Queuing theory applied to the optimal management of bank excess reserves

Taufemback, Cleiton and Da Silva, Sergio

Federal University of Santa Catarina

2011

Queuing theory applied to the optimal management of bank excess reserves

Cleiton Taufemback, Sergio Da Silva

Graduate Program in Economics, Federal University of Santa Catarina,
Florianopolis SC 88049-970, Brazil

Abstract. Although the economic literature on the optimal management of bank excess reserves is age-old and large, here we suggest a fresh, more practical approach based on queuing theory.

Keywords: commercial bank excess reserves; queuing theory

PACS: 89.65.Gh Economics; econophysics, financial markets, business and management; 89.20.-a Interdisciplinary applications of physics

1. Introduction

Commercial bank reserves are funds held by depository institutions, which can be used to meet the institution's legal reserve requirement. In the United States these funds are held either as balances on deposit at the Federal Reserve or as cash in the bank's vault. Reserves that are applied toward an institution's legal requirement are called required, while additional reserves, if any, are called excess.

Banks face a trade-off when deciding their levels of excess reserves. The more the excess reserves, the less the risks of bankruptcy in case of a bank run; but this also means lower profits in terms of reduced loans [1].

Strictly speaking, banks do not create money based on the reserve of cash they keep. Banks create electronic money based on legal reference, not the reserve. For example, for any amount of cash on deposit, banks are legally allowed to create, say, 90 percent of that amount as new electronic money. Thus, if one deposited \$100 in a bank, the bank would be allowed to create (through loans) \$90 new dollars. It is impossible for everyone to possess all their money as cash at the same time because the total amount of cash is less than the total amount of the two types of money, cash and electronic.

Figure 1 shows the excess reserves of American depository institutions. The monthly data are in billions of dollars and range from 1st January 1959 to 1st December 2010. As can be seen, the quantity of reserves in the U.S. banking system has risen dramatically since September 2008. However, because the quantity of bank reserves is determined by the size of the Federal Reserve's policy initiatives, the recent rise is unlikely to be related to commercial bank lending. Also, such large increase in bank reserves need not be inflationary, because the payment of interest on reserves allows the Federal Reserve to adjust short-term interest rates independent of the level of reserves [2]. For decades, holders of liabilities of banks in the United States had felt secure with the protection of the modest equity-capital cushion at 10 percent, allowing banks to lend lavishly. From September 2008 onwards, however, investors seemed to require 14 percent capital rather than the long-standing 10 percent [3].

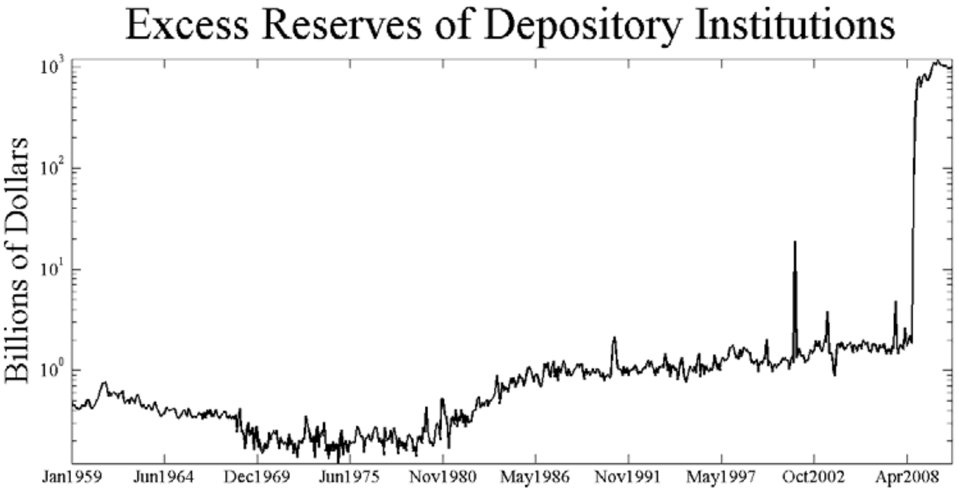

Figure 1. Historical values of excess reserves of depository institutions in the United States. Monthly data from January 1959 to December 2010. Values are shown in logarithms. Source: Board of Governors of the Federal Reserve System.

The economic literature on the optimal management of excess bank reserves is age-old and large, the issue being debated since the banking crisis of the early 1930s [4]. The problem is usually tackled using econometrics (see for instance Refs. [5−7] and further references therein). Here in this paper, we contribute to the literature by suggesting a new, more practical approach based on queuing theory. The probabilistic tools originated in queuing theory have many traditional applications in areas, such as telecommunications and traffic engineering but, as far as we know, they have not been applied to economic and financial problems hitherto.

