DELAY ANALYSIS IN A PACKET SWITCHED NETWORK BY METHODS
BASED ON A QUEUEING NETWORK MODEL WITH BURSTY TRAFFIC

Karl LINDBERGER and Sven-Erik TIDBLOM

Swedish Telecommunications Administration
Network Department
S-123 86 Farsta Sweden

A queueing network model with batches arriving to the primary queues as Poisson processes, with the batch sizes geometrically distributed and with exponentially distributed service times is considered. Departure streams can be superposed with other streams and offered to later queues in the network. Results from this special model are used to select key parameters of the streams and construct simple approximative expressions for important delay quantities in the network for a method to be used more generally, e.g. in packet switching applications.

1. INTRODUCTION

In a packet switched network data is always delayed when transferred from DTE-A to DTE-B and vice versa. The end-to-end delay of a data packet is influenced by a lot of factors. The main factors with impact on this delay are: the number of nodes on the route chosen for the virtual call and the traffic situation along this route. These main factors are always present no matter how the network is designed or who is the manufacturer of the equipment. However, when one tries to make a detailed analysis of delay one naturally gets focused on the particular national solution chosen for packet switching. E.g. when data is transferred through a node it has to pass through a number of service stages and these stages are always dependent on the particular equipment chosen. For our national X.25 network we have constructed a formula for the mean transit delay of packets passing through a node and out on a link. According to this formula the mean transit delay consists of three parts: one constant part, which is always present even if the traffic is zero and one part which depends on the total load on the node and finally a part which has to do with the load on the considered outgoing link. Thus, with given loads on the links and on the nodes constituting a route for a call we have tools to forecast the expected end-to-end delay along that route. However, since these formulas are closely tied to our national X.25 network and also because of the limited space allowed for the paper we have decided to omit the detailed presentation of these results.

Now, our experience of packet switching is on one hand based on the rather careful study of the particular system just mentioned and on the other hand on a mathematical treatment of a more general queueing network model, with results which could be of greater interest for a broader audience. We find it important to have a simple tool to study how certain qualities such as burstyness of the traffic streams in this queueing network may influence on the delay of a packet through the network. The model may be applied to a situation where the rather high load on the transmission links is the reason for the major part of the total delay.

Even though most of our experiences of packet switching comes from a special type of low bit rate data network an essential reason for trying to find these rather general methods to handle bursty traffic streams in a queueing network is to be able to discuss traffic problems in future techniques such as ATD (ATM).
Before we know the final details of such a system we must be able to answer questions e.g. on the behaviour in the network when a typical traffic stream consists of small packets from a call of a service with very high bandwidth.

The method is based on a model with batch arrivals in the primary queues as an original reason for burstyness. Thus in the applications the most essential difference from an M/M/1-queueing network should be in the arrival processes. In the paper we will find that a departure stream from one queue which is possibly superposed with other streams and offered to a next queue should be represented by three traffic parameters. A goal has been that the formula for a key quantity in the method such as the mean number of customers in a node should be a simple and very explaining expression in those three parameters for each of the offered streams.

Our type of network and the purpose of the method is of course related to the QNA in [1] though not quite as general. Departure processes from queueing systems discussed e.g. in [3] are not generally of renewal type. However, the superposition of streams including departure ones offered to a later queue is in many methods (see [2]) replaced by something simpler e.g. a renewal process with certain characteristics. In our case it is rather described by its effect on a key quantity of the queue to which it is offered. We have not yet found any other paper which has solved exactly our problem but of course it is related to several papers on tandem queues. In most papers it is the service time distribution that has been the generalization of the M/M/1-case and particularly a deterministic one, e.g. as in [4], could be interesting also in a correspondence to our case. Normally a great number of papers should have been referred to, but the comprehensive reference lists of [1]-[4] cover a lot.

