Staffing Requirements Based on Stochastic Arrivals and Infinite Server Model (M/M/s)

Abstract

The optimal staffing scheme is to allocate the minimum number of servers required to maintain a certain standard of service. This article presents the methodology used in creating a model that would find the best accommodation between the demand and the feasibility of the staffing schedule. However, the results are not perfectly optimal due to the stochastic nature of call arrival and the non-stationary nature of servers.

Special Thanks to: Professor Ingolfsson and Fernanda Campello, University of Alberta

Shuang E
Derek Hewitt
Jenny Morrow
Scott Van Bolhuis
EXECUTIVE SUMMARY
Currently CCC runs on a simple staffing model that was introduced early on in the existence of its operations. The current model was adopted to forecast the expected workload throughout a typical week; however the calculations were not overly robust and cannot account for the high variability of demand that a typical EMS centre would see. These high fluctuations had a negative impact on staffing utilization thus contributes a less than optimal staffing scheme. Program managers required the ability to properly plan shift schedules as well as to understand the impacts of alternative workflows. It was our job to create a model that attends to the needs of the managers and include the optimal staffing schedule in every situation and provide a user friendly interface that the managers can use in the future to alter the staffing level as CCC takes on extra demand from the other communities.

Our analysis and scope included:

- Analyzing and mapping the process flow
- Examining call center telephony and CAD data
- Analyzing arrival and service rate
- Developing a queuing model
- Integrating the analysis result onto a queuing model using Markovian queuing theory
- Simulating the results

This project was unique due to the vicissitudes of demand after the Edmonton based Central Communication Center taking over Edmonton Emergency Response Communication Center and Strathcona County Emergency Services. CCC is currently coordinating Edmonton metro inter facility transfer and medical flight transportation.

This paper presents the optimal staffing levels to achieve the tasks of CCC without violating regulations and service quality requirements.

BACKGROUND
A common feature of many call centre service systems is that the demand for service often varies greatly by time of day and that the length of identical service are not constant. In this report we discuss ways to cope with the varying demand when setting staffing and scheduling requirements.

Setting staffing requirements is one in a hierarchy of decisions that must be made in the design and management of a service system. In a long term planning horizon, managers set the
system capacity of the call centre and work around the number of available service positions as well as the capacity of the supporting equipment. The daily staffing decision specifies the number of service representatives needed to work during each staffing interval over the day. After the staffing requirements are set, managers make scheduling decisions, specifying the number of service agents to work on specific periods, in conformance with the previously determined staffing levels. Managers often make further adjustments called flexing decisions, which move servers in and out of the line of duty, and if this flexibility can be achieved, it is often possible to provide a high quality of service.¹

Service systems can also have several other characteristics. In light of the call centre service system that will be focused on, the service must be performed relatively soon after the request is made. In this report you will find the amount of servers required to meet specific standards such as answering 95 percent of calls within 5 seconds, 10 seconds, and 15 seconds. With higher rates of service, obviously more servers will be required.

Using the telephony and computer-aided-dispatch data from various call centres throughout the province we are able to construct a model to satisfy the staffing and scheduling demands of the Central Communications Centre in Edmonton. Our model will be focused on a basic queuing model (M/M/s) to determine the minimum servers required to meet certain standards mentioned above. The model will account for certain inputs such as the inter arrival time, service time, number of servers, system capacity, calling population, and the service disciplines assumed and will be discussed in depth throughout this report.

**The Goal in Staffing**
The number of agents required to be on duty as a function of time is the staffing requirement. This is based on our initial M/M/s model. A major complication arises from the constraints in the facility, for example staffing changes often occur only once every hour or once every 8 hours. This has to be treated as a constraint in the model depending on the union agreement and company guidelines.²

---

¹ Green, Kolesar, and Whitt, “Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System” 16.
² Green, Kolesar, and Whitt, “Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System” 15.
Overall, we want to meet the service target of the call evaluators at CCC. We want the evaluators to meet specific service levels while minimizing the staff needed to successfully meet demand at all times. Our service level constraint will control for the zero abandonment constraint needed in the 911 call center at CCC.

**Client Company**

Alberta Health Services EMS Dispatch is responsible for delivering high quality dispatch and medical services throughout the provinces EMS system. The three entities included in the scope of EMS Dispatch are: i) emergency ground ambulance evaluation and dispatch, ii) inter facility patient transfer scheduling and coordination, and iii) a brand new entity of air ambulance flight dispatch and evaluation that was integrated into CCC on April 1st 2009.

AHS is currently undergoing a major transformation with the number of communication centres throughout the province. In the past AHS had 35 communication centres and its plan is to consolidate all of these into three main centres located in Calgary, Edmonton, and Peace River. The Central Communications Centre (CCC) is the Edmonton location and is what this project will be focused on. CCC performs all of the tasks described above, being ground emergency, inter facility transfers, and air ambulance and will be absorbing significant demand over time when it takes over the duties of other communities within the CCC zone of operation.

