Performance analysis and decomposition results for some dynamic priority schemes in 2-class queues

Prathamesh Mayekar, J. Venkateswaran, Manu K. Gupta and N. Hemachandra
Industrial Engineering and Operations Research, IIT Bombay, Powai, Mumbai - 400076, India

***Abstract*— Many device to device communication networks can be modelled by multi-class tandem queues. In many applications, it is desired to have different quality of service for various classes. This can be achieved by implementing dynamic priority across classes. Performance analysis is an important aspect in such multi-class tandem queueing models for resource allocation. In this paper, we analyse two important, relatively complex and analytically intractable performance measures, tail probability and switching frequency, for two class queueing system with two different (relative and earliest due date based) dynamic priority schemes across classes. Such a two class queueing system can be used to model voice and data calls in communication networks. A simulator is built to analyse such queueing systems and various observations are made. Based on computational evidence, it is conjectured that two stage exponential queueing network with two classes of customers is decomposable as far as mean waiting times are concerned when relative priority is used across classes to schedule the customers. Based on further experiments, it is conjectured that departure processes with relative dynamic priority are indeed Poisson in two class exponential queue. We also conduct relevant statistical analysis in support of the conjectures.**

***Index Terms*— Dynamic priority, simulation, tandem queues, multi-class queues, queuing network, tail probability, switching frequency, optimal control**

I. INTRODUCTION

Device to device (D2D) communication in cellular networks is defined as direct communication between two mobile users without traversing the Base Station (BS) or core network. D2D communication was first introduced in [1] to enable multihop relays in cellular network. Huge literature has evolved since then with other potential D2D use cases such as multi-casting, peer-to-peer communication, video dissemination, machine-to-machine (M2M) communication and so on (See [2], [3], [4] and [5]). A recent survey on such an important topics can be found in [6].

Many D2D communication problems can be modelled by multi-class tandem queueing network (See [7]). For example, packets can be either transmitted via the one-hop route or two-hop route and the resulting system can be formulated as a two stage tandem queueing model. In particular, two class tandem queueing network is useful to model voice and data calls in communication network. Performance analysis of such models is of significant interest in designing, implementation and resource allocation of these systems. In this paper, we analyse some relatively complex but important performance measures of such systems.

Quality of service is another important aspect in D2D communication problems. When quality of service has to be differentiated among customers, multi class queueing system appears to be the natural choice for modelling, where customers are arriving over time. By quality of service differentiation, we mean that one type of customers have preferential treatment over others. A well known queue discipline to provide service discrimination is to give strict priority. Strict priority discipline has a drawback as it does not provide enough degree of freedom for service discrimination to lower priority customers which may starve for service. There are different types of dynamic priorities proposed in literature to avoid such starvation. There are dynamic priorities based on delay [8], on numbers in each class [9], on due date [10] (see [11], [12] also). A two class discrete time queuing system with priority jumps is studied in [13]. These dynamic priorities can alleviate the starvation problem of one class of customers relative to the other by suitably choosing the dynamic priority parameter.

Mean waiting time expressions are known in the form of recursion for different types of dynamic priority. Other performance measures (tail probability, switching frequency, etc.) of such an important queueing system are not known analytically. It will be interesting to pursue research in analysing various other performance measures for such multi-class dynamic priority queueing system. Simulation is one of the important tool significantly used in literature to analyse complex dynamic random systems (see [14] and references therein). In this paper, we analyse tail probability and switching frequency for certain dynamic priority systems using simulation in case of two classes, single stage, single server queue.

Note that above discussion on service discrimination is with respect to single server queue. A natural extension of above problem will be a tandem queueing network which is more suitable for D2D communication. Some decomposition results for mean waiting time are known in single class exponential queueing network. These networks are called Jackson network in literature (see [15]). Some results with multi-class queues are also explored in literature (see Kelly network in [16]). In this paper, we conjecture (based on computations) a decomposition

Prathamesh Mayekar is an M.Tech. and Manu K. Gupta is a Ph.D. student at IE&OR, IIT Bombay and they are partially supported by a Teaching Assistantship offered by Government of India.

J. Venkateswaran and N. Hemachandra are with department of Industrial Engineering and Operations Research, IIT Bombay, Mumbai-400076, India.

E-mail addresses: {prathamesh220, jayendran, manu.gupta, nh}@iitb.ac.in

result for mean waiting time in a two stage queueing network when relative dynamic priority is implemented to schedule customers across classes.

Another community of researchers focused on dynamic control over multi-class queueing systems due to its various applications in computers, communication networks, and manufacturing systems. One of the main tools for such control problems is to characterize the achievable region for performance measure of interest, then use optimization methods to find optimal control policy (see [17], [18], [19]). Optimal control policy for certain non linear optimization problems for two class work conserving queueing systems is derived in [20]. In this paper, we solve certain optimal control problem with linear cost objective in two stage tandem queueing network assuming that the conjecture is true.

Departure process has attracted significant attention of researchers due to its applications in analysis of queueing networks. Earliest results are due to Burke [21] who showed that departure process of $M/M/1$ queue is Poisson. Results on departure process of multi class queues with strict priority are explored in literature (see [22]). Based on simulation, we further conjecture that the departure process of two class exponential queue is Poisson when relative dynamic priority is implemented across classes. Relevant statistical analysis (histograms and Q-Q plots) is performed to strengthen the conjecture.

