Insensitivity of a network of symmetric queues with balanced service rates

J. Virtamo

June 12, 2003

Abstract

Previous insensitivity results on queueing systems include on one hand, e.g., symmetric queues in isolation and, on the other hand, networks of processor sharing queues with globally dependent balanced service rates, i.e. so-called Whittle networks. In this paper we show that insensitivity property of a queueing network with balanced service rates is retained when each of the processor sharing nodes is replaced by any kind of symmetric queue.

1 Introduction

Kelly [1] has defined the concept of a symmetric queue and proven that symmetric queues are insensitive, i.e. the state distribution of the system and all the performance metrics depend on the service requirement distribution only through the mean. Symmetric queues, though a very specific class of queues, constitute still a large family of queueing systems and include, e.g. processor sharing (PS) and last in first out (LIFO) scheduling disciplines. Kelly's proof for the insensitivity made use of a representation of the general service requirement distribution in terms exponential phases with mean of a fixed size and of the construction of the reverse process to verify the global balance conditions.

A Whittle network is a queueing network consisting of PS nodes with service rates that depend on the global state of the network in such a way that so-called balance conditions are satisfied. An exposition of Whittle networks is provided in the book by Serfozo [2]. An important property of these networks is that the state distribution and performance metrics are insensitive and depend on the traffic characteristics only through the loads of different nodes. Again, the insensitivity can be proven by representing a general distribution as a phase type distribution and by noting that the enlarged network, where each node is divided into a set of subnodes, constitutes a Whittle network.

Whittle networks have recently been further explored by Bonald and Proutière [3]. They showed that also the converse is true: any queueing network with PS nodes possessing the insensitivity property necessarily satisfies the balance condition, i.e. Whittle networks are the only insensitive PS networks. They also used Whittle networks as a tool for modelling resource sharing between flows in a network like the Internet. By imposing the additional

requirement that in each state of the network at least one link of the network is saturated they arrived at the notion of balanced fairness [4]. Balanced fairness defines the most efficient resource sharing between the flows in the network such that the performance metrics are insensitive and depend only on the loads of different flow classes.

In this paper we combine the concepts of symmetric queues and Whittle networks and consider queueing networks where the nodes have balanced global state dependent service rates and the service discipline within each node is any symmetric discipline in the sense of Kelly's definition. We prove that insensitivity holds also for these generalized systems. Further, we note that the proof of Bonald and Proutière for the converse relation covers this more general case too: insensitivity of a network of symmetric queues implies that the service rates are balanced.

Our insensitivity proof is more direct in that it deals with a general, continuous service requirement distribution from the outset without relying on a representation as a phase type distribution. The proof is also direct in the sense that it is based on simply showing that the global balance equations are satisfied by a trial distribution; no construction of the reverse process is needed.

2 Balanced system with symmetric queues

The system consists of a network of K queues indexed by k. Customers are allowed to move from one queue to another; a customer leaving queue l moves next to queue k with the probability $p_{l,k}$. Two special nodes are defined: the source s and the destination d. The total exogenous arrival rate, i.e. arrivals from node s is λ; the rate of exogenous arrivals to node k is $\lambda\, p_{s,k}$. The effective customer rates λ_k through different queues satisfy the flow equations,

$$\lambda_k = \lambda\, p_{s,k} + \sum_{l=1}^{K} \lambda_l p_{l,k}, \quad \forall k. \tag{1}$$

We assume that from each node there is a path leading to destination d with a positive probability, i.e. there are no absorbing groups of states. Then the total rate of departures from the network equals the total arrival rate,

$$\lambda = \sum_{k=1}^{K} \lambda_k p_{k,d}. \tag{2}$$

The number of customers in queue k is denoted X_k, and the occupancy state of the whole system is specified by the vector $X = (X_1, \ldots, X_K)$. A point in the state space is denoted by x. For brevity, we use the notation $|x|$ for the total occupancy, $|x| = \sum_k x_k$.

The service requirement S_k of a customer in queue k is a random variable independent of the service requirements of other customers and of the arrival process. The distribution is assumed general with mean s_k. We define $G(s) = \mathrm{P}\{S_k \geq s\}$ so that $G(0) = 1$. The following notations are adopted: $f_k(s) = -G'(s)$ (delta function components are allowed in $f_k(s)$ to account for atomic probabilities), $\rho_k = \lambda_k s_k$ is the load of queue k, and $\rho = (\rho_1, \ldots, \rho_K)$.