Figure 1.0

**Staffing Capacity as of February 22, 2011**
OBSERVATIONS

Data Used in Analysis

- North Communication Center (NCC) Telephony Call Activity from April 12, 2007 to December 15, 2010
- Edmonton Response Communication Center (ERCC) Telephony 911 Call Stats which includes all 911 calls redirected from Telus from February 4, 2010 to January 19, 2011
- Computer-aided-dispatch data (CAD) of CCC’s call activity for the months between December, 2010 to February 2011
- Dissection of 911, IFT, and admin calls included. MS Access database includes a sample of the 911 call history and the CallActivity worksheet which includes all telephone transactions coming and going from NCC.
- Call activity of Strathcona County Emergency Services
- CAD activity that relates to this data, as well as the raw data from CCC and SCC.
- Time on task at CCC, AIR ambulance and IFT activity

Descriptive Statistics

A typical staffing model for service centres, such as the CCC, is centred on past data and statistics, which in turn can be used to account for the randomness in the demand for service and predict future demand within certain variability. To start the model and determine the optimal staffing level for CCC, the two main inputs that need to be determined are the arrival rate and the service rate. These two figures are the basis of the model and it would be impossible to move forward without determining them. However, because of the three different entities and the highly seasonal nature of the industry, one service or arrival rate would not be optimal. To move forward we have to sort the statistics to determine the different arrival and service rates for ground emergency inter facility transfer and air ambulance and then further break down the statistics to account for the seasonality. The two main seasonality’s to account for are the differences between each hour of the day, and each day of the week.

Stochastic Demand

As shown in the graph below there are certain hours where numerous staff will have to be working and certain hours where few staff will be optimal.

When determining the seasonality for arrival rates, it is apparent that we would only have to incorporate a weekday schedule and a weekend schedule to satisfy the demand. There are minor differences between weekdays but the only significant impact affecting the model is the
sharp drop in demand on weekends. Further breaking the statistics down to account for the differences of the arrival rates for the three entities, the distributions for the most part follow the same pattern as the aggregate weekday or weekend arrival rates. The only setback experienced occurred when there were too few calls from a certain entity, which in turn showed no significant pattern at any time. This however would not have a large enough impact to affect the model.

The CCC call centre has recently taken on the demand from ERCC and in the near future will also absorb the demand from Strathcona County. It was a important step to account for this demand when discovering the inputs for the model so the model does not become outdated. It was also important that we make the model robust and easily changeable for the managers at CCC to stay useful as they absorb calls from other communities in the long term. Analyzing CCC’s call demand for the past three months determines the scale factor for CCC’s call demand in comparison to NCC’s call demand.

In our ex-ante analysis, call demand is observed to be significantly different between hours of the day as well as days of the week.

Figure 1.0

![Daily Aggregate Call Demand](image)

However, call demand throughout the day is expected to be similar between weekend days as well as between most of the weekdays.
Statistical Test for Arrival Rate Difference between the Days of the Week

Hypothesis tests are conducted between each of the weekly days to examine whether the call demand is significantly different between the days.

- $H_0 =$ Call demand between the two days are the same through out the day
- $H_a =$ Call demand between the two days are different through out the day

If t-statistics is greater than 2 or less than -2, reject the null hypothesis. If not, do not reject the null hypothesis.
The follow are the t-statistics of the hypothesis tests.

Table 1.0

<table>
<thead>
<tr>
<th></th>
<th>Saturday</th>
<th>Friday</th>
<th>Thursday</th>
<th>Wednesday</th>
<th>Tuesday</th>
<th>Monday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>0.934</td>
<td>-4.615</td>
<td>-5.734</td>
<td>-5.677</td>
<td>-4.075</td>
<td>-3.524</td>
</tr>
<tr>
<td>Monday</td>
<td>3.107</td>
<td>-0.138</td>
<td>-1.844</td>
<td>-1.924</td>
<td>-0.538</td>
<td></td>
</tr>
<tr>
<td>Tuesday</td>
<td>-3.359</td>
<td>-0.948</td>
<td>-4.064</td>
<td>5.092</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wednesday</td>
<td>4.582</td>
<td>2.226</td>
<td>0.207</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thursday</td>
<td>4.651</td>
<td>2.056</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friday</td>
<td>3.840</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The daily arrival rates are not significantly different between the days corresponding to the red cells. Therefore, staffing level is expected to be similar between those days.

**Methodology**

The optimal staffing level is the minimum number of staff required to meet the call demand while maintaining a certain waiting time threshold. While different call types have different standards in waiting times, the arrival rates and service rates for different call types are examined separately. As a rough examination of the call demand between the weekdays, call arrivals of all call types are aggregated for each hour. (See Appendix A)

The hourly arrival rates of each call types are examined for each hour and day of the week.

The distribution of hourly arrival rates based on the raw data is almost identical to lognormal distribution. However, there is not a significant difference between lognormal and Poisson distribution when applying to determine the minimum number of servers required maintaining a certain threshold.

Based on the M/M/S queuing model, the minimum number of servers required is determined for each hour of the week. The scheduling model is applied to fit such demand of servers onto a feasible staff schedule.