This paper is organised as follows. Section II describes mechanism of relative priority and earliest due date based dynamic priority scheme. Section III presents performance analysis of tail probability and switching frequency. Section IV conjectures a decomposition result on mean waiting time for two stage exponential queueing network based on computations. An application of this decomposition result along with another conjecture on departure process is presented in same section. Conclusions and future avenues are discussed in Section V.

II. Dynamic Priority Schemes

A. Relative priority

This type of dynamic priority was first proposed by Moshe Haviv and Van Der Wal [9]. In this multi-class priority system, a *positive* parameter p_i is associated with each class i. Arrivals are assumed to be independent Poisson with rate λ_i. Service times can have any general distribution. If there are n_j jobs of class j on service completion, the next job to commence service is from class i with following probability:

$$\frac{n_i p_i}{\sum_{j=1}^{N} n_j p_j}, \quad 1 \le i \le N \tag{1}$$

Here N is the number of customer classes.

Mean waiting time expression for class i, $E(W_i)$, under this discipline is given by following recursion [9]:

$$E(W_i) = W_0 + \sum_j E(W_j)\rho_j \frac{p_j}{p_i + p_j} + \tau_i E(W_i), \quad 1 \le i \le N. \tag{2}$$

where $\tau_i = \sum_{j=1}^{N} \rho_j \frac{p_j}{p_i + p_j}$, $1 \le i \le N$, ρ_i is the load factor of class i and $W_0 = \sum_{i=1}^{N} \lambda_i \bar{x}_i^2/2$. $\bar{x}_i^2$ is the second moment of service time of class i. In particular for $N = 2$ and (without loss of generality) $p_1 + p_2 = 1$,

$$E(W_i) = \frac{1 - \rho p_i}{(1 - \rho_1 - p_2\rho_2)(1 - \rho_1 - p_1\rho_1) - p_1 p_2 \rho_1 \rho_2} W_0, \tag{3}$$

for $i = 1, 2$ where $\rho = \rho_1 + \rho_2$.

B. Earliest Due Date (EDD) Dynamic Priority

This type of dynamic priority across multiple classes was first proposed by Henry M. Goldberg [10]. Consider a single server queueing system with N number of classes similar to relative priorities. Each class i has a constant urgency number u_i (weights) associated with it. Without loss of generality, classes are numbered so that $u_1 \le u_2 \le \cdots \le u_N$. If a customer from class i arrives at the system at time t_i, he is assigned a real number $t_i + u_i$. Server chooses the next customer to go into service out of those present in the queue as the one with minimum value of $\{t_i + u_i\}$. The server is busy so long as customers are present in the system. Let W_r denotes the waiting time of a class r jobs. In steady state, $E(W_r)$ is given by [10]:

$$\begin{aligned} E(W_r) \quad &= \quad E(W) + \sum_{i=1}^{r-1} \rho_i \int_0^{u_r - u_i} P(W_r > t)dt \\ &\quad - \sum_{i=r+1}^{k} \rho_i \int_0^{u_i - u_r} P(W_i > t)dt \end{aligned} \tag{4}$$

for $r = 1, \ldots, N$. Here $E(W) = \frac{W_0}{(1-\rho)}$ and ρ_i is the traffic due to class i. In case of two classes, expected waiting time is given by Theorem 2 in [10]:

$$E(W_h) = E(W) - \rho_l \int_0^u P(T_h[W] > y)dy \tag{5}$$

$$E(W_l) = E(W) + \rho_h \int_0^u P(T_h[W] > y)dy \tag{6}$$

where index h and l are for higher and lower class priority. u_l and u_h are the weights associated with lower and higher classes, where $u = u_l - u_h \ge 0$. $T_h[W]$ is the limit of $T_h[W(t)]$ as $t \to \infty$ which is defined below. Let $W(t)$ be the total uncompleted service time of all customers present in the system at time t, regardless of priority. $W(t) \to W$ as $t \to \infty$.

$$T_h[W(t)] = \inf\{t' \ge 0;\ \hat{W}_h(t + t' : W(t)) = 0\}$$

where $\hat{W}_h(t + t' : W(t))$ is the residual workload of the server at time $t + t'$ given an initial workload of $W(t)$ at time t and considering input workload from class h only after time t.

III. Performance Analysis

In this section, we discuss the performance analysis of some relatively complex performance measures for two class single server queue when dynamic priority is implemented across classes. Note that only mean waiting time expressions are known analytically for dynamic priorities discussed in previous section. Other performance measures are also important for such multi-class queueing systems. We present the analysis of tail probability and switching frequency using simulation.

We built a simulator in SimPy for single server queue with two classes of customers. Relative priority or earliest due date based dynamic priority is implemented across classes. SimPy is process-based discrete-event simulation framework based on standard Python language. Logic of the simulator, simulation code and validation of simulator can be seen in technical report (See [23]). All simulation results presented in this section are for exponential service time and Poisson arrivals.

Simulation run time: Total simulation run time for relative priority and earliest due date based dynamic priority simulator is 11,000 time units with warm up period of 1,000 time units in all the experiments of this section.

