

A Practical View of Queues with Lane Switching

Coleen R. Wilder

Nick T. Thomopoulos

Stuart School of Business

Illinois Institute of Technology

565 W. Adams Street, Chicago, Illinois 60661

crwilder@yahoo.com, thomop@stuart.iit.edu

Presented at the MBAA International Conference on March 24, 2010 in Chicago

Recipient of the Distinguished Paper Award in the OMEA Category (Operations Management & Entrepreneurship)

Abstract

This paper is concerned with the analysis of queuing systems when there are two distinct populations. Its pragmatic nature is both intentional and far-reaching. Two significant contributions concern configurations with arrival dependent service rates, one of which examines lane switching objectives. Linear algebra and calculus are the primary disciplines used to generate various performance metrics.

Introduction

The focus of this research is to examine a class of queuing problems frequently found in real world applications. The primary value of which is to aid practitioners in the design and management of queues. The class of problems to be addressed herein are those in which the customer population can be separated into two distinct groups, which will be referred to as type-one and type-two customers. The criteria used to define the two groups is immaterial to this study; a requirement of the partitioning, however, is that there is a difference in either the arrival rate or service rate or both rates.

When there is more than one service facility to service the two customer groups, a manager is confronted with the challenge of choosing the best allocation of customer groups to service facilities. Three options are to be explored and compared to a base case. Standard performance measures are generated for selected scenarios to assess the tradeoffs between various configurations.

This paper is based in part on the forthcoming Ph.D. dissertation and research, *The Queuing Theory of Two Populations with Lane Switching* [1]. The dissertation includes a full set of tables along with the underlying methodologies for their creation. These efforts are an extension of previous work conducted by Thomopoulos [2] and Plumchitchom [3] in which tables are produced for specific queuing configurations.

The value of these tables cannot be undermined. The inspiration for this study was from a project undertaken over a decade ago for a large steel manufacturer. One of its Basic Oxygen Furnaces (BOF) was in the capital funding stages of getting a new chemical laboratory. A critical step in the steelmaking process is to test samples of molten steel from suspended ladles.

Processing cannot continue until the test results are known. The test samples were to be sent to the new lab through pneumatic tubes and the testing performed in the order in which the samples arrived. Current plans included one machine to perform the tests due to its million dollar price tag. Back-ups in the lab, however, caused delays which were equated to financial losses and too many delays resulted in ruined careers. The lead chemist wanted to know the probability that a sample would wait for testing and the expected time spent waiting for service. There were two types of tests performed, each with their respective processing times. If a second lab machine was added to the proposal, without solid numbers to justify the additional million-dollar expenditure, it risked being denied. Should a second machine be included in the proposal? Should it have a dedicated pneumatic tube? Could the two machines share a tube (queue)? Was there an option between the two that offered the same or better service? An answer was needed in two weeks. Instead the company paid for an analyst to crunch numbers for two weeks providing the best answers in the given time frame. The model found herein would have been able to answer all the questions from the lead chemist in a manner of minutes.

Any establishment that creates queues should view the performance tables in the cited works as a source for competitive advantage. Managers will be able to cut analyst costs, decrease lead time, and may also achieve a higher level of customer satisfaction by reducing the time customers wait for service. These tables give practitioners the necessary tools to predict expected performance metrics as well as to gain insights from various queuing configurations.

Model Description

The following four queuing systems (A through D) will examine the dynamics of queues that service two distinct populations each with their own set of independent parameters. The interarrival and service times are both exponentially distributed with respective means of $1/\lambda$, and $1/\mu$. All arrivals are equal in status; there are no priority rules in practice. By design, the arrival rate is less than the service rate to ensure stability; this restriction satisfies the condition $\rho = \lambda/\mu < 1$.

(A) Base Case

This configuration considers the case of only one service facility. Although it is a classic example that appears in most texts on queuing theory, it is included to serve as a point of reference. An example of this configuration is the current proposal for the BOF chemical lab: two types of samples (populations), one lab machine, and each sample service time is exponentially distributed. Since the two populations are merged into one, their combined service distribution becomes a hyper-exponential distribution. Using Kendall's notation the system is a M/G/1 system for which closed form equations are available.

(B) Shared Queues

This configuration merges the two populations into one queue. The two service facilities operate in parallel and each is capable of serving either type of arrival. Customers enter the queue when both facilities are occupied and the next customer in line will be serviced by the next available server. A classic example of this configuration is a post office with two clerks and one common line. The first customer in line is serviced by the next available clerk; a customer is able to choose equally between clerks when both are available.

