WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

CAPACITY ALLOCATION IN GENERALIZED JACKSON NETWORKS

Lawrence M. Wein

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

#2018-88

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

CAPACITY ALLOCATION IN GENERALIZED JACKSON NETWORKS

Lawrence M. Wein

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02139

#2018-88

CAPACITY ALLOCATION IN GENERALIZED JACKSON NETWORKS

Lawrence M. Wein

Sloan School of Management, Massachusetts Institute of Technology

Cambridge, Massachusetts 02139

We consider a capacity allocation problem for a generalized Jackson network (a Jackson network with general interarrival time and service time distributions). The problem is to determine the service rate (or capacity) that minimizes the expected equilibrium customer delay subject to a linear budget constraint on the capacities. The problem is analyzed using a Brownian approximation of the generalized Jackson network. The resulting capacity allocation is a generalization of the classical square root capacity allocation for Jackson networks. A numerical example is provided that demonstrates the allocation's effectiveness.

1. Introduction

Consider a queueing network with K single-server stations, each of which has an infinite capacity waiting room. Customers arrive at station k according to a renewal process, and the interarrival times have finite mean λ_k^{-1} and finite squared coefficient of variation (variance divided by the square of the mean) c_{ak}^2. Customers, upon completion of service at station k, are next routed to station j with probability P_{kj}, independent of previous history. It is assumed that the Markov routing matrix $P = (P_{kj})$ has spectral radius less than one, so that all customers eventually exit the network. We also assume $P_{kk} = 0$ for $k = 1, ..., K$, so no immediate feedback of customers is allowed. The service times at station

$k = 1, ..., K$ are independent and identically distributed random variables with finite mean μ_k^{-1} and finite squared coefficient of variation c_{sk}^2. The sequences of interarrival times, service times, and routing decisions are assumed to be mutually independent. Customers are served FIFO (first-in first-out) at each station. Let the $K-$vector $\gamma = (\gamma_k)$ of effective arrival rates be defined by $\gamma = \lambda(I - P)^{-1}$, where $\lambda = (\lambda_k)$ is the $K-$vector of arrival rates. Then it is assumed that the traffic intensity $\rho_k = \gamma_k/\mu_k < 1$ for each $k = 1, ..., K$, so that the system is stable. This model is called a generalized Jackson network, since it reduces to a standard Jackson network when the interarrival times and service times at each station are exponentially distributed.

Queueing networks are very useful models of computer, communication, and manufacturing systems. One of the basic design issues for such a network is that of capacity allocation. The classical result on this subject is the square root capacity assignment derived by Kleinrock [3]. The problem he considers is to choose the vector of service rates $\mu = (\mu_k)$ in a Jackson network to minimize the expected equilibrium sojourn time per customer (or, equivalently, the expected equilibrium customer delay or the expected equilibrium number of customers in the system) subject to the budget constraint $\sum_{k=1}^{K} d_k \mu_k = D$, where d_k is the unit cost of capacity at station k, and D is the total available budget. The solution to this problem is to choose

$$\mu_k^* = \gamma_k + \frac{\sqrt{d_k \gamma_k}}{\sum_{j=1}^{K} \sqrt{d_j \gamma_j}} \left(\frac{D - \sum_{j=1}^{K} d_j \gamma_j}{d_k} \right) \text{ for } k = 1, ..., K. \tag{1}$$

In the case where the unit cost of capacity is the same at each station, this assignment first allocates just enough capacity to each station to satisfy its effective arrival rate, and then allocates the excess capacity among the stations in proportion to the square roots of their effective arrival rates.

Notice that Kleinrock assumes capacity is continuous, whereas in most applications capacity is actually discrete (for example, parallel machines at a manufacturing work station); see Shanthikumar and Yao [5], [6] for recent work on server allocation in Jackson

networks. However, in many applications, such as communication networks, the capacity is allocated in large enough numbers that the continuity assumption is quite reasonable. Furthermore, even when the continuity assumption is not realistic, this assignment still offers a back-of-the-envelope calculation that can add some insight to the allocation problem.

Kleinrock also assumes that all interarrival times and service times are exponentially distributed. Thus, it is assumed that each station, in effect, exhibits the same amount of variability. However, in most queueing systems, some stations exhibit more variability than other stations. In a manufacturing environment, for example, this variability may stem from the occurance of machine breakdowns or operator unavailability. Recently, Bitran and Tirupati [1] have analyzed capacity allocation in multiclass queueing networks with general interarrival time and service time distributions. However, they do not obtain closed form results, and so it is very difficult to gain any insight from their model without making extensive numerical calculations.

In this paper, we derive a simple capacity assignment for the generalized Jackson network. The problem we analyze is the same as Kleinrock's problem, except that general interarrival and service time distributions are allowed. As in Kleinrock's problem, the squared coefficients of variation for all service time distributions are assumed to be constant, and independent of the value of the service rates chosen. Our optimal assignment is a square root allocation that reduces to Kleinrock's assignment when $c_{ak}^2 = c_{sk}^2 = 1$ for $k = 1, ..., K$. The result is obtained by using an approximation based on results from heavy traffic theory developed by Reiman [4] and Harrison and Williams [2]. Although our results are rooted in heavy traffic theory, they appear to be quite robust with regard to the heavy traffic assumptions, as can be seen in the numerical example of Section 4.

2. The Brownian Approximation

In this section, the Brownian model proposed by Harrison and Williams [2] will be briefly summarized. This model is a refinement of the heavy traffic approximation of a generalized Jackson network derived by Reiman [4]. The Brownian approximation requires that the total load imposed on each station is approximately equal to its capacity. More precisely, we assume there exists a large integer n such that

$$\max_{1 \le k \le K} \sqrt{n}|1 - \rho_k| < 1. \tag{2}$$

As a canonical example, one may think of ρ_k being between 0.9 and 1.0 for each station k, in which case $n = 100$ satisfies the *balanced heavy loading* condition (2). The primary process of interest is the *scaled vector queue length* process $Q^* = (Q_k^*)$ defined by

$$Q_k^*(t) = \frac{Q_k(nt)}{\sqrt{n}}, \ t \ge 0, \text{ for } k = 1, ..., K, \tag{3}$$

where $Q_k(t)$ is the number of customers queued and in service at station k at time t, and n is the large integer specified in (2). Under the balanced heavy loading condition (2), Harrison and Williams [2] show that the scaled queue length process Q^* is well approximated by a process Z that has a unique stationary distribution. They also show that this stationary distribution has a product-form density function if and only if a certain skew-symmetry condition holds among the network data. It has been suggested in Harrison and Williams [3] that the stationary distribution of Z may be approximated by the product form distribution even when the skew-symmetry condition does not hold exactly, and this suggestion is incorporated into our approximation scheme. Under this assumption, Harrison and Williams show that the expected number of customers at station k in equilibrium is

$$\frac{\sigma_k^2}{2(\mu_k - \lambda)} \text{ for } k = 1, .., K, \tag{4}$$

where, from equation (27) of Reiman [4] and equation (2.23) of Harrison and Williams [2],

$$\sigma_k^2 = \lambda_k c_{ak}^2 + \gamma_k c_{sk}^2 + \sum_{j=1}^{K} \gamma_j P_{jk}(c_{sj}^2 P_{jk} + 1 - P_{jk}) \text{ for } k = 1, ..., K. \tag{5}$$

3. The Capacity Allocation

Using the Brownian approximation of the previous section, the capacity allocation problem is to choose the vector $\mu = (\mu_k)$ so as to minimize

$$\sum_{k=1}^{K} \frac{\sigma_k^2}{2(\mu_k - \lambda)} \tag{6}$$

subject to

$$\sum_{k=1}^{K} d_k \mu_k = D. \tag{7}$$

To solve this problem, we define the Lagrangian L by

$$L = \sum_{k=1}^{K} \frac{\sigma_k^2}{2(\mu_k - \lambda)} + \pi(\sum_{k=1}^{K} d_k \mu_k - D), \tag{8}$$

where π is the Lagrange multiplier, and set $\frac{dL}{d\mu_k} = 0$ for $k = 1, ..., K$. This yields

$$\mu_k^* = \gamma_k + \frac{\sqrt{\sigma_k^2}}{\sqrt{2\pi d_k}} \text{ for } k = 1, ..., K. \tag{9}$$

Substituting these values of μ_k into the constraint (7), solving for π, and substituting the value of π back into equation (9) yields the square root capacity assignment:

$$\mu_k^* = \gamma_k + \frac{\sqrt{d_k \sigma_k^2}}{\sum_{j=1}^{K} \sqrt{d_j \sigma_j^2}} \left(\frac{D - \sum_{j=1}^{K} d_j \gamma_j}{d_k}\right) \text{ for } k = 1, ..., K. \tag{10}$$

Notice that when $c_{ak}^2 = c_{sk}^2 = 1$ for $k = 1, ..., K$, then, by equation (5),

$$\sigma_k^2 = \lambda_k + \gamma_k + \sum_{j=1}^{J} \gamma_j P_{jk} = 2\gamma_k, \tag{11}$$

and thus our assignment (10) reduces to Kleinrock's assignment (1) under the exponential assumptions.

The capacity assignment (10) allocates the excess capacity to station k in proportion to the square root of the variability parameter σ_k^2. The quantity σ_k^2 in equation (5) has three terms. The first term measures the amount of variability arriving to station k from outside of the queueing system, the second term measures the amount of variability from the server at station k, and the third term measures the total variability arriving to station k from the other stations in the system. The main insight gained from this paper is the following: the optimal capacity assignment compensates for the highly variable stations by adding more capacity to them.

4. An Example

Consider a three station network with $\lambda = (6, 2, 0), c_a^2 = (.54, 4, 0), c_s^2 = (.54, 2, .08)$, and

$$P = \begin{pmatrix} 0 & 1/3 & 2/3 \\ 0 & 0 & 0 \\ 0 & 1/2 & 0 \end{pmatrix}. \tag{12}$$

Assume all capacity costs are the same, so that $d_1 = d_2 = d_3 = 1$. It can easily be shown that the skew-symmetry condition in Harrison and Williams [2] is not satisfied by this problem data. The vector of effective arrival rates for our problem is $\gamma = (6, 6, 4)$. We will consider four different cases of our problem, with the total budget D varying in each case. Let us define the traffic intensity ρ of the problem to equal $\sum_{k=1}^{3} \gamma_k / D$, which is the sum of the effective arrival rates divided by the sum of the service rates. The four values of D were chosen to correspond to values of ρ=.6, .7, .8, and .9, respectively. For all four cases, we performed simulation experiments on three capacity assignment rules: the Brownian approximation (denoted by B in Table 1) from equation (10), Kleinrock's assignment (denoted by K) from equation (1), and the proportional capacity assignment

(denoted by P), where capacity is allocated in direct proportion to the effective arrival rates; that is,

$$\mu_k^* = \frac{\gamma_k}{\sum_{j=1}^{K} \gamma_j} D \text{ for } k = 1, ..., K. \tag{13}$$

(INSERT TABLE 1 HERE)

The results are displayed in Table 1. Each combination of capacity assignment and traffic intensity was tested by 200 independent runs of 10,000 customers each. For each case, we computed the average customer sojourn time (and 95% confidence interval), the average server utilization levels, and the average queue lengths at the three stations. For all values of the traffic intensity ρ, the Brownian approximation outperformed the proportional assignment rule, which in turn outperformed Kleinrock's assignment rule. Since exponential distributions were not assumed in this example, it is not surprising that the proportional rule outperformed Kleinrock's rule. It is interesting to note that, although our results were derived under heavy traffic conditions, the Brownian allocation performed well at all levels of the traffic intensity. For $\rho = .6, .7, .8$, and $.9$, the percentage improvements of the Brownian assignment over the proportional assignment were 5.0%, 6.5%, 8.0%, and 6.4%, respectively. Notice that, of the three rules, the Brownian assignment produced the most *imbalanced* average utilization levels, but the most *balanced* average queue lengths.

References

[1] G. R. Bitran and D. Tirupati, "Trade-off Curves, Targeting and Balancing in Queueing Networks," submitted to *Oper. Res.*, 1977.

[2] J. M. Harrison and R. Williams, "Brownian Models of Queueing Networks with Homo-

geneous Customer Populations," to appear in *Stochastics*, 1987.

[3] L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, Dover Publications, Inc., New York, 1964.

[4] M. I. Reiman, "Open Queueing Networks in Heavy Traffic," *Math. of Oper. Res.* 9, 441-458, 1984.

[5] J. G. Shanthikumar and D. D. Yao, "Optimal Server Allocation in Systems of Multi-Server Stations," Management Science 33, 1173-1180, 1987.

[6] J. G. Shanthikumar and D. D. Yao, "On Server Allocation in Multiple Center Manufacturing Systems", to appear in *Oper. Res.*, 1987.

CAPACITY ALLOCATION	TRAFFIC INTENSITY	MEAN SOJOURN TIME (95% CI)	MEAN SERVER UTILIZATIONS			MEAN QUEUE LENGTHS		
			1	2	3	1	2	3
B	*.9*	*2.80 (±.05)*	*92.3*	*86.4*	*91.8*	*5.77*	*10.0*	*3.94*
P	*.9*	*2.99 (±.07)*	*90.0*	*89.8*	*89.9*	*4.22*	*14.0*	*3.03*
K	*.9*	*3.10 (±.08)*	*90.4*	*90.3*	*88.5*	*4.45*	*15.2*	*2.53*
B	*.8*	*1.26 (±.01)*	*84.2*	*74.0*	*84.0*	*2.31*	*3.75*	*1.64*
P	*.8*	*1.37 (±.02)*	*80.0*	*80.0*	*79.9*	*1.63*	*5.78*	*1.17*
K	*.8*	*1.40 (±.02)*	*80.9*	*80.8*	*77.5*	*1.75*	*6.12*	*0.96*
B	*.7*	*.749 (±.004)*	*75.8*	*62.4*	*75.0*	*1.19*	*1.88*	*0.81*
P	*.7*	*.803 (±.007)*	*70.0*	*70.0*	*70.0*	*0.80*	*2.96*	*0.58*
K	*.7*	*.822 (±.007)*	*71.1*	*71.1*	*67.0*	*0.86*	*3.16*	*0.48*
B	*.6*	*.490 (±.002)*	*66.8*	*51.7*	*66.1*	*0.65*	*0.99*	*0.45*
P	*.6*	*.516 (±.003)*	*60.0*	*60.0*	*60.0*	*0.42*	*1.62*	*0.31*
K	*.6*	*.529 (±.003)*	*61.3*	*56.5*	*56.5*	*0.45*	*1.75*	*0.24*

Table 1