Adaptive Planning of Staffing Levels in Health Care Organisations

Harini Kulatunga[1], W.J. Knottenbelt[1], and V. Kadirkamanathan[2]

[1] Department of Computing, Imperial College London,
180 Queen's Gate, London SW7 2AZ, United Kingdom
{hkulatun,wjk}@doc.ic.ac.uk

[2] Department of Automatic Control and Systems Engineering,
University of Sheffield,
Mappin Street, Sheffield S1 3JD, United Kingdom
visakan@sheffield.ac.uk

Abstract. This paper presents a new technique to adaptively measure the current performance levels of a health system and based on these decide on optimal resource allocation strategies. Here we address the specific problem of staff scheduling in real-time in order to improve patient satisfaction by dynamically predicting and controlling waiting times by adjusting staffing levels. We consider the cost of operation (which comprises staff cost and penalties for patients waiting in the system) and aim to simultaneously minimise the accumulated cost over a finite time period. A considerable body of research has shown the usefulness of queueing theory in modelling processes and resources in real-world health care situations. This paper will develop a simple queueing model of patients arriving at an Accident and Emergency unit and show how this technique provides a dynamic staff scheduling strategy that optimises the cost of operating the facility.

Keywords: Adaptive staff scheduling, staffing cost minimisation, integrated health care systems.

1 Introduction

Health care systems are gradually evolving from disparate general practices and hospitals to integrated care delivery systems [1]. In this context developing system-wide integration of administration, clinical care, information technology (IT), and financing is the ultimate goal. It has been found that highly centralized networks had better financial performance than did those in more decentralized networks [2]. One of the important operational issues in health care involves resource planning such that the goals of high resource utilisation, meeting patient response times and minimising cost are met [3]. A general modelling and solution methodology found in the existing literature is a *steady-state* queueing network model and an optimisation framework to guide resource planning decisions [5], [4]. The authors believe that most methods are more appropriate either when

P. Kostkova (Ed.): eHealth 2009, LNICST 27, pp. 88–95, 2010.

patient arrival rates are expected to remain relatively constant or when all possible uncertainties in patient arrivals are known in advance so that an allocation policy for a fixed time period can be predetermined. However in an Integrated Health Care (IHC) system such fixed optimisation strategies are not the most efficient since total knowledge of all possible scenarios will be impossible. Such a policy for all times can lead to insufficient/wasted resources over some time periods and ultimately increase in cost.

For management of a IHC system, queueing models can provide an analysis capability which can improve the timeliness of interventions and be helpful in the process of external scrutiny (e.g. by the Health Care Commission). Furthermore they can be a starting point for verifying the credibility of reported performance by different IT system solutions. This paper presents a solution methodology based on *transient* queueing models as opposed to steady-state models which can only provide information about the real-world system under the assumption that the system is stationary. On the other hand a transient queueing model is a more robust approach in providing approximate real-time information about a health care system. In this paper the queueing model is combined with a dynamic optimisation technique which determines a resource allocation policy based on instantaneous system performance measurements and guarantees a minimised accumulated cost over a finite time period. To be specific, the model of the health care system *adaptively* determines the optimal staffing level (i.e. number of clinicians) required to achieve a target level of customer service (i.e. target waiting times) and minimises the staffing related cost of operating the facility. In essence this is an attempt to answer the following question,

> At a given time which resource allocation strategy gives the optimal cost?

and provide a financial planning model for IHC administrators and financial managers to study and evaluate the economic impact of changes in a organization's resources at a given time.

The rest of the paper is organised as follows. The first section explains the general staffing model considered in this paper. Next section which describes the solution method for the model, presents the relation between accumulated cost, staffing level and waiting times of patients. The section also describes the adaptive technique to determine the optimal staffing policy over a finite time duration. This is followed by a section on how this general model relates to the specific problem of an Accident and Emergency (A&E) department. Finally simulation results are generated based on actual input data gathered from an NHS centre in London. This paper concludes with some remarks on future directions.

