*Proceedings of the 2002 Winter Simulation Conference*
*E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.*

# APPLICATION OF SIMULATION AND MEAN VALUE ANALYSIS TO A REPAIR FACILITY MODEL FOR FINDING OPTIMAL STAFFING LEVELS

G. Boyer
A. N. Arnason

Department of Computer Science
University of Manitoba
Winnipeg, MB R3T 2N2, CANADA

## ABSTRACT

Staffing problems arise in a wide range of applications including job shops, call centres, and hospital emergency departments. They are characterised by the need to allocate shift workers with varying skills to handle an arrival stream of tasks having different sub-task routings and (sub-task) skill requirements. The Manitoba Telecom Service Trouble Diagnosis and Repair System (TDRS) has 3 skill-levels of staff handling multiple types of faults occurring in telephone switching equipment. TDRS is a pure staffing problem having no equipment constraints: the only resource constraint is staff itself. The object of this study is to show how this can be modelled as an open network of queues with feedback and allowing for temporal and fault-class heterogeneity. Analytic mean value analysis then facilitates validation and selecting feasible staffing strategies for closer examination by simulation. The purpose of experiments using simulation is to find effective performance visualisations and "optimal" staffing allocations.

## 1 INTRODUCTION

The deregulation of the telecommunications industry has created increased competition among the providers of these services, which in turn has created intense pressure for them to find ways to cut costs to maintain profitability, as stated by Regnier and Cameron (1990). Increased competition also means the customer expects a highly reliable service. The objectives of high reliability and low cost conflict; in order to increase the reliability of the network, more money must be spent on such things as better equipment and increased maintenance. Thus the challenge for managers of a telecommunications network is to try to maintain high reliability while at the same time lowering cost.

Manitoba Telecom Services (MTS) was, until 1997, a Crown (i.e., government-owned) Corporation. It is the sole provider of local telephone service to the million or so residents of Manitoba. The lucrative long-distance market has has been open to competition for several years, though, so MTS has found itself exposed to the same sort of competitive pressure that has motivated other telecommunications companies to cut costs. One of the areas they examined in the hope of finding some costs savings is the network fault diagnosis and repair process.

During the normal course of operations some network components will fail or otherwise function unsatisfactorily. These failures (called troubles) must be detected, diagnosed and repaired. Of course these troubles have an impact on network reliability, so although an equilibrium of unrepaired troubles exists at most times, it is desirable to have low numbers of these unrepaired troubles. Having a large number of repairpersons available at all times will certainly help to keep the number of unrepaired troubles low, but it is an unwarranted expense. Providing more training for the personnel who diagnose the troubles should also help to decrease the number of unrepaired troubles by enabling them to repair troubles more quickly, but is itself another expense. Staffing can also be reduced during non-business hours to reduce overtime costs provided unacceptable backlogs don't accumulate. So the question the MTS managers want answered is: What is the best mix of personnel staffing levels and its allocation to shifts that will result in minimum cost and still maintain the number of unrepaired troubles in the network at an acceptable level?

MTS has 3 levels of support staff who are responsible for analyzing and resolving troubles.

Level 1: Network Operations Centre (NOC) staff handle the bulk of the trouble resolution process for the Winnipeg area and Provincial Network Operations Centre (PNOC) are responsible for the province-wide network as well as handling Winnipeg when NOC staff are not available. Both NOC and PNOC determine the severity of incoming troubles and dispatch Craft personnel or call upon DSG staff when needed.

Level 2: Digital Support Group (DSG) staff have the highest level of training and experience. They are normally only available in business hours, but may be called in during off-hours if a service-threatening trouble arises.

Craft: Craftspersons perform testing and repair under the direction of NOC/PNOC and DSG staff. The craft pool is staffed around the clock.

In addition to the MTS staff, equipment manufacturers also maintain their own troubleshooting teams who are called upon when critical network components fail (Level 3 support or OSO "Outside Support Organisation").

Trouble reports are received at NOC/PNOC. The staff analyze the trouble by performing automated diagnostics from the Centre. If the cause of the trouble can be determined, action can be taken to correct it ranging from rebooting switching software to dispatching a Craftsperson to a remote site to repair damaged equipment. If the NOC/PNOC staff cannot determine the cause of the trouble, it will be passed to the DSG staff, and if they are not available and/or the trouble is serious enough, the manufacturer's troubleshooting staff will also be called.

Troubles are prioritized into classes according to their potential for causing disrupted or degraded service and are designated by codes (e.g., NS=Non-service affecting). They can be classified in decreasing order of criticality as:

- (E1) critical troubles,
- (E2) major troubles,
- (S1/S2) customer-reported troubles,
- (NS) minor troubles.

Critical troubles are ones that result in immediate loss or degradation of service to customers, while major troubles have the potential to do so; they are given immediate attention. The customer-reported troubles affect only one or a few customers while minor troubles cause no service failure (e.g., diagnostic test warnings or failure of backup equipment). Both are given lower priority and must wait for service until higher-priority troubles are resolved. E1/E2 always receive immediate service if personnel are available but NS/S1/S2 may be "ticketed" for service on the next business day. Accumulations of the lower priority troubles are not desirable though, since they may be symptomatic of a fault that could disrupt service if left unresolved.

