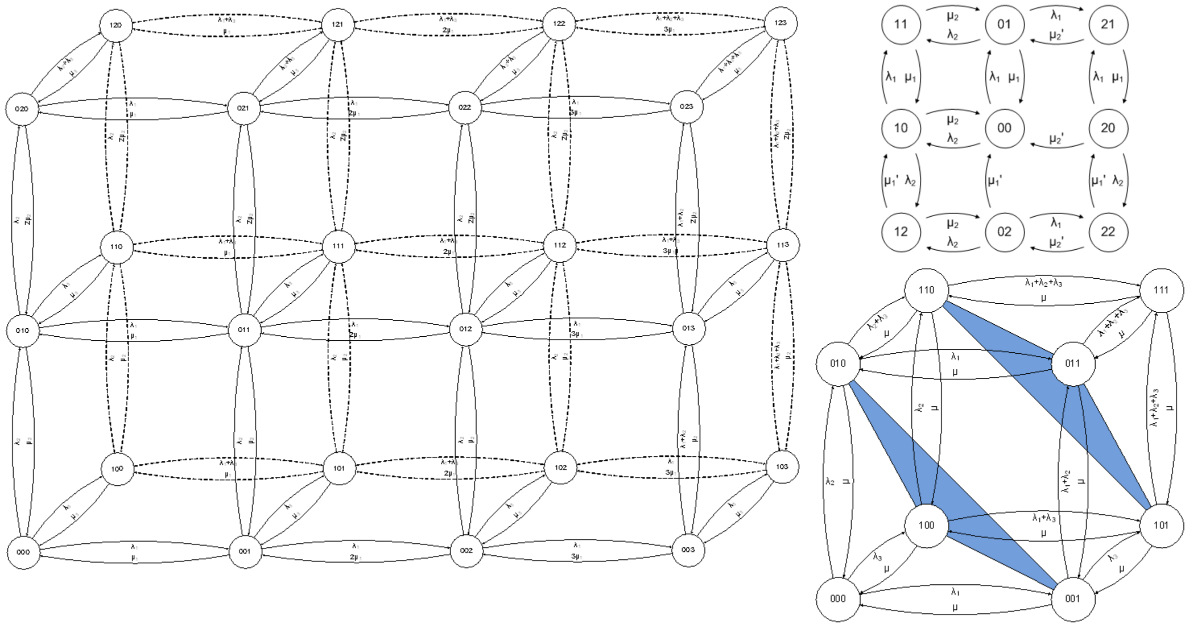
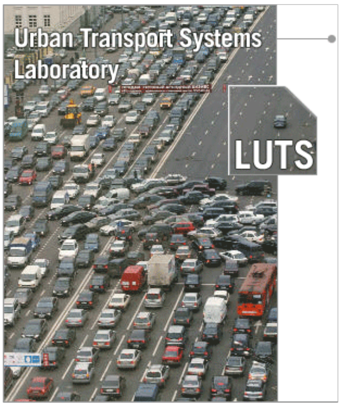


Extended Hypercube Models for Large Scale Spatial Queueing Systems

Burak Boyacı

Nikolas Geroliminis

STRC 2011 **May 2011**



ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

STRC 2011

Extended Hypercube Models for Large Scale Spatial Queueing Systems

Burak Boyacı
Urban Transport Systems Laboratory
Ecole Polytechnique Fédérale de Lausanne (EPFL)
GC C2 406, Station 18, 1015 Lausanne, Switzerland
phone: +41-21-69-32486
fax: +41-21-69-35060
burak.boyaci@epfl.ch

Nikolas Geroliminis
Urban Transport Systems Laboratory
Ecole Polytechnique Fédérale de Lausanne (EPFL)
GC C2 389, Station 18, 1015 Lausanne, Switzerland
phone: +41-21-69-32481
fax: +41-21-69-35060
nikolas.geroliminis@epfl.ch

May 2011

Abstract

Different than the conventional queueing systems, in spatial queues, servers travel to the customers and provide service on the scene. This property makes them applicable to emergency response systems (e.g. ambulances, police, fire brigades) and on-demand transportation systems (e.g. shuttle bus services, paratransit, taxis). The difference between the spatial queues and conventional queueing systems is various types of customers and servers and different service rates for different customer-server pairs. For the Markovian arrival and service characteristics, one of the methods to find system performance measures is to model and calculate steady state probability of the Markov chain for the hypercube queueing model (HQM) (Larson, 1974).

One of the obstacles on the way to apply HQMs to real life problems is the size of the problem; it grows exponentially with the number of servers and a linear system with exponential number of variables should be solved for each instance. In this research, in order to increase scalability of the problem, we propose two new models. In addition to that, we modeled the problem by using Monte Carlo simulation and tested the convergence and stability properties of the simulation results and compare them with stationary distributions.

Keywords

emergency response, spatial queues, hypercube queueing models, location models

1 Introduction

Emergency response systems are important for modern societies. They protect public health, provide assistance and ensure safety. Response areas of ambulances, design of police-beats or locations of fire brigades are important decisions for these systems. Although the demand rate is low on average for emergency response systems, the service availability is important when they are needed. In other words, in addition to adequate coverage, rapid and reliable response times are also important for emergency response systems.

One of the main issues that emergency response systems should cope with is the level of congestion in the service requests. Although on average the system utilization is not close to one, the probabilistic nature of the demand and service times can build congestions. The amount of congestion is directly related to the number of servers and the budget that is dedicated to these systems. But the smart strategic decisions have great effect on them as well. That is the reason, why we need scientific approaches consistent but also applicable. Clever allocation of the resources can improve the level of service without increasing the dedicated budgets.

On-demand transportation (also known as demand responsive transport, dial-a-ride transit) is an advanced, user-oriented form of public transport with flexible routing and scheduling of vehicles operating in shared-ride mode between pick-up and drop-off locations according to passengers needs. These systems provide service in areas with low passenger demand where regular bus service is not applicable. Shuttle bus services, paratransit, shared taxis and taxicabs are some types of on-demand transportation systems.

Although intelligent transportation systems technologies (e.g. signal priority, exclusive lanes, route guidance information) help on-demand transportation systems to work better, there is still need for efficient scheduling and dispatching strategies for these highly variant and congested systems. For instance, deciding the borders of sub-regions and number of paratransit vehicles needed in each region to maximize service rate with limited number of vehicles is an interesting and important question for these kinds of systems.

2 Literature Survey

The early models dealing with the location of emergency response systems assume deterministic demand. They ignored stochastic nature of the problem and dealt on coverage and median models.

