Analysis and Dynamic Routing Implementation of Hierarcical Healthcare Referral System

Khairurizal Alfathdyanto, Prof. Ir. Abdullah Alkaff M.Sc., Nurlita Gamayanti ST. MT.
Department of Electrical Engineering, Faculty of Electrical Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
e-mail: khairurizal12@mhs.ee.its.ac.id, alkaff@ee.its.ac.id, nurlita@ee.its.ac.id

***Abstract*—Hierarchical Healthcare Referral System (HHRS) is implemented by National Insurance Providing Agency (BPJS) as part of the healthcare insurance policies. Patients who want to get health insurance in a hospital should get a referral from the community health center in which they are registered. Congestion of patients happens in certain hospital as there is no policy implemented to govern the referral system. In this paper, HHRS is modeled as a network of queuing system and is analyzed for its queue performances. Analysis of queuing network performances shows the influence patient preferences to buildup congestion of patients in hospitals. Referral is then controlled by means of dynamic routing with considering patient preferences. Estimation of arrival rate is done with hypercube queuing theory which concerns user preference. Simulation shows that patient preferences affect the arrival rate at each hospital, the application of dynamic routing can reduce the maximum utility and reduce the average waiting time, prioritization of patients improve dynamic routing performance on systems with a high workload.**

Keywords—dynamic routing; hierarchical healthcare referral system; patient priority; queueing networks;

I. INTRODUCTION

Hierarchical Healthcare Referral System (HHRS) is implemented by National Insurance Providing Agency (BPJS) as one of the procedure in its healthcare insurance. In the system, patients who want a health insurance must be examined in a first-level healthcare. Patient who needs further treatment (after been determined so) obtained a referral to hospital. The application of this system is being constrained by problem such as congestion of patients in famous hospitals. This phenomenon happens because there is no regulation that limits the destination of the referral which results in higher referral ratio to better hospital. As noted in [1], Ilir 5th Community Healthcare Center, in West Sumatera, had referral ratio up to 60%. The direction of referral also can be altered by patient preference. This would push the referral ratio into famous healthcare facilities further.

The implemented HHRS doesn't have strict regulation. Patients can still freely choose the direction of their referral. This resulted in congestion of patients on famous hospitals which has more complete equipment and more satisfactory service. Although the freedom of choice is important, it results a bigger problem of congestion on favorite hospital which affect the quality of service.

A referral allocating system would be needed to clear this problem. The system should consider patient's preference as it becomes the basis of initial selection for them. On other hand, the system needs to estimate the condition of destination hospital. This is to avoid potential congestion that will be arisen by assigning a patient. Then, it gives an alternative hospital to the patient which falls not far from their criteria of selection.

Congestion of patients on HHRS can be seen as a queuing network problem in public services. As mentioned in [2], HHRS seen as a queuing network with "blocking" properties as hospital assumed to have finite queue capacity. Such case happened as the hospital gets more crowded. In [3], queuing problem on healthcare facility can be solved by implementing appointment scheduling. Patient assigned a schedule beforehand so they will come on pre-assigned time, minimizing waiting time on the place.

In reality, patient can come any time and hospital can never refuse to treat them. It will be too late if too many patient assigned into a single hospital. Referral policy is needed to regulate patient flow beforehand which prevents upcoming congestion. This problem can be similarly seen as data packet regulation on communication network. In [4], Adaptive Virtual Delay (AVD) is a convenient solution that also includes delays of travel and waiting time. In this paper, HHRS to be implemented with this kind of routing algorithm so it will dynamically consider travel time in addition to waiting time on the facilities.

In this paper, HHRS will be observed based on patient flows within the network. In addition, different patient treatment will be implemented by assuming the severity of the disease. Therefore, determination of proper routing regulation on the system will be essential. Author also considers patient's preferences and priorities other than queue performances. A queuing network model of HHRS then is obtained and the routing regulation implemented on it to improve queue performances and solve the problem.

