Resource allocation in general queueing networks with applications to data networks

P.K. Pollett*

Abstract

We consider the problem of how best to assign the service effort in a queueing network in order to minimise the expected delay under a cost constraint. We shall study systems with several types of customers, general service time distributions, stochastic or deterministic routing, and a variety of service regimes. For such networks there are typically no analytical formulae for the waiting time distributions. Thus, we shall approach the optimal allocation problem using approximation techniques, in particular, the *residual-life approximation* for the distribution of queueing times. This work generalises results of Kleinrock, who studied networks with exponentially distributed service times. We illustrate our results with reference to data networks.

1 Introduction

Since their inception, queueing network models have been used to study a wide variety of complex stochastic systems involving the flow and interaction of individual items: for example, "job shops", where manufactured items are fashioned by various machines in turn [7]; the provision of spare parts for collections of machines [17]; mining operations, where coal faces are worked in turn by a number of specialised machines [12]; delay networks, where packets of data are stored and then transmitted along the communications links that make up the network [18, 1]. For some excellent recent expositions, which describe these and other instances where queueing networks have been applied, see [2, 6] and the important text by Serfozo [16].

In each of the above-mentioned systems it is important to be able to determine how best to assign service capacity so as to optimise various performance measures, such as the expected delay or the expected number of items

*Department of Mathematics, University of Queensland, Queensland 4072, Australia.

(customers) in the network. We shall study this problem in greater generality than has previously been considered. We allow different types of customers, general service time distributions, stochastic or deterministic routing, and, a variety of service regimes. The basic model is that of Kelly [8], but we do not assume that the network has the simplifying feature of quasi-reversibility [9].

2 The model

We shall suppose that there are J queues, labelled $j = 1, 2, \ldots, J$. Customers enter the network from external sources according to independent Poisson streams, with type u customers arriving at rate ν_u (customers per second). Service times at queue j are assumed to be mutually independent, with an arbitrary distribution $F_j(x)$ that has mean μ_j^{-1} (units of service) and variance σ_j^2. For simplicity we shall assume that each queue operates under the usual first-come-first-served (FCFS) discipline and that a total effort (or capacity) of ϕ_j (units per second) is assigned to queue j. We shall explain later how our results can be extended to deal with other queueing disciplines.

We shall allow for two possible routing procedures, that of *fixed routing*, where there is a unique route specified for each customer type, and *random alternative routing*, where one of a number of possible routes is chosen at random. (We do not allow for *adaptive* or *dynamic routing*, where routing decisions are made on the basis of the observed traffic flow.)

For fixed routing we define $R(u)$ to be the (unique) ordered list of queues visited by type u customers. In particular, let $R(u) = \{r_u(1), \ldots, r_u(s_u)\}$, where s_u is the number of queues visited by a type-u customer and $r_u(s)$ is the queue visited at stage s along its route ($r_u(s)$, $s = 1, 2, \ldots, s_u$, are assumed to be distinct).

It is perhaps surprising that random alternative routing can be accommodated within the framework of fixed routing (see for example Exercise 3.1.2 of [10]). If there are several alternative routes for a given type u, then one simply provides a finer type classification for customers using these routes. We label the alternative routes as (u, i), $i = 1, 2, \ldots, N_u$, where N_u is the number of alternative routes for type-u customers, and we replace $R(u)$ by $R(u, i) = \{r_{ui}(1), \ldots, r_{ui}(s_{ui})\}$, for $i = 1, 2, \ldots, N_u$, where now $r_{ui}(s)$ is the queue visited at stage s along alternative route i and s_{ui} is the number of stages. We then replace ν_u by $\nu_{ui} = \nu_u q_{ui}$, where q_{ui} is the probability that alternative route i is chosen. Clearly $\nu_u = \sum_{i=1}^{N_u} \nu_{ui}$, and so the effect is to thin the Poisson stream of arrivals of type u into a collection of independent Poisson streams, one for each type (u, i). We should think of customers as being identified by their type, whether this be simply u for fixed routing, or the finer classification (u, i) for alternative routing. For convenience, let us denote by T the set of all types, and suppose that, for each t in T, customers of type t arrive according to a Poisson stream with rate ν_t and traverse the route $R(t) = \{r_t(1), \ldots, r_t(s_t)\}$, a collection of s_t distinct queues. This is the *network of queues with customers of different types* described in [8]. If all service times have a common exponential distribution with mean μ^{-1} (and hence $\mu_j = \mu$), the model is analytically tractable. In equilibrium the queues behave *independently*: indeed, *as if they were isolated*, each with independent Poisson arrival streams (independent among types). For example, if we let

