Proceedings of the 1990 Winter Simulation Conference
Osman Balci, Randall P. Sadowski, Richard E. Nance (eds.)

A SYSTEMS ANALYSIS AND MODEL OF A PARALLEL MULTI-SERVER QUEUEING SYSTEM

Thomas D. Clark, Jr.
Kathy L. McCommon

Department of Information
and Management Sciences
The Florida State University
Tallahassee, Florida 32306-1042

Donald H. Hammond

Department of Management
University of New Orleans
New Orleans, Louisiana 70148

ABSTRACT

A feasible set of management policies to handle multi-tandem queueing systems characterized by stochastic demand is presented in this paper. This type of system is common in many service related organizations. The research employs a previously developed simulation model of a State of Florida driver licensing office to manipulate policy variables in an effort to provide options for increased system efficiency. A key consideration is the ability to balance customer waiting time against idle server time. The use of simulation analysis enables the modeler to incorporate complex characteristics of this type of system into the model, therefore providing a realistic representation of possible management actions.

1. INTRODUCTION

The structure of our modern society could be characterized as an endless series of waiting lines and servers. Almost every interaction between the public and those who wish to provide either a free or commercial service involves a "waiting" process. This phenomenon has spawned numerous management science paradigms which attempt to find methods for allocating scarce resources to meet changing demands for that resource while optimizing one of the variables of the process. The optimized variable may be maximum profit in a commercial venture, it may be minimum cost in a public utility, or it may be today's most valuable resource: time.

Most of the analytical techniques thus far developed (mathematical programming) manipulate deterministic demand estimates or statistical averages of demand in an attempt to find the optimum expected return, an estimate of what may happen if all the probability functions are correct and remain unchanged for the duration of the time involved. However, when the demand is generated by a stochastic process such as when independent consumers arrive to buy or use a service, the analytical models either become too complex to solve or the assumptions become so simplifying that they remove all reality from the problem. In these situations, the modern computer has become the systems science laboratory where complex management systems can be modeled and manipulated through the techniques of computer simulation.

In the described study, the operations of an actual business office are modeled and analyzed to determine the most appropriate management procedures to handle stochastic demand. The goal of this research is to determine a set of policies applicable to a range of management systems.

The referent system for this research is a State of Florida driver licensing office, a system of thirteen parallel multiple server-multiple queue operations that provides motor vehicle operator licenses to the public. Originally analyzed and modeled by Clark [1986], a SLAM II [Pritsker 1984] model forms the basis for policy analysis. Utilizing variations of this model, Clark investigates fifty combinations of management policies affecting customer waiting time. An analysis of the variance in this performance variable demonstrates significantly shorter waiting times when the demand is dampened by utilizing a "block scheduling" appointment system. This study also reports a significant interaction between labor scheduling and job flexibility. These two factors become the focus of the follow-on experiments reported here. In the remainder of this paper the nature of service operation systems is discussed followed by a brief description of the system modeled, the research methodology, and the results of the study.

2. SERVICE OPERATING SYSTEMS

Service systems have a high degree of interaction with their environment. Customers are direct inputs from the environment that actually become a part of the system. For this reason, service facilities must be accessible to customers. This is generally interpreted as a requirement for small, decentralized facilities located in close proximity to areas of high demand [Fitzsimmons & Sullivan 1982]. What is actually required is a system with the ability to quickly adapt to accommodate varying demand.

Customers are an integral part of the system because production and consumption of the service occur simultaneously. For the licensing offices in this study, these services include such operations as personal information processing, testing, and photographing which all require the customer's presence. Due to the necessity of customer-server interaction, a service is considered a "perishable" commodity [Voss, Armistead, Johnson and Morris 1985]. For instance, if a server is not in demand for a period of time, that service capacity is lost. Service capacity cannot be stored and used at a later time. Instead, service capacity, facility utilization, and idle server time must be balanced against customer waiting time. To accomplish a satisfactory balance, capacity and demand must be matched as closely as possible.

Overall, service organizations are considered inherently more inefficient than manufacturing operations (where tangible products are produced) due to the uncertainty of customer demand [Chase and Tansik 1983]. Therefore, the management of resources, especially labor, is extremely important in a service operation. The two feasible strategies for dealing with demand fluctuations are to manage demand or to manage resources to match demand.