Median problems locate the facilities on discrete candidate location set in such a way that minimizes average response time or distance. Hakimi (1964) proposed p-median problem in which the main aim is to locate p facilities on a finite set of candidate locations in such a way that minimizes total transportation cost of n customers. Although it is a combinatorial optimization problem, there are some exact algorithms (Galvão and Raggi, 1989; Körkel, 1989; Avella and Sassano, 2001) and heuristics (Daskin and Haghani, 1984; Schilling *et al.*, 1993) as well. There exists a recent survey, Mladenović *et al.* (2007), which covers most of the literature on meta-heuristics about this subject.

Coverage models are used to locate limited number of facilities (i.e. emergency response systems) in such a way to maximize total coverage. Toregas *et al.* (1971) proposed the *location set covering problem* in which the objective is to cover the entire area within a desired distance by minimum number of facilities. The *maximal covering location problem* (MCLP) which is proposed by Church and ReVelle (1974), maximizes coverage within a desired distance S by locating a fixed number of facilities. The probabilistic version of this problem, namely *maximum availability location problem* (MALP), the maximized value is the regions which are covered with α-reliability (Marianov and ReVelle, 1996). Daskin and Stern (1981) altered the MCLP and proposed a model named *backup coverage model* that maximizes the number of regions that are covered more than once. Gendreau *et al.* (1997) modified the backup coverage model with two time limits.

Although the literature mainly covers static and deterministic location models, in recent models uncertainty is also taken into account. This uncertainty can be either related to planning future periods (dynamic models) or input model parameters (probabilistic models). *Dynamic models* are suitable for models which, are considering the relocation of vehicles. The first article on this subject is written by Ballou (1968) in which the main aim is to relocate a warehouse in such a way that maximizes the profit in a finite horizon. Scott (1971) works with the extension of this problem with more than one facilities. Schilling (1980) extends MCLP with additional time constraint.

For urban problems, it is obvious that *probabilistic models* are the most suitable ones. For location and allocation of emergency response systems and other service on-demand vehicles (e.g. taxis), it is more convenient to model both the demands and the duration of the time the facility serving these demands with probabilistic models. In these models, with some probability, it is always possible to have demand which cannot be intervened by any facility, because of stochasticity in both demand and service times. Manne (1961), Daskin (1983), ReVelle and Hogan (1989) and, Marianov and ReVelle (1996) are some of the important articles written in this literature.

Larson (1974) proposed a *hypercube queueing model* (HQM) which is the first model that embeds the *queueing theory* in facility location allocation problems. This model analyzes systems such as emergency services (e.g. police, fire, ambulance, emergency repair), door-to-door pickup and delivery services (e.g. mail delivery, solid waste collection), neighborhood service centers (e.g. outpatient clinics, libraries, social work agencies) and transportation services (e.g. bus and subway services, taxicab services, dial-a-ride systems) which has response district design and service-to-customer mode (Larson and Odoni, 1981). The solution of this model provides state probabilities and associated system performance measures (e.g. workload, average service rate, loss rate) for given server locations. "The HQM is not an optimization model; it is only a descriptive model that permits the analysis of scenarios" (Galvão and Morabito, 2008). HQM models the current state as a continuous-time Markov process but does not determine the optimal configuration. Police patrolling (Sacks and Grief, 1994) and ambulance location (Brandeau and Larson, 1986) are two applications modeled by HQM. Marianov and ReVelle (1996) extended the MALP and developed *queueing maximum availability location problem.*

The first model proposed by Larson (1974) assumes that the service time is independent of the locations of the calls for service and the dispatched unit. This argument was supported by the idea that time spend on the road is negligible compared to service time. This can be a fact for fire brigades but not for the ambulances and on-demand vehicles. However even with this simplification, as number of servers (n) is increased, number of states (2^n) grows exponentially. That's why Larson (1975) also proposed a heuristic method because of this exponential behavior. As an extension, Atkinson *et al.* (2008) assume different service rates for each server in the system with equal interdistrict or intradistrict responses which increases number of states (3^n) significantly. Recently, Iannoni and Morabito (2007); Iannoni *et al.* (2008) embedded hypercube in a genetic algorithm framework to locate emergency vehicles along a highway. They extend the problem to enable multiple dispatch (e.g. more than one server can intervene for the same incident).Geroliminis *et al.* (2009, 2011) integrate the location and distracting decisions in the same optimization and solve the problem by using steepest descent (Geroliminis *et al.*, 2009) and genetic algorithms (Geroliminis *et al.*, 2011).

3 Hypercube Queueing Model

The conventional HQM models (Larson, 1974) include *hypercube* in the name since the transition graph of the continuous time Markov chain used to model this structure has a hypercube structure when the number of servers is more than three. State variables contain n binary variables which shows if server i is available (0) or busy (1). In other words, each state is a number in base 2 and each digit shows the state of the corresponding server. For each region which are called *atoms* (j) there exist a priority list of servers. Incidents in each region are served by the available server with the highest priority for this atom. If it does not exist any available server, either the call is lost (i.e. call for ambulance is dispatched by a backup) or joins a queue to be served (i.e. taxi customers are asked to be waited until there is one available), depending on the problem assumptions. Service requests arrive from each atom according to an independent

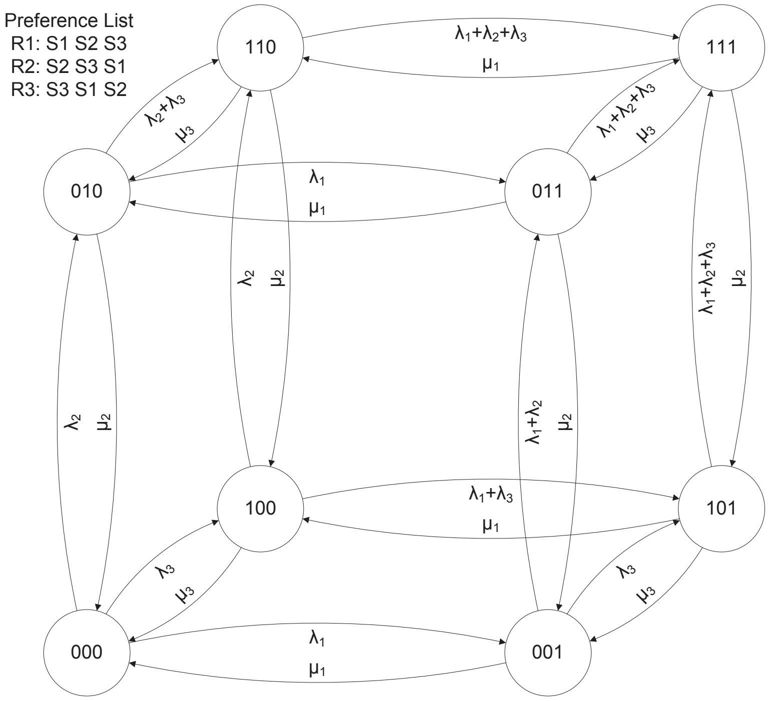


Figure 1: Larson (1974)'s HQM for three servers

Poisson process with parameter λ_j. Larson (1974) assumes each server has exponentially distributed equal service rates μ_i for any region. The transition graph of HQM with three states for this model can be seen in Figure 1. Note that, as the system gets full, in other words more servers get busy, the burden on the free servers increases. For instance in state "110" all the servers but the first are busy. That's why the next incident in any region will be served by the first server. This is also the reason of having high transition rate ($\lambda_1 + \lambda_2 + \lambda_3$) from state "110" to "111".

