International Journal of Production Economics 196 (2018) 211–225

Contents lists available at ScienceDirect

International Journal of Production Economics

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

# Exploiting the characteristics of serial queues to reduce the mean and variance of flow time using combined priority rules

Rodrigo Romero-Silva [a,*], Sabry Shaaban [b], Erika Marsillac [c], Margarita Hurtado [a]

[a] Faculty of Engineering, Universidad Panamericana, Augusto Rodin 498, Insurgentes Mixcoac, 03920, Mexico City, Mexico
[b] Department of Strategy, ESC La Rochelle, 102 Rue de Coureilles, 17024 La Rochelle, France
[c] Information Technology, Decision Sciences and Maritime Supply Chain Management Department, Old Dominion University, 5115 Hampton Boulevard, Norfolk, VA 23529, USA

ARTICLE INFO

*Keywords:*
Dispatching rules
Dynamic scheduling
Balanced flow-shop
Tandem queues
Discrete event simulation

ABSTRACT

This paper addresses the trade-off challenge from reducing either the mean or variance of flow time when using simple sequencing rules in balanced, multi-class, serial queues. Study results show that instead of the expected zero sum situation, a balance between the two objectives can be achieved by assigning different priority rules to different queues. The order of priority rule assignments in different queues is shown to be relevant because variability along the line creates unbalanced queue lengths for each station, depending on the characteristics of the line. Thus, it was found that a simple heuristic for reducing both the mean and the variance of flow time in non-heavy traffic environments is to assign the first queue a priority rule that reduces its mean queue length while assigning the other queues a priority rule that reduces the variance of flow time. Conversely, for very-high traffic environments, performance improvements are shown from assigning the first queue a priority rule that reduces the variance of flow time while assigning the other queues a priority rule that reduces queue length.

## 1. Introduction

Priority rules have been widely used to solve sequencing and scheduling problems in dynamic production environments, where static schedules are difficult to implement due to the dynamic arrival rate of jobs and the stochastic nature of the production process.

The Shortest Expected Processing Time (SEPT) rule is a useful heuristic that reduces total completion time and jobs' mean flow time (*MFT*) for deterministic and stochastic machine shop environments, respectively (Blackstone et al., 1982; Pinedo, 2012). It is also useful to minimise the mean queue length in multi-class, non-preemptive M/G/1 queues (Buzacott and Shanthikumar 1993; Gross et al., 2008).

Despite this, in stochastic environments, as *MFT* is reduced through the application of SEPT, the variance of flow time (*VFT*) is also increased when compared to other sequencing rules (Holthaus and Rajendran, 1997). Consequently, increased *VFT* can result in reduced predictability, directly impacting business performance, since customers may prioritise delivery dependability over delivery speed in certain environments (Wierman et al., 2005; Christensen et al., 2007).

To address this trade-off, some authors (Mahmoodi et al., 1996; Barman, 1997; Barman and Laforge 1998) suggest using a combination of different simple priority rules (PR) throughout the flow-shop's stations to address opposing objectives in dynamic and stochastic flexible flow shop environments. For instance, Barman and Laforge (1998) found that a combination of PR reduced delivery speed and increased dependability. However, Barman (1997) found that the order of assignment of PR to different stations in a balanced line did not produce significantly different results.

In a similar study, when considering more sophisticated PR, Jayamohan and Rajendran (2000) suggest that the use of a single rule could produce better results (simultaneously reducing *MFT* and *VFT*, mean tardiness and number of tardy jobs) than a combination of simple rules. They concluded that Barman's results lacked a direct comparison with better performing rules. Furthermore, the study of Barman and Laforge (1998) found that assigning SEPT to the second station in a 'flow-dominant shop', where the intermediate work centre serves as a bottleneck station, diminished the mean flow time while also limiting the increase of flow time variance.

* Corresponding author.
*E-mail addresses:* rromeros@up.edu.mx (R. Romero-Silva), shaabans@esc-larochelle.fr (S. Shaaban), emarsill@odu.edu (E. Marsillac), mhurtado@up.edu.mx (M. Hurtado).

https://doi.org/10.1016/j.ijpe.2017.11.023
Received 8 May 2017; Received in revised form 13 November 2017; Accepted 29 November 2017
Available online 1 December 2017
0925-5273/

This suggests that the approach of assigning different PR to different stations and assigning them according to the characteristics of the line could show some promise in a practical sense, as it uses and applies simple rules constantly throughout the line stations.

### 1.1. Objectives of the paper

The study of PR in multi-class, non-pre-emptive serial queues has value for companies that cannot determine in advance the actual processing time of *each* job (a necessary condition to apply other sequencing rules) but can model each *class* of jobs, using a probability distribution. Hence, this paper investigates the impact of assigning simple PR to different stations of balanced serial queues to reduce both *MFT* and *VFT*.

Moreover, this analyses whether the specific stochastic characteristics of the line can be exploited to produce a good heuristic for assigning PR to specific stations. This paper will address the following research questions:

(1) Which approach is best to address multi-criteria problems (such as the reduction of *MFT* and *VFT*): using single PR or using a combined assignment of PR?
(2) Does the assignment order of PR influence *MFT* and *VFT* in balanced serial queues?
(3) Which is the best simple heuristic for assigning PR to individual stations?

The answers to these questions can provide practitioners a simple static procedure for sequencing their operations to reduce *MFT* and *VFT*. To comprehensively assess these scenarios, a simulation study of various balanced, single-server, multi-class serial queues was developed.

