Journal of Applied Mathematics and Stochastic Analysis, **11**:3 (1998), 377-390.

MULTISERVER QUEUEING NETWORKS AND THE TANDEM QUEUE MODEL

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

(Received November, 1997; Revised May, 1998)

Using a tandem queue model we evaluate the local "*endogenous*" (= internal) queueing delay in single server and multiserver queueing networks. The new concept of the *apparent* overall upstream queueing delay(as perceived by the downstream network) allows us to analyze the distribution of this local queue by interpolating between the distributions of the *tandem queue* (generated by a *concentration tree*) and the isolated G/G/s *queue*. The interpolation coefficients depend on the proportion of "*premature departures*", typically interfering in the upstream stage and leaving the considered path without being offered to the considered local queue. On the other hand, local "*exogenous*" arrivals (from outside the network) require the introduction of the "*interference delay*" concept. Finally, in the case of *single server* queueing networks, we stress the need to *extend the capacities of the buffers*, by considering the "*worst case*" scenario and by using an "*equivalent tandem queue*" model.

Key words: Queueing Networks, Concentration Tree, Tandem Queues, Local Queueing Delay, Jitter Delay, Endogenous and Exogenous Arrivals, Interference Delay.

AMS subject classifications: 60K25, 90B22.

1. Introduction

In this paper, we analyze the local queueing delay of single server and multiserver queueing networks. We assume that customers only gain access to a downstream queue after completion of of the upstream service. Service discipline at all queues is "*first come-first served*". The traditional approximation of networks by means of an isolated G/G/s queue relies too heavily on the use of a *local* traffic source, as if an output stream from the upstream stage could be considered as a simple traffic source for the downstream stage with no regard to any other influence from the environment. Unfortunately, this modeling approach of the "local traffic source" hides the effects of a number of queueing phenomena which can be significant, especially in the

Printed in the U.S.A. ©1998 by North Atlantic Science Publishing Company

case of single server networks. For instance, models usually ignore the local *interference delay* caused to an internal ("*endogenous*") customer by another "*exogenous*" customer arriving directly (from outside the network) at the considered local queue, during the upstream service time of the preceding *endogenous* customer. This supplementary queueing phenomenon may be accumulated in the other upstream stages. In addition, the impact of "*premature departures*" is not completely taken into account in meshed networks. These departures interfere with the upstream progression of customers, even though these departures are not offered to the considered local queue, since they leave the considered path just before this local queue. Even though the effect of these phenomena is attenuated in multiserver queueing networks, as shown in the results of traffic simulations, this is not a sufficient reason to justify inaccurate models, notably in the case of single server networks.

To take into account the interference between partial traffic streams handled by the considered local queue we consider the *concentration tree*, over which these traffic streams converge towards that queue. When we introduce the new concept of "*apparent*" overall upstream queueing delay (as perceived by the local queue), the preceding method of the concentration tree will allow us to consider only the "*premature departures*" present at the previous stage. It follows that we only need to analyze a simplified network as shown in Figure 1. This can be reduced to the "truncated" network of Figure 2, where partial traffic stream A_i interferes with "premature departures" of stream B_i [not offered to final stage $(m+1)$] throughout the *m-stage* tandem queue $\boldsymbol{i}$ $(i = 1,\ldots,n)$. The case of tandem queues of different lengths $\boldsymbol{m_i}$ will be included in the definition of an overall equivalent tandem queue.