The normal flow of resolution of a trouble is shown in Figure 1. It is first analyzed by NOC/PNOC, where there are three possible outcomes: resolution of the trouble, determination of a hardware fault or failure to resolve the trouble. In the first case, the trouble is cleared and leaves the system. In the second case, the trouble is passed onto Craft for hardware repairs. In the third case the trouble is referred to DSG for analysis. Troubles referred to DSG follow the same flow, except unresolved troubles are referred to OSO. When a trouble is passed onto Craft, there are two possible outcomes: the hardware fault is repaired, in which case the trouble leaves the system, or the fault is not repaired, in which case the trouble is referred back to the previous support level for further analysis.

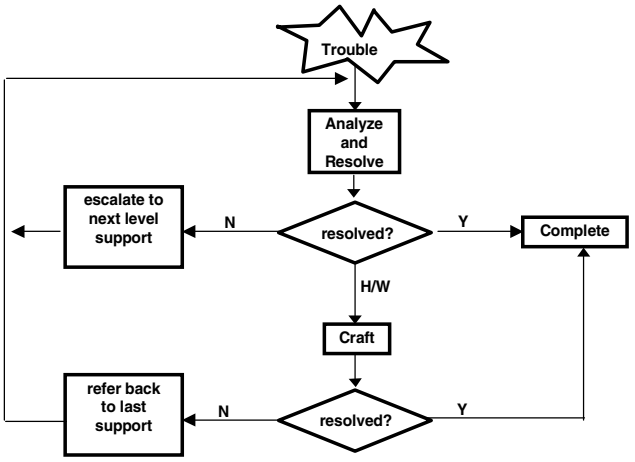


Figure 1: Normal Trouble Flow

This normal trouble flow can be modified by bypassing the Analyze and Resolve stage and directly escalating to the next level support. The trouble flow will be modified according to three factors:

1. support level involved,
2. criticality level of trouble,
3. time of occurrence (i.e., business or off-business hours).

By discussion with MTS, details of trouble flow for each support level were determined from the moment a trouble arrives. For example, the detailed flow for troubles arriving at Level 1 (NOC/PNOC) is shown in Figure 2.

It became apparent from study of these flows that the system is essentially like a job-shop, with troubles as the tasks to be routed among (machine) groups of level staff, but with the following significant differences:

### 1.1 Preemption

Level 3 and 4 troubles always receive immediate service, either when they arrive in the system or when they move to a new service level according to the normal trouble flow. If such a trouble requires service at a time when all the support staff are busy with troubles of lower criticality, work on one of these lower criticality troubles will have to be delayed in order to free up staff to work on the higher criticality trouble. The work done prior to preemption is not lost, so this is preemptive-resume.

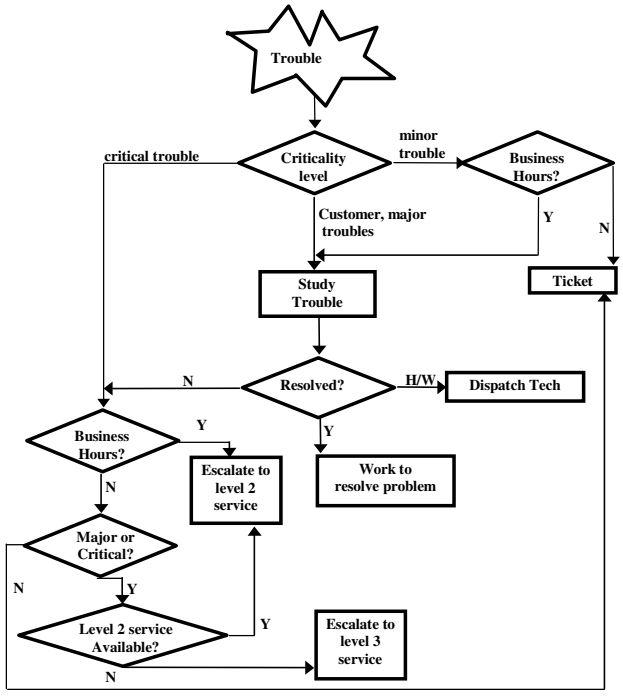


Figure 2: Level 1 (NOC/PNOC) Trouble Flow

### 1.2 Temporal Effects

There are both time-of-day effects and day-of-week effects. These effects are largely attributable to variations in staffing levels that occur on a daily and weekly basis. For example, both DSG and NOC are staffed during business hours (8:00 a.m. to 4:30 p.m. Monday to Friday) only. Consequently there is less staff available during non-business hours, so accumulations of troubles occur during the evenings and on weekends. These accumulations are then reduced when staff is available during business hours, although it may take several days for the weekend accumulation to be reduced.

If a Level 3 or 4 trouble occurs during off-business hours an attempt will be made to obtain DSG staff on a call-out basis; however, there is no guarantee this will be successful. The likelihood of obtaining on-call staff varies with time-of-day and day-of-week. This on-call staff will work on the Level 3 or 4 trouble until it is resolved and then return home, so their presence will not affect the accumulation of troubles during off-hours.