State probabilities are defined by p_{nij} where n denotes the number of units in the system, i denotes the type of customer in facility one, and j the type of customer in facility two; valid

4. Using the unity condition

$$\sum_{n=0}^k P_n = 1 = P_{000} + P_{101} + \dots + P_{322}$$

and substitution, a value can be found for P_{000} as follows:

$$P_{000} + q_{101} \cdot P_{000} + \dots + q_{322} \cdot P_{000} = 1$$

$$P_{000}(1 + q_{101} + \dots + q_{322}) = 1$$

$$P_{000} = \frac{1}{(1 + q_{101} + \dots + q_{322})}$$

5. The probability there are k units in the system is $P_k = P_{k,11} + P_{k,12} + P_{k,21} + P_{k,22}$. If P_k is larger than the tolerance, a larger k is chosen and the previous steps are repeated; once P_k is within the selected tolerance the algorithm stops.

(C) Dedicated Queues

This configuration dedicates a queue to each population with no sharing of service facilities. An example of this configuration is an airline with separate check-in lines for clients flying first-class and coach. Agents working the first-class counter are reserved for first class customers only. If no one is in line, they are not allowed to service customers from coach since it may delay service to a first-class customer. Performance metrics for each channel are generated the same as for the standard M/M/1 queue. An overall system metric is calculated as the weighted average of the two independent queues. Although this configuration is not interesting in its own right, it is included for comparison purposes.

(D) Mixed Queue, Arrival dependent service times

A mixed queue is defined in this study as a queue in which arrivals from an originating queue switch to another queue and are merged or mixed with arrivals from the destination queue. An example of this configuration is a small grocery store with an express line for customers with less than ten items and another line for any customer. Customers with one item take the same amount of time to service regardless of the register they use. This change converts the service time distribution for the general lane from an exponential to a hyper-exponential distribution. It should be noted that in this study the service rate for type-one arrivals may be greater than, less than, or equal to the service rate for type-two arrivals. Type-one arrivals are allowed to switch queues with the objective of equalizing the workload between the two servers. For reference purposes, the first channel will be referred to as the specialized queue and the second as the general queue.

The challenge to calculating performance metrics for lane switching models is to quantify the frequency at which customers switch lanes. An obvious circumstance under which customers switch lanes is when a type-one arrival moves to a general lane after observing their facility is occupied and the general facility is available. This event, however, is not the only circumstance under which a type-one arrival will switch to a general lane. One strategy arrivals may use is to balance the workload between the two facilities; in this manner, an arrival makes an assessment based on the percentage of time the general lane is busy. The performance metric that incorporates both the arrival and service rates is the utilization factor, ρ ; the effective workloads of each facility are equalized by transferring the unbalanced workload from the specialized queue to the general queue. A key assumption of this methodology is that customers

are able to make good decisions when they have the option to balance the workload between two service facilities.

Results

Consider the dilemma posed by the lead chemist in the introduction. Suppose the combined workload for the two types of samples generate a value for ρ of 0.7 ($\rho_1 + \rho_2 = 0.7$); in other words, the current system is moderately busy. Table 1 lists various combinations of arrival and service rates that produce the desired workload. Subscripts one and two refer to type-one and type-two samples respectively. Arrival and service rates without subscripts are the average for the two types. The ratio is the ratio of $\rho_1:\rho_2$. Suppose again that the ratio of the two workloads is 3:2; in other words, type-one samples generate more of a workload than type-two samples. The service time for type-one samples is used as a base. For example, let the service time for type-one samples equal one minute. The service time for type-two samples then is 1.5 minutes since

$$\tau_{s_2} = 1/\mu_2 = \frac{1}{2/3} = 1.5$$

where τ_{s_2} is the time to service one customer and μ_2 is the service rate for a type-two customer. Using this combination, server one therefore is faster than server two.