2. THE MODEL AND THE RESULTS

We are here going to suggest an approximation method for a certain type of queueing network with single server nodes, bursty traffic streams and service time distributions with not too much variation. The purpose of the method is, with use of simple formulae, to recognize which types of mixtures of traffic streams, including departure ones, that could cause too long (mean) delays somewhere in the network. We could also express it in the following way. If a bursty traffic stream is offered to a queue and is there "extra" delayed due to this burstyness, then "how much" of that burstyness is still left in the departure process?

The approximation will be based on results from a more specific type of problem. Three versions of this problem in limited networks will here be discussed in detail and then from those cases some general conclusions for the corresponding cases with arbitrary network size can be drawn. The first version, I, is the following one.

Consider two queues in tandem. Arrivals to the primary one take place in batches at time points described by a Poisson process with intensity $\lambda$ and the probability of a batch size $k$ is $f_k = pq^{k-1}$, where $q = 1-p$. The two queues are FCFS ones with independent exponentially distributed service times with service intensities $1$ and $\mu$ respectively.

Departures from the primary queue are with probability $\gamma$ offered to the secondary queue to which is also offered an independent Poissonian arrival stream with intensity $\alpha$. It is the behaviour of this secondary queue with arrivals which are the superposition of two streams, where one is of departure type, that we are interested in at first. By putting $\lambda' = p\rho_1$ and $\mu\rho_2 = \alpha + \gamma\rho_1$ we can note that $\rho_1$ and $\rho_2$ are service factors in the two queues.
Let $X_1$ and $X_2$ be the number of customers at each node respectively at an arbitrary time point when the system can be regarded as stationary.

With

$$P(j,k) = P(X_1=j, X_2=k)$$

we can write down the state equations of the system

$$(\lambda+1+\alpha+\mu)P(j,k) = \sum_{i=1}^{j} P(j-i,k)\lambda f_i + P(j+1,k-1)\gamma + \alpha P(j,k-1) + \mu P(j,k+1) + (1-\gamma)P(j+1,k) \quad j,k \geq 1 \tag{1}$$

$$\hspace{2cm} (\lambda+1+\alpha)P(j,0) = \sum_{i=1}^{j} P(j-i,0)\lambda f_i + \mu P(j,1) + (1-\gamma)P(j+1,0)$$

$$\hspace{2cm} (\lambda+\alpha+\mu)P(0,k) = \gamma P(1,k-1) + \alpha P(0,k-1) + \mu P(0,k+1) + (1-\gamma)P(1,k)$$

$$\hspace{2cm} (\lambda+\alpha)P(0,0) = \mu P(0,1) + (1-\gamma)P(1,0)$$

$$\sum_{j=0}^{\infty} \sum_{k=0}^{\infty} P(j,k) = 1$$

By letting the generating function of $(X_1,X_2)$ be defined by

$$G(s,u) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} s^j u^k P(j,k)$$

and

$$F(s) = \sum_{k=1}^{\infty} s^k f_k = ps/(1-qs)$$

where $0 \leq s,u \leq 1$, we obtain from (1) the relation

$$G(s,u)\left\{u(1-s)(1 - \frac{\rho_1 ps}{1-qs}) + (1-u)(\mu s - \gamma u - us\alpha)\right\} =$$

$$= G(0,u)u(1-s-\gamma(1-u)) + G(s,0)\mu s(1-u) \tag{2}$$

This equation should contain all information about the state probabilities but a simple explicit solution seems to be difficult to find.