For different service rates for inter and intradistrict responses, the size of the model will increase. In this model we have three different possibilities for each server: available (0), busy with intradistrict response (1) and busy with interdistrict response (2) (Atkinson *et al.*, 2008). Figure 2 is a transition graph of an example with two servers. As an example "20", in which both regions are served by the servers of the other regions, represents the state where the first server is available and the second server intervenes an incident outside its own region.

It is good to note here that, the intradistrict server has always priority for the incidents inside its own region. We can see this in the transition diagram. When the system is empty, if there is an incident in a region, we cannot assign server from another region. This is also the case for Larson (1974)'s model. However, we should also note here that, this does not prevent having states such as "22". Although, practically it is rare for lightly congested systems, it has always a nonnegative probability in theory.

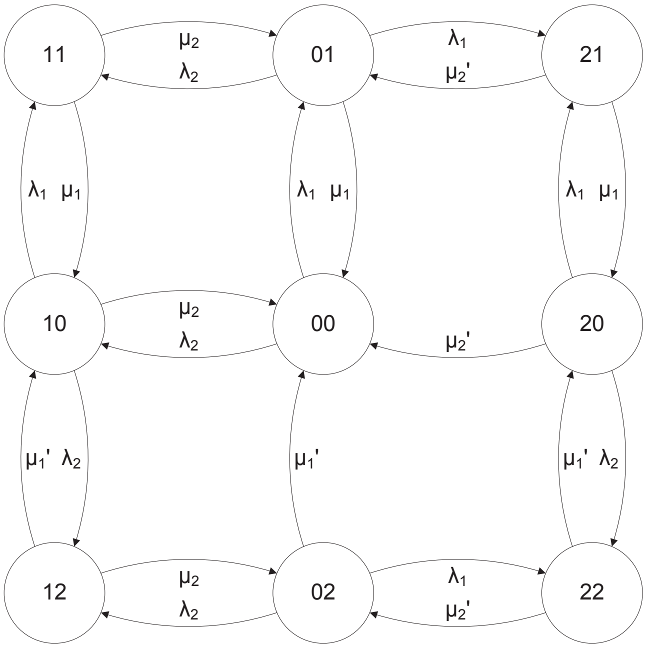


Figure 2: HQM for two servers with different inter (μ_i) and intradistrict (μ_i') service rates

4 Extended Hypercube Queueing Models

In the models proposed by Larson (1974) and its counterpart with different inter and intradistrict service rates, there are 2^n and 3^n states respectively which makes each limiting probability impossible to calculate, for even very small cases. For instance, if there are 20 servers, number of states for 2^n model has around one million whereas for the 3^n case this is more than three billions. In other words, we need to solve a system of equations with over million unknowns. One of the ways to get rid of this problem can be looking at the problem in an aggregate level.

In extended HQMs, we alter the conventional model such a way that more than one servers can be assigned to each region. We model the problem as a discrete location problem with queueing characteristics. In this new model, there are I types of servers which we call them *bins*. Each server in these bins serves their own customers and the rest with different service rates (μ_i and μ_i'). There is only one queue and this queue works in first in first out (FIFO) manner. Each customer who enters the system or leaves the queue to have service chooses the server which serves him with the maximum rate (which is given as priority list). We have n servers and our aim is to decide how many servers should we assign for each bin (n_i for $\forall i$) which will optimize the average performance of the system (e.g. minimize interdistrict response, loss rate and/or maximize average service rate). If the interarrival time of each customer and service times of each server had deterministic distributions, this model would be modeled as a simple discrete location allocation problem and could be solved by a linear programming formulation. However, we are interested in stochastic systems and this model is more appropriate for the probabilistic demand and service times.

At first glance the proposed model can be seen suboptimal for emergency response systems because, making responsible regions smaller and assigning one emergency response system for each region would give better results. We are restricting the region sizes and assigning more servers in each region which can give suboptimal solutions. However in the conventional HQM, number of states increases extremely fast and with this extension, model can be used to solve real life instances. Furthermore, for deciding the location and allocation of on-demand transportation systems, this approach is more convenient. Cities can be partitioned into regions and number of on-demand vehicles which will optimize the overall system performance can be assigned to these regions accordingly. Although this model has exponential number of states $((n_1 + 1)(n_2 + 1) \dots (n_I + 1))$ it is far less than the conventional hypercube models. As an example, a system of 3 bins with 9, 6 and 5 servers in each for different inter and intradistrict service rate case ($\mu_i \neq \mu'_i$ for $\forall i$) number of states would be 32340 whereas this number would be 420 if we assume equal service rates for inter and intradistrict responses ($\mu_i = \mu'_i$ for $\forall i$). Please note that for the same total number of servers, the conventional two models need over million states.

The first *extended hypercube queueing model* (EHQM) that we are proposing assumes equal intra and interdistrict service rates. Each number in the state name presents number of busy servers in this *bin*. For instance "132" stands for 2, 3 and 1 busy servers in the first, second and third bins respectively. An EHQM model contains $((n_1 + 1)(n_2 + 1) \dots (n_I + 1))$ states in which n_i is assigned number of servers in bin i and I is the total number of bins. The transition graph of an example with three bins which has 3,2 and 1 servers respectively in each bin is shown in Figure 3. Note that this model has 24 states, which is 64 for Larson (1974)'s hypercube model. By using the following transition graph we can write the transition equations for each state and can calculate steady state probabilities. For instance for the state "012" in which there are 2, 1 and 0 busy servers in the first, second and third bins we can write the following transition equation:

$$P_{012}(\lambda_1 + \lambda_2 + \lambda_3 + 2\mu_1 + \mu_2) = \lambda_1 P_{011} + \lambda_2 P_{002} + 3\mu_1 P_{013} + 2\mu_2 P_{022} + \mu_3 P_{112} \tag{1}$$

For different intra and inter-arrival service rates EHQM is not descriptive enough. For this reason, we are proposing another model (EHQM′) which has more states but differentiates the intra and interdistrict responses from each other. Note that it is also possible to construct a model which is more detailed with different service rates for each customer-server pairs but this might create excessive amount of states. That's why we have only two different service rates for each server, μ_i (intradistrict) or μ'_i (interdistrict). In addition to that, in the states, there are two variables for each server. A variable pair for each server shows number of intra and interdistrict responses separately. This new model has $\prod_i \binom{n_i + 2}{2}$ states.

