6

Multi-media Call Centres

M. Heiliö, M.J. Luczak, J.C.W. van Ommeren, W.R.W. Scheinhardt

Abstract

We consider a multi-media call centre answering customer requests as a queueing network with priority classes. We use differential equations to analyse its performance as the system becomes large and approaches a 'fluid limit'. We also suggest how to approximate average delays using existing algorithms for the classical $M/G/c$ queueing model.

Keywords

Multi-media call centre, queueing system, differential equation, waiting time, priority class.

6.1 Introduction

A call centre is a location employing a certain number of agents to deal with customer queries concerning specific types of product or service. At a multi-media call centre people submit their requests using a variety of distant communication means, such as telephone, fax, email or the Internet. Occasionally a single round of communication (that is, a single call) is insufficient: the operator may have to first search for an answer to the query and then return the call.

Imagine such a busy call centre that handles numerous requests of various types each day. A good example would be a telephone exchange that looks up domestic and international phone numbers, facilitates telephone connections, provides information on telephone bills, deals with customer complaints, and is engaged in network trouble-shooting. The key issue is that of performance, often measured as the amount of time that customers need to wait for service. One would like to ensure

that all requests are dealt with by an operator within a 'reasonable' time. The term 'reasonable' will typically assume different meanings depending on the type of task or query involved. A fax or email could take longer to process than a telephone call. Incoming telephone calls are given priority over most other tasks; e-mails or faxes may be interrupted in order to deal with fresh phone calls, though one would not normally interrupt an outgoing (that is, 'return') call. There may also be certain absolute quality levels, imposed to ensure that each request be served within a specific time limit.

In Section 6.2 of this report we construct a performance model for a multi-media call centre. The setting is a system of servers (operators) and tasks queueing at the servers. Given a number of operators with different skills, several request types characterised by different arrival rates, durations, and service level requirements, the problem is to decide the schedule and priorities so as to minimise the operating cost in the system. The cost function can be the number of operators, their total salary (operators able to deal with more kinds of tasks are paid more money), costs of training operators to enable them to perform more tasks, fines paid to customers when quality requirements are not satisfied, or a combination of all of the above. We choose to focus solely on the mean waiting time as a function of the number and speed of servers and task arrival rates.

In Section 6.3 we describe an alternative approach to studying the performance of multi-media call centres, however only as a possible onset of future research. The basis for this section is the well-known $M/G/c$ queueing system where, unlike in the first model, task service times need not be exponentially distributed.

6.2 Call centre as a large queueing network

In this section we analyse a multi-media call centre as a queueing network. There are n servers, representing people who work at the call centre. Under appropriate scaling, let $n\lambda_1$ be the arrival rate of phone calls; $n\lambda_2$ the arrival rate of faxes; $n\lambda_3$ the arrival rate of e-mails. In our model the scheduler sends each arriving task to a randomly chosen server. If there are several types of agents, each trained to deal with only a subclass of tasks, every new request is handed over to a randomly chosen server among those who are able to work on it.

We do not explicitly distinguish between inbound calls (that is, fresh requests and queries) and outbound calls (calls made in response to earlier requests and queries). The distinction is incorporated implicitly into the total serving time. It is possible to model interruptions in service as 'pre-emptions' by higher priority jobs, and this is what we shall do in the sequel. We could also deal with performance issues such as different delay requirements for different types of requests, and we could model several different priority scheduling strategies. Here we only consider the case where all delay requirements are the same, and in fact we do not overtly try to minimise the delay as such; we look exclusively at the mean waiting time. We further concentrate

on one particular scheduling strategy, described in the subsequent paragraph.

Assume that telephone calls have absolute priority over faxes and e-mails, and faxes have absolute priority over e-mails. This means that whenever a telephone call arrives at a server currently dealing with a fax or email, this task immediately gets postponed ('pre-empted') while the server answers the phone call. Similarly, if a fax arrives while a server is busy answering an email query, this job gets pre-empted, and the agent answers the fax instead. We further suppose that arrivals of each type of task form a Poisson process; the arrival rates are $\lambda_1, \lambda_2, \lambda_3$ for phone calls, faxes and e-mails respectively. Task durations are assumed to be exponentially distributed with means $1/\mu_1, 1/\mu_2, 1/\mu_3$ respectively.