$$\alpha_j(t,s) = \begin{cases} \nu_t & \text{if } r_t(s) = j \\ 0 & \text{otherwise,} \end{cases}$$

so that the arrival rate at queue j is given by

$$\alpha_j = \sum_{t \in T} \sum_{s=1}^{s_t} \alpha_j(t,s)\,,$$

and the demand (in units per second) by $a_j = \alpha_j/\mu$, then, provided the system is stable ($a_j < \phi_j$ for each j), the expected number of customers at queue j is given by

$$\bar{n}_j = \frac{a_j}{\phi_j - a_j} \tag{1}$$

and the expected delay $\overline{W}_j$ by

$$\overline{W}_j = \frac{1}{\alpha_j}\left(\frac{a_j}{\phi_j - a_j}\right) = \frac{1}{\mu\phi_j - \alpha_j}\,.$$

For further details, see Section 3.1 of [10].

3 The residual life approximation

Under our assumption that service times have *arbitrary* distributions, the model is rendered intractable. In particular, there are no analytical formulae for the delay distributions. We shall therefore adopt one of the many approximation techniques. Consider a particular queue j and let $Q_j(x)$ be the distribution function of the *queueing time*, that is, the period of time a customer spends at queue j *before* its service begins. The *residual-life approximation*, developed by the author [14], provides an accurate approximation for $Q_j(x)$:

$$Q_j(x) \simeq \sum_{n=0}^{\infty} \Pr(n_j = n) G_j^{(n)}(x)\,, \tag{2}$$

where

$$G_j(x) = \mu_j \int_0^{\phi_j x} (1 - F_j(y))\, dy$$

and $G_j^{(n)}(x)$ denotes the n-fold convolution of $G_j(x)$. The distribution of the number of customers n_j at queue j, which appears in (2), is that of a corresponding *quasi-reversible network* [10]: specifically, a network of *symmetric* queues obtained by imposing a symmetry condition at each queue j. In the present case, this amounts to replacing the existing FCFS discipline by a preemptive-resume last-come-first-served discipline at every queue in the network. The term *residual-life approximation* comes from renewal theory; $G_j(x)$ is the *residual-life distribution* corresponding to the (lifetime) distribution $F_j(x/\phi_j)$.

One immediate consequence of (2) is that the expected queueing time $\overline{Q}_j$ is approximated by

$$\overline{Q}_j \simeq \frac{1 + \mu_j^2 \sigma_j^2}{2\mu_j \phi_j}\, \bar{n}_j\ ,$$

where $\bar{n}_j$ is the expected number of customers at queue j in the corresponding quasi-reversible network. Hence, the expected delay at queue j is approximated as follows:

$$\overline{W}_j \simeq \frac{1}{\mu_j \phi_j} + \frac{1 + \mu_j^2 \sigma_j^2}{2\mu_j \phi_j}\, \bar{n}_j\ . \qquad (3)$$

Under the residual-life approximation, it is only $\bar{n}_j$ which changes when the service discipline is altered. For the existing FCFS discipline, $\bar{n}_j$ is given by (1) with $a_j = \alpha_j/\mu_j$.

Simulation results presented in [14] justify the approximation by assessing its accuracy under a variety of conditions. Even for relatively small networks with generous mixing of traffic, it is accurate, and the accuracy improves as the size and complexity of the network increases. (The approximation is very accurate in the tails of the queueing time distributions and so it allows an accurate prediction to be made of the likelihood of extreme queueing times.) For moderately large networks the approximation becomes worse as the coefficient of variation $\mu_j \sigma_j$ of the service time distribution at queue j deviates markedly from 1, the value which obtains in the exponential case.