2.1 Managing Demand

One approach to demand management is to smooth the variability through the use of marketing tactics. Marketing promotions such as off-peak pricing discounts provide incentive for customers to shift from a high demand period to one with lower demand. These procedures generally smooth demand fluctuations, but do not completely level the pattern. Since a state license is a form of "use tax" collection, this approach may be more useful for motivating customers to pay within a thirty-day time frame than it would be for motivating them to arrive at the business office during low demand hours. For this reason, a marketing approach is not considered feasible for the referent office and is not included in this study.

Job scheduling takes a more forceful approach to the control of service demand by requiring a reservation or appointment system [Sasser 1976]. This approach has been successful for airline companies, lawyers, and many professional service providers. For other business operations such as medical care,

fine restaurants, and hairdressers, a combination of reservation/ appointment and walk-in service is more applicable. Although reservation systems are not feasible in situations such as emergency services, they may certainly be a feasible option for a routine service such as obtaining or renewing a civil license. Clark [1986] demonstrates that waiting time is significantly reduced when service arrivals are block scheduled into a uniform distribution.

If job scheduling techniques do not succeed in dampening the fluctuations and large demand variations persist, the feasible choices are limited. The system must be willing to tolerate the waiting lines which occur or adopt one or the following resource scheduling policies to adjust service capacity to demand.

2.2 Managing Resources

Labor scheduling is an approach to labor resource management that attempts to match server capacity to customer demand. To be effective, this approach requires an accurate demand forecast. Managers may choose to match capacity to demand forecasts using traditional schedules with overtime/ undertime, supplemental part-time schedules, shiftwork, or some form of alternative work schedule [Buffa 1983]. Alternative work schedules in Monday through Friday daytime operations may include flex-time (employee has some flexibility as to which hours are worked), and compressed work schedules (employees work a 40 hour week in four 10-hour days) [Thomas 1982].

Although traditional schedules combined with overtime/ undertime are most prevalent, the part-time component of national employment has been continually growing throughout the past decade [Smith 1986]. Employing part-time workers to meet fluctuating demand reduces the need to pay premium overtime wages, and may also reduce the error caused by work related fatigue [Werther 1976]. Disadvantages to part-time employment do exist, however. Extra employees result in an increased workload for both the personnel department and supervising management [Curry and Haerer 1981].

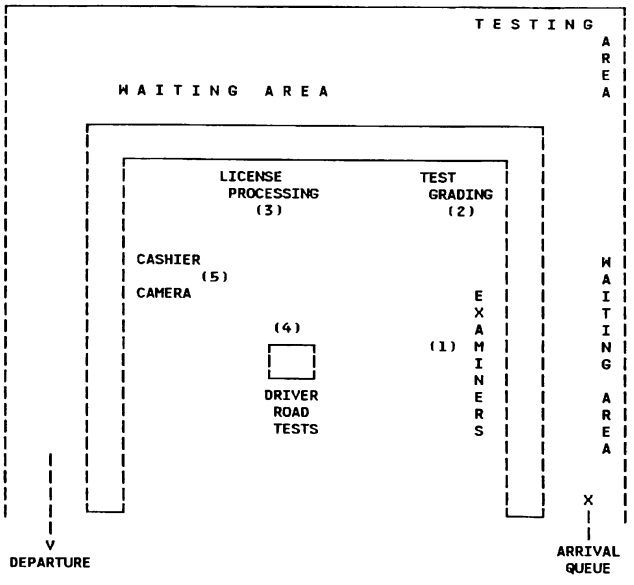
If part-time work is not feasible, flexibility for full time workers may be possible using alternative work schedules. Alternative work schedules permit the worker to have a choice in the time periods worked. In flex-time, the worker may choose how to schedule his or her eight hour day given certain time constraints for beginning work, taking lunch, and finishing work. Another alternative schedule, the compressed work week, allows the employee to work 40 hours in four 10-hour days, as opposed to five 8-hour days. Both types of alternative schedules reduce tardiness and absenteeism by providing the flexibility for employees to attend to personal business during normal working hours [Albright 1979; Kramer and Baily 1982; Cunningham 1981]. The availability of alternative scheduling is considered a major benefit in recruiting and retention of employees; and often leads to increased productivity, morale, and satisfaction [Curry and Haerer 1981]. Disadvantages to alternative schedules include planning, supervision and communication difficulties resulting from the variety of employee schedules within the organization [Thomas 1982].

