QUEUING THEORY APPLIED IN HEALTHCARE

This report surveys the contributions and applications of queuing theory applications in the field of healthcare. The report summarizes a range of queuing theory results in the following areas: Waiting time and utilization analysis, system design and appointment systems. The goal is to study the information used by analysts to model the healthcare process on queuing theory model.

Shreyas [04020518]

11/15/2007

QUEUING THEORY APPLIED IN HEALTHCARE

Introduction

Organizations that provide Health care processes can be viewed as Queuing systems in which the patients arrive , wait for the service , obtain service and then depart. The healthcare processes , varying in complexity and scope , consist of a set of activities and procedures (both medical and non- medical) that the patient must undergo in order to receive the required treatment. The resources (or servers) in these queuing systems are the trained personnel and specialized equipment that these activities and procedures require.

Applied Queuing Theory *(a summary of printed papers)*

- **McClain (1976) :** Impact of bed assignment policies on utilization , waiting time and probability of turning away patients.
- **Noseck & Wilson (2001) :** Improving customer satisfaction by predicting and reducing waiting times and adjusting staffing.
- **Green (2006a) :** relationship amongst delays , utilization and no. of servers. : the basic M/M/s model applied in healthcare.

WAITING TIME & UTILIZATION ANALYSIS

- **Broyles & Cochran (2007) : [RENEGING]** calculates the percentage of patients who leave an emergency department without getting help using arrival rate , service rate , utilization capacity, hence determining the resulting revenue loss
- **Worthington (1991) : [VARIABLE ARRIVAL RATE]** presents an M(λq)/G/S model for service times of any fixed probability distribution and for arrival rates that decreases linearly with queue length and expected waiting time.
- **McQuarrie (1983) : [PRIORITY QUEUING DISCIPLINE]** shows that when utilization is high , it is possible to minimize the waiting time by giving priority to clients with shorter service time

SYSTEM DESIGN

- **Bailey (1954) :** establishes the existence of threshold capacity (when service=demand)and hence argues that the system be designed where service exceeds expected demand by a factor of 1 or 2.
- **Bruin (2005) :** [**BLOCKING]** no. of beds reqd. to achieve a maximum turn away rate of 5% at an emergency cardiac department.
- **Keller & Laughhunn (1973) : [COST MINIMIZATION]** determining the minimum cost of serving patients at Duke University Medical Center.
- **Young [1962a,b] : [COST MINIMISATION]** proposes an incremental analysis approach in which the cost of an additional bed is compared with benefits it generates , until they are equal.

APPOINTMENT SYSTEMS

[systems with appointments reduce variability and waiting times at the facility.]

- **Bailey (1952,1954)** proposes (a) appointment interval , (b) consultant arrival time as two variables that determine the efficiency of an appointment system. The ratio of total time wasted by all patients to the consultant's idle time should equal consultant's time relative to patients'.
- **Bottlenecks :** Nodes at which services are dispensed. Ratio of demand to available service is very high.

MODELLING A HEALTHCARE SYSTEM AS A QUEUEING NETWORK THE CASE OF A BELGIAN HOSPITAL

[Stefan Creemers & Marc R. Lambrecht]

Abstract

The performance of healthcare systems in terms of patient flow times and utilization of critical resources can be assessed through queuing simulation models. Modeling is focused on impact of outages (preemptive and non preemptive outages) on the effective utilization of resources and on the flow time of patients. Queueing network solutions like Decomposition and Brownian motion approaches are developed.

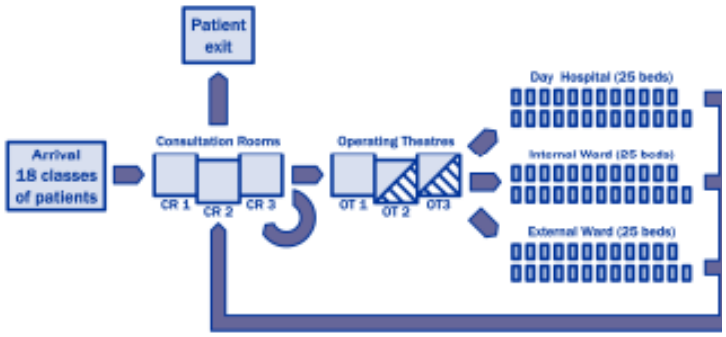
Modeling the healthcare system as a queuing model

Two approaches:

- Parametric Decomposition ; using the Kingman Equation and the approximation derived by Whitt (1993) to assess performance.
- A Brownian queueing model (Harrison 1998)

PARAMETRIC DECOMPOSITION

Fig. gives the queueing network representation of the orthopaedic department. The queueing model represents an open re-entry network that consists of five G/G/m workcentres (consultation , surgery and three wards representing the locations at which the recovery takes place. It is modeled on the decomposition technique pioneered by Jackson (1957). Queue discipline is FCFS and any variation in the arrival of patients is presumed to be absorbed in the variance of the arrival process.



The paper infers

$$E\begin{bmatrix}1\\ \omega\end{bmatrix} = \begin{matrix}1 & (\tau i + \tau r)\\ \nu 1 & \tau r\end{matrix}$$

where $\frac{1}{\omega}$ *is the average service time* $\frac{1}{\nu 1}$ *is the natural service time* (τr) *is the resolve times of interrupts during service ·*

Later on the paper also observes that the probability of encountering and interrupt increases as the service time increases , and hence exists a positive correlation between service times and repair times induced by interrupts.