A. Tail probability of waiting time

Tail probability of waiting time is one of the important performance measure for any queueing system. Tail probability of waiting time is the probability that the waiting time W_i is greater than some given tail value t, $P(W_i > t)\ \forall\ i \in\{1, 2\}$. This performance measure is often used for defining quality of service in various optimal control problems. We calculate the tail probability for waiting time of class 1 and class 2.

We use the random seed to be 1 in our simulators for all the analysis related to tail probability.

1) Relative priority: We study the change in tail probability while varying three different input parameters (tail value, arrival rate and dynamic priority parameter). Change in tail probability of waiting time for class 1 and class 2 customers with different tail values is shown in Figure 1. We consider parameter setting $\lambda_1 = 6,\ \lambda_2 = 8,\ \mu = 15,\ p_1 = 0.3$ and $p_2 = 0.7 = 1 - p_1$ and t is varying from 0.3 to 3. It is noted that tail probability of waiting time decreases monotonically as t increases.

Change in tail probability of waiting time of class 1 and class 2 customers by varying dynamic priority parameter can be seen in Figure 2. We consider the parameter setting $t = 0.8,\ \lambda_1 = 6,\ \lambda_2 = 8,\ \mu = 15$ and p_1 is varied from 0.1 to 0.9. It is noted that tail probability for class 1 customers decreases while it increases for class 2 customers. Tail probabilities of class 1 and class 2 customers are equal at $p_1 = 0.5$.

This phenomenon is intuitive. Increasing the dynamic priority parameter for class 1, will lead to lesser waiting time for class 1 and hence smaller tail probability. Similar comments can be made for class 2.

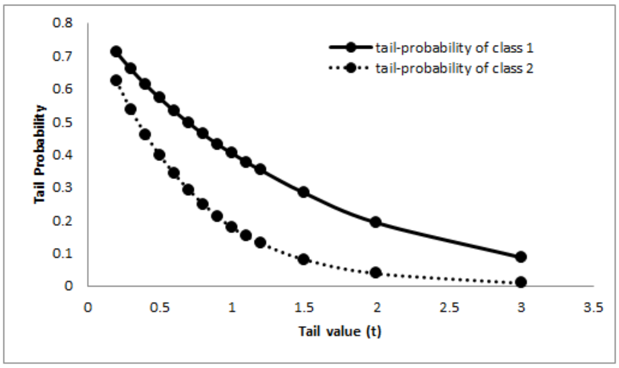

Fig. 1. Change in $P(W_1 > t)$ and $P(W_2 > t)$ by varying t

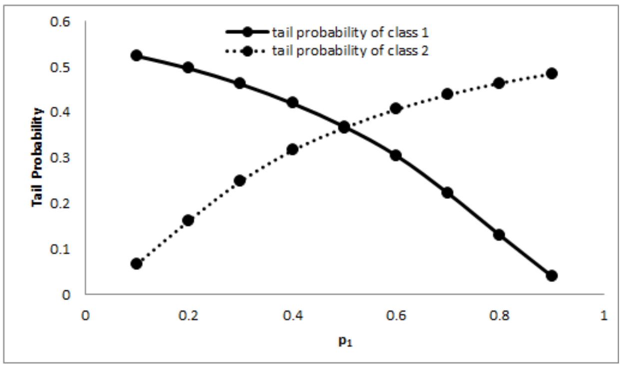

Fig. 2. Change in $P(W_1 > t)$ and $P(W_2 > t)$ by varying p_1

Change in tail probability of class 1 and class 2 customers by varying arrival rate of class 1 customers can be seen in Figure 3. We consider the parameter setting $t = 0.5,\ \lambda_2 = 8,\ \mu = 15,\ p_1 = 0.3\ p_2 = 0.7$ and λ_1 is varied from 1 to 6. It is noted that tail probability increases with λ_1. Tail probability for both classes increases with higher rate for higher values of λ_1.

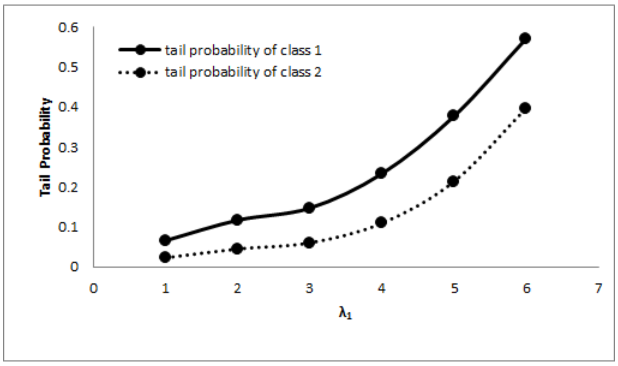

Fig. 3. Change in $P(W_1 > t)$ and $P(W_2 > t)$ by varying λ_1

2) EDD dynamic priority: In this section, we study the change in tail probability while varying the various parameters discussed in previous section for EDD based dynamic priority.