2 Staffing Model

Planning staffing levels requires balancing two conflicting goals: minimising cost and maintaining high service levels. Here the service level is defined by a waiting time target. Increasing staff will reduce patient waiting times but will result in high resource costs. A health care system will exhibit stochastic behaviour such

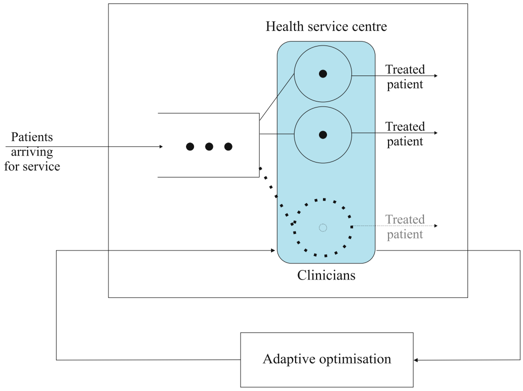

Fig. 1. Adaptive Staffing Model

as the uncertainty in how long it takes to treat individual patients and arrival rates of patients to different departments within a hospital. To account for this randomness in behaviour the health system is modelled using queueing theory principles (Fig. 1). The system is modelled as a network of $c(n)$ service stations or number of clinicians at n-th time instant. Patients enter the system and either wait in a queue or go for treatment to one of the service stations. This paper considers a First-Come-First-Served (FCFS) discipline and no priority levels between customers even though more complex definitions are possible (and should be incorporated to reflect phenomena found in real systems). The staff planning involves determining how many clinicians are needed to provide a target level of service at minimum cost. The stochastic nature of arrivals and service times are represented by probability distributions. It is assumed that the arrivals are random and Poisson distributed and the service rate is denoted as $\mu(n) = \mu \cdot c(n)$. μ is the rate of service and $c(n)$ is the number of servers and the staff allocation amounts to deciding the level $c(n)$. A maximum limit to the number of servers that can be added is set to K. A $GI/G/c(n)$ queueing model is used here to provide a good approximation of time-varying patient arrival rates and general service time distribution. The results section will consider patient arrival intensities to follow a non-homogeneous Poisson distribution, that is a Poisson process with a time-varying rate parameter to indicate a sudden rush of patients. We will then show that this approach will adaptively allocate staff levels to deal with these dynamics and ensure that patient waiting times do not increase and miss targets.

2.1 Transient Approximations for GI/G/c(n) Model

Based on Pollazek-Khintchine formulas and 'transient Little's Laws' [6] the distribution of the patient delay $W(n)$ in the system when he or she arrives at time n is given by,

$$F_n(x) = G \star [1 - \rho(n)] \sum_{i=0}^{\infty} \rho(n)^n H_e^{i\star}(x) \tag{1}$$

where $\star$ represents the convolution operation. The system performance parameters at n-th time are calculated as follows,

$$\begin{aligned}
\text{Staff utilisation level: } \rho(n)^{-1} &= 1 + \frac{\mu_e}{\mu_1 p(n) + \mu(Q(n) - B(n))} \\
\text{Patient waiting time in system: } E[W(n)] &= m + \mu_1 p(n) + \mu[Q(n) - B(n)] \\
\text{Prob. the patient has to wait in queue: } p(n) &= \alpha(B(n)) \\
\alpha(B(n) &= \frac{P(c; B(n)) - P(c-1; B(n))}{P(c; B(n)) - (B(n)/c)P(c-1; B(n))} \\
B(n) &= \max[0, \mu c[\lambda(n) - Q'(n)]] \\
Q'(n) &= (Q(n) - Q(n-1))/\delta
\end{aligned}$$

where the additional parameter $B(n)$ is the number of clinicians seeing patients at a given time and $P(c; B) = \sum_{i=0}^{c} e^{-B} B^i / i!$ [7]. The number of patients waiting for treatment is $Q(n)$ and δ is the duration of the n-th time interval. The mean and equilibrium mean of the service time distribution is given by m and m_e respectively. The means μ and μ_1 are the means of service time and residual service time distributions when $Q(n) \geq c$ at the time of a new arrival,

$$\begin{aligned}
H(x) &= 1 - \left[\bar{G}_e(x)\right]^{c-1} \bar{G}(x) \\
H_1(x) &= 1 - \left[\bar{G}_e(x)\right]^{c}
\end{aligned} \tag{2}$$

and $H_e(x)$ is the equilibrium distribution associated with $H(x)$ and its mean is denoted μ_e.

