Approximate Queueing Network Analysis of Patient Treatment Times

S.W.M. Au-Yeung, P.G. Harrison and W.J. Knottenbelt
Department of Computing, Imperial College London, SW7 2AZ, UK

Keywords

Queueing Theory, Analytical Models and Approximation Methods

ABSTRACT

We develop an approximate generating function analysis (AGFA) technique which approximates the Laplace transform of the probability density function of customer response time in networks of queues with class-based priorities. From the approximated Laplace transform, we derive the first two moments of customer response time. This technique is applied to a model of a large hospital's Accident and Emergency department for which we obtain the mean and standard deviation of total patient service time. We experiment with different patient-handling priority schemes and compare the AGFA moments with the results from a discrete event simulation.

1. INTRODUCTION

Many complex processing systems in the real world (including telecommunications networks, manufacturing systems, emergency services and healthcare systems) have stringent response time requirements. Such requirements typically relate not only to mean response times, but also to variability of response times. For example, the National Health Service in the UK aims to have an ambulance at the scene of 75% of life-threatening incidents within 8 minutes [15]. In such contexts, it is important to develop effective performance models that can give insights into how different customer processing strategies affect moments and/or distributions of response time.

Over many decades, extensive use has been made of multiclass queueing networks as an effective modelling abstraction. For certain classes of queueing networks, Mean Value Analysis (MVA) [20, 19] and a plethora of related techniques (e.g. [1, 16, 21, 17, 22]) provide an efficient and elegant route to mean values of measures of interest (such mean waiting time and throughput), but not higher moments. For closed queueing networks with underlying (semi-)Markov chains, recent much more computationally-intensive methods based on numerical Laplace transform inversion can be applied to determine exact moments and, where tractable, probability distributions, of customer service times [10, 3, 6]. However, in general, this method suffers from the well-known *state space explosion problem* and so is limited to models with of the order of 10 million states at most. Since accurate models of real life systems typically have much larger state spaces, especially when modelling large numbers of customers, performance analysts must often resort to simulation. And while simulation can be used to model complex systems at arbitrary levels of detail, it typically requires a high cost and effort to construct an accurate model. Further, long execution times are often required to produce reliable results bounded by narrow confidence intervals.

The approximate generating function analysis (AGFA) technique that is the focus of the present paper provides an efficient analytical way to approximate the mean and variance of response time in networks of multiclass queues with blocking and class-dependent priorities. Support for the latter two phenomena allow the technique to be applied in diverse modelling scenarios, as illustrated by the hierarchical queueing network model of an Accident and Emergency department presented later as a case study. The remainder of this paper is organised as follows. Section 2 presents technical details of the AGFA technique. Section 3 describes our case study, a queueing network model of a hospital A&E department. Section 4 presents the numerical results and graphs from both the AGFA method and a Java simulation. Section 5 concludes and considers future research directions.

2. APPROXIMATE GENERATING FUNCTION ANALYSIS

The essence of the technique is that of Cobham's formula for calculating mean values of response times in $M/G/1$ queues. This uses the fact that the mean value of a sum of random variables is equal to the sum of the corresponding means, whether or not the variables are independent. Furthermore, given the mean sojourn time of a low priority customer in a queue, the mean number of higher priority arrivals in that time can be calculated. This analysis is adapted to the calculation of the Laplace transform of response time probability density, which is the expectation of the exponential function of a sum of random variables. Single nodes are analysed in this way, after which sub-networks are solved and aggregated according to the hierarchical MVA approach [20, 7].

Valuetools '07, October 23-25, 2007, Nantes, France

2.1 Notation

We consider a network with two customer classes and M multi-server nodes, with m_i constant rate exponential servers with rates μ_{ir} at node i for class r ($1 \le i \le M, r = 1, 2$). Class 1 has non-pre-emptive priority over class 2. In particular, we consider the passage of a special 'tagged' customer through queueing node i and define the following random variables at equilibrium:

- $\mathbf{K} = (K_1, K_2)$ class population vector, i.e. there are K_r customers of class r in the network ($r = 1, 2$);
- B_{ir} class r service time of a single server at the node, exponential with parameter μ_{ir};
- L_{ir} number of class r customers in the queue waiting to start service;
- Q_{ir} time spent by a class r customer waiting to *start* service;
- N_{ir} number of class r customers in the queue, including any in service, at a random instant of time (i.e. the class r queue length);
- W_{ir} response time of a class r customer, i.e. the sum of queueing time and service time, $Q_{ir} + B_{ir}$;

Let the steady state probability that the (joint) queue length at node i is $\mathrm{n} = (n_1, n_2)$ be $\pi_i(\mathrm{n} \mid \mathbf{k}) = \mathbb{P}(N_{i1} = n_1, N_{i2} = n_2 \mid \mathbf{K} = \mathbf{k})$. We will make use of the probability that an arriving customer has to queue, q_{ir}. In a network with processor sharing servers and no priorities, this is just the probability that the equilibrium queue length is less than m_i when the population is reduced by one in the arriving customer's class, by the arrival theorem (see [11] for example). Thus, $1 - q_{ir} = \mathbb{P}(N_{i1} + N_{i2} < m_i \mid \mathbf{K} = \mathbf{k}) = \sum_{u=0}^{m_i-1} \sum_{v=0}^{m_i-1-u} \pi_{ir}(u, v \mid \mathbf{k}^{r-})$ for appropriate $\mathbf{k}, \mathbf{k}^{r-} (r = 1, 2)$.

For a continuous random variable X, we denote its probability distribution function by $X(t) = \mathbb{P}(X \le t)$ and the Laplace-Stieltjes transform of this distribution (the LSTD) by $X^*(\theta) = E[e^{-\theta X}]$ [1] We denote the density function by $x(t) = X'(t)$, the derivative of the distribution function, with Laplace transform $X^*(\theta)$. We also denote the nth moment of X by $X_{;n} = E[X^n] = (-1)^n X^{*(n)}(0)$ (where the parenthesized superscript denotes differentiation n times). Thus, for example, $S_{2;1}$ is the mean of S_2.

For a discrete random variable Y, we denote its probability generating function (pgf) by $G_Y(z) = E[z^Y]$ and the nth *factorial moment* of Y by $Y_{;fn} = E[Y(Y-1)\dots(Y-n+1)] = G_Y^{(n)}(1)$.

2.2 An approximate MVA algorithm

2.2.1 Class 1

The high priority class 1 customers are straightforward to handle since the tagged customer only has to wait for those class 1 customers already queueing and the customer in service (of either class), if any. Consider a generic node i in a closed network of M queues. Dropping the subscripts ir for brevity, we have for class 1:

$$Q_1^*(\theta) = E[\, E[e^{-\theta(S_1+\dots+S_L+UR)} \mid N_1, N_2]\,]$$

[1] $E[\cdot]$ and $E[\cdot|\cdot]$ denote the expectation and conditional expectation operators.

where the random variable U is defined by:

$$U = \begin{cases} 1 & \text{if } N_1 + N_2 \ge m \\ 0 & \text{if } 0 \le N_1 + N_2 < m \end{cases}$$

The random variables S_l are independent and identically distributed (i.i.d.) as the minimum of the m service time random variables at the individual servers. Therefore each is exponential with parameter $m\mu$ in a single class node. In the multiclass case, they are still exponential but have parameter $m_1\mu_1 + (m - m_1)\mu_2$ when there are m_1 class 1 and $m - m_1$ class 2 customers in service. We make the approximating assumption that, given the class of the tagged customer, the network's population vector $\mathbf{k}$ and the state encountered on arrival $\mathbf{n}$, this rate remains the same throughout the tagged customer's sojourn in the queue, Q_1, viz. $n_1\mu_1 + n_2\mu_2$ if $n_1 + n_2 \le m$ and $(n_1\mu_1 + n_2\mu_2)m/(n_1 + n_2)$ if not. Of course, this result is exact if the service rate is the same for both classes ($\mu_1 = \mu_2$) and in the single class case ($n_2 = 0$). Note too that we would not have this problem if the priority discipline were pre-emptive, whereupon no class 2 customer could be in service if m or more class 1 customers were present.