We define the following family of random variables. Let n_{a,i_1,i_2,i_3} be the number of servers with at least i_1 tasks of type a (phone calls), at least i_2 tasks of type b (faxes), at least i_3 tasks of type c (e-mails), and currently busy serving a type a task. The variables $n_{b,i_1,i_2,i_3}, n_{c,i_1,i_2,i_3}$ are defined similarly. Then the vector

$$\mathbf{n} = (n_{a,i_1,i_2,i_3}, n_{b,i_1,i_2,i_3}, n_{c,i_1,i_2,i_3} : i_1, i_2, i_3 \geq 0)$$

forms a Markov process. The transitions of the process can be written down as vectors $\pm n^{-1}\mathbf{e}_{a,i_1,i_2,i_3}, \pm n^{-1}\mathbf{e}_{b,i_1,i_2,i_3}, \pm n^{-1}\mathbf{e}_{c,i_1,i_2,i_3}$, where $\mathbf{e}_{a,i_1,i_2,i_3}$ is a vector with a one in the position corresponding to co-ordinate n_{a,i_1,i_2,i_3} and zeros everywhere else, and the other definitions are analogous. One could write down explicitly the Kolmogorov forward equations [5] determining the time evolution of this process. This would in turn yield 'balance equations' [5] for the stationary distribution of the process. However, the size of the state space increases rapidly with the number of servers, which makes it impractical to compute this stationary distribution, even for relatively small systems. The key problem is that the probabilities for each state include a normalising factor which ensures that all the probabilities sum to one. As the system grows this factor quickly becomes hard to compute efficiently.

Therefore it is useful to look at what happens in the limit as the number of servers n tends to infinity. The parameters λ_i, μ_i embody the (appropriately scaled) relation between the number of servers, their speed, and the arrival rate of tasks; varying λ_i, μ_i will tell us how many servers we need in large finite systems and how fast they should operate. The deterministic system obtained as the limit of the finite systems as the size n tends to infinity is a 'law-of-large numbers' type of limit. Earlier work and simulations of related models [4, 6, 7, 8, 9, 11, 12, 13] show that this sort of approximations are quite accurate even for relatively small n.

We shall now establish limiting differential equations for the pre-emptive priority service model outlined above. Similar equations could be obtained for other kinds of scheduling policies. Define $\mathbf{x} = \frac{1}{n}\mathbf{n}$, a re-scaled version of the process $\mathbf{n}$. Then $\mathbf{x}$ is also a Markov process, and moreover we have

$$0 \leq x_{a,i_1,i_2,i_3}, x_{b,i_1,i_2,i_3}, x_{c,i_1,i_2,i_3} \leq 1, \qquad i_1, i_2, i_3 \in \mathbb{N}.$$

Note also that $x_{b,i_1,i_2,i_3} = 0$ whenever $i_1 > 0$ and $x_{c,i_1,i_2,i_3} = 0$ whenever $i_1 > 0$ or $i_2 > 0$.

In the limit as $n \to \infty$ one would expect the process $\mathbf{x}$ to become deterministic and approach the solution of the system of differential equations presented below.

$$\begin{aligned}\frac{dx_{a,i_1,i_2,i_3}}{dt} &= \lambda_1(x_{a,i_1-1,i_2,i_3} - x_{a,i_1,i_2,i_3}) \\ &\quad +\lambda_2(x_{a,i_1,i_2-1,i_3} - x_{a,i_1,i_2,i_3}) \\ &\quad +\lambda_3(x_{a,i_1,i_2,i_3-1} - x_{a,i_1,i_2,i_3}) \\ &\quad -\mu_1(x_{a,i_1,i_2,i_3} - x_{a,i_1+1,i_2,i_3}), \qquad i_1, i_2, i_3 > 0,\ i_1 \neq 1.\end{aligned}$$

When $i_1 > 1$, and $i_2 = 0$ or $i_3 = 0$, the above is also valid, with the convention that $x_{a,i_1,-1,i_3} = x_{a,i_1,i_2,-1} = 1$.