II. HIERARCHICAL HEALTHCARE REFERAL SYSTEM

Patients sometime require an additional treatment after an appointment with a doctor. The doctor then gives a referral for them obtaining a treatment on a better healthcare facility. In HHRS, healthcare facilities are divided into groups in

hierarchical order. Better healthcare facility is assigned in higher level, vice versa. Referral then advances patient from a healthcare facility to a higher-level facility on the hierarchy.

HHRS divides healthcare facilities into three different levels. First level consists of community health centers. Second level consists of hospital with general doctors. Third level consists of hospital with specialized doctor. The third level usually handles a severe patient and congestion rarely happens so only two levels are considered in the model.

A. *Queuing Network in HHRS*

Queuing network by definition is a set of connected queue systems. Note that each healthcare facility can be viewed as a queue system. Patient waits to be served by doctor which is server in this case. This queue usually is an M/M/n queue with n doctors. Patients depart from the queue either terminates from the system or given a referral to gain additional treatment in the hospital.

According to its hierarchical order we can group the facility into two groups, CHC and hospital. The queue systems of those two are connected through the referral, as a patient gets a referral he/she departs from CHC to a hospital. In other words, HHRS can be viewed as a queuing network. This is illustrated in Fig. 1

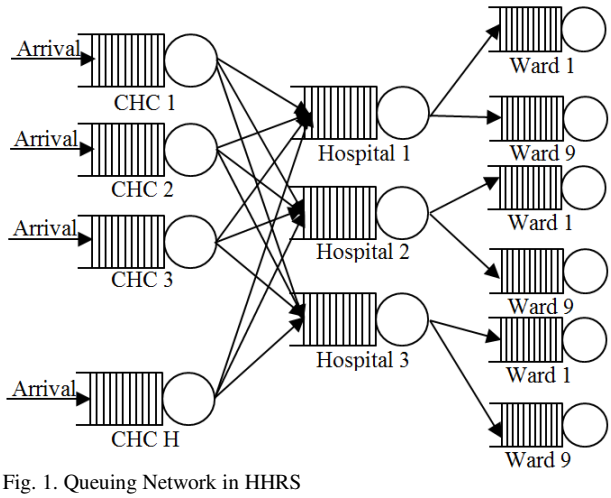

Fig. 1. Queuing Network in HHRS

B. *Dynamic Routing in HHRS*

Dynamic routing is user placement policy which considers the condition of server in the system according to specific criteria. Dynamic routing can be divided into three main components[4]:

1) Information policy, determines which state information of the system be collected and the method gaining those information.

2) Transfer policy, determines whether an incoming user is to be processed or transferred.

3) Location policy, determines which node that will process the transferred user.

Dynamic routing policy that becomes the basis of dynamic routing in this paper is Adaptive Virtual Delay (AVD). AVD aims to balance the virtual delay on each server. Virtual delay consists of virtual services delay and virtual transfer delay. Denote the virtual service delay of a server-i as F_i and virtual transfer delay between servers as G_{ij}. The value of F_i and G_{ij} can be expressed in the equation (1) and (2):

$$F_i = s_i (n_i + 1) \quad (1)$$

$$G_{ij} = \frac{t_{ij}}{1 - \rho'_{ij}} \quad (2)$$

The variable s_i represents average service time of server-i. Its value can be calculated by $1/\mu_i$. Next, the variable n_i represents number of queue on server-i. Last, the variable t_{ij} represents transfer time from server-i to server-j

Virtual service delay is the amount of time user must settle on a server-i if the user had entered the server-i at the time of calculation. While the virtual transfer delay is the travel time between servers when a user is determined to be moved to the other server. [4]

HHRS is a network of queue which spread on a region. Therefore, it can be seen as a spatially distributed queue (SDQ). One of model that concerns about this kind of queuing network is Hypercube Queuing System (HQS). System information like expected arrival rate can not be measured easily. An estimate is needed and can be obtained by using HQS model.