The random variable R is the time to the next service completion from the arrival instant of the tagged customer. By the memoryless property, R is distributed as S and is also independent by hypothesis.

We therefore obtain (for class 1):

$$\begin{aligned} Q_1^*(\theta) &= E[\, S^*(\theta)^{L_1}\, E[e^{-\theta UR} \mid N_1, N_2]\,] \\ &= E[\, S^*(\theta)^{L_1}\, R^*(\theta U)] \\ &= G_{L_1}(S^*(\theta)) R^*(\theta) - \\ &\quad (1 - q_1) R^*(\theta) + 1 - q_1 \end{aligned} \tag{1}$$

(noting that $R^*(\theta U) = R^*(\theta) + (1 - R^*(\theta))(1 - U)$) where $G_{L_1}(z) = 1 - q_1 + \sum_{u=m}^{k_1} \sum_{v=m-u}^{k_2} \pi(u, v) z^{u-m_1}$ and m_1 is the number of class 1 customers in service when all servers are busy. This can approximated for non-pre-emptive priority, as above, but in our calculation of moments it comes from the application of Little's result in the extended MVA algorithm. Note that $R^* = S^*$ and, in the calculations of q and $G_{L_1}(z)$, we assume a population vector $\mathbf{k}$ in which the component corresponding to the class of the arriving customer has been reduced by one (in accordance with the Arrival Theorem) [2].

The moments of the class 1 queueing time follow by differentiation at $\theta = 0$. For the first few moments this is a straightforward process, but the n-fold differentiation of the term $G_{L_1}(S^*(\theta))$ for arbitrary n leads to ever-increasing

[2] Notice that in the case of a single class $M/M/m$ queue with constant arrival rate λ, we have $S^*(\theta) = m\mu/(m\mu + \theta)$ and

$$\begin{aligned} G_L(z) &= 1 - q + (1 - \rho)q + (1 - \rho)q \sum_{l=1}^{\infty} \rho^l z^l \\ &= 1 - q + \frac{(1 - \rho)q}{1 - \rho z} \end{aligned}$$

where $\rho = \lambda/(m\mu)$. Consequently, we obtain the well known result

$$Q_1^*(\theta) = 1 - q + \frac{(1 - \rho)qm\mu}{m\mu + \theta - \rho m\mu} = 1 - q + q\frac{(m\mu - \lambda)}{m\mu - \lambda + \theta}$$

complexity. It can be obtained simply using a programming language that supports higher-order functions – here differentiation with respect to θ – and otherwise using an auxiliary recursive definition [8]. Here we obtain the first two moments explicitly.

Mean queueing time for class 1.

Differentiating the class 1 queueing time LSTD given by equation 1, we find

$$\begin{aligned} Q_1^{*\prime}(\theta) &= G'_{L_1}(S^*(\theta))S^{*\prime}(\theta)S^*(\theta) + \\ &\quad G_{L_1}(S^*(\theta))S^{*\prime}(\theta) - (1-q_1)S^{*\prime}(\theta) \end{aligned} \quad (2)$$

At $\theta = 0$, we obtain

$$Q_{1;1} = L_{1;1}S_{1;1} + S_{1;1} - (1-q_1)S_{1;1} = (L_{1;1} + q_1)S_{1;1}$$

as could have been obtained by a simple direct argument.

Second moment of queueing time for class 1.

Differentiating equation 2 at $\theta = 0$ we find similarly

$$Q_{1;2} = L_{1;f2}S_{1;1}^2 + L_{1;1}S_{1;2} + 2L_{1;1}S_{1;1}^2 + S_{1;2} - (1-q_1)S_{1;2}$$

which simplifies to

$$Q_{1;2} = (L_{1;2} + L_{1;1})S_{1;1}^2 + (L_{1;1} + q_1)S_{1;2}$$

2.2.2 Class 2

A class 2 customer has to wait, not only for the service completion of any customer in service at its arrival instant and all class 1 and 2 customers already waiting, but also for all class 1 customers that arrive during its queueing time. As with class 1 customers, we assume that the total service rate remains constant throughout a class 2 customer's sojourn time in the queue, so that service times are the same random variables S, R that depend only on the state of the queue on arrival, (n_1, n_2).

Let C be the number of class 1 arrivals during the tagged customer's queueing time Q_2. Since these arrivals are assumed to be Poisson with rate λ_1 (and so have pgf $e^{-\lambda_1 t(1-z)}$ for a time period t), C has pgf defined by

$$\begin{aligned} G_C(z) &= E[z^C] = E[\,E[z^C \mid Q_2]\,] \\ &= E[e^{-\lambda_1 Q_2(1-z)}] = Q_2^*(\lambda_1(1-z)) \end{aligned}$$

Writing $H = \max(N_1 + N_2 - m, 0)$, we therefore have:

$$\begin{aligned} Q_2^*(\theta) &= E[\,E[e^{-\theta(S_1+\ldots+S_H+S_{H+1}+\ldots+S_{H+C}+UR)} \mid N_1, N_2, Q_2]] \\ &= E[S^*(\theta)^H R^*(\theta U) E[S^*(\theta)^C \mid Q_2]] \\ &= E[S^*(\theta)^H R^*(\theta U) e^{-\lambda_1 Q_2(1-S^*(\theta))}] \end{aligned} \quad (3)$$

Mean queueing time for class 2.

Setting out as for class 1, we first differentiate the class 2 queueing time LSTD given in equation 3 to find

$$\begin{aligned} Q_2^{*\prime}(\theta) &= E[HS^*(\theta)^{H-1}R^*(\theta U)e^{-\lambda_1 Q_2(1-S^*(\theta))}S^{*\prime}(\theta)] + \\ &\quad E[S^*(\theta)^H R^{*\prime}(\theta U)Ue^{-\lambda_1 Q_2(1-S^*(\theta))}] + \\ &\quad E[S^*(\theta)^H R^*(\theta U)\lambda_1 Q_2 e^{-\lambda_1 Q_2(1-S^*(\theta))}S^{*\prime}(\theta)] \\ &= E[S^*(\theta)^{H-1}e^{-\lambda_1 Q_2(1-S^*(\theta))}\big(HR^*(\theta U)S^{*\prime}(\theta) + \\ &\quad US^*(\theta)R^{*\prime}(\theta U) + \lambda_1 Q_2 S^*(\theta)R^*(\theta)S^{*\prime}(\theta)\big)] \end{aligned} \quad (4)$$

At $\theta = 0$, we therefore obtain

$$Q_{2;1} = H_{2;1}S_{2;1} + q_2 R_{2;1} + \lambda_1 Q_{2;1}S_{2;1}$$

since $E[U] = q$. This gives Cobham's familiar result for mean values (see for example [11]):

$$Q_{2;1} = \frac{(H_{2;1} + q_2)S_{2;1}}{1 - \lambda_1 S_{2;1}}$$

The mean values $S_{2;1}, q_2$ and $H_{2;1}$ depend on the state existing just before an arrival instant, as discussed above, and can be computed as part of the standard variable rate MVA algorithm that we use.

Second moment of queueing time for class 2.

