

東海大學

資訊工程系研究所

碩士論文

指導教授：江輔政博士

Advisor : Prof. Fuu-Cheng Jiang, Ph.D.

醫院急診部門醫療資源成本最佳化之佈署

設計與製作---採用排隊理論模型

The Optimization Deployment of Provider
Staffing in Hospital Emergency Department
using Queuing Theory

研究生：鍾佳儒

中華民國一零七年六月

東海大學

資訊工程系研究所

碩士論文

指導教授：江輔政博士

Advisor : Prof. Fuu-Cheng Jiang, Ph.D.

醫院急診部門醫療資源成本最佳化之佈署
設計與製作---採用排隊理論模型

The Optimization Deployment of Provider
Staffing in Hospital Emergency Department
using Queuing Theory

研究生：鍾佳儒

中華民國一零七年六月

東海大學碩士學位論文考試審定書

東海大學資訊工程學系 研究所

研究生 鍾佳儒 所提之論文

醫院急診部門醫療資源成本最佳化之佈署

設計與製作-採用排隊理論模型

經本委員會審查，符合碩士學位論文標準。

學位考試委員會

召集人

時文中 簽章

委員

陳金鈴

黃宜豐

指導教授

江輔政 簽章

中華民國 107 年 6 月 30 日

致謝

謹以此篇論文，作為兩年碩士生涯之記錄與紀念。此篇論文的完成，首先要感謝我的指導教授，江輔政教授不辭辛勞地悉心指導，以及口試委員：時文中老師、陳金鈴老師以及黃宜豐老師，使這篇論文更加嚴謹完備，同時也要感謝一路上給予我幫助及鼓勵的所有東海大學資工系師長們，以及提供良好求學環境的東海大學。

還要感謝實驗室夥伴們：陳永鑫學長以及黃緯滔學長給予我寶貴的經驗傳承，使我們的研究能夠延續並更加深入。縱使兩位已經畢業，也願意在百忙之中抽空給予我們協助，也使得研究生生活更加多采多姿。

最後要感謝我的家人：我的雙親鍾立聖先生以及陳美鈴女士，一路上對我的精神與心靈上的支持，讓我可以無後顧之憂地全心完成碩士學程。也要感謝我的姑姑鍾立雯女士、姑丈劉耀仁先生，這段時間給我的關心、建議以及經驗分享和未來方向的探討，使我能夠不再徬徨，並更全面的掌握自己擁有的選擇性。

網路管理實驗室 鍾佳儒 於 2018 年 6 月

摘要

運作一個有效率的醫院急診部門(HED, Hospital Emergency Department)，基本上須經由其成本支出改善及急診醫療資源(Provider Staffing)最佳化、動態配置等核心要素，給予系統模式來設計。隨著 HED 來客(病患)率動態變化，以排隊理論作為設計之核心重點，本研究提出一套系統最佳化規劃，提供 HED 管理部門作決策參考。

方法: 本研究論文針對 HED 醫療資源需求變異、動態佈署與 HED 醫療成本支出最佳化之間的互動，採用排隊理論中之 M/M/S/K 模型 (多重伺服器、病患有限等候容量的排隊模型)，作為設計之核心重點。原則上，醫院急診部門(HED)的來客(病患)率有其階段性或可預估之週期性，來客率多量或量少期間兼而有之。以排隊理論中之 M/M/S/K 模型為主，推導相關數理方程式，醫療成本支出最佳化為核心重點，來動態佈署其 HED 醫療資源。

結果: 本研究論文主要之創意，在於將醫院急診部門的醫療資源視同於 M/M/S/K 模型中的多重伺服器(Multi-server)，將急診部門有限等候空間視同於伺服器之 Queue Buffer。此種新的設計模式，以排隊理論為數理基礎，建構出醫療資源平台的後勤體系，同時推導相關數理方程式，建立系統必要性能函數，完成成本最佳化之計量研究；最後吾人也使用模擬工具驗證成本模式的可行性。

結論: 因此針對 HED 的醫療資源體系，以 HED 來客(病患)率有其階段性或可預估之週期性，有效率佈署寶貴的醫療資源，為經營成本最佳化為導向，本研究成果確實提供了一個有效可行的系統規劃，作為醫院急診部門管理部門作決策參考。

關鍵字：醫院急診部門，醫療資源，排隊理論模型，成本最佳化

Abstract

As hospital administrators evaluate potential approaches to improve cost, quality, and throughput efficiencies in the hospital emergency department (HED), the deployment or distribution of valuable medical resources emerges to be the dominate challenge to health policy makers. An improvement of emergency services is an important stage in the development of healthcare system. The research on optimal deployment of medical resources appears to be an important issue of HED long-tern management. The HED's performance in terms of patient-flow and of available resources can be studied using the queue-based approach. The kernel point of this research is to approach the optimal cost on distributing HED resources using queuing theory. To model the proposed approach for qualitative profile, a generic HED system is mapped into the M/M/S/K queue-based model for application. On quantitative work, a comprehensive mathematical analysis on cost pattern has been made in detail. Relevant simulations have also been conducted to validate the proposed optimization model. The design illustration is presented to demonstrate the application scenario in HED platform. Hence the proposed approach indeed provides a feasibly cost-oriented decision support framework to adapt HED management requirement.

Keywords: optimization, hospital emergence department, queue-based methodology.

Table of Contents

致謝.....	i
摘要.....	ii
Abstract.....	iii
List of Figures.....	v
List of Table.....	vi
Chapter 1 Introduction.....	1
Chapter 2 Related Work.....	4
Chapter 3 The Proposed Model of Medical Emergency Service.....	6
3.1 The generic platform of medical emergence service.....	6
3.2 Mapping profile between HED service platform and the M/M/S/K QS.....	11
Chapter 4 Quantitative Modeling and System Measures for the HED Platform.....	16
4.1 Theoretical Analysis.....	16
4.2 System Performance Measures.....	19
Chapter 5 Performance Evaluation.....	21
5.1 Introduction to MATLAB Simulation Tool.....	21
5.2 Issue on Decision Support for HED Management.....	22
5.3 Optimization Evaluation.....	24
5.4 Issues on Cost Profile under the Constraint of Average Waiting Time.....	27
Chapter 6 Conclusion.....	31
References.....	32

List of Figures

Fig.3.1The functional deployment on the ground floor of TVGH-HED building.....	7
Fig.3.2The functional deployment on the ground floor of CGMH-HED building.....	10
Fig.3.3The generic service platform of emergency department	12
Fig.3.4An M/M/S/K queue system mapped by the HED service platform	14
Fig.4.1State-transition-rate diagram for the proposed model	17
Fig.5.1 Optimal cost patterns shown in terms of three average arrival rates.....	25
Fig.5.2 An enlarged diagram showing optimal cost data from Fig. 5-1.	26
Fig.5.3Decision support on optimal cost at $S^*=7$ under the constraint of reduction of AWT by 68.9%.....	29

List of Table

Table 5.1 Numerical data on AWT and the corresponding cost values with range of S from unity to 12 for a fixed arrival rate of patients ($\lambda = 3.5$).....	30
---	----

Chapter 1 Introduction

Hospital plays an important role for the healthcare system of society. They have changed rapidly in parallel with improvements in the science and technology of medical instruments and medicine. The research domain of health services operation management (HSOM) appears to be an increasing challenge to hospital administrators and health policy makers. The goal of HSOM should have spared no efforts to develop strategies that will enable the provision of high-quality services, while operating with increased costs and under pressure resulting from competition [1]. One of the most demanding departments in terms of economic resources consumption and programming is the hospital emergency department (HED). To this extent its operational profile should be monitored and optimized in order to provide the optimal quality of medical service subject to the budget constraint.

The operation of HEDs must be available 24 hours, and moreover, it should respond to multiple demands requiring sophisticated technical equipment, and the manpower to operate these, all of which imply higher costs. Large HEDs have even higher costs because they offer a wide range of services that would be unavailable in a small rural HED [2]. The HEDs pose traditionally a crucial issue concerning hospitals' cost containment and management. The optimization of patient flow and bottleneck elimination in key departments could be a viable way at policy maker disposal to decrease operational cost and boost the quality of care [3]. In the interest of patient throughput and resource utilization, appropriate key performance measures are selected like the deployment on the size of staffing providers, HED patient arrival patterns, and service rate of staffing providers, waiting time, etc. To explore the tradeoff study among them, the proposed queue-based optimization technique on cost may provide the hospital management with decision support on the deploying the number of staffing providers in HED under constraints of kernel performance parameters.