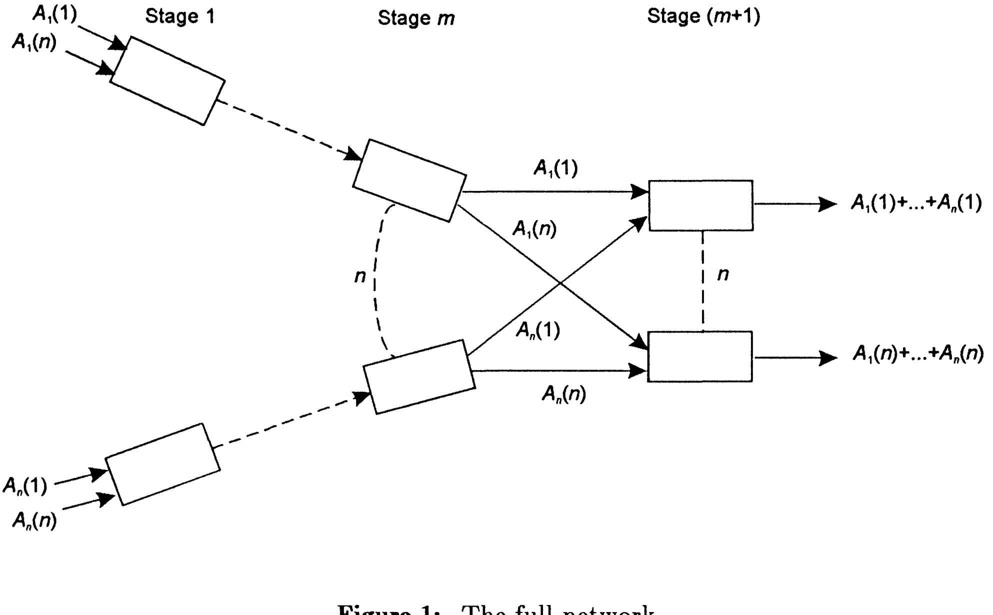

Figure 1: The full network

To simplify the analysis and to eliminate the influence of the above mentioned *interference delay* between local "*endogenous*" arrivals (within the network), we assume that *busy periods are not broken up*. This assumption is reasonable in the case of heavily loaded networks (when service times do not vary too widely) and is adequate in the case of packet switched networks with the same transmission speed at every stage. It follows (as we will see) that all partial traffic streams have the same distribution of local queueing delay, due to some "agglutination effect" that makes the use of the "interference delay" concept unnecessary. Consequently, this distribution will be evaluated for an *arbitrary "endogenous"* traffic stream. In other words, the above mentioned case of incoming tandem queues with distinct lengths m_i can be replaced by the symmetric case of Figure 2 for a given total traffic intensity at the final stage.

In the recent paper [2], we analyzed the case of networks of *single server queues* showing that the distribution of the *local endogenous queue* could be derived as an interpolation between the distributions of the isolated G/G/1 queue and the *overall tandem queue* (equivalent to the above defined concentration tree, jitter delay excluded). It is the *tandem queue concept* which allows us to make use of the upstream part of the network. The interpolation coefficients depend on the degree to which the network is meshed. The latter determines the proportion of "premature departures". These coefficients also depend on the heavy load appearing in the considered partial traffic stream. This may explain the impossibility (at the busy hour) of a very high local queueing delay, which is more realistic, compared to the G/G/1 queue, which would lead to a very large queueing delay! In the present work we extend this analysis to the case of multiserver networks and show to what extent the approximation by the isolated G/G/s queue is more easily justified.

In Section 2 we outline the earlier study of single server queueing networks before extending the analysis in Section 3 to the case of multiserver networks. Finally, in Section 4 we conduct a numerical study and compare the values obtained with the results of traffic simulations in the case of packet switched networks with two links arbitrarily handling two populations of packets of very different lengths leading therefore to widely differing service times. In Section 2 and 3 the arrival processes at the network input are assumed to be governed by a general probability distribution in the stationary regime. In Section 4 the study is limited to the case of Poisson arrivals and we evaluate the traffic handled in the *buffers*, particularly in the case of *single links*. In the latter case, we stress the need to *extend the capacities of the buffers* by using an *"equivalent tandem queue" model.*

2. Single Server Queueing Networks

2.1 Notation and Assumptions

Let us consider the truncated network of Figure 2. We recall that the queueing discipline (in each successive queue) is *"first come-first served"*. The system is assumed to be in the *stationary regime.*

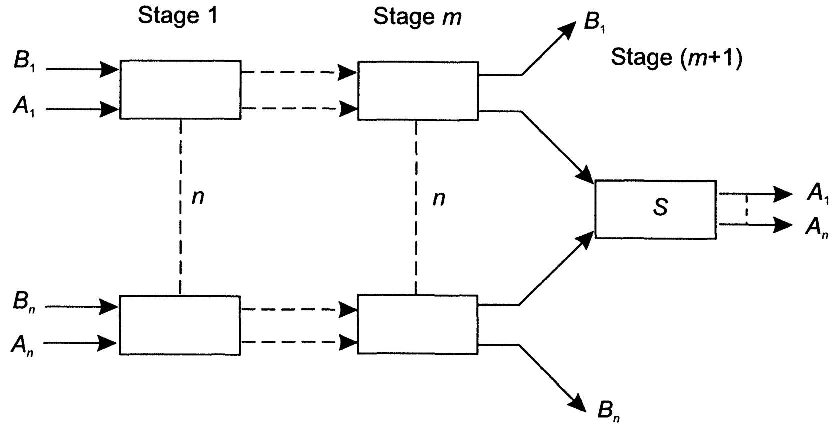

Figure 2: The truncated network

For each of the n *identical and independent tandem queues* of Figure 2, at *stage* k $(k = 1,\dots,m)$ and for the j^{th} customer at the considered queue, we set:

- *arrival epoch* at stage k: X'^k_j;
- *interarrival time* between customers $(j-1)$ and j: $Y'^k_{j-1} = X'^k_j - X'^k_{j-1}$;
- *service time*: T^k_j;
- *local queueing delay*: w^k_j:
- *sojourn time*: $s^k_j = w^k_j + T^k_j$;
- *occasional idle period* during the interarrival interval Y'^k_{j-1}: e'^k_j.