The scheduling model consists of all the possible combinations of the staffing shifts. (See Appendix B) The blue cells correspond to the number of servers to schedule for the corresponding shift. The row with red numbers has the number of servers to be scheduled at
each hour of the day. It is the sum product between the blue column and the hourly columns. The row with green numbers has the number of servers demanded to maintain a certain waiting time threshold based on the demand of the corresponding hour. It is determined by the M/M/S queuing model. The yellow cell is the total staffing hour required for the day. It is the sum product between the blue column and the “Length” column. The “Length” column is the length of the corresponding shift.

Figure 2.0

<table>
<thead>
<tr>
<th>Shift</th>
<th>Length</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>71</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>188</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

The objective is to minimize the total staffing hours for the day under the constraint that the number of servers scheduled is either equal or greater than the minimum number of servers required to maintain the waiting time threshold for every hour of the day.

Figure 2.1

[Diagram showing the number of servers over the hours of the day, with blue line representing the minimum server hours and red line representing the scheduled server hours.]
SOLVING FOR THE OPTIMAL STAFFING SCHEDULE

1. Determining inputs and outputs
We began the model creation process by determining what the desired inputs and outputs would be. Deciding on the output was easy as it was defined for us in the problem; we needed the output to be an optimal staffing schedule. Determining the inputs was more difficult as we had to decide what was necessary to produce the output we required, as well, we needed to make the model user friendly so the inputs had to be minimal. Through analyzing the data we determined that, after adjusting for holidays, the percentage of calls arriving in each hour of the week remained approximately the same across each week; this assumption became the basis of our model.

2. Determining minimum servers required
In discovering that the calls arrived in approximately the same pattern each week, we were able to reduce the inputs of our model to two; expected weekly demand for two types of calls, air and ground as these types have different service rates. We began by finding the percentage of weekly calls arriving in each of the hours of the week. In the next column was headed by an input cell where the operator would enter the expected weekly demand for that call type. That input cell was linked to the rest of the column and was multiplied by the percentage of calls arriving in each hour to get hourly arrival rate. To find the minimum number of servers required we also needed to account for service rate as well as service level and threshold time. From the data we had already received we determined the call centre’s current service level and threshold time to be approximately 96 percent of calls answered within 10 seconds; we also conducted sensitivity testing for the client. Threshold and service level are also easy to adjust in the model if the client so chooses. Using the data we were able to obtain the average service as the final input to our formula. Although we found the service rates to approximately follow a log normal distribution, we elected to use an M/M/s queueing model. Since the solution is stepwise and the number of required servers only increases when it reaches a certain threshold, we found there to be no significant differences in the results from the M/M/s model and the ARENA model, which used log normal.

3. Creating a Scheduling model
Once we had the minimum required servers for each call type we had to aggregate them to get the minimum number of servers for the whole call centre. After determining the minimum
number of servers needed to fulfill the service level and threshold requirements, we needed to find some way to create an optimal schedule. Our original approach was to make a table in excel for each day of the week, with hours across the top and shift length to the left. At the far left we had a column of variable cells to decide the optimum number of each shift required to fill the demand; these were constrained to be integers. A decision cell which calculated the total number of hours used was created by taking the sumproduct of the variable column and the shift length column; this was minimized. We then used binary variables to create shifts, where 1s represented an hour of work and 0s represented time not working. In the first draft we had three different shift lengths, 12 hours, 6 hours, and 4 hours, and we had the option to begin each shift length in every hour of the day. Along the bottom we placed the minimum required servers in each hour and below that we calculated the number of servers scheduled by taking the sumproduct of the binary code in that column and the variable cells. Once the preliminary model worked out we began to take into consideration other constraints which complicated the model.

4. Incorporating breaks
After reviewing union documents were able to determine appropriate shift lengths and start times. We also found rules for incorporating breaks, thus requiring us to split the hours into 15 minute intervals. The limits of solver prevented us from listing all the possible shifts with varying break times so we had to fix the breaks. Adding breaks to the shifts increased the amount of over staffing experienced by the call centre, however, this is unavoidable since they need to have enough staff working so that one person can take a break without causing the call centre to be understaffed for the duration.

5. Finalizing the Model
Turning this model into a useful schedule required us to piece together the optimal schedules for each day making sure to connect any shifts that continue into the following day. We then compared the minimum servers required to the number of servers scheduled in each of the 15 minute intervals and adjusted the breaks, within the union guidelines, to reduce the amount of over staffing in each interval. We then used the final schedule to create a more useful end product for the client, turning the binary model into a weekly scheduling table. We repeated this process for many levels of weekly call demand. Piecing together the model for each day of the week is a cumbersome process and places unreasonable expectations on the operator, so from here we developed a new model that included all the days of the week. From this new model we
were able to create an easily readable schedule output as the final output, with the inputs being the expected demand for each call type for the week.

**Forecasting**

In forecasting future demand we found that there is more than one seasonality cycle for arrival rates. As expected, there are both intraday cycles and intraweek cycles. The intraday cycle is expected since inter facility transfer calls will not likely be experienced in the early morning hours, as well, emergency calls will be reduced while the general population is asleep. Intraweek cycles are to be expected as well since, as previously mentioned, inter facility transfer call arrivals are far lower on weekends. The call arrival distributions from emergency also vary on weekend evenings as people stay awake longer.

To accommodate for both of these seasonality cycles we adopted a Holt-Winters method that had been adapted for double seasonality. This method was extended to account for the double seasonality experienced when forecasting electricity demand. Since the call centre arrival data approximated that of the electricity demand data we were able to apply this method to forecasting future call demand.

Much like the electricity study conducted by James W. Taylor and Patrick E. McSharry we found that the intraday cycles for weekdays were approximately the same so we chose to use the method presented by Gould et al. that incorporates a smoothing parameter for each type of day to adjust for the intraweek cycle in demand. The three types of days were defined as weekday, Saturday, or Sunday.\(^3\)

As we got further into the project and began developing our model we realized that it would be unrealistic for the client to be required to enter a new demand for each hour as demand increases. With this in mind we decided upon a new approach to forecasting demand. To do this we needed to incorporate the intraday and intraweek cycles right into our model. Our method for accomplishing this was to find, on average, the percentage of weekly calls for each call type that arrived in each hour. From looking at the data we were able to determine that while demand will likely increase over time, the percentage of calls arriving in each hour of the week remains

---

\(^3\) Taylor and McSharry, “Short-Term Load Forecasting Methods: An Evaluation Based on European Data,” 2216.
relatively stable. Knowing this we were able to make the sole input to our model be the weekly demand. The demand for the week would then be multiplied by the percentage of calls arriving in each.

SIMULATING THE RESULTS

Importance of Simulation
The simulation approach is a powerful tool for assessing the staffing levels in a facility like CCC. Anton, Bapat and Hall 1999; Brigandi et al. 1994; Kwan et al 1988 exemplified this in their study. Fortunately, CCC has a vast collection of data as we have mentioned, this is ideal for a simulation which can ultimately convert all of this input data into useful outputs that can aid management and measure performance of the processes in the facility. The key is to make the correct assumptions and analysis based on the input data; in our case arrival rates and service rates throughout the different call types and times of day and week.

Simulation provides the flexibility that CCC requires. Ultimately, simulation provides the versatility to explore a variety of staffing options and incorporate constraints we may face, such as union requirements. We can then measure the output in real time and form a concise and powerful deliverable for the optimal evaluator staffing scheme and be confident with our recommendation. Examples of performance and output will be shown below in our ARENA section. Overall, Simulation answers the questions that we need to know without the cost and use of tangible resources.

Description of Finite and Infinite-server Queuing
Because of the high priority nature of calls received by the EMS call centre, it is crucial that calls are serviced as close to the time they entered the system as possible. Keeping with this criterion, we began our analysis of the call centre data assuming an infinite server queuing system. Under this assumption the virtual wait time is zero, theoretically reducing the occurrence of customer balking to zero as well. It is vital for balking in such a high priority line of work to be as low as possible to ensure that no potentially life threatening calls are missed.

---

4 Ger Koole, *Call Center Mathematics: A scientific method for understanding and improving contact centers*, 3 - 37
5 Ger Koole, *Call Center Mathematics: A scientific method for understanding and improving contact centers*, 10-37
We used the results from our initial assumption of infinite servers to evaluate the utilization. We then used our findings to compare against the more accurate assumption of a finite queuing system. Using service time distributions for the different call types and arrival rates for the different hours of the week we determined the minimum number of call evaluators required per hour to answer 95 percent of calls within 10 seconds. We then used this optimal result to compare with the results from the infinite server queuing system. Using ARENA we simulated with the minimum number of servers and evaluated the output parameters Wq (time spent in waiting in the queue) and PrBalk (probability that a customer will balk).

**Markov Chain (Birth-and-Death Process)**

**Steady-state or stationary distribution**

It is important to note the assumptions behind our analysis and model. Once the CCC queueing system begins operation, the state of the system or the number of customers within the system is largely distorted by the initial operation and time (t) that has elapsed; this is referred to as a transient state. For the purpose of analysis, practicality and results we have determined that our model has reached a steady state or stationary state that is subject to the drastic fluctuations outlined in a transient state. For the most part queueing theory has relied on the steady-state distribution and the theories and analysis behind our model reflect that.