Thus, in Section 2 we present a brief introduction to queuing theory and show the deduction of the famous Erlang B and C formulas. In Section 3 we show how queuing theory can be applied for the excess reserves problem and in Section 4 we conclude the study.

2. Queuing theory

Working with telephone lines, Agner K. Erlang showed in 1917 [8] that traffic requests follow a Poisson process, that is, calls are independent and the activity of a channel is exponentially distributed (the longer the duration of a call, the likelier for the user to finish the connection), and the probability of a user to reach a situation where no more channels are available can be determined by some characteristics of the system, such as the number of channels and the average duration of a call. Erlang's work prompted the development of queuing theory [9].

The most important tools of queuing theory are the Erlang B and Erlang C formulas. The Erlang B and C formulas are generally used to define the probability that a user cannot make use of a resource at a given time. In the case of telephone lines, this means the probability of no free lines, that is,

$$P(\text{blocking}) = P(\text{all channels are in use})\,. \tag{1}$$

Using Kendall's [9] notation, the Erlang B is described as an $M/M/C/C$ system and the Erlang C is an $M/M/C/\infty$ (Table 1). The main difference between the Erlang B and C formulas is that in the Erlang C the user is allowed to wait for sometime to access the resource, so there is no limit of users.

Table 1. Kendall's notation for the Erlang B and Erlang C.

M	Exponentially distributed inter-arrival times (Poisson process)
M	Exponential service time distribution
C	Number of servers in the queue
C/∞	Maximum number of customers who can be there in the system

As for the Erlang B for telephone systems it is possible to describe a continuous system with discrete observations under particular conditions using Markov chains [10]. Here, the idea is to sample in a δ time interval, where δ is a small positive number. If N_k, or $N(k\delta)$, is the number of busy channels at time $k\delta$, then N_k can be described as a discrete Markov chain, and $N_k \in [0, C]$. Thus, the probability of state transition $P_{i,j}$ is given by

$$P_{i,j} = P\{N_{k+1} = j \mid N_k = i\}. \qquad (2)$$

Figure 2 shows the state diagram of this system where λ is the average arrival rate, H is the average call length, $\mu = \frac{1}{H}$, and $A = \lambda H$ is total traffic intensity measured in Erlangs, which are a dimensionless quantity.

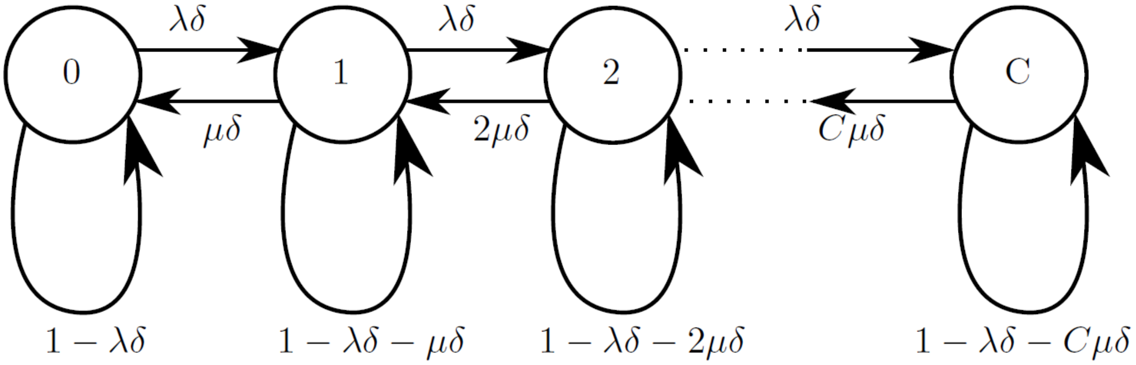

Figure 2. State diagram for the Erlang B of telephone line systems.