Figure 4, 5 and 6 show the variation in tail probability of waiting time with change in t, u_1 and λ_1 respectively. All the experiments are performed with $\lambda_1 = 6,\ \lambda_2 = 8,\ u_1 = 2,\ u_2 = 5,\ t = 0.2$ and $\mu = 15$. Tail value t, λ_1 and u_1 are

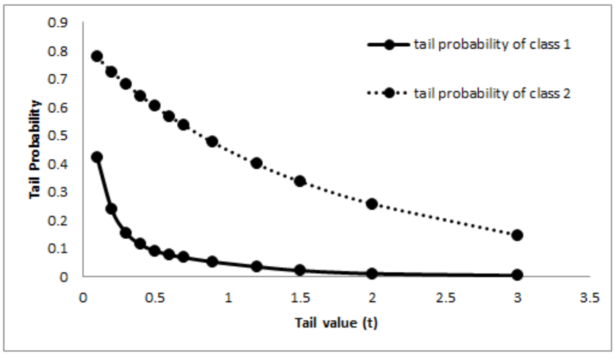

Fig. 4. Change in $P(W_1 > t)$ and $P(W_2 > t)$ by varying t

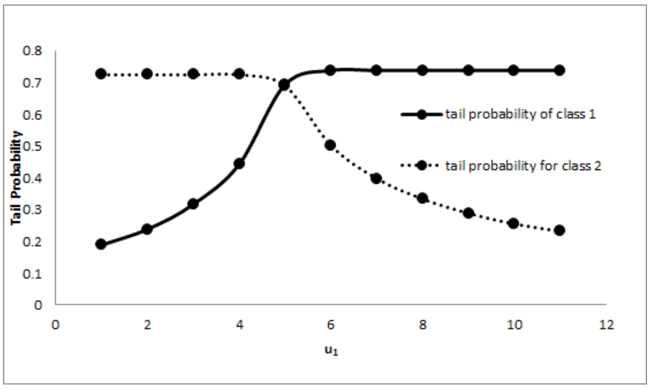

Fig. 5. Change in $P(W_1 > t)$) and class 2 $P(W_2 > t)$ by varying u_1

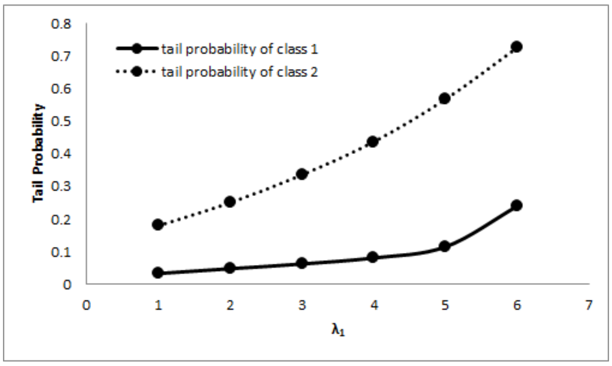

Fig. 6. Change in $P(W_1 > t)$ and class 2 $P(W_2 > t)$ by varying λ_1

varied appropriately as shown in different figures. Observations similar to the case of relative priority about tail probability can be noted. One of the interesting observation in Figure 5 is tail probability remains constant for certain range of u_1. Figure 4 and 6 show the monotonic behaviour of tail probability with change in tail value t and λ_1 respectively.

B. Switching frequency

An important aspect of multi-class queueing system is switching frequency. This performance measure is important when there is cost associated with each switch. Hence we investigate the switching frequency of relative and earliest due date based dynamic priority. Switching frequency is defined as the ratio of number of service switches among different classes the server undergoes divided by the total number of customers served. Thus switching frequency always lies in $[0, 1]$. We kept the random seed to be 1 for all the analysis concerning switching frequency in our simulators.

1) Relative priority: We study the change in switching frequency while varying dynamic priority parameter p_1. We consider parameter setting $\lambda_1 = 4$, $\lambda_2 = 8$, $\mu = 15$. It is noted in Figure 7 that switching frequency poses a nice geometric structure (concave function) for relative dynamic priority. Switching frequency is maximum at $p_1 = 0.5$.

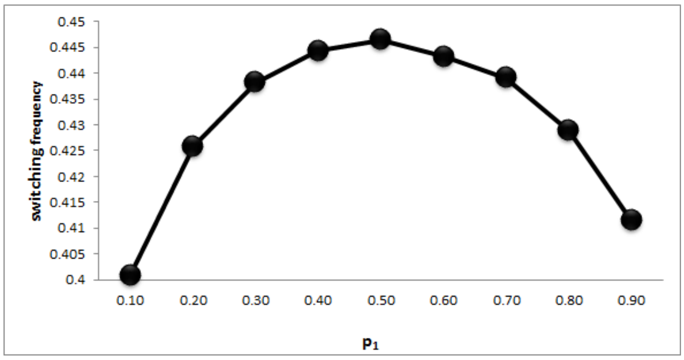

Fig. 7. Switching frequency Vs p_1

2) Earliest Due Date: We study the change in switching frequency while varying early due date parameter u_1 and keeping $u_2 = 50$ (constant). We consider parameter setting $\lambda_1 = 4$, $\lambda_2 = 8$, $\mu = 15$. Switching frequency initially remains constant as u_1 increases, after a certain threshold value (less than u_2) switching frequency starts monotonically increasing. Switching frequency is maximum at $u_1 = u_2$ (global first come first serve), then it starts monotonically decreasing till a threshold value after which it remains constant (see Figure 8).