3 Adaptive Staff Planning Approach

Lets formalise the staff allocation problem now. Under a specific policy π a sequence of decisions are defined which include at a specific time t a decision on the number of staff $1 \leq c(t) \leq K$ to be allocated. In order to achieve both a target waiting time and minimise the accumulated cost of staff resources, penalties on missing targets and cost of patient queueing there is a need to add/remove part of the staff depending on the workload and queue length. This is investigated here in terms of a queueing model with many service stations. Given that $W_{\max}$ is the target maximum waiting time and C_r - the individual staff cost per unit time, C_p - the penalty cost per unit time, C_s - cost associated with the queue length per unit time, the objective function is given as,

$$V_\pi = E_\pi \left[\int_{t=0}^{\infty} \left(C_r c(t) + C_p \max(E[W(t)] - W_{\max}, 0) + C_s Q(t) \right) \, dt \right] \tag{3}$$

The above optimisation problem can be solved as a finite horizon dynamic programming problem. The goal of this paper is to determine self-adaptive policies in order to minimise the accumulated cost of operation. The cost criterion is minimised based on the mean waiting time, queue length and number of staff. In this framework staff allocation is altered at the end of discrete-time equidistant time intervals $\delta = 1/\gamma$ with $\gamma = \max(\lambda) + K\mu$ where λ is the patient arrival rate. The cost at epoch n is given by,

$$C(i, c(n)) = (C_r c(n) + C_p \max(E[W(n)] - W_{\max}, 0) + C_s i)/\gamma \tag{4}$$

when $Q(n) = i$. The goal of this problem is to recursively find the decision vector $\mathbf{c} = (c(1), \ldots, c(N))$ which finds the minimum cost path using dynamic programming as follows,

$$V(Q(n) = i) = \max_{\mathbf{c}(n) \in (1,\ldots,K)} \left[C(i, c(n)) + \sum_{\forall j} p_{ij}^{c(n)} V(Q(n+1) = j) \right] \tag{5}$$

The complexity of this type of solution can be very high therefore we use a technique called neuro-dynamic programming [8] to derive a near-optimal solution. Specifically we consider a one-step look ahead approach to approximate the cost-to-go function such that,

$$V_{n+1}(Q(n+1), c) = \sum_{\forall j} p_{ij}^{c(n)} (C_r c + C_p \max(E[W(n+1)] - W_{\max}, 0) \\ + C_s j)/\gamma \tag{6}$$

for $c \in \{1, \ldots, K\}$. The sojourn time for an arrival at the next epoch $E[W(n+1)]$ is calculated using matrix analytic methods when the number of servers is equal to c. Then the resource allocation decision for epoch $(n+1)$ is given by,

$$c(n+1) = \min_{c \in \{1,\ldots,K\}} \{C[Q(n), c(n)] + V_{n+1}(Q(n+1), c)\} \tag{7}$$

4 Staff Planning in an A&E Unit

A&E departments are being placed under increasing pressure to process a growing number of patients safely and quickly. This is evidenced by the national government target whereby 98% of patients must spend 4 hours or less from arrival to admission, transfer or discharge, as well as an increase in the number of attendances to A&E departments and walk-in centres in England. Concurrently, in $2003-6$, seven hospital trusts reported one or more A&E departments closed or downgraded, with one new A&E department opening. For the remaining open A&E departments it is important to be able to predict the changes in patient arrivals and take optimal decisions that meet cost and service level targets. There are many studies describing simulations of A&E departments in which either a Poisson arrivals process is assumed or historical attendance is replicated. There has been some success in application of queueing models to A&E departments in particular the studies by [9,10,11] which will have limited success in complex