For $i_1 = 1$, $i_2 > 0$, the equation becomes

$$\begin{aligned}\frac{dx_{a,1,i_2,i_3}}{dt} &= \lambda_2(x_{a,1,i_2-1,i_3} - x_{a,1,i_2,i_3}) + \lambda_3(x_{a,1,i_2,i_3-1} - x_{a,1,i_2,i_3}) \\ &\quad -\mu_1(x_{a,1,i_2,i_3} - x_{a,2,i_2,i_3}) + \lambda_1 x_{b,0,i_2,i_3}.\end{aligned}$$

The term $\lambda_1 x_{b,0,i_2,i_3}$ accounts for the increase in $x_{a,1,i_2,i_3}$ when there are no type a tasks in the system, a type a task arrives, and 'pre-empts' a type b task currently being served.

For $i_1 = 1$, $i_2 = 0$, $i_3 > 0$ we have

$$\begin{aligned}\frac{dx_{a,1,0,i_3}}{dt} &= \lambda_3(x_{a,1,0,i_3-1} - x_{a,1,0,i_3}) - \mu_1(x_{a,1,0,i_3} - x_{a,2,0,i_3}) \\ &\quad +\lambda_1 x_{b,0,0,i_3} + \lambda_1 x_{c,0,0,i_3}.\end{aligned}$$

For $i_1 = 1$, $i_2 = i_3 = 0$ we obtain

$$\frac{dx_{a,1,0,0}}{dt} = -\mu_1(x_{a,1,0,0} - x_{a,2,0,0}) + \lambda_1 x_{b,0,0,0} + \lambda_1 x_{c,0,0,0}.$$

Similarly we derive equations governing the time development of the fraction of servers busy working on type b tasks:

$$\begin{aligned}\frac{dx_{b,0,i_2,i_3}}{dt} &= \lambda_2(x_{b,0,i_2-1,i_3} - x_{b,0,i_2,i_3}) + \lambda_3(x_{b,0,i_2,i_3-1} - x_{b,0,i_2,i_3}) \\ &\quad -\mu_2(x_{b,0,i_2,i_3} - x_{b,0,i_2+1,i_3}) - \lambda_1 x_{b,0,i_2,i_3}, \qquad i_2, i_3 > 0, i_2 \neq 1.\end{aligned}$$

For $i_2 = 1$, $i_3 > 0$, the equation becomes

$$\begin{aligned}\frac{dx_{b,0,1,i_3}}{dt} &= \lambda_2(x_{b,0,0,i_3} - x_{b,0,1,i_3}) + \lambda_3(x_{b,0,1,i_3-1} - x_{b,0,1,i_3}) \\ &\quad -\mu_2(x_{b,0,1,i_3} - x_{b,0,2,i_3}) + \lambda_2 x_{c,0,0,i_3}.\end{aligned}$$

For $i_2 = 1$, $i_3 = 0$, we get

$$\frac{dx_{b,0,1,i_3}}{dt} = \lambda_2(x_{b,0,0,0} - x_{b,0,1,0}) - \mu_2(x_{b,0,1,0} - x_{b,0,2,0}) + \lambda_2 x_{c,0,0,0}.$$

Finally the appropriate equations for servers busy serving type c tasks are as follows:

$$\begin{aligned}\frac{dx_{c,0,0,i_3}}{dt} &= \lambda_3(x_{c,0,0,i_3-1} - x_{c,0,0,i_3}) - \mu_3(x_{c,0,0,i_3} - x_{c,0,0,i_3+1}) \\ &\quad -\lambda_1 x_{c,0,0,i_3} - \lambda_2 x_{c,0,0,i_3}, \qquad i_3 \geq 0.\end{aligned}$$

In the above, we assume an infinite 'waiting room', that is no request gets lost from the system. With this assumption the obvious stability condition is

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} + \frac{\lambda_3}{\mu_3} < 1.$$

Equally, we could assume that the queue size at each server can be at most B for some $B \in \mathbb{N}$. This would result in truncating the equations to the appropriate polytope where the capacity constraints $i_1 + i_2 + i_3 \leq B$ are satisfied (arrival rates would be zero at servers where $i_1 + i_2 + i_3 = B$).

The above constitutes a system of linear differential equations. A natural strategy would be to calculate its fixed points and examine the stability of these fixed points. This would enable one to analyse the equilibrium (stationary or long-term) behaviour of the system. When the stability conditions are met, then each finite n-server queueing system is an irreducible Markov process with a unique stationary distribution, and converges to that distribution as time tends to infinity. (Observe that a finite capacity system is always stable.) Therefore under these conditions one would expect the limiting differential equations to have a unique, globally attractive, fixed point. However, this needs further theoretical study as well as computer simulations. Suitable ideas and techniques can be found in [2, 4, 7, 8, 9, 11, 12, 13]. Additionally it should be possible to prove that the finite random process with n servers converges as n tends to infinity, in a suitable sense, to the deterministic differential equations [1, 3, 4, 8, 9, 11, 12, 13].