Although the analysis of mean values is straightforward, not actually needing generating functions at all, the situation is much more complex for higher moments because of the dependence amongst the random variables concerned – Q_i, L_i, U. In particular, this leads to covariance terms in the second moments.

We therefore define the two-variable generating function $A(z, \theta)$ by

$$\begin{aligned} A(z,\theta) &= E[z^H e^{-\theta Q_2}] = E[E[z^H e^{-\theta Q_2} \mid N_1, N_2]] \\ &= E[z^H S^*(\theta)^H R^*(\theta U) e^{-\lambda_1 Q_2(1-S^*(\theta))}] \end{aligned}$$

by the same reasoning as in the previous section. Taking the expectation w.r.t. U,

$$\begin{aligned} A(z,\theta) &= 1 - q_2 + \\ &\quad R^*(\theta)\left\{E\Big[\big(zS^*(\theta)\big)^H e^{-\lambda_1 Q_2(1-S^*(\theta))}\Big] - (1-q_2)\right\} \\ &= (1-q_2)\big(1 - S^*(\theta)\big) + \\ &\quad S^*(\theta)\,A\big(zS^*(\theta), \lambda_1(1-S^*(\theta))\big) \end{aligned}$$

Now let $y = zS^*(\theta) and \phi = \lambda_1(1 - S^*(\theta))$ so that $y = 1$ and $\phi = 0$ when $z = 1$ and $\theta = 0$. Using primes to denote differentiation of a function of a single variable and the facts that $\frac{\partial y}{\partial z} = S^*(\theta)$, $\frac{\partial y}{\partial \theta} = zS^{*\prime}(\theta)$, $\frac{\partial \phi}{\partial z} = 0$, $\frac{\partial \phi}{\partial \theta} = -\lambda_1 S^{*\prime}(\theta)$, so that $\partial/\partial\theta = zS^{*\prime}(\theta)\partial/\partial y - \lambda_1 S^{*\prime}(\theta)\partial/\partial\phi$, we obtain:

$$\begin{aligned} \frac{\partial A}{\partial \theta} &= \big(A(y,\phi) + q_2 - 1\big)S^{*\prime}(\theta) + \\ &\quad S^*(\theta)S^{*\prime}(\theta)\left(z\frac{\partial A}{\partial y} - \lambda_1\frac{\partial A}{\partial \phi}\right) \end{aligned} \quad (5)$$

Thus, at $z = 1, \theta = 0$, we obtain $-Q_{2;1} = -q_2 S_{2;1} - S_{2;1}(H_{2;1} + \lambda_1 Q_{2;1})$ so that

$$Q_{2;1}(1 - \lambda_1 S_{2;1}) = H_{2;1}S_{2;1} + q_2 S_{2;1}$$

as obtained already. Differentiating again at $z = 1$ and $\theta = 0$, omitting the arguments of functions for brevity, where the meaning is clear, and noting that $\frac{\partial z}{\partial y} = 1/S^*$, $\frac{\partial z}{\partial \phi} = z/(\lambda_1 S^*)$ so that $z\frac{\partial z}{\partial y} = \lambda_1 \frac{\partial z}{\partial \phi}$, we now find

$$\begin{aligned} \left.\frac{\partial^2 A}{\partial \theta^2}\right|_{1,0} &= q_2 S_{2;2} + 2S_{2;1}^2[H_{2;1} + \lambda_1 Q_{2;1}] + S_{2;2}[H_{2;1} + \lambda_1 Q_{2;1}] - \\ &\quad S_{2;1}^2\left[z\left(z\frac{\partial^2 A}{\partial y^2} - \lambda_1\frac{\partial^2 A}{\partial y \partial \phi}\right) - \lambda_1\left(z\frac{\partial^2 A}{\partial \phi \partial y} - \lambda_1\frac{\partial^2 A}{\partial \phi^2}\right)\right]_{1,0} \\ &= q_2 S_{2;2} + (H_{2;1} + \lambda_1 Q_{2;1})(S_{2;2} + 2S_{2;1}^2) + \\ &\quad S_{2;1}\left[H_{2;f2}S_{2;1} - 2\lambda_1 S_{2;1}\frac{\partial^2 A}{\partial y \partial \phi} + \lambda_1^2 S_{2;1}Q_{2;2}\right] \end{aligned} \quad (6)$$

We compute the covariance term $\frac{\partial^2 A}{\partial y \partial \phi}$ at $z = 1, \theta = 0$ as follows. First,

$$\frac{\partial A}{\partial z} = S^* \frac{\partial A}{\partial y}$$

since $\frac{\partial \phi}{\partial z} = 0$. Differentiating w.r.t. θ now gives

$$\frac{\partial^2 A}{\partial z \partial \theta} = S^{*\prime} \frac{\partial A}{\partial y} + S^* \left[\frac{\partial^2 A}{\partial y^2} z S^{*\prime} + \frac{\partial^2 A}{\partial y \partial \phi} (-\lambda_1 S^{*\prime}) \right]$$

At $z = y = 1, \theta = \phi = 0$, this yields

$$\left. \frac{\partial^2 A}{\partial z \partial \theta} \right|_{1,0} = -\frac{(H_{2;1} + H_{2;f2}) S_{2;1}}{1 - \lambda_1 S_{2;1}}$$

Finally, substituting into equation 6 at $z = 1, \theta = 0$, we obtain

$$\begin{aligned} Q_{2;2} &= \frac{q_2 S_{2;2} + (H_{2;1} + \lambda_1 Q_{2;1})(S_{2;2} + 2S_{2;1}^2) + H_{2;f2} S_{2;1}^2}{1 - \lambda_1^2 S_{2;1}^2} \\ &\quad + \frac{2\lambda_1 S_{2;1}^3 H_{2;2}}{(1 - \lambda_1 S_{2;1})(1 - \lambda_1^2 S_{2;1}^2)} \end{aligned} \tag{7}$$

$Q_{2;1}$ was computed in the previous subsection and, again, the expected value $H_{2;2}$ is computed in the MVA-based algorithm, considering the superposition of the two classes. The second moment $S_{2;2}$ is approximated as the average of the square of the service time of a single server, estimated at equilibrium when all servers are busy. This double approximation is a potentially major source of error in our model; however, it is exact when the two classes have identical service time random variables.

2.2.3 The MVA-based hierarchical model

Apart from the aforementioned moments of the time to the next service completion after the arrival instant of the tagged customer, the only state-dependent parameters that are needed for constant-rate, multi-server queues are the queueing probabilities q_1, q_2. In the case of a single server at equilibrium, this is just the utilisation, which is known to be the product of the arrival rate and the mean service time, by the usual steady-state argument or Little's result. However, multiple servers or state-dependent service times require that every (significant) queue length probability be computed in order to find q_1, q_2 and the first two moments of S_1, S_2 – at each queue and for each network population vector in a closed network. This is the main expense of the algorithm. It also goes some way to explaining why the problem has for long been solved for $M/G/1$ queues but remains open for $M/G/m$ for $m > 1$. [3]

Network decomposition and aggregate servers.

In our hierarchical modelling methodology, we successively decompose a queueing network of multi-servers, where each individual server has constant rate, into a collection of sub-networks. This is a common approach to modelling large systems, pioneered to a considerable extent by Woodside and others in their analysis of layered queueing networks; see for example [7]. The sub-networks we identify as most appropriate are each solved, using the AGFA approach described in the previous subsections, for the first two moments of their response time, given each (multi-server) node's service time moments, the network's routing probabilities and the constant populations of its customer-classes. No class transitions are allowed within a sub-network (which constrains the choice of sub-networks, of course). Each node in a sub-network is analysed using the results of the previous section and Little's result (for both the first and second moments of queue length and waiting time) at class populations increasing from 0 to the maximum required. This is done in a straightforward modification of the standard MVA algorithm with state-dependent parameters to yield the required first two moments [11]. At the next level up, these moments are assigned to those of the individual service times at the corresponding multi-server nodes. The number of parallel servers at each node is set to the maximum population specified for each class in the sub-network at the lower level – recall that no class transitions occur. Hence the class population maxima are preserved all the way up the hierarchy. The higher-level network is then fully parameterised by its routing probabilities, easily obtainable from the initial, flat network's specification.