The name HQS is derived from the structure of its state transition diagram. In HQS, state is defined as condition of servers whether it is busy or idle. Then, state is represented as n-digit binary number. If server-i is busy then i^{th} digit of the binary number is 1, vice versa. State transition diagram of the system with 3 server forms a cube as shown in Fig. 2. For n>3, the state transition diagram will form a hypercube.[6]

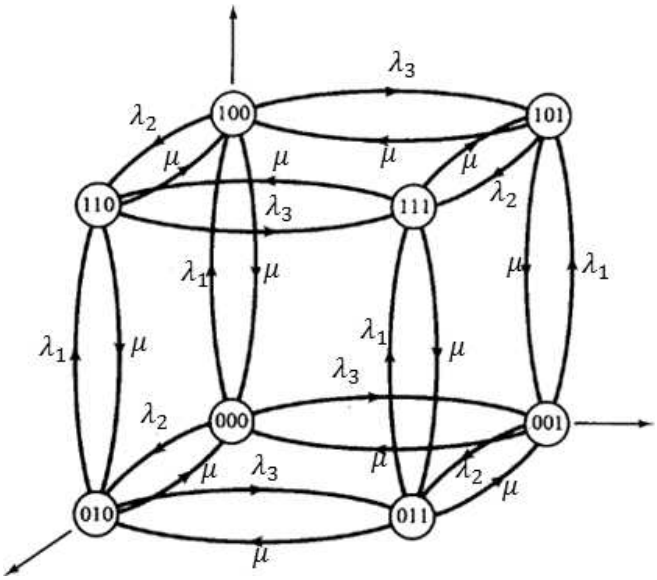

Fig. 2. Cube Shape Formed by State Transition Diagram of HQS

State equation that can be derived from the state transition diagram of HQS is as many as 2^n equations. Solution of the system for the number of servers that is less than or equal to 3

is not complex. However, exceeding that number may cause the search of system solutions to become complex so that it takes some approximation algorithm to estimate the value of such solutions. [7]

C. Priority in HHRS

Patient sometime has different degree of importance that we may see from the severity of the disease. Some of those can take precedence over the others in the queue. Such case can be called as queuing system with priority. Priority in queuing system can be implemented into 2 ways, preemptive and non-preemptive [7].

Non-preemptive priority is most likely used in healthcare facilities. This policy allows an ongoing service to proceed until it finishes before higher priority user gets a service. Likewise, in healthcare treatment, an ongoing treatment can not be interfered until the process is finished. In emergency case, hospital mat not implement this policy. This case is omitted because referral always proceed to outpatient department.

III. Model Formulation

Queuing network of HHRS can be divided into three sub-sections, healthcare facility that will be viewed as M/M/n queuing system which is interconnected by referral, routing of referral that can be seen as HQS, and network of queues that follow BCMP model to categorize patient based on their disease.

A. Queuing Network Modeling Scheme

Referral system is modeled by using referral ratio of first-level healthcare system over East Surabaya and being referred into 3 most frequently referred hospital in East Surabaya, Haji hospital, Islamic hospital and Airlangga university hospital. After getting referral from first-level healthcare level patients will have different path according to their disease, i.e. patient with coughing disease will not enter surgical ward. Those disease then categorized into 8 wards, surgical (D-1), eye (D-2), ear, nose and throat abbreviated as ENT (D-3), internal disease (D-4), lung (D-5), cardiac (D-6), nerve (D-7), skin (D-8). The categorization follows BCMP network method.

