STAFFING LEVELS AT THE AUCKLAND POLICE COMMUNICATION CENTRE

Bert P. K. Chen

Department of Engineering Science

University of Auckland

Private Bag 92019

Auckland, New Zealand

Abstract

The Auckland Police Communication Centre accepts both emergency and some non-emergency calls from the upper half of the North Island, and dispatches police units to those calls. The centre would like to select staffing levels so that certain customer service performance criteria are met. We investigate the use of queuing models to assist in setting these staffing levels. An approach that ignores the fact that emergency calls receive higher priority (the standard Erlang delay formula) tends to yield slightly higher staffing levels than a second approach that explicitly considers the two priority levels of the calls.

The existing shift structures at the centre are perhaps not as flexible as one might like to deal with variable demand. With smaller shift lengths, and more flexible starting times, one would expect to need less staff to cover the required levels, and our analysis supports this.

1 Introduction

The Auckland Police Communication Centre in New Zealand fields calls requiring police response from the upper North Island. The call takers of the communication centre simply answer the incoming calls and pass them to the dispatchers who dispatch jobs according to the level of emergency and the current policing resources.

Calls are classified into three groups: emergency (111) calls, allied emergency agencies calls and non-emergency (general) calls. Allied Emergency Agencies calls are calls requiring ambulance or fire service in which police help is also needed.

The Police Communication Centre aim requires that their call takers meet the following performance target: 90% of emergency (111) calls and allied emergency agencies calls to be answered within 10 seconds, and 80% of non-emergency (general) calls to be answered within 30 seconds. Owing to these different call service requirements, the three types of calls could be divided further into two priority classes: emergency (111) calls and allied emergency agencies calls are classified as priority 1 calls since they have the same performance target, and non-emergency (general) calls are rated priority 2 because of the lower performance requirement.

When a call arrives to the communication centre and all call takers are busy, the call is queued. If a priority 1 call waits for more than 10 seconds in the queue, then it is passed to another communication centre in another part of the country. According to the

Police Communication Centre staff, this occurs only very rarely. Queued priority 2 calls are answered only after any queued priority 1 calls have been cleared.

The Police Communication Centre has a budget for their staffing and therefore avoidance of exceeding it is encouraged. Appropriate staffing levels, corresponding service performance and the cost-effectiveness of the existing roster structure for the call takers are the most important concerns.

In this paper we consider the problem of determining the number of call takers required in each hour to ensure that the service requirements of calls are satisfied at all times of the day. We also investigate potential savings by moving from the existing roster structure of call takers to a different roster structure.

Currently, the Auckland Police Communication Centre uses an Erlang *c* calculator to estimate the number of call takers required (see M/M/c later) which is the most common queuing model adopted in a call centre environment [6]. Our approach is to study the existing model (M/M/c) the communication centre adopts and compare it with a new model that takes into account the priorities of calls. The new model applied here is the non-preemptive M/M/c queue [4]. It is easy to calculate the exact performance of priority 1 calls, however, difficulty with priority 2 service estimation was encountered while implementing this new model as it requires considerable computation that is not easily performed in the spreadsheet environment that the Police Communication Centre prefers. We used Markov's inequality [2] to obtain a bound on the service performance of priority 2 calls. The queuing models are discussed in detail in Section 2.

We are very lucky in that a large amount of data was available to us. In particular, the communication centre has hourly data going back several years on the number of calls (broken down by type) that were received, together with service times. The service times for non-emergency calls appear to be somewhat longer than those for emergency calls. In addition, we identified two distinct seasons from the data of call arrivals. The period from March to August was defined for the purposes of this study as the "winter" season and the period from September to February was defined as the "summer" season. Each season is treated separately for the remainder of this paper.

It is often assumed that calls arrive as a Poisson process at call centres for several reasons [5]. Suppose we assume that the calls in a given period are generated by a Poisson process with a fixed arrival rate l. A Chi-Square Test of arrivals shows that some times are not well modelled by a Poisson process, while other times are. Recent research [3] suggests that this may be because the arrival rate at certain times is random, not fixed as our previous analysis assumed.

By assuming a fixed arrival rate, we may underestimate the number of servers required in our queuing models in Section 2. This can be partially addressed by decreasing the service rate.

