Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

OPTIMAL QUEUE LENGTH-BASED SERVER SHARING DECISIONS IN FIELD SERVICES

Saligrama Agnihothri

School of Management, Binghamton
University, State University of New York,
Binghamton, NY 13902, USA

Suman Niranjan

College of Business Administration, Savannah
State University,
Savannah, GA 31404, USA

ABSTRACT

We consider a field service system with equipment located in a geographic area. The area is divided into two territories, each with a single server who provides on-site service. The arrival of requests for service calls, the travel time to customer location, and on site repair time are all random variables. Since minimizing response time (defined as the time between equipment failure and the arrival of the repairperson to the site) is one of the primary objectives in field services, it is a common practice to re-deploy servers between territories to reduce large response times. In this paper, we consider a queue length-based, threshold type, redeployment policy. We use simulation to find an optimal server sharing policy so as to minimize the average response time. In particular, we explore the impact of demand arrival rates and travel time between territories on server sharing decisions.

1 INTRODUCTION

The quality of after-sales support directly impacts the success of companies producing complex equipment such as machines in medical electronics and office products. When customers have service level agreements (SLA), repair should be completed within a specified (guaranteed) time. It is well known from queueing theory that providing a low system delay requires low utilization of servers leading to high service cost. When there are multiple servers, each serving a heterogeneous customer group, the uncertainty in demand and service time leads to greater server workload imbalance. Sharing servers between territories when necessary is one way to reduce workload imbalance.

2 MODELING ASSUMPTIONS

- There are two service territories. Each territory has a single server who attends to service calls on site. Each server is assigned a home territory. Home territory of server i is T_i, i = 1, 2. The inter-arrival times of service calls are assumed to be independent and exponentially distributed random variables. The arrival rates of calls in territory λ_i is , $i = 1, 2$. Let $p_i = \lambda_i / (\lambda_1 + \lambda_2)$ = proportion of call arrivals to territory i, $i = 1, 2$. Define $\Lambda = \lambda_1 + \lambda_2$. Hence, $\lambda_i = \Lambda p_i$, $i = 1, 2$.
- We assume that on-site repair time is S (independent of territory), have general distribution with a mean of one. Travel time within a territory is negligible, and between territories is T and has a gamma distribution. Hence, total service time is S+T if a server has to travel between territories to complete a repair.
- We assume that service discipline is first-come-first-served. A two-threshold type policy is used to redeploy servers between the two territories. Threshold values for queue-length in territory i is (L_i, U_i) ,$1 \le L_i \le U_i$; $i = 1, 2$. For example, when the queue-length in T_2 increases to U_2, server 1 is in T_1 , and queue-length in T_1 is less than U_1, server 1 helps server 2 by travelling to T_2 and

978-1-4799-7486-3/14/$31.00 ©2014 IEEE

serving the head-of-the-line customer in T_2. Server 1 continues to serve customers in T_2 until either (a) queue length in T_2 drops below L_2 and there is at least one customer waiting in T_1 or (b) the queue length in T_1 reaches U_1. Note that when $L_i = U_i = 1, i = 1, 2$; whenever a server in T_1 is idle and there is at least 1 customer in T_2, server 1 travels to T_2 and serves the head-of-the line customer and vice versa.

- The performance measures of interest is aggregate average waiting time (down time) in the system which is defined as $WS(AG) = p_1*AWT_1 + p_2*AWT_2$ where AWT_i is Average Waiting Time in the system for customers in territory $i; i = 1, 2$.

- The system is in steady state.

3 INSIGHTS FROM SIMULATION

Note that a full sharing system could be modelled as *M/G/2* system. We consider several examples of full and partial sharing systems by varying threshold values and find conditions under which partial sharing system is optimal. Threshold type policies are beneficial because (i) customers whose current response time is nearing the threshold value should be given higher priorities so that their response time may not be significantly larger than the promised response time; and (ii) threshold policies minimize inter-territory travels which are unproductive. Note that as upper threshold limit increases, server sharing decreases. One of the insights obtained from this paper is that at low arrival rates (low server utilization), proportion of time a server is idle is high and hence inter-territory travel would not increase server utilization and customer delay. Hence, a full sharing system is optimal. However, at high arrival rates (high utilization) if servers are busy traveling and hence unproductive, the queue length and hence wait time increases significantly. Since partial sharing system reduces the inter-territory travels, it also reduces average wait time in the system. An example of the above insight is depicted in Figure 1 below.

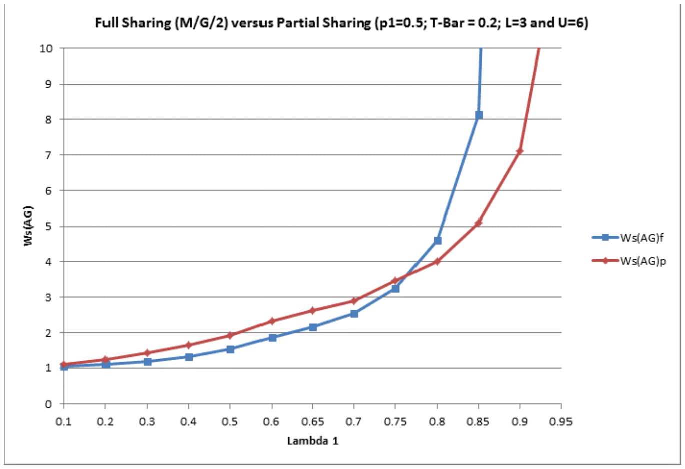

Figure 1: Comparing mean wait in the system of full and partial server sharing system.

4 CONCLUSION

In this research we investigate the impact of redeploying servers between two single-server territories and find an optimal server sharing policy so as to minimize the average response time. In particular, we study (i) optimal threshold limits under a given set of parameters, (ii) the impact of general distributions of inter-arrival time, service time, and travel time distributions on server sharing policy, and (iii) the optimal server sharing rules under asymmetric demand arrival rates and batch arrivals.