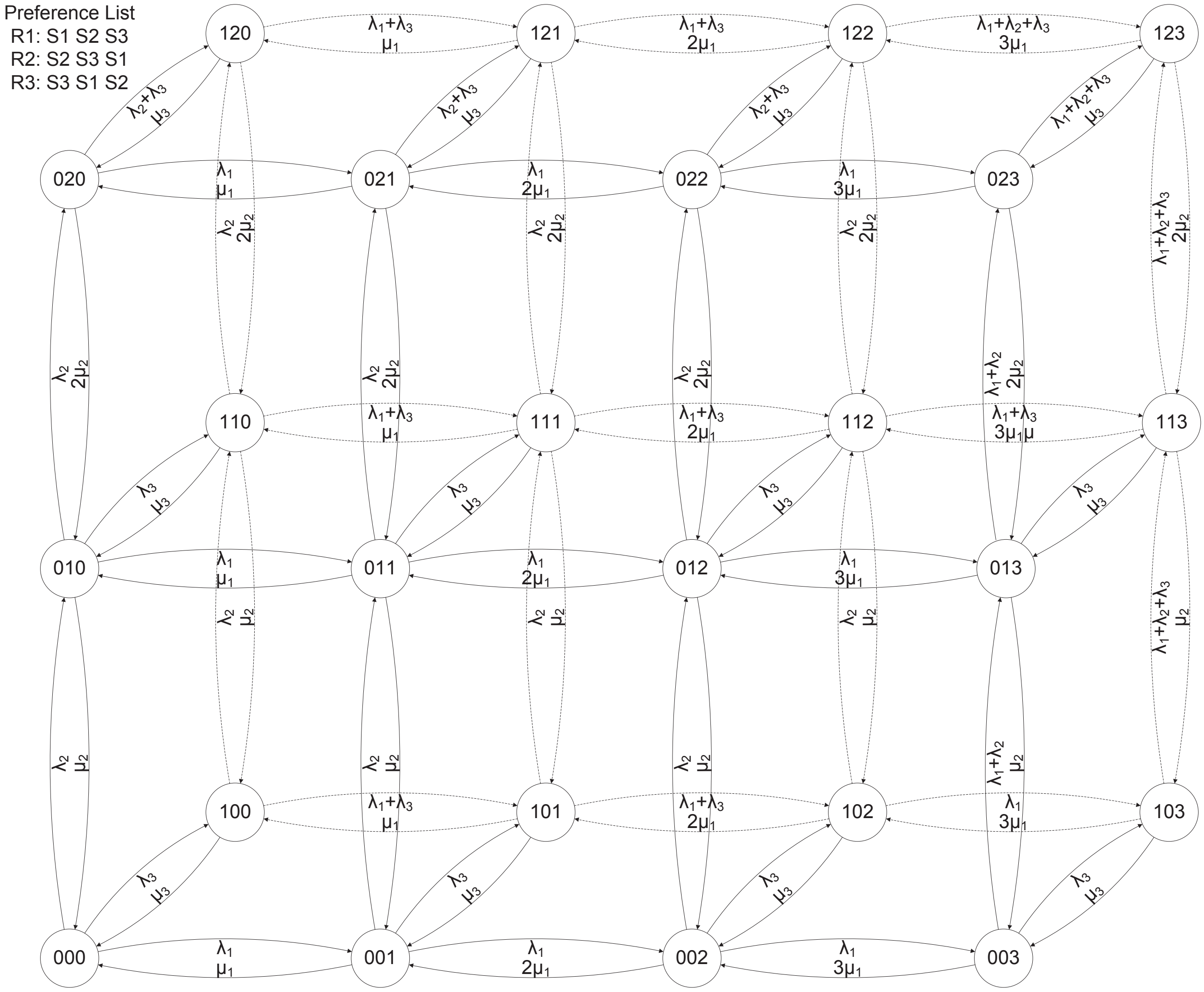


Figure 3: EHQM for three bins with equal inter and intradistrict service rates (μ_i)

In Figure 4 one can see a transition matrix of an (EHQM′) model which has 2 bins with 2 servers in each. Each row of the state name gives information about one of the servers. For instance in the state name "$\genfrac{}{}{0pt}{}{11}{10}$ ", first row shows the properties of the first server. Both of its servers are busy where the number on the left shows one of the servers serves intradistrict response and the number on the right shows one of them serves an interdistrict response. Only the number on the right is nonzero in the second raw which means there exists only one busy server in second bin which serves an intradistrict response. This state can also be seen separately with the states connected to it in Figure 5. The transition equation for the same state can be written as:

$$P_{\substack{11\\10}}\left(\lambda_1+\lambda_2+\mu_1+\mu_1'+\mu_2\right)=\lambda_1 P_{\substack{01\\10}}+\lambda_2 P_{\substack{11\\00}}+\mu_2' P_{\substack{11\\11}}+2\mu_2 P_{\substack{11\\20}} \qquad (2)$$