The operation of HEDs must be available 24 hours, and moreover, it should respond to multiple demands requiring sophistical equipment, and the manpower to operate these, all of which imply higher costs. Large HEDs have even higher costs because they offer a wide range of services that would be unavailable in a small rural HED [2]. The HEDs pose traditionally a crucial issue concerning hospitals' cost containment and management. The optimization of patient flow and bottleneck elimination in key departments could be a viable way at policy maker disposal to decrease operational

cost and boost the quality of care [3]. In the interest of improving patient throughput and resource utilization, the appropriate key performance measures are selected like the deployment on the size of staffing providers, HED patient arrival patterns, and service rate of staffing providers, waiting time etc.

This novel idea in this work is originated from the theory of an M/M/S/K queuing system, which is used to estimate the optimal number of providers needed during each staffing interval [4]. At some pre-configured period (say a shift, or a day), there exists a finite quantity of staffing providers to provide medical service under limited waiting rooms in HED for patients. On application modeling, such finite quantity of staffing providers (i.e., S) can be regarded as the term: “server” in the M/M/S/K model of queuing theory. The quantity of $(K - S)$ can be considered to be the rather limited waiting rooms in HED regulated by each hospital.

The research goal for this work is to explore the issue: On the cost-based deployment, how many sets of staffing providers in the HED schedule would be optimal if a certain level of the server availability is kept? To explore the tradeoff study on them, the optimization technique may provide the HED management with decision support on the number of staffing providers. The key contributions of this paper are threefold: (1) this work provides HED administrator with an efficient deployment of staffing providers for the HED platform to optimize the cost improvement. On management aspect, the proposed system can be adopted to be a decision-making methodology approaching predictive management other than reactive or chaotic management. (2) On quantitative analysis, the M/M/S/K queue model has been applied and derived, and then relevant system metrics has been established in a brand-new manner. The mathematical expression for cost function has been established for evaluation requirement. (3) On verification aspect, relevant experimental results are conducted and obtained in terms of configurations on cost optimization and average waiting time. The simulated results indicate that the proposed approach may provide a feasible decision support for deployment on quantities of standby servers.

The rest of the paper is organized as follows: Chapter 2 describes related work. To demonstrate the framework qualitatively, an M/M/S/K model of queuing theory is adopted and the mapping profile is demonstrated in Chapter 3. On quantitative work in Chapter 4, the mathematical analysis has been conducted in detail and also relevant system performance measures like the expected number of online servers, the expected number of spares, etc. have been derived. Following this, in Chapter 5, the queue-based model is further addressed in terms of cost function, which simulations

are conducted as well for the feasibility of the proposed scheme. Finally, some concluding remarks are made in Chapter 6.

Chapter 2 Related Work

The HED crowding represents an important issue that may affect the quality and access of health care. Accordingly, the optimization on average waiting times in the HED has become a focus across many mainstream hospitals [5][6][7]. As defined by the Canadian Association of Emergency Physicians [8], HED overcrowding is a situation in which the demand for services exceeds the ability of health care professions to provide care within a reasonable length of time. As stated in [9], significant variation in emergency department (HED) patient arrival rates necessitates the adjustment of staffing patterns to optimize the timely care of patients. The authors in [9] collected detailed HED arrival data from an urban hospital and used a queue-based analysis to gain insights on how to change provider staffing to decrease the proportion of patients who leave without being seen. However none of optimization materials in terms of mathematical theory were addressed at all for these open literatures[5][6][7].

Finamore et al.[8] described an innovative use of a satellite clinic to divert returning to HED for care on a scheduled basis. Their strategy allows patients returning for following-up diagnostics or treatment to by pass the main HED. The proposed HED satellite clinic may shorten the waiting times by multiply ways like: (1) Increasing capacity to remove returning patients from the pool of patients requiring care in the HED. (2) Creating a separated staffed treatment area. Emergency department (HED) visit data were used to measure crowding and completion of waiting room time, treatment time, and boarding time for all patients treated and released or admitted to a single HED during 2010. In the work of [10], the authors conduct HED relevant statistical analysis and concluded that HED census at arrival demonstrated variation in crowding exposure as time-varying HED census. In the work of Wiler et al. [11], the authors developed an agent-based simulation model for the evaluation of FTT (Fast Track Strategies) scheme applied in the hospital HED to reduce patient waiting time. By and large, the issues regarding cost optimization on the HED management cost are not a concern for these open literatures [8][10][11].

Vass and Szabo [12] evaluated 2195 questionnaires in the HED situated in Mures County, Romania for a period three years (2010-2013). Their research reported that long waiting times was the most important complaint in patient's satisfaction surveys. To perceive the quantitative profile, a specific M/M/3 queuing model had only been considered in their work to demonstrate the computation details. Motivated by the work of [12], an interesting issue inspires us: Is it possible to provide an effective and

feasible approach to be the decision support on the optimization of provider staffing under cost constraint for the hospital HED with more elaborative queue-based framework? This research generalizes the queuing model of [12] into M/M/S/K queuing framework in terms of three practical aspects: (1) Numbers of medical server (provider staffing) can be configured to one of system parameters instead of a fixed quantities. Such a dynamic staffing level enables a hospital to quantify the cost patterns and the alleviation of HED waiting times as well. (2) The space available in the HED capacity would be limited for every hospital management. The fourth factor (K) in the notation of M/M/S/K model symbolizes the fact that there are only K patients can be allowed to enter the waiting rooms of the HED in order not to exacerbate the issue of overcrowding. (3) The exact mathematical expressions would be derived in an elaborative manner and the relevant cost formulation would be used to provide the generic decision support for the hospital administrators.

Chapter 3 The Proposed Model of Medical Emergency Service

3.1 The generic platform of medical emergence service

This research explores feasible decision support proposed to optimize running cost under the constraint of the waiting time at a hospital HED using queue-based models. The exemplified HED is in a metropolitan hospital (Taichung Veterans General Hospital or TVGH) located in central Taiwan. It began offering medical services on September 16, 1982. Since 1991 it has been accredited as a “Medical Center and First- Class Teaching Hospital” by the Department of Health, Taiwan. Taichung Veterans General Hospital is a 1,500-bed hospital with more than 3,500 employees. According to the statistical average data of registration each year in TVGH (2013-2015), it has offered the capacity of medical services composed of more the 6,000 outpatients, 130 inpatients and 180 patients in the emergency room daily [13].As a public medical center, it provides safe, high-quality medical service with advanced facilities and training programs as well as outstanding research and development programs.