Finally, for the considered *final stage server*, the sequence of arrivals will be indexed by i with the processes:

- *interarrival interval at the input* to all n tandem queues, for the customers offered to the considered final server: Y_{i-1};
- *arrival rate* (stationary regime): λ;
- *interarrival interval* at the considered final stage server: Y''_{i-1};
- *service time*: $T_i[T = E(T_i)]$;
- *load* (= traffic intensity in stationary regime): $\rho = \lambda \cdot T$;
- *local queueing delay*: w_i;
- *sojourn time*: $s_i = w_i + T_i$;
- *occasional idle period* during the interarrival interval Y''_{i-1}: e_i.

Note the difference between the arrival process Y'^1_{j-1} for a given tandem queue and the process Y_{i-1} relating to all tandem queues but only for customers destined for the considered final stage server. The couple $[Y_{i-1}, T_i]$ defines an *isolated G/G/1 server* handling the same partial traffic streams as those handled by the considered final stage server. Its queueing delay will be denoted W_0 in the *stationary regime*. In this regime, the arrival processes Y'^1_{j-1} and Y_{i-1} are *governed by some general probability distribution.*

2.2 The Jitter Effect

To be able to use the concept of *equivalent tandem queue* in order to evaluate the local queueing delay at the final stage, we need to replace the actual process Y''_{i-1} by a process keeping the same arrival order as at the entry to the network, since a tandem queue keeps the same arrival order at each stage. The difference between the two processes comes from the mutual independence of the $\boldsymbol{n}$ incoming tandem queues in Figure 2.

This difference generates a *local jitter effect.* All the customers of the same local busy period experience the same *local jitter delay* $\boldsymbol{J}$, whose distribution function has been approximately evaluated in Le Gall [2], Subsection 3.1, for large n (Figure 2):

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

This jitter effect is only significant for heavily loaded networks.

2.3 The Equivalent Tandem Queue

Firstly, we assume that there are no "premature departures" (i.e., no traffic streams B_i in Figure 2) and no local "exogenous" (coming from outside) arrivals. The local queue at the final stage of the concentration tree (already mentioned in the introduction) may then be equivalent to that of an *equivalent tandem queue* as defined in Le Gall [2], Subsections 2.1, 2.5 and 2.7. To understand this, consider the relations (at stage $k-1 < m$):

$$s_{j-1}^{k-1} - Y'^{k-1}_{j-1} = \left(s_{j-1}^{k-1} - Y'^{k-1}_{j-1}\right) + \left(s_{j-1}^{k-1} - Y'^{k-1}_{j-1}\right)^{-} = w_j^{k-1} - e'^{k-1}_j. \tag{2}$$

At stage k we may write:

$$w_j^k = \left(s_{j-1}^k - Y'^k_{j-1}\right)^+, \text{ with } Y'^k_{j-1} = Y'^{k-1}_{j-1} + \left(s_j^{k-1} - s_{j-1}^{k-1}\right), \tag{3}$$

which leads to the following relation:

$$s_j^{k-1} + w_j^k = \text{Max}\left[s_j^{k-1}, \left(s_{j-1}^{k-1} - Y'^{k-1}_{j-1}\right) + s_{j-1}^k\right]. \tag{4}$$

If we use relation (2) and if we subtract w_j^{k-1}, this relation becomes

$$T_j^{k-1} + w_j^k = \text{Max}\left[T_j^{k-1}, s_{j-1}^k - e'^{k-1}_j\right]. \tag{5}$$

If the busy periods are not broken up, the downstream busy period corresponds to the upstream busy period. Consequently, during the downstream busy period (at stage k) we have no idle period, $e'^{k-1}_j = 0$, and

$$T_j^{k-1} + w_j^k = s_{j-1}^k. \tag{6}$$

The arrival process disappears during the local busy period. Its influence only appears to initiate the busy period through the jitter effect. Finally, we do not change the local queueing delay (excluding the jitter effect) if we replace a nonsymmetrical network by the symmetrical network of Figure 1, provided we keep the same total traffic intensity and the same length for the equivalent tandem queue.

From (6), we deduce that the busy period is not broken up if the following condition is satisfied:

$$T_j^{k-1} \leq s_{j-1}^k. \tag{7}$$

In the case of packet switched networks, we have $T_j^{k-1} = T_j^k$, if the transmission speed is the same at each stage. Relation (6) becomes:

$$s_j^k = s_{j-1}^k = \ldots = s_{j_0}^k, \tag{8}$$

where j_0 corresponds to a customer initiating the busy period. The sojourn time has the same value for all the customers of the same busy period: it corresponds to some *agglutination phenomenon*, which also appears at the final stage of the equivalent tandem queue.

Moreover, relation (7) may be satisfied even in the case of mutually independent successive service times *for heavily loaded networks* when the extended delays at the final stage tend to cause busy periods to amalgamate, if service times do not vary too widely. (See Hypothesis 2 in Section 2.7 of [2].)

2.4 The Impact of Premature Departures

Now we suppose that "premature departures" take place at the upstream stage. They are not offered to the final stage server considered and correspond to traffic streams B_i in Figure 2.