At the top level, we analyse an open network of aggregated nodes, at any of which there may be interaction between the classes. In particular, the service rate of each class may depend on the joint population of both classes currently at the node. Such a node is solved by a direct Markov model with state space truncation. This is not excessively expensive for a single node with just two classes, and in fact no more than about 2000 states are needed in practice. Nevertheless, the calculation requires the major share of the computation time of the whole algorithm.

The only remaining quantities needed for the hierarchical algorithm are the mean and second moments of the numbers of visits a task makes to each node. These are derived in the next section. This whole decomposition was implemented on a Macintosh G5 computer running Mathematica 5.1 under OS X version 10.4. Clearly a lower level implementation in a language such as C would be orders of magnitude more efficient numerically, and benefit especially the single node, direct Markov models discussed above.

2.2.4 Moments of visit counts

The MVA algorithms, open or closed, require the same moments of the nodes' visit counts as those required for response time. To this end, let the random variable V_{ir} denote the number (or rate) of visits a task of class r makes to node i and let X_{ir} be the visit count (or rate) of external arrivals, $1 \le i \le M, 1 \le r \le R$. Then we have

$$E[z^{V_{ir}}] = E[z^{X_{ir} + \sum_{j,s} N_{js;ir}}]$$

where $N_{js;ir}$ is the number of class s service completions at node j that go to node i as class r. Thus

$$\begin{aligned} E[z^{V_{ir}}] &= E[E[z^{X_{ir} + \sum_{j,s} N_{js;ir}} \mid V_{js}]] \\ &= E[z^{X_{ir}}] \prod_{j,s} E\left[\left(1 - p_{js;ir}(1 - z)\right)^{V_{js}} \right] \end{aligned}$$

since the random variables $N_{js;ir}$ are independent and binomially distributed with parameters $(V_{js}, p_{js;ir})$. Hence we have

$$G_{V_{ir}} = G_{X_{ir}} \prod_{j=1}^{M} \prod_{s=1}^{R} G_{V_{js}}\left(1 - p_{js;ir}(1 - z)\right)$$

[3] Notice too the subtle dependence that arises when considering non-exponential servers that precludes simply setting the moments of S_1, S_2 to those of the residual service time [9].

Differentiating once, then twice at $z = 1$ then yields:

$$V_{ir;1} = X_{ir;1} + \sum_{j=1}^{M}\sum_{s=1}^{R} p_{js;ir} V_{js;1} \quad (8)$$

$$V_{ir;f2} = V_{ir;1}^2 + X_{ir;f2} - X_{ir;1}^2 + \sum_{j=1}^{M}\sum_{s=1}^{R} p_{js;ir}^2 \left(V_{js;f2} - V_{js;1}^2\right) \quad (9)$$

3. ACCIDENT AND EMERGENCY MODEL

3.1 Description

To illustrate the use of the combined AGFA-MVA technique, which we abbreviate to just AGFA, we apply it to a model of an Accident and Emergency (A&E) department [2]. The AGFA technique is especially suited to the analysis of healthcare systems as these often have limited resources, contain blocking phenomena and have class-based priorities.

The idea of using queueing theory and networks to model health service departments is, of course, by no means new. Several studies have been made of patient flow in hospitals in general [18, 4, 12] and Emergency departments in particular [5, 13, 14]. However, these studies have had limited success and subsequent impact for two main reasons. Firstly, there has been a lack of sophistication in the models with very simple high-level queueing models being used which typically do not tie system performance to the underlying resources (so we cannot, for example, assess the response time impact of employing an extra nurse or purchasing an extra x-ray machine). Secondly, many of these models do not take into account phenomena that occur in the corresponding real life systems such as blocking and class-based priority queueing.

Figs. 1 and 2 show the simplified hierarchical queueing network model of patient flow we have developed in conjuction with an A&E consultant. The model takes the form of a network of $M/M/m$ queues with two forms of patient arrivals: walk-in patients who come into A&E via their own transport and patients that arrive by ambulance.

3.1.1 Walk-in Patients

These patients enter via the A&E waiting room where they are registered at reception. The receptionists route each patient into one of three queues: patients with a clear case of minor trauma or illness are placed in the minors queue; patients with a clear case of a serious trauma or illness are sent to the majors queue; all others are sent for nurse assessment.

Minors Queue.

Patients in the minors queue must first wait for a minors cubicle to become free; the patient then waits there for a minors practitioner (either a minors doctor or a nurse practitioner) to see them. The minors practitioner can decide to:

- perform investigative tests such as blood tests and x-rays, or
- ask for a specialist opinion, or
- treat (if necessary) and discharge the patient (to home, their GP or to the pharmacy to pick up medication), or
- send the patient to be admitted to a (surgical) ward, or the MAU (Medical Assessment Unit) which assesses the need for medical admissions.

Majors Queue.

Patients in the majors queue wait for a bed in a majors bay to become free; once there, a nurse performs tests (vitals, bloods, x-ray etc.) so that essential information is ready for a doctor. Tests for both majors and minors are processed in the same laboratory facility. When (s)he has assessed the patient, the doctor may require a specialist opinion, require more tests, or send the patient out of A&E (possibly after treatment) via the routes mentioned above in the minors queue. Occasionally a patient may suffer a sudden and rapid deterioration; in such a case the patient is transferred to a resuscitation bay and is attended to by a resuscitation team.

Nurse Assessment.

Patients in the nurse assessment queue wait for an assessment room to become available; they then wait there for a nurse who assesses the severity of their illness or injury. The nurse can send the patient either to the minors queue, the majors queue or discharge them out of A&E to a specialist clinic, ward, GP etc.

Specialists.

Specialists may be called in by a minors practitioner or majors doctor. Minors patients are only referred to "other" specialists which encompass ENT (ear, nose and throat), Gynaecology and Orthopaedics. Majors patients may be seen by medical, surgical and "other" specialists. After assessment, patients are discharged from A&E, either being sent to a clinic for a more thorough investigation, being admitted to a ward or being sent to the MAU.

3.1.2 Ambulance Arrivals

Ambulance arrivals are split into two types: Standard arrivals who are patients that do not require immediate medical treatment and Blue Call arrivals who are very seriously ill or injured patients that require urgent medical attention.

Standard Arrivals.

These patients are handed over to a nurse from the ambulance. The nurse assesses the patient, decides which queue to assign them to, and either sends them to reception to be registered or straight to a majors bay (as appropriate).

Blue Call Arrivals.

These patients are assigned a resuscitation bed and are attended to by a resuscitation team. Once stable, the patient leaves A&E, being sent either to an operating theatre, to the ITU (Intensive Treatment Unit), or to a ward. Patients who cannot be resuscitated are sent to the mortuary.