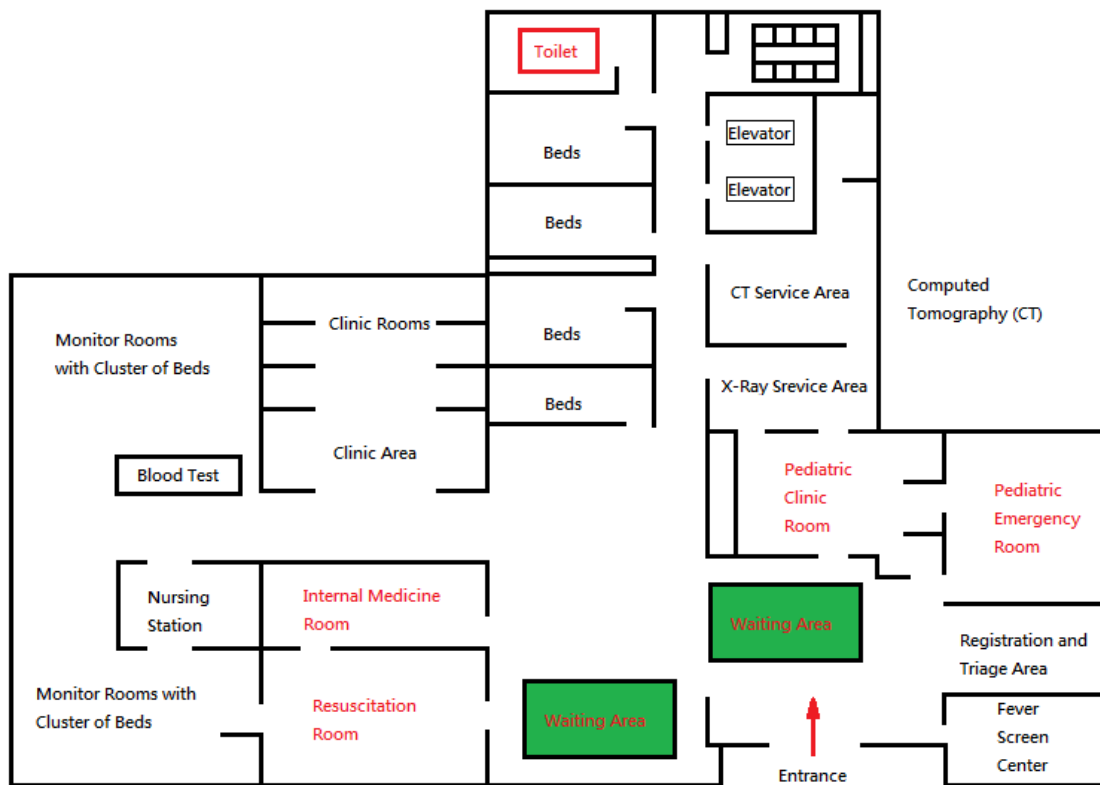


Fig.3.1 The functional deployment on the ground floor of TVGH-HED building

The HED building with eight floors at the TVGH (TVGH-HED) provides comprehensive emergency services 24 hours a day. The functional deployment on the ground floor of TVGH-HED building, as shown in Fig.3.1, is composed of different zones including Registration Triage Area, Resuscitation Areas, Internal Medicine Areas, Pediatric Treatment Area, Waiting Area, Clinic Area, Monitor Rooms and Fever Screen Center, and relevant auxiliary service units like X-ray service and nursing stations. Since the hospital HED is a rather complicated service system due to random arrivals, various disease chains, uncertain service times of care, and randomness in human decision-making, it is difficult to model the whole HED with a single operational model. From the perspective of model genericity [14], a generic operational model is defined as a formal description of operations performed to deliver a health service that is applicable in a wide range of health service delivery settings. For the sake of simplicity, this research concentrates the optimization issue on a specific platform of medical service which are used to model staffing providers for a single disease chain hereafter.

Another exemplified hospital is the famous Chang Gung Memorial Hospital (CGMH) which is composed of 8 medical centers in Taiwan. The CGMH was founded in 1976 by the well-known entrepreneur Mr. Yung-Ching Wang whose vision was to reform healthcare environment to ensure that both the rich and the poor would have equal access to good hospital healthcare. One of the eight CGMH medical centers, the Linkou CGMH, was visited for the research thesis to gain the medical resource profile. The Linkou CGMH-HED serves 15,000 patients each month and hope for improve efficiency and reduce error rate[15]. Considering vicissitude of social structure, the elevation of life quality, the demand for high medical quality, those contribute to the importance of emergency medical service. Due to the multi-dimensional and complexity of emergency medical service, physicians and nurses encountered more challenges while facing emergent patients.

The Linkou CGMH-HED was established officially since 1992, namely devotes to the promotion of emergency medical service quality, and had completed each system's impetus under the hospital support, for example the establishment of 24 hours all time attending physicians care, the establishment of none separate fields emergency medical system, the assistance for government to create pre-hospital emergency service medical system, the impetus of emergency training courses, the strengthening of personnel training and the promotion of specialist system of emergency medical department[15]. The Linkou CGMH-HED provides comprehensive emergency services 24 hours a day. The functional deployment on the two floors of Linkou

CGMH-HED building, as shown in Fig.3.2, is composed of three zones including Registration Triage Area, Resuscitation Areas, Adult Treatment Area, Waiting Area, Clinic Area, Monitor Rooms and Fever Screen Center, and relevant auxiliary service units like X-ray service and nursing stations.

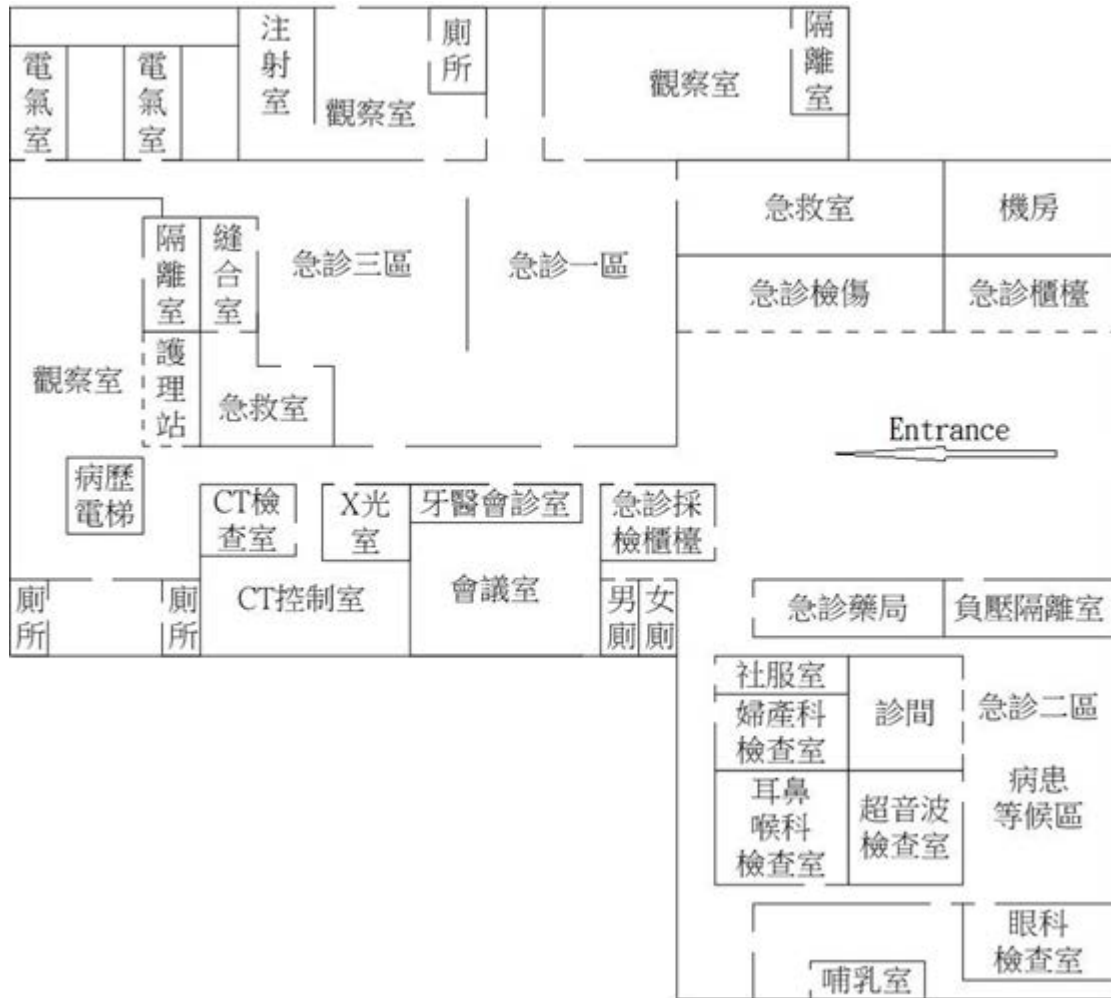
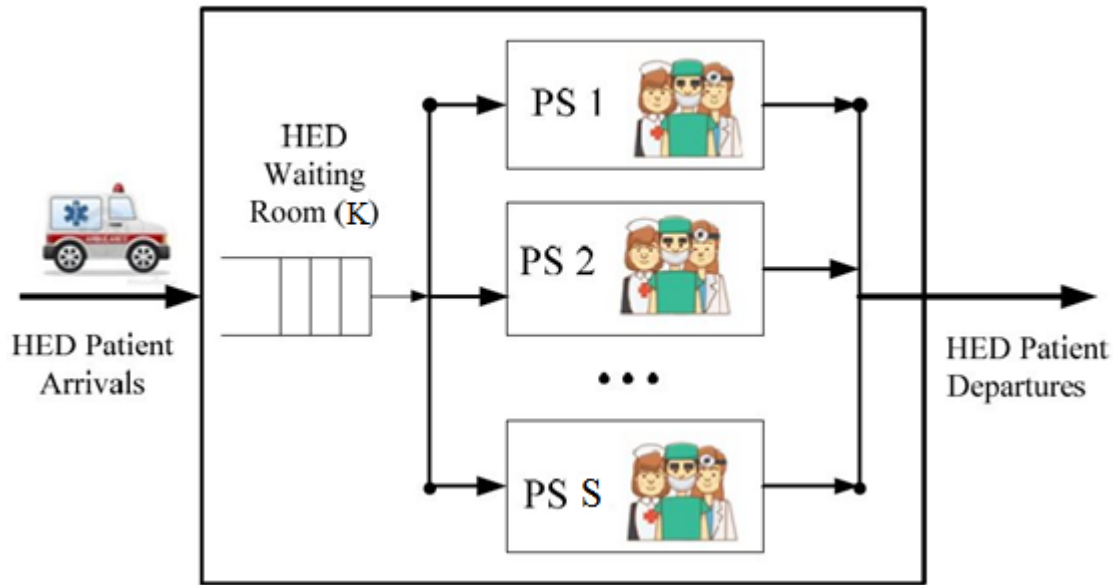