Another resource management approach aimed at meeting varying consumer demand is job flexibility, the ability to adapt work assignments to unexpected demand shifts. If an employee is able to perform more than one type of service, the system will be more adaptable [Bracken, Calkin, Sanders and Thesen 1985]. When a bottleneck occurs in one area, an employee in another designated area with slack time is able to assist in the congested area to clear the backlog. For instance in the licensing office, if the line in front of the examination station is long, a cashier with idle time can help at the testing desk - if the cashier has been properly cross-trained. Training an employee in more than one work skill is a form of job enrichment which may increase employee productivity, satisfaction, job involvement, and internal motivation while decreasing absenteeism and turnover [Umstot, Bell and Mitchell 1976]. Labor scheduling and job flexibility policies are the focus of the two follow-on experiments in this study. A description of the actual system and the procedures for developing the model will be discussed in the next section.

3. SYSTEM AND MODEL DEVELOPMENT

An analysis of the various driver licensing offices in Florida was conducted [Clark 1986] to aid the Florida State Auditor General in evaluating staffing and management for the offices. The Concord office in Miami was selected because it well represented the problems faced by many of the offices statewide. At the time of the first study, the system was characterized by long waiting lines and extended service periods, both of which were growing steadily.

A diagram of the Concord office is shown in Figure 1. The basic flow of the system is typical of many bureaucratic offices. The hours of operation are from 7:00 AM until 6:00 PM with the servers taking rest and lunch breaks during the day. Customers arrive before the office opens and form a line at the door. The arrival process after opening has a non-stationary mean, causing the demand to vary with the time of day. The needs of the customer dictate one of the thirteen patterns through the office stations depending upon whether he or she needs information only, a written test, a renewed license, an original certificate, or a combination of these. All customers begin at station one where they gather information, complete forms, have their vision checked and, if necessary, receive a written test. Between station one and two, those taking the written test move to the testing area, returning to the counter at station two when finished. At station two the tests are graded and oral testing is administered. Personal identification data are typed onto blank license certificates at station three, and customers requiring driving tests wait at station four for an examiner. When all processing is completed, customers are photographed and pay fees at station five. After this station, each customer exits; however, many customers exit between stations due to test failures, termination of requirements, or simply due to time conflicts.



Station (1). Examiners greet applicants, determine eligibility, complete license renewal forms, check vision, distribute written test forms, and handle financial responsibility cases.
Station (2). Examiners grade written tests and assign road and oral tests.
Station (3). Examiners process license reinstatements and type original license forms.
Station (4). Examiners administer road tests.
Station (5). Examiners operate the cash register and camera.

Figure 1. The Driver License Office

At the Concord station, four examiners work at station one, one examiner works at station two, two examiners type licenses at station three, two examiners administer outside driving tests, and one examiner is the cashier and camera operator. Work shifts are staggered to provide coverage for the eleven hours of operation. Sufficient job flexibility exists to permit coverage of all positions during rest and lunch breaks.

In this discrete event simulation, a customer arrival is generated according to the applicable Poisson distribution. A transaction type and station visitation pattern are then determined and assigned. The customer entity cycles through the system network obtaining service or waiting in a queue when a server is not available. Once the customer is paired with a server, the length of service is determined from the appropriate time distribution. Data are collected on the length of each queue and the time required to transit the system. The model validation is based on a comparison of its behavior with the actual system data collected [Clark 1986].

Clark's initial experimentation with the model suggests threefactors that materially affect the processing of customers. As discussed above, the pattern of customer arrivals (job scheduling), the scheduling of examiner work shifts and the allocation of examiners to work stations (labor scheduling), and the manner in which examiners assist one another (job flexibility). These management policy approaches are selected for separate investigation, and modifications to the basic model represent these policies.