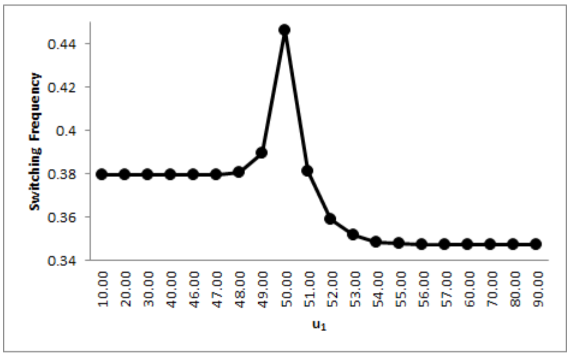

Fig. 8. Switching frequency Vs u_1

IV. A Decomposition Result

In this section, we describe a decomposition result for mean waiting time in two stage exponential queueing network based on simulation. We discuss applicability of this decomposition result in certain optimal control problems. Analytical results of departure process are known only for static priority in multi class queues, to the best of our knowledge. Based on computations, we further conjecture a result on departure process of two class queueing system when relative dynamic priority is implemented across classes.

Consider a two stage queueing network with two classes of customers. Let λ_1 and λ_2 be the independent Poisson arrival rates for class 1 and class 2. Let μ_1 and μ_2 be the exponential service rates at stage 1 and stage 2 respectively as shown in Figure 9. Note that service rate is same for both the classes in each stage. Consider the relative priority parameter as p_{ij} for $i = 1,\ 2$ and $j = 1,\ 2$. Here index i and j represent the class and stage respectively in a two class two stage queueing network. It follows from definition of relative priority that $p_{1j} + p_{2j} = 1$ for $j = 1,\ 2$.

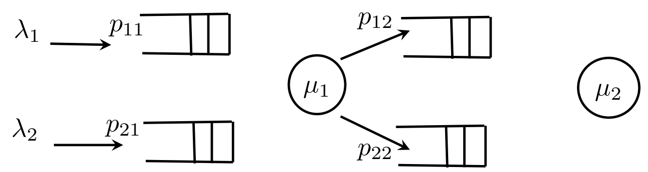

Fig. 9. A two class two stage queueing network

A simulator for two stage queueing network described above is build in SimPy (simulation package in python) [24]. SimPy code and logic for simulator can be seen in technical report (see [23]). Simulation run time is taken as 11,000 time units with a warmup period of 1,000 time units in all the experiments except for the heavy traffic in Table VI. It is noted in experiments that performance measures do not change significantly if one increases the run length. Hence this fixed run length is enough to produce accurate results. Five replications of each experimental setting are performed to capture the effect of random numbers used in simulator. Half width is calculated using t-distribution at 95% level of confidence in all experiments.

Simulator is validated using following observations for two stage tandem queueing network. Details are available in technical report [23].

- We compare simulated mean waiting time with theoretical mean waiting time (see [9]) in the first stage and note that the difference is not statistically significant.
- Arrival and departure rates are same in any stable queueing system. This phenomenon is observed in simulation and it is a necessary condition for the validation of two stage queueing network simulator.

A. Main result

We conjecture the following result based on our computation. We did experiments with low, moderate and heavy traffic rate with different relative dynamic priority parameters.

Conjecture 1: A two class two stage queueing network with Poisson arrivals is decomposable in terms of mean waiting time when relative priority is used across classes in both stages for exponential service time distribution at first stage.

Above decomposability result implies that the mean waiting time at second stage is independent of relative priority parameters (p_{11}, p_{21}) and service rate at first stage (μ_1). Mean waiting time of the second stage is the mean time customer spends in the second stage queue after finishing service from first stage. Mean waiting time at the second stage can be calculated by using theoretical mean waiting time expression (see Equation (3)) while using the parameters (p_{12}, p_{22} and λ_1, λ_2) at second stage only. Note that theoretical mean waiting time value at second stage is calculated without knowing if the arrival process at second stage is Poisson which is necessary for using Equation (3). It turns out that simulated mean waiting time matches with the theoretical mean waiting time at second stage (see simulation results in this section). We further conjecture in Section IV-B that the departure process from stage 1 is Poisson.

First of all, we present some simulation results in the support of above conjecture. In these results, we compute the difference between simulated and theoretical mean waiting at both stages and for class 1 and class 2.

We vary the relative priority parameter while keeping moderate traffic intensity. Input parameter setting is as: $\lambda_1 = 4,\ \lambda_2 = 8,\ \mu_1 = 15$ and $\mu_2 = 16$ (see Table I, II and III).

We further did the experiment by varying arrival rates λ_1 and λ_2 while keeping the dynamic priority parameters fixed at $p_{11} = 0.5$, $p_{12} = 0.4$ and service rates fixed at $\mu_1 = 15$, $\mu_2 = 16$ (See Table IV and V).

We also did some experiments with heavy traffic with parameters as $p_{11} = 0.8$, $p_{12} = 0.9$, $\mu_1 = 15$, $\mu_2 = 14.5$, $\lambda_1 = 6$ and $\lambda_2 = 8$. Simulation run time is increased from 11000 to 100000 with a warmup period of 20000 (See Table VI). Some more simulation experiments are performed with different parameter settings and similar phenomenon as in Table I-V is noted (see technical report [23]).