Denote i as the index of disease categorization, i={1,2,…,9}, and h as the index of community healthcare center (CHC), h ={1,2,…,H}. Let r_h be referral ratio of the CHC-h which arrived at a rate of λ_h. p_{ih} is the proportion of patient diagnosed with disease-i on CHC-h. Define λ_i as arrival rate of referred patient with disease-i from all Surabaya. Then, by using BCMP method we get equation (3)

$$\lambda_i = \lambda_{i0} + \sum_{h=0}^{H} \lambda_h r_h p_{ih} \tag{3}$$

B. HQS model of referral system

Denote H_i as a hypercube model of ward i. Arrival rate of user into H_i is λ_i. The area of East Surabaya is divided into 18 area of districts as seen in Fig. 3. Denote j as the index for each district, j = {1,2,3,…,18}. Subset A_j of set H is defined as follows, the index h is a member of A_j if CHC-h is located within district-j. Define λ_{ij} as arrival rate of patient with disease-i from district-j then the value of λ_{ij} can be expressed as equation (4) and λ_i is a sum of λ_{ij} for all j added with λ_{i0}, external arrival rate from outside the region.

$$\lambda_{ij} = \sum_{h \in A_j} \lambda_h r_h p_{ih} \tag{4}$$

Denote k as the index of hospital, k = {1,2,3,…,K}. Only three hospital of east Surabaya is considered within the model. So the value of k is set to be 3. Hospital-1 refers to Haji hospital. Whereas, Hospital-2 refers to Islamic hospital. Lastly, Hospital-3 refers to Airlangga university hospital.

For each model H_i, the number of doctor handling with disease-i in hospital-k is represented as d_{ik}. If a hospital doesn't have any doctor handling disease-i or d_{ik} = 0 then hospital-k isn't considered as a server in model H_i.

In Larson, HQS is usually used in SDQ with moving server and static user. HHRS is an SDQ with moving user into static servers. Modification from previously stated HQS in his work must be done, especially in server location probability, preference and dispatch policy.

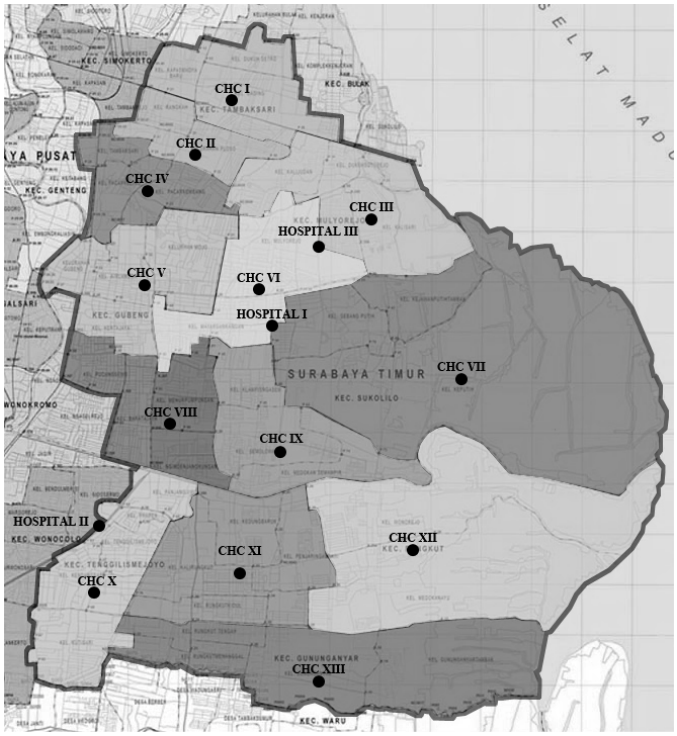

Fig. 3. Map of East Surabaya with Position of Each Healthcare Facility

Since those servers are static, probability of server location is 1 in the district where that server take place and 0 otherwise. This affects determination of travel time throughout region into those servers. A fixed travel time estimation can be obtained thus simplifying the model.

Preference of queue assignment can now be defined as the rank of hospitals according to user from district-k. The rank can be obtained from historical data of referrals and usually based on quality of service, doctor, equipment or travel time. A different unique rank can be assigned to each hospital.