Fig.3.2 The functional deployment on the ground floor of CGMH-HED building

3.2 Mapping profile between HED service platform and the M/M/S/K QS

The proposed generic framework on the HED service platform is considered to be modeled as an M/M/S/K queuing system (QS), which is used to estimate the optimal number of providers needed during each staffing interval. An input-throughput-output framework of HED operations is used as the prototype shown in the Fig.3.3 for generic profile [16]. The icon of ambulance symbolizes the arrivals of HED patients. Practically, patient arrivals are hard to be scheduled or even controlled significantly. Arrival may surge on some unpredictable time windows due to short-term disaster, car accidents, and seasonal influences [17]. On modeling language, the busy and regular time windows can be associated with high and normal arrival rates respectively. Patient arrivals in the proposed model are assumed to be Poisson processes [18], with average hourly rates that are forecasted for each future hour in question (say a shift, or a day) [19].


Generic HED Service Platform



Legend:

HED: Hospital Emergency Department

PS: Provider Staffing

 : The icon of ambulance symbolizes the arrival of HED patients

 : This icon represents the HED waiting facility for treatment-phase


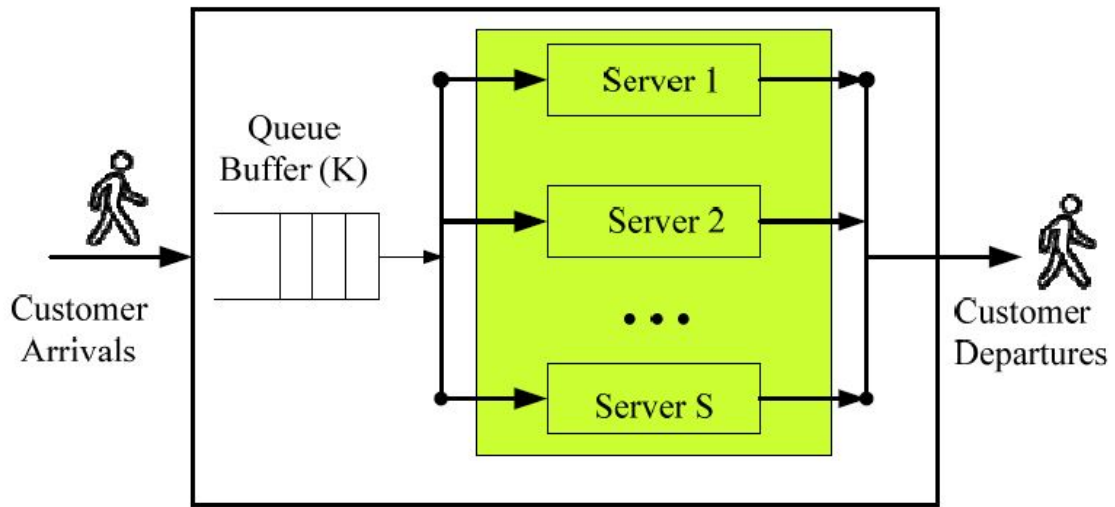
 : This icon symbolizes medical treatment resources composing of nurses, physicians, and medical instruments, etc.

Fig.3.3 The generic service platform of emergency department

The itinerary for HED patients from arrival to exit can conceptually be divided into three phases [11]. The first phase named as “Waiting for treatment phase (waiting phase)” as symbolized by the icon: HED Waiting Rooms of Fig.3.3. In the waiting phase the patient goes through some standard process that assist the HED to grasp the record of patients and current medical condition. These are termed as registration and Triage process, respectively. Registration guaranteed administratively that patient demographics are captured accurately for billing and maintaining record. Triage is the first assessment conducted by a healthcare professional after the patient arrives in the HED. The second phase (treatment phase) begins when the patient is placed in bed. For the reason of simplicity, the treatment phase is represented by the icon of provider staffing (medical servers) in Fig.3.4 for generic profile. The whole medical service depends largely in patient acuity and the physician activities. On modeling language, the duration of treatment can be regarded as the service rate of (medical) server mathematically. The treatment phase is followed by the post-treatment phase which is represented by the expression: HED patient departures in Fig.3.4. Exiting from the treatment area of HED, it is reasonably assumed that the patients are discharged either as outpatient or into hospital for the HED patient departures.

HED Platform modeled as an M/M/S/K Queue



Legend:



: This icon symbolizes customers entering the M/M/S/K queuing system



: This icon represents the queue buffer for customers waiting for service



: The QS service facility is depicted by this icon

Fig.3.4 An M/M/S/K queue system mapped by the HED service platform

The mapping scenarios are illustrated between Fig.3.3 and Fig.3.4 for theoretical approach. An M/M/S/K queuing model was used to estimate the number of providers needed during each staffing time window. In Fig.3.4, the proposed model assumes a single queue with regulated and finite waiting rooms that feeds into S identical servers with blue highlights which mapped to providers in hospital HED. The icons of walking-man (customers) symbolize the HED patient arrivals. Based on the proposed queuing model, relevant system metrics like expected number of customers in the queue buffer, and the probability that all servers are busy can be analyzed and derived mathematically [9]. For instance, a patient's total length-of-stay from arrival to departure from the HED platform is termed as the patient throughput time, which is equivalent to the waiting times in the QS. Patient throughput time has a significant impact on operational and economic efficiency as well as overall patient satisfaction, which is a measure of medical service quality [20].

Generally, the performance metric on average waiting times may provide the HED administrator with a decision support in how to alleviate patients' complaint. To avoid the deterioration of average patient throughput time (i.e., the average wait times in the QS), the optimization approach on the average waiting times under some constraints like limited amount of servers in the QS (i.e., mapped counterpart: level of staffing in the HED platform) would be explored further in this article. The metric on the probability that all servers can be used to reveal the possibility and scenario in which the notorious HED crowding may occur. How to reduce HED crowding phenomenon in some specific times windows? Such a metric can provide decision support to properly configure/ deploy hospital resources for the HED administrator as well.