TABLE I
RESULTS WITH DYNAMIC PRIORITY PARAMETERS $p_{11} = 0.1$ AND $p_{12} = 0.8$

Replication	Differences at stage 1		Differences at stage 2	
	Class 1	Class 2	Class 1	Class 2
1	0.0057	0.0006	0.0159	-0.0139
2	-0.0172	-0.0003	0.0121	-0.0145
3	0.0031	0.0018	0.0147	-0.0077
4	-0.0016	-0.0003	0.0174	-0.0054
5	0.0299	0.0059	0.0160	-0.0039
mean	0.0040	0.0016	0.01527	-0.0063
halfwidth	0.0212	0.0032	0.00246	0.0119

It is noted that zero lies in the confidence interval (mean ± half width) of the difference between simulated and theoretical mean waiting time in all tables except Table III. However, all simulated mean waiting times are very close to theoretical mean waiting time and the maximum difference between simulated

TABLE II

RESULTS WITH DYNAMIC PRIORITY PARAMETERS $p_{11} = 0.3$ AND $p_{12} = 0.6$

Replication	Differences at stage 1		Differences at stage 2	
	Class 1	Class 2	Class 1	Class 2
1	0.0078	0.0064	0.0014	-0.0112
2	-0.0073	-0.0002	0.0088	-0.0009
3	0.0068	0.0051	0.0082	-0.0003
4	-0.0046	-0.0003	-0.0025	-0.0139
5	-0.0058	-0.0014	0.0155	0.0065
mean	-0.0006	0.0019	0.0063	-0.0040
halfwidth	0.0091	0.0045	0.0088	0.0105

TABLE III

RESULTS WITH DYNAMIC PRIORITY PARAMETERS $p_{11} = 0.8$ AND $p_{12} = 0.9$

Replication	Differences at stage 1		Differences at stage 2	
	Class 1	Class 2	Class 1	Class 2
1	0.0056	0.0217	-0.0049	0.0037
2	0.0007	0.0045	-0.0056	0.0117
3	0.0023	0.0050	-0.0045	0.0010
4	-0.0002	0.0065	-0.0039	-0.0039
5	-0.0015	-0.0027	-0.0026	0.0079
mean	0.0014	0.0070	-0.0044	0.0041
halfwidth	0.0035	0.0112	0.0014	0.0075

TABLE IV

RESULTS WITH ARRIVAL RATES $\lambda_1 = 1$ AND $\lambda_2 = 1$

Replication	Differences at stage 1		Differences at stage 2	
	Class 1	Class 2	Class 1	Class 2
1	0.00026	0.00009	0.00039	0.0002
2	-0.00029	-0.00053	-0.00045	-0.0001
3	0.00043	0.00067	0.00025	0.0010
4	0.00053	0.00071	0.00002	-0.0039
5	0.0002	-0.00042	-0.00038	0.0079
mean	0.000229	.000105	-0.000034	0.000147
halfwidth	0.000400	0.000729	0.000469	0.000216

TABLE V

RESULTS WITH ARRIVAL RATES $\lambda_1 = 4$ AND $\lambda_2 = 1$

Replication	Differences at stage 1		Differences at stage 2	
	Class 1	Class 2	Class 1	Class 2
1	-0.0015	0.0002	0.0004	-0.0003
2	0.0012	-0.0006	0.0011	0.0009
3	0.0009	0.0013	0.0002	-0.0003
4	-0.0009	-0.0009	-0.0001	-0.0012
5	0.0014	0.0016	0.0016	0.0019
mean	0.00019	0.00056	0.00066	0.00020
halfwidth	0.00162	0.00124	0.00087	0.00147

and theoretical waiting time is of the order 10^{-1} with respect to theoretical waiting time. Hence the difference between theoretical and simulated waiting time is statistically insignificant in all table except Table III. We conclude that such a two stage tandem queueing network is decomposable with mean waiting time as described in conjecture (except for moderate traffic cases similar to the example in table III in which one particular customer class has high priority across both stages).

TABLE VI

RESULTS WITH $p_{11} = 0.8$, $p_{12} = 0.9$, $\mu_1 = 15$, $\mu_2 = 14.5$, $\lambda_1 = 6$ AND $\lambda_2 = 8$ (HEAVY TRAFFIC)

Replication	Differences at Stage 1		Differences at Stage 2	
	Class 1	Class 2	Class 1	Class 2
1	-0.0033	-0.0265	-0.0207	-0.0894
2	-0.0053	-0.0133	-0.0403	-0.2671
3	-0.0046	-0.0193	-0.0020	0.0547
4	0.0008	0.0048	0.0175	0.2585
5	-0.0106	-0.0448	-0.0252	-0.1466
mean	-0.0046	-.0198	-0.0141	-0.0380
halfwidth	0.0050	0.0225	0.0277	0.2510

