MODELS FOR ASSIGNMENT OF 911 EMERGENCY TELEPHONE OPERATORS

Peter Kolesar[1], with Albert Pedrinan[2]
and Peter Stein[3]

[1]Columbia University

[2]The City College (CUNY)

ABSTRACT

We discuss the design and implementation of a SIMSCRIPT simulation of the 911 emergency telephone system in New York City. The simulation was created to aid the New York City Police Department in scheduling operators and controlling the flow of calls through the system so that delays could be kept to tolerable levels while the system was run economically. A second goal was to use the simulation to validate an approximate analytic queueing model of the complex system. The queueing model based on the M/M/C queue can be used to provide quick responce to management questions.

I. INTRODUCTION

The 911 emergency telephone system in New York City handles about three million calls for police, fire, and ambulance service per year. At peak hours as many as 50 operators will be receiving these calls. It is important to the citizens of the City to economically maintain a rapid response to emergency calls. The study described here was part of an effort by the New York City Police Department (NYPD) which operates the 911 system to improve management of its operation. Several questions were concerning the Department the most pressing of which was scheduling operators. The main goal of our study was to model the call receipt and handling process so that the interrelations between call rates, the number of operators assigned to different borough stations, and the delay probabilities could be better understood. The complexities of the system which consists of separate but interrelated borough stations made a simulation study appear desireable. Our goals in conducting the simulation study were to

-create an accurate model of the system that could imitate the complex message routing possibilities of the system so that alternative operating protocols and schedules could be evaluated.

-use the model to determine whether --and under what conditions-- simpler analytic queueing models could be used as approximations to the actual system.

-use the results of the study to develop planning tools that were easy to use by NYPD planners and managers.

AN OVERVIEW OF THE 911 TELEPHONE SYSTEM

We present here an overview of the 911 telephone system by which emergency calls from the public reach the Police Department. We will ignore details and fine points which are not relevant to the purpose of this study.

When a citizen dials 911, he is connected to the Police Department via one of a fixed number of trunk lines connecting the borough from which he is calling to the Communications Center at Police Headquarters in lower Manhattan. (At current call rates, it is rarely the case that all trunk lines will be busy, but should that happen, telephone company operators may route the call to the Communications Center via special lines.) Each borough has a fixed number of trunk lines and a fixed number of Automatic Call Director stations (ACD stations) which are facilities at which telephone calls can be answered and processed by Police Department personnel called turret operators. Not all of the ACD stations are manned at all times.

If a turret operator is free when a call is received, it is routed to a free operator with essentially no delay. This routing is done by switching devices which select one of the free operators at random thus avoiding an overload to any one operator.

The function of the turret operator is to obtain from the caller information about the place and nature of the emergency, the name or phone number of the caller, etc. While obtaining this information, the turret operator transmits it to the SPRINT computer via a terminal at which he sits. The information is processed by the computer and then transmitted to the appropriate dispatcher who will contact and assign patrol cars to the call. The service or processing time of a call by the turret operator is the interval between receipt of the call and the moment when the operator is ready to handle another call. This is usually somewhat longer than the duration of the telephone call as the operator may wish or need to take a brief pause between calls.

If the call arrives when all operators are busy, it must wait to be serviced. While waiting it occupies its trunk line. If, when the call arrived all the operators were busy but there were no other calls waiting, it is passed through a gating device and admitted to a set we call the corral. (The term corral is our own usage and describes the set of calls which have passed the gate but have not yet been served.) The gate now closes and stays closed until all the calls in the corral have been assigned to operators. In the case in question, there is only one call in the corral and the gate will stay closed until the call assigned to the first operator who becomes free. The gate now opens and admits to the corral all calls that are waiting to be served. That is, all calls that arrived since the gate last closed. The gate stays closed until all these calls are assigned to operators. It then opens again, and this process of gating continues.

If several calls are admitted to the corral at once, they are selected in random order and connected to the turret operators as the operators become free. Priority is not given to the calls that arrived first. Each and every borough operates in this fashion. In the simplest mode of operation calls wait until they can be served by an operator in their home borough. This simple mode of operation can be modified in two ways: by using the "interflow" option and/or by using the "overflow" option.

We describe the interflow option next. There are switches which we call interflow switches whereby it is possible to "share" operators and and/or calls between the paired boroughs of Manhattan and the Bronx, and between Brooklyn and Queens. For each borough pair there are two switches. So, for example, we can route Manhattan calls to Bronx operators by turning that switch on, and Queens calls to Brooklyn operators and also Brooklyn calls to Queens operators by turning both of those switches on. When an interflow switch is on, say from Manhattan to the Bronx, Manhattan operators are available to serve both Manhattan and Bronx calls, but the switching mechanisms are such that Manhattan operators give priority to Manhattan calls. At any rate, selection from the corral of the next call to be served is always in random order.