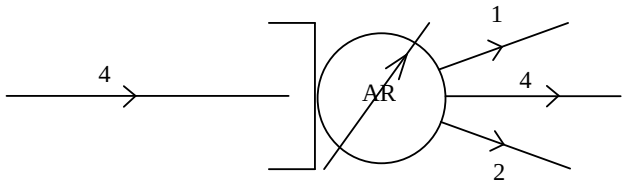
3.1.3 Passive resources

In many cases a patient needs to obtain a (passive) resource before they can progress along a treatment path, holding the resource until the patient no longer needs it. It is this resource possession that can lead to blocking phenomena occuring in the system. An example is the nurse assessment rooms (of which there are 5 in our A&E department). A patient must wait for one to become free before

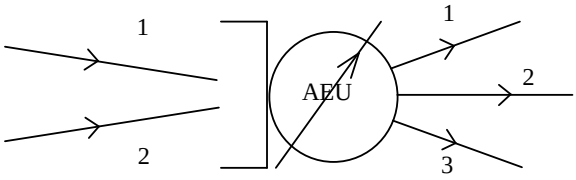
Patient Classes

1. Minors
2. Majors
3. Resusc
4. Assessment

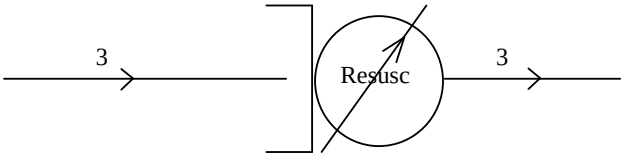
Aggregated Server AR (Assessment room)



Aggregated Server AEU (Whole medical unit)



Aggregated Server Resusc (Blue Call)



Top-Level Model

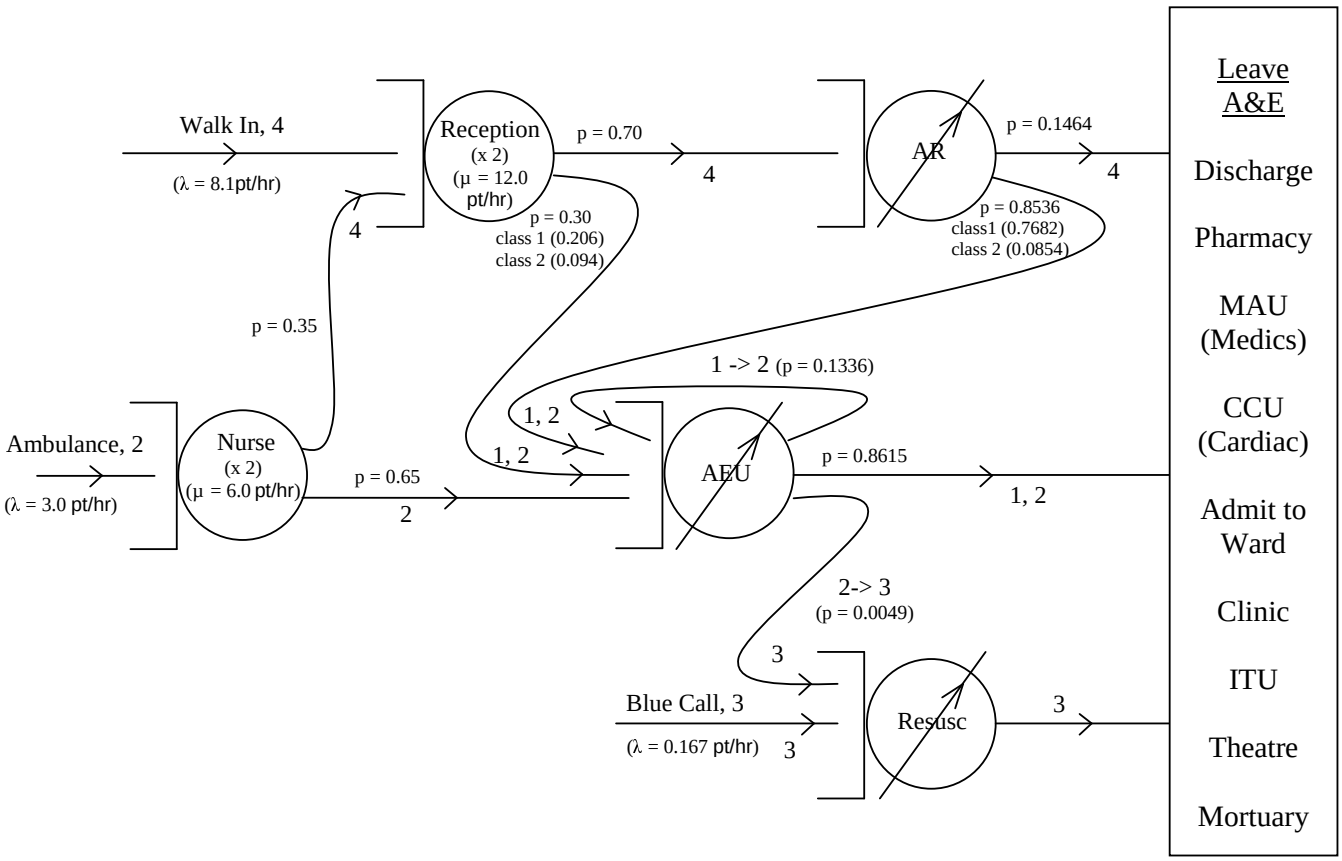
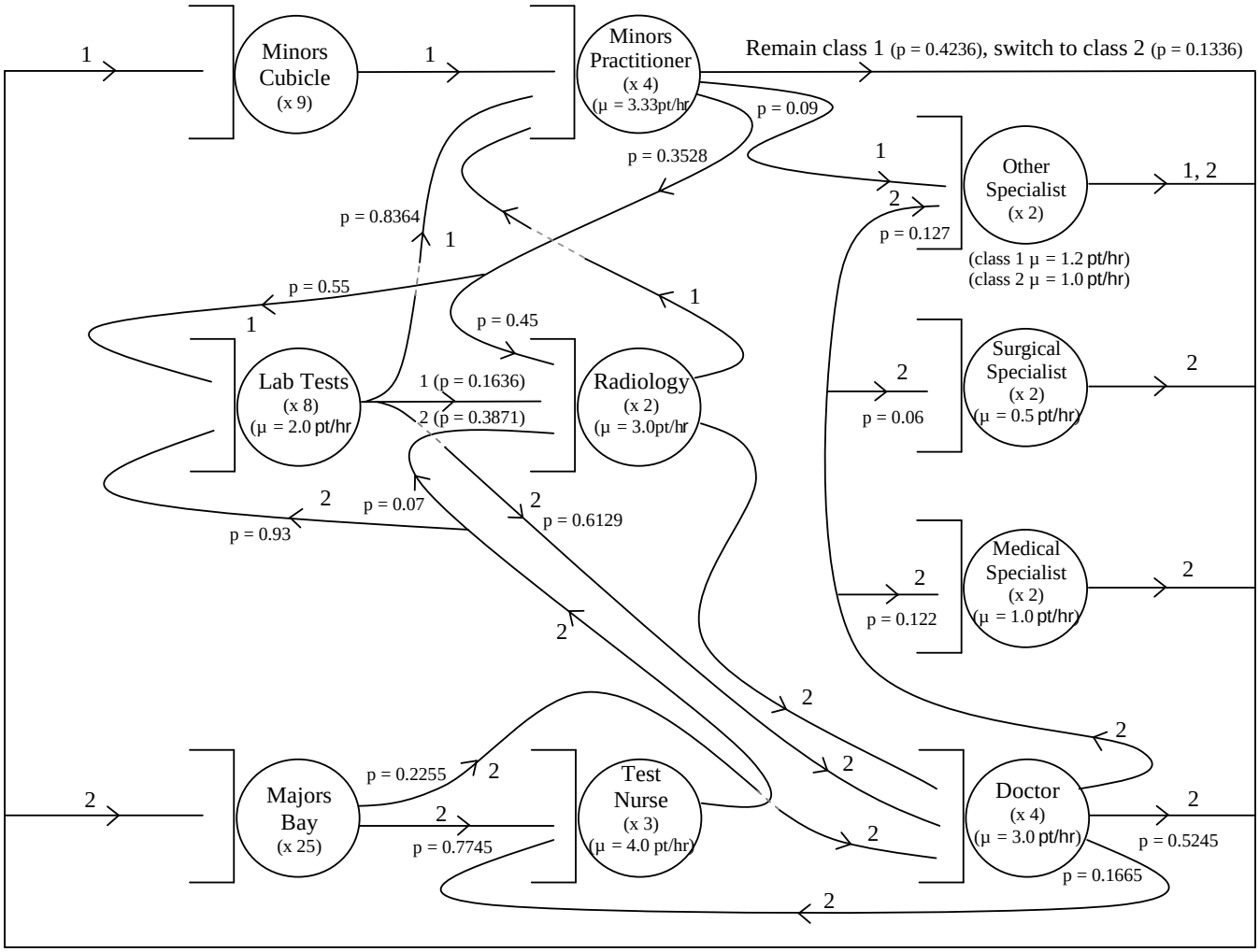
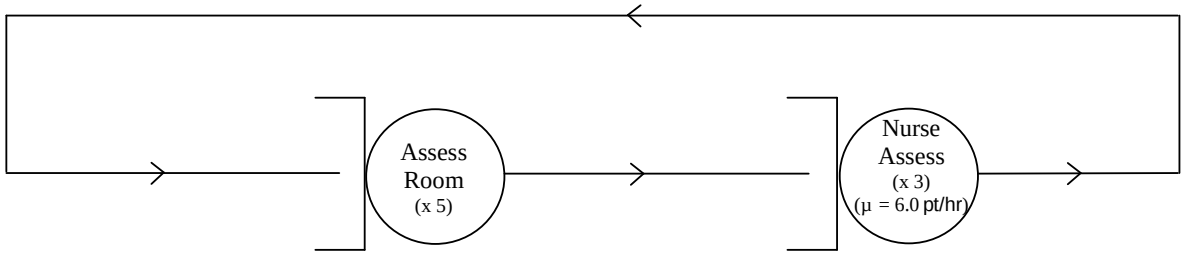


