Waiting time based routing policies to parallel queues with percentiles objectives

Benjamin Legros

PII: S0167-6377(17)30523-0
DOI: https://doi.org/10.1016/j.orl.2018.04.001
Reference: OPERES 6354

To appear in: *Operations Research Letters*

Received date : 15 September 2017
Revised date : 3 April 2018
Accepted date : 5 April 2018

Waiting time based routing policies to parallel queues with percentiles objectives

Benjamin Legros

EM Normandie, Laboratoire Métis, 64 Rue du Ranelagh, 75016 Paris, France

benjamin.legros@centraliens.net

Abstract

We develop a method to obtain near-optimal routing policies to parallel queues with decisions based on customers' wait and performance objectives which include percentiles of the waiting time. We formulate and explicitly derive a value function where the waiting time is used as a decision variable. This allows us to apply a one-step policy improvement method to obtain an efficient routing solution. Numerical illustrations reveal that classical monotone policies are not always optimal.

Keywords: Markov decision process; performance evaluation; waiting time; one-step improvement; relative value function.

1 Introduction

Context and motivation. The historical problem of routing jobs/customers at arrival among a set of parallel queues to achieve some performance objective has received a high interest in the research literature. The motivation is the complexity of the theoretical problem together with its usefulness in practice. For computers application, communication network environments, or automatic call distributors in call centers, the possibility of determining efficient or optimal policies can lead to enhanced performance and cost reduction.

[23] is the first to show that the intuitive "Join the shortest queue" policy is optimal to minimize the expected sojourn time with identical exponential servers. Using a dynamic programming approach, [9] extend the analysis to different service rates and show that the "Shorter Queue Faster Server Policy" is optimal. When the exponential assumption for services is relaxed the above intuitive policies are no longer necessarily optimal. Counterexamples can be found in [22] and [14]. Several related articles focus on routing optimization, performance evaluation, for individual or social optimization, with observable queues or not (see, e.g., [8], [10], [3], [2]).

In the routing studies to parallel queues, either the system state is unknown or the given information is the number of jobs in the system. Thus, the number of jobs is the decision variable to build a deterministic policy. It provides optimal policies in contexts where the expected sojourn time has to be minimized. However, if the cost function is based on percentiles of the waiting time, the quantity in the system may no longer be the best decision variable. With a percentile of the waiting time, the system pays a penalty per