**Terminology and Notation**

Unless stated otherwise, the following notation applies:

**State of the system** = number of customers in the queueing system

**Queue length** = number of customers waiting for service (=state of system minus number of customers being served)

N(t) = number of customers in queueing system at time t (t>=0)

P_n(t) = probability that exactly n customers are in queueing system at time t, given number at time 0

s = number of servers in queueing system

---

6 Hillier and Lieberman, *Introduction to Operations Research*, 525 - 527


\( \lambda_n = \) mean arrival rate (expected number of arrivals per unit time) of new customers when \( n \) customers are in the system

\( \mu_n = \) mean service rate (expected number of customers completing service per unit time)

As we have discovered, queueing models contain inputs, or arriving entities and then outputs, or leaving entities. There is a specific context for the terms birth and death in queueing. Birth refers to when an entity or new customers arrives into the system and death refers to the customer or entity leaving or being disposed from the queueing system. The birth-and-death process in queueing states how the number of customers in the queueing system changes over time. Overall, the process states that births and deaths of individual entities occur stochastically.\(^9\)

Below is a rate-diagram for the birth-and-death process (continuous Markov chain) of a queueing system.

\[\text{STATE:} \]

\[ \begin{array}{cccc}
0 & 1 & 2 & 3 \\
\lambda_0 & \lambda_1 & \lambda_2 & \lambda_{n-2} \\
\mu_1 & \mu_2 & \mu_3 & \mu_{n-1} \\
\end{array} \]

In the above diagram, \( \lambda \) represents mean arrival rate from our notation and terminology. The numbers inside the circles are representative of the number of customers in the queueing system at time \( t \). \( \mu \) represents the service rate for the customers in the queue. Depending on how many customers are in the respective chain of the queueing system we must multiply the \( n \) by the service rate to accommodate the increased customers being served. As \( n \) becomes large we start to observe generalizing properties related to the service rates.\(^{10}\)


\(^{10}\) Hillier and Lieberman, *Introduction to Operations Research*, 539
**M/M/s Model and Markov Chains**

In our queueing model we have assumed our interarrival times are independently and identically distributed according to exponential distribution (Poisson).\(^\text{11}\) We did observe in our data analysis and distributions fitting that the service times were closer to a lognormal distribution than an exponential one but because our servers are stepwise incremented, it is sufficient to assume the exponential service times inherent in an M/M/s model and our results were not significantly different. This model is an extension of the general Birth-and-Death Process.

Rate diagram for M/M/s (multiple server case)

The deviation from the original model is evident in the above M/M/s rate diagram. As shown, the mean arrival rate is constant in the multi-server model. As \(s\) becomes large the service times remaining increase linearly with \(s\), demonstrated by \(s\lambda\).\(^\text{12}\)

The properties are summarized below:

\[
\begin{align*}
\lambda_n &= \lambda, \text{ for } n=0,1,2,\ldots \\
\mu_n &= (s\lambda), \text{ for } n = 1, 2, \ldots, s, \quad n = s, s+1
\end{align*}
\]

The average service rate \((s\lambda)\) M/M/s when it exceeds the mean arrival rate \((\lambda)\) we can derive the following relationship. This is prevalent in our analysis and was a factor in our queueing model. In our ARENA results we observed the utilization of resources increased as this relationship held.

\(^\text{11}\) Hillier and Lieberman, *Introduction to Operations Research*, 543

The following formula also known as Utilization Factor describes the property:

\[ \rho = \frac{\lambda}{s\mu} < 1 \]  

The following are M/M/s queueing formulas based on the steady-state or stationary distribution of the queueing system. The notation is based on the equation:

\[ \rho = \frac{\lambda}{s\mu} < 1 \]

If \( \lambda < s\mu \)

\[ P_0 = \frac{1}{\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} + \left( \frac{\lambda}{s\mu} \right)^s \sum_{n=s}^{\infty} \frac{\lambda^n}{n!}} \]

\[ = \frac{1}{\sum_{n=0}^{\infty} \frac{\lambda^n}{n!} + \left( \frac{\lambda}{s\mu} \right)^s \frac{1}{1 - \left( \frac{\lambda}{s\mu} \right)}} \]

And

\[ P_n = \begin{cases} \frac{\lambda^n}{n!} P_0 & \text{if } 0 \leq n \leq s \\ \frac{\lambda^n}{s!} P_0 & \text{if } n \geq s \end{cases} \]

Based on the notation:

\[ \rho = \frac{\lambda}{s\mu} \]

\[ L_q = \sum_{n=s}^{\infty} (n-s) P_n \]

\[ = \sum_{j=0}^{\infty} j P_{s+j} \]

\[ = \sum_{j=0}^{\infty} \left( \frac{\lambda^j}{s!} \right) \rho^j P_0 \]

\[ = P_0 \left( \frac{\lambda^j}{s!} \right) \rho \sum_{j=0}^{\infty} \frac{d}{dp} (p^j) \]

\[ = P_0 \left( \frac{\lambda^j}{s!} \right) \rho \frac{d}{dp} \left( \sum_{j=0}^{\infty} p^j \right) \]

\[ = P_0 \left( \frac{\lambda^j}{s!} \right) \rho \frac{d}{dp} \left( \frac{1}{1-p} \right) \]

\[ = \frac{P_0 (\lambda^j/s!) \rho}{s! (1-p)^2} \]

\[ W_q = \frac{L_q}{\lambda} \]

---


Above formulas cited and consulted\textsuperscript{16}

Our analysis is centred around a single queueing model: M/M/s Poisson arrival process with a time-varying arrival rate function, independent and identically distributed service times following a general probability distribution, a possibly time-varying number of call evaluators on shift in an infinitely large waiting room, Servers’ duties are assigned on a first come first serve cyclical basis. Customer abandonment constraint is replaced by the waiting time threshold. Customers are assumed to be homogeneous and may wait in an invisible queue where they have no knowledge of the current state of the system. The services times are observed to be lognormal.