Figure 1: Top level of queueing network model of patient flow

AEU Submodel



AR Submodel



Resusc Submodel

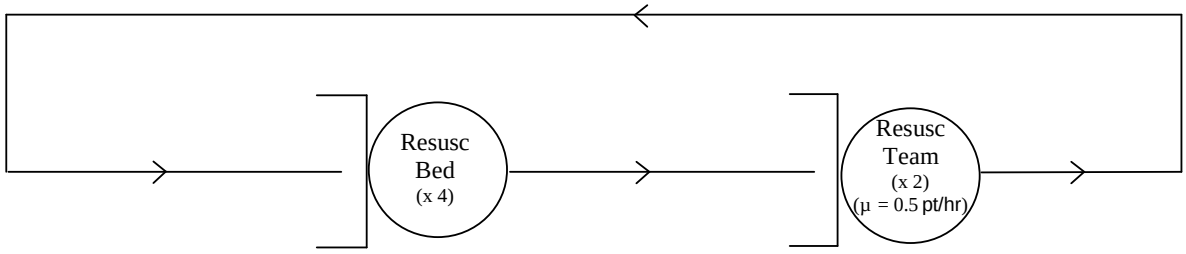


Figure 2: Lower levels of queueing network model of patient flow

entering the room for assessment by a nurse. Once the assessment has been completed, the patient leaves the room, freeing it up for the next patient. Other passive resources include minors cubicles (of which there are 9), majors bays (of which there are 25) and resuscitation beds (of which there are 4). There is an added deterministic delay of 1 minute in acquiring a passive resource to account for the time it takes for a patient to move to the resource when it becomes available.

In the AGFA-MVA model, we aggregate each passive resource and all its associated *active* resources, i.e. those providing a service that actually progresses a patient through treatment, into a single node in the higher level model. Where these active resources include one shared with another class that is also associated with another passive resource, the union of the two sets of resources, associated with each passive resource, is aggregated – essentially giving a transitive closure. In our model, this leads to the AEU aggregate node, which includes majors bays, minors cubicles and all their associated resources described above. This is solved using a closed, two-class queueing network model, incorporating AGFA.

4. NUMERICAL RESULTS

4.1 Mean and variance of patient response time

Tables 1, 2, 3, 4, 5, 6, 7 and 8 compare the mean and standard deviation of patient response time as calculated using our discrete-event simulation and the AGFA technique for various types of patient arrival (Walk-in, Ambulance and Blue call arrivals) – under 25%, 50%, 75%, 80%, 85%, 90%, 95% and 100% arrival rates for both ambulance and walk-in arrivals; the blue call arrival rate is held constant. Figures 3, 4 and 5 compare the simulation and AGFA results shown in these tables, illustrating clearly the loading levels at which patient response times start to increase rapidly as the department approaches saturation.

The simulation results presented are the average of ten runs. Each run includes a transient period, during which 2 000 000 patients move through the system (and during which passage time statistics are not collected), followed by a measurement period which lasts long enough to observe 10 000 passages of Blue Call arrivals through the system; in this period around 485 000 passages of Walk-in arrivals and 180 000 passages of Ambulance arrivals are also observed.

Three different patient priority schemes are analysed:

- *No Priority* in which First In First Out (FIFO) queues are implemented,
- *Majors Priority* in which majors patients are given priority at the shared resources (i.e. lab tests, radiology and "other" specialist), and
- *Minors Priority* in which minors patients are given priority at the shared resources.

>From our results, we see that the mean values obtained via the AGFA technique show very good agreement with our simulation results, especially under minors priority and for workloads up to 80% under the other priority systems. It seems that the mean values diverge as the system becomes more highly utilised, as is illustrated by the results under the majors priority system in particular; here patients remain in the department for longer, resulting in greater saturation. It is well known that both approximate analytical methods and simulation tend to suffer from loss of accuracy in saturated systems.

As expected, the AGFA results for the standard deviations are generally not as good a match against the simulation, but they are still within 25% for workloads up to 80% for minors priority and within 90% under the other priorities. This is partly because, although aggregation can be shown to preserve many expected values of random variables associated with the queueing processes concerned, the same cannot be said for higher moments. Furthermore, the approximations pointed out in the AGFA analysis of Section 2 become more significant at higher moments.

In terms of run times, each simulation run required approximately 30–40 minutes wall clock time (depending on the priority scheme and the PC cluster workstation used), with results for each priority scheme and workload combination being averaged over 10 runs to obtain confidence intervals on the means. By contrast, AGFA required between 30 seconds and 5 minutes (in the saturated Majors priority case) wall clock time for each (single-run) priority scheme and workload combination.

Finally we consider some interesting insights into how differing workloads and priority systems affect the walk-in and ambulance arrival patient service times in our A&E model. We can see that under low loading (25%) the priority scheme makes no difference to the results for either of the arrival types. Under medium loading (50% to 75%) we can see that the ambulance arrivals perform better under majors priority (as expected). The walk-in arrivals are not much affected by the priority scheme used, with the no priority scheme giving the same results as for minors priority (i.e. a mean response time of 1.42 hrs) and the majors priority scheme giving only slightly higher results (i.e. mean response time of 1.46 hrs). When we get to the higher workloads (80%) we start to see the walk-in arrivals having the best mean response time under minors priority; however, majors priority still provides the lowest mean response times for ambulance arrivals. As the system reaches saturation (90% to 100%) we can see that the minors priority scheme gives the best mean response times for all the arrival types. This is particularly interesting in light of the stringent treatment-time targets introduced by the UK Government for A&E patients. These require 98% of all A&E patients to be seen, treated and discharged in under 4 hours, the practical effect of which is to encourage A&E departments to prioritise the treatment of minors patients.