To understand the diagram in Figure 2, first assume there is no channel in use (state 0). After a while the probability of continuing to be in state 0 is $1-\lambda\delta$. Starting from state 1, the probability of returning to state 0 is $\mu\delta$ and the probability of continuing to be in state 1 is $1-\lambda\delta-\mu\delta$. Of course, the sum of all probabilities must equal one. When the system reaches a state k the probability that k channels are in use equals $k-1$ channels times $\lambda\delta$. Then

$$\lambda\delta P_{k-1} = k\mu\delta P_k, \quad k \le C. \qquad (3)$$

Equation (3) is known as the global balance equation because

$$\sum_{k=0}^{C} P_k = 1. \qquad (4)$$

For $k=1$, equation (3) becomes

$$P_1 = \frac{\lambda P_0}{\mu}. \tag{5}$$

For many values of k one has

$$P_k = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \tag{6}$$

or

$$P_0 = \left(\frac{\mu}{\lambda}\right)^k P_k k! = 1 - \sum_{i=1}^{C} P_i. \tag{7}$$

Inserting equation (6) into (7) yields

$$P_0 = \frac{1}{\sum_{k=0}^{C} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}. \tag{8}$$

According to equation (6), the probability of blocking for C channels is given by

$$P_C = P_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}. \tag{9}$$

Plugging equation (8) in (9) gives

$$P_C = \frac{\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}{\sum_{k=0}^{C} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}. \tag{10}$$

Because total traffic is given by $A = \lambda H = \frac{\lambda}{\mu}$, equation (10) can be rewritten as

$$P_C^{\text{Erlang B}} = \frac{A^C \frac{1}{C!}}{\sum_{k=0}^{C} A^k \frac{1}{k!}}, \tag{11}$$

which is the Erlang B formula for C channels.

The derivation of the Erlang C formula is similar to that of the Erlang B apart from the fact that there is no user limit after the system has overblown its capacity.

Figure 3 shows the state diagram for the corresponding Erlang C of telephone line systems.

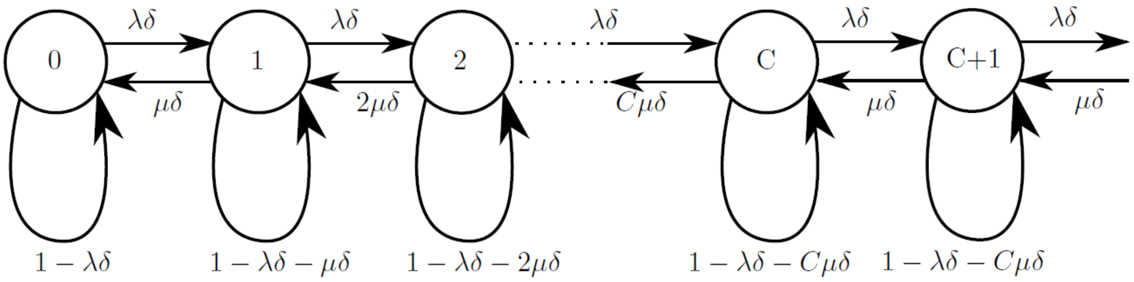

Figure 3. State diagram for the Erlang C of telephone line systems.

For $k \le C$, equation (3) becomes

$$P_k = \frac{\lambda}{\mu}\frac{1}{k}P_{k-1} \tag{12}$$

and, for $k > C$,

$$P_k = \frac{\lambda}{\mu}\frac{1}{C}P_{k-1}\,. \tag{13}$$

Thus,

$$P_k = \begin{cases} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}P_0, & k \le C \\ \left(\frac{\lambda}{\mu}\right)^k \frac{1}{C!}\frac{1}{C^{k-C}}P_0, & k > C \end{cases} \tag{14}$$

The global balance equation now becomes

$$\sum_{k=0}^{\infty} P_k = 1\,. \tag{15}$$

Thus,

$$P_0\left(1+\frac{\lambda}{\mu}+\ldots+\frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^{C+1}\frac{1}{C^{(C+1)-C}}+\ldots\right)=1$$

$$P_0\left(1+\sum_{k=1}^{C-1}\left(\frac{\lambda}{\mu}\right)^k\frac{1}{k!}+\sum_{k=C}^{\infty}\frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^k\frac{1}{C^{k-C}}\right)=1$$

$$P_0 = \frac{1}{\sum_{k=1}^{C-1}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^C \frac{1}{1-\frac{\lambda}{\mu C}}}\,. \tag{16}$$

Using equation (14), the probability that all the channels are in use when a new call occurs is given by

$$\begin{aligned} P(\text{all channels are in use}) &= \sum_{k=C}^{\infty} P_k = \sum_{k=C}^{\infty} \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{C^{k-C}} P_0 \\ &= P_0 \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C \sum_{k=C}^{\infty}\left(\frac{\lambda}{\mu}\right)^{k-C} \frac{1}{C^{k-C}} \\ &= P_0 \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C \frac{1}{1-\frac{\lambda}{\mu C}}\,. \end{aligned} \tag{17}$$

Equation (17) is valid for $\frac{\lambda}{\mu C} < 1$ only, that is, $C > A$. This is expected since $C < A$ would mean total traffic greater than the quantity of channels available, this meaning that the customers will wait for service indefinitely.