Staff who are still working on a trouble when the end of their shift comes will generally remain working on it until the service is complete. This results in more than the normal number of staff working for a brief period at the beginning of a new shift. Exceptions to this policy would be made if the trouble requires a long period of further service; in this case the staff working on the trouble on the old shift would familiarize the staff taking over the work before leaving.

The arrival rate and severity-class mix of the troubles themselves is typically free from temporal effects. The reliability of digital switching equipment is such that there is no corresponding increase in non-customer-reported trouble arrivals during times of increased network usage. Customer-reported troubles will increase during business hours since this is the time when most customers will notice them. This does not have a significant impact on the overall arrival rate since the major source for trouble arrivals is network monitoring activity (Chen *et al* 1988).

### 1.3 Task Sharing

Staff from 2 service levels may work together on a trouble. This type of interaction may be continuous or intermittent (e.g., with a DSG staff providing a consultative role). Such task sharing is hard to quantify and to incorporate into any tractable analytic model.

## 2 LOGICAL MODEL: COMPONENTS AND GOALS

The TDRS can be adequately represented as a network of queues with feedback, in which troubles are the (customer) entities and the staff pools by skill level are the (server) resources (Figure 3).

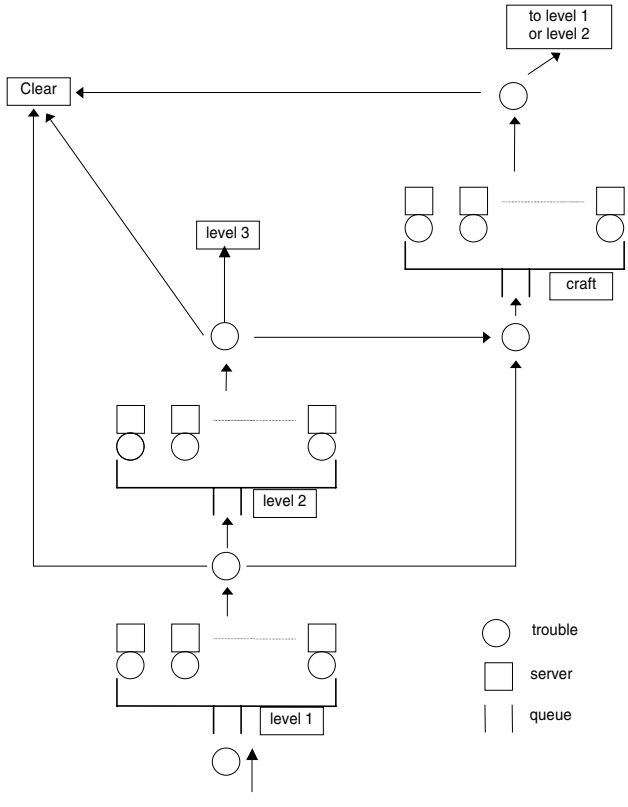


Figure 3: Logical Trouble Flow

Each criticality class has a different routing and has different service times. The input mechanisms and performance measures for this system are summarized in Table 1 and discussed further below.

Table 1: Input and Performance Variables

| Variable | Description |
|---|---|
| *(a) Input mechanisms* | |
| Trouble arrival rate | Poisson |
| Trouble criticality mix | multinomial |
| Trouble study time | beta |
| Trouble repair times | beta |
| Trouble service outcome | multinomial |
| Call-out success | Bernoulli |
| *(b) Performance measures* | |
| true queues (by skill group) | time integral |
| troubles-in-system (by criticality) | time integral |
| utilization (by skill per week) | time integral |
| receipt-to-close | observational |

### 2.1 Performance Measures

The network operator is faced with a tradeoff between network reliability and the cost of the repair process. Some measure of reliability is needed in order to answer the question "How reliable is reliable enough?" Typically reliability measures for telephone networks are expressed from the customer's point of view. Fagerstrom and Healy (1993) propose two measures of telephone network reliability, namely availability (the probability the network is available when a user at a random time attempts to make a call) and the probability a customer does not experience an outage longer than 5 minutes in a year.

This model requires some different measures of reliability. Troubles may or may not cause disruptions in service, which was what the two quantities above were concerned with. Since the repair process is concerned with clearing troubles as quickly as possible, measuring the average time to clear troubles from the network, broken down by criticality level, is a good measure of the efficiency of the repair process. Chen *et al* (1988) report this measure, which they call *receipt-to-close time*, in their simulation study of one aspect of the telephone network repair process. Although it is only an indirect measure of the reliability of the network, one would expect that as receipt-to-close time increases, network reliability decreases.

The second measure reported by Chen *et al* (1988) is queue lengths of troubles awaiting service. Long queues of troubles awaiting service mean that there are more network elements not functioning properly, increasing the probability that a customer will experience a disruption or degradation in service. Observing the behaviour of the queues over time can also reveal the sort of transient effects mentioned earlier, such as certain times of the day or week when queues are large. We measured true queues (by staff service level) and total troubles-in-system (by criticality).