Combining the preemptive and non preemptive outages, further simplification yields

$\frac{1}{\psi} = \frac{1}{\omega} + \frac{1}{n\mu s}$, where $\frac{1}{\psi}$ is the effective service time,

while it service time experienced by patient (and as such includes the impact of outages).

The variance of the effective service times at the consultation workstation may be approximated as :

$$\sigma_{\psi}^2 = \sigma_{\omega}^2 + \frac{\sigma_s^2}{n} + \frac{1}{\mu_s^2}\left(\frac{n-1}{n^2}\right)$$

SQUARED COEFFICIENT OF VARIATION OF AGGREGATE ARRIVAL PROCESS

Albin (1984) pointed out that if at least one of the inter arrival time distributions is not a Poisson process , the resulting aggregate interarrival times do no longer hold the property of ***independence***. So approximations were adopted. And a system of ***linear equations*** [Shanthikumar (1981)] was used to solve for the unknown squared coefficients of variation.

Flow time expressions

The total waiting times (or flow times) incorporate both waiting time in the queue as well as actual processing , and hence w.r.t to the Kingman equation , one can define the expected flow time of a patient as the workstation i as follows:

Kingman Equation

$$E\left[W_{Kingman}\right]=\left(\frac{C_{a_i}^2+C_{s_i}^2}{2}\right)\left(\frac{\rho_i^{\sqrt{2\left(m_i+1\right)}-1}}{m_i\left(1-\rho_i\right)}\right)\left(\frac{1}{\mu_i}\right)+\frac{1}{\mu_i} \quad \text{and}$$

Whitt equation

$$E\left[W_{whitt}\right]=\gamma_i\frac{C_{a_i}^2+C_{s_i}^2}{2}E\left[W_{M|M|m_s}\right]+\frac{1}{\mu_i}$$

allows subsequent approximation of $G/G/m_i$

BROWNIAN QUEUEING MODEL

(alternate queueing modeling technique)

Brownian approach is recent , and deeply rooted in the heavy traffic theory and they hold the advantage that they study queueing theory as a whole (i.e. Brownian models do not use decomposition approach). It has been shown that queueing processes (workloads , number in queue , waiting time..) often have semi martingale Brownian Motion (SRBM) as a limiting Assuming SRBM , the parameters defined are :

- A drift vector θ
- A covariance matrix Γ
- A reflection matrix R_f

The dimensionality of Brownian motion is the no. of workstations on network (denoted by *i)*

The most notable difference between Brownian model and Decomposition approach is the Routing mechanism.

In Brownian Queueing model author has assumed existence of 6 treatment processes of a patient.

- Consultation phase prior to surgery
- Surgery
- Recovery (division day hospital)
- Recovery (division internal ward)
- Recovery (division external ward)
- Consultation phase after surgery.

Hence in Brownian model we make a distinction between consultation phase prior to surgery and after surgery, whereas in decomposition approach we consider only a single consultation phase in which initial and followup consultations are combined. So let c (c Є {1,…,C}) be the no. of patient class.

For analysis , the arrival process is split into 2 separate streams of initial and followup consultations. For class *k* patients ,

$$\therefore\ \gamma_k=\gamma_{\alpha_k}+\gamma_{\beta_k}$$

Hence, $\lambda_{b_1}=\sum_{k=1}^{K}\eta_k\gamma_{\alpha_k}$ and *arrival rate at stage 6 is* $\lambda_{b_6}=\sum_{k=1}^{K}\eta_k\gamma_{\beta_k}$

Then applying ***law of conservation of flows*** , no. of arriving patients is equal to no. of patients leaving at each time unit at each stage

$r_{b_{11}} = 1 - \frac{\eta}{\lambda_{b_1}}$ *and* $r_{b_{12}} = \frac{\eta}{\lambda_{b_1}}$ where η is the average no. of patients leaving each phase.

$\therefore r_{b_{66}} = 1 - \frac{\eta}{\lambda_{b_6}}$

Then theory of SRBM is applied and computations are done thro' *QNET software* , the **Stationary Means of** SRBM (z_i) are calculated.

Therefore , the no. of patients present at workstation *i* (in queue and process) is:

$$\bar{Q}_i = z_i \mu_i ,$$

and then

using Little's formula $E\left[W_{brownian}\right] = \frac{\bar{Q}_i}{\lambda_i}$

The paper further explores the validation of the model generated, thro' **Simulation Analysis**. The authors also define

PERFORMANCE INDICATORS:

- Patient total expected waiting time (i.e. the expected flow time of an average patient).
- Ratio of time spent on absences (κ_s)
- Ratio of time spent on resolving interrupts (κ_s)
- The effective utilization rate at the consultation workstation. (ρ_1)

Also , average effective service time at the consultation workstation can be divided into three components:

- Natural service time [$\frac{1}{\nu_1}$]
- Outage time due to unscheduled absences [$\frac{1}{\mu_s}$]
- Outage time due to service interruptions. [$\frac{1}{\mu_g}$]

Therefore,

Average effective service time $\left(\frac{1}{\varphi_s}\right) = \frac{1}{\nu_1} + \frac{1}{\mu_s} + \frac{1}{\mu_g}$

Ratio of time spent on absences $\kappa_s = \frac{\nu_1}{n\mu_s}$

Ratio of time spent on resolving interrupts $\kappa_f = \frac{\tau_r}{\tau_i - \tau_r}$

CONCLUSION

The contributions of this paper can be summarized as follows :

- Two different modeling approaches *viz.* Parametric Decomposition and Brownian Queueing models , were discussed and compared when applied to the same problem
- New expressions assessing the impact of service outages were developed.
- A comparison between various scenarios , evaluating the impact of service outages in heavily loaded systems , provides several managerial insights.