The interflow switches are turned on or off at the discretion of the NYCPD supervisor of the ACD operations. As far as we can determine there are no hard and fast rules detailing when this is to be done, but switching is used to avoid long delays. Interflow is not possible between Richmond and the other boroughs.

We now describe the overflow option. There is a group of ACD stations, called the overflow or backup positions, which are not assigned to any particular borough but instead are reserved for special uses. Some of these positions are usually manned by Spanish-speaking turret operators, to which calls received by the regular borough operators can be routed via special switches. In addition, emergency calls received directly by the telephone company can be routed to the overflow operators via one of several "direct" lines. This might happen if a citizen dials "operator" instead of 911 and has a very serious emergency to report, or if all the trunk lines for a borough are occupied. These two classes of calls --Spanish language and direct calls-- get priority in assignment to free overflow operators.

In addition, there is an "overflow" switch for each borough. If this switch is on for a given borough, calls being held in that borough's corral may be routed to an overflow operator when one becomes free. If the overflow switch is on for several boroughs simultaneously, the selection of the borough that gets served next when an overflow operator becomes free is made at random. The call to be served is also selected at random from among those in that corral. A delay of 600 microseconds is built into the switching operation for all boroughs except Richmond. This is so that Richmond -- which does not have the interflow option -- gets priority in usage of the oveflow operators. It thus has a kind of interflow with the overflow group. The effect is that Richmond calls will be selected before any other borough call. Recall, however, that direct and Spanish language calls still get priority over Richmond calls. Calls received over the direct lines to the overflow operators are "gated" and "corraled" in the same way as borough calls.

A SIMSCRIPT SIMULATION MODEL OF THE 911 SYSTEM

The simulation models the 911 police emergency phone system and produces, for each group of borough operators and for the overflow operators, statistics concerning the queue lengths, number of busy operators, the number of trunk lines occupied, the number of calls in the corral, the amount of time calls must wait before being answered (served), and the number of calls received and served. The major purpose of the simulation was to test rules for allocating operators among the boroughs, focusing on the utilization of a minimum number of operators while keeping less than 2 percent of the calls delayed more than 30 seconds. Since the ACD Simulation is a very general and flexible model it may well serve other uses of the Police Department in analyszing communications problems. The simulation was written in SIMSCRIPT II.5 and the following description assumes that the reader has a basic familiarity with that language.

Each group of operators (one for each borough plus the overflow group) is represented in the simulation by a permanent entity called

BOROUGH. Each call coming into the system is represented by a temporary entity called a JOB. Each BOROUGH has three sets associated with it: a set called LINES--all calls occupying trunk lines (being served or waiting for service), a set called QUEUE--all calls waiting for service, and a set called CORRAL--calls waiting for service which have gone through the gate. Any JOB for a particular BORO may belong to one or all of these SETS.

EVENT FOR ARRIV. CALL

This event is externally triggered by the reading of a tape or disk, and represents the arrival of a call into the system. The DURATION and BORO of the call are read and its ENTRY. TIME is set to the current "clock time". The program updates a counter for the number of jobs received by that BORO. If so, the call is placed in the set LINES of its BORO, if no, the call is considered lost and is no longer processed-- but it is counted. If a line was available, the program checks in the following order to see if an operator is available: in its own BORO, or , if INTERFLOW.SWITCH is on, in its PARTNER.BORO, or, if OVERFLOW.SWITCH is on, in the overflow (BOROUGH 6). If a free operator is found, a routine ASSIGN.OPER is called. Its arguments are the identity of call being processed and the BOROUGH with which the free operator is associated. If no free operator is found, the call is placed in the QUEUE for its BORO, and in the CORRAL for its BORO if the corral is empty. Counters are updated for number of calls lost, the number of calls delayed and which borough a call is received from and served by.

ROUTINE TO ASSIGN. OPER GIVEN JJOB AND BOR

This routine models what occurs when an operator answers a call --either when the call has just arrived, or after it has been waiting for an operator to become available. Counters for the number of operators busy and the number of jobs served for that BOROUGH are updated, and if the call has been waiting more than 30 seconds, that counter is also updated. An EVENT FOR CALL.END is scheduled with the call being processed (JJOB) as an argument at the current "clock time" plus the DURATION of JJOB (the time it takes the call to be served).

This routine is called from the event for ARRIV.CALL and from the routine to START.JOB.