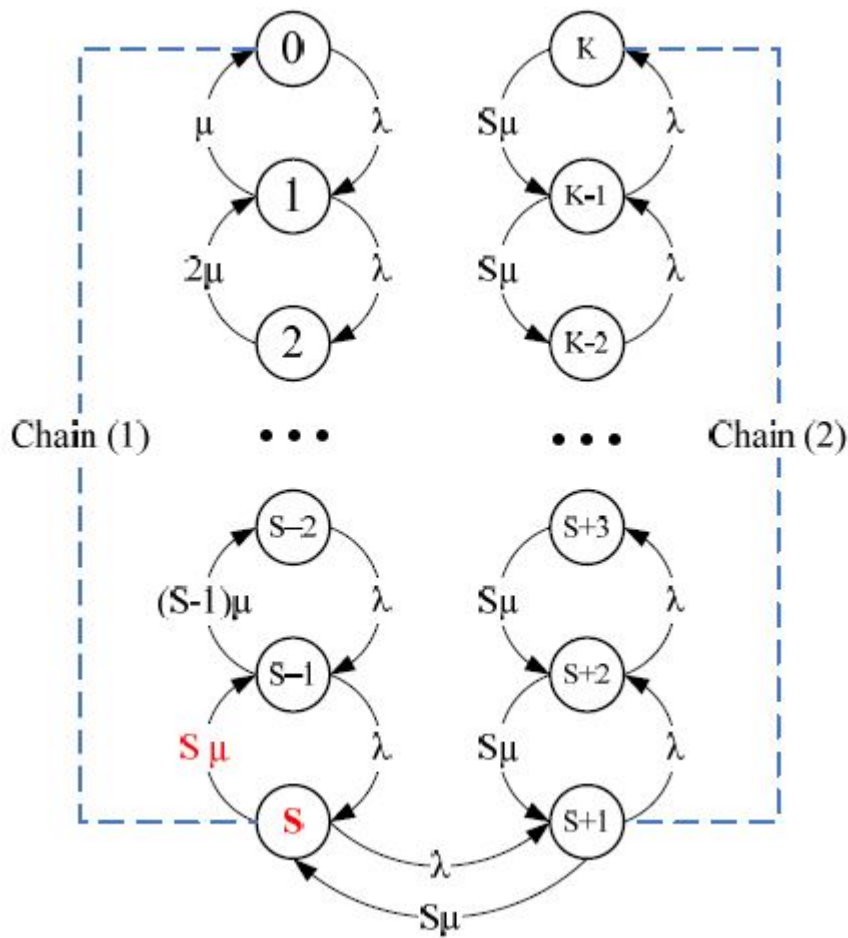
Chapter 4 Quantitative Modeling and System Measures for the HED Platform

4.1 Theoretical Analysis

The service-oriented model on the HED platform in Fig.3.4 is considered to have S servers with adequate level of staffing and the finite size (K) of waiting room for HED arrivals. The birth-and-death process is adopted to derive analytic steady-state solutions to the $M/M/S/K$ queuing system (QS). Let the states n ($n = 0, 1, 2, \dots, K$) represent the number of customers in the QS. McManus et al. [18] studied all admissions to the medical–surgical ICU of a large, urban children’s hospital during a 2-year period. Their statistical analysis had confirmed that the arrival rate of patients to ICUs follows a Poisson distribution. And the durations of stay (service times) were found to follow an exponential distribution as well. Hence, it is reasonably assumed that the customers arrive according to a Poisson process with mean arrival rate $\lambda_n = \lambda$ if $0 \leq n \leq K$ and $\lambda_n = 0$ if $n > K$ due to finite system capacity. The QS has S servers, each having the service times an exponential distribution with an identical service rate $\mu_n = \mu$. The service volume can be classified into two parts as follows:

$$\text{Mean Service Rate: } \mu_n = \begin{cases} n\mu, & \text{if } 1 \leq n \leq S \\ S\mu, & \text{if } (S + 1) \leq n \leq K \end{cases} \quad (1)$$

To approach analytic steady-state results for the proposed model, we first construct the state-transition-rate diagram depicted in Fig.4.1. The number inside the circle represents the number of customers (patients) in the system. Each circle in Fig.4.1 shows the steady-state probability scenario that may happen during service period in the system. For each circle except the first one ($n = 0$) and the last one ($n = K$), there are four arrows marked with the corresponding values of state-transition rate. The quantity marked along each arrow implies either flow-in probability into that state or flow-out probability off that state.



Legend:

\textcircled{n} : denoting the state that there are n customers in the QS.

S : number of servers deployed in the QS

λ : mean arrival rate for the ED customers into the QS.

μ : mean service rate of one server .

K : the queue capacity regulated by the administration.

Fig.4.1 State-transition-rate diagram for the proposed model

Let the notation: $P(n)$ = the probability that there are n customers in the system where $n=0,1, 2, \dots, K$. Hence. For steady-state case, the state probability functions $P(n)$ can be obtained from the birth-and-death formula in association with the state-transition-rate diagram shown in Fig.4.1. We define notations $\rho = \lambda/\mu$ for the server utilization and $\rho_w = \rho/S = \lambda/(S\mu)$ for the whole system utilization. According to the value n (number of customers in the QS) may happen, two segments are defined by the vector: [Chain (1), Chain (2)] = [$1 \leq n \leq S$, $(S+1) \leq n \leq K$]. the state probability functions $P(n)$ can then be derived in terms of two segments as follows:

Chain (1): $1 \leq n \leq S$

$$\begin{aligned} P(n) &= \frac{\lambda_0 \cdot \lambda_1 \cdot \lambda_2 \cdots \lambda_{n-1}}{\mu_1 \cdot \mu_2 \cdot \mu_3 \cdots \mu_n} P(0) = \frac{\lambda^n}{\mu (2\mu)(3\mu) \cdots (n\mu)} P(0) = \frac{\lambda^n}{\mu^n n!} P(0) \\ &= \frac{\rho^n}{n!} P(0) \quad (2) \end{aligned}$$

Chain (2): $(S+1) \leq n \leq K$,

$$\begin{aligned} P(n) &= \frac{\lambda_0 \cdot \lambda_1 \cdot \lambda_2 \cdots \lambda_{n-1}}{(\mu_1 \cdot \mu_2 \cdots \mu_R)(\mu_{R+1} \cdots \mu_n)} P(0) \\ &= \frac{\lambda^n}{[\mu \cdot (2\mu) \cdots (S\mu)](S\mu \cdots S\mu)} P(0) \\ &= \frac{\lambda^n}{S! \mu^S (S\mu)^{n-S}} P(0) = \frac{\rho^n}{S! (S)^{n-S}} P(0) \quad (3) \end{aligned}$$

There are $(n-S)$ terms of $S\mu$ in the parenthesis: $(S\mu \dots S\mu)$ of the above denominator. Equations (2) and (3) are the closed-forms for the state probability functions $P(n)$ in terms of two chains in which the number of customers may happen. To obtain $P(0)$, we substitute expressions (2) and (3) in normalizing equation : $\sum_{n=0}^S P(n) = 1$ and it yields:

$$\begin{aligned} \sum_{n=0}^S \frac{\rho^n}{n!} P(0) + \sum_{n=S+1}^K \left(\frac{\rho^n}{S! S^{n-S}} \right) P(0) &= 1 \\ P(0) &= \left[\sum_{n=0}^S \frac{\rho^n}{n!} + \sum_{n=S+1}^K \left(\frac{\rho^n}{S! S^{n-S}} \right) \right]^{-1} = \left[\sum_{n=0}^S \frac{\rho^n}{n!} + \frac{\rho^S (1 - \rho_w^{K-S+1})}{S! (1 - \rho_w)} \right]^{-1} (4) \end{aligned}$$

4.2 System Performance Measures

Mathematical expectations are crucially important for the long-run theoretical average values of relevant parameters in the system. To formulate the expressions regarding system performance metrics, it is necessary to construct average-based functions such as expected number of customers in the queue, expected number of busy servers in the system and the like. The following mathematical analyses are all necessity of system performance measures of an M/M/S/K QS.

Let

L_s = expected number of customers in the system,

L_q = expected number of customers in the queue buffer,

$E[I]$ = expected number of idle servers,

$E[B]$ = expected number of busy servers,

P_B = Probability that all servers are busy,

W_s = average waiting times in the system,

W_q = average waiting times in the queue buffer.

With steady-state probability functions (2~4), it yields

$$L_s = \sum_{n=0}^K n P(n) \quad (5)$$

$$L_q = \sum_{n=S}^K (n - S) P(n) \quad (6)$$

$$E[I] = \sum_{n=0}^{S-1} (S - n) P(n) \quad (7)$$

$$E[B] = S - E[I] \quad (8)$$

$$P_B = \sum_{n=S}^K P(n) \quad (9)$$

To express above parameters in terms of $(S, K, \rho, \rho_w, P_0)$, the system performance measures can be derived as follows:

$$\begin{aligned} L_s &= \sum_{n=0}^K n P(n) = \sum_{n=0}^{S-1} n P(n) + \sum_{n=S}^K n P(n) = \sum_{n=0}^{S-1} n \cdot \frac{\rho^n}{n!} P(0) \\ &\quad + \sum_{n=S}^K (n - S + S) P(n) \\ &= \sum_{n=0}^{S-1} n \cdot \frac{\rho^n}{n!} P(0) + \sum_{n=S}^K (n - S) P(n) + S \sum_{n=S}^K P(n) = \sum_{n=0}^{S-1} n \cdot \frac{\rho^n}{n!} P(0) + L_q + \\ &R P_B \quad (10) \end{aligned}$$

$$P_B = \sum_{n=S}^K P(n) = \sum_{n=S}^K \frac{\rho^n}{S! S^{n-S}} P(0) = \frac{\rho^S}{S!} \frac{[1-(\rho_w)^{K-S+1}]}{(1-\rho_w)} P(0) \quad (11)$$

By the change of indices: $j = n-S$ on expression (6), it yields:

$$L_q = \sum_{n=S}^K (n - S) P(n) = \sum_{n=S}^K (n - S) \frac{\rho^n}{S! S^{n-S}} P(0) = \frac{\rho^S P(0)}{S!} \sum_{j=0}^{K-S} [j \cdot (\rho_w)^j] \quad (12)$$

The average waiting times in the system and in the queue buffer (W_s , W_q) can be derived by applying the Little's formula, that is, $W_s = \frac{L_s}{\lambda}$ and $W_q = \frac{L_q}{\lambda}$ respectively.