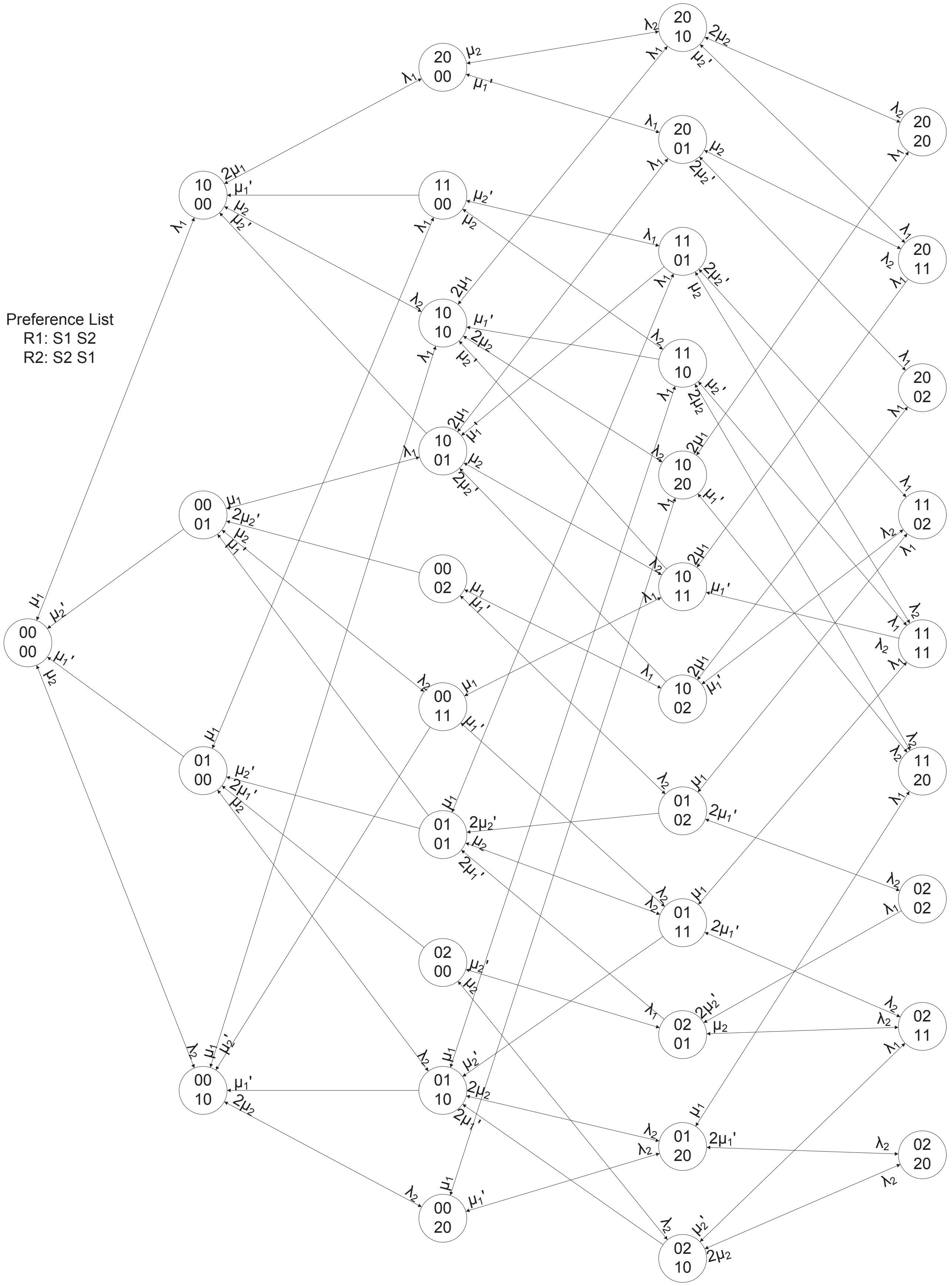


Figure 4: EHQM$'$ for two bins containing two servers in each bin with different intra (μ_i) and interdistrict(μ_i') service rates

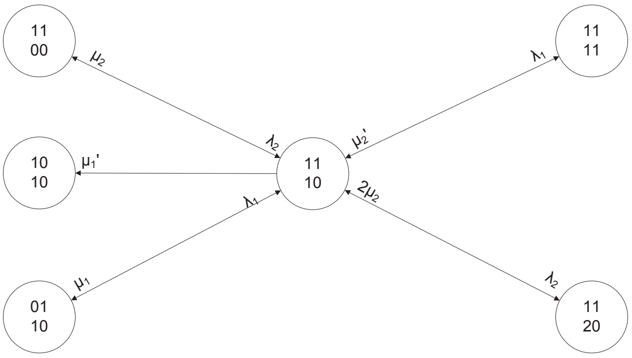


Figure 5: Single state with its connected states from EHQM′

5 Monte Carlo Sampling

Monte Carlo sampling is a class of computational algorithm that depends on repeated random sampling to compute the sample results. In our problem, we plan to utilize this algorithm to generate random arrival samples. Our main aim is to model spatial queueing systems as mixed integer linear programming problems. We plan to use the generated arrival samples with deterministic service time estimation to find where to locate the servers.

Before modeling the spatial queueing systems by using a mixed integer linear programming formulation, we need to be sure that the properties of the HQM can be represented by Monte Carlo sampling. For this purpose, a discrete time event simulation is created and the results acquired from the simulator are compared with the results of the HQMs. We compared both convergence and stability properties of the simulation. For the convergence check, we compared the convergence rate of the simulation to the probabilities calculated by solving Markov model of the HQM. Similar to that, for the stability performance of the method, we investigated the probabilities of the simulation after sharp and tense changes. The experiments give promising results and show that this method is applicable for any HQMs. Before going into details of these experiments, let us start with describing the simulation environment.

In these experiments we tested the 3^n aggregate HQM without queue. The main reason to work with a system without queue is the computation difficulties of 3^n aggregate models with queue. Different than 2^n aggregate model, 3^n aggregate model with queue has more than one tails. In other words, if we want to implement a 3^n aggregate HQM with queue, our problem size is hardly limited. In a system without queue, each customer is either started to have service from a server or left unserved as soon as they arrive the system. The interarrival distribution of each customer type and the service time distribution of each server for given customer type are predetermined. We have two events in the system: arrival and departure of a customer. As it is mentioned before, arrival is either served right after arrival occurs or lost and left the system without service.

At the beginning of the simulation we created arrivals from each type of customer. When simulation clock hits an arrival a new arrival event is created by using the given arrival distribution for the customer type and added to the event list. If it is served by a server, service time of the customer is also calculated and a departure event is added to the event list. For $\overline{*}$ is a value generated from the distribution $*$, the pseudocode of the simulation can be seen in Figure 6.

```
input: T (arrival end time), $A_j$ (random variable for interarrival of customer type $j$),
       $S_{ij}$ (random variable for service time of server type $i$ to customer type $j$),
       $P_j$ (priority list of server types for customer $j$), $n_j$ (number of servers of type $j$)

1. Initialize the event list $E = \{\}$, available server list $C_j = 0$ for $\forall j$
2. For each $j$
   (a) Generate arrival event $e$ of type $j$ with occurrence time $\overline{A_j}$
   (b) Add event $e$ to the event list
3. Repeat while there is an event in the list
   (a) Take the earliest event $e$ from the event list with time stamp $t$ and type $j$
   (b) If $e$ is an arrival event
       i. Generate arrival event $e^{\text{new}}$ of type $j$ with occurrence time $t + \overline{A_j}$
       ii. Add $e^{\text{new}}$ to the event list if its occurrence time is less than $T$
       iii. Let $i$ be the server who has the highest priority for customer type $j$
       iv. If $i$ is a number
           A. Generate a departure event $e^{\text{new}}$ of server type $i$ serving customer type $j$
              with occurrence time $t + \overline{S_{ij}}$
           B. Decrease $n_j$ by 1
       v. Else
           A. Increase loss customers count by 1
   (c) Else (if $e$ is a departure event of server type $i$ serving customer type $j$ )
       i. Increase $n_j$ by 1
```