5. CONCLUSION AND FUTURE WORK

We have introduced an efficient and novel approximate generating function analysis (AGFA) technique and have compared its results against those from a discrete-event simulation. We have shown that the technique works well for mean response times under a number of different priority schemes although discrepancies were noted when the system modelled starts to become saturated under high workloads. The corresponding standard deviations – equivalent to second moments – show generally adequate agreement but (not atypically) are less accurate. This is because higher moments lack the linearity properties of first moments (means) and so greater care and precision is required in their analy-

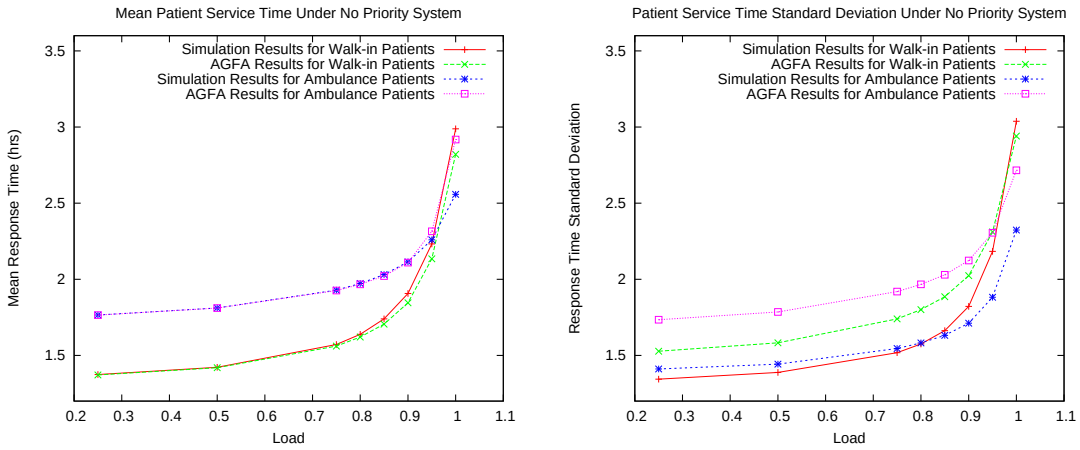


Figure 3: Simulated and AGFA mean (left) and standard deviation (right) of the service time for walk-in and ambulance arrivals for differing workloads under the no priority system

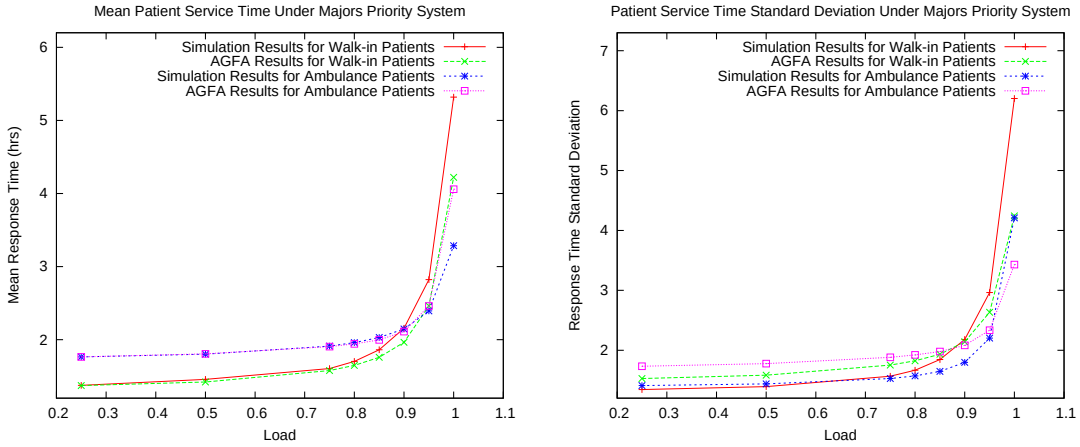


Figure 4: Simulated and AGFA mean (left) and standard deviation (right) of the service time for walk-in and ambulance arrivals for differing workloads under the majors priority system.

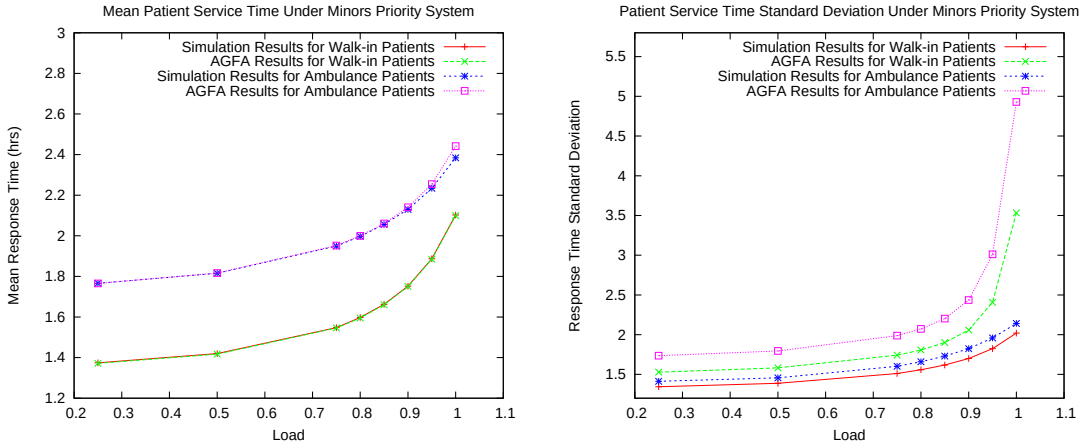


Figure 5: Simulated and AGFA mean (left) and standard deviation (right) of the service time for walk-in and ambulance arrivals for differing workloads under the minors priority system.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	1.37	1.34	1.37	1.53	1.77	1.41	1.76	1.73	2.08	2.04	2.08	2.04
Majors Priority	1.38	1.34	1.37	1.53	1.77	1.41	1.76	1.73	2.08	2.04	2.08	2.04
Minors Priority	1.37	1.34	1.37	1.53	1.77	1.41	1.77	1.74	2.08	2.03	2.08	2.04

Table 1: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 25% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	1.42	1.39	1.42	1.58	1.81	1.44	1.81	1.79	2.08	2.04	2.08	2.04
Majors Priority	1.46	1.39	1.42	1.58	1.80	1.44	1.81	1.78	2.08	2.04	2.08	2.04
Minors Priority	1.42	1.39	1.42	1.58	1.82	1.46	1.82	1.79	2.08	2.04	2.08	2.04

Table 2: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 50% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

sis. Although the AGFA method provides this well in open queues, its approximation becomes worse when it is applied in closed systems with constrained class populations at individual nodes.

As future work we intend to further refine the AGFA method in order to get better agreement at greater loads and higher moments and also to adapt the technique to incorporate more complex queueing disciplines such as time-based queueing priorities (i.e. queues with ageing). There is also great potential in using the AGFA technique in optimising more complex queueing network models where the mean and standard deviations of customer response time is optimised by finding the optimal resource allocations; a prime example would be to find the optimal staff and resource mix in order to minimise the mean and standard deviation of patient treatment times in the case study A&E model.