Queue k is served at rate $\phi_k(x)$, which in general depends on the global state of the system. We restrict this dependency by considering only systems where the service capacities can be derived from a balance function $\Phi(x)$ as follows,

$$\phi_k(x) = \frac{\Phi(x - e_k)}{\Phi(x)} = \frac{\Phi(T_k x)}{\Phi(x)}, \quad \forall x \text{ s.t. } x_k > 0, \tag{3}$$

where e_k is a K-vector with 1 in component k and 0 elsewhere, and where we have defined the "lowering operator" T_k such that $T_k x = x - e_k$. Similarly, for later use, we define the "raising operator" T^k such that $T^k x = x + e_k$. When (3) holds, the capacities are said to be balanced by $\Phi(x)$. This definition immediately implies that the following balance property is satisfied,

$$\frac{\phi_i(T_j x)}{\phi_i(x)} = \frac{\phi_j(T_i x)}{\phi_j(x)}, \quad \forall i, j \text{ s.t. } x_i > 0,\ x_j > 0. \tag{4}$$

Conversely, it is easy to show that if this balance property is satisfied, then there exists a balance function $\Phi(x)$ such that (3) is true. Note that any positive function $\Phi(x)$ defines a balanced system.

Each of the K queues is assumed symmetric. The queueing disciplines may be different in different queues in so far as each of them can be described as a symmetric queue. Customers in each queue are ordered, with queue k containing customers in positions $i = 1, \ldots, x_k$. The crucial property that characterizes a symmetric queue is the following (cf. [1]):

- A proportion $\gamma_k(i, x_k)$ of the total service capacity $\phi_k(x)$ of queue k is directed to the customer in position i ($i = 1, \ldots, x_k$). When the customer in position i of queue k departs from the system, customers previously in positions $i+1, \ldots, x_k$ of that queue move to positions $i, \ldots, x_k - 1$ respectively.

- A customer arriving at queue k with x_k customers moves into position i ($i = 1, \ldots, x_k + 1$) of that queue with probability $\gamma_k(i, x_k + 1)$; customers previously in positions $i, \ldots, x_k$ move to positions $i + 1, \ldots, x_k + 1$ respectively.

3 Stationary distribution and insensitivity

The state of the system is described by the vector (X, Y), where $X = (X_1, \ldots, X_K)$, and $Y = (Y_1, \ldots, Y_K)$ with $Y_k = (Y_{k,1}, \ldots, Y_{k,X_k})$. $Y_{k,i}$ specifies the amount of service already obtained by a class-k customer in position i, for $i = 1, \ldots, X_k$. It is easy to see that (X, Y) contains enough information to render it a Markov process, whereas X alone is not generally a Markov process. Again, corresponding points of the state space are denoted with lower case letters, e.g. (x, y).

In order to simplify notation, we introduce further "raising" and "lowering" operators as follows: $U^{i,z}$ inserts a new customer with the amount z of service already received into position i of a given queue. For instance for queue k with x_k customers we have,

$$U^{i,z} y_k = (y_{k,1}, \ldots, y_{k,i-1}, z, y_{k,i}, \ldots, y_{k,x_k}), \quad i = 1, \ldots, x_k + 1.$$

$T^{k,i,y}$ inserts a new customer with the amount z of service already received into position i of queue k . Thus

$$T^{k,i,z}(x,y) = (T^k x, y_1, \ldots, y_{k-1}, U^{i,z} y_k, y_{k+1}, \ldots, y_K), \quad \forall k,\ i = 1, \ldots, x_k + 1.$$

Correspondingly, U_i removes a customer from position i of a given queue. For instance for queue k with x_k customers we have,

$$U_i y_k = (y_{k,1}, \ldots, y_{k,i-1}, y_{k,i+1}, \ldots, y_{k,x_k}), \quad i = 1, \ldots, x_k.$$

Finally, $T_{k,i}$ removes a customer from position i of queue k,

$$T_{k,i}(x,y) = (T_k x, y_1, \ldots, y_{k-1}, U_i y_k, y_{k+1}, \ldots, y_K), \quad \forall k,\ i = 1, \ldots, x_k.$$

We assume (X, Y) has a stationary distribution expressed as a mixed point probability and joint probability density function

$$\pi(x,y) = \frac{\partial}{\partial y} \mathrm{P}\{X = x,\, Y \leq y\},$$

where $\frac{\partial}{\partial y}$ denotes partial derivative with respect to all the components of the vector y.