In this paper, the arrival rate for each hour for both seasons was obtained by taking the average of the actual calls received in the corresponding hour of the entire season (26 weeks).

Each of the three types of calls has a different service time. For our model we need a common service time. This is calculated by taking the weighted average of service times for the three different types of calls. Furthermore, given the fact that we cannot isolate the service time distributions, it is reasonable to assume that they are exponential.

As for the rostering aspects of call takers, we looked at the workload allocation models that use the existing roster structure, as well as a new and more flexible structure.

These two models with different roster structures should give the Police Communication Centre some indications in regard to a better shift structure that also reduces staffing hours. The detail of the workload allocation models is discussed in Section 3.

In Section 4, we provide and discuss the results of our analysis on a subset of the Police Communication Centre data set

2 The Queuing Models

2.1 M/M/c Queue

Consider a queuing system where customers arrive according to a Poisson process with rate l. Service times for all customers are exponentially distributed with common mean m^{-1} and there are c servers.

For traffic intensity $\rho < 1$, the state probabilities are denoted by p_n = probability that n calls are in the system. Let W be a random variable having the limiting wait in queue distribution. The distribution of waiting time, given that one has to wait, is exponential with mean $(cm - l)^{-1}$ [7].

The corresponding fraction of time that all servers are busy is

$$P(W > 0) = 1 - \sum_{j=0}^{c-1} p_j . \tag{2.1.1}$$

The probability that one does not wait is

$$P(W = 0) = P(N \le c-1) = \sum_{j=0}^{c-1} p_j , \tag{2.1.2}$$

where N is a random variable taking values of number of callers in the system.

Then the probability of the wait in the queue being less than w units of time is

$$P(0 \le W \le w) = P(W = 0) + P(W > 0) \times \left(1 - \exp\left(-(cm - l)w\right)\right). \tag{2.1.3}$$

For a stable system, we must have $r < 1$, so that the minimal number of servers required to get a stable system is

$$c = \lceil l/m \rceil . \tag{2.1.4}$$

Subject to the service requirements, we can increment on the minimal number of servers found from (2.1.4) until the desired level of performance is reached. The level of performance is checked by (2.1.3) to see if the required proportion of calls that has to be answered within the target time is reached.

The M/M/c queue ignores the priority of calls, which means that no calls receive priority over any other calls. The arrival rate we used includes both priority 1 and priority 2 calls. To ensure that 90% of priority 1 calls wait within 10 seconds, we choose c so that 90% of all calls wait within 10 seconds. This ensures that both priority 1 and priority 2 calls will satisfy their service requirements.

2.2 Non-Preemptive Priority M/M/c Queue

In this priority queue model, we choose a staffing level per hour, c_1, that gives 90% of priority 1 calls an answer within 10 seconds and a staffing level per hour, c_2, that gives 80% of priority 2 calls an answer within 30 seconds. We then take the maximum of c_1 and c_2 as the ultimate staffing level that will satisfy the requirements for both priority 1 and priority 2 calls.

Consider a non-preemptive M/M/c priority queuing system with n priority classes [4] where customers of class i arrive according to a Poisson process with rate l_i, $1 \le i \le n$. Service times for all customers are exponentially distributed with parameter m.

Customers of class i have non-preemptive priority over customers of class j whenever $i < j$, and service within each class follows the first come first serve rule.

Based on the assumption that an arriving customer waits for service if and only if all servers are busy, Kella and Yechiali defined the following:

The overall arrival of the system,

$$l = \sum_{i=1}^{n} l_i . \tag{2.2.1}$$

The traffic intensity of each priority class i,

$$r_i = \frac{l_i}{cm} . \tag{2.2.2}$$

s_j denotes the sum of the traffic intensities of class 1 up to class j,

$$s_j = \sum_{i=1}^{j} r_i , \tag{2.2.3}$$

and the overall traffic intensity of the system is

$$r = s_n = \frac{l}{cm} . \tag{2.2.4}$$

Consequently, the probability of all servers being busy in a non-preemptive M/M/c queue [4] with the same service rate for all classes is

$$p = \frac{(l/m)^c}{c!(1-r)} \left[\sum_{i=0}^{c-1} \frac{(l/m)^i}{i!} + \frac{(l/m)^c}{c!(1-r)} \right]^{-1} . \tag{2.2.5}$$