Considering equation (16) in (17) yields

$$P(\text{all channels are in use}) = \frac{\frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C}{\left(1-\frac{\lambda}{\mu C}\right)\left(\sum_{k=0}^{C-1}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^C \frac{1}{1-\frac{\lambda}{\mu C}}\right)}$$

$$P(\text{all channels are in use}) = \frac{\left(\frac{\lambda}{\mu}\right)^C}{\left(\frac{\lambda}{\mu}\right)^C + C!\left(1-\frac{\lambda}{\mu C}\right)\sum_{k=0}^{C-1}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}}\,. \tag{18}$$

Using the fact $A = \lambda H = \frac{\lambda}{\mu}$ in equation (18) yields the Erlang C formula:

$$P_C^{\text{Erlang C}} = \frac{A^C}{A^C + C!\left(1-\frac{A}{C}\right)\sum_{k=0}^{C-1}\frac{A^k}{k!}}\,. \tag{19}$$

Accordingly, the probability that a delay will exceed a given time t is given by

$$P(\text{delay} > t) = P(\text{delay} > 0)P(\text{delay} > t \mid \text{delay} > 0) \tag{20}$$

or

$$P(\text{delay} > t) = P(\text{delay} > 0)\; e^{-\frac{(C-A)t}{H}} . \tag{21}$$

3. Using the Erlang B and C formulas to manage excess reserves

Now we can adapt the variables in Section 2 used for telephone line systems to consider the problem faced by a commercial bank to manage its excess reserves. The main difference between the telephone system and the bank transactions is that in the telephone system each user uses only one channel per time while in the bank transaction system a user can occupy many channels (represented by cash) at the same time. To address this issue, we have two alternatives: (1) to multiply the number of users by the bank average withdrawals, or (2) to divide the total amount of cash by the bank average withdrawals, and then treat each lump of cash as a single channel. Here, we decide to consider the second alternative.

As a result, the new meanings of the variables are shown in Table 2.

Table 2. Definition of the variables.

λ_U	Total withdrawal requests per time unit over all users and channels.
H	Average time between withdrawals and deposits, or average time to return the cash to the bank
μ	$1/H$
w	The bank average withdrawals
C	Fraction of the bank liabilities over w
U	Users

A more proper notation is $C = \gamma L / w$, $0 \le \gamma \le 1$, where L stands for the bank liabilities. As long as H and λ_U are stable in time, one has $A_U = \lambda_U H$ and $A = A_U U$, where A and A_U are, respectively, the amount of cash flow for all the users and for a single user. As a result, equations (11) and (19) can be rewritten as

$$P_{\gamma L/w}^{\text{Erlang B}} = \frac{\dfrac{A^{\gamma L/w}}{(\frac{\gamma L}{w})!}}{\sum_{k=0}^{\gamma L/w} \dfrac{A^k}{k!}} \tag{22}$$

and

$$P_{\gamma L/w}^{\text{Erlang C}} = \frac{A^{\gamma L/w}}{A^{\gamma L/w} + (\frac{\gamma L}{w})!\left(1 - \dfrac{wA}{\gamma L}\right)\sum_{k=0}^{\gamma L/w-1} \dfrac{A^k}{k!}} . \tag{23}$$

Customers' demands for withdrawals of cash are memoryless, that is, demands are random. Also, the probability of a customer to keep cash on hand is exponentially distributed, that is, it is less likely the customer will keep cash for a long time vis-a-vis for a short time.

Thus, given U and A_U, the bank can set the proper amount of cash on hand to meet their customers' demands for withdrawals of cash. As a result, the proper amount destined to loans is also set.

How the Erlang B and Erlang C formulas can be used here is now illustrated with the help of two examples.

Example 1. The T-account of the bank is as shown in Table 3, and the customers' characteristics are depicted in Table 4. For that amount of money on deposit with the bank, how much money is allowed to be lent such that the probability that a customer is caught blocked from any demands for withdrawal of cash is less than one per cent?

Table 3. T-account of the bank prior to the optimal management of excess reserves, millions of dollars.