Chapter 5 Performance Evaluation

5.1 Introduction to MATLAB Simulation Tool

MATLAB® is a high-level language and interactive environment for numerical computation, visualization, and programming [21][22]. Using MATLAB, we can analyze data, develop algorithms, and create models and applications. The new version of MATLAB is R2018a. It has new tools for building apps, writing scripts, and team-based software development. More options for data analytics, machine learning, and deep learning[21].

The language, tools, and built-in math functions enable user to explore multiple approaches and reach a solution faster than with spreadsheets or traditional programming languages, such as C/C++ or Java®. The MATLAB can be used for a range of applications, including signal processing and communications, image and video processing, control systems, test and measurement, computational finance, and computational biology. Over one million of engineers and scientists in industry and academia use MATLAB tool package, the language of technical computing [23].

The key features are stated briefly as follows:

- High-level language for numerical computation, visualization, and application development
- Interactive environment for iterative exploration, design, and problem solving
- Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, numerical integration, and solving ordinary differential equations
- Built-in graphics for visualizing data and tools for creating custom plots
- Development tools for improving code quality and maintainability and maximizing performance
- Tools for building applications with custom graphical interfaces
- Functions for integrating MATLAB based algorithms with external applications and languages such as C, Java, .NET, and Microsoft® Excel®.

Numeric Computation

MATLAB provides a range of numerical computation methods for analyzing data, developing algorithms, and creating models. The MATLAB language includes mathematical functions that support common engineering and science operations. Core math functions use processor-optimized libraries to provide fast execution of vector and matrix calculations.

Available methods include:

- Interpolation and regression
- Differentiation and integration
- Linear systems of equations
- Fourier analysis
- Eigen values and singular values
- Ordinary differential equations (ODEs)
- Sparse matrices

MATLAB add-on products provide functions in specialized areas such as statistics, optimization, signal analysis, and machine learning.

5.2 Issue on Decision Support for HED Management

The strategy to minimize the total cost of the operating horizon is referred to as the optimal policy. Like all institutions of medical business, the cost pattern attracts much attraction for approaching long-term running steadily on the hospital management. To optimize the cost, we develop a steady-state expected cost function per unit time for an M/M/S/K queuing system, in which the parameter vector of $[S, K, \lambda, \mu]$ and the average waiting times (W_q) are considered to be decision variables. The cost element CW is defined as the waiting cost per unit time (or called cost rate) per customer (HED patient) present in the system. Our goal is to provide the decision support on determining the optimal number of servers S, say S^* , so as to optimize the cost function. To formulate the cost function, some cost parameters are defined in the following vector form as follows:

C_q = cost rate (cost per unit time) when one customer present is waiting for service,
 C_s = cost rate when one customer joins the system and is served,

$[C_B, C_I]$ = cost rate when one server is [busy, idle].

Using the definitions of each cost element with its corresponding feature, the cost function $F(S, K)$ can be developed in association with system metrics: $L_s, P_B, L_q, E[I]$, and $E[B]$ of which are given in equations (10), (11), (12), (7) and (8) respectively. It is noted that the steady-state probabilities for two segments are given in expressions (2) and (3). The probability that there are no customer in the system: $P(0)$ is given by expression (4).

$$\begin{aligned}
F(S, K) &= C_q L_q + C_s (L_s - L_q) + C_B E[B] + C_I E[I] \\
&= (C_q - C_s) L_q + C_s L_s + C_B E[B] + C_I E[I] \\
&= (C_q - C_s) \left[\frac{\rho^S P(0)}{S!} \sum_{j=0}^{K-S} j \cdot (\rho_w)^j \right] \\
&+ C_s \left[\sum_{n=0}^{S-1} n \cdot \frac{\rho^n}{n!} P(0) + \frac{\rho^S P(0)}{S!} \sum_{j=0}^{K-S} [j \cdot (\rho_w)^j] \right] + R \frac{\rho^S P(0) [1 - (\rho_w)^{K-S+1}]}{S! (1 - \rho_w)} \\
&+ C_B [S - \sum_{n=0}^{S-1} (S - n) P(n)] + C_I \sum_{n=0}^{S-1} (S - n) P(n) \tag{13}
\end{aligned}$$

The cost function $F(S, K)$ in equation (13) is expressed in terms of basic parameters like $[S, K, \lambda, \mu]$ and cost elements. It is noted that the utilization parameters of the unit server and system is given by $[\rho, \rho_w] = [\lambda/\mu, \lambda/(S\mu)]$ respectively. The state probability functions $P(n)$ for two segments are given in expressions (2) and (3), which are quite complex for the control parameter S . To find the optimal profile on the cost function, it is of necessity to show the existence of convexity or unimodality function of $F(S, K)$. However, this mathematical task is difficult to implement. The cost function $F(S, K)$ is unimodal; that is, it has a single relative minimum.

5.3 Optimization Evaluation

Examining equation (13), it is noted that the parameter R occurs not only at the location of in-line items but also of upper limit of summation symbol Σ , which makes $F(S, K)$ a highly nonlinear and complex function. Instead, practical numerical examples are presented and intensively studied by applying the proposed models. The optimization evaluation is firstly probed in terms of cost patterns in this subsection. For illustrative purpose, we firstly study the effect of varying S while keeping K constant, and then varying N while keeping S constant. All simulations are performed with MATLAB platform with custom MATLAB scripts. The exemplified system parameters are listed to be vector forms as follows:

- (a) Average arrival rate of patients (λ) = 2.5, 3.0, and 3.5,
- (b) Average service rate of a server (μ) = 1,
- (c) Cost rate: $[C_q, C_s, C_B, C_I,] = [200, 150, 120, 100]$,
- (d) $K = 15$ for the emergency department of small-and-medium size.

