A Quantitative Approach to Measure Organisation Performance

Trong B. Tran and Steven R. Davis

***Abstract*—This paper applies a mathematical model for dealing with staffing levels in organisations. The model in the present study is the deterministic reneging queuing model (M/M/c–D) where customers lose patience after a deterministic time and leave the queue without being served. Inter–arrival times of customers and service times follow exponential distributions. Staff in an organisation here are treated as 'servers', while incoming work is considered as 'customers' of the system. Staffing levels are examined from an output or production orientation in situations where a moderate number of workers function as a pool of staff that performs similar operations. Experiments on different staff sizes were conducted to confirm the validity of the model. Both the experimental results and the models indicate that there is an optimum level of staff which maximises the 'profit' of an organisation. The paper provides a practical tool for managers in identifying staffing size for his or her crew. However, the research is limited to the cases where the inter–arrival time of customers and the service time follow exponential distributions, and the maximum acceptable waiting time of customers is deterministic.**

***Index Terms*—Human resource management, overstaffing, queuing model, reneging, understaffing**

I. Introduction

Workforce sizes have different effects on performance of an organisation. However, the literature has not reached an agreement on the relationship between staff levels and organisation performance. Researchers have approached the issue in both qualitative and quantitative methods and come up with different results. Some scholars conclude that moderate understaffing has positive effects on outcomes, while others prove that slight overstaffing performs better. They all agree that, however, both great overstaffing and extreme understaffing conditions have negative effects on organisation performance.

The present study aims to investigate the staff – performance relationship in organisations by conducting experiments on different work group sizes to see the effect on output. The results are then compared with the deterministic reneging queuing model (M/M/c–D). Both the experiments and the model indicate that there is an optimum staffing size that maximises the 'profit' of an organisation.

The paper is organised as follows. Section 2 gives some background of the research including a review of staffing issues and performance in organisations as well as an introduction of queuing theory. In Section 3, lists the basic concepts and equations necessary to evaluate organisation performance through the M/M/c–D model. The experiment is presented in Section 4, while Section 5 gives the comparison between the experimental results and the theory. In Section 6, optimisation is examined. Conclusions are given in Section 7.

Manuscript received September 21, 2011; revised October 10, 2011.

Authors are with the School of Civil and Environmental Engineering, The University of New South Wales, Sydney, Australia (e-mail: tran.trong@student.unsw.edu.au, Tel: +61 2 9385 4290; fax: +61 2 9385 6193)

II. Background

A. Staffing Issues and Performance in Organisations

The choice of staffing level in organisations has occupied the minds of researchers for a long time. However, there is still debate over the effect of overstaffing and understaffing on organisation performance. On one hand, some people advocate a moderate level of understaffing, arguing that this leads to increased organisation output per head [1]–[5]. On the other hand, there is a number of researchers that argue that slight overstaffing leads to improved organisation outcomes [6]–[10].

Proponents of understaffing argue that workers are likely to work more efficiently [1] and experience higher motivation [3]–[5]; hence they produce more in a slight understaffing condition. In understaffed groups, Barker and Gump [1] point out that each person engages in a wider variety of tasks, expends greater effort to achieve organisation goals and takes on more responsibility. Oxley and Barrera [2] and Greenberg [3] suggest that worker motivation may be improved with slight understaffing. Echoing this point, Bechtel [4] and Vecchio and Sussmann [5] conduct empirical studies and show that employee levels and worker motivation have a curvilinear relationship. Their studies conclude that moderate understaffing and worker satisfaction are related.