EVENT FOR CALL.END GIVEN JJOB

This event represents what occurs when a call is finished being served and an operator becomes free. The counter for the number of operators busy for the serving borough of the call (SBORO of JJOB) is updated, and the call is removed from the LINES of its own BORO. The operator then searches for a call to serve. It checks the CORRAL of its own BOROUGH and if this corral is empty, the CORRAL of its PARTNER.BORO (if its INTERFLOW.SWITCH is on). If both these CORRALS are empty and the operator is an overflow operator, it then checks the CORRALS of any BOROUGH whose OVERFLOW.SWITCH is on. If more than one CORRAL is not empty, one is chosen randomly. The selection of the CORRAL is done using internally generated random numbers. When a nonempty CORRAL is found, a ROUTINE TO START.JOB is called with arguments identifying the BOROUGH the CORRAL is associated with and the BOROUGH the operator is associated with. If no nonempty CORRAL is found, the operator waits for more calls to enter the system.

Event notices for the event CALL.END are created in the routine ASSIGN.OPER.

ROUTINE TO START.JOB GIVEN B AND S

This routine represents what occurs when an operator finds a call that has been waiting for service. The first call is removed from the CORRAL associated with BOROUGH B (calls were placed in the CORRAL randomly) and the same call is removed from the QUEUE associated with BOROUGH B. The ROUTINE TO ASSIGN.OPER is called with arguments identifying the call just removed and the BOROUGH (s) associated with the free operator. If this CORRAL is now empty, ROUTINE TO OPEN.GATE is called with BOROUGH B as an argument. A counter is updated for the number of calls from BOROUGH B as an argument. A counter is updated for the number of calls from BOROUGH B served by BOROUGH S.

ROUTINE TO OPEN.GATE GIVEN BOR

This event represents what occurs when calls that are queued up but not yet admitted to a CORRAL are allowed into the CORRAL in order depending on random numbers that are generated for each call in the QUEUE.

The filing of the jobs in the CORRAL takes advantage of the fact the CORRAL is defined as a set ranked by the attribute ORDER. This automatically places jobs in the set according to the aforementioned random numbers and not in the order of their arrival in the system. In this way the simulation imitates the random order of call selection that occurs in the real ACD system.

EVENT FOR END.OF.SIMULATION GIVEN SW

This event represents either the end of the initialization period or the end of the simulation period depending on whether SW=0 or SW=1. If SW=0, various counters and statistics are reset to 0 and the simulation will be resumed. Regardless, of whether SW=0 or 1 the proportions of jobs lost, jobs delayed and jobs delayed over 30 seconds are calculated and a report containing the results of the simulation for the appropriate period is printed. If SW=0, the simulation is continued, but if SW=1, processing is terminated.

CALL GENERATION PROGRAM

An External Event File, consisting of the stream of calls that is input to the simulation, must be created prior to the running of the simulation. This is done by a Fortran program that generates a record for each call. The call arrival times are generated according to a Poisson process with a mean equal to the sum of the borough arrical rates. Service times are exponentially distributed. Calls are distributed randomly among the boroughs according to individual borough arrival rates.

AN APPROPRIATE QUEUEING MODEL

Because of its simplicity and the ease with which estimates of queue lengths and delays can be computed with it, we attempted to build an approximate queueing model of the 911 system based on the M/M/C queue. Indeed if one ignores the overflow capability and is content with delay estimates for the pairs of boroughs that are linked by overflow the M/M/C model is appropriate (for average delays).

The use of the overflow option creates analytic problems. To model the system directly is difficult since a multidimensional state space would be necessary which we believed would make computations very difficult, if possible at all. An approximation could be built if one could estimate how the overflow operators are shared by the boroughs or borough pairs. For example, suppose that there are ten overflow operators, and that Manhattan and the Bronx have their interflow switches on and are therefore paired, and that Brooklyn and Queens have their interflow switches on and are therefore paired, and that Richmond stands alone. Depending upon the call rates for each of the boroughs and the number of operators assigned to them, the overflow operators will handle a certain number of calls from each borough. Conversely, one might imagine that a certain number of overflow operators are "effectively" assigned to the boroughs or borough pairs. We built an approximation model of this as follows. This work is due to Richard Urbach and is reported in (2).