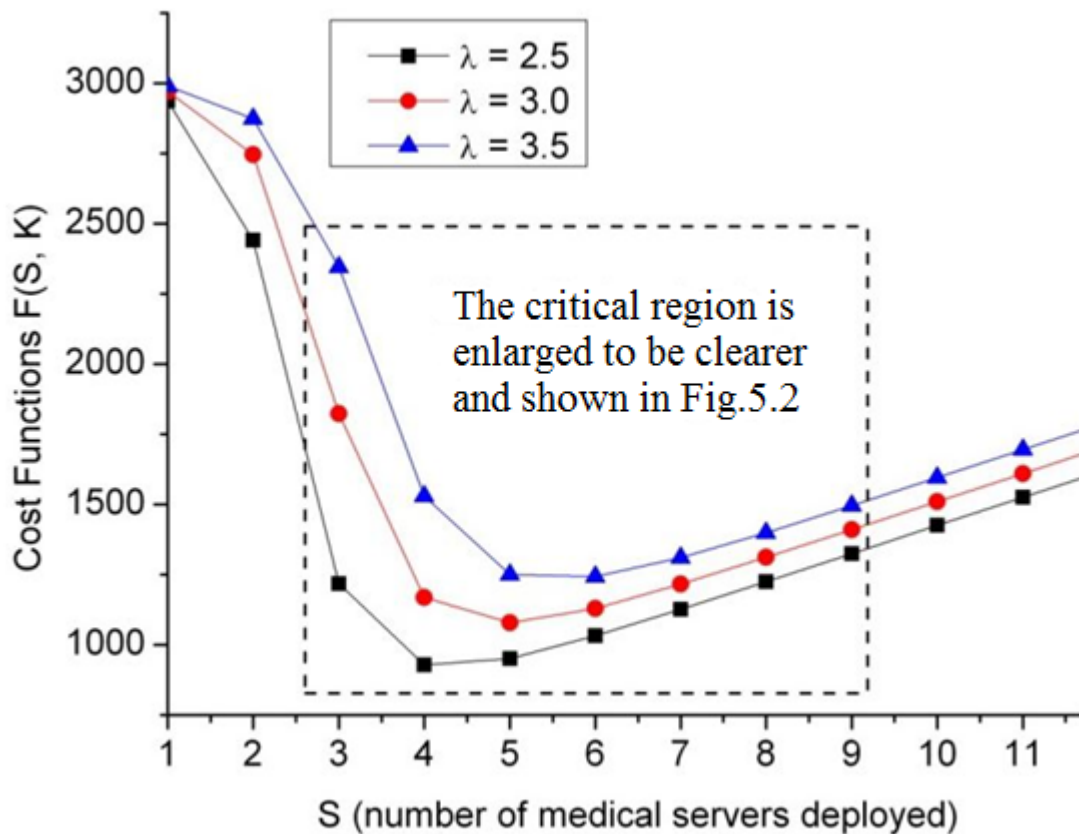


Fig.5.1 Optimal cost patterns shown in terms of three average arrival rates

Since contour plots provide the best graphical representation of the optimization problem, and also possess a powerful visualization that permits the solutions of the optimization problem by inspection. To validate the analytical solution, the graphical results are obtained and shown in Fig.5.1, which three cost contours with icons: black box, red circle, and blue triangle are depicted along the Y-axis in terms of $\lambda = 2.5$, 3.0, and 3.5, respectively. Generally, larger arrival rate of patients implies that medical service cost would be higher so that the blue line marked with triangle icon ($\lambda = 3.5$) is situated over the red line marked with circle ($\lambda = 3.0$). To be clearer on the crucial region surrounded by dash-line rectangle in Fig.5.1, the enlarged detail is depicted in Fig.5.2. In Fig.5.2, the X-axis is from $S = 3$ to $S = 9$ for critical region with an enlarged view. Each contour is attached by its optimal cost value with the corresponding optimal S^* .

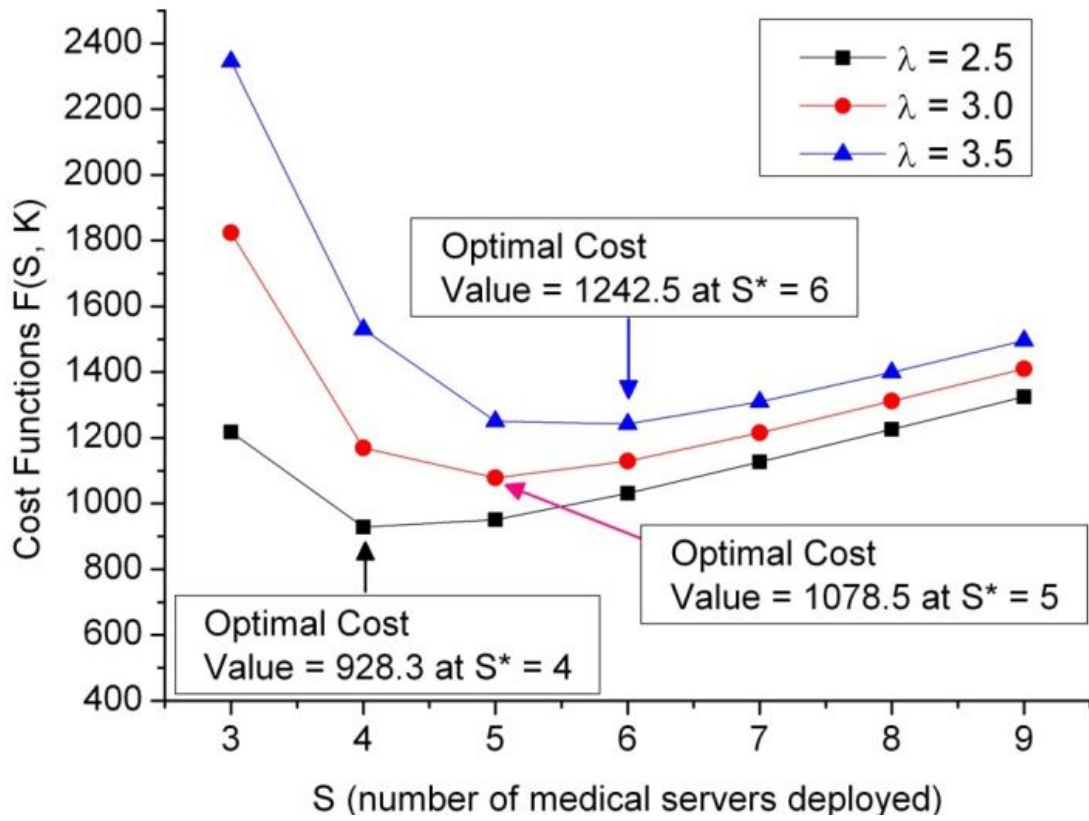


Fig.5.2 An enlarged diagram showing optimal cost data from Fig.5.1.

5.4 Issues on Cost Profile under the Constraint of Average Waiting Time

In view of performance evaluation on HED, the average waiting time (AWT) can be also regarded as a measure of performance committed to the HED patients, and of a yardstick for comparing the effectiveness in the deployment of the staffing providers on a quantitative manner. Practically, it is reasonable for HED management experts to guarantee target level on the AWT when they want to alleviate the sense of worry for potential HED patients. Logically the more the amount of staffing providers is deployed, the more expense the HED management would cost. Hence, the proposed approach would like to explore the issue on decision support for optimal cost under the constraint of the AWT in some target level.

In Fig.5.3 with double-Y axis, the left Y-axis and the right Y-axis are set to be the cost values and the average waiting time (AWT) respectively. Observing the solid-line contour marked with black rectangles (i.e., the left Y-axis), the optimal cost value $F(S,K) = 1234.5$ occurs at $S^* = 6$ based on similar parameters in Fig.5.2 with the average arrival rate $\lambda = 3.5$. However, the corresponding AWT approaches 6.84 units, which is a reference matrix for decision-marking. The proposed generic model could be used for general insights into the issue faced in deploying multiple staffing providers for a disease chain or a single department like the Department of Pediatrics shown in middle right-handed location of the ground floor in Fig.3.1. On the right Y-axis, the dash-curve marked with red star shows the variation profile on the performance metric for AWT.

During busy time window for a specific disease chain in HED, patients may spend hours crowded waiting rooms before seeing a doctor. Those who choose to tolerate larger waiting time expose themselves to others who may have contagious illness. To alleviate such an occasion impact, one straight approach of reducing the waiting time is to deploy more staffing providers for that specific disease chain. The simulation results in Fig.5.3 provider an exemplified decision reference on re-deploying the amount of staffing providers to alleviate the waiting time under a fixed average arrival rate of HED patient.

Then an issue emerges from the judgement: how many extra staffing providers are needed to gain the reduction on the AWT by some level (for example, 50%) without over-provisioning? Observing the red-star contour with the right Y-axis in Fig.5.3, it is

found that the AWT can be reduced by 68.9% at $S^* = 7$ (pointed by the red dash-line) at the expense of only adding one staffing provider and cost values $F(S^*=7,N) = 1309.8$ than the minimum cost $F(S=6,N) = 1242.5$. The detailed numerical data are listed in Table 1 with range of S from unity 12. In other words, the proposed approach can provide a quantitative decision support on the trade-off study between the cost profile and the amount of staffing providers on HED deployment.