Figure 6: Pseudocode for the discrete event simulation for HQM model without queue

The convergence rate of the system is investigated by comparing the simulation results with the calculated steady state probabilities from the Markov chain of the same HQM model. We have done the comparisons with plenty different scenarios and random number seeds. Here you can see the results from three scenarios with three different demand intensity. In the first scenario, we checked the convergence rate on conventional 3^n HQM with 8 servers. In the second and third examples, 3^n aggregate models with 4 bins are taken into consideration. The demand is almost equally distributed in the second scenario. Opposite to that, in third scenario we simulated an instance with different demand rates. The parameters of each scenario with average demand intensity can be seen in Table 1. In order to create heavily (lightly) congested systems, the demand rates are multiplied (divided) by two.

	bin	number of	intradistrict	interdistrict	region	demand	priority list
	ID	servers	service rate	service rate	ID	rate	of the region
Scenario 1	1	1	17	8	1	10	1 2 3 4 5 6 7 8 9
	2	1	16	9	2	11	2 3 4 5 6 7 8 9 1
	3	1	17	8	3	2	3 4 5 6 7 8 9 1 2
	4	1	18	7	4	10	4 5 6 7 8 9 1 2 3
	5	1	16	9	5	2	5 6 7 8 9 1 2 3 4
	6	1	19	6	6	12	6 7 8 9 1 2 3 4 5
	7	1	22	3	7	7	7 8 9 1 2 3 4 5 6
	8	1	21	4	8	6	8 9 1 2 3 4 5 6 7
	9	1	18	7	9	7	9 1 2 3 4 5 6 7 8
Scenario 2	1	3	21	4	1	28	1 2 3 4
	2	3	17	8	2	32	2 3 4 1
	3	4	22	3	3	44	3 4 1 2
	4	5	21	4	4	52	4 1 2 3
Scenario 3	1	1	21	4	1	12	1 2 3 4
	2	2	17	8	2	16	2 3 4 1
	3	5	22	3	3	60	3 4 1 2
	4	7	21	4	4	68	4 1 2 3

Table 1: Parameters of the three scenario

In Table 1, the first four columns are parameters related to bins and their servers and the rest three columns are related to regions (customer types). The last column related to the regions shows the priority list of the servers for given region. In other words, these numbers show in which order the servers are assigned to the customer given in the row. In order to keep problems understandable and not to confuse reader, we use simple priority lists.

The convergence graphs for the given scenarios can be seen in Figure 7. In these graphs, the x-axes show number of incidents (arrival or departure) whereas y-axes are the average absolute difference between steady state probability and the probability calculated from the beginning of the simulation. In each scenario we used two different service time distributions: Markovian and deterministic.

From the convergence graph given in Figure 7, it is seen that, the convergence rate of the scenario to the exact solution is quite fast. After 10000 incidents the difference between exact solution and the simulation solution becomes less than 1%. One of the other observations from these graphs is the effect of service time distribution to the system. It can be seen clearly that, the simulations with Markovian and deterministic service times have almost the same convergence curves to the exact solution. These two findings tell us that these spatial queueing problems can be modeled as mixed integer linear programming problems with an incident list of 5000 arrivals. Note that, this behavior is not specific to the scenarios represented, we observed the same results in almost all of the scenarios that we have tested.