4 Optimal allocation of effort

We now turn our attention to the problem of how best to apportion resources so that the expected network delay, or equivalently (by Little's Theorem) the expected number of customers in the network, is minimised. We shall suppose that there is some overall network budget F (dollars) which cannot be exceeded, and that the cost of operating queue j is a function f_j of its capacity. Suppose that the cost of operating queue j is proportional to ϕ_j, that is, $f_j(\phi_j) = f_j\phi_j$ (the units of f_j are dollars per unit of capacity, or dollar-seconds per unit of service). Thus, we should choose the capacities subject to the cost constraint

$$\sum_{j=1}^{J} f_j\phi_j = F\,. \tag{4}$$

We shall suppose that the average delay of customers at queue j is adequately approximated by (3), so that

$$\overline{W}_j \simeq \frac{1}{\mu_j\phi_j} + \frac{1+\mu_j^2\sigma_j^2}{2\mu_j\phi_j}\left(\frac{\alpha_j}{\mu_j\phi_j - \alpha_j}\right).$$

Using Little's Theorem, we obtain an approximate expression for the mean number $\bar{m}$ of customers in the network. This is

$$\begin{aligned} \bar{m} &\simeq \sum_{j=1}^{J}\alpha_j\left\{\frac{1}{\mu_j\phi_j} + \frac{\alpha_j(1+\mu_j^2\sigma_j^2)}{2\mu_j\phi_j(\mu_j\phi_j-\alpha_j)}\right\} \\ &= \sum_{j=1}^{J}a_j\left\{\frac{1}{\phi_j} + \frac{a_j(1+c_j)}{2\phi_j(\phi_j-a_j)}\right\}, \end{aligned}$$

where $c_j = \mu_j^2\sigma_j^2$ is the squared coefficient of variation of the service time distribution $F_j(x)$. We seek to minimise $\bar{m}$ over $\phi_1,\ldots,\phi_J$ subject to (4).

To this end, we introduce a lagrange multiplier λ^{-2}; our problem then becomes one of minimising

$$L(\phi_1,\ldots,\phi_J;\lambda^{-2}) = \bar{m} + \frac{1}{\lambda^2}\left(\sum_{j=1}^{J} f_j\phi_j - F\right).$$

Setting $\partial L/\partial\phi_j = 0$ for fixed j yields a quartic polynomial equation in ϕ_j, namely

$$2f_j\phi_j^4 - 4a_jf_j\phi_j^3 + 2a_j(a_jf_j-\lambda^2)\phi_j^2 - 2\epsilon_ja_j^2\lambda^2\phi_j + \epsilon_ja_j^3\lambda^2 = 0, \tag{5}$$

where $\epsilon_j = c_j - 1$, and our immediate task is to find solutions such that $\phi_j > a_j$ (recall that this latter condition is required for stability). The task is simplified by observing that the transformation

$$\phi_jf_j/F \to \phi_j,\ a_jf_j/F \to a_j,\ \lambda^2/F \to \lambda^2, \tag{6}$$

reduces the problem to one with unit costs $f_j = F = 1$, whence the polynomial equation (5) becomes

$$2\phi_j^4 - 4a_j\phi_j^3 + 2a_j(a_j - \lambda^2)\phi_j^2 - 2\epsilon_j a_j^2\lambda^2\phi_j + \epsilon_j a_j^3\lambda^2 = 0, \tag{7}$$

and the constraint becomes

$$\phi_1 + \phi_2 + \ldots + \phi_J = 1\,. \tag{8}$$

If service times are exponentially distributed ($\epsilon_j = 0$ for each j), it is easy to verify that (7) has a unique solution on (a_j, ∞) given by $\phi_j = a_j + |\lambda|\sqrt{a_j}$. Upon application of the constraint (8) we arrive at the optimal capacity assignment

$$\phi_j = a_j + \left(1 - \sum_{k=1}^{J} a_k\right)\frac{\sqrt{a_j}}{\sum_{k=1}^{J}\sqrt{a_k}}\,,$$

for unit costs. In the case of general costs this becomes

$$\phi_j = a_j + \frac{1}{f_j}\left(F - \sum_{k=1}^{J} f_k a_k\right)\frac{\sqrt{f_j a_j}}{\sum_{k=1}^{J}\sqrt{f_k a_k}}\,,$$

after applying the transformation (6). This is a result obtained by Kleinrock [11] (see also [10]): the allocation proceeds by first assigning enough capacity to meet the demand a_j, at each queue j, and then allocating a proportion of the affordable excess capacity,

