{6a}$$

In particular, we get:

$$E(w_n^2) = C_1, \qquad Var(w_n^2) = C_2. \tag{7a}$$

c) Case of Negative Exponential Service Time Distributions: In the case of negative exponential service time distributions, we may write from (3*a*):

$$\begin{cases} E(T_n^1) = \frac{1}{\mu_1}, & \rho_1 = \frac{\lambda}{\mu_1};\ \gamma_2(z) = \frac{\mu_1}{\mu_1 + z} \cdot \frac{\lambda + (1 - Q_0)z}{\lambda + z}; \\ E(T_n^2) = \frac{1}{\mu_2}, & \rho_2 = \frac{\lambda}{\mu_2}. \end{cases} \tag{8a}$$

Expression (3*a*) becomes:

$$\gamma_2(z) = \frac{\lambda}{\lambda + z}. \tag{9a}$$

We again find the same Poisson input at stage 2 as at stage 1. We may conclude:

Proposition 2a: (Case of identical successive service time distribution functions) *For the tandem queue M/M/1→M/1, in stationary regime, the* actual *queueing delay of an arbitrary customer at stage* 2 *is governed by an M/M/1 server.*

For more general tandem queues, the *actual* queueing delay and the *virtual* queueing delay (at stage 2) are different. *Proposition* 2*a* is not valid at stages > 2 (see Annex 2), which is not consistent with Jackson's theory: the concept of traffic source is not valid in this downstream stages (see our comments in Section 4.1, Note (a)). Finally, *Proposition* 1 is much easier to apply.

Notes: In this text, to avoid some misunderstanding, we used the term "*queueing delay*" because the term "*waiting time*" sometimes includes the service time.

Annex 2. $(m+1)$-Stage Tandem Queues with Poisson Input

In this case of single server tandem queue with Poisson input, we want to evaluate (Section *a*) the second component $\boldsymbol{V_n^{m+1}}$ in *Proposition* 1, at stage $(m+1)$ with $m > 1$, and we will give a simple extension to the more general case of correlated successive service times (Section *b*).

a) Evaluation of the 2nd Component $\boldsymbol{V_n^{m+1}}$ in Proposition 1 ($\boldsymbol{m > 1}$): We want to extend expression (3*a*) of $\gamma_2(z)$. In the *stationary regime*, we let:

$$Q_{m+1} = Prob(w_n^{m+1} = 0), \tag{1b}$$

corresponding to an *actual* idle period and given by the first component $\boldsymbol{U_n^{m+1}}$ in *Proposition* 1. Due to this component, the actual idle period at stage $(m+1)$ corresponds to the actual idle period at upstream stages. Thus, we have (at stationary regime):

$$Q_{m+1} = W_{m+1}(0), \tag{2b}$$

where $W_{m+1}(t)$ is the distribution function of the unitary queueing delay per stage (i.e., the overall queueing delay divided by the number of stages, excluding the first stage). In Le Gall [2], Annex B, for the distribution of the corresponding sojourn time, when the $lim_{t\to\infty} t[1 - F_{m+1}(t)] = 0$, or when T_n^{m+1} has a finite support (notations of Section 3.2.2), we gave:

$$S_{m+1}(t) \cong \left[\frac{1-\rho}{1-\rho F_0(t)}\right]^{m+1} \cdot F_{m+1}(t), \tag{3b}$$

with

$$\begin{cases} \rho = \lambda \cdot E(T_n^{m+1}), \\ F_0(t) = \dfrac{1}{E(T_n^{m+1})} \displaystyle\int_0^t [1 - F_{m+1}(u)] \cdot du. \end{cases} \tag{4b}$$

Due to the stochastic relation between the sojourn time s_n^{m+1} and the queueing delay w_n^{m+1}, we get:

$$w_n^{m+1} = s_n^{m+1} - T_n^{m+1}.$$

We deduce the expression:

$$W_{m+1}(t) \cong \int_0^{\infty} \left[\frac{1-\rho}{1-\rho F_0(t+u)}\right]^{m+1} \cdot dF_{m+1}(u), \tag{5b}$$

and, consequently from (2*b*)

$$Q_{m+1} = \int_0^{\infty} \left[\frac{1-\rho}{1-\rho F_0(u)}\right]^{m+1} \cdot dF_{m+1}(u). \tag{6b}$$

To evaluate the arrival process we note that, not considering the existence of broken up busy periods by considering an interarrival time θ_n^m during busy periods, the arrival process is still regenerative as in Annex 1. Let

$$\psi_m(z) = Eexp(-z\theta_n^m). \tag{7b}$$

From (3*a*), the L-S transform of the interarrival distribution, at stage $(m+1)$ becomes:

$$\gamma_{m+1}(z) = \psi_m(z) \cdot \frac{\lambda + (1-Q_{m+1})z}{\lambda + z}, \tag{8b}$$

with Q_{m+1} being given by expression (6*b*). The actual local queueing delay of the second component $\boldsymbol{V_n^{m+1}}$ corresponds to the L-S transform of this delay, deduced from (4*a*):

$$\begin{cases} \Phi_{m+1}(z) = Eexp(-zw_n^{m+1}) \\ = Exp\left(-\frac{1}{2\pi i} \cdot \int_{-0} \left[\frac{1}{z-u} + \frac{1}{u}\right] \cdot log[1-\gamma_{m+1}(-u) \cdot \beta_{m+1}(u)] \cdot du\right), \end{cases} \tag{9b}$$

where $\beta_{m+1}(z) = Eexp[-z\tau_n^{m+1}]$ is defined by (28).

b) Case of Correlated Successive Service Times: *Proposition* 1 is still valid when successive service times T_n^{m+1} (n fixed) are *correlated*. But now, T_n^{m+1} and θ_n^m are correlated. In expression (9*b*), using expression (8*b*), we have to make the substitution

$$\psi_m(-z) \cdot \beta_{m+1}(z) \rightarrow Eexp(-z[\tau_n^{m+1} - \theta_n^m]). \tag{10b}$$

References

[1] Le Gall, P., The overall sojourn time in tandem queues with identical successive service times and renewal input, *Stoch. Proc. and their Appl.* **52** (1994), 165-178.

[2] Le Gall, P., Traffic modeling in packet switched networks for single links, *Annales des Telecom.* **49**:3-4 (1994), 111-126.

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

[4] 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.

[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] Le Gall, P., Single server queueing networks with varying service times and renewal input, *J. of Appl. Math. and Stoch. Anal.* **13**:4 (2000), 429-450.

[7] Pollaczek, F., Application d'opérateurs intégro-combinatoires dans la théorie des intégrales multiples de Dirichlet, *Ann. de l'Institut Henri Poincaré* **XI**:III (1950), 113-133.

[8] 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).