1) An application: In this section, we discuss an application of above decomposition result in solving an optimal control problem for two stage queueing network. Consider the linear waiting time cost minimization problem as follows:

$$\textbf{P1:} \quad \min_{\mathcal{F}} \quad c_1\bar{W}_1 + c_2\bar{W}_2$$

where $\bar{W}_1$ and $\bar{W}_2$ are mean waiting time of class 1 and class 2 respectively. Let c_1 and c_2 be the cost associated with class 1 and class 2 respectively and $\mathcal{F}$ be the set of all non pre-emptive, non anticipative and work conserving scheduling policies. It is proved in [25] that relative dynamic priority is *complete* for single stage and two class queues. In other words, optimizing over relative priority is equivalent to optimizing over set of all non pre-emptive, non anticipative and work conserving scheduling policies. Hence optimizing over $\mathcal{F}$ is equivalent to optimizing over relative priority and equivalently problem **P1** can be rewritten as:

$$\min_{p_{11},\ p_{12}} \quad c_1\bar{W}_1 + c_2\bar{W}_2$$

Note that $\bar{W}_1$ and $\bar{W}_2$ are functions of p_{11} and p_{12}. By using above decomposition result and assuming that the conjecture 1 is true, we have

$$\bar{W}_1(p_{11}, p_{12}) = \bar{W}_{11}(p_{11}) + \bar{W}_{12}(p_{12})$$

$$\bar{W}_2(p_{11}, p_{12}) = \bar{W}_{21}(p_{11}) + \bar{W}_{22}(p_{12})$$

Here $\bar{W}_{ij}$ for $i, j = 1, 2$ is mean waiting time for class i in stage j. Above optimization problem can be rewritten as:

$$\min_{p_{11},\ p_{12}} \quad c_1(\bar{W}_{11}(p_{11}) + \bar{W}_{12}(p_{12})) + c_2(\bar{W}_{21}(p_{11}) + \bar{W}_{22}(p_{12}))$$

which can be simplified as

$$\min_{p_{11},\ p_{12}} \quad c_1\bar{W}_{11}(p_{11}) + c_2\bar{W}_{21}(p_{11}) + c_1\bar{W}_{12}(p_{12}) + c_2\bar{W}_{22}(p_{12})$$

or

$$\min_{p_{11}} \quad c_1\bar{W}_{11}(p_{11}) + c_2\bar{W}_{21}(p_{11}) +$$

$$\min_{p_{12}} \quad c_1\bar{W}_{12}(p_{12}) + c_2\bar{W}_{22}(p_{12})$$

It is clear that the original optimization problem **P1** can be decomposed in above simpler optimization problem. Also note that solution of above two decomposed optimization problems is well known as $c\mu$ rule in literature (see [26]). Hence optimal control policy for optimization problem **P1** can be simply obtained.

B. Departure process of relative priority queue

We perform some more experiments with inter departure time and further conjecture a result on departure process which is stronger variant of the Conjecture 1. We did experiments with low, moderate and heavy traffic.

Conjecture 2: A two class single stage exponential queueing system has Poisson departure process when relative dynamic priority is implemented across classes.

Histograms: We present some simulation results in the support of above conjecture in Figure 10 and 11. We plot the histograms of inter-departure time at first stage for both classes. Histograms are plotted for a bin width of 0.01 time units. We consider the parameter setting as $\lambda_1 = 6$, $\lambda_2 = 8$, $\mu_1 = 15$, $\mu_2 = 16$, $p_{11} = 0.8$, $p_{12} = 0.9$ It can be seen from these figures that inter departure times are exponentially distributed. Also it is know that departure rate is same as arrival rate for any stable queueing system. Hence the departure process will be Poisson with parameters λ_1, λ_2 for classes 1 and 2 respectively as conjectured.

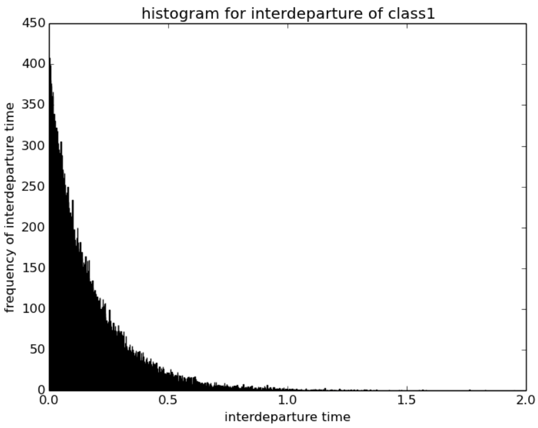

Fig. 10. Estimated density of interdeparture time for class1 customers

Q-Q plot: We use Q-Q plot to further strengthen this conjecture. In a Q-Q plot, we order the data points and plot them against the theoretical data quantiles of the distribution. If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line $y = x$. We sketch Q-Q plot for the parameter setting $\lambda_1 = 4.0$, $\lambda_2 = 8.0$, $\mu_1 = 15.0$, $\mu_2 = 16.0$, $p_{11} = 0.8$, $p_{12} = 0.9$ for both class 1 and class 2 customers (See Figure 12 and 13). We also sketch Q-Q plot for the parameter setting $\lambda_1 = 1.0$, $\lambda_2 = 2.0$, $\mu_1 = 15.0$, $\mu_2 = 16.0$, $p_{11} = 0.6$, $p_{12} = 0.8$ for both class 1 and class 2 customers (See Figure 14 and 15). It is noted from Q-Q plots that points approximately lie in line $y = x$. Hence results of conjecture 2 are accurate.