$$\frac{1}{f_j}\left(F - \sum_{k=1}^{J} f_k a_k\right)$$

(that which could be afforded to queue j), in proportion to the square root of the cost $f_j a_j$ of meeting that demand. In the case where some or all of the ϵ_j, $j = 1, 2, \ldots, J$, deviate from zero, (7) is difficult to solve analytically. We shall adopt a perturbation technique, assuming that the lagrange multiplier and the optimal allocation take the following forms:

$$\lambda = \lambda_0 + \sum_{k=1}^{J} \lambda_{1k}\epsilon_k + O(\epsilon^2) \tag{9}$$

$$\phi_j = \phi_{0j} + \sum_{k=1}^{J} \phi_{1jk}\epsilon_k + O(\epsilon^2)\,, \quad j = 1, \ldots, J, \tag{10}$$

where $O(\epsilon^2)$ denotes terms of order $\epsilon_i\epsilon_k$. The zero-th order terms come from Kleinrock's solution: specifically, $\phi_{0j} = a_j + \lambda_0\sqrt{a_j}$, $j = 1, \ldots, J$, where

$$\lambda_0 = \frac{1 - \sum_{k=1}^{J} a_k}{\sum_{k=1}^{J}\sqrt{a_k}}\,.$$

On substituting (9) and (10) into (7) we obtain an expression for ϕ_{1jk} in terms of λ_{1k}, which in turn is calculated using the constraint (8) and by setting $\epsilon_k = \delta_{kj}$ (the Kronecker delta). We find that the optimal allocation, to first order, is

$$\phi_j = a_j + \lambda_0\sqrt{a_j} - \frac{\sqrt{a_j}}{\sum_{k=1}^J \sqrt{a_k}} \sum_{k\neq j} b_k\epsilon_k + \left(1 - \frac{\sqrt{a_j}}{\sum_{k=1}^J \sqrt{a_k}}\right) b_j\epsilon_j \,, \tag{11}$$

where

$$b_k = \frac{1}{4}\lambda_0 a_k^{3/2} \frac{a_k + 2\lambda_0\sqrt{a_k}}{(a_k + \lambda_0\sqrt{a_k})^2} \,.$$

For most practical applications, higher-order solutions are required. To achieve this we can simplify matters by using a single perturbation $\epsilon = \max_{1\leq j\leq J} |\epsilon_j|$. For each j we define a quantity $\beta_j = \epsilon_j/\epsilon$ and write ϕ_j and λ as power series in ϵ:

$$\lambda = \sum_{n=0}^{\infty} \lambda_n \epsilon^n \,, \qquad \phi_j = \sum_{n=0}^{\infty} \phi_{nj}\epsilon^n \,, \qquad j = 1, \ldots, J. \tag{12}$$

Substituting as before into (7), and using (8), gives rise to an iterative scheme, details of which can be found in [13]. The first-order approximation is useful, none-the-less, in dealing with networks whose service time distributions are all 'close' to exponential in the sense that their coefficients of variation do not differ significantly from 1. It is also useful in providing some insight into how the allocation varies as ϵ_j, for fixed j, varies. Let ϕ_i', $i = 1, 2, \ldots, J$, be the new optimal allocation obtained after incrementing ϵ_j by a small quantity $\delta > 0$. We find that to first order in δ

$$\phi_j' - \phi_j = \left(1 - \frac{\sqrt{a_j}}{\sum_{k=1}^J \sqrt{a_k}}\right) b_j\delta > 0$$

and, for $i \neq j$,

$$\phi_i' - \phi_i = -\frac{\sqrt{a_i}}{\sum_{k=1}^J \sqrt{a_k}} (\phi_j' - \phi_j) < 0 \,.$$

Thus, if the coefficient of variation of the service time distribution at a given queue j is increased (respectively decreased) by a small quantity δ, then there is an increase (respectively decrease) in the optimal allocation at queue j which is proportional to δ. All other queues experience a complementary decrease (respectively increase) in their allocations and the resulting deficit is reallocated in proportion to the square root of the demand.

In [13] empirical estimates were obtained for the radii of convergence of the power series (12) for the optimal allocation. In all cases considered there, the closest pole to the origin was on the negative real axis outside the physical limits for ϵ_i, which are of course $-1 \leq \epsilon_j < \infty$. The perturbation technique is therefore useful for networks whose service time distributions are, for example, Erlang (gamma) ($-1 < \epsilon_j < 0$) or mixtures of exponential distributions ($0 < \epsilon_j < \infty$) with not too large a coefficient of variation.