Supporters of overstaffing emphasise that extra staff can be used to support an organisation in absorbing any environmental turbulence [6], [7], facilitate organisation strategic directions [11], and enhance organisation performance [12]. The existence of extra staff in an organisation allows the organisation to experience new postures in relation to a changing environment. Hence the organisation is more likely to support special projects [6]. Cheng and Kesner [7] state that organisations with additional staff are more likely to respond aggressively to shifting environmental demands. Having human resource slack in an organisation is argued to enable creative behaviour and experimentation with new strategies, such as introducing new products or entering new markets [11]. Knight [13] supports this point and shows that new processes, new products, and new ideas follow when adding staff to an organisation. Researchers acknowledge that increasing staff levels for an organisation means more cost. Hence they suggest that an

organisation should not overstaff greatly [6], [9], [10]. George [10] and Tan [9] conduct studies and confirm the existence of an optimal level of human resource slack in an organisation. In summary, additional staff can be a source of competitive advantage, but too many additional staff may reduce organisation performance [9].

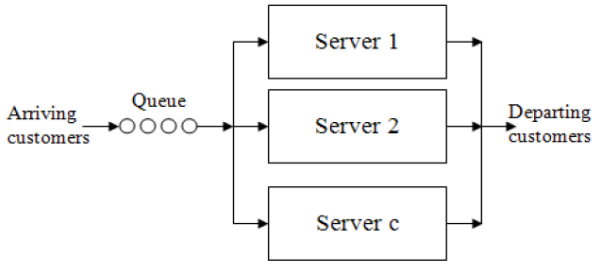

Fig. 1. A queuing system

Criteria for measuring organisation performance are various. Different types of organisations and/or different managers have different requirements for performance. A company may focus on business performance such as gaining profit, extending markets, inventing new products; a government agency may concentrate on satisfying customers; or a rescue team may direct attention to short response times for saving lives are some examples of different aspects of organisation performance [14], [15].

It is hypothesized in this study that overstaffing leads to improved organisation performance compared to understaffing. The research looks at slight overstaffing and moderate understaffing levels only since other conditions, great overstaffing and extreme understaffing, are proven to destroy performance of an organisation [5], [9], [10]. The present work investigates the performance of different sized groups according to two criteria: the probability of service and the 'profit' of the groups. The 'profit' here is defined as the difference between outputs and inputs of an organisation. Extra staff in the study is seen as redundant staff in the reliability sense, not the sacking/hiring sense.

B. Benefits and Costs Associated with Overstaffing

Having extra staff also means having extra benefits as well as extra costs associated with that workforce for an organisation. Such benefits and costs are in both tangible and intangible forms. Some of these benefits/costs are listed as follows.

Benefits:

1) Extra staff creates extra products for an organisation
2) Excess staff helps to buffer an organisation from environmental shocks and gives them freedom in responding to competitors strategies. As a result, extra personnel acts as a resource cushion that stabilises the organisation's operations [10].
3) Using extra staff may provide an organisation with a competitive advantage because it is difficult for competitors to obtain the same resource configurations and copy the organisation's strategy [16].

Costs

1) Extra staff means extra costs, including payroll costs and administration costs [17].
2) Employees tend to have less contribution to their group performance [18].

C. Queuing Models

A queuing system can be described as customers arriving for a service, waiting for service if not immediately served, and leaving the system after being served. Fig. 1 illustrates a queuing system with a single queue and multiple servers. The term *customer* is used in general sense and does not necessarily imply a human customer or that the operation is a purchase.

When observing a queuing system, the following parameters need to be considered [19], [20].

1) *Arrival pattern of customers*: the average rate of customers entering the queuing system λ (and hence the inter–arrival time $t_a = 1/\lambda$) and, if available, its statistical distribution
2) *Service time*: the average time that a server processes a request $t_s = 1/\mu$ (and hence μ is the service rate) and, if available, its statistical distribution.
3) *Queuing capacity*: finite or infinite.
4) *Queuing disciplines*: first in first out (FIFO), last in first out (LIFO), service in random order (SIRO), etc.

When describing a queuing model, part or all of the form [(*a*/*b*/*c*):(*d*/*e*/*f*)] is used, where *a* is the inter–arrival time distribution, *b* is the service time distribution, *c* is number of parallel servers, *d* is the service discipline, *e* is the maximum number of customers allowed in the system, and *f* is the input source.