By assigning different rank, for each state one can find exactly one best hospital for each upward state transition. When a patient about to be referred, dispatch policy selects the best ranked hospital for the current state. A tie cannot occur in this situation because every hospital ranked differently. Thus the rate of the transition equals to sum of departure rate from district which the hospital is most preferred on the previous state.

The modification did not change the model much as the state still represent busy/idle condition of server within the region. Transition between states occurs identically with fixed rates as stated before. However, the dispatch policy gives exact optimal server for every condition. Futhermore, overall performance analysis is simplified with fixed parameter.

C. Dynamic Routing With Patient Preferences

The patient is referred to a hospital based on the preference expressed on preference matrix. Denote P_i as the preference matrix for HQS model H_i. Preference matrix P_i have a size of 18x3. Entry P_{ijk} of the matrix contained in row-j and column-k represents preference rank of hospital-k from district-j users. Preference matrix contains the degree of user preference value, simply we use its rank according to each area. This is obtained from the proportion of referral from historical data.

Patients prefer a hospital which will give them faster services. That is, the waiting time on the system expressed on equation (1). This became the foundation of the preference function. As patients travel from their area to referred hospital, number of patient will increase according to the arrival rate of the hospital from the previously recorded value. The expected increase in waiting time is proportional to the increase in number of patient. Expected increase of patient equals to the product of travel time, T_{jk}, and arrival rate, λ_{ik}. This number is added into the function.

Users preference for a hospital is obtained from historical data so the value on P_i could be a referrence of it. Then, entry P_{ijk} represents preference rank of hospital-k from district-i users. In their preferred hospital, patients are assumed in willing to wait for a time constant of T hours. Time constant T is multiplied to the rank of the hospital, P_{ik}. This is then added to the function which expresses a contribution of preference in routing algorithm. In summary the preference function can be expressed as equation (5),

$$F_{ik} = \left(\frac{n_k + \lambda_{ik} T_{jk}}{\mu_{ik}} + P_{ik} T\right)^{-1} \quad (5)$$

Expected waiting time added with willingful preference waiting time will make a greater value on preferred hospital but still considering the expected waiting time. Author defines the three component of dynamic routing for referral system as follows,

1) Information policy, Information collected form the system are the number of queue. Arrival rate estimate will be estimated with hypercube approximation method.

2) Transfer policy, Patient will be referred to their first preference hospital. If congestion happens in said hospital then he/she will be transferred.

3) Position policy, Patient are to be referred to the hospital which has the greatest preference function value.

D. Priority on Patients

Patient is grouped into two category that is, severe and normal. Severe patient needs more service time but needs faster treatment. In this case, severe patient can pass the queue for a faster treatment. The priority type is non-preemptive as the doctor can't leave the patient currently undergoing a treatment.

Denote p_0 as proportion of severe patient. Denote μ_s as the initial service rate average of the doctor. Assume that severe patient is serviced with a Markovian service process which has rate $\mu_0 = c.\mu_s$,with some constant c. The average service rate of normal patient which is denoted as μ_1 must satisfy equation (6)

$$\mu_1 = \frac{(1 - p_0) c \mu_s}{c - p_0} \quad (6)$$

IV. Implementation

A. Simulation Model of HHRS in East Surabaya

Simulation model of HHRS is made on Matlab by using SimEvent toolbox of Simulink. He dynamic routing is inscribed on a script and used on user-defined function block. For simplicity one Simulink model is used solely for one disease. This is valid because a patient on one kind of disease wouldn't queue in a ward that his/her disease doesn't belong. The simulation can be divided into three parts, entity generation, dynamic routing and server on each hospital.