5 State-dependent capacity and nonlinear constraints

So far we have assumed that the capacity does not depend on the state of the queue (as a consequence of the FCFS discipline), and, that the cost of operating a queue is a linear function of its capacity. Let us briefly consider some other possibilities. Let $\phi_j(n)$ be the effort assigned to queue j when there are n customers present. If, for example, $\phi_j(n) = n\phi_j/(n+\eta-1)$, where η is a positive constant, the zero-th order allocation, optimal under (4), is precisely the same as before (the case $\eta = 1$). For values of η greater than 1 the capacity increases as the number of customers at queue j increases and levels off at a constant value ϕ_j as the number becomes large. If we allow η to depend on j we get a similar allocation but with the factor

$$\frac{\sqrt{f_j a_j}}{\sum_{k=1}^{J} \sqrt{f_k a_k}} \quad \text{replaced by} \quad \frac{\sqrt{f_j \eta_j a_j}}{\sum_{k=1}^{J} \sqrt{f_k \eta_k a_k}}.$$

See [10] for further details. The higher order analysis is very nearly the same as before. The factor $1 + c_j$ is replaced by $\eta_j(1 + c_j)$; for the sake of brevity, we shall omit the details.

As another example, suppose that the capacity function is linear, that is $\phi_j(n) = \phi_j n$, and that service times are exponentially distributed. In this case, the total number of customers in the system has a Poisson distribution with mean $\sum_{j=1}^{J} a_j/\phi_j$ and it is elementary to show that the optimal allocation subject to (4) is given by

$$\phi_j = \frac{\sqrt{f_j a_j}}{f_j \sum_{k=1}^{J} \sqrt{f_k a_k}} F, \qquad j = 1, \ldots, J.$$

It is interesting to note that we get a *proportional* allocation, $\phi_j/\phi_k = a_j/a_k$, in this case if (4) is replaced by $\sum_{j=1}^{J} \log \phi_j = 1$. More generally, we might use the constraint

$$\sum_{j=1}^{J} f_j \log(g_j \phi_j) = F$$

to account for 'decreasing costs': costs become less with each increase in capacity. Under this constraint, the optimal allocation is $\phi_j = \lambda a_j/f_j$, where

$$\log \lambda = \frac{F - \sum_{k=1}^{J} f_k \log(g_k a_k/f_k)}{\sum_{k=1}^{J} f_k}.$$

6 Data networks

One of the most interesting and useful applications of queueing networks is in the area of telecommunications, where they are used to model (among other things) data networks. In contrast to circuit switched networks (see for example [15]), where one or more circuits are held simultaneously on several links connecting a source and destination node, only one link is used at any time by a given transmission in a data network (message or packet switched network); a transmission is received in its entirety at a given node before being transmitted along the next link in its path through the network. If the link is at full capacity, packets are stored in a buffer until the link becomes available for use. Thus, the network can be modelled as a queueing network: the queues are the communications links and the customers are the messages. The most important measure of performance of a data network is the total delay, the time it takes for a message to reach its destination. Using the results presented above, we can optimally assign the link capacities (service rates) in order to minimise the expected total delay. We shall first explain in detail how the data network can be described by a queueing network.

Suppose that there are N switching nodes, labelled $n = 1, 2, \ldots, N$, and J communications links, labelled $j = 1, 2, \ldots, J$. We assume that all the links are perfectly reliable and not subject to noise, so that transmission times are determined by message length. We shall also suppose that the time taken to switch, buffer, and (if necessary) re-assemble and acknowledge, is negligible compared with the transmission times. Traffic entering the network from external sources is assumed to be Poisson, and that which originates from node m and is destined for node n is offered at rate ν_{mn}; the origin-destination pair determines the message type. Message lengths are assumed to be mutually independent and arbitrarily distributed with common mean μ^{-1} (bits, say). We shall assume that each link operates under a FCFS discipline and that a total capacity of ϕ_j (bits per second) is assigned to link j.