The third important measure is staff utilizations. Utilization will differ among skill-level groups and shifts. However, because staffing patterns follow recurrent weekly patterns and arrivals are time-homogeneous, utilizations averaged over the week will stabilise if utilizations are under 100%. Calculating utilizations for given routings and staff allocations is an important means of determining if the system can cope with the total long-run trouble stream.

### 2.2 Service and Routing Mechanisms

Each staff group is modelled as a resource that provides service to trouble entities in 3 stages:

1. study the trouble;
2. work on the trouble;
3. route to another resource or clear the trouble.

The first stage, studying the trouble, models the initial phase of the repair process in which the staff attempts to determine the cause of the trouble and the necessary action to resolve it. This stage has three possible outcomes:

- the trouble can be resolved immediately;
- the trouble is caused by a hardware fault;
- the trouble is unresolved.

Trouble study may or may not result in a delay. If the trouble is to be studied, then a delay will occur. However, trouble study may also simply be a decision to escalate the trouble to a higher service level or to ticket it, which does not result in a delay.

The second stage, working on the trouble, is performed only if the outcome of studying the trouble is that the trouble can be resolved immediately. The other two outcomes of the study stage will result in service proceeding immediately to stage 3, routing the trouble to another resource for further analysis or hardware repair. If the result of studying the trouble is that it can be resolved immediately, then the service in stage two results in a delay while the trouble is resolved. After the delay is complete the trouble proceeds to stage three where it is cleared.

The third stage of service routes the trouble to the appropriate queue for the next service required, or marks it as cleared as appropriate. Once marked as cleared, the trouble leaves the system. Level 3 (OSO) support is not modelled so those troubles requiring Level 3 support also leave the system. This has the effect of biasing receipt-to-close times for those troubles requiring Level 3 service low, since in reality they do not leave the system. Utilizations are also biased low since staff are still involved in the service of troubles receiving Level 3 service.

The situation where staff from two service levels work together on a trouble is not explicitly modelled. Rather, a

trouble may return back to a previous service level, say from Level 2 to Level 1 or craftsperson to Level 2, in order to approximate the type of intermittent consultation that occurs when two service levels work on a trouble. Since the model never has two staff actually working on a trouble simultaneously, this approximation will bias low the utilizations of the staff involved.

The system is therefore a network of queues as shown in Figure 3, where the queues are priority queues and time dependent server numbers (skill level staffing) implemented as follows:

Temporal effects occur on both a daily and weekly basis caused by changing staff levels during the three weekday shifts and reduced staff on the weekend shifts. In the logical model shift changes are accomplished by updating the number of available (i.e., non-busy) servers and server capacity to equal the new shift strength at shift-change time. This is equivalent to all the non-busy staff leaving at quitting time and the new shift's staff arriving. Servers that are busy at shift-change time remain working on the trouble until its service is completed, which is exactly what happens in the MTS repair process.

If there are troubles waiting for service at shift-change time, they immediately begin service with the arriving servers. If the system is congested with troubles this can result in more than the shift strength of servers working (all the new servers and those continuing a service from the previous shift) for some time after the beginning of a shift.

Generally, Level 2 staff work during business hours only. However, critical and major troubles may be escalated from Level 1 to Level 2 service immediately. If such a trouble arrives on a weekend or during off-hours, it would have to wait some time before it could obtain service from a Level 2 server. When a critical or major trouble requires Level 2 service at a time when none are available, the model will attempt to provide the Level 2 service on a call-out basis. Obtaining Level 2 service during off-hours is probabilistic; if a Level 2 server is obtained, service begins immediately, and when it is finished, the Level 2 server leaves the model. If a Level 2 server is not obtained, the trouble must wait until the next shift change. If the shift is a business hours shift then Level 2 servers will be available, else the probabilistic process is repeated. The probability of obtaining off-hours Level 2 service depends on both the criticality of the trouble and the current shift.

If a Major or Critical trouble requires service at a time when all the servers are busy with troubles of lower criticality, one of these lower criticality troubles will have to be preempted in order to free up staff to work on the higher criticality trouble. A Major or Critical trouble requiring service and finding no available staff will cause the preemption of service for the minor or customer trouble whose service was started most recently. A Critical trouble may preempt the service of a Major trouble if there are no minor or customer troubles to preempt. Preempted troubles are returned to the queue in FIFO order, but ahead of non-preempted troubles with the same criticality.

### 2.3 Service Delay and Trouble Arrival Models

Service time models for study and repair times are modelled using transposed and scaled Beta distributions (Law and Kelton, 2000, section 6.11). The state space thus has a minimum ($a$), a maximum ($b$) and by choosing a mode ($m$) that exceeds the mean ($\mu$), the distribution has a positive skew over this range. Maguire (1994) justifies the use of this distribution for hospital emergency department patient treatment times in terms that are compelling for this application too: there is a minimum amount of time the service will take. This time is usually not far from the time the treatment will most likely take (mode). This is because most practitioners are competent at what they do, so unless unexpected delays occur, the elapsed time will probably be much closer to the minimum time than it will be to the maximum time. The average time will usually be larger than the most likely time because of occasional long delays that can occur. Moreover, the beta distribution is easily fit from summary data or expert guesses on the likely parameter ($a, b, m, \mu$) values for a given service.