Let W_k denote the steady-state waiting time in the queue for priority k customers. Kella and Yechiali also give the Laplace Transform of W_k for each $k \geq 1$. It is easy to invert the transform for $k = 1$ (priority 1 customers) to get the probability that the waiting time in the queue is less than or equal to w units of time,

$$P(W_1 \leq w) = 1 - p \exp(-(cm - l_1)w) . \tag{2.2.6}$$

On the contrary, there is no easy way to invert the transform for $k > 1$ (lower priority classes).

Alternatively, we can use Markov's inequality [2], or numerical transform inversion [1] to calculate $P(W \leq w)$, the probability that the waiting time in the queue before a server is available is at most w units of time for lower priority classes.

In the approach where Markov's inequality is employed to bound $P(W_k \leq w)$ for any lower priority class k, we need the first two moments of W_k.

These two moments of the waiting time of a class k customer [4] are given by

$$EW_k = \frac{p}{cm(1 - s_k)(1 - s_{k-1})} , \tag{2.2.7}$$

and

$$EW_k^2 = \frac{2p(1 - s_k s_{k-1})}{(cm)^2(1 - s_k)^2(1 - s_{k-1})^3} . \tag{2.2.8}$$

Markov's inequality [2] for a non-negative random variable X and constants x, $a > 0$ states that

$$P(X \geq x) \leq \frac{EX^a}{x^a}. \tag{2.2.9}$$

Hence, we may immediately conclude that

$$P(W_k > w) \leq \min\left(\frac{EW_k}{w}, \frac{EW_k^2}{w^2}\right). \tag{2.2.10}$$

From (2.2.10), we are able to get an upper bound of $P(W_k > w)$. Thereafter, a lower bound on $P(W_k \leq w)$ can be evaluated by subtracting the bound (2.2.10) from 1. This lower bound is then used as an estimate of $P(W_k \leq w)$. However, there is a possibility that this approximation of $P(W_k \leq w)$ is an underestimation of the true $P(W_k \leq w)$; and this could lead to a possible overestimation of required server numbers. In practice, this problem is often addressed by ignoring the performance requirements for the lower priority classes and concentrating solely on satisfying the performance target for the top priority customers. This would bring down the prediction of the number of servers required subject to the performance requirements and hopefully yield an estimation closer to the true value.

Note that determination of the number of servers needed to satisfy performance targets is very similar to the process outlined in the last few paragraphs of Section 2.1. First of all, the minimum number of servers required to yield a stable system is calculated by using (2.1.4). Then this minimum value is incremented by one at a time until the desired level of performance is reached for all priority classes.

3 The Workload Allocation Models

These workload allocation models select shift starting times and calculate the number of staff required in the selected shifts so that required staffing levels are met. However, these models do not generate full lines of work (shift for workers over a week is not considered).

3.1 Model Using the Existing Shift Structure

Currently, the call takers in the Police Communication Centre work according to a 5-week cyclic roster. The features of this cyclic roster are summarised in Table 1 below.

	Week 1	Week 2	Week 3	Week 4	Week 5
Mon	N	x	x	L	e
Tue	N	x	x	L	e
Wed	N	x	e	L	x
Thu	N	s	e	x	x
Fri	X	s	e	x	n
Sat	X	s	x	e	n
Sun	X	s	x	e	n

Shift Hours & Lengths	
e = 0600 to 1600	10 early
s = 1600 to 0200	10 swing
L = 1200 to 2200	10 late
n = 2100 to 0600	9

Table 1. The existing cyclic roster for the call takers in the Police Communication Centre.

In Table 1, e, s, and L represent the early, swing and late shifts respectively with a shift length of 10 hours; n is the special 9-hour shift, and x denotes a day-off. Call takers are divided into 5 groups. Each group works in a different shift everyday or has a day off. For example, on any Monday, two groups would have the day-off while the other three work in the early, late or special 9-hour shift accordingly. Every call taker would have been through all the allocated shifts in Table 1 every 5 weeks.