Table 1: Input Parameters for $\rho = 0.7$

| λ_1 | μ_1 | ρ_1 | λ_2 | μ_2 | ρ_2 | λ | μ | % type1 | % type2 | ratio |
|-------------|---------|----------|-------------|---------|----------|-----------|-------|---------|---------|-------|
| 0.350 | 1.000 | 0.350 | 0.700 | 2.000 | 0.350 | 1.050 | 1.500 | 33% | 67% | 1:1 |
| 0.467 | 1.000 | 0.467 | 0.233 | 1.000 | 0.233 | 0.700 | 1.000 | 67% | 33% | 2:1 |
| 0.525 | 1.000 | 0.525 | 0.117 | 0.667 | 0.175 | 0.642 | 0.917 | 82% | 18% | 3:1 |
| 0.560 | 1.000 | 0.560 | 0.070 | 0.500 | 0.140 | 0.630 | 0.900 | 89% | 11% | 4:1 |
| 0.233 | 1.000 | 0.233 | 0.933 | 2.000 | 0.467 | 1.167 | 1.667 | 20% | 80% | 1:2 |
| 0.350 | 1.000 | 0.350 | 0.350 | 1.000 | 0.350 | 0.700 | 1.000 | 50% | 50% | 2:2 |
| 0.420 | 1.000 | 0.420 | 0.187 | 0.667 | 0.280 | 0.607 | 0.867 | 69% | 31% | 3:2 |
| 0.467 | 1.000 | 0.467 | 0.117 | 0.500 | 0.233 | 0.583 | 0.833 | 80% | 20% | 4:2 |
| 0.175 | 1.000 | 0.175 | 1.050 | 2.000 | 0.525 | 1.225 | 1.750 | 14% | 86% | 1:3 |
| 0.280 | 1.000 | 0.280 | 0.420 | 1.000 | 0.420 | 0.700 | 1.000 | 40% | 60% | 2:3 |
| 0.350 | 1.000 | 0.350 | 0.233 | 0.667 | 0.350 | 0.583 | 0.833 | 60% | 40% | 3:3 |
| 0.400 | 1.000 | 0.400 | 0.150 | 0.500 | 0.300 | 0.550 | 0.786 | 73% | 27% | 4:3 |
| 0.140 | 1.000 | 0.140 | 1.120 | 2.000 | 0.560 | 1.260 | 1.800 | 11% | 89% | 1:4 |
| 0.233 | 1.000 | 0.233 | 0.467 | 1.000 | 0.467 | 0.700 | 1.000 | 33% | 67% | 2:4 |
| 0.300 | 1.000 | 0.300 | 0.267 | 0.667 | 0.400 | 0.567 | 0.810 | 53% | 47% | 3:4 |
| 0.350 | 1.000 | 0.350 | 0.175 | 0.500 | 0.350 | 0.525 | 0.750 | 67% | 33% | 4:4 |

Using Table 2 under the current proposal that both types of lab samples share the same machine, the lead chemist should expect there to be on average 1.7 samples waiting for service at any given time. Performance improvements for each configuration are found in their respective columns and, not surprisingly, the shared configuration has the smallest expected queue length of 0.227. Another factor not previously mentioned in the introduction is that the lab technicians are union employees; each technician has a job description that specified the tests they are allowed to perform. With this constraint, the lead chemist should expect a queue length 0.413 samples.

The mixed configuration's performance is between that of the prior two configurations but it allows some specialization of technicians which may be acceptable to the union representatives.

Table 2: Queue Length Comparison for two populations with $\rho = 0.7$

| | One Server | Two Servers, Dedicated | | | Two Servers, Shared | Two Servers, Mixed | | |
|-------|------------|------------------------|----------|----------|---------------------|--------------------|-----------|-----------|
| ratio | L_q | L_q | L_{q1} | L_{q2} | L_q | L_{qe} | L_{q1e} | L_{q2e} |
| 1:1 | 1.838 | 0.377 | 0.188 | 0.188 | 0.234 | 0.377 | 0.188 | 0.188 |
| 2:1 | 1.633 | 0.479 | 0.408 | 0.071 | 0.222 | 0.377 | 0.188 | 0.188 |
| 3:1 | 1.684 | 0.617 | 0.580 | 0.037 | 0.226 | 0.385 | 0.188 | 0.196 |
| 4:1 | 1.764 | 0.736 | 0.713 | 0.023 | 0.231 | 0.400 | 0.188 | 0.211 |
| 1:2 | 1.815 | 0.479 | 0.071 | 0.408 | 0.232 | 0.479 | 0.071 | 0.408 |
| 2:2 | 1.633 | 0.377 | 0.188 | 0.188 | 0.222 | 0.377 | 0.188 | 0.188 |
| 3:2 | 1.699 | 0.413 | 0.304 | 0.109 | 0.227 | 0.382 | 0.188 | 0.193 |
| 4:2 | 1.815 | 0.479 | 0.408 | 0.071 | 0.236 | 0.398 | 0.188 | 0.209 |
| 1:3 | 1.786 | 0.617 | 0.037 | 0.580 | 0.230 | 0.617 | 0.037 | 0.580 |
| 2:3 | 1.633 | 0.413 | 0.109 | 0.304 | 0.222 | 0.413 | 0.109 | 0.304 |
| 3:3 | 1.701 | 0.377 | 0.188 | 0.188 | 0.227 | 0.377 | 0.188 | 0.188 |
| 4:3 | 1.833 | 0.395 | 0.267 | 0.129 | 0.238 | 0.388 | 0.188 | 0.200 |
| 1:4 | 1.764 | 0.736 | 0.023 | 0.713 | 0.229 | 0.736 | 0.023 | 0.713 |
| 2:4 | 1.633 | 0.479 | 0.071 | 0.408 | 0.222 | 0.479 | 0.071 | 0.408 |
| 3:4 | 1.700 | 0.395 | 0.129 | 0.267 | 0.227 | 0.395 | 0.129 | 0.267 |
| 4:4 | 1.838 | 0.377 | 0.188 | 0.188 | 0.238 | 0.377 | 0.188 | 0.188 |

Another factor of interest to the lead chemist is the expected time a sample would spend in the queue. These results are given in Table 3. The units for the time are dependent on the units from the input parameters. Using the example for Table 1, with the current proposal, samples are expected to wait 2.8 minutes. An extra machine with the worst configuration reduces the expected wait time by a factor of four.