4. RESEARCH DESIGNS AND RESULTS

Experiment 1 is a single factor design with four levels. This test compares three alternative labor scheduling methods to the existing labor schedule. The null hypothesis states that non-traditional (alternative) labor scheduling will have no effect on customer waiting time. The alternative schedules include a compressed work week schedule as level one, flex-time work schedules with part-time augmentation as level two, and a combination of traditional schedules and part time workers as level three. The original labor schedule at the Concord office is tested as level four. The performance variable for this experiment is the mean "time in the system" or the average time a customer would expect to spend transiting the office. Although all of the alternative schedules produce lower transit times, no labor scheduling technique examined produced a significant performance variance. Therefore, the null hypothesis is not rejected. See Figure 2.

ANALYSIS OF VARIANCE ON AVERAGE TIME IN SYSTEM

SOURCE	DF	SS	MS	F	p
SCHEDULE	3	339.7	113.2	1.33	0.277
ERROR	44	3747.9	85.2		
TOTAL	47	4087.6			

LEVEL	N	MEAN	STDDEV
1	12	51.575	8.425
2	12	54.633	7.619
3	12	55.900	8.045
4	12	58.992	12.123

POOLED STDDEV = 9.229

Figure 2. One-way ANOVA Results Experiment One

Experiment 2 is a two factor design which also uses mean "time in system" as the performance variable. The first factor is the arrival pattern as described by Clark [1986]: the current pattern and a uniform pattern as might be expected with block scheduling. The second factor represents degrees of job flexibility. See Figure 3.

Factor 1. Levels of Arrival Patterns.
- a. The current arrival pattern (control group)
- b. A uniform arrival pattern (block scheduling jobs).

Factor 2. Levels of Job Flexibility
- a. The current staffing (10) with present assistance patterns.
- b. Seven workers pooled at station 1, Drive testers assisted until required at station 4. (Moving service)
- c. The seven pooled workers from station 1 also gave driving tests when no line existed at station 1. (Moving service)
- d. Same as level b. except that one station 1 worker moved to the Drive Tester pool after the lunch break. (Moving Service)
- e. All ten workers were pooled as general purpose workers at station 1. (Moving service)

Figure 3. Factor Levels for Experiment 2

The basic management policy examined in Experiment 2 is the moving server concept. This method of paring customers with servers is thought to produce a more efficient use of the labor resources. This is accomplished by having the server/customer pair move as far through the station network as possible, leaving the customer alone only when the customer is performing an individual act such as a written examination. A server returning to the pool takes a customer from the most advanced queue each time, thus clearing any bottleneck that would prevent a smooth flow through the system. The Null Hypothesis is that moving service would have no effect on system transit time. This hypothesis was easily rejected.

To reconfirm the results shown in the earlier experiments with this model [Clark 1986], a one-way analysis of variance was conducted using the transit times and arrival patterns. The results (Figure 4) demonstrate that the arrival pattern significantly affects the time in transit regardless of the labor patterns tested.

ANALYSIS OF VARIANCE ON AVERAGE TIME IN SYSTEM

SOURCE	DF	SS	MS	F	p
ARRIVAL	1	505	505	4.78	0.031
ERROR	118	12468	106		
TOTAL	119	12973			

LEVEL	N	MEAN	STDDEV
Uniform	60	44.33	9.75
Current	60	48.44	10.78

POOLED STDDEV = 10.28

Figure 4. One-way ANOVA Results Experiment Two

The results of a two-way ANOVA (Figure 5) indicate a significant difference in the mean transit time of a customer attributable to both the arrival pattern and to the labor assignment patterns. Since the two-way ANOVA does not show which models cause the variance, a one-way ANOVA with a Scheffe multiple contrasts test is performed. This test groups the eight moving server models as different from the fixed server models at the p = .05 level. This analysis indicates large reductions in the average time required to transit the system where labor

TWO-WAY ANALYSIS OF VARIANCE - MEAN TIME IN SYSTEM BY WORK ASSIGNMENT TYPE AND ARRIVAL PATTERN.

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Main Effects	7370.308	5	1474.062	29.261	.000
Work Assignment Type	6865.188	4	1716.297	34.070	.000
Arrival Pattern	505.120	1	505.120	10.027	.002
2-way Interactions	60.962	4	15.241	.303	.876
TYPE x Arrival	60.962	4	15.241	.303	.876
Explained	7431.270	9	825.697	16.391	.000
Residual	5541.397	110	50.376		
Total	12972.667	119	109.014		

Figure 5. Two-way ANOVA Results Experiment Two

resources are cross-trained, pooled, and move with the customer. Indeed, all of the moving server models exhibit significantly better performance than the present staffing model, and this is true for both customer arrival patterns.