Arrivals can be taken as random Poisson processes with time homogeneous arrival rates. This implies independence in trouble occurrence as well as time-constant arrivals. Most of the faults are software related and, in the experience of MTS, appear to follow a random process. Some faults, like hardware faults, can cause a cascade of trouble reports but these are recognised and "stapled together" at the study stage and treated as a single trouble. Some faults, like customer reported faults, are not time homogeneous, being more likely to be reported in business hours and early in the day, but these compose a small proportion of the troubles.

### 2.4 Study Goals

Part way through this study, a planned data mining and analysis exercise was cancelled due to Corporate reorganisation. The plan was to use the extensive trouble reports to fit the exogenous mechanisms (Table 1a) and summarise some performance measures (Table 1b) to use for model validation. Instead, we had to rely on less precise information from staff interviews that yielded a wealth of detail on flows but only rough approximations of the precise flow rates and service times. Thus the focus of the study changed from a classic Study-Model-Fit-Simulate-Validate-Experiment cycle (Law and Kelton, 2000, section 5.1) to a more methodological study with the following aims:

- develop a graphical user interface to permit specification of exogenous variable parameters and routings so that model is easily configured;

- develop dynamic graphics output so end-users can "see" if behaviours are "feasible" (i.e., all weekly staff utilisations are less than 1 and queue lengths settle down to a recurrent weekly pattern);
- develop deterministic analytic tools to show when a configuration is "feasible". The analytic tools should be driven by the same model configuration data structures that drive the simulation so they can also be used to validate correctness of model implementation;
- find a "base configuration" using current staffing levels that produces convincing output (i.e., that is feasible and that MTS staff deem to give plausible queue size dynamics and receipt-to-close averages);
- show that manipulation of staffing levels by re- allocation to shifts and skill levels can produce improved trouble resolution (reduced receipt-to close times) while retaining a feasible configuration.

## 3 IMPLEMENTATION

The TDRS model was implemented in SIMSCRIPT II.5 on a Sun/Solaris workstation using the X-Windows graphical system. Simscript's SimGraphics interface provides a graphical user interface that supports both the user input and the dynamic graphics for output display. Its process-resource language is well suited to implementing complex network-of-queues models; among other advantages, support for arrays of resources permits a more succinct network representation, and the ability to cancel processes on the (pending) event set is vital to implementing preemption. Simscript's language constructs for observational and time-integral statistics gathering (TALLY and ACCUMULATE) provide a succinct language for specifying arrays of statistics by service levels, trouble criticalities, and shift as well as overall. Simulations were replicated using the batch means method to obtain confidence intervals for the performance measures.

Users set the input parameters using text boxes arranged in arrays. They can modify the study outcomes as well through a 3 (service-levels) by 4 (trouble criticalities) by 2: (business/off-business shifts) array of outcomes: for each array element, the user can choose from a drop-down menu to study/ escalate/ ticket. The user then specifies run parameters (simulation duration, number of replications etc.; defaults from the base model are provided). Output from the runs was in the form of dynamic graphics for queue sizes and staff utilisations (illustrated below) and extensive table output of the statistical summaries.

Temporal effects are implemented by a shift-change event that causes the number of servers to change. If the shift increases staff, queues are checked for start of service. Services in progress are not terminated but are allowed to complete. The call-out success probability changes with shift and all enqueued critical and major troubles are (re)tried to find a service on call-out.

Validation of correct implementation was by careful examination of trace output and use of the MV analysis as described below.

## 4 MEAN VALUE ANALYSIS

A number of simplifying assumptions were made in deriving a mean value analysis (MVA). All of them tend to increase the staff utilisations relative to the more realistic simulation results, so the MVA acts as an upper bound on utilisation: if the MVA yields a feasible solution, the simulation will also, but the converse is not true. The simulation can be run with equivalent simplifying mechanisms so that the MVA and simulation can be checked against one another for validation.

- each shift is considered separately to remove time heterogeneities; results over shifts are obtained using summation or weighted averaging. Services in the simulation may carry over a shift change, providing "extra" servers temporarily and thus reducing utilisations;
- call-outs are ignored (since this creates a variable number of servers) in the MVA but will reduce utilisations and queue sizes of major and critical troubles in the simulation.

Under suitable assumptions on service time, the system is a BCMP multi-class open network of queues. The BCMP class is an extension of the Jackson product-form class of models that also has a product-form solution. Bolch et al. (1998) state the BCMP theorem (conditions on service time distributions and queuing disciplines) and give (in Chapter 8) MVA algorithms for deriving expected queue size, delay-in-queue, and server utilisation. However, the beta service time distributions do not satisfy the BCMP theorem and it is well-know that the queue size and delay results of MVA analysis are very sensitive to the service time assumption. For example, if the analysis assumes exponential service times and they are in fact deterministic with the same mean, then delays and queue sizes will be over-stated by the MVA. However, throughput, visit ratios, and server utilisations depend only on the expected arrival and service rates, given the routing matrix and number of servers. The utilisation results are independent of queuing discipline provided it is work conserving, and this includes the preemptive-resume mechanism used here as preemption does not create extra service demand.