Assets		*Liabilities*	
Required reserves	\$10	Checkable deposits	\$100
Excess reserves	\$90		

Table 4. Customers' characteristics

λ_U	0.025 requests per day, per customer, and per channel
H	3 days
U	1,000,000 customers
w	\$ 100

What we are looking for here is the value of parameter C (or $\gamma L / w$) which satisfies equation (22) at the one percent probability of blocking. Thus, using the customers' characteristics in Table 4, C was found using a Matlab recurrence routine (available on http://www.mathworks.com/matlabcentral/fileexchange/824). We found $C = 74,340$ and $\gamma L = Cw = 7,434,000$. This means the bank should use approximately \$82.5 million for loans. The new T-account for the bank after the optimal management of excess reserves is shown in Table 5.

Table 5. T-account of the bank after the optimal management of excess reserves using the Erlang B formula, millions of dollars.

Assets		*Liabilities*	
Required reserves	\$10	Checkable deposits	\$100
Loans	\$82.5		
Excess reserves	\$7.5		

Example 2. In Example 1 above, if it happens for a customer to be blocked, the best that a banker can do is to ask him to come back later. However, the service could be improved if the banker decided that no more than 10 percent of the customers blocked will have to wait for more than 15 minutes to make their withdrawals.

Using equation (21) for the situation of blocking, one has

$$P(\text{delay} > t \mid \text{delay}) = e^{-\frac{(C-A)t}{H}}$$

$$P(\text{delay} > 15 \text{ min} \mid \text{delay}) = e^{-\frac{(C-A)t}{H}}. \tag{24}$$

Isolating C in equation (24) yields

$$C = A - \frac{H \ln P(\text{delay} > 15 \text{min} \mid \text{delay})}{t}. \tag{25}$$

Solving equation (25) by considering the values in Table 4 produces $C = 75,663$ and $\gamma L = 7,566,300$. Inserting this result into equation (23) yields $P(\text{delay} > 0) = 0.0089$. Finally, the probability that a customer is made to wait for more than 15 minutes is given by the joint probability

$$P(\text{delay} > 15 \text{ min}) = P(\text{delay})P(\text{delay} > 15 \text{ min} \mid \text{delay}) \tag{26}$$

$$= 0.0089 \times 0.1 = 8.9 \times 10^{-4}$$

or 0.089 percent.

In the examples above the bank balance sheets considered checkable deposits only. Equity has been omitted from the liabilities because we are focused on liquidity, that is, excess reserve management. Thus, we do not consider the possibility of insolvency, which is incidentally what Figure 1 illustrates. Bank runs are not considered either. A bank facing a run may collapse even if still solvent. An example is that of the Icelandic banks in October 2008. Considering bank runs compromises our model assumption that withdrawals are random. This is because in a liquidity crisis withdrawals become highly correlated. Future work should then try to overcome this limitation by considering arrival times which are not exponentially distributed.

4. Conclusion

We put forward a new, more practical approach for the optimal management of excess reserves of commercial banks. Although the economic literature on the issue is age-old and large, there is no attempt so far to tackle the problem using queuing theory which is commonly employed in other more established areas such as telecommunications and traffic engineering. We show how the Erlang B and Erlang C formulas of queuing theory applied to the telephone line system can be properly adapted to the problem faced by a bank in its choice of the optimal amount of excess reserves. Two examples were shown for the T-account of a bank.

References

[1] F.S. Mishkin, The Economics of Money, Banking & Financial Markets, 9th Edition, Upper Saddle River, Prentice Hall, 2009.

[2] T. Keister, J. McAndrews, Why are banks holding so many excess reserves?, Federal Reserve Bank of New York staff reports, 380, 2009.

[3] Greenspan, A., Banks need more capital, The Economist, December, 18th 2008.

[4] P.A. Frost, Banks' demand for excess reserves, Journal of Political Economy 79 (1971) 805.

[5] G. Selvaretnam, Regulation of reserves and interest rates in a model of bank runs, University of St Andrews working paper, CDMA07/14, 2007.

[6] E. Jallath-Coria, T. Mukhopadhyay, A. Yaron, How well do banks manage their reserves? NBER working paper, 9388, 2002.

[7] D.R. Skeie, Banking with nominal deposits and inside money, Journal of Financial Intermediation 17 (2008) 562.

[8] A.K. Erlang, Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges, Electroteknikeren 13 (1917) 5.

[9] H.C. Tijms, A First Course in Stochastic Models, New York, John Wiley & Sons, Ltd, 2003.

[10] T.S. Rappaport, Wireless Communications: Principles and Practice, 2nd edition, Upper Saddle River, Prentice Hall, 2002.