Paper organisation is as follows: Section 2 provides a literature review of sequencing and PR. The study methodology is described in Section 3, while Section 4 shows the study results. Section 5 presents the discussion and implications of the results and Section 6 provides study conclusions.

## 2. Literature review

Scheduling and Queueing Theory experts have studied the use of sequencing and PR under different philosophies. Thus, it is felt that the "*integration of methodologies from these two fields may be valuable*" (Terekhov et al., 2014).

Scheduling aims to find the best sequencing rule(s) to order a set of *n* jobs in a specific station to reach certain levels of shop performance, where each job has its own characteristics (deterministic or stochastic) and is assigned a particular order in the sequence of each queue. Typical environments where sequencing rules are applied include single and parallel machine environments, simple and hybrid flow-shops, and simple and hybrid job-shops.

Alternately, Queueing Theory studies stochastic systems, where jobs are categorised into classes instead of one single entity, and aims to find the steady-state performance of each customer class and of the overall queue regarding different PR. Typical environments studied under this paradigm include single and multiple server queues and serial (tandem) queues with single or multiple servers.

Such complementary goals lead us to review different sequencing and PR options concerning the reduction of *MFT* and *VFT*.

### 2.1. Sequencing rules

The ability of the SPT (Shortest Processing Time) rule to reduce *MFT* has been widely recognised (Blackstone et al., 1982; Shafaei and Brunn, 1999; Gonzalez-R et al., 2010), as it minimises the total weighted completion time for the deterministic single-machine problem, as well as the expected sum of the weighted completion times for the stochastic single-machine problem in non-pre-emptive environments (Pinedo, 2012). Nevertheless, SPT does not adequately address maximum flow time or *VFT* (Eilon et al., 1975).

To overcome this limitation, three main approaches have been developed: (a) alternating and combining the use of sequencing rules, (b) considering information about actual work content of the following shop station to increase job flow, and (c) estimating the remaining flow time (*FT*) in shop for a job.

Some of the earliest proponents of Approach (a) were Conway and Maxwell (1962), who alternated between SPT and FIFO in different time periods to limit the time that jobs would wait to be processed. Other studies (Subramaniam et al., 2000; Singh et al., 2007) suggested selecting the best sequencing rule to use at certain points in time, depending on current shop conditions.

Approach (b) proposes a number of specifically designed rules to reduce *MFT* and *VFT*. For example, WINQ (the least total work in the queue of its next operation) was found (Haupt, 1989) to be efficient when combined with SPT and with a limit on the time that jobs wait in a queue, resulting in the PT+WINQ+AT (Processing time + WINQ + Arrival Time) and PT/TIS rules (Processing Time divided by Time in system) which perform well in hybrid flow-shop environments (Holthaus and Rajendran, 1997, 2002; Mizrak and Bayhan 2006).

Now, Approaches (a) and (b) simultaneously address multi-criteria problems in hyper-heuristic development (see, e.g., Korytkowski et al., 2013; Pickardt et al., 2013). Hyper-heuristic research focuses on the automatic creation of heuristics, i.e., combined dispatching rules that use different attributes, and the automatic assignment of those heuristics to the different work centres on the shop floor (Branke et al., 2016).

Some authors (Lu et al. 1994; Mittler and Schoemig, 2000; Mönch et al., 2013) also suggest Fluctuation Smoothing (FS) rules, or Approach (c), to reduce *MFT* and *VFT* by estimating the remaining *FT* of a job in complex environments.

Each of the three approaches has shown promise, but each also has drawbacks. For instance, Approach (a) has implementation issues since a continuously dynamic selection of sequencing rules can be as challenging to implement as a fixed schedule. Moreover, some proposals of Approaches (a) and (b) require a continuous evaluation of the shop-floor state to assess the best sequencing rule, or of load conditions in downstream stations, which can be challenging to implement since most companies do not possess updated and accurate information about the status of the shop-floor (Romero-Silva et al., 2016).

Finally, the practicality of Approach (c) is questionable, since not all companies have the capacity to create simulation models and perform constant simulation runs to build the iterative estimates of the remaining *FT* per station. This deficiency can also hinder the practicality of hyper-heuristics (Li et al., 2016).

### 2.2. Priority rules

The SEPT rule has been proven useful for single-server queues as applying the $c\mu$ rule (Baras et al., 1985; Buyukkoc et al., 1985) minimises the weighted mean number of customers in queue (Avrachenkov et al., 2010). This rule considers the associated waiting cost per customer and divides it by the mean service rate of each customer class. This rule is highly practical as its informational needs are minimal, requiring only the expected cost of waiting and the expected mean service time per class.

Buzacott and Shanthikumar (1993) proposed two "truncated" SEPT (2C-NP – Two Class Non-Preemptive, and the SPTT – Shortest Processing Time Truncated) rules with the objective of reducing both *MFT* and *VFT*. The 2C-NP rule separates the customers into two classes, depending on a mean service time 'threshold' termed as $\alpha$; thus, customers with a lower mean service time than $\alpha$ will have higher priority, while other customers will have lower priority. Likewise, the SPTT rule assigns a nominal mean service time of $\alpha$ to all customers with a higher mean service time than $\alpha$, while all other classes are served under a simple SEPT rule. The 2C-NP rule performed well in reducing both *MFT* and *VFT* (Shanthikumar