To evaluate the local queueing delay we have to subtract the *overall upstream queueing delay* $\boldsymbol{V(1;m)}$, from stage 1 to stage m, from the *overall queueing delay* $W(1;m+1)$ from stage 1 to stage $(m+1)$. In fact, in Le Gall [2], we showed that we do not have to use the observed value $\boldsymbol{V(1;m)}$ related to all the upstream customers. Observe that the upstream customers considered are offered to the final stage server considered, with a probability $(1/n)$ in Figure 2. Finally, we use the "*apparent*" *over all queueing delay* as perceived by the local queue in the form

$$V'(1;m) = \boldsymbol{h} \cdot V(1;m), \tag{9}$$

where $\boldsymbol{h}$ is the random variable $= 1$ with probability $(1/n)$ and $= 0$ with probability $[1-(1/n)]$. We will also introduce this "*apparent*" delay in the evaluation of $W(1;m+1)$, since we wish to evaluate the local queueing delay.

In the general, nonsymmetrical case, we define at *upstream stage* $\boldsymbol{m}$ (including all the incoming tandem queues):

- *total load* (= traffic intensity): $\boldsymbol{a}$;
- *part* of this total load corresponding to the *customers offered to the considered final stage server*: $\boldsymbol{a'}$;
- *part* of this total load corresponding to "*premature departures*": $\boldsymbol{a''}$.

We have:

$$a = a' + a''. \tag{10}$$

Following our comment after relation (6), we may replace this nonsymmetrical case by the symmetrical case of Figure 2 where the number of identical incoming tandem queues is:

$$\boldsymbol{n} = \text{integer part of } a/a'. \tag{11}$$

The length $\boldsymbol{m}$ of these identical tandem queues was defined in Le Gall [2], Section 2.7. Now, in relation (9) the random variable $\boldsymbol{h}$ is defined by expression (11).

2.5 The Local Queueing Delay Without Local Exogenous Arrivals

We suppose that all local customers come from an upstream stage ("*endogenous* *arrivals*"). No customer arrives directly from outside the network ("*exogenous arrivals*"). In Le Gall [2], Theorem 5, we derived the key formula for the local queueing delay $\boldsymbol{w}$ of an *arbitrary customer of any partial traffic stream*, at the final stage server considered in the *steady state*:

$$\boxed{Ee^{zw} = \left(\tfrac{a'}{a}\right) \cdot Ee^{z\boldsymbol{D(m)}} + \left(1 - \tfrac{a'}{a}\right) \cdot Ee^{z\boldsymbol{W}_0}}, \tag{12}$$

where $\boldsymbol{D(m)}$ is the local queueing delay of the "*concentration tree*", i.e., of the $(m+1)$*-stage equivalent tandem queue including the jitter delay* at the final stage. Finally, expression (12) combines the case of *no premature departures* or *heavy load* in the considered partial traffic stream [local queueing delay $\boldsymbol{D(m)}$ with actual upstream load] and the case of *many premature departures* corresponding to a local queueing delay $\boldsymbol{W}_0$ of the isolated G/G/1 server. This simple result is *due to the fact that the busy periods are not broken up.* Moreover, *it explains why a heavy endogenous load cannot generate a very large local queueing delay* (as opposed to an isolated G/G/1 queue).

2.6 Case of Local Exogenous Arrivals

We suppose the existence of local *exogenous* arrivals at the final stage server (i.e. not coming from the upstream stage) with the following notation in the *stationary mode*:

- *total* load (= traffic intensity): ρ;
- *endogenous* load: ρ_0;
- *exogenous* load: ρ_1.

We have:

$$\rho = \rho_0 + \rho_1. \tag{13}$$

We include the existence of exogenous arrivals for the evaluation of $\boldsymbol{W}_0$ and $\boldsymbol{D(m)}$ in expression (12) by supposing that the exogenous arrivals are offered to stage 1 with zero service time in successive stages 1 to m and with the actual service time at the final stage $(m+1)$. But to apply relation (9) we have to take care, as we will explain below.

In the evaluation of $\boldsymbol{W}_0$ and $\boldsymbol{D(m)}$, we suppose that the interferences are only due to the arrivals (at the entry to the network) and to the service times. Now, we note that the exogenous arrivals may be immediately served during the overall upstream queueing delay of some endogenous arrivals generating a new *interference delay.* To take this phenomenon into account in relation (9) we will distinguish between the two kinds of local arrivals.

For an arbitrary *endogenous arrival*, we note that an upstream customer may only be offered to the considered final stage server if this server is not busy due to an exogenous arrival (probability of occupancy: ρ_1). Finally, in expression (12) we have to make the substitution:

$$\tfrac{a'}{a} \rightarrow \tfrac{a'}{a} \cdot (1 - \rho_1). \tag{14}$$

For an arbitrary local *exogenous arrival*, expression (9) says that the considered final stage server is busy due to an arbitrary endogenous arrival. In expression (12) we have to make the substitution:

$$\tfrac{a'}{a} \rightarrow \tfrac{a'}{a} \cdot \rho_0. \tag{15}$$

This method of evaluation could not be a part of the classical concept of local traffic source.