However, we have that $P(X_2=0)=1-\rho_2$ and the generating function for $X_1$ is

$$G(s,1) = (1-\rho_1)(1-qs)/(1-qs-\rho_1 ps) \tag{3}$$

With $a=EX_2$, $b=E(X_1|X_2=0)$ and $c=E(X_2|X_1=0)$ we also have the two simple relations in $a$, $b$ and $c$,

$$a(\mu+1-\gamma-\alpha-\rho_1) = \mu(1-\rho_2)b + (1-\gamma)(1-\rho_1)c + \alpha+\rho_1/p-\rho_1(\mu+1-\gamma-\alpha-\rho)/p(1-\rho_1) \tag{4}$$

and

$$a(\mu-\alpha-\gamma-\alpha) = -(1-\rho_1)\gamma(1+c) \tag{5}$$
Further derivations only imply either identities or relations with at least one new moment of higher order. Thus one simple relation seems to be missing to determine $a$ exactly, but with some inequalities and certain qualities of $G(s,u)$ in addition to (4) and (5) we can draw the following conclusions. We can write

$$a = EX_2 = T_1 + T_2 + T_3 = \rho_2/(1-\rho_2) + T_2 + T_3$$

i.e. $T_1$ is the term we would have got by itself in the M/M/1-case. The most interesting term is

$$T_3 = \frac{q \gamma \rho_1}{\rho \alpha + \gamma - \mu}(\alpha + \gamma - \mu)^2 \mu(1-\rho_2)$$

since it includes the factor $q/p$, but also in some cases vanishes. The term $T_2$ is a little problem since we at this point can not give an exact expression for it. An approximation (or an exact guess) is however the following expression where the $q$-factor is the uncertain one.

$$T_2 = \frac{q \gamma \rho_1}{(1-\rho_1)}(\alpha + \gamma - \mu) \geq 0$$

$$T_2 = \frac{q \rho_1(1-\rho_1)\gamma^2}{\mu^2(1-\rho_2)^2} \alpha + \gamma - \mu < 0$$

Simulations of some special cases also show that the factor is in the right magnitude and thus the $T_2$-term is not dominating. Compare also with

$$EX_1 = \rho_1/p(1-\rho_1) = (1+q/p)\rho_1/(1-\rho_1)$$

We shall now study the first extension of the system, II, where the Poissonian stream with intensity $\alpha$ offered directly to the secondary queue in I is replaced by a departure process from the same type of queue as the primary one in I i.e. we have here in II two parallel primary queues with parameters $(\rho_i, p_i) i = 1,2$. For simplicity all service intensities are 1 here and departures from each queue are offered to the secondary one with probabilities $\gamma_1$ and $\gamma_2$ respectively i.e. the service factor there is $\rho = \gamma_1 \rho_1 + \gamma_2 \rho_2$

The corresponding equation to (2) is here

$$G(s,t,u)\{ut(1-s)(1 - \frac{\rho_1 p_1 s}{1-q_1 s}) + us(1-t)(1 - \frac{\rho_2 p_2 t}{1-q_2 t}) +$$

$$+ (1-u)(st-\gamma_1 ut-\gamma_2 us)\} = G(0,t,u)(1-s-\gamma_1(1-u))ut +$$

$$+ G(s,0,u)(1-t-\gamma_2(1-u))us + G(s,t,0)(1-u)st$$

where $G(s,t,u)$ is the generating function of $(X_1,X_2,X)$ and $X$ is the number of customers in the secondary node. Also in this case we have some problems to find the exact expression of $EX$ but we could write

$$EX = T_1 + T_{21} + T_{22} + T_{31} + T_{32}$$

where $T_1$ is $\rho/(1-\rho)$,

$$T_{3i} = \frac{\gamma_i q_i \rho_i}{p_i(1-\rho_i)}(\rho_1 - \gamma_i \rho_i - 1 + \gamma_i)^+(1-\rho)$$

and $T_{2i}$ could be approximated by $\gamma_i q_i \rho_i/(1-\rho_i)$. How the extension to more than 2 primary groups will turn out is obvious and also the generalization to different service intensities $(\mu_i)$ can be done. Compare with I.
The special conclusion of this case II is that each individual primary queue with batch arrivals can cause its own $T_3$-type of term in $EX$ of each later node that a part of its stream will pass through. The term including the factors $q_i/p_i$ depends on the other streams offered to that node only through their contribution to the service factor of the node. A similar conclusion seems to be possible for the $T_2$-type of terms and anyway they are not dominating. Thus it is no particular loss of generality in the next case to let all parallel streams offered to non-primary queues to be Poissonian.