Note that the mean number of customers per server in the system is given by

$$\begin{aligned}&\sum_{i_1 \geq 1} i_1(x_{a,i_1-1,0,0} - x_{a,i_1,0,0}) + \sum_{i_2 \geq 1} i_2(x_{a,0,i_2-1,0} - x_{a,0,i_2,0}) \\ &+ \sum_{i_3 \geq 1} i_3(x_{a,0,0,i_3-1} - x_{a,0,0,i_3}) + \sum_{i_2 \geq 1} i_2(x_{b,0,i_2-1,0} - x_{b,0,i_2,0}) \\ &+ \sum_{i_3 \geq 1} i_3(x_{b,0,0,i_3-1} - x_{b,0,0,i_3}) + \sum_{i_3 \geq 1} i_3(x_{c,0,0,i_3-1} - x_{c,0,0,i_3}).\end{aligned}$$

One could calculate the mean waiting time for each customer class using Little's formula [14].

6.3 Call centre as an $M/G/c$ queueing system

In this section we introduce another line of research that may be pursued when modelling multi-media call centres. This certainly is not an in-depth study, but rather an

outline of some general ideas and possible direction(s) where one could expect at least some results.

As in the previous section we concentrate on delay performance; in particular on approximating the average delay experienced by various types of calls. At this stage it is impossible to determine the degree to which our results are practically relevant and contribute to a deeper understanding of the problem.

The basic building block of our analysis will be the multi-server queueing system with Poisson inputs, also known as the $M/G/c$ system. Here M stands for the Poisson arrival process, G denotes the general service time distribution, and the constant c denotes the number of servers active (all servers are assumed identical). In the light of earlier studies of various communication systems, the Poisson assumption (or approximation) appears plausible. It aids the analysis considerably, if only because dividing and combining of various customer classes (see below) makes no impact on the nature of the arrival processes involved.

Presently we do not look at the asymptotic behaviour as the number of servers becomes very large. We do not use differential equations to estimate call waiting times. Instead, we suggest that the expected waiting times of calls or customers in the system be approximated with the help of powerful algorithms contained in Tijms [10].

One of the key features characterising a multi-media call centre is the presence of several different types of customers, who thus fall into different priority classes. In particular, we shall consider an $M/G/c$ system with two types of customers. If there is a need to distinguish between more than two customer types, initially several classes may be combined into one 'multi-class'. Such a multi-class may be partitioned and analysed in more detail at a later stage.

In what follows we shall assume that whenever necessary 'high priority' customers (calls) may interrupt the service of 'low priority' customers (calls). Typically, high priority calls correspond to service requests that come in by telephone or other 'real-time' media, such as Internet chat sessions or real-time video. Low priority calls represent requests that do not require immediate service, since they arrive in the form of fax or email. Both high and low priority calls are characterised by their own Poisson arrival process and service time distribution.

6.3.1 High priority calls

Assume that a high priority call interrupts a low priority call if on its arrival all the servers are busy and at least one low priority call is in service. Thus high priority calls may now be regarded in complete isolation, separate from low priority calls, since obviously low priority calls have no influence on them. This implies that we can analyse high priority calls using the standard single-class $M/G/c$ model. Real data may be used to fit in an appropriate service time distribution G. When more than one call types together constitute the high priority class, the total service distribution

is a mixture of the individual service time distributions. Algorithms described in Tijms [10] should yield reasonable approximations to the expected waiting time of high priority calls.

As an aside we note that it might be useful to include server 'vacations' (that is, idle periods) into the model. In this way one would account for short-term variations in the number of available servers due to coffee breaks, etc. Vacations could be modelled as 'extra' customers of the highest priority type.

6.3.2 Low priority calls

We now outline very briefly three different approaches that may be used to analyse low priority calls. The reader is referred to Tijms [10] for more details.