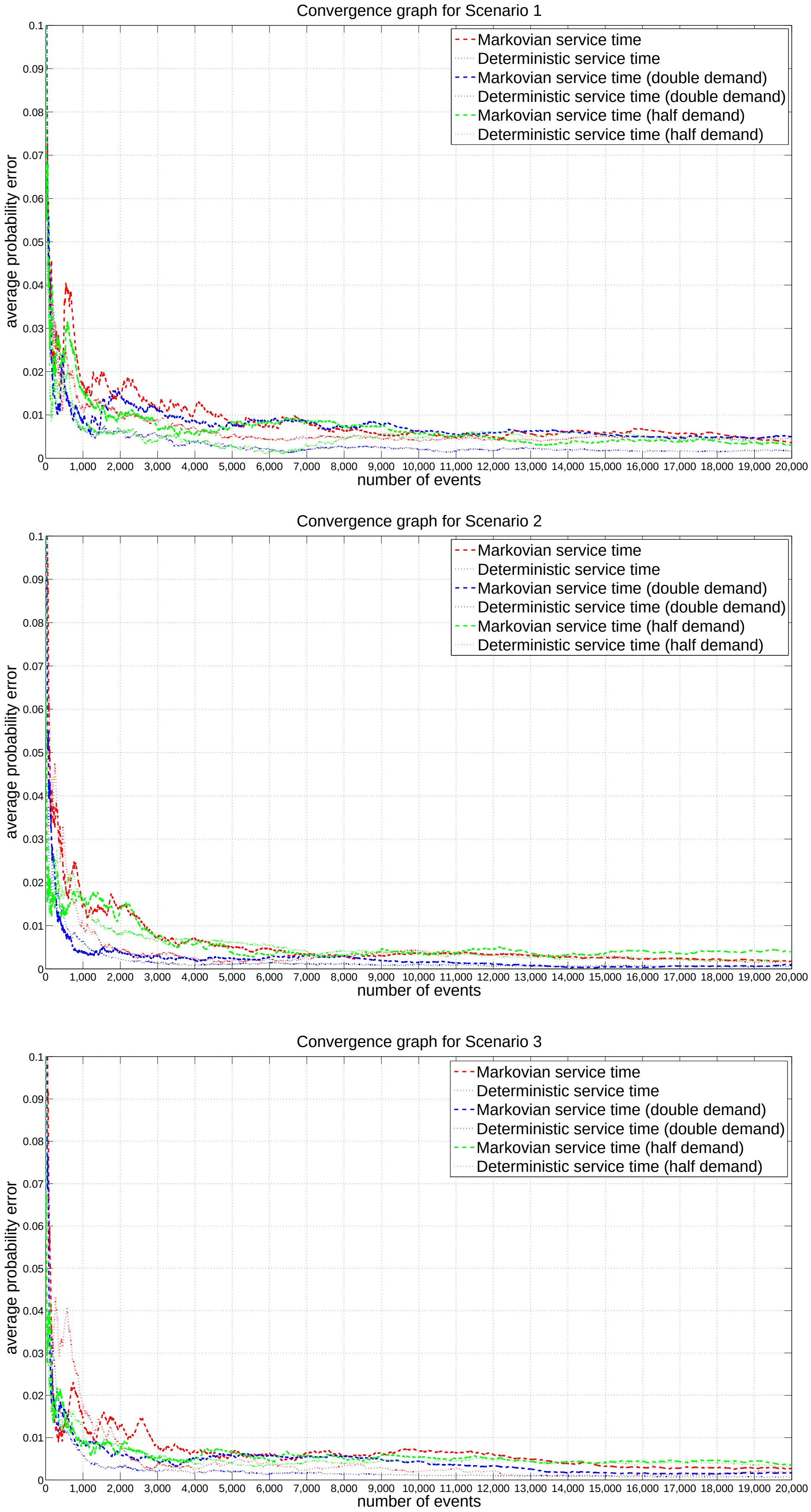


Figure 7: Convergence graphs of the three scenarios

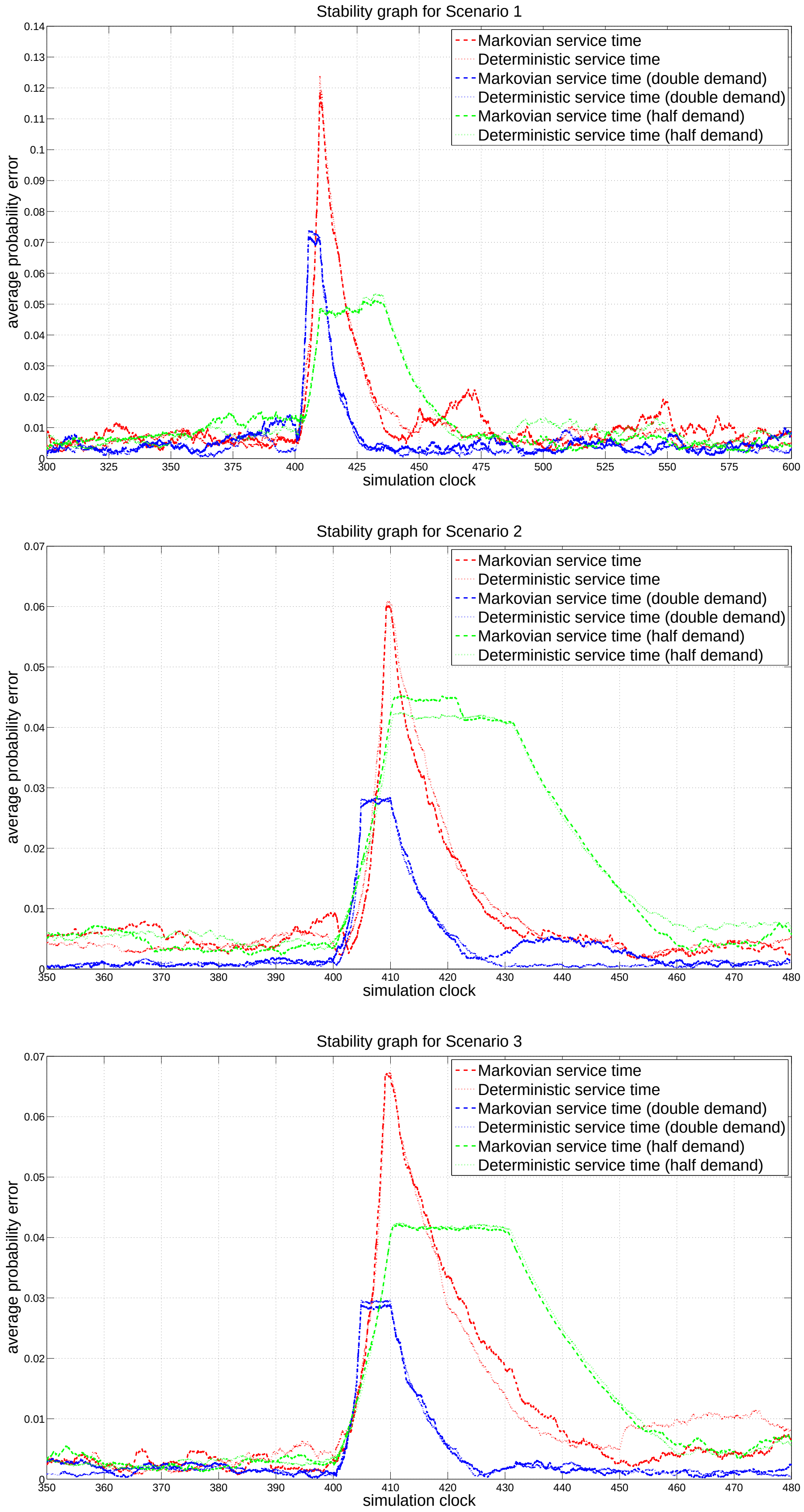
Not only convergence but also stability of the simulation for rapid changes is also investigated. In all three scenarios, demand rate of each region is multiplied by 5 between time interval 400 and 410. Different than the convergence graphs, in stability graphs, x-axes represent the time and the y-axes are the probability differences between the exact value and the value calculated from the scenario by taking recent time interval that contains 8000 incidents. The stability graphs for the three cases can be seen in Figure 8.

Note that the duration of the three peaks in Figure 8 is different. The explanation for this phenomenon is the following: In order to show all three scenarios in the same graph, we use simulation clock in x-axes. However convergence rate does depend on number of incidents not the simulation clock. If we check the peak durations in each graph, we will realize that, half demand intensity peak is twice as long as the normal demand intensity. This relationship is similar for normal and double demand intensity comparisons as well. In other words, if we convert the x-axes into number of iterations all three peaks have equal lengths. The convergence rate is a function of number of incidents not simulation clock.

We can also see from Figure 8 that the convergence of the system to the steady state probabilities after rapid changes is fast. In other words, simulation reacts the changes in the demand as soon as they occurs but recovers with the same pace; fluctuation in the demand effects the simulation but compensates this dramatic increase in the demand quite fast as well. This is a beneficiary property of the spatial queueing systems. Because of this property, HQM can be applied to demands with fluctuations.When a time-dependant smooth demand is applied, an HQM can be solved at each time step to identify the steady state performance measures, which will be close to reality as fast convergence shows. Last but not least, in these graphs, it can also be seen that the deterministic service time assumption gives very close results to the Markovian service time model.