In order to apply the above results, we shall need to make a further assumption. It is similar to the celebrated *independence assumption* of Kleinrock [11]. We shall suppose that transmission times at any given link are independent and identically distributed, and that transmission times at different links are independent. Clearly a message maintains its length as it passes through the network. However, numerous simulation results (see for example [11]) suggest that, even so, the network behaves *as if* successive transmission times at a given link are independent. This phenomenon can be explained by observing that the arrival process at a given link is the result of the superposition of a generally large number of streams, and the approximation can then be justified on the basis of limit theorems concerning the superposition of marked point processes; see [4, 5], and the references therein. The assumption that independence is apparent at the queues *themselves* can be justified on the basis of the corresponding results on thinning of marked point processes (see for example [3]). Kleinrock's independence assumption differs from ours in that he assumes the transmission time distribution at a given link j is exponential with common mean μ^{-1}, a natural consequence of the usual teletraffic modelling assumption that messages emanating from outside the network are independent and identically distributed exponential random variables. However, although the exponential assumption is usually valid in circuit switched networks, we should not expect it to be appropriate in the present context of message/packet switching, since packets are of similar length. Thus, it is more realistic to assume, as we do here, that message lengths have an arbitrary distribution. In order that this be reflected in our independence assumption, we shall allow successive messages requesting transmission along a given link j to be arbitrarily distributed. Although this distribution might be the same at each link, we find it no less convenient to assume that it differs from one to another. Thus, in accordance with the notation established in Section 2, we can assume that at link j message lengths have a distribution function $F_j(x)$ which has mean μ_j^{-1} and variance σ_j^2.

For each origin-destination (ordered) pair (m, n), let $R(m, n) = \{r_{mn}(1), r_{mn}(2), \ldots, r_{mn}(s_{mn})\}$, be the ordered sequence of links used by messages on that route; s_{mn} is the number of links and $r_{mn}(s)$ is the link used at stage s. Let

$$\alpha_j(m, n, s) = \begin{cases} \nu_{mn} & \text{if } r_{mn}(s) = j, \\ 0 & \text{otherwise,} \end{cases}$$

so that the arrival rate at link j is given by

$$\alpha_j = \sum_m \sum_{n \neq m} \sum_{s=1}^{s_{mn}} \alpha_j(m, n, s),$$

and the demand (in bits per second) by $a_j = \alpha_j / \mu_j$. Assume that the system is stable ($a_j < \phi_j$ for each j). Then, for units costs, the optimal allocation of capacity (constrained by $\sum_j \phi_j = 1$) is given by $\phi_j = a_j + \lambda_0 \sqrt{a_j}$, $j = 1, \ldots, J$,

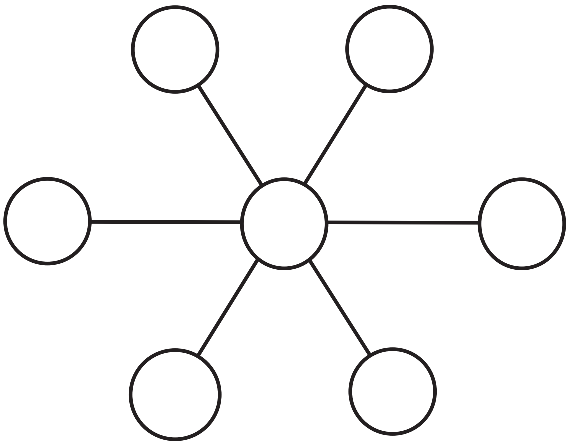

Figure 1: A symmetric star network (6 outer nodes and 6 links)

where

$$\lambda_0 = \frac{1 - \sum_{k=1}^{J} a_k}{\sum_{k=1}^{J} \sqrt{a_k}}\,,$$

in the case of exponential transmission times. More generally, in the case where the transmission times have an arbitrary distribution with mean μ_j^{-1} and variance σ_j^2, the optimal allocation is, to first order,

$$\phi_j = a_j + \lambda_0 \sqrt{a_j} - \frac{\sqrt{a_j}}{\sum_{k=1}^{J} \sqrt{a_k}} \sum_{k \neq j} b_k \epsilon_k + \left(1 - \frac{\sqrt{a_j}}{\sum_{k=1}^{J} \sqrt{a_k}}\right) b_j \epsilon_j\,, \tag{13}$$

where

$$b_k = \frac{1}{4} \lambda_0 a_k^{3/2} \frac{a_k + 2\lambda_0 \sqrt{a_k}}{(a_k + \lambda_0 \sqrt{a_k})^2}\,,$$

and $\epsilon_j = \mu_j^2 \sigma_j^2 - 1$.