The symbols *a* and *b* may be M (Markovian distribution), D (deterministic distribution), E_l (Erlang distribution with parameter *l*, $l = 1, 2 \ldots$), *G* (general distribution). The symbol *d* may be FIFO, LIFO or SIRO; and symbols *c*, *e*, and *f* are either a finite number of infinite. When (*d*/*e*/*f*) is omitted, the default meaning is taken as (FIFO/∞/∞).

When a customer arrives at a system it enters the queue and waits until it is served. Alternatively the customer may decide not to join the queue if it is too long. Also, a customer may decide to leave the queue before being served because it will only tolerate a limited waiting time. Avoiding a system because of a long queue is called balking. Losing patience and leaving the queue before being served is called reneging. Leaving after service begins, but before it finishes, is called interruption.

The most common stochastic queuing models assume that the inter–arrival time and service time follow the exponential distribution or, equivalently, that the arrival rate and service rate obeys a Poison distribution [20].

Queuing models are applied widely in human organisations. M/M/c, M/M/c/K, M/G/1, G/G/c, for example, are used in setting up staffing size of a call centre or identifying the number of beds required for a hospital [15], [21]. These models assume that customers do not leave the system until service is finished.

Queuing models for different behaviours of customers also have been developed and applied widely. Boots and Tijms [22], for example, prove an application of an M/M/c model for systems with impatient customers. Barrer [23] and Bocquet [24] suggest the use of an M/M/c queue with impatient customers to model the military situations of selecting the number of artillery pieces for a battle and the number of surface to air missiles required to protect an

airfield. Haight [25] and Rao [26] introduce models that combine reneging, balking and interruption.

When applying a queuing model to a given system, the following information needs to be paid attention to [19]:

1) *Waiting* time for customers in the queue and in the system
2) Queue *length* and number of customers in the system
3) Server idle time
4) *Number* of idle servers
5) *Output* of the operation

III. Modelling

Reneging queuing models are used in evaluating systems with impatient customers [23], [24], [27]. This paper applies the M/M/c–D reneging queuing model, where customers lose patience after a deterministic time τ, and leave the queue without being served [28]. Such customers are lost to the system.

Assumptions:

1) The system is a single queue and multiple servers (as in Fig. 1).
2) Servers are independent of each other.
3) There is no limit in waiting room for incoming customers.
4) There is only one customer arriving at the system at a time.
5) A *server* can serve only maximum one customer at a time.
6) Once a customer enters a server for being served, the customer does not leave the system until the service is finished.
7) The *inter*–arrival time and the service time follow exponential distribution.

Symbols used in following are:

1) $1/\lambda$ average inter–arrival time of customers; λ is the average arrival rate.
2) $1/\mu$ average service time per customer; μ is the average service rate.
3) ρ servicing factor; $\rho = \lambda/\mu$.
4) c number of servers.
5) τ maximum time between arrival and start of service that a customer will wait before reneging.
6) P_0 probability that there are zero customers in the system at a particular time.
7) P_L probability that a customer will not be offered service before time τ and hence renege.
8) P_S probability of service; $P_S = 1 - P_L$
9) Θ system output (customers per unit time).
10) B_p benefit to the organisation of serving one customer
11) C_w cost of one worker
12) N_w number of workers
13) F 'profit'

Several parameters may be considered when measuring system performance: probability of no customers in the system P_0; probability of losing a customer P_L; probability of service P_S; output of the system Θ [24], [27]. The relevant formulae for these follow.