3. Multiserver Queueing Networks

Now, when considering the networks of Figure 1 and 2, we replace each single server by a *multiserver with a capacity of L single servers.* We retain the same load for each new server as in the previous case and we keep the same service discipline (FC-FS) for each queue. Finally the arrival process $\boldsymbol{Y}'^{1}_{j-1}$ and $\boldsymbol{Y}_{i-1}$, as considered at the last paragraph of Subsection 2.1, are *still governed by some general probability distribution.* We want to extend expression (12), which needs some preliminary remarks to use the concept of the equivalent tandem queue.

3.1 The Equivalent Tandem Queue

Firstly, we need to figure out to what extent can we go when using relations similar to (2)-(7) above. Let us consider a G/G/L queue with the vectorial equation given by Borovkov [1]. At stage k ($\leq m$), the interarrival time is still denoted $\boldsymbol{Y}'^{k}_{j-1}$. Let $\boldsymbol{w}^{k}_{j,h}$ be the queueing delay until (at least) $\boldsymbol{h}$ servers are free for arrival $\boldsymbol{X}^{k}_{j}$. The *vector* $\boldsymbol{w}^{k}_{j}$ has coordinates $w^{k}_{j,h}$ $(h = 1,\ldots,L)$. Let $R(x)$ be the vector with entries $(x_1,\ldots,x_L)$, where $x_1 = \mathrm{Min}_h(x_h)$, and $\boldsymbol{e} = (1,0,\ldots,0)$ and $\boldsymbol{i} = (1,\ldots,1)$. Borovkov derived the following general local equation (at stage k):

$$w^{k}_{j} = \mathrm{Max}\Big[0, R(w^{k}_{j-1} + T^{k}_{j-1}\cdot e) - \boldsymbol{Y}'^{k}_{j-1}\cdot \boldsymbol{i}\Big]. \tag{16}$$

Analogous to relations (2), we may write at stage $(k-1)$:

$$\begin{aligned} A &= \mathrm{Min}_h(w^{k-1}_{j-1,h}) + T^{k-1}_{j-1} - \boldsymbol{Y}'^{k-1}_{j-1} = [A]^{+} + [A]^{-} \\ &= \mathrm{Min}_h(w^{k-1}_{j,h}) - e'^{k-1}_{j}. \end{aligned} \tag{17}$$

e'^{k-1}_{j} is positive on an *idle period or due to a partial occupancy* at stage $(k-1)$ during the time $\boldsymbol{Y}'^{k-1}_{j-1}$. Consequently, we have:

$$\begin{aligned} Y'^{k}_{j-1} &= \boldsymbol{Y}'^{k-1}_{j-1} + \Big[\mathrm{Min}_h(w^{k-1}_{j,h}) + T^{k-1}_{j}\Big] - \Big[\mathrm{Min}_h(w^{k-1}_{j-1,h}) + T^{k-1}_{j-1}\Big] \\ &= \Big[\mathrm{Min}_h(w^{k-1}_{j,h}) + T^{k-1}_{j}\Big] - \mathrm{Min}_h(w^{k-1}_{j,h}) + e'^{k-1}_{j} \\ &= T^{k-1}_{j} + \mathbf{e}'^{k-1}_{j}. \end{aligned} \tag{18}$$

Let us apply this expression to (16). We deduce the following relation, similar to (5):

$$T^{k-1}_{j}\cdot \mathrm{i} + w^{k}_{j} = \mathrm{Max}\Big[T^{k-1}_{j}\cdot \boldsymbol{i}, R(w^{k}_{j-1} + T^{k}_{j-1}\cdot e) - \boldsymbol{i}\cdot e'^{k-1}_{j}\Big]. \tag{19}$$

As for the single server case, *we will suppose that the busy periods are not broken up* and, consequently, $e'^{k-1}_{j} = 0$ during the local busy period, leading to:

$$T_j^{k-1} \cdot \boldsymbol{i} + w_j^k = R(w_{j-1}^k + T_{j-1}^k \cdot \boldsymbol{e}). \tag{20}$$

The sojourn time is:

$$s_{j-1}^k = \mathrm{Min}_h(w_{j-1,h}^k) + T_{j-1}^k. \tag{21}$$

Taking this expression into account, relation (20) gives:

$$T_j^{k-1} + \mathrm{Min}_h(w_{j,h}^k) = s_{j-1}^k. \tag{22}$$

This expression is consistent with condition (7) to avoid breaking up the busy periods and, consequently, the same customer initiates the busy period in each successive upstream stage. The same arguments as for the single sever case may be used to introduce the concept of equivalent tandem queues if there are no "premature departures". But there are two apparent differences:

- a multiserver tandem queue cannot keep the same arrival order at each successive stage: consequently, *the local jitter effect disappears*;
- unless successive service times are constant, we cannot define the string of successive servers used by the same customer (particularly in the case of interferences with "premature departures") and, consequently, we cannot evaluate with accuracy the number $(m+1)$ of stages for the *single server* equivalent tandem queue. Due to the slight impact of $\boldsymbol{m}$, we will approximate the equivalent tandem queue by taking $m = 1$.