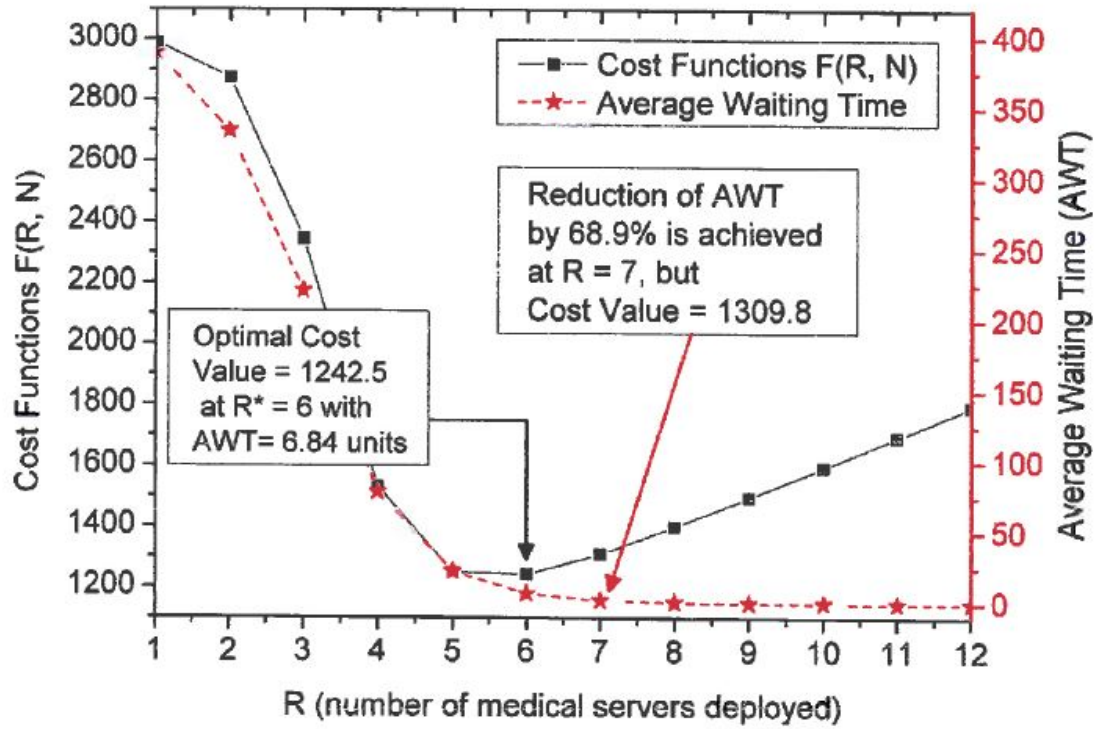


Fig.5.3 Decision support on optimal cost at $S^*=7$ under the constraint of reduction of AWT by 68.9%.

S	Cost Values F(S,K)	AWT
1	2990.0	388.57
2	2873.9	333.42
3	2345.8	220.77
4	1530.2	78.57
5	1250.7	22.51
6	1242.5	6.84
7	1309.9	2.13
8	1399.5	0.65
9	1496.3	0.19
10	1595.4	0.05
11	1695.1	0.01
12	1795.0	0

Table 5.1 Numerical data on AWT and the corresponding cost values with range of S from unity to 12 for a fixed arrival rate of patients ($\lambda = 3.5$).

Chapter 6 Conclusion

In terms of patients flow and available resources, an efficient generic methodology to optimize the performance of HED platform has been addressed in this research. The proposed queue-based approach provides ED administrator with an efficient deployment of staffing providers to optimize the cost profile. Conceptually the HED service platform is mapped into an M/M/S/K queuing system, and illustrated using appropriate figures and materials in the work. To gain the insight on the queuing model, the mathematical derivation is detailed for application need as well.

Based on the quantitative analysis, the M/M/S/K queue model has been applied and derived, and then relevant system metrics has been established in a brand-new manner. The mathematical expression for cost function has been established for evaluation requirement. On verification aspect, relevant experimental results are conducted and obtained in terms of integration configurations on cost optimization and patient arrival rates. Instead of chaotic management, the proposed generic methodology may provide feasible applications to approaching an effective decision support in terms of deploying appropriate staffing providers to alleviate the impact on HED crowding.

References

- [1] S. Zengin, R. Güzel, B. Al, S. Kartal, E. Sarcan, and C. Yildirim, “Cost analysis of a university hospital’s adult emergency service,” *The Journal of Academic Emergency Medicine*, vol. 12, pp. 71–75, 2013.
- [2] D. Simonet, “Cost reduction strategies for emergency services: Insurance role, practice changes and patients accountability,” *Health Care Annals*, vol. 17, pp. 1–19, 2009.
- [3] P. Cremonesi, E. Di Bella, and M. Montefiori, “Cost analysis of emergency department,” *Journal of Preventive Medicine and Hygiene*, vol. 51, pp. 157–163, 2010.
- [4] R.B. Cooper, *Introduction to Queuing Theory*, 2nd edition.1981, Elsevier Science Publishing Co., Inc. 52 Vanderbilt Avenue, New York 10017.
- [5] N. R. Hoot and D. Aronsky, “Systematic review of emergency department crowding: causes, effects, and solutions,” *Annals of Emergency Medicine*, vol. 52, no. 2, pp. 126–136, August 2008.
- [6] J. Kennedy, K. Rhodes, C. A. Walls, and B. R. Asplin, “Access to emergency care: restricted by long waiting times and cost and coverage concerns,” *Annals of Emergency Medicine*, vol. 43, no. 5, pp. 567–573, May 2004.
- [7] R. Derlet, J. Richards, R. Kravitz, “Frequent overcrowding in US emergency department,” *Academic Emergency Medicine*, vol. 8, pp. 151–155, 2001.
- [8] S. R. Finamore and S. A. Sheila, “Shorting the wait: A strategy to reduce waiting times in the emergency department,” *Journal of Emergency Nursing*, vol. 35, no. 6, pp. 509–514, November 2009.
- [9] L. V. Green, J. Soares, J. F. Giglio, R. A. Green, “Using queuing theory to increase the effectiveness of emergency department provider staffing,” *Academic Emergency Medicine*, vol. 13, no. 1, pp. 61–68, 2006.
- [10] M. L. McCarthy, R. Ding, J. M. Pines, and S. L. Zeger, “Comparison of methods for measuring crowding and its effects on length of stay in the emergency department,” *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1269–1277, 2011.

- [11] A. Kaushai, Y. Zhao, Q. Peng, T. Strome, E. Weldon, M. Zhang, and A. Chochinov, "Evaluation of fast track strategies using agent-based simulation modeling to reduce waiting time in a hospital emergency department," *Socio-Economic Planning Sciences*, vol. 50, pp. 18–31, 2015.
- [12] H. Vass and Z. K. Szabo, "Application of queuing model to patient flow in emergency department. Case Study," *Procedia Economics and Finance*, vol. 32, pp. 479–487, 2015.
- [13] Taichung Veterans General Hospital: Department s of Medical services
<http://www.vghtc.org.tw/>
- [14] A. Fletcher and D. Worthington, "What is a generic hospital model? --- a comparison of generic and specific hospital models of emergency patient flows," *Health Care Management Science*, vol. 12, pp. 374–391, 2009.
- [15] CGMH Emergency Medicine websites
<https://www.cgmh.org.tw>
- [16] N. R. Hoot, L. J. LeBlanc, I. Jones, S. R. Levin, C. Zhou, C. S. Gadd, and D. Aronsky, "Forecasting emergency department crowding: a discrete event simulation," *Annals of Emergency Medicine*, vol. 52, no. 2, pp. 116–125, August 2008.
- [17] J. K. Cochran and K. T. Roche, "A multi-class queuing network analysis methodology for improving hospital emergency department performance," *Computers & Operations Research*, 36 (2009) 1497–1512.
- [18] M. L. McManus, M. C. Long, A. Cooper, and E. Litvak, "Queuing theory accurately models the need for critical care resources," *Anesthesiology*, vol. 100, no. 5, pp. 1271–1276, May 2004.
- [19] Y. N. Marmore, S. Wasserkrug, and A. Shtub, "Toward simulation-based real-time decision-support systems for emergency department," *Proceedings of the 2009 Winter Simulation Conference*.
- [20] C. Oh, A. M. Novotny, P. L. Carter, R. K. Ready, D. D. Campell, and M. C. Leckie, "Use of a simulation-based decision support to improve emergency department throughput," *Operations Research for Health Care*, vol. 9, pp. 29–39, 2016.
- [21] MATLAB website
<http://www.mathworks.com/products/matlab/>
- [22] MATLAB wiki

<https://en.wikipedia.org/wiki/MATLAB>

[23]Online MATLAB Training

<http://www.mathworks.com/academia/tah-training.html>