TABLE I. Queue Parameter of Hospitals in East Surabaya

Variable	Value of Ward-*i* (D-*i*)							
	D-1	***D-2***	***D-3***	***D-4***	***D-5***	***D-6***	***D-7***	***D-8***
μ_i	6.16	8.03	8.77	10.17	6.93	11.37	8.56	8.70
d_{i1}	4	2	2	2	1	2	2	2
d_{i2}	2	1	1	2	1	1	1	1
d_{i3}	8	2	2	3	3	2	2	2
d_i	14	5	5	7	5	5	5	5

Entities are generated based on referral proportion of the disease then multiplied to departure rate of the CHC. Server is modeled based on the average doctor treating patient in corresponding hospitals with service rate obtained from observation on the hospital. The condition of queue becomes a feedback to the dynamic routing that allocates patient to suitable hospital. Patient priority is simulated by randomly giving a label to a generated entity with probability of p_0 which is started from 5% of the population.

B. Parameter Used in Simulation

Parameters needed in the simulation consist of queue parameter of the hospital in east Surabaya, overall arrival rate which is departure rate from CHC and travel time between from districts into hospitals. Departure rate data is obtained

from BPJS historical records in 2015. Queue parameter such as, mean of service rate and number of doctor is obtained from author's survey. Table 1 shows the value obtained from the survey.

TABLE II. TRAVEL TIME FROM CHC TO HOSPITAL

District CHC	Travel time to … (minutes)		
	H1	*H2*	*H3*
CHC-1	13	28	13
CHC-2	19	7	21
CHC-3	16	21	19
CHC-4	21	28	14
CHC-5	5	21	3
CHC-6	14	20	15
CHC-7	6	17	10
CHC-8	11	20	13
CHC-9	22	26	24
CHC-10	8	25	8
CHC-11	17	29	18
CHC-12	8	15	15
CHC-13	27	25	30

Travel time is estimated with assistance of Google Maps. The location of each district is represented by the center point of that district. Set the point as origin and each hospital as the destination. The travel time is then obtained on the site. Table II shows the value obtained in this process.

V. RESULT AND DISCUSSION

Simulation shows that implementation of dynamic routing clearly improves the behavior of healthcare queue systems, especially in the waiting time and server utilization. Fig. 4 below shows the number of queue in those hospitals after implementation of dynamic routing. Timely condition of queue is also seen to be maintained.

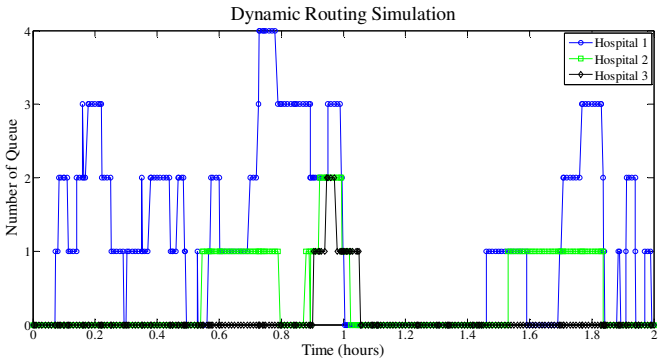

Fig. 4. Queue Condition of Dynamic Routing Simulation

Author then interested in the behavior of dynamic routing responding changes on the system. This raises a question of whether the routing still gives patients their preferred hospitals when their preference changed. Will the routing give valid referral in more crowded system. Is the routing applicable when handling differently prioritized patient. Sections below are discussing such interest.

A. *Effect of User Preference on System*

Patient preference is our main concern in creating dynamic routing thus it should affect the referral assignments accordingly. Preference of patients is altered to see this phenomenon. Initially, Hospital-2 is more preferred by patients because of its advanced equipment and experienced personnel. Then, we change the rank of Hospital-2 to be under Hospital-3 in certain districts numbered 1 until 9 because its distance is nearer. The arrival rates of both hospitals and in each ward are then compared. The comparison of change percentage is shown on table 1.