1. One possibility is to consider the standard $M/G/c$ queueing system in which the speed of each server is decreased by a factor $1 - \rho_{\text{hp}}$. Here ρ_{hp} is the stationary mean fraction of time that each server is busy serving high priority calls (when the system is stable, then of course $0 < \rho_{\text{hp}} < 1$). The rationale follows easily from the analysis of high priority calls presented in the previous subsection. However, this is a rather crude approximation, likely to lead to over-optimistic estimates, since it takes no account of variations in the service rate available to low priority calls.

2. Alternatively, we could analyse low priority calls as an $M/G/c$ queueing system with vacations caused by the service of high priority calls. The main drawback of this approach lies in its total ignorance of the fact that vacations do not arrive according to a Poisson process independent of the state of the queue.

3. Finally, one could try to do more justice to the individual nature of various types of low priority calls. However, for this we need an assumption that contradicts the basic assumption made in the analysis of high priority calls, namely that high priority calls should *not* be allowed to interrupt low priority calls already being served. That is we now impose a non-preemptive priority service discipline. Therefore, once again one can only hope for approximate results; in practical situations one may decide whether the error size is satisfactory. Another necessary condition is that high and low priority calls must have the same service time distribution. When this holds, then the waiting time distribution (for both high and low priority calls) is the same as the waiting time distribution of a customer in an ordinary $M/G/c$ queueing system without priorities and with first come, first served service discipline. The arrival rate is simply the sum of the arrival rates over all customer classes. This model will provide approximations for EW, the expectation of the waiting time of a call whose type is unknown. Then, by conditioning on the type of call, we obtain that

$$EW = \frac{\rho_{\text{lp}}}{\rho_{\text{lp}} + \rho_{\text{hp}}} EW_{\text{lp}} + \frac{\rho_{\text{hp}}}{\rho_{\text{lp}} + \rho_{\text{hp}}} EW_{\text{hp}}.$$

Hence, using the expression for EW_{hp} found under the pre-emptive resume service regime, we get an approximate formula for EW_{lp}. Earlier studies of related systems suggest that the resulting error is unlikely to be excessively large, particularly if the system consists of many servers.

Bibliography

[1] Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

[2] Deimling, K. (1977). *Ordinary Differential Equations in Banach Spaces*. Springer-Verlag. Lecture Notes in Mathematics. Vol. 96.

[3] Ethier, S.N. and Kurtz, T.G. (1986). *Markov Processes: Characterisation and Convergence*. Wiley, New York.

[4] Gibbens, R.J., Hunt, P.J., and Kelly, F.P. (1990). Bistability in Communication Networks. In *Disorder in Physical Systems* (G.R. Grimmett and D.J.A. Welsh, eds.) 113-128. Oxford University Press.

[5] Grimmett, G.R. and Stirzaker, D. (1992). *Probability and Random Processes*. Oxford University Press, Oxford.

[6] Hunt, P.J. (1990). Limit Theorems for Stochastic Loss Networks. Ph.D. thesis, Cambridge.

[7] Kelly, F.P. (1991). Loss Networks. *The Annals of Applied Probability* **1** 319-378.

[8] Luczak, M.J. (2000). Probability, Algorithms and Telecommunication Systems. Ph.D. thesis, Oxford.

[9] Mitzenmacher, M. (1996). The Power of Two Choices in Randomised Load Balancing. Ph.D. thesis, Berkeley.

[10] Tijms, H.C. (1994). *Stochastic models – an algorithmic approach*. Wiley, Chichester.

[11] Turner, S.R.E. (1998). The Effect of Increasing Routing Choice on Resource Pooling. *Probability in the Engineering and Informational Sciences* **12** 109–124.

[12] Vvedenskaya, N.D., Dobrushin, R.L., and Karpelevich, F.I. (1996). Queueing System with Selection of the Shortest of Two Queues: An Asymptotic Approach. *Problems of Information Transmission* **32** 15–27.

[13] Whitt, W. (1985). Blocking When Service is Required From Several Facilities Simultaneously. *AT & T Technical Journal* **64** 1807-1856.

[14] Wolff, R. (1989). *Stochastic Modelling and the Theory of Queues*. Prentice-Hall, New Jersey.

[13] Watt, W. (1985) Blocking When Service is Required From Several Facilities Simultaneously. AT & T Technical Journal 64 1807-1856.

[14] Wolff, R. (1989) Stochastic Modelling and the Theory of Queues. Prentice-Hall, New Jersey.