The variability of the observed lognormal distribution was not great, and the squared coefficient of variation was between 1 and 2. Variance should be replaced by SCV, because it is independent of the mean; the SCV of a random variable is unchanged if it is multiplied by a constant. In such cases it is reasonable to use exponential distributions. We indicate what can be done if exponential distributions are not appropriate. A high level of customer abandonment may be a sign of poor service; indeed, it often implies lost sales. On the other hand, a low level of abandonment, such as 1 percent, in a large call center may be a sign of proper staffing, where supply appropriately balances demand. Regardless of the interpretation, it can be useful to recognize the presence of customer abandonment and explicitly include its impact on performance and hence on staffing. Even a small amount of customer abandonment can significantly impact system performance and staffing requirements. In the past, abandonments were not included in staffing models, primarily because they appeared to make the model too complicated.\textsuperscript{16}

\begin{align*}
W &= W_q + \frac{1}{\mu} \\
L &= \lambda(W_q + \frac{1}{\mu}) = L_q = \frac{\lambda}{\mu}
\end{align*}

\textsuperscript{15} Hillier and Lieberman, \textit{Introduction to Operations Research}, 546-547

\textsuperscript{16} Gans, Koole, and Mandelbaum, “Telephone Call Centers: Tutorial, Review, and Research Prospects”
1) **Emergency**

We describe in detail what the adjacent diagram represents for 911 calls:
Step 1) Emergency incident occurs, caller calls 911 EPS (EPS primary, AHS secondary)
Step 2) Call evaluator verifies location of caller, while location information from TELUS and phone companies is populated
Step 3) Call evaluator prealerts dispatcher and paramedics
Step 4) Conduct ProQA (roughly 45sec) while paramedics are getting ready
Step 5) Information populates into CAD
Step 6) After scenario is confirmed and acuteness identified, paramedics are notified
Step 7) Time stamp recorded
Note: Children stay on the phone for the entire duration
Roughly 10 percent of the time the call evaluator stays on the line to conduct pre-arrival instructions

2) **IFT (Inter-facility transfer)**

Step 1) Call or fax from AHS entity or other contracting company
• Fax is pre-booked days in advance, Calls are within hours or the same day
Step 2) Multiple calls can be made and modifications to facility transfer route
• IFT transfer planning is one of the most cognitively demanding positions

Notes:
• Dispatching for IFT is not linear and static like 911 calls, can be pushed backed and modified
• Seven or more radio calls are used for each IFT (inter-facility transfer call), CCC deals with roughly 150 IFT calls per-day
• One person is designated for time-stamping and another person is designated for radio receiving.

3) **Air Ambulance**

The flight portion of the incoming calls is also a diverse entity. Flight call evaluators have to be fluent in both inter facility transfer coordination as well as emergency because the incoming flight calls could be either. The IFT’s are pre-booked and the info waits in CAD for 3-4 days before the transfer takes place. Flight calls are also much longer than the normal emergency call and can be overly demanding.

**Emergency Flight Calls:**

Step 1) Helicopter takes off within 30min of call
Step 2) Many more calls are made to coordinate activities between ground crew, paramedics and critical care teams, multiple events must be coordinated

Notes:
• Time of a call may be double, triple or even longer than a regular ground 911 call
• Two flight call evaluators, also take regular calls (one dispatcher)
• Heavy call demand for time stamping from multiple areas (STARS etc)

**Arena-Theory, Assumptions and Understanding**

Arena simulation software provides forward visibility of the staffing scheme by analyzing the impact of potential contingencies, rules and strategies. When lives are at stake, Arena simulates the decisions before causing any impact on the service.

The arriving calls in the simulations model are assumed to have heterogeneous groups differentiated by the types of calls (Emergency, Inter Facility Transfer, etc) with different arrival
rate and service rate distributions. The calls within each group are assumed to be homogeneous with identical arrival rate and service rate distributions.

Mapping CCC's Process Flow onto Arena
As a call comes into CCC, it is redirected to the call evaluator. If the call is an emergency, the call evaluator would pre-alert the paramedics by default and conduct a 45 second pro QA with the caller to examine the acuteness of the situation. Interfacility transfer calls requires the call evaluator to stay on the line for an extended amount of time, and its service time have greater variation in length.

Call evaluators are responsible for taking every incoming call except for air inter facility transfer and air emergency calls. The flight requests are received and coordinated by the flight coordinators. All call evaluators and flight coordinators are assumed to be homogeneous within their group.

Figure 3.0
After running 100 simulations, the expected, maximum and minimum waiting times in hours for different call types are obtained. The simulated waiting times based on the suggested staffing scheme does qualify with respect to their corresponding thresholds.