On the other hand, from (22), we may deduce again that the local queueing delay (at the final stage) does not change if we replace a nonsymmetrical network of multiservers by a *two-stage* symmetrical network of *single servers* as illustrated in Figure 2. *The local queueing delay at the final stage of the equivalent* <u>*two-stage*</u> *tandem queue will be denoted* $\boldsymbol{D}(\mathbf{1})$.

3.2 The Local Queueing Delay Without Local Exogenous Arrivals

Provisionally, we assume *no local exogenous arrivals.* But now, "*premature departures*" exist, as illustrated in Figure 2. At the final stage, a customer, served just after another customer from the same incoming *single server* tandem queue (see above), perceives only a final stage *single server* due to the property (22) of the local busy period. On the contrary, he perceives a final stage multiserver if he is disturbed by upstream "premature departures". We introduce a slight modification in the notation of Section 2.4, at the *upstream stage* (for the total number of servers):

- $\boldsymbol{a}$: *total* load of the network;
- $\boldsymbol{a'}$: *part* of this total load corresponding to the customers *offered to* the considered final stage *multiserver*;
- $(\boldsymbol{a'/L})$: *part* of this total load handled by the considered equivalent *single server* tandem queue.

In the equivalent, symmetrical case, the number $\boldsymbol{n}$ of incoming single server tandem queues is, instead of (11):

$$\boldsymbol{n} = \text{integer part of} \left[\frac{a}{(a'/L)}\right]. \tag{23}$$

Relation (9) has to be applied with this new value of $\boldsymbol{n}$ and, instead of (12), the *local queueing delay* $\boldsymbol{w}$ for an *arbitrary* customer of *any partial traffic stream*, at the final stage multiserver in the *stationary regime*, is defined by:

$$\boxed{Ee^{zw} = \left(\frac{a'}{a.L}\right)\cdot Ee^{z.D(1)} + \left[1 - \left(\frac{a'}{a\cdot L}\right)\right]\cdot Ee^{z\cdot W_0}}, \tag{24}$$

where $D(1)$ is the local queueing delay at the final stage of the equivalent *two-stage* single server *tandem queue* ($m = 1$, without any jitter effect), and W_0 is the local queueing delay of the isolated G/G/L *server*. But now, the interpolation coefficient is L times smaller than in the case $L = 1$. In practice, with more than three incoming paths of multiservers (and $L > 2$), we have approximately:

$$Ee^{z\cdot w} \cong Ee^{z\cdot W_0}. \tag{25}$$

In that case, the result is consistent with the concept of a local traffic source. We may mention that the explicit distribution of W_0 has been recently given in Le Gall [3].

Note: In the case of constant service times, $D(1) = 0$ in (24), but the jitter effect appears again.

3.3 Case of Local Exogenous Arrivals

Now, we suppose the existence of local *exogenous* arrivals at the final stage multiserver. The notation of Section 2.6 is still valid for each server of the considered final stage multiserver: case of the loads ρ_0 and ρ_1 per server. The concept of equivalent single server tandem queue allows us to substitute (14) and (15) in expression (24) to define the local queueing delay for an endogenous arrival and for an exogenous arrival, respectively, provided that $D(1)$ and W_0 include the exogenous arrivals.

4. Case of Multilink Packet Switched Networks

We will apply the preceding considerations to the case of multilink packet switched networks with *Poisson arrivals.*

4.1 The Traffic Model ($L \geq 1$)

We consider the symmetrical network with n branches of m successive multilinks (= multiservers), of capacity L, and a final stage multilink (of capacity L) as illustrated in Figure 2. The system load ρ is the same in each link of the multilinks in successive stages. The *arrival rate* (in each link) for each individual traffic stream A_i and B_i is λ. The transmission speed is the same at each stage, i.e., the successive packet lengths (= service times) for the same customer are identical: $T_n^1 = T_n^2 = \ldots = T_n^m = T_n$, with $T = E(T_n)$.

Each individual traffic stream (in each link) is the mixture of two partial traffic streams of category j ($j = 1, 2$), corresponding to *packets of constant length* T_j ($T_1 < T_2$), λ_j being the arrival rate. We let:

$$\rho_j = \lambda_j \cdot T_j,\ \lambda = \lambda_1 + \lambda_2,\ \rho = \rho_1 + \rho_2,\ T = \rho/\lambda. \tag{26}$$

In expression (24) we have, from formula (37) in Le Gall [2]:

$$\frac{a'}{a\cdot L} = \frac{1}{nL},\ \overline{D(1)} = \frac{\lambda_1}{\lambda}\cdot(T_2 - T_1)\cdot\left[1 - \frac{1-\rho}{1-\rho_1}\cdot\frac{1-e^{-K}}{K}\right],$$

with

$$K = \frac{\lambda_2 \cdot (T_2 - T_1)}{1 - \rho_1}. \tag{27}$$

4.2 Traffic Simulations ($L = 2$)

To detect discrepancies between formulas (24) and (25), some traffic simulations were carried out [4] for the network of Figure 2, with $n = 3$, 10, $m = 10$, $L = 2$, $T_1 = 1$ and $T_2 = 10$. The number of packets ran for each example was 10^8, giving excellent accuracy.