The next extension of I, called III, is the case with three nodes in series. We have the same conditions as in I but with $(\gamma, \alpha, \mu)$ replaced by $(\gamma_1, \alpha_2, \mu_2)$ and in addition we have a third node to which customers from the secondary node are offered with probability $\gamma_2$ together with those from a Poissonian stream with intensity $\alpha_3$. The service intensity in the third node is $\mu_3$. With $G(s,u,v)$ as the generating function of $(X_1, X_2, X_3)$ we can obtain the correspondence to (2).

\[ G(s,u,v) = uv(1-s)(1 - \frac{p_1s}{1-q_2s} ) + v(1-u)(\mu_2s - \gamma_1u - \alpha_2su) + \\
+ s(1-v)(\mu_3u - \gamma_2mu_2v - \alpha_3uv) = G(0,u,v)uv(1-s - \gamma_1(1-u)) + \\
+ G(s,0,v)s\mu_2(1-u - \gamma_2(1-v)) + G(s,u,0)\mu_3(1-v) \]  
(12)

Also in this case derivation leads to interesting conclusions on $EX_3$ but not information enough to determine $EX_3$ exactly. However $T_{13}$, the $T_1$-term of $EX_3$ is as usual

\[ \rho_3/(1-\rho_3), \text{ where } \mu_3\rho_3 = \alpha_3 + \gamma_2\alpha_2 + \gamma_1\gamma_2\rho_1 \]

With the substitutions

\[ B_1 = 1-\rho_1, \gamma_1B_2 = \mu_2(1-\rho_2), \gamma_1\gamma_2B_3 = \mu_3(1-\rho_3) \]  
(13)

$T_{33}$ and also $T_{32}$ i.e. the $T_3$-terms of $EX_3$ and $EX_2$ can be written in a simple form

\[ T_{33} = \frac{\rho_3}{p} \gamma_1\gamma_2\rho_1 \left( \frac{1}{B_3} - \frac{1}{B_2} \right), \quad B_1 \geq B_2 \geq B_3 \]  
(14)

\[ T_{33} = \frac{\rho_3}{p} \gamma_1\gamma_2\rho_1 \left( \frac{1}{B_3} - \frac{1}{B_1} \right), \quad B_2 \geq B_1 \geq B_3 \]  
(15)

\[ T_{32} = \frac{\rho_3}{p} \gamma_1\rho_1 \left( \frac{1}{B_2} - \frac{1}{B_1} \right), \quad B_1 \geq B_2 \]  
(16)

and the terms are equal to zero otherwise. We also note that $T_{31}$, the $T_3$-term of $EX_1$ is $\rho_1q/pB_1$. Also the $T_2$-term, $T_{23}$, is depending on the order of the $B_i$:s but is not greater than for the case $B_1 \geq B_2 \geq B_3$ where it in a similar way as in (8) can be approximated by $q\gamma_1\gamma_2\rho_1/B_1$ and again the $q$-factor is a bit uncertain.

One more result that can be useful is

\[ E(X_3|X_2=0) - \frac{\rho_3}{1-\rho_3} = \frac{\mu_3-\alpha_3}{\mu_3\rho_3-\alpha_3} \left( E(X_3) - \frac{\rho_3}{1-\rho_3} \right) \]  
(17)

and its correspondence for $E(X_2|X_1=0)$.
3. THE INTERPRETATION OF THE RESULTS

We are now going to discuss whether some of the results above can be used to describe some typical effects in a packet switched queueing network.