Theorem. *The stationary distribution of* (X, Y) *is given by*

$$\pi(x,y) \;=\; \frac{\Phi(x)}{\theta(\rho)} \prod_{k=1}^{K} \lambda_k^{x_k} \prod_{i=1}^{x_k} G_k(y_{k,i}) \;=\; \frac{\Phi(x)}{\theta(\rho)} \prod_{k=1}^{K} \rho_k^{x_k} \prod_{i=1}^{x_k} \frac{G_k(y_{k,i})}{s_k}, \tag{5}$$

where $\theta(\rho)$ *is the normalization constant*

$$\theta(\rho) \;=\; \sum_x \Phi(x) \prod_{k=1}^{K} \rho_k^{x_k}. \tag{6}$$

Proof. One has to show that (5) satisfies the global balance equation. To this end, we consider all types of transition affecting the probability density at the state (x, y). First note that using the balance property (3) and the trial (5) one obtains the auxiliary relations

$$\pi(T^{k,i,z}(x,y)) \;=\; \frac{\lambda_k G_k(y_{k,i})}{\phi_k(T^k x)}\, \pi(x,y), \qquad i = 1, \ldots, x_k + 1, \tag{7}$$

$$\pi(T_{k,i}(x,y)) \;=\; \frac{\phi_k(x)}{\lambda_k G_k(y_{k,i})}\, \pi(x,y), \qquad x_k > 0,\ i = 1, \ldots, x_k. \tag{8}$$

Now, the density of in-flow due to divergence of the flow field $u(x)\pi(x,y)$, i.e. the sink density of the flow field, is

$$-\nabla_y \cdot (u(x)\pi(x,y)) \;=\; -u(x) \cdot \nabla_y \pi(x,y) \;=\; -\sum_{k=1}^{K} \sum_{i=1}^{x_k} \gamma_k(i, x_k) \phi_k(x)\, \frac{\partial}{\partial y_i} \pi(x,y), \tag{9}$$

where $u(x)$ is the velocity vector of a system point in the y-space due to service execution when the number of customers in different classes are specified by the vector $x = (x_1, \ldots, x_K)$. The vector $u(x)$ is doubly indexed by the queue and position indices (k, i), with $k = 1, \ldots, K$, and $i = 1, \ldots, x_k$, so that $u_{k,i}(x) = \gamma_k(i, x_k)\phi_k(x)$ defines the rate at which a queue-k customer in position i receives service. Note that the vector $u(x)$ depends on x only and can be interchanged with the ∇_y operator.

With trial (5) expression (9) becomes

$$\pi(x, y) \sum_{k=1}^{K} \sum_{i=1}^{x_k} \gamma_k(i, x_k)\phi_k(x) \frac{f_k(y_i)}{G_k(y_i)}$$

cancelling the density of out-flow due to completed service requirements which always has this expression since the hazard rate is $f_k(y_i)/G_k(y_i)$.

Density of in-flow at (x, y) due to completed service requirements of customers departing from the system at occupancy level $|x| + 1$, i.e. due to transitions $T^{k,i,z}(x, y) \to (x, y)$ for any $k = 1, \ldots, K,\ i = 1, \ldots, x_k + 1$ and $z \in (0, \infty)$, is

$$\sum_{k=1}^{K} \sum_{i=1}^{x_k+1} \int_{z=0}^{\infty} \gamma_k(i, x_k + 1)\phi_k(T^k x) \frac{f_k(z)}{G_k(z)} \pi(T^{k,i,z}(x, y))\, dz \;=$$
$$\pi(x, y) \sum_{k=1}^{K} \lambda_k\, p_{k,d} \sum_{i=1}^{x_k+1} \gamma_k(i, x_k + 1) \int_{z=0}^{\infty} f_k(z)\, dz \;=\; \sum_{k} \lambda_k\, p_{k,d}\, \pi(x, y) \;=\; \lambda\pi(x, y),$$

where in the last step equation (2) was used. This expression cancels the density of out-flow due to exogenous customer arrivals, i.e. transitions to occupancy level $|x| + 1$, which always equals $\lambda\pi(x, y)$.

Then, there is an in-flow due to exogenous customer arrivals, i.e. transitions from occupancy level $|x| - 1$. In particular, for a state (x, y) such that $y_{k,i} = 0$ the density of in-flow due to transitions $T_{k,i}(x, y) \to (x, y)$ triggered by exogenous arrivals at position i of queue k is

$$\lambda p_{s,k} \gamma_k(i, x_k) \pi(T_{k,i}(x, y)) \tag{10}$$

since the rate of exogenous arrivals at queue k is $\lambda p_{s,k}$ and an arriving customer takes position i with the probability $\gamma_k(i, x_k)$ and has so far received no service, $y_{k,i} = 0$.