5. CONCLUSIONS

Clark's initial analysis of fifty different policies indicates that more efficient performance could be obtained if the arrival rate fluctuations could be dampened. In fact, of the initial data set (54 policies), 65 percent of the policies that yield significantly better performance measures have deterministic arrivals as a common attribute. Follow-on experiments replicate the result that deterministic customer arrival rates produce significantly shorter mean transit times. An appointment or block scheduling system would improve the customer service efficiency of this office and reduce waiting time.

The comparison of three alternative work schedules with the current traditional work schedules demonstrates that all four produce approximately the same transit times. Since these management policies may produce better morale and productivity among the labor force without significant increases in transit time, they appear to be good candidates for labor management policies.

The dramatic results produced by the moving server models indicate that this method may provide the most efficient pairing of server resources with customer demand. A significant amount of time is saved by simply pooling the cashier, test grader, and license preparers with the first station personnel, and having these servers move through the stations with the customer. Modifying this assignment model to require idle station one examiners to occasionally assist in driving tests saves even more time. Both of these situations are not great departures from the current practice of covering all stations while workers are at lunch or on a rest break. The policy of serving the most advanced customer first helps expedite the transit of those being served and minimizes the time required waiting in queues.

While the specific results of this inquiry may not be universally applicable, the situation is general enough to have applications beyond the specific models and this referent office. The general decrease in flow time produced by a leveled demand pattern and by the moving server with advanced station priority may certainly be realized in other queuing systems such as fast food service, university registration, and most license office situations.

REFERENCES

Albright, B. (1979), "The Joys of Flex Time for Both Managers and Operators," *Word Processing World 6*, 2, 61.

Bracken, J., J. Calkin, J. Sanders, and A. Thesen (1985), "A Strategy for Adaptive Staffing of Hospitals under Varying Environmental Conditions," *Health Care Management Review 10*, 4, 43-53.

Buffa, E.S. (1983), *Modern Production/Operations Management*, Seventh Edition, John Wiley & Sons, Inc., New York, NY.

Chase, R.B. and D.A. Tansik (1983), "The Customer Contact Model for Organization Design," *Management Science 29*, 9, 1037-1050.

Clark, T.D., Jr. (1986) "A Systems Analysis and Model of Driver Licensing in the State of Florida," in *Proceedings of the 1986 Winter Simulation Conference*, J.R. Wilson, J.O. Henrikson, and S.D. Roberts, IEEE, Piscataway, NJ, 842-849.

Cunningham, J.B. (1981) "Exploring the Impact of a Ten-Hour Compressed Shift Schedule," *Journal of Occupational Behaviour 2*, 3, 217-222.

Curry, T.E., Jr. and D.N. Haerer (1981), "Flexi-Time: Is It For You?," *Public Relations Journal 37*, 3, 54-57.

Fitzsimmons, J.A. and R.S. Sullivan (1982), *Service Operations Management*, McGraw-Hill, Inc., New York, NY.

Kramer, O.P. and R.L. Bailey (1982), "Flexible Work Schedules Reap Financial Returns," *Modern Office Procedures*, 110-122.

Pritsker, A.A.B. (1986), *Introduction to Simulation and SLAM II*, Third Edition, Halsted Press, New York, NY.

Sasser, W.E. (1976), "Match Supply and Demand in Service Industries," *Harvard Business Review 54*, 6, 133-140.

Smith, S.J. (1986), "The Growing Diversity of Work Schedules," *Monthly Labor Review 109*, 11, 7-13.

Thomas, E.G. (1982), "Update on Alternative Work Methods," *Management World 11*, 1, 30-32.

Umstot, D.D., C.H. Bell, Jr., and T.R. Mitchell (1976), "Effects of Job Enrichment and Task Goals on Satisfaction and Productivity: Implications for Job Design," *Journal of Applied Psychology 61*, 4, 379-394.

Voss, C., C. Armistead, B. Johnson, and B. Morris (1985), *Operations Management in Service Industries and the Private Sector*, John Wiley and Sons, Inc., New York, NY.

Werther, W.B. (1976), "Mini-Shifts: An Alternate to Overtime," *Personnel Journal 55*, 3, 130-133.