1) The probability of no customers in the system, P_0, is obtained *from* [23]

$$P_0^{-1} = \begin{cases} \sum_{k=0}^{c} \frac{\rho^k}{k!} + \frac{\rho^{c+1}}{(\rho - c)c!}(e^{\mu\tau(\rho-c)} - 1) & (\rho \neq c) \\ \sum_{k=0}^{c} \frac{\rho^k}{k!} + \frac{\mu\tau\rho^{c+1}}{c!} & (\rho = c) \end{cases} \quad (1)$$

2) The probability of losing customers after time τ [23], [24]

$$P_L = \begin{cases} \frac{\frac{\rho^c}{c!}e^{-\mu\tau(c-\rho)}}{\sum_{k=0}^{c-1}\frac{\rho^k}{k!} + \frac{\rho^c\rho e^{-\mu\tau(c-\rho)} - c}{c!}} & (\rho \neq c) \\ \frac{\frac{c^c}{c!}}{\sum_{k=0}^{c-1}\frac{c^k}{k!} + \frac{c^c}{c!}(1+c\mu\tau)} & (\rho = c) \end{cases} \quad (2)$$

3) The *probability* of service

$$P_S = 1 - P_L \quad (3)$$

4) The *output* of the system

$$\Theta = \lambda P_S \quad (4)$$

5) Organisation 'profit' is considered to be made up of two components, namely cost associated with the workers and benefit associated with the output [19].

$$F = B_p\Theta + C_wN_w \quad (5)$$

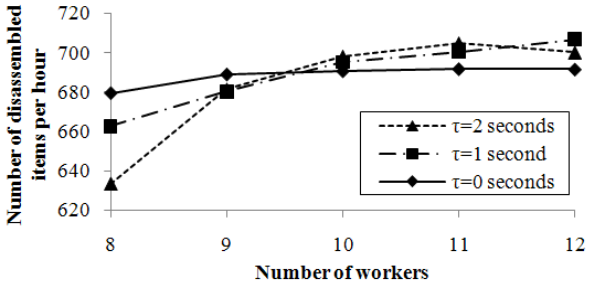

Fig. 2. Hourly outputs of experimental groups (for the cases where interarrival times follwed exponential distribution).

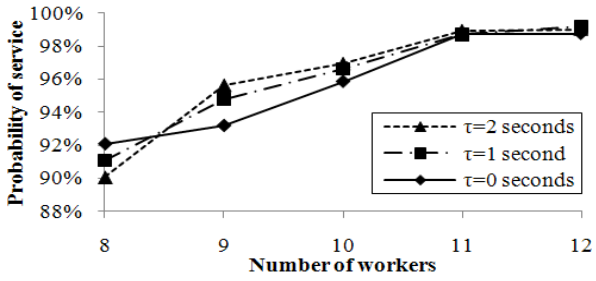

Fig. 3. Probabilities of service in of experimental groups (for the cases where interarrival times follwed exponential distribution).

Given B_p, Θ, and C_w, the optimisation problem becomes one of finding the number of workers N_w that maximises the 'profit' F.

IV. Experiments

A. Experimental Design

Experiments were conducted on work groups where

workers independently disassembled electronic components from circuit boards. The number of workers in experimental groups ranged from 8 to 12.

Inter–arrival times of incoming work were regulated to follow an exponential probability distribution, so as to simulate variable work input. Only one piece of work arrived at a time. Each piece of arriving work was served by a single randomly selected worker. In the case that all workers were busy at the time that work arrived three different scenarios were tested. In the first experiment the work left the system immediately. In the second experiment it waited to be served for up to 1 second before leaving. In the third experiment it waited for up to 2 seconds before leaving. Work leaving was regarded as lost work.

B. Data Collection and Data Validation

Data collected from the experiments included arrival times, waiting times (if applicable), proposed starting times of service, real starting times of service, service times, and finishing times of service. In the experiments for 8–, 9– and 10–worker groups, the experiments stopped when the number of lost customers reached 100. The 11– and 12–worker group experiments stopped when this number reached 25. After being stopped, the experiments restarted from scratch.

C. Results and Analysis

The average inter–arrival time ($1/\lambda$) and the average service time ($1/\mu$) respectively in the three experiments were (in seconds): 5.04, 26.92; 5.07, 27.71; and 5.12, 28.08. As a result the servicing factors, $\rho=\lambda/\mu$, are respectively for the three experiments: 5.34; 5.47; 5.48. This indicates that the optimum number of workers would be 5 or 6 if there was no variability in work input.