Rather than use the complex MVA analysis just to derive utilisations, we developed a simple method that could be carried out on a spreadsheet. We do this by finding the single-class Markov (memoryless) equivalent model of the multi-class model and derive utilisations from it. In a Markov model, where an entity goes next after leaving a

service is independent of its past history. For example, on leaving the technician, all jobs would have the same trinomial distribution of leaving the system, or returning to a NOC/PNOC, or returning to a DSG. But in fact, jobs that came from NOC/PNOC can only return to NOC/PNOC staff. Nevertheless, if service means are not history dependent (which is true for TDRS), then a Markov system with the same net transition rates between services will have the same server utilisations.

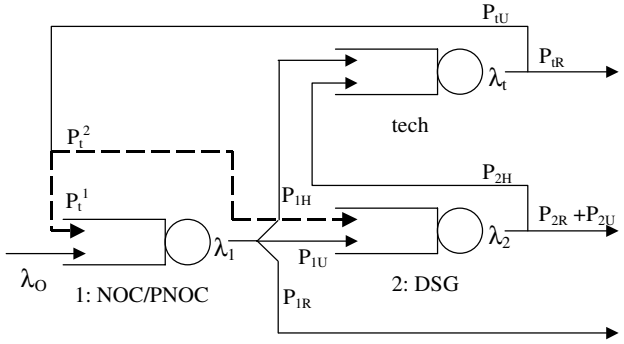


Figure 4: TDRS Throughput Model

The model for the TDRS system for transitions of any one of the criticality levels is shown in Figure 4. The non-Markovian transitions are shown by dotted lines. Of the proportion $P_{tU}$ that are routed back from a tech for more service, a proportion $P_t^1$ goes back to Level 1 service and a proportion $P_t^2 = 1\text{-}P_t^1$ goes back to Level 2. All the other transition rates ($P$) are known (as input parameter values). Input and throughput rates must balance at each queue and are given by $\lambda_0$, $\lambda_1$, $\lambda_2$, and $\lambda_t$ for the rates from the outside, at Level 1 servers, at Level 2 servers and at the technicians, respectively. Balance prescribes that, for the system in Figure 4:

$$[\lambda_O,\lambda_1,\lambda_2,\lambda_t]\times\begin{bmatrix} 0 & 1 & 0 & 0 \\ P_{1R} & 0 & P_{1U} & P_{1H} \\ P_{2R}+P_{2U} & 0 & 0 & P_{2H} \\ P_{tR} & P_t^1 P_{tU} & (1-P_t^1)P_{tU} & 0 \end{bmatrix} = [\lambda_O,\lambda_1,\lambda_2,\lambda_t]$$

We need to solve for $P_t^1$ given that:

$$P_t^1 = \frac{P_{1H}\lambda_1}{P_{1H}\lambda_1 + P_{2H}\lambda_2} \qquad (1)$$

but $\lambda_1$ and $\lambda_2$ are not known *a priori* because of the feedback from the technician queue to both Level 1 and Level 2 queues. This feedback adds to both $\lambda_1$ and $\lambda_2$. The iteration starts by ignoring feedback ($\lambda_1 = \lambda_0$ and $\lambda_2 = P_{1U}\lambda_0$). Then the value of $P_t^1$ is calculated using (1) and substituted into the matrix balance equation. This is solved for new values of $\lambda_1$ and $\lambda_2$ and the process is iterated until it converges. The resultant throughputs are those of the single-class Markov system with equivalent transition and utilisation rates. Each criticality class has its own balance matrix and arrival rate ($\lambda_0$) and is solved for separately. Utilisations can then be calculated using the utilisation form of Little's law:

$$\rho = \lambda\, E[S]/N \qquad (2)$$

where E[S] is the expected service time at a server. This is a mixture of the study and repair times but is calculable as:

$$E[S] = E[\text{study time}] + P_R\, E[\text{repair time}] \qquad (3)$$

A feasible solution does not require all utilisations within a week to be below 1; excess server demand in one shift may be resolved by server availability in the next. Therefore only weekly utilisations are obtained using summation as follows. First the expected service times for each criticality class are calculated using (3). Next, the throughputs for each criticality level are multiplied by the expected service times during each shift to obtain the numerator in (2), the service demanded. Finally the service demanded for each service level is summed over all shifts, yielding the weekly service demanded for each service level. The calculation of weekly service available is more straightforward, involving multiplying the number of staff available per shift by the number of shifts and summing. Finally the ratio of service demanded to service available gives the desired analytic utilization. Agreement between MVA and simulation model was typically excellent (once some errors in simulation logic were detected and fixed!) as shown in Table 2 for one test run:

Table 2: Utilisations Used to Validate

*(a) MVA from spreadsheet*

Total Service Hours

| service | demanded | available | utilization |
|---|---|---|---|
| NOC | 540.5196 | 752 | 0.7188 |
| DSG | 75.6107 | 168 | 0.4501 |
| tech | 287.4166 | 456 | 0.6303 |