Table 3: Comparison of time in the queue for two populations, $\rho = 0.7$

| | One Server | Two Servers, Dedicated | | | Two Servers, Shared | Two Servers, Mixed | | |
|-------|------------|------------------------|----------|----------|---------------------|--------------------|----------|----------|
| ratio | w_q | w_q | w_{q1} | w_{q2} | w_q | w_{qe} | w_{q1} | w_{q2} |
| 1:1 | 1.750 | 0.359 | 0.538 | 0.269 | 0.223 | 0.359 | 0.538 | 0.269 |
| 2:1 | 2.333 | 0.685 | 0.875 | 0.304 | 0.317 | 0.538 | 0.538 | 0.538 |
| 3:1 | 2.625 | 0.962 | 1.105 | 0.318 | 0.352 | 0.600 | 0.538 | 0.673 |
| 4:1 | 2.800 | 1.167 | 1.273 | 0.326 | 0.367 | 0.634 | 0.538 | 0.754 |
| 1:2 | 1.556 | 0.411 | 0.304 | 0.438 | 0.199 | 0.411 | 0.304 | 0.438 |
| 2:2 | 2.333 | 0.538 | 0.538 | 0.538 | 0.317 | 0.538 | 0.538 | 0.538 |
| 3:2 | 2.800 | 0.681 | 0.724 | 0.583 | 0.374 | 0.630 | 0.538 | 0.754 |
| 4:2 | 3.111 | 0.822 | 0.875 | 0.609 | 0.405 | 0.682 | 0.538 | 0.897 |
| 1:3 | 1.458 | 0.504 | 0.212 | 0.553 | 0.188 | 0.504 | 0.212 | 0.553 |
| 2:3 | 2.333 | 0.590 | 0.389 | 0.724 | 0.317 | 0.590 | 0.389 | 0.724 |

| | | | | | | | | |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| 3:3 | 2.917 | 0.646 | 0.538 | 0.808 | 0.389 | 0.646 | 0.538 | 0.808 |
| 4:3 | 3.333 | 0.719 | 0.667 | 0.857 | 0.432 | 0.706 | 0.538 | 1.000 |
| 1:4 | 1.400 | 0.584 | 0.163 | 0.636 | 0.182 | 0.584 | 0.163 | 0.636 |
| 2:4 | 2.333 | 0.685 | 0.304 | 0.875 | 0.317 | 0.685 | 0.304 | 0.875 |
| 3:4 | 3.000 | 0.697 | 0.429 | 1.000 | 0.400 | 0.697 | 0.429 | 1.000 |
| 4:4 | 3.500 | 0.718 | 0.538 | 1.077 | 0.454 | 0.718 | 0.538 | 1.077 |

Conclusion and Remarks

The lead chemist now has the numbers to compare the purchase price for an additional machine to the delay costs. The table values are not surprising and confirms intuition; a shared configuration with two machines performs best. The quandary, however, is typically over the magnitude of improvement; just how much better is one configuration over the other? Real world applications don't always lend themselves to choosing the best configuration. Customer perceptions, budget constraints, union rules, and many other factors overrule what would seem to be a cogent argument.

Research to expand the configurations and performance metrics is in progress. Of particular note is a configuration analogous to the mixed queue with arrival dependent service times in which all other factors are equal, except that the service times are a function of the service facility and not an arrival.

References

- [1] Wilder, Coleen., *The Queuing Theory of Two-Populations with Lane Switching*. Work-in-progress Ph.D. dissertation. Chicago, IL: Illinois Institute of Technology Stuart School of Business, 2009.
- [2] Thomopoulos, Nick T., *Strategic Inventory Management and Planning*, Carol Stream, IL. Hitchcock Publishing Company, 1990.
- [3] Plumchitchom, Noin., *The Queuing Theory of the Erlang Distributed Interrival and Service Time*, 2006. Academic Source Premier. EBSCO. IIT Galvin Library. 15 June 2009 <<http://web.ebscohost.com.ezproxy.gl.iit.edu/>>