The experimental results illustrate what might be anticipated, namely that the probability of service and the output increase when the number of workers increases. Adding an extra worker to a smaller group has a bigger effect than adding an extra worker to an already large group. For example, in the third experiment when the group was enlarged from 8 to 9 workers, the probability of service increased by nearly 6% and the output increased by more than 50 per hour. However, when the number of workers changed from 10 to 11, the increase in probability of service was approximately 2% and the output increased by about 10 per hour. The probability of service and hourly output are plotted in Fig. 3 and Fig. 2 respectively as follows.

V. Comparison of the Experiment and the Model

Given the arrival rate of work λ and the service rate μ, (1) to (4) can be used to calculate P_0, P_L, P_S, and Θ. Fig. 4 and Fig. 5 show the comparison between the experimental results and the queuing model results for P_S and Θ for the τ = 1 second experiment. The results from the other experiments show similar trends.

Correlation coefficients were used to measure how well the queuing model fits the experimental results for both the probability of service and hourly output. Statistical tests for all three experiments produced correlation coefficients in the range 0.91 to 0.997. The p–values were all less than 3% for both parameters implying good agreement with the experimental results at the 95% confidence level.

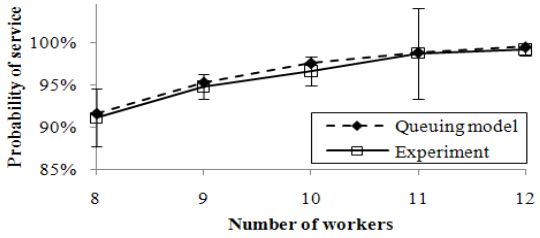

Fig. 4. Comparing probability of service between the experiment and the model for groups in the τ = 1 second and interarrival times followed exponential distribution experiment.

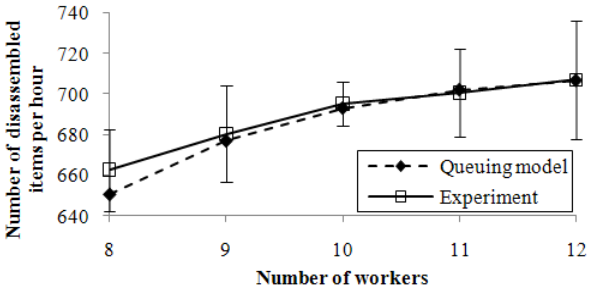

Fig. 5. Comparing hourly output between the experiment and the model for groups in the τ = 1 second and interarrival times followed exponential distribution experiment. Error bars refer to experimental results.

VI. Optimising Staffing Level

When adding workers to a group, the probability of service and the output of the group increase. However, the cost of adding workers may be in excess of the benefits that the additional workers bring to the organisation. The optimisation study in this section aims at finding the number of workers in a group that bring the maximum 'profit' to the group.

The output Θ can be calculated from (4), while B_p and C_w can be estimated from the market. Fig. 4 presents results where the number of workers in the groups range from 1 to 20, λ = 0.1972, μ = 0.0361, τ = 1, B_p= \$0.8 and C_w= –\$25 (the negative sign here indicates that the group needs to outlay money to hire workers). λ, μ and τ are the same values as were used in the second experiment above. C_w was assessed from the labour market for appropriate semi–skilled workers. B_p was estimated as the average price of the electrical components on the second hand market. Figure 5 confirms that there is a size for which the group gains the maximum benefit and beyond that point, the 'profit' will decrease. In this case the optimum number of workers is 8. This is quite different to the value of 5 or 6 that would be derived from the servicing factor, as discussed in Section 4c, for the case of no variability in work input.