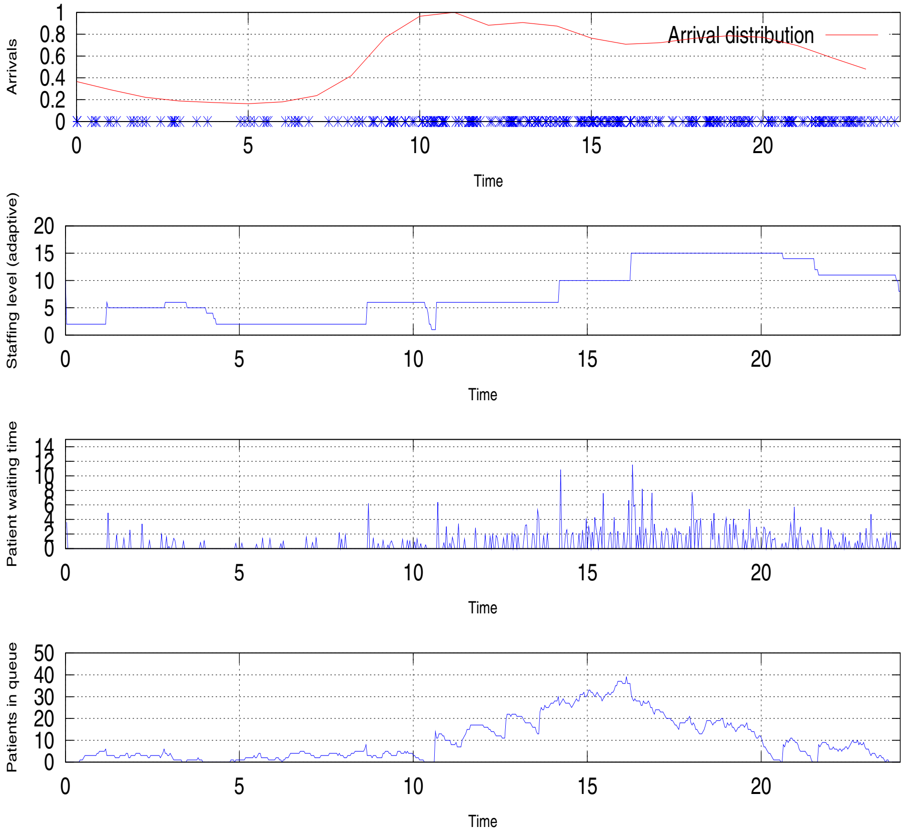

Fig. 2. Adaptive staff management

systems. Firstly, the high-level queueing models used typically do not tie system performance to the underlying resources. Secondly, many of these models do not take into account phenomena that occur in the corresponding real life systems such as time-varying arrivals. This paper goes beyond these models and presents two new aspects for the first time. Firstly, the problem formulation jointly finds the optimal set of resources that achieve waiting time targets and minimise cost. Secondly, a transient queueing model and an adaptive performance optimisation technique is developed that allocates resources on-line. Simulation results are presented that shows the adaptive staff planning capabilities of the technique described. A discrete event simulation (DES) of a multi-server queueing system (Fig. 1) is built to model the complete emergency service centre. Here each server represents one staff member. Actual arrivals data from an A&E unit in London obtained over a one year period is used as input to the DES. The arrival rates are computed by averaging over all the days in the year and the arrival rate distribution is given in the first sub-plot in Fig. 2. The service rate of each staff member is assumed to be 1.379 (patients/hour) and the cost parameters are