The parameters $\gamma_i$ are meant as a possibility to split the streams, but it is a random splitting they represent. In the packet switching application a more realistic assumption is perhaps that packets from the same message are going the same route through the network and thus that a split stream should be regarded as two (or more) individual streams with the same $p$-value but with the original $\rho_1$-value divided into parts proportional to how many customers are going the different ways. This means that in the formulae for $B_i$ all $\gamma_i$ should be replaced by 1 which makes quite a difference. With this idea defining very small individual streams we can also let them pass through an individual primary queue with service intensity 1 before they are superposed with other streams. This means that the $B_1^{-1}$-factor will be very close to 1 and the total sum of all $T_2$-terms in a node will be $\approx q\mu\rho$. Anyway the $T_2$-terms are not particularly interesting compared to the other terms. The approximation sum is less than 1. Let us now interpret the general $T_3$-terms in a node by looking at $T_{33}$ and $T_{32}$.

To a certain node somewhere in the network a number of small individual streams are offered on their way through the network. Each stream $i$, should be described by three parameters at this point $(\rho_i, p_i, C_i)$, where $\rho_i$ and $\rho_1$ are the original parameters of load (when $\mu = 1$) and burstyness of that particular stream of customers going the same way through the network. (Partial sharing of the route is regarded as different streams). The parameter $C_i$ is updated as the stream is going through the network. If the node in question is of order $k(i)$ for stream $i$ the $C_i$-value should here be

$$1/C_i = \min_{j \leq k(i)-1} B_j; \quad B_j = \mu_j(1-\hat{\rho}_j)$$

where $\mu_j$ and $\hat{\rho}_j$ are the service intensity and the total service factor on a node that stream $i$ has been passing through.

Now if our special node has the total service factor $\hat{\rho}$ and the service intensity $\mu$ we have

$$\mu\hat{\rho} = \sum \rho_i, \quad B = \mu(1-\hat{\rho}) , \quad C = 1/B$$

and the mean number of customers in the node is approximately

$$EX = \mu\hat{\rho}(C+q) + \sum \rho_i(q_i/p_i)(C-C_i)^+$$

(18)

corresponding to the $T_1$-, $T_2$- and $T_3$-type of terms. The departure stream $i$ has the three parameters

$$\rho_i, p_i, \max (C, C_i)$$

This must be regarded as a very simple method for such a complicated problem and it is even simpler if we as in packet switching networks may have reason to assume that all service intensities $\mu_i$ are equal (to unity). Then the third parameter is a function of the maximum $\hat{\rho}$-value that stream $i$ has experienced so far. Possibly one could simplify the method even more by using a standard $p$-value for all streams i.e. a general mean message length.
Now in the case of all $\mu_i=1$, (18) gives the mean virtual waiting time. By using a correspondence to (17) on the $T_3$-terms of the streams coming from the same previous node as stream $i$ and leave the other $T_3$-terms as they are we obtain an approximation also for the individual mean waiting times at each node.

Having established the key parameters of a stream $(\rho_i, p_i, C_i)$ in the network we can use the equivalence in a smaller system with only primary and secondary queues to find expressions for other interesting quantities such as variances, covariances and percentiles of waiting time distributions. Thus the decomposition is possible and such a smaller system is easier to study in detail.

4. CONCLUSIONS

Future work will be further studying of the equation (2) trying to derive an explicit expression for $G(s,u)$ (or possibly find a missing reference offering it). The most interesting modification of the model would be to change the exponential service time distributions to a deterministic one and study the corresponding problem in discrete time with the constant packet length as a time unit. Possibly some simple expressions could be found also in that case showing similar effects in the network as in our model.

Anyway we have given rather simple expressions showing that an originally bursty traffic stream, after departure from a primary node, could cause almost as little effect on the queue length of a secondary, less loaded node, as if it were Poissonian, but then in a more loaded third queue cause essentially longer queues than in a corresponding M/M/1-case with the same service factor. Such a quality of a departure process which we think has some generality, is difficult to express with only two parameters.

REFERENCES