For $\bar{w}$ (case of an *arbitrary* customer), Tables 1a and 1b give comparative results between simulations and calculations [formulae (24) and (27)] for $\rho = 0.9$ (Table 1a) and $\rho = 0.6$ (Table 1b) with two cases: $\rho_2 = \rho_1$ and $\rho_2 = 0.3\rho_1$. For $n = 10$ there does not appear to be any significant discrepancy between $\bar{w}$ and $\overline{W_0}$. For $n = 3$ and $\rho = 0.6$, on the contrary, $\bar{w}$ is 30% (for $\rho_2 = \rho_1$) and 40% (for $\rho_2 = 0.3\rho_1$) higher than $\overline{W_0}$ (M/G/2 queue). But even in this case, there appears to be a good agreement between simulations and calculations [formula (24)].

Table 1a: Case $L = 2$

	$\rho = 0.9$			
	$\rho_2 = \rho_1$		$\rho_2 = 0.3\rho_1$	
	$\bar{w}$		$\bar{w}$	
n	C	S	C	S
3	10.8	10.6	6.5	7.0
10	11.4	11.5	6.5	6.7
$\overline{W_0}$	11.6		6.5	

Table 1b: Case $L = 2$

	$\rho = 0.6$			
	$\rho_2 = \rho_1$		$\rho_2 = 0.3\rho_1$	
	$\bar{w}$		$\bar{w}$	
n	C	S	C	S
3	1.9	2.0	1.1	1.1
10	1.6	1.6	0.9	0.9
$\overline{W_0}$	1.5		0.8	

***n** identical tandem queues of* **10** *successive multilinks* ($L = 2$) *converging on one final stage multilink* ($L = 2$), *among* 10 [*see Figures* 1 *or* 2].

1) *Traffic mix* (in each queue): 2 populations of packet lengths (= service times).

	short packets	long packets	Total
packet length:	$T_1 = 1$	$T_2 = 10$	
arrival rate (per link):	λ_1	λ_2	$\lambda = \lambda_1 + \lambda_2$
load (= traffic intensity per link)):	$\rho_1 = \lambda_1 \cdot T_1$	$\rho_2 = \lambda_2 \cdot T_2$	$\rho = \rho_1 + \rho_2$

2) *Mean local queueing delay* (for an *arbitrary* packet in the final stage queue): $\bar{w}$.
" for the isolated M/G/2 queue: $\overline{W_0}$.

3) *Columns*: C = *calculated* value; S = *simulated* value.

4.3 The Occupancy in the Buffers

a) *Case of the multilink networks* ($L > 1$)

We use the traffic model as described in Subsection 4.1 for the local queue considered, but we drop the hypothesis: $(a/a') = n$. Expressions (24) and (27) give for the *mean local sojourn time* of an *arbitrary* packet:

$$\left\{\begin{aligned} \bar{s} &= \frac{a'}{a}\frac{1}{L}\overline{D(1)} + \left(1 - \frac{a'}{a}\frac{1}{L}\right)\overline{W_0} + T, \\ &\text{with: } T = \frac{\lambda_1}{\lambda}T_1 + \frac{\lambda_2}{\lambda}T_2. \end{aligned}\right. \tag{28}$$

In expression (27) of $\overline{D(1)}$ we may use the approximation: $\frac{1-e^{-K}}{K} \cong 1$. We rewrite:

$$\left\{\begin{aligned} \overline{D(1)} &\cong \frac{\lambda_1}{\lambda} \cdot \frac{\rho_2}{1-\rho_1} \cdot (T_2 - T_1), \\ \bar{s} &= \frac{a'}{a}\frac{1}{L}[(1-H)T_2 + HT_1] + \left(1 - \frac{a'}{a}\frac{1}{L}\right)(\overline{W_0} + T), \\ &\text{with: } H = \frac{\lambda_1}{\lambda}\frac{1-\rho}{1-\rho_1}. \end{aligned}\right. \tag{29}$$

$\bar{s}$ gives the *occupancy in the buffer* per *arbitrary* packet, when the *environment* is at *busy hour*. When the *environment* is *at off-peak hour* (i.e. very small traffic intensity), we have: $a' \cong a$ and expression (29) becomes closer to the tandem queue model:

$$\bar{s} \cong \frac{1}{L}[(1-H)T_2 + HT_1] + \left(1 - \frac{1}{L}\right)(\overline{W_0} + T). \tag{30}$$

Practically, for dimensioning purpose, we have to consider this "*worst case*". At the same time for the local queue considered *in the network*, the *occupancy* (per *arbitrary* packet) is the smaller value T, given by expression (28).

b) *Case of single link networks* ($L = 1$)