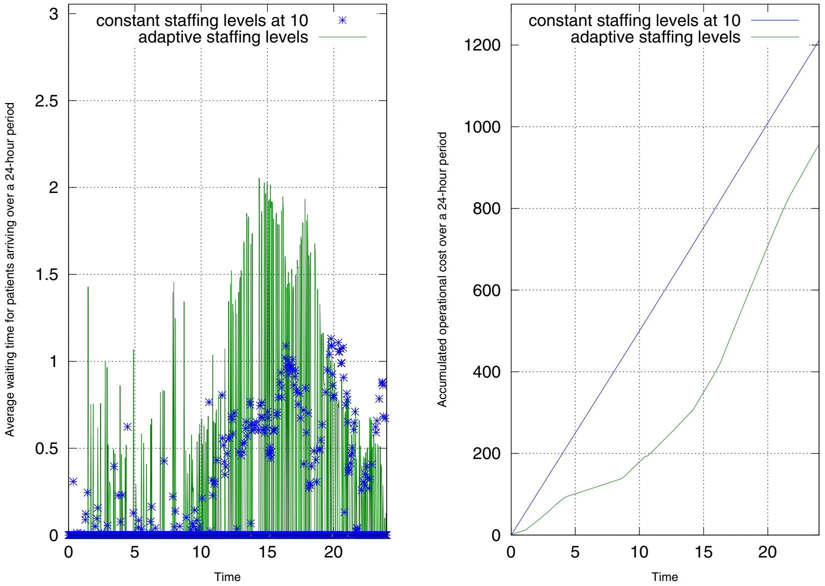

Fig. 3. Accumulated cost

assumed to be, (1) cost of patients queueing for treatment is 3 cost units per hour (2) penalty of missing the 4 hour waiting time target is 2 cost units per hour and (3) resource cost is 5 cost units per hour. More realistic parameters can be used and more complex models can be built when these techniques are applied to an actual system. Results are obtained under two conditions over a 24 hour period from $0000 - 2400$ hours. That is once when the staffing level is changed dynamically and the other when the staffing level is maintained at a constant level of 10 clinicians during the day. The adaptive staffing policy takes values from the set $\{1, 2, 3, 4, 5, 6, 10, 15, 20\}$ such that accumulated cost of operation over the day is minimised. The results of the adaptive staff allocation techniques is given in Fig. 2 where measures from the system are plotted at time slots of 2 minutes. Additionally it is assumed here that when a staff reduction is recommended by the optimisation technique the DES takes the maximum between the suggested level and currently busy number of staff. This is to represent that treatment in progress is not interrupted. Sub-plots $2 - 3$ shows that when ever the overall waiting time during a time slot increases sharply the staff allocation policy recommends an increase of staff leading to a subsequent drop in the overall waiting time over the next time slot. Sub-plot 4 gives the number of patients waiting in a queue at each time slot during the day. The targeted performance results of using this techniques is given in Fig. 3 and shows the performance management capabilities that can be obtained by adaptively controlling the staffing level. The first sub-plot indicates the average waiting time a patient arriving at the emergency unit would face at any given time during the day. The second sub-plot gives the overall accumulated cost of operating the facility for the day based on the cost parameters given earlier. This highlights the benefits of reduced overall cost when adapting the staffing level dependent upon the current workload and waiting times. This in effect also reduces the under utilisation of expensive staff resources when the emergency unit is not busy.

5 Conclusions

This paper presented a new technique to adaptively allocate staff in a IHC system in order to meet quality of service targets and minimise cost. The results obtained show the benefit to management and administration of a health centre to evaluate the effects of staffing policies on performance and cost. Furthermore the unique adaptive nature of this technique means that decisions can be taken in real-time in response to changes in workload.

References

1. Darzi, A.W.: Ideas from Darzi:polyclinics. NHS Confederation Publications (2008)
2. Bazzoli, B.J., Chan, B., Shortell, S., D' Aunno, T.: The financial performance of hospitals belonging to health networks and systems. Inquiry 37(3), 234–252 (2000)
3. Smith-Daniels, V.L., Schweikhart, S.B., Smith-Daniels, D.E.: Capacity management in health care services: Review and future research directions. Decision Sciences 19, 889–918 (1988)
4. Brailsford, S.C., Lattimer, V.A., Tamaras, P., Turnbull, J.C.: Emergency and on demand health care: modelling a large and complex system. Journal of the Operational Research Society 55, 34–42 (2004)
5. Gorunescu, F., McClean, S.I., Millard, P.H.: A queueing model for bed occupancy management and planning of hospital. Journal of the Operational Research Society 53, 19–24 (2002)
6. Riaño, G.: Transient behaviour of stochastic networks: Application to production planning with load dependent lead times, Ph.D. Thesis, Georgia Institute of Technology (2002)
7. Grassmann, W.K.: Finding the right number of servers in real-world queueing systems. Interfaces 2, 94–104 (1988)
8. Bertsekas, D.P., Tsitsiklis, J.: Neuro-dynamic programming. Athena scientific, Belmont (1996)
9. Coats, T.J., Michalis, S.: Mathematical modelling of patient flow through an Accident and Emergency department. Emergency Medicine Journal 18, 190–192 (2001)
10. Mayhew, L., Carney-Jones, E.: 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)
11. Mayhew, L., Smith, D.: Using queueing 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, Actuarial Research Paper No. 177 (2006)