Assuming that we have partner boroughs i and j with K_i and K_j operators on duty respectively, and in addition boroughs i and j have interflow on and overflow off, we thus have virtually $M/M/(K_i + K_j)$ queueing model for that borough pair. The major differences between the 911 system and $M/M/(K_i + K_j)$ is that calls are picked from the queue (corral) at random while in the $M/M/(K_i + K_j)$ model calls are picked on a first in first out basis. In test results comparing the standard $M/M/(K_i + K_j)$ with FIFO and a $M/M/(K_i + K_j)$ queueing model having calls picked from the queue completely at random, it was found that there was virtually no difference in the models at the utilization rates likely to be encountered in the actual system.

This analysis falls short when overflow is operating. Now we estimate the share of overflow operators used by a borough or a borough pair by the equation

$$E_i = K_i + \alpha_i C$$

where E_i is the effective number of operators on duty in borough i, K_i the actual number of operators on duty in borough i, C the number of overflow operators, and α_i a multiplier that assigns borough i the effective proportion of overflow operators it utilizes during that time period. In a sense, we are attempting to force 911 to look like an $M/M/E_i$ model. Now to estimate α_i for each borough or partner boroughs.

ρ_i = the utilization rate of borough i's operators

λ_i = the mean arrival rate of borough i

μ_i = the mean service rate

Using our expression of E_i we get that:

$$\rho_i = \lambda_i / (K_i + \alpha_i C) \mu$$

We estimate α_i by

$$\hat{\alpha}_i = \frac{\lambda_i / K_i}{\sum_j \lambda_i / K_j}$$

We can then get an estimator for ρ_i as follows. Let

$$\gamma_i = \lambda_i / (K_i + \hat{\alpha}_i C) \mu$$

Then γ_i can be used to estimate ρ_i by

$$\rho_i = a_i \gamma_i + b_i$$

To obtain sample values for the utilization, the SIMSCRIPT simulation was used for different call rates and operator assignments. Since the values of λ_i, K_i, α_i, μ, and C are all known, a_i and b_i were found by linear regressions.

Given the estimated values for a_i and b_i we have:

$$E_i = \lambda_i / (a_i \gamma_i + b_i) \mu$$

A FORTRAN program has been written using this approximation which:

1) Calculates the number of operators that should be on duty to meet the requirement that a not more than given fraction of the calls wait t seconds or more.
2) Given the number of borough operators and overflow operators calculates the probability of delay of t seconds or greater.
3) Given a utilization rate for the bo-

rough pairs and Richmond and the number of overflow operators, the distribution of operators is calculated along with the probability of waiting t seconds.

In actually working with the simulation, we found that a good heuristic for allocating R operators between paired boroughs i and j is to assign $(\lambda_i / \lambda_i + \lambda_j)R$ operators for borough i and $(\lambda_j / \lambda_i + \lambda_j)R$ operators for borough j. If there were any operators left over due to rounding we assign them to the borough with the highest call rate.

This program has been used by the NYPD in analyzing operator allocations.

STATISTICAL ANALYSIS

An extensive series of simulations were run under widely varying conditions in order to obtain data for our statistical analysis. The goal of this analysis was to run regressions of of the model

$$\rho_i = a_i + b_i \gamma_i$$

Such analyses were carried out for individual boroughs or boroughs pairs using the subprograms SCATTERGRAM and REGRESSION of the Statistical Package for the Social Sciences (SPSS). The results for each configuration confirmed that there was indeed a significant and strong linear relationship between ρ and γ with values of r^2 generally over 0.9.

In addition a_i was generally nearly zero and b_i was generally nearly one, suggesting that to a good approximation $\rho = \gamma$. In support of this we offer results of a regression of pooled data from all borough pairs:

ANOVA Table
(pooled data)

Source	D.F.	S.S.	M.S.	Overall F
Regression	1	10.7004	10.7004	2537.046
Residual	152	0.64109	0.0042	
Total	153	11.34152		

	(S.E.)
Estimate of Slope (b) = 0.9605	(0.014)
Estimate of Intercept (a) = 0.0453	(0.010)

ACKNOWLEDGEMENTS

This research was partially supported by the NYPD through a contract with the New York City-Rand Institute, by research funds of the Rand Corporation, and by the Sloan Foundation through a grant to the City College of New York (CUNY). Dr. Kenneth Rider, then of the New York City- Rand Institute provided valuable guidance throughout the project. The encouragement of Dean Paul Karmel of City College is gratefully recorded.

References

1. E. Ignall, P.Kolesar, and W. Walker, "The Use of Simulation Models for Emergency Services", Proceedings of the 1974 Winter Simulation Conference, pp. 528-536, January 1974.

2. R. Urbach, "Estimating In-Queue Waiting Times of Emergency 911 Telephone Calls: A System of Parallel Queens" Interval Note IN-22757-NYC, The New York City-Rand Institute, March 1975.