6 Conclusion

In this paper we have investigated different models and solution methods for spatial queueing systems which have wide range of usage in the urban systems such as deciding the response areas of ambulances or paratransit vehicles. Although there are some hypercube queueing models exist in the literature, they are not applicable to real life problems because of their computational complexity. In Section 4, two new aggregate hypercube queueing models are proposed to tackle that issue. In Section 5, we investigate the applicability of Monte Carlo sampling and discrete event simulation to the hypercube queueing problems. The results showed that the system modeled by the simulation converges to the exact values quite fast and gives opportunity to use mixed integer linear programming formulations for problems related to spatial queueing models. It is also seen that the effect of service time distribution is minimal on the system and



Stability graph for Scenario 1

Stability graph for Scenario 2

Stability graph for Scenario 3

Figure 8: Stability graphs of the three scenarios

exponential distribution can be substituted with deterministic distribution for the service time. As a next step we have started to work on modeling of the hypercube queueing models with a method as such. Preliminary results are promising and highly applicable to real life problems.

References

Atkinson, J., I. Kovalenko, N. Kuznetsov and K. Mykhalevych (2008) A hypercube queueing loss model with customer-dependent service rates, *European Journal of Operational Research*, **191** (1) 223 – 239, ISSN 0377-2217.

Avella, P. and A. Sassano (2001) On the p-median polytope, *Mathematical Programming*, **89**, 395–411, ISSN 0025-5610. 10.1007/PL00011405.

Ballou, R. H. (1968) Dynamic warehouse location analysis, *Journal of Marketing Research*, **5** (3) pp. 271–276, ISSN 00222437.

Brandeau, M. and R. Larson (1986) Extending and applying the hypercube queueing model to deploy ambulances in boston, *TIMS Studies in Management Science*, **22**, 121–153.

Church, R. and C. ReVelle (1974) The maximal covering location problem, *Papers in Regional Science*, **32**, 101–118.

Daskin, M. (1983) A maximum expected covering location model: Formulation, properties and heuristic solution, *Transportation Science*, **17** (1) 48–70.

Daskin, M. and E. Stern (1981) A hierarchical objective set covering model for emergency medical service vehicle deployment, *Transportation Science*, **15** (2) 137–152.

Daskin, M. S. and A. Haghani (1984) Multiple vehicle routing and dispatching to an emergency scene, *Environment and Planning A*, **16** (10) 1349–1359.

Galvão, R. and R. Morabito (2008) Emergency service systems: The use of the hypercube queueing model in the solution of probabilistic location problems, *International Transactions in Operational Research*, **15** (5) 525–549, ISSN 1475-3995.

Galvão, R. and L. Raggi (1989) A method for solving to optimality uncapacitated location problems, *Annals of Operations Research*, **18**, 225–244, ISSN 0254-5330. 10.1007/BF02097805.

Gendreau, M., G. Laporte and F. Semet (1997) Solving an ambulance location model by tabu search, *Location Science*, **5** (2) 75–88.

Geroliminis, N., M. Karlaftis and A. Skabardonis (2009) A spatial queuing model for the emergency vehicle districting and location problem, *Transportation Research Part B: Methodological*, **43** (7) 798 – 811, ISSN 0191-2615.

Geroliminis, N., K. Kepaptsoglou and M. Karlaftis (2011) A hybrid hypercube - genetic algorithm approach for deploying many emergency response mobile units in an urban network, *European Journal of Operational Research*, **210** (2) 287–300, ISSN 0377-2217.

Hakimi, S. (1964) Optimum locations of switching centers and the absolute centers and medians of a graph, *Operations Research*, **12** (3) 450–459, ISSN 0030364X.

Iannoni, A., R. Morabito and C. Saydam (2008) A hypercube queueing model embedded into a genetic algorithm for ambulance deployment on highways, *Annals of Operations Research*, **157**, 207–224, ISSN 0254-5330. 10.1007/s10479-007-0195-z.

Iannoni, A. P. and R. Morabito (2007) A multiple dispatch and partial backup hypercube queuing model to analyze emergency medical systems on highways, *Transportation Research Part E: Logistics and Transportation Review*, **43** (6) 755 – 771, ISSN 1366-5545. Challenges of Emergency Logistics Management.

Körkel, M. (1989) On the exact solution of large-scale simple plant location problems, *European Journal of Operational Research*, **39** (2) 157 – 173, ISSN 0377-2217.

Larson, R. (1974) A hypercube queuing model for facility location and redistricting in urban emergency services, *Computers & Operations Research*, **1** (1) 67 – 95, ISSN 0305-0548.

Larson, R. (1975) Approximating the performance of urban emergency service systems, *Operations Research*, **23** (5) 845–868, September-October 1975.

Larson, R. and A. Odoni (1981) *Urban Operations Research*, Prentice-Hall, Englewood Cliffs, N.J.

Manne, A. (1961) Capacity expansion and probabilistic growth, *Econometrica*, **29** (4) 632–649, ISSN 00129682.

Marianov, V. and C. ReVelle (1996) The queueing maximal availability location problem: A model for the siting of emergency vehicles, *European Journal of Operational Research*, **93** (1) 110–120, ISSN 0377-2217.

Mladenović, N., J. Brimberg, P. Hansen and J. Moreno-Pérez (2007) The p-median problem: A survey of metaheuristic approaches, *European Journal of Operational Research*, **179** (3) 927 – 939, ISSN 0377-2217.

ReVelle, C. and K. Hogan (1989) The maximum availability location problem, *Transportation Science*, **23** (3) 192–200, August 1989.

Sacks, S. and S. Grief (1994) Orlando police department uses or/ms methodology new software to design patrol district, *OR/MS Today*, 30–42.

Schilling, D., V. Jayaraman and R. Barkhi (1993) A review of covering problems in facility location, *Location Science*, **1** (1) 25–55.

Schilling, D. A. (1980) Dynamic location modeling for public-sector facilities: A multicriteria approach, *Decision Sciences*, **11** (4) 714–724, ISSN 1540-5915.

Scott, A. (1971) Dynamic location - allocation systems: some basic planning strategies, *Environment and Plann*, **3** (1) 73–82.

Toregas, C., R. Swain, C. ReVelle and L. Bergman (1971) The location of emergency service facilities, *Operations Research*, **19** (6) 1363–1373.