TABLE III. ARRIVAL RATE CHANGES RESPECT WITH PATIENT PREFERENCE

No	Ward	Percentage of Change	
		Hospital-2	*Hospital-3*
1	Surgical (D-1)	0.34	-1.30
2	Eye (D-2)	2.06	-4.05
3	ENT (D-3)	23.80	-36.43
4	Int. Disease (D-4)	8.42	-9.54
5	Lung (D-5)	59.62	-26.88
6	Cardiac (D-6)	357.76	-86.88
7	Nerve (D-7)	451.74	-93.35
8	Skin (D-8)	288.83	-85.43

Arrival rate to hospital-3 is increased and oppositely is for hospital-2, this complies with our assumption. It gives positive change to hospital-3 and negative change to hospital-2 due to rank drop. The dynamic routing is created considering patient rank of preference so it allocates the patient to their preferred hospital. This gives a proof of the ability of the dynamic routing to follow change of patient preference given a correct interpretation of preference rank.

On second note, different percentage of changes occurs between wards. Arrival on nerve ward gets the biggest change whereas surgical ward gets the smallest change. Initially, nerve ward is never a busy ward thus it has less patient to be referred. In this condition, the routing has a freedom to assign the patient according to their preference. On contrary, surgical ward is a busy ward with more patients to be referred. Thus, the routing can't freely assign those patients to their preferred hospital.

B. *Effect of Referral Ratio on the System*

An increase in referral ratio means more patients depart to hospitals from community health centers. Thus, overall arrival rate Referral ratio affects directly the arrival rate to the hospital. Testing is done by gradually increasing the parameter with a step of 10%. Workload or utility of the hospital increases as testing commences. Fig.5 shows the comparison on lung ward of the 3 hospital.

Ward utilization increases as the referral ratio increases. It follows because the system gets busier with an increasing patient departure. However, it maintains a proportional ratio throughout the change. This means the dynamic routing implemented in the referral system allocates patient equally maintaining a workload balance.

It can be seen that hospital-3 experiences greater increase than hospital-2 even though the latter should be higher in preference rank. This happens because hospital-3 has 3 lung

doctors whereas hospital-2 has only 1. Dynamic routing allocates the patient to capable hospital. Dynamic routing can allocate user to capable server regardless of preferences in order to maintain an optimal workload.

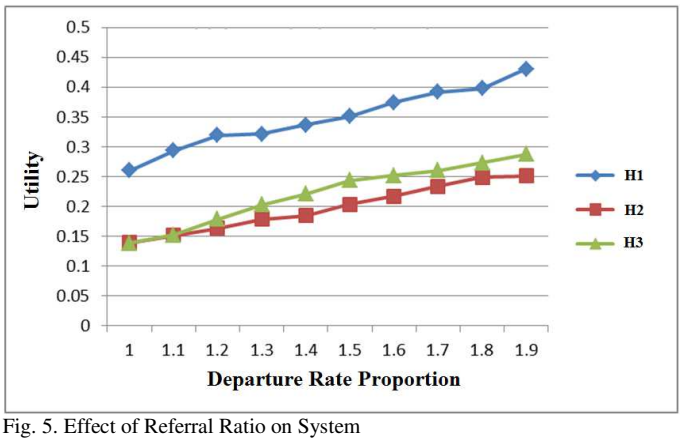

Fig. 5. Effect of Referral Ratio on System

C. *Effects of Patient Priority on System*

Priority on healthcare system will alter the order of the queue. By giving the same dynamic routing, it must give a bad allocation to the hospitals. Thus, the waiting time should be longer for certain patient. We are interested in how far such a dynamic routing withstands the proportion of severe patient.

Testing is done by increasing the proportion of severe patients gradually in 5% steps. Change is observed on surgical ward that have more patients and high activity level. The average waiting time of patient in the ward is then observed. Fig. 6 shows a comparison of waiting time in surgical ward of the 3 hospitals.

Increase of severe patient percentage also increases average waiting time on the ward. This happens because severe patient takes precedence over other. Thus, the waiting time for other patient is increased. Fig. 6 shows a big increase on hospital-3. Hospital-3 initially has faster waiting time. This shows the capability of dynamic routing for allocating patient with considering patient waiting time.