6. ACKNOWLEDGEMENTS

We are grateful for the help and advice of many members of staff at our case study hospital and associated institutions, including John Knottenbelt, Rick Juniper, Raj Singh, Sunil Johal, Sharon Ahearn and Ken Walton. We would also like to thank Tony Field for the use of his JINQS Java queueing network simulation library.

7. REFERENCES

[1] I.F. Akyildiz. Mean Value Analysis for blocking queueing networks. *IEEE Transactions on Software Engineering*, 14(4):418–428, April 1988.

[2] S.W.M. Au-Yeung, P.G. Harrison, and W.J. Knottenbelt. A Queueing Network Model of Patient Flow in an Accident and Emergency Department. In *Proc. 20th Annual European and Simulation Modelling Conference*, pages 60–67, Toulouse, France, October 2006.

[3] J.T. Bradley, N.J. Dingle, W.J. Knottenbelt, and P.G. Harrison. Performance queries on semi-Markov stochastic Petri nets with an extended Continuous Stochastic Logic. In *Proc. Petri Nets and Performance Models (PNPM'03)*, pages 62–71, 2003.

[4] T.J. Chaussalet, H. Xie, and P. Millard. A closed queueing network approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*, 45(5):492 – 497, 2006.

[5] T.J. Coats and S. Michalis. Mathematical modelling of patient flow through an Accident and Emergency department. *Emergency Medicine Journal*, 18:190–192, 2001.

[6] N.J. Dingle, P.G. Harrison, and W.J. Knottenbelt. Response time densities in Generalised Stochastic Petri Net models. In *Proc. 3rd Int. Workshop on Software and Performance (WOSP 2002)*, pages 46–54, 2002.

[7] G. Franks and C.M. Woodside. Multiclass multi-servers with deferred operations in layered queueing networks, with software system applications. In *Proc. 12th IEEE/ACM Intl. Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems (MASCOTS 2004)*, Volendam, The Netherlands, October 2004.

[8] N. Gulpinar, P.G. Harrison, B. Rustem, and L.-F. Pau. Optimization of a tandem M/GI/1 router network with batch arrivals. In *Proc. Workshop on Performance Modelling, Evaluation and Optimization of Parallel and Distributed Systems (PMEO-PDS 2005)*, 2005.

[9] P.G. Harrison. On teaching M/G/1 theory with extension to priority queues. *IEE Proceedings - Computers and Digital Techniques.*, 147(1):23–26, January 2000.

[10] P.G. Harrison and W.J. Knottenbelt. Passage time distributions in large Markov chains. In *Proc. ACM SIGMETRICS 2002*, pages 77–85, Marina Del Rey, California, June 2002.

[11] P.G. Harrison and N.M. Patel. *Performance Modelling of Communication Networks and Computer Architectures*. International Computer Science Series.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	1.57	1.52	1.56	1.74	1.93	1.55	1.93	1.92	2.09	2.04	2.09	2.04
Majors Priority	1.61	1.57	1.58	1.75	1.91	1.53	1.91	1.88	2.08	2.04	2.09	2.04
Minors Priority	1.55	1.51	1.55	1.74	1.95	1.60	1.95	1.99	2.08	2.04	2.09	2.04

Table 3: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 75% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	1.64	1.58	1.62	1.80	1.97	1.58	1.96	1.97	2.09	2.04	2.09	2.04
Majors Priority	1.70	1.67	1.65	1.82	1.96	1.57	1.94	1.92	2.09	2.04	2.09	2.04
Minors Priority	1.60	1.56	1.60	1.81	2.00	1.66	2.00	2.07	2.09	2.04	2.09	2.04

Table 4: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 80% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

Addison Wesley, 1993.

[12] N. Koizumi, E. Kuno, and T.E. Smith. Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, 8(1):49 – 60, February 2005.

[13] L. Mayhew and E. Carney-Jones. Evaluating a new approach for improving care in an Accident and Emergency department: The NU-care project. Technical report, Cass Business School, City University, 2003.

[14] L. Mayhew and D. Smith. Using queuing theory to analyse completion times in Accident and Emergency times in the light of the government 4-hour target. Technical report, Cass Business School, City University, 2006. Actuarial Research Paper No. 177.

[15] NHS Information Centre. Ambulance services, England: 2005-06. Online report, October 2006. Available at `http://www.ic.nhs.uk/pubs/ambulanceserv06`.

[16] K.R. Pattipati, M.M. Kostreva, and J.L. Teele. Approximate Mean Value Analysis algorithms for queueing networks: existence, uniqueness and convergence results. *Journal of the ACM*, 37(3):643–673, July 1990.

[17] D.C. Petrui and C.M. Woodside. Approximate Mean Value Analysis based on Markov chain aggregation by composition. *Linear Algebra and its Applications*, 386:335–358, July 2004.

[18] J. Preater. Queues in health. Keele mathematics research report, University of Keele, Mathematics Department, 2001.

[19] M. Reiser. Mean Value Analysis: A personal account. In *Performance Evaluation: Origins and Directions; Lecture Notes in Computer Science 1769/2000*, pages 491–504, 2000.

[20] M. Reiser and S. Lavenberg. Mean-Value Analysis of closed multiclass queuing networks. *Journal of the ACM*, 27(2):313–322, April 1980.

[21] H. Wang and K.C. Sevcik. Experiments with improved approximate Mean Value Analysis algorithms. *Performance Evaluation*, 39(1–4):189–206, 2000.

[22] E.M. Winands, I.J. Adan, and G.J. van Houtum. Mean Value Analysis for polling systems. *Queueing Systems: Theory and Applications*, 54(1):35–44, September 2006.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	1.74	1.66	1.71	2.03	1.63	2.02	2.03	2.09	2.04	2.09	2.05	2.05
Majors Priority	1.86	1.84	1.76	1.93	2.03	2.0	1.98	2.09	2.04	2.09	2.05	2.05
Minors Priority	1.66	1.62	1.66	1.90	2.06	1.73	2.06	2.02	2.09	2.04	2.09	2.05

Table 5: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 85% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	1.91	1.82	1.75	2.06	2.11	1.71	2.14	2.44	2.09	2.05	2.09	2.04
Majors Priority	2.16	2.18	1.96	2.14	2.15	1.80	2.11	2.08	2.09	2.05	2.09	2.04
Minors Priority	1.75	1.70	1.75	2.06	2.13	1.82	2.14	2.44	2.09	2.04	2.09	2.04

Table 6: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 90% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	2.23	2.18	2.13	2.31	2.26	1.88	2.31	2.31	2.09	2.04	2.09	2.05
Majors Priority	2.82	2.96	2.46	2.63	2.39	2.20	2.47	2.33	2.09	2.05	2.09	2.05
Minors Priority	1.89	1.82	1.88	2.41	2.23	1.96	2.25	3.01	2.09	2.04	2.09	2.05

Table 7: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 95% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.

	Walk-In arrivals				Ambulance arrivals				Blue call arrivals			
	Sim.		AGFA		Sim.		AGFA		Sim.		AGFA	
	Mean	S.D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.	Mean	S. D.
No Priority	2.99	3.04	2.82	2.94	2.56	2.32	2.92	2.72	2.09	2.05	2.09	2.05
Majors Priority	5.32	6.20	4.22	4.24	3.29	4.21	4.06	3.43	2.09	2.04	2.09	2.05
Minors Priority	2.10	2.02	2.10	3.53	2.38	2.14	2.44	4.93	2.09	2.04	2.09	2.05

Table 8: Mean and standard deviation (S. D.) of response times for different classes of arriving patient under two different priority schemes, with 100% ambulance and walk-in arrival rates, calculated by simulation and our AGFA technique.