To illustrate this, consider a *symmetric star network* such as the one depicted in Figure 1. A collection of outer nodes communicate via a single central node. Suppose that there are J outer nodes and thus J communications links. Figure 2 depicts the queueing network corresponding to the communications network in Figure 1. It is a fully-connected symmetric network, with the queues representing the communications links. Clearly there are $J(J-1)$ routes, a typical one being $R(m,n) = \{m,n\}$, where $m \neq n$. Suppose that transmission times have a common mean μ^{-1} and variance σ^2 (for simplicity, set $\mu = 1$), and, to begin with, suppose that transmission times are exponentially distributed and that all traffic is offered at the same

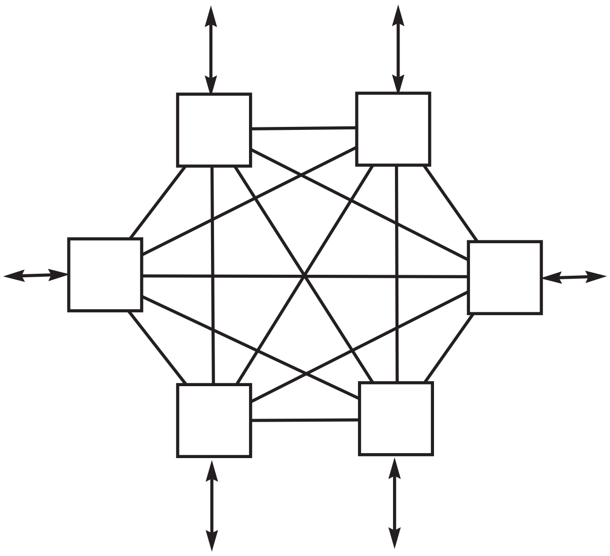

Figure 2: The corresponding queueing network (6 queues)

rate ν. Clearly the optimal allocation will be $\phi_j = 1/J$, owing to the symmetry of the network. What happens to the optimal allocation if we alter the traffic offered on one particular route by a small quantity? Suppose that we alter ν_{12} by setting $\nu_{12} = \nu + e$. The arrival rates at links 1 and 2 will then be altered by the same amount e. Since $\mu_j = 1$ we will have $a_1 = a_2 = \nu + e$ and $a_j = \nu$ for $j = 3, \ldots, J$. The optimal allocation is easy to evaluate. We find that, for $j = 1, 2$,

$$\phi_j = \nu + e + \frac{(1 - J\nu - 2e)\sqrt{\nu + e}}{(J-2)\sqrt{\nu} + 2\sqrt{\nu + e}} = \frac{1}{J} + \frac{1}{2}(J-2)\frac{(J\nu + 1)}{J^2\nu}\, e + O(e^2),$$

and, for $j = 3, \ldots, J$,

$$\phi_j = \nu + \frac{(1 - J\nu - 2e)\sqrt{\nu}}{(J-2)\sqrt{\nu} + 2\sqrt{\nu + e}} = \frac{1}{J} - \frac{J\nu + 1}{J^2\nu}\, e + O(e^2).$$

Thus, to first order in e, there is an $O(1/J)$ decrease in the capacity at all links in the network, except at links 1 and 2, where there is an $O(1)$ increase in capacity.

When the transmission times are not exponentially distributed, similar results can be obtained. For example, suppose that the transmission times have a distribution whose squared coefficient of variation is 2 (such as a mixture of exponential distributions). Then, it can be shown that the optimal allocation is given by

$$\phi_j = \frac{1}{J} + \frac{1}{2}\frac{(J^2\nu^2 - J\nu + 2)(J^2\nu^2 - 2J\nu - 1)}{J^2\nu}\, e + O(e^2),$$

for $j = 1, 2$, and

$$\phi_j = \frac{1}{J} - \frac{1}{4}(J-2)\frac{(J^2\nu^2 - J\nu + 2)(J^2\nu^2 - 2J\nu - 1)}{J^2\nu} e + O(e^2),$$