Simulation Results

| Queue                |                      |                      |                      |                      |                      |
|----------------------|----------------------|----------------------|----------------------|----------------------|
|                      | Average              | Half Width           | Minimum Average      | Maximum Average      | Minimum Value        | Maximum Value        |
| Air Emergency Evaluation and Dispatching Queue | 0.00001645 | 0.00 | 0.00 | 0.000016445 | 0.00 | 0.01058338 |
| Air IFT Evaluation and Dispatching Queue | 0.00007436 | 0.00 | 0.00 | 0.000038219 | 0.00 | 0.06191403 |
| Emergency Call Evaluation Queue | 0.00019906 | 0.00 | 0.00 | 0.00009409 | 0.00 | 0.03429195 |
| IFT Call Evaluation Queue | 0.00018588 | 0.00 | 0.00 | 0.000325766 | 0.00 | 0.03912238 |
| Non-Emergency Call Evaluation Queue | 0.00022392 | 0.00 | 0.00 | 0.00011506 | 0.00 | 0.04522525 |
| Other Calls Queue | 0.00017384 | 0.00 | 0.00 | 0.00036300 | 0.00 | 0.01317478 |

The utilization rates of the servers are relatively low at a glance. However, the fact that enough servers must be scheduled to prevent the call waiting times to breach the thresholds, the individual servers tends to be not overly busy. Given the low call waiting time threshold our client has in place, callers are not expected to balk.

<table>
<thead>
<tr>
<th>Number Busy</th>
<th>Average</th>
<th>Half Width</th>
<th>Minimum Average</th>
<th>Maximum Average</th>
<th>Minimum Value</th>
<th>Maximum Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluator 1</td>
<td>0.08728586</td>
<td>0.00</td>
<td>0.06185913</td>
<td>0.0960</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Evaluator 2</td>
<td>0.08201742</td>
<td>0.00</td>
<td>0.07842957</td>
<td>0.0857365</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Evaluator 3</td>
<td>0.06158340</td>
<td>0.00</td>
<td>0.05875516</td>
<td>0.06527445</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Evaluator 4</td>
<td>0.05732785</td>
<td>0.00</td>
<td>0.05333591</td>
<td>0.06092667</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Evaluator 5</td>
<td>0.02918684</td>
<td>0.00</td>
<td>0.02641988</td>
<td>0.03210141</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Evaluator 6</td>
<td>0.03944401</td>
<td>0.00</td>
<td>0.03418299</td>
<td>0.04240855</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Flight Coordinator 1</td>
<td>0.03081575</td>
<td>0.01</td>
<td>0.01643792</td>
<td>0.05523834</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
<tr>
<td>Flight Coordinator 2</td>
<td>0.02714658</td>
<td>0.00</td>
<td>0.01803143</td>
<td>0.03741694</td>
<td>0.00</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Long-Term Issues and Cross-Training

The inputs and considerations within our queueing model are all related to human behaviour. Understanding this human behaviour is a key component to effectively making staffing and scheduling decisions.
A major challenge with modeling is incorporating these human factors. It would be important for AHS to incorporate or understand the complicated aspects of cross-training, customer psychology and its effect on efficiency. Gans, Koole and Mandelbaum have shown that customers and employees have the ability to adjust behaviour based on expectations.\textsuperscript{17}

In “Call Center Mathematic”, Koole emphasized the efficiency of specialized agents, or in our case evaluators.\textsuperscript{18} First of all, they need less training and are often more efficient due to their specialization. The major downside of having specialized evaluators is a lack of flexibility. Another disadvantage is that specialization reduces economies of scale.

**Conclusion**

The simulated call waiting time result does meet the standard set by the threshold. The staffing optimization model based on the methodology discussed above is necessary in determining the most efficient schedule for call centres. Such scheduling scheme minimizes the number of servers required and the number of servers allocated.

\textsuperscript{17} Ger Koole, *Call Center Mathematics: A scientific method for understanding and improving contact centers*, 3 - 37
\textsuperscript{18} Ger Koole, *Call Center Mathematics: A scientific method for understanding and improving contact centers*, 3 - 37
Works Cited and Consulted:


<table>
<thead>
<tr>
<th>Hour</th>
<th>Sunday</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
<th>Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:00:00 AM</td>
<td>14.255</td>
<td>11.318</td>
<td>12.904</td>
<td>14.215</td>
<td>15.056</td>
<td>12.677</td>
<td>17.406</td>
</tr>
<tr>
<td>1:00:00 AM</td>
<td>12.415</td>
<td>9.737</td>
<td>11.754</td>
<td>11.423</td>
<td>11.533</td>
<td>11.699</td>
<td>15.528</td>
</tr>
<tr>
<td>2:00:00 AM</td>
<td>12.030</td>
<td>8.192</td>
<td>10.812</td>
<td>11.754</td>
<td>12.167</td>
<td>10.011</td>
<td>14.351</td>
</tr>
<tr>
<td>3:00:00 AM</td>
<td>9.906</td>
<td>7.560</td>
<td>6.369</td>
<td>7.669</td>
<td>8.922</td>
<td>7.687</td>
<td>11.575</td>
</tr>
<tr>
<td>4:00:00 AM</td>
<td>8.675</td>
<td>6.224</td>
<td>5.604</td>
<td>5.673</td>
<td>5.845</td>
<td>6.526</td>
<td>8.785</td>
</tr>
<tr>
<td>5:00:00 AM</td>
<td>7.440</td>
<td>6.908</td>
<td>6.688</td>
<td>6.608</td>
<td>6.815</td>
<td>6.901</td>
<td>8.011</td>
</tr>
<tr>
<td>6:00:00 AM</td>
<td>8.253</td>
<td>11.848</td>
<td>9.108</td>
<td>11.285</td>
<td>10.615</td>
<td>11.707</td>
<td>7.467</td>
</tr>
<tr>
<td>7:00:00 AM</td>
<td>10.827</td>
<td>16.246</td>
<td>14.627</td>
<td>16.915</td>
<td>17.254</td>
<td>13.299</td>
<td>11.119</td>
</tr>
<tr>
<td>8:00:00 AM</td>
<td>13.926</td>
<td>23.355</td>
<td>22.535</td>
<td>22.650</td>
<td>22.332</td>
<td>19.473</td>
<td>12.891</td>
</tr>
<tr>
<td>9:00:00 AM</td>
<td>18.847</td>
<td>26.658</td>
<td>26.477</td>
<td>26.738</td>
<td>27.092</td>
<td>26.491</td>
<td>15.896</td>
</tr>
<tr>
<td>10:00:00 AM</td>
<td>19.354</td>
<td>25.793</td>
<td>27.481</td>
<td>28.712</td>
<td>30.291</td>
<td>31.447</td>
<td>16.466</td>
</tr>
<tr>
<td>11:00:00 AM</td>
<td>19.375</td>
<td>30.066</td>
<td>28.965</td>
<td>28.588</td>
<td>29.113</td>
<td>30.118</td>
<td>20.884</td>
</tr>
<tr>
<td>12:00:00 PM</td>
<td>21.247</td>
<td>27.590</td>
<td>25.485</td>
<td>28.062</td>
<td>27.339</td>
<td>28.803</td>
<td>19.451</td>
</tr>
<tr>
<td>1:00:00 PM</td>
<td>21.811</td>
<td>31.620</td>
<td>29.604</td>
<td>29.985</td>
<td>27.843</td>
<td>31.413</td>
<td>20.895</td>
</tr>
<tr>
<td>2:00:00 PM</td>
<td>24.951</td>
<td>42.957</td>
<td>29.115</td>
<td>33.800</td>
<td>34.549</td>
<td>32.190</td>
<td>22.500</td>
</tr>
<tr>
<td>3:00:00 PM</td>
<td>20.343</td>
<td>29.357</td>
<td>29.815</td>
<td>31.754</td>
<td>31.972</td>
<td>30.328</td>
<td>18.468</td>
</tr>
<tr>
<td>4:00:00 PM</td>
<td>23.705</td>
<td>27.906</td>
<td>34.412</td>
<td>33.692</td>
<td>32.051</td>
<td>27.519</td>
<td>19.105</td>
</tr>
<tr>
<td>5:00:00 PM</td>
<td>17.829</td>
<td>25.972</td>
<td>25.200</td>
<td>29.304</td>
<td>29.136</td>
<td>23.969</td>
<td>17.175</td>
</tr>
<tr>
<td>6:00:00 PM</td>
<td>16.845</td>
<td>22.972</td>
<td>21.808</td>
<td>24.792</td>
<td>24.041</td>
<td>21.525</td>
<td>17.128</td>
</tr>
<tr>
<td>7:00:00 PM</td>
<td>17.494</td>
<td>18.357</td>
<td>20.319</td>
<td>24.958</td>
<td>25.659</td>
<td>19.912</td>
<td>17.359</td>
</tr>
<tr>
<td>8:00:00 PM</td>
<td>18.187</td>
<td>20.709</td>
<td>20.623</td>
<td>23.296</td>
<td>24.072</td>
<td>19.683</td>
<td>16.830</td>
</tr>
<tr>
<td>9:00:00 PM</td>
<td>16.411</td>
<td>19.722</td>
<td>18.396</td>
<td>20.415</td>
<td>19.647</td>
<td>19.344</td>
<td>15.572</td>
</tr>
<tr>
<td>10:00:00 PM</td>
<td>16.832</td>
<td>16.308</td>
<td>18.454</td>
<td>20.538</td>
<td>19.497</td>
<td>19.659</td>
<td>16.842</td>
</tr>
</tbody>
</table>
## Appendix B

<table>
<thead>
<tr>
<th>Shift</th>
<th>Length</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Shift Length Hours

<table>
<thead>
<tr>
<th>Shift</th>
<th>Length</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Summary

<table>
<thead>
<tr>
<th>Demand</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supply</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