*(b) Simulated utilisations over 10 Reps*

| service | mean | 95% c.i. |
|---|---|---|
| NOC | 0.7162 | (0.7080, 0.7244) |
| DSG | 0.4433 | (0.4206, 0.4660) |
| tech | 0.6252 | (0.6166, 0.6338) |

## 5 EXPERIMENTS AND RESULTS

The baseline experiment used current staffing levels and best estimates of input parameters derived from staff interviews. MVA results indicated the system was feasible and dynamic queue results shown to staff were pronounced reasonable. Utilisations were displayed by staff type (i.e., service level) vs. shift type (weekend/weekday etc.) and were mostly below 80%. Queue traces could be displayed either by fault criticality level (Figure 5) or by staff service

level (Figure 6). Figure 5 gives the trace for Major troubles in the system. A weekly effect is evident on the weekends where there are periods of at least one day in length in which the number of troubles in the system is never zero. This is to be expected since major troubles that require Level 2 service on the weekends are less likely to obtain service by call-out and must wait until the next week. No clear daily weekday effect is visible. The number of troubles in the system is more variable than for critical troubles (not shown), which is to be expected since they occur more frequently than critical ones and have lower priority.

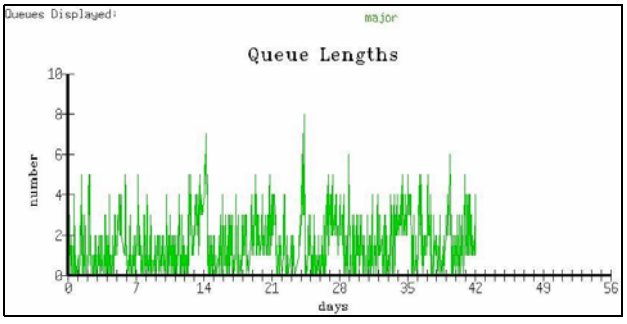


Figure 5: Baseline Trace for Major Troubles in System

On the other hand, queue size of all troubles for Level 2 Service (Figure 6) show very strong daily and weekly cycles because Level 2 (DSG) staff do not work weekends or nights. The spikes on the rising weekend queue sizes result from successful call-outs.

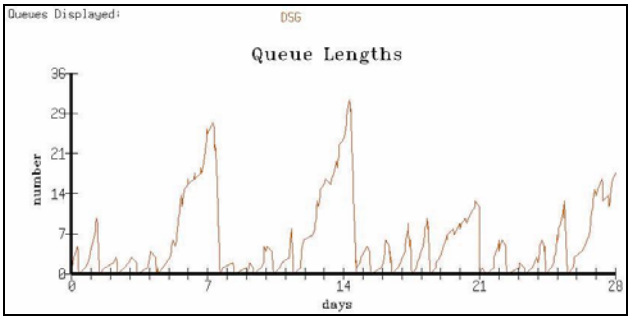


Figure 6: Baseline Trace for Troubles Queued for Level 2

Receipt to close times were also plotted against shift for each criticality level and were thought to be typical. Times of around 2 hours occurred during the weekday shifts and rose to 5 to 6 hours on weekends.

We carried out experiments to show the effects of reallocating staff per shift. We did not have data on the cost of staff by level and by shift, or the overtime costs for call-outs. An objective function for optimising staff allocation would have to take account of these costs as well as quantifying the benefits of improved performance as measured by utilisations, receipt-to-close times, etc. (Table 1). Instead, we did a number of experiments to show the effect of reallocating the most expensive resource, Level 2 (DSG) staff, to show its effects on performance without attempting formal optimisation. The purpose was to demonstrate the sensitivity of performance to such allocation.

There are 6 shift types to which DSG staff can be allocated: weekday night, day, and evening; weekend night, day and evening. A DSG allocation can be designated by the numbers per shift: e.g., the baseline allocation with 2 DSG staff during the weekday day-shift only is d020000. Experiments involved manipulating the allocation of DSG staff to shifts using

- the same total staffing hours (80) per week as in the baseline (d020000 and d011000);
- 2 experiments with 16 additional hours (96 hours) allocated to shifts in 2 different ways (d020010 and d011010);
- a staff reduction to 56 hours (d010010); and
- a "luxury case" experiment (d111111) with all shifts having 1 DSG (168 hours).

Common random numbers were used between runs to generate the identical stream of arrivals and criticality types. The reduced staff model fails the feasibility criterion in MVA and indeed, the utilisation of Level 2 staff is over 1 in the simulation (Figure 7). Re-allocating DSG staff to other shifts improves utilisation and reduces callout hours (second column in Figures 7 and 8). Adding staff may give less improvement than reallocating staff (column 3) but can give large improvements if allocated properly (column 4 in Figures 7 and 8).