In this model where the existing shift structure of the communication centre is kept, we are seeking a weekly workforce allocation that minimises the total staffing hours. We assume that call takers can work in 4 different shifts, e, s, L and n as outlined in Table 1, and everyday in a week from Monday to Sunday.

The aim here is to find the number of people required in each shift which would satisfy the hourly staffing requirements suggested by the queuing models in the previous section and keep the total staffing hours to a minimum.

This workforce allocation problem can be formulated as an integer program of the form given by:

$$\text{Minimise} \quad z = c^T x \tag{3.1.1}$$

$$\text{Subject to} \quad Ax \geq b \tag{3.1.2}$$

$$x_j \in \{0, 1, 2, ...\} \qquad j = 1, ..., n \tag{3.1.3}$$

where $A = (a_{ij})$ is a $m \times n$ matrix of zeros and ones;

$a_{ij} = 1$ if shift j exists in hour i,

$= 0$ otherwise;

$x = (x_j)$ is a $n \times 1$ vector of variables representing the number of people in shift j;

$b = (b_i)$ is a $m \times 1$ vectors containing the hourly staffing requirements of a typical week;

$c = (c_j)$ is a $n \times 1$ cost vector representing the length of shift j.

In this model, $m = 168$ because there are 168 hours in a week and $n = 28$ as there are 28 shifts in a week with 4 shifts in each day. The cost vector of the integer program is the vector containing the relevant shift lengths for each shift, either 9 or 10, since we are minimising the total staffing hours of a week. A matrix contains the information on shift structure.

3.2 Model Using a Flexible Shift Structure with a Small Shift Length

Instead of adopting the current shift structure, this model examines the case where a smaller shift length and more flexible starting times of shifts are implemented. In particular, we are looking at a shorter shift length of 8 hours and shifts that can begin at the start of any hour in a week.

Again we are seeking a weekly workforce allocation that minimises the total staffing hours, subject to the hourly staffing levels suggested by our queuing models.

The number of people required in each shift can be found by solving the same integer program, (3.1.1), (3.1.2) and (3.1.3) in Section 3.1.

In this model, $m = n = 168$ since there are 168 hours in a week and 168 possible starting times if shifts start at the beginning of the hour. The cost vector, c, is now a vector of 8's since the shift length is kept constant at 8 hours. A is a 168×168 square matrix containing all the shift information.

4 Results and Discussion

4.1 Results and Interpretation from the Queuing Models

Figure 1 shows the required hourly staffing levels in a typical week (168 hours, from midnight Monday to midnight Sunday) suggested by our queuing models that satisfy the performance requirements of the Police Communication Centre for winter. We detected quite a large amount of variation in the predicted staffing levels. In winter, required staffing levels vary from 3 to 12 servers. This variation might be due to the time of day effect. In particular, we observed that the predicted staffing levels reached their peaks roughly around the period 4 to 7pm in weekdays whereas in weekends the staffing levels predicted are less variable. We suspect that the time of day effect plays an important role in weekdays. Plot for summer is not included here because it shows similar results.

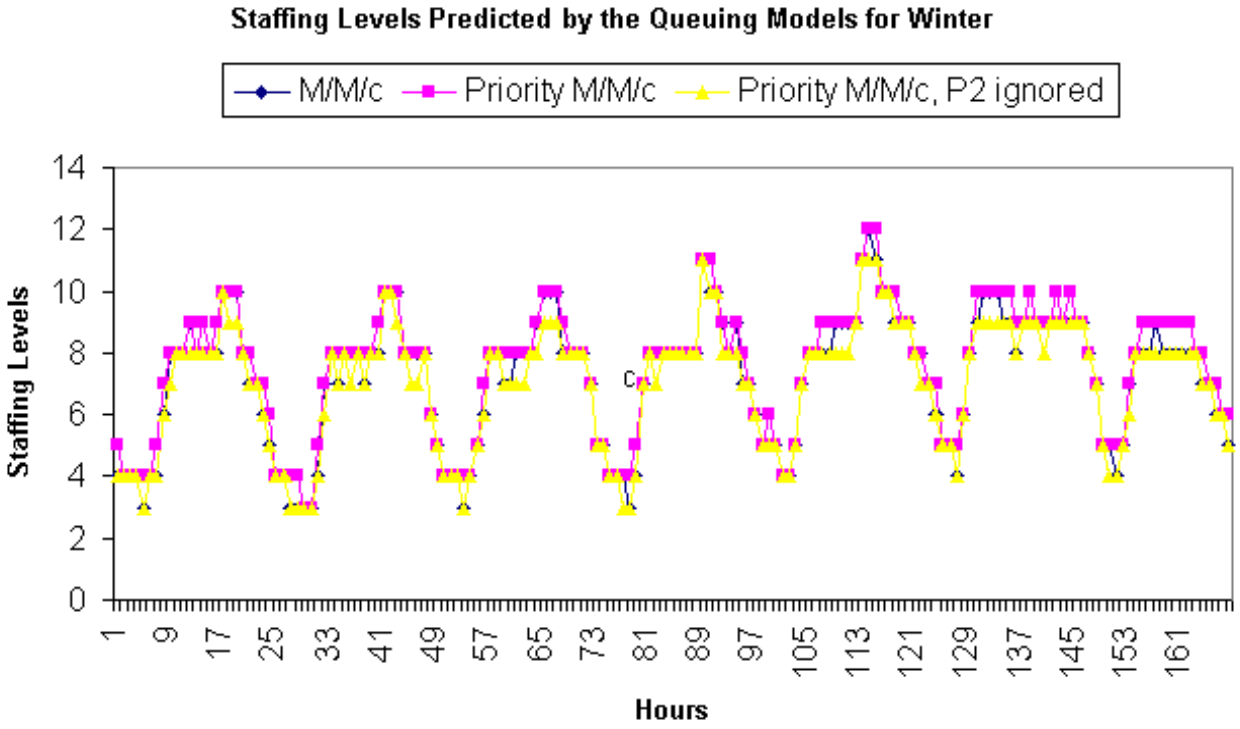

Figure 1. Plot of the staffing levels suggested by our queuing models for winter.

The plot shows the results obtained from the M/M/c queue and the non-preemptive priority M/M/c queue. Since priority 2 calls are less urgent than priority 1 calls, they can tolerate greater delays. In the priority queue model, we used Markov's inequality to bound the priority 2 performance. We looked at two cases where case 1 ensures the bound is satisfactory and a second case that reports bounds without enforcing the priority 2 service requirement. From Figure 1, we can see that case 1 where priority 1 and priority 2 service requirements are enforced has the highest predicted staffing levels. Case 2 has the lowest predicted staffing levels among the three sets of staffing levels. The standard M/M/c queue produce results in between these 2 cases.

Theoretically, staffing levels predicted in the priority model should be less than that of the standard non-priority model. This is because of the fact that all calls are treated as priority 1 in the standard model, whereas in the priority model, calls are divided into priority 1 and priority 2 with the latter having a lower performance target.

Hence we can deduce that Markov's inequality overestimates the required staffing levels in modelling priority queues and better results can be obtained without enforcing the priority 2 service requirement. In fact, the results obtained from the priority queuing model when ignoring P2 service requirement look quite good. It appeared that the performance target of priority 2 calls is met most of the time even though we attempted to ignore it.

4.2 Results and Interpretation from the Workload Allocation Models

From the workload allocation model that used the existing shift structure of the Police Communication Centre, Figure 2 shows the staffing coverage of workforce allocation solved subject to the staffing levels suggested by the priority M/M/c queue, ignoring the priority 2 service requirement in a typical week in winter. The total weekly staffing hours are 1565 for winter and 1593 for summer.

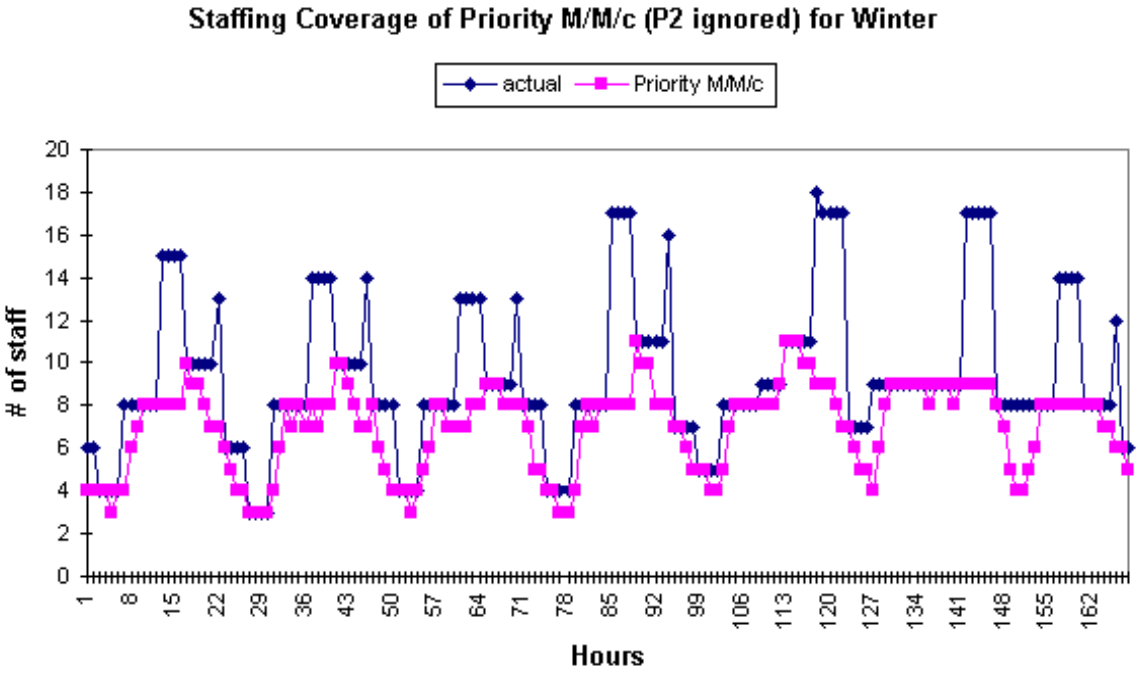

Figure 2. Plot of the staffing coverage of the priority M/M/c queue ignoring the P2 service requirement in a typical week in winter.

We observed a large amount of "over-cover" in the plot. This indicates that the existing shift structure leads to an excessive amount of paid staffing hours if the Police Communication Centre is determined to satisfy their service goals at all times of the day.

From the workload allocation model that considers a more flexible shift structure and a shorter shift length, Figure 3 shows the staffing coverage of workforce allocation solved subject to the staffing levels suggested by the priority M/M/c queue, ignoring the priority 2 service requirement in a typical week in winter. The total weekly staffing hours are 1296 for winter and 1312 for summer.

We observed a much tighter coverage in these plots as compared to the model that uses the existing shift structure. This suggests that moving from the existing shift structure to a more flexible one would lead to a potential reduction in paid staffing hours.

Plots for summer are not shown in here as they produce similar results to winter.

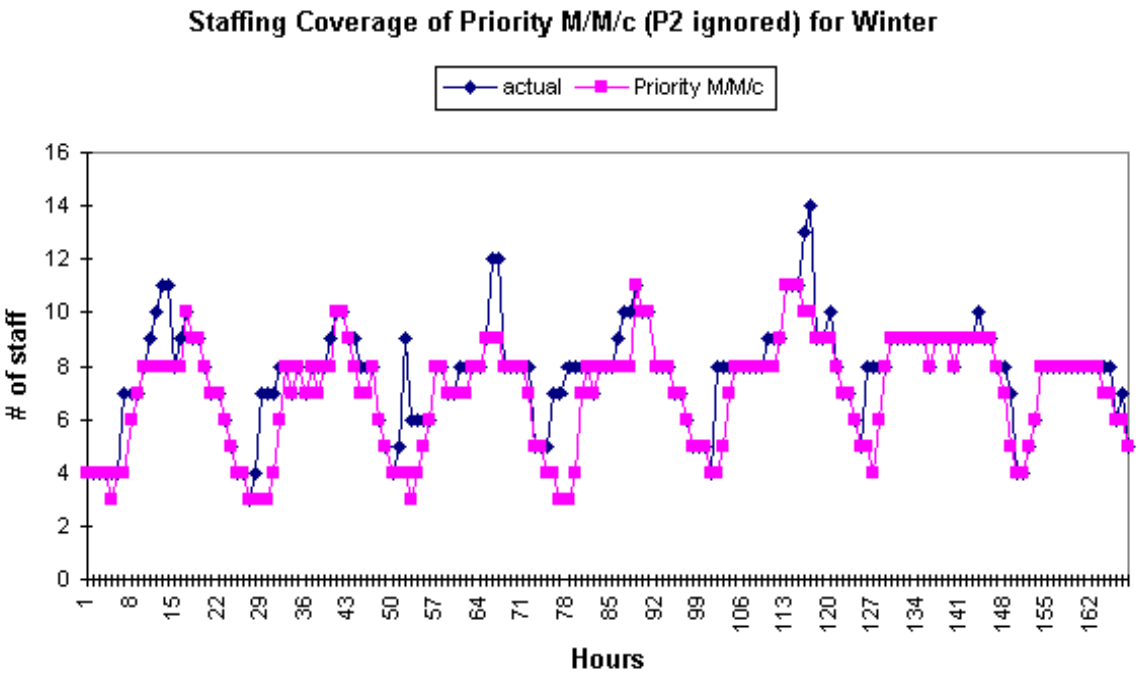

Figure 3. Plot of the staffing coverage of the priority M/M/c queue ignoring the P2 service requirement in a typical week in winter.

Comparing the two workload allocation models of different shift structures, Table 2 summarises the total weekly paid hours for each model. Percentage paid staffing hour reductions for moving to a roster with a more flexible shift structure are also calculated.

	Winter	Summer	Winter	Summer
	M/M/c	M/M/c	Priority M/M/c	Priority M/M/c
Existing	1605	1613	1565	1593
Flexible	1320	1344	1296	1312
% Reductions	17.76%	16.68%	17.19%	17.64%

Table 2. Total staffing hours of rosters and the percentage paid hour reduction moving to a flexible roster.

Notice that the last two columns of Table 2 are the ones we see as most relevant since they contain the results which were obtained from our best model.

We observe that paid staffing hour reductions of up to 17% are possible if we move to a roster that has more flexible starting times and a shorter shift length of 8 hours.

5 Impact and Conclusions

The results obtained in this project have been reviewed with great interest by the Auckland Police Communication Centre. The management is suitably impressed and keen to establish ongoing collaboration with the University of Auckland.

The two queuing models, the M/M/c queue and the non-preemptive priority M/M/c queue, produced reasonable results in finding the hourly staffing levels of a typical week.

The non-preemptive priority M/M/c queue is the model that best simulates the call operations in the Police Communication Centre.

Markov's inequality was shown to be acceptable in modelling priority queues but exact transform inversion would be a better choice. Using Markov's inequality, the results showed a prediction of staffing levels of the priority M/M/c queue slightly greater than that of the standard M/M/c queue. As a result, the number of servers

required to satisfy the performance target was slightly overestimated. Priority 2 service requirement was ignored to remedy this problem.

Subject to the staffing levels as revealed by our queuing models, it was found that the existing roster structure of the Police Communication Centre leads to a large amount of "over-cover". A move to smaller shift lengths and more flexible shift starting times could yield reductions in paid staffing hours of up to 17%.

Acknowledgements

I would like to thank Michael Mann and Roly Williams of the Auckland Police Communication Centre for their support of this project. This research was partially supported by New Zealand Public Good Science Fund grant number UOA 803.

References

[1] Abate, J. and Whitt, W. (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing.* **7** 36–43.

[2] Billingsley, P. (1986) *Probability and Measure.* Second edition. Wiley, New York.

[3] Henderson, S. G. (1999) Setting staffing levels in call centers with random arrival rates. *Under Submission.*

[4] Kella, O. and Yechiali, U. (1985) Waiting times in the non-preemptive priority M/M/c queue. *Commun. Statist. - Stochastic Models.* **1**(2) 257–262.

[5] Ross, S. M. (1983) *Stochastic Processes.* Wiley, New York.

[6] Segal, M. (1974) The operator scheduling problem: a network flow approach. *Operations Research.* **22**(4) 808–823.

[7] Wolff, R. W. (1989) *Stochastic Modeling and the Theory of the Queues.* Prentice Hall, Englewood Cliffs, New Jersey.