In the case of *single link* queueing networks ($L = 1$), we deduce that the *buffers have to be overdimensioned, on using the "equivalent tandem queue" model* and not on using the traditional G/G/1 model. Due to expression (29) of $\overline{D(1)}$, *a proportion*

$[\rho_2/(1-\rho_1)]$ *of short packets has the same high occupancy in the buffer as long packets*, when the environment is not at busy hour. It is due to the *agglutination phenomenon* of short packets behind long packets as already mentioned after relations (8). Practically, the "*worst case*" corresponds to *off-peak hours* in the network, except for some traffic streams where each incoming path corresponds to only one outgoing path, this path handling the same traffic stream and only this traffic stream. For this "*worst case*", *the capacity of buffers has to be multiplied by* 3 (*and even by* 4) compared with the use of the traditional G/G/1 model, when $T_2 \geq 30 \cdot T_1$.

For an accurate calculation we have to consider the *tandem queue modeling* $(n = 1)$. Due to expression (12), we use the *mean local queueing delay* $\overline{\boldsymbol{D(m)}}$ instead of $\overline{\boldsymbol{D(1)}}$, because *the phenomenon of agglutination is amplifying with the number of stages already crossed.* Expression (27) above has to be replaced by formulas (37) and (38) in Le Gall [2]:

$$\left\{\begin{array}{r}\overline{\boldsymbol{D(m)}} = \overline{\boldsymbol{V(2,m+1)}} - \overline{\boldsymbol{V(2,m)}},\\ \text{with: } \overline{\boldsymbol{V(2,m+1)}} = \dfrac{\lambda_1}{\lambda} \cdot m \cdot (T_2 - T_1)\left[1 - \dfrac{1-\rho}{1-\rho_1} \cdot \dfrac{1-e^{-K(m)}}{K(m)}\right],\\ \text{and: } K(m) = m \cdot \dfrac{\lambda_2}{1-\rho_1} \cdot (T_2 - T_1).\end{array}\right. \tag{31}$$

For $n = 1$, the jitter delay does not exist: $\overline{J} = 0$. With the traditional M/G/1 server model, the buffer is dimensioned for an occupancy equal to $(\overline{W_0} + T)$. With the tandem queue model the occupancy becomes $\overline{D(m)} + T$. It follows an *overload coefficient* (of the buffer) equal to:

$$C = \frac{\overline{D(m)} + T}{\overline{W_0} + T}. \tag{32}$$

As an example, consider the case of "*intelligent networks*" where signaling packets ($\cong 50$ bits $\rightarrow T_1 = 1$) of the ITU-T CCS N°7 interferes with intelligent packets ($\cong 1500$ bits$\rightarrow T_2 = 30$). For $m = 5$, $\rho_2 = 0.25 \cdot \rho_1$, $\rho = \rho_1 + \rho_2 = 0.6$, expressions (31) and (32) give:

$$\overline{W_0} = 5.1, \quad \overline{D(m)} = 20.6, \quad C = 3.5. \tag{33}$$

As a consequence, it is *absolutely necessary to dimension the buffer on using the tandem queue model*, because the buffer may be congested before any detection by the time-outs in the network: in our example $\rho = 0.6$ only! And during all this time congestion of the buffer, all the links under the control of this buffer are not accessible. For example, as we mentioned above, it is the case of the occurrence of some occasional overloads in some traffic streams at off-peak hours in the network. Finally, the phenomenon is significant because *the busy hour in the buffers is quite different of the busy hour in the network.*

5. Conclusion

Based on the method in [2], we used the tandem queue model. We exploited the concept of *apparent* upstream queueing delay as perceived by the downstream network, to be able to introduce simply an interpolation linking the distribution of the local queueing delay of multiserver queueing networks to those of the tandem queue and the isolated G/G/s queue. This result justifies the use of the usual G/G/s approximation for networks which are sufficiently meshed. This approximation *suggests that the busy periods are not broken up*: a usual case for packet-switched networks and for heavily loaded networks when the service times are not too much varying.

While this result has an apparent simplicity, it in fact illustrates the potential danger in designing and managing large queueing networks in spite of concerns of the uselessness of standardization of the network and its management. Multiserver networks do not benefit much from the tandem queue effect which, in single server queueing networks, tends to smooth the traffic offered and make a network more resistant to overloads. To manage such networks at a load high enough for economic and "*interactive*" efficiency, it is necessary to introduce out-of-band signaling, using ITU signaling system N°7 for example, in order to limit response times. In other words, for *interactive networks*, requiring low response times at high loads, it is necessary to control the multiserver queueing network by means of a *single server network for signaling*, using only short packets to avoid some significant overdimensioning in the buffers, which *needs to be dimensioned with the "equivalent tandem queue" model.*

References

[1] Borovkov, A.A., *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York 1976.

[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., The stationary G/G/s queue with non-identical servers, *J. of Appl. Math and Stoch. Anal.* **11**:2 (1998).

[4] Romoeuf, L., Traffic simulations in the control plane, Note *CNET/DT/LAA/EIA/EVP/95-08LR*, May 22, 1995.