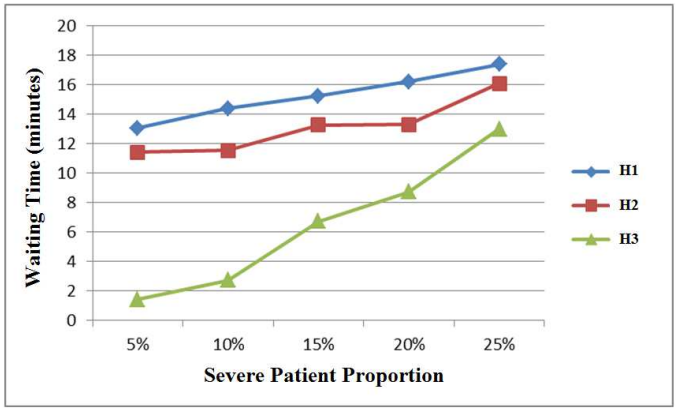

Fig. 6. Effects of Patient Priority on System

The dynamic routing needs an improvement in case a different priority imposed on the patient. Although the routing allocates the patient into less crowded hospital, a big increase in waiting time happens so the system behaves poorly. In non-emergency case, this patient prioritization is not implemented at least in Indonesia. Every referral patient in outpatient department is of the same priority. But in case a different priority imposed on the patient, some new routing algorithm should be developed.

VI. Conclusion

Healthcare referral system can be seen as a spatially distributed queue and its performance is estimated with HQS. Throughout the simulation author concludes that the implementation of dynamic routing improves waiting time of patient and maintains workload balance of the hospital. The dynamic routing took patients that initially would be referred into busy hospital and allocate them into idle hospitals.

The dynamic routing still preserves its capability to distribute patients in case of patient prioritization. Even so, individual queue performance became worse as the priority patient increased. This leaves a room for improvement in case such condition happened. Future work may include comparison of several routing algorithm on HHRS in term of efficiency. Furthermore, a suitable algorithm will be formulated to accommodate priority variation of patients.

References

[1] Dede S. 2015, 'Analisis Pelaksanaan Sistem Rujukan Rawat Jalan Tingkat Pertama (RJTP) pada Peserta BPJS Kesehatan di Puskesmas 5 Ilir dan Puskesmas Merdeka', Repositori Universitas Sriwijaya.

[2] Mengyu G., dkk. 2011, 'Effectiveness of Referral Incentive Policy: Exploring Using Queueing Network Model with Blocking', dalam International Conference on Service Systems and Service Management (ICSSSM), 2011, hal. 1-6.

[3] Cayirli, T. dkk. 2006, 'Designing Appointment Scheduling Systems for Ambulatory Care Services', Journal of Health Care Management Sciences, no. 9, hal. 47-58.

[4] Zhang Y., dkk., "A Performance Comparison of Adaptive and Static Load Balancing in Heterogeneous Distributed Systems," Proc. IEEE 28th Ann. Simulation Symp., pp. 332-340, Phoenix, Ariz., April 1995.Oliver, C.I. 2011, Fundamentals of Stochastic Network, Wiley, New Jersey.

[5] Larson, R.C. 1981, Urban Operation Research, Prentice-Hall, New Jersey.

[6] Trivedi, K.S. 2001, Probability and Statistics with Reliability, Queuing, and Computer Science Applications, 2nd Ed, Wiley, New Jersey.

[7] Watts, J. dan Taylor, S. "A Practical Approach to Dynamic Load Balancing," IEEE Trans. Parallel and Distributed Systems, vol. 9, no. 3, pp. 235-248, March 1998.

[8] Larson, R.C. 1974, Urban Emergency Service System: An Iterative Procedure for Approximating Performance Characteristics, The Rand Institute, New York.