Similarly, there is an in-flow at (x, y) with $y_{k,i} = 0$ due to service completions in queue l followed by a move to queue k, i.e. due to transitions $T^{l,j,z}T_{k,i}(x, y) \to (x, y)$, with $l = 1, \ldots, K$, $j = 1, \ldots, x_l + 1$, and $z \in (0, \infty)$. The density of the in-flow is

$$\begin{aligned}
&\sum_{l=1}^{K} \sum_{j=1}^{x_l+1} \int_{z=0}^{\infty} \gamma_l(j, x_l + 1)\phi_l(T^l T_k x) \frac{f_l(z)}{G_l(z)} p_{l,k} \gamma_k(i, x_k) \pi(T^{l,j,z} T_{k,i}(x, y)) dz \\
&\quad = \sum_{l=1}^{K} \sum_{j=1}^{x_l+1} \int_{z=0}^{\infty} \gamma_l(j, x_l + 1) f_l(z) \lambda_l p_{l,k} \gamma_k(i, x_k) \pi(T_{k,i}(x, y)) dz \\
&\quad = \sum_{l=1}^{K} \lambda_l p_{l,k} \gamma_k(i, x_k) \pi(T_{k,i}(x, y)).
\end{aligned} \tag{11}$$

By equation (1) the in-flows (10) and (11) can be combined as

$$\lambda_k \gamma_k(i, x_k)\pi(T_{k,i}(x, y)) = \gamma_k(i, x_k)\phi_k(x)\pi(x, y), \tag{12}$$

where the second form is obtained with the aid of (8) noting that $G(y_{k,i}) = G(0) = 1$. This is seen to cancel the density of out-flow $u_{k,i}(x)\pi(x, y)$ emanating from the coordinate plane $y_{k,i} = 0$. □

Integration of (5) with respect to all the components of y leads to the following corollary:

Corollary 1. *The stationary distribution of X is*

$$\pi(x) = \frac{\Phi(x)}{\theta(\rho)} \prod_{k=1}^{K} \rho_k^{x_k}. \tag{13}$$

This is the main result of the paper: the stationary distribution of X in a network of symmetric queues with balanced service rates depends on the traffic characteristics only through the loads ρ_k of the queues $k = 1, \ldots, K$.

Dividing (5) by (13) we obtain another corollary:

Corollary 2. *Conditioned on $X = x$, the amounts of service received by different customers are independent and for all customers in queue k, irrespective of their position, the received service has the pdf $G_k(s)/s_k$.*

In fact, by making the state description even more detailled (as in [1]) by including the total service requirement $S_{k,i}$ of a customer in position (k, i) to the state for all k and $i = 1, \ldots, x_k$, one can find the stationary distribution in an analogous fashion and concludes again that given $X = x$ the pairs $(S_{k,i}, Y_{k,i})$ are independently distributed with the pdf of $S_{k,i}$ being $sf_k(s)/s_k$ and, further conditioned on $S_{k,i} = s$, the received service $Y_{k,i}$ has uniform distribution in $(0, s)$. The marginal distribution of $Y_{k,i}$, conditioned solely on $X = x$, then has the pdf $G_k(y)/s_k$ in accordance with Corollary 2.

Remark 1. *Also the following converse relation is true. If the stationary distribution of X in a network of symmetric queues is insensitive to the distributions of the service requirements at different queues, then the service rates are necessarily balanced.*

This result was derived for a network of processor sharing nodes in [3, Theorem 2] but the proof does not at all rely on any assumption on the service discipline other than that it must be work conserving or, in other words, that the $\phi_k(x)$ represent the actual rates at which queues are served, not just the service capacities.

References

[1] F.P. Kelly, *Reversibility and Stochastic Networks*, (Wiley, New York, 1979).

[2] R. Serfozo, *Introduction to Stochastic Networks* (Springer, Berlin, 1999).

[3] T. Bonald, A. Proutière, Insensitivity in processor-sharing networks, Performance Evaluation 49 (2002) 193-209.

[4] T. Bonald, A. Proutière, Insensitive bandwidth sharing in data networks, Queueing Systems 44 (2003) 69-100.