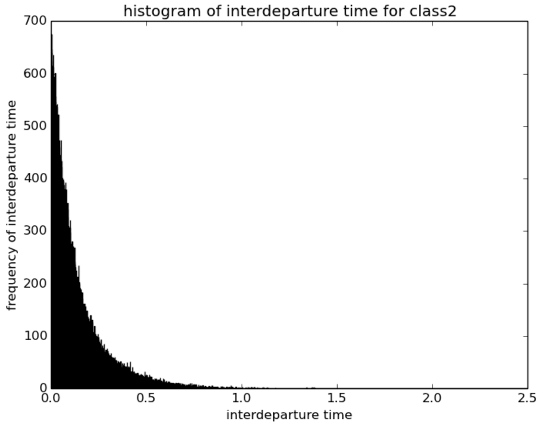

Fig. 11. Estimated density of interdeparture time for class2 customers

We also did experiments with different arrival rates and similar phenomenon is noted for both histograms and Q-Q plots (see [23] for details).

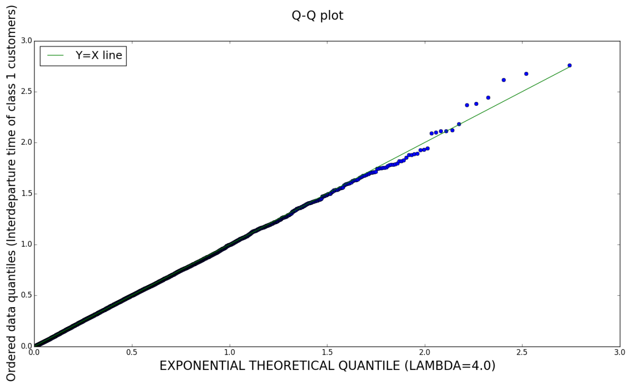

Fig. 12. Q-Q plot for class 1 inter-departure time distribution.

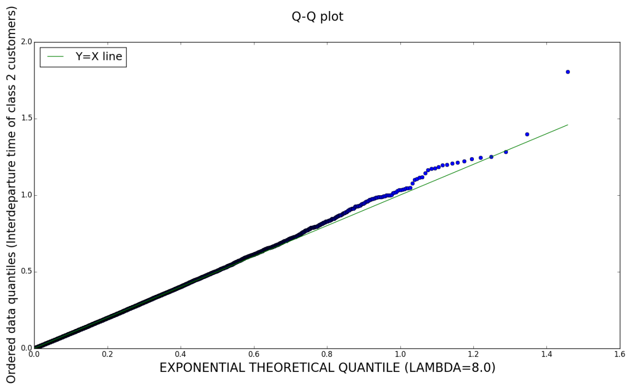

Fig. 13. Q-Q plot for class 2 inter-departure time distribution.

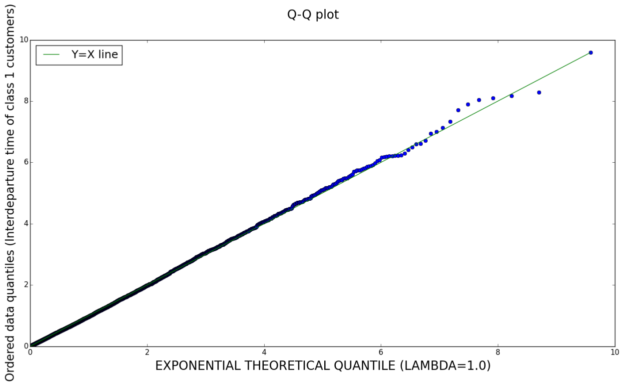

Fig. 14. Q-Q plot for class 1 inter-departure time distribution.

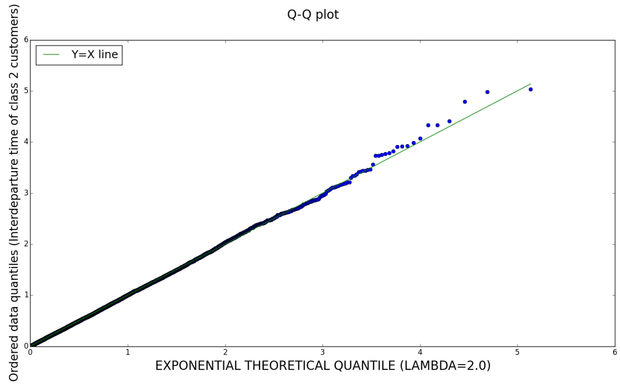

Fig. 15. Q-Q plot for class 2 inter-departure time distribution.

V. Conclusions and Future Work

Simulations of relative priority and earliest due date dynamic priority provide an insight into the behaviour of complex performance measures such as tail probability and switching frequency. It is noted that switching frequency poses a nice geometric shape (concave) with change in relative priority parameter. The decomposition result of mean waiting times for a two stage relative priority queuing network is useful in solving optimal control problems for queuing network, which often arise in D2D communication network design.

It will be interesting to analyse the other complex performance measure for example variance of waiting time. Extending the results obtained here for N classes is another interesting future avenue. Validity of conjectures 1 and 2 for other dynamic priority queuing systems can also be studied. Proving these conjectures analytically remains an open problem.

References

[1] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: A new architecture for wireless communications," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3. IEEE, 2000, pp. 1273–1282.

[2] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, "Intracluster device-to-device relay algorithm with optimal resource utