European Journal of Operational Research 230 (2013) 596–603

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Server allocation for zero buffer tandem queues

Mohammad H. Yarmand, Douglas G. Down *

Department of Computing and Software, McMaster University, Hamilton, ON, Canada L8S 4K1

ARTICLE INFO

Article history:
Received 3 August 2012
Accepted 17 May 2013
Available online 25 May 2013

Keywords:
Server allocation
Zero buffer
Tandem queue

ABSTRACT

In this paper we consider the problem of allocating servers to maximize throughput for tandem queues with no buffers. We propose an allocation method that assigns servers to stations based on the mean service times and the current number of servers assigned to each station. A number of simulations are run on different configurations to refine and verify the algorithm. The algorithm is proposed for stations with exponentially distributed service times, but where the service rate at each station may be different. We also provide some initial thoughts on the impact on the proposed allocation method of including service time distributions with different coefficients of variation.

1. Introduction

Consider a tandem line consisting of N stations ($N \geqslant 2$) where the service rate of a server assigned to station i is μ_i ($i = 1, 2, \ldots, N$). The service times at each station follow exponential distributions and are independent and identically distributed with rate μ_i (i.e. the rate can depend on the station). There are M servers available to be allocated to the stations. The servers are capable of working at any station and can process only one job at a time. The servers are homogeneous meaning that servers assigned to the ith station each work at rate μ_i.

We assume that there are always jobs waiting to be served at the first station. Jobs served at the last station immediately leave the system. The throughput of the system is equal to the departure rate from the last station. We assume there are no buffers between stations. Our problem of interest is: *given M servers, allocate them to the N stations, such that the throughput is maximized.* We could define a similar problem in terms of blocking and starvation probabilities. In that case, the goal would be to minimize an aggregated measure of these probabilities over all stations.

We propose an algorithm that has as a primary goal to roughly equalize the workloads at each of the stations, meaning that the number of servers is proportional to the mean service time at a station. However, the heterogeneous mean service times and lack of buffers introduce additional complexity beyond making the workloads equal (further discussed in Section 2). We use simulation to obtain insights about the nature of the system and later to measure the performance of our algorithm for a number of configurations. In addition to exponentially distributed service times, we extend the algorithm by considering service times with coefficients of variation other than 1. We illustrate that the algorithm performs well if the coefficients of variation of all stations are increased or decreased equally. Based on a number of simulations, we infer that the algorithm also works well on configurations where the majority of stations have service times with coefficient of variation near one and the remaining stations have service times with coefficient of variation less than one.

The stated problem is motivated by the bed management issue in hospitals. In short, bed management is the problem of assigning a number of beds to different departments of a hospital, such that patient flow is optimized [6]. Patients need to go through these departments to complete their treatment cycle (e.g. the emergency, express, medicine, and alternative level of care departments). The fact that patients must be assigned to a bed at all times, represents the zero-buffer nature of this problem.

A zero-buffer environment arises either from characteristics of the processing technology itself, or from the absence of storage capacity between stations. The bed management problem is caused by the absence of storage capacity. Another example is the allocation of facilities/workers to the stations of an assembly line. As a concrete example, Hu et al. [14] consider a car assembly line in which each car is carried by a specific conveyor with no extra conveyors between stations.

An example of a case where the technology itself requires a zero-buffer environment is the canning process in which delays should be avoided to keep the food fresh. In particular, no buffer space is allowed between the cooking operation and the canning operation [3]. Another example is the production of steel, where molten steel undergoes a series of operations such as molding into ingots, unmolding, reheating, soaking, and preliminary rolling [21]. To maintain the molten steel's temperature, each operation should

* Corresponding author. Tel.: +1 905 525 9140.
E-mail addresses: yarmanmh@mcmaster.ca (M.H. Yarmand), downd@mcmaster.ca (D.G. Down).

0377-2217/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2013.05.029

follow the previous operation, immediately. Such applications are closely related to the problem of scheduling jobs in a no-wait setting.

Optimization modeling is typically used to formulate general allocation problems in this research domain (see Hillier and So [11], for example). Throughput is denoted by $R(q,s,w)$, where $q = (q_1, q_2, \ldots, q_N)$, $s = (s_1, s_2, \ldots, s_N)$, and $w = (w_1, w_2, \ldots, w_N)$ denote the allocation of buffers, servers, and workload to stations respectively. Q is the total number of available buffer spaces and W is the total mean service time over all stations. The optimization problem is expressed as:

$$\begin{aligned}
&\text{maximize } R(q,s,w)\\
&\text{subject to } \sum_{j=2}^{N} q_j = Q,\\
&\qquad\qquad \sum_{j=1}^{N} s_j = M,\\
&\qquad\qquad \sum_{j=1}^{N} w_j = W,\\
&\qquad\qquad q_j \text{ is an integer greater than or equal to } 0, \quad j \in \{2,3,\ldots,N\},\\
&\qquad\qquad s_j \text{ is an integer greater than } 0, \quad j \in \{1,2,\ldots,N\},\\
&\qquad\qquad w_j > 0, \quad j \in \{1,2,\ldots,N\},
\end{aligned}$$

where q, s, and w are decision vectors (q has entries q_j, etc.). Note that workload allocation (w) is the problem of determining the mean service time at each station, given that the mean service times sum to a fixed value W.

Hillier and So [10] aim to maximize throughput for tandem queues with equal workloads (w_i equal for all i) and small or no buffers ($q_i = 0$ or 1). They claim the optimal server allocation (s) assigns extra servers rather uniformly to the interior stations and refine this claim based on the number of servers and stations at hand. They introduce the *bowl phenomenon*: with single server stations, different mean service times, and equal buffers, the optimal workload allocation (w) assigns less work to the interior stations than to the end stations. It appears that the interior stations (especially the center stations) are critical in determining system performance and so should be given preferential treatment when making design decisions. Alexandros and Chrissoleon [1] extend Hillier and So's method [10] and perform server allocation in large production lines with multiple parallel stations. They employ simulated annealing to solve the optimization problem which models the allocation problem.

Magazine and Stecke [15] consider a three station tandem queueing system with no buffers ($q_i = 0$). They follow the results of Hillier and So [10] and as the number of servers increases, the unbalancing in favour of the middle station is increased. This behavior continues until the unbalancing becomes too severe. At this point, a server is taken away from the middle station and a server is added to the first and third stations. They also state that if unbalancing and distributing servers (w,s) are left to our control, both should be as balanced as possible.

Avi-Itzhak and Yadin [4] study single server stations with no or finite buffers in between stations. For two-station lines, they calculate the mean response time in terms of probabilities of the first station being empty/busy, queue sizes, and the number of jobs in stations.

Cheng and Zhu [5] state that when assigning M heterogeneous servers to M stations with no buffer between the first two ($q_2 = 0$) (resp. the last ($q_M = 0$)) stations and possible buffers for interior stations, it is better to allocate the slower server to the first (resp. the last) station.

Van Woensel et al. [19,24] move a step further and consider any possible acyclic multi-server configuration with arbitrary service and inter-arrival time distributions. They model the joint buffer and server allocation problem (q,s) as a non-linear optimization problem with integer decision variables. They use the Generalized Expansion Method to evaluate throughput. They further use Powell's algorithm (detailed in Himmelblau [13]) for allocation purposes. Smith et al. [20] also model the buffer allocation problem (q) as an optimization problem and use the Generalized Expansion Method to estimate the throughput.

Andriansyah et al. [3] study zero-buffer multi-server general queueing networks. They use the Generalized Expansion Method to evaluate the throughput for a class of acyclic networks. They employ genetic algorithms to solve a multi-objective optimization problem to provide the trade-off between the total number of servers used and the throughput. van Vuuren et al. [23] study multi-server tandem queues with finite buffers with generally distributed service times. They decompose lines to two-station subsystems by a spectral expansion method.

Andradóttir et al. [2] study server allocation (s) in infinite buffer settings ($q_i = \infty$) with flexible servers using a linear programming approach. We would like to contrast the two extremes (in terms of buffer sizes) in tandem lines for allocation of fixed servers. Namely, in Section 2 we compare our configuration of interest (zero-buffer) with a configuration with infinite buffers between stations.

There has also been work done on the effect of variability of service times for tandem lines. El-Rayah [7] studies the optimal arrangement of single server tandem lines (s) with no buffer spaces ($q_i = 0$) and where servers have different coefficients of variation. They discover that assigning servers with higher coefficients of variation to the exterior stations leads to higher throughput. Muth and Alkaff [16] study the effect of independent changes in the mean service time and the service time variance on a tandem line's throughput. They study single-server tandem lines with three stations and no buffers and offer a method to compute the throughput. Papadopoulos et al. [9,17,18] examine specific production lines (with feedback or unreliable stations) by generating sparse transition matrices and solving them using the Successive Over Relaxation (SOR) method. They consider single-server tandem lines with finite buffers and Erlang or exponential service times.

Futamura [8] studies the effect of service time variability in systems with and without buffers. Futamura suggests that server allocation should follow the inverted bowl phenomenon except that more servers are assigned to stations with higher coefficient of variation to alleviate the impact of higher variance. Hillier et al. [12] define the *inverted bowl phenomenon*: when the total amount of storage space is a decision variable and workloads are equal (w_i equal for all i), the optimal buffer allocation (q) commonly follows an inverted bowl pattern. In other words, the allocation provides the stations toward the center of the line with more buffer storage space than the other stations.

The problem we consider is different in the following respects. The models in [5,12,22,24] include buffers in their configurations. Avi-Itzhak and Yadin [4] study small single server lines, however it is not clear how to generalize their results to longer multi-server lines. Hillier and So [10] consider tandem queues with small buffers and perform simulations for the case with no buffers. They assume that workload is balanced and the numbers of servers at stations differ by at most two (i.e. there is a limited number of extra servers). In other words, starting from a balanced system, they study how to allocate extra servers. We will apply their allocation method to more generic cases to discover its potential shortcomings. Futamura [8] studies the same tandem queues that Hillier and So [10] consider. The tandem line that Magazine and Stecke [15] targets is limited as all rates are equal and there are only three stations. Andriansyah et al. [3] focus on a system with arrivals, with better results achieved when the arrival rate is somewhat