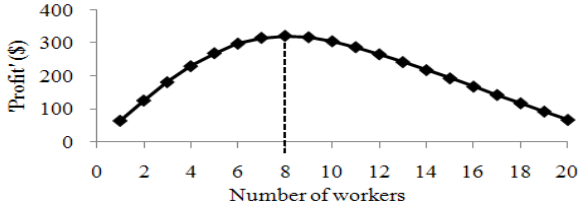

Fig. 6. 'Profit' of groups using base case values

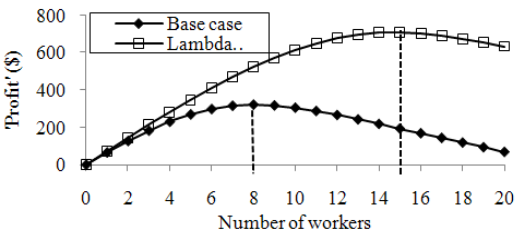

Fig. 7. Sensitivity of 'profit' of groups to doubling λ.

A sensitivity analysis was carried out by varying each of the variables λ, μ, τ, B_p and C_w separately to observe the effect on the optimum number of workers. The example in the previous paragraph is used as the base case. Fig. 7 shows, for example, the effect of doubling λ. In this case the optimum number of workers increases from 8 to 15 and the maximum 'profit' correspondingly nearly doubles.

Fig. 8 shows the sensitivity diagram relating the five variables to the optimum number of workers. The optimum number of workers is most sensitive to λ and least easing λ or B_p, or by decreasing μ or C_w. However, it is unaffected by changing τ.

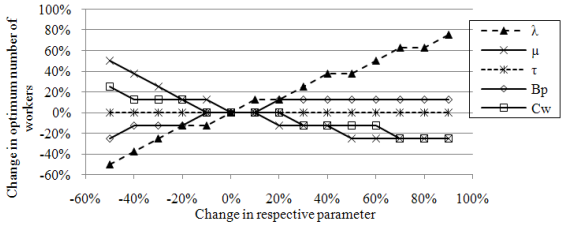

Fig. 8. Sensityvity diagram.

VII. Findings and Conclusions

The study shows that the M/M/c–D model can be used to describe the probability of service and output of working groups when two conditions occur simultaneously. Firstly, the arrival rate of work obeys an exponential distribution. Secondly, the maximum time that a customer will wait between arrival and start of service before reneging is a fixed value. The model may be used to optimise the staffing levels of work groups.

Results from the research also point out that a group should not be greatly overstaffed. Increasing the number of workers creates more output and hence brings more benefits for the organisation. However, the cost of additional workers is constant, while the benefits that each worker brings decreases as the number of workers increases. An optimisation study is recommended for establishing the appropriate number of employees in an organisation.