for $j = 3, \ldots, J$. Thus, to first order in e, there is an $O(J^3)$ decrease in the capacity at all links in the network, except at links 1 and 2, where there is an $O(J^2)$ increase in capacity. Indeed, the latter is true whenever the squared coefficient of variation c is not equal to 1, for it is easily checked that

$$\phi_j = \frac{1}{J} + g_J(c)e + O(e^2), \qquad j = 1, 2,$$

and

$$\phi_j = \frac{1}{J} - \frac{1}{2}(J-2)g_J(c)e + O(e^2), \qquad j = 3, \ldots, J,$$

where

$$g_J(c) = \frac{J\nu(J\nu - 1)^3 c - (J^4\nu^4 - 3J^3\nu^3 + 3J^2\nu^2 + J\nu + 2)}{2J^2\nu}.$$

Clearly $g_J(c)$ is $O(J^2)$. It is also an increasing function of c, and so this accords with our previous general results on varying the coefficient of variation of the service time distribution.

7 Conclusions

We have considered the problem of how best to assign service capacity in a queueing network so as to minimise the expected number of customers in the network subject to a cost constraint. We have allowed for different types of customers, general service time distributions, stochastic or deterministic routing, and, a variety of service regimes. Using an accurate approximation for the distribution of queueing times, we derived an explicit expression for the optimal allocation to first order in the squared coefficient of variation of the service time distribution. This can easily be extended to arbitrary order in a straightforward way using a standard perturbation expansion. We have illustrated our results with reference to data networks, giving particular attention to the symmetric star network. In this context we considered how best to assign the link capacities in order to minimise the expected total delay of messages in the system. We studied the effect on the optimal allocation of varying the offered traffic and the distribution of transmission times. We showed that for the symmetric star network, the effect of varying the offered traffic is far greater in cases where the distribution of transmission times deviates from exponential, and that more allocation is needed at nodes where the variation in the transmission times is greatest.

Acknowledgements: I am grateful to Tony Roberts for suggesting that I adopt the perturbation approach described in Section 4. I am also grateful to Erhan Kozan for helpful comments on an earlier draft of this paper. The support of the Australian Research Council is gratefully acknowledged.

References

[1] H. Akimaru and K. Kawashima. *Teletraffic: Theory and Applications*. Springer-Verlag, London, 2nd edition, 1999.

[2] G. Bloch, S. Greiner, H. de Meer and K. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications*. Wiley, New York, 1998.

[3] T. Brown. *Some Distributional Approximations for Random Measures*. PhD thesis, University of Cambridge, 1979.

[4] T. Brown and P. Pollett. Some distributional approximations in Markovian networks. *Adv. Appl. Probab.*, 14:654–671, 1982.

[5] T. Brown and P. Pollett. Poisson approximations for telecommunications networks. *J. Austral. Math. Soc.*, 32:348–364, 1991.

[6] X. Chao, M. Miyazawa and M. Pinedo. *Queueing Networks: Customers, Signals and Product Form Solutions*. Wiley, New York, 1999.

[7] J. Jackson. Jobshop-like queueing systems. *Mgmt. Sci.*, 10:131–142, 1963.

[8] F. Kelly. Networks of queues with customers of different types. *J. Appl. Probab.*, 12:542–554, 1975.

[9] F. Kelly. Networks of queues. *Adv. Appl. Probab.*, 8:416–432, 1976.

[10] F. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

[11] L. Kleinrock. *Communication Nets*. McGraw-Hill, New York, 1964.

[12] E. Koenigsberg. Cyclic queues. *Operat. Res. Quart.*, 9:22–35, 1958.

[13] P. Pollett. *Distributional Approximations for Networks of Queues*. PhD thesis, University of Cambridge, 1982.

[14] P. Pollett. Residual life approximations in general queueing networks. *Elektron. Informationsverarb. u. Kybernet.*, 20:41–54, 1984.

[15] K. Ross. *Multiservice Loss Models for Broadband Telecommunication Networks*. Springer-Verlag, London, 1995.

[16] R. Serfozo. *Introduction to Stochastic Networks*. Springer-Verlag, New York, 1999.

[17] J. Taylor and R. Jackson. An application of the birth and death process to the provision of spare machines. *Operat. Res. Quart.*, 5:95–108, 1954.

[18] W. Turin. *Digital transmission systems: performance analysis and modelling*. McGraw-Hill, New York, 2nd edition, 1998.