Optimal Server Allocation in Queues with Holding Cost and Customer Abandonment

Anna Park | Faculty Advisor: Dr. Hayriye Ayhan

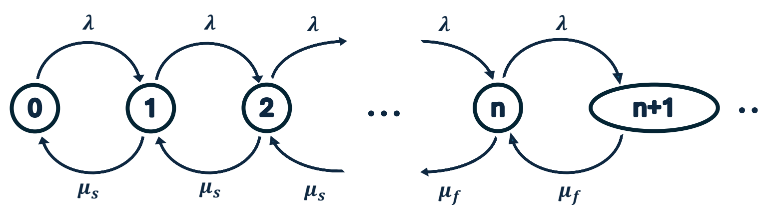
Georgia Tech College of Engineering
H. Milton Stewart School of Industrial and Systems Engineering

Abstract

We consider a single server infinite capacity Markovian queue. Our objective is to determine the optimal server allocation policy to minimize the long run average cost. There are two types of servers: a fast server and a slow server. It is more costly to hire the fast server. We consider two versions of the problem. In the first version, there is holding cost for customers in the system. In the second version, the customers waiting in line can get impatient and leave without being served incurring an abandonment cost in addition to the holding cost incurred by customers waiting in line. Our objective is to determine the dynamic allocation of the servers.

Notation

$\boldsymbol{\mu_f}$ (service rate of fast server), $\boldsymbol{\mu_s}$ (service rate of slow server), $\boldsymbol{c_f}$ (cost of using fast server), $\boldsymbol{c_s}$ (cost of using slow server), $\boldsymbol{\lambda}$ (arrival rate), $\boldsymbol{\theta}$ (abandonment rate), $\boldsymbol{h}$ (holding cost per unit time per customer).



Infinite Queue with Threshold *n*

$$h * \sum_{i=0}^{n=i} i\pi_i + \sum_{i=1}^{n-1} c_s\pi_i + \sum_{k=n}^{\infty} c_f\pi_k$$

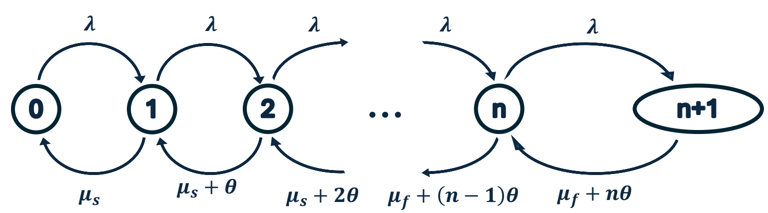
Expected Long-Run Cost in the Holding Cost Model

$$h * \sum_{i=0}^{n=i} i\pi_i + \sum_{i=1}^{n-1} c_s\pi_i + \sum_{k=n}^{\infty} c_f\pi_k + \sum_{k=2}^{\infty} c(k-1)\theta\pi_k$$

Expected Long-Run Cost in the Model with Holding Cost and Abandonments

Threshold Policy

It is known that in the holding cost model, the optimal policy is a threshold policy.

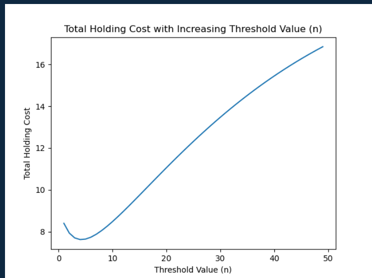


Infinite Queue with Threshold *n* and Abandonment

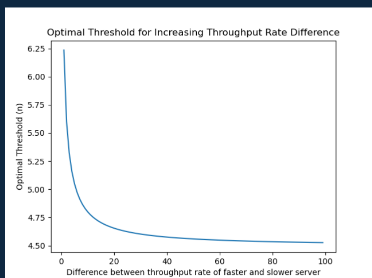
We compute the optimal threshold and see how it changes with respect to system parameters.

Holding Cost with Threshold Policy

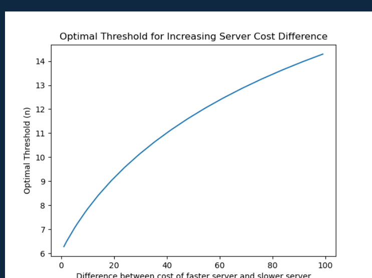
The following figures visualize the behavior of infinite queues and threshold values (n) with respect to changing parameters.



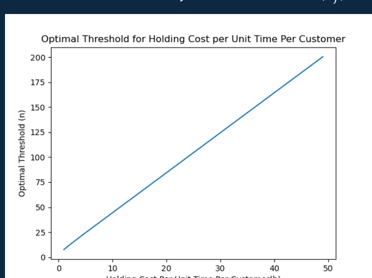
Total holding cost is calculated for various potential threshold values (n) of an infinite server when all other parameters ($\lambda, \mu_s, \mu_f, c_s, c_f$) remain constant.



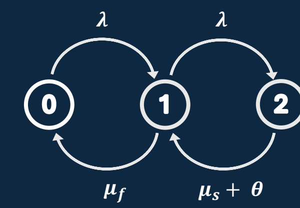
The threshold value (n) is calculated for an increasing difference between the faster and slower server. Only the faster server, μ_f , is manipulated to calculate this difference.



The figure above exhibits varying thresholds for increasing difference between the cost of the faster and slower server. Only cost of the faster server, c_f, is manipulated.



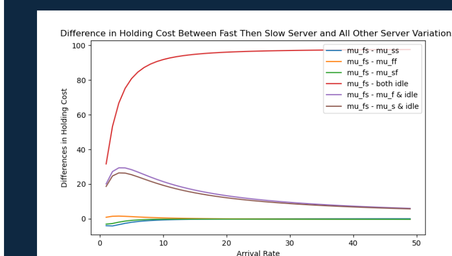
A range of potential threshold values (n) for an increasing holding cost per unit time per customer (h). All other variables are held constant.



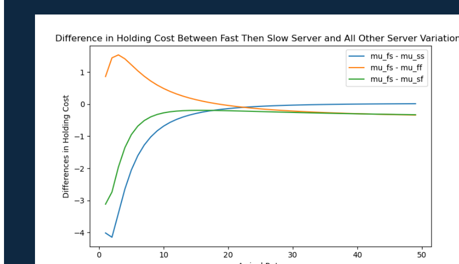
Ongoing Research

We start our study of systems with abandonments by considering a system that can hold at most 2 customers (N=2).

We consider three actions: use the fast server, use the slower server, idle both servers. By creating various configurations of the system using these three types of servers, we compare the differences in total holding cost by adjusting various parameters.



The figure above shows the difference between the fast then slow server and all other policy combinations mentioned in the previous paragraph. When this difference is positive, it denotes that the latter policy is optimal with respect to the arrival rate (λ). Negative difference in holding cost indicate that the counterintuitive, former policy (fast then slow server) is the optimal policy.



Taking a closer look at the queues without idle servers shows that it may be more optimal to use the fast/slow server policy with abandonment when certain conditions are met. Although brief, indication that this policy may be optimal is present for arrival rate between $\lambda = 20$ and $\lambda = 30$. We are currently in the process of testing for the validity of our findings and will continue researching this topic.