References

[1] R. G. Barker and P. V. Gump, *Big School, Small School: High School Size and Student Behavior*: Stanford University Press, 1964.
[2] D. Oxley and J. M. Barrera, "Undermanning Theory and the Workplace – Implication of setting size for job satisfaction and social support," *Environment and Behavior*, vol. 16, pp. 211–234, 1984.
[3] C. I. Greenberg, "Toward an integration of ecological psychology and industrial psychology: Undermanning theory, organization size, and job enrichment," *Journal of Nonverbal Behavior*, vol. 3, pp. 228–242, 1979.
[4] R. B. Bechtel, "The undermanned environment: A universal theory?," in *Man environment interactions (Part VIII)*, D. H. Carson, Ed., ed Stroudsburg, Penn.: Dowden, Hutchinson and Ross, 1974.
[5] R. P. Vecchio and M. Sussmann, "Staffing sufficiency and job enrichment: Support for an optimal level theory," *Journal of Occupational Behaviour*, vol. 2, pp. 177–187, 1981.
[6] L. J. Bourgeois, "On the Measurement of Organizational Slack," *The Academy of Management Review*, vol. 6, pp. 29–39, 1981.
[7] J. L. C. Cheng and I. F. Kesner, "Organizational Slack and Response to Environmental Shifts: The Impact of Resource Allocation Patterns," *Journal of Management*, vol. 23, pp. 1–18, February 1, 1997.
[8] J. R. Galbraith, *Designing complex organizations*. Boston, MA, USA: Addison–Wesley Longman Publishing Co, Inc, 1973.
[9] J. Tan, "Curvilinear Relationship Between Organizational Slack and Firm Performance: Evidence from Chinese State Enterprises," *European Management journal*, vol. 21, pp. 740–749, 2003.
[10] G. George, "Slack Resources and the Performance of Privately Held Firms," *Academy of Management Journal*, vol. 48, pp. 661–676, 2005.
[11] J. D. Thompson, *Organisations in action*. New York: McGraw–Hill, 1967, D. C. Hambrick and C. C. Snow, "A contextual model of strategic decision making in organizations," *Academy of Management Proceedings*, vol. 37, pp. 109–112, 1977.
[12] K. G. Rust and J. P. Katz. (2002, Organizational Slack and Performance: The Interactive Role of Workforce Changes. [Working Paper]. Available: http://www.midwestacademy.org/Proceedings/2002/papers/Rust.doc
[13] K. E. Knight, "A Descriptive Model of the Intra–Firm Innovation Process," *The Journal of Business*, vol. 40, pp. 478–496, 1967.
[14] J. T. Delaney and M. A. Huselid, "The Impact of Human Resource Management Practices on Perceptions of Organizational Performance," *The Academy of Management Journal*, vol. 39, pp. 949–969, 1996, P. M. Wright, T. M. Gardner, L. M. Moynihan, and M. R. Allen, "The relationship between HR practices and firm performance: examining casual order," *Personnel Psychology*, vol. 58, pp. 409–446, 2005.
[15] R. W. Hall and L. Green, "Queueing Analysis in Healthcare," in *Patient Flow: Reducing Delay in Healthcare Delivery*. vol. 91, ed: Springer US, 2006, pp. 281–307.
[16] Y. Mishina, T. G. Pollock, and J. F. Porac, "Are more resources always better for growth? Resource stickiness in market and product expansion," *Strategic Management Journal*, vol. 25, pp. 1179–1197, 2004.
[17] G. B. Voss, D. Sirdeshmukh, and Z. G. Voss, "The Effects of Slack Resources and Environmental Threat on Product Exploration and Exploitation," *Academy of Management Journal*, vol. 51, pp. 147–164, 2008.
[18] A. Wicker, S. Kirmeyer, L. Hanson, and D. Alexander, "Effects of manning levels on subjective experiences, performance, and verbal interaction in groups," *Organizational Behavior and Human Performance*, vol. 17, pp. 251–274, 1976.
[19] D. G. Carmichael, *Engineering queues in construction and mining*. Chichester: Ellis Horwood Limited, 1987.
[20] D. Gross, *Fundamentals of queueing theory*: Wiley–India, 2008.
[21] G. Koole and A. Mandelbaum, "Queueing models of call centers: An introduction," *Annals of Operations Research*, vol. 113, pp. 41–59, 2002, S. Fomundam and F. Herrmann, "A survey of queuing theory applications in healthcare," The institute for system research2007.
[22] N. K. Boots and H. Tijms, "An M/M/c queue with impatent customers," *Sociedad de Estadística e Investigación Operativa*, vol. 7, pp. 213–220, 1999.
[23] D. Barrer, "Queuing with impatient customers and ordered service," *Operations Research*, vol. 5, pp. 650–656, 1957.
[24] S. Bocquet, *Queueing Theory with Reneging*. Victoria: Defence Science and Technology Organisation, 2005.
[25] F. Haight, "Queuing with reneging," *Metrika*, vol. 2, pp. 186–197, 1959.
[26] S. S. Rao, "Queuing models with balking, reneging, and interruptions," *Operations Research*, vol. 13, pp. 596–608, 1965.
[27] B. V. Gnedenko and I. N. Kovalenko, *Introduction to queueing theory*. Jerusalem: S. Monson, 1968.
[28] D. Worthington, "Reflections on queue modelling from the last 50 years," *Journal of the Operational Research Society*, vol. 60, pp. S83–S92, 2009.