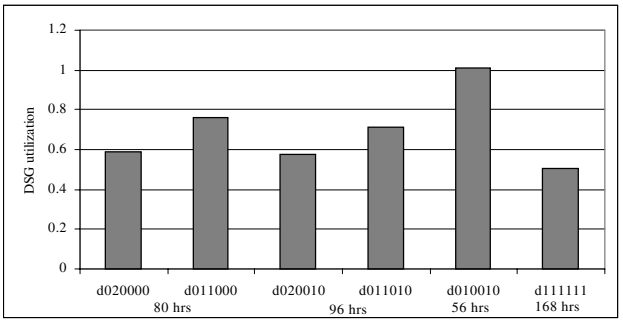


Figure 7: Level 2 Staffing Effects on Utilisation

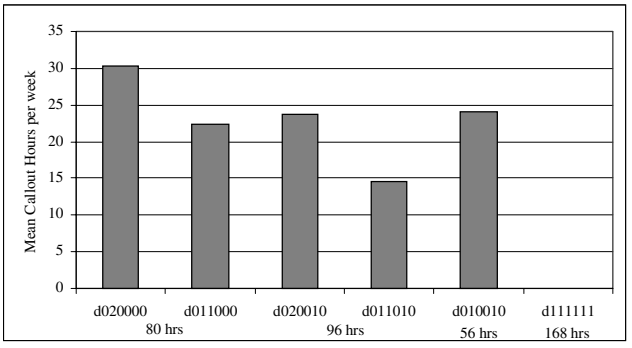


Figure 8: Level 2 Staffing Effects on Call-out Hours

## 6 CONCLUSIONS

The lessons learned in this research included the value of MVA for validation and optimisation. Without the confirmation of MVA on utilisation and routing transition rates we would have had little confidence in the results, even had they seemed plausible. Moreover, MVA would have been very valuable if we had attempted formal optimisation of staffing levels. Given a suitable objective function, there are too many possible allocations of 3 levels of staff to 6 shift types to evaluate by simulation alone. MVA should be used to identify a near-optimal subset of allocations that can be explored more fully by simulation.

We found that the dynamic display is a necessity for studying the behaviour of a heterogeneous system for two reasons. Firstly it provides insights about the system's behaviour over time that are not apparent from mean value and table outputs. Secondly the people involved in the real-world system are better able to interpret the model's behaviour and will give better feedback than if presented with tables full of numbers. Thus support for dynamic graphics is essential in a simulation package. Similar experience was reported by Tanir and Booth (1999) in their study of staffing call centers.

The problem of choosing an appropriate level of granularity for collection of model outputs is one we struggled with. Consequently model outputs are reported with levels of granularity ranging from each individual shift of the week for receipt-to-close times, to an aggregate of 6 shift types over the week for queue lengths, to a single value for total call-out hours. Choices were based on which aspect of the system's behaviour we wished to highlight, from shift effects to comparison of alternate scenarios. Because choices of granularity changed with the questions being asked, it indicates that there is no correct choice for granularity of model outputs, and granularity is an aspect of the simulation the user should be able to select according to the kind of questions he is asking. A simulation language needs powerful and flexible support for statistics specification and reporting by various attribute classes. Dumping disaggregated statistics to a file and reorganising them with a general statistics package does not seem a feasible solution especially given the importance of dynamic displays as well as static summaries.

## ACKNOWLEDGMENTS

Work on TDRS was made possible by a grant from MTS.

## REFERENCES

Bolch, G., S. Greiner, H. deMeer, and K. S. Trivedi. 1998. *Queuing networks and Markov chains*. New York: Wiley Interscience.

Chen, A. L. P., E. J. Cameron, G. F. Shuttleworth, and E. C. Anderson. 1988. A simulation approach for network operations performance studies. In *Proceedings COMPSAC 88: The twelfth international computer software and applications conference*, ed. G. J. Knafl, 105-112. Washington D. C.: Institute of Electrical and Electronics Engineers.

Evans, G. W., T. B. Gor, and E. Unger. 1996. A simulation model for evaluating personnel schedules in a hospital emergency department. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 1205-1209. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Fagerstrom, R., and J. Healy. 1993. The reliability of LEC telephone networks. *IEEE communications magazine* 31 (6): 44-48.

Law, A. M., and W. D. Kelton. 2000. *Simulation modeling and analysis*. 3rd ed. New York: McGraw-Hill.

McGuire, F. 1994. Using simulation to reduce length of stay in emergency departments. In *Proceedings of the 1994 Winter Simulation Conference*, ed. J. D. Tew, S. Manivannan, D. A. Sadowski and A. F. Seila, 861-867. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Regnier, J., and W. H. Cameron. 1990. State-dependent dynamic traffic management for telephone networks. *IEEE communications magazine* 28 (10): 52-53.

Tanir, O., and R. J. Booth. 1999. Call center simulation in Bell Canada. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P.A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 1640-1647. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers. Available online via <http://www.informs-cs.org/wsc99papers/237.PDF> [accessed August 21, 2002].

## AUTHOR BIOGRAPHIES

**GORD BOYER** is an Instructor in the Department of Computer Science, University of Manitoba. His Master's thesis was on the TDRS problem, and he continues work with Neil Arnason on simulation and population analysis software. His email address is <gboyer@cs.umanitoba.ca.>.

**NEIL ARNASON** is a Professor in the Computer Science Department at the University of Manitoba. His Ph D. (Edinburgh 1971) was in population modelling and estimation. His research interests are mainly in animal population models and survey techniques and in simulation methods applied to computer and network systems. His email address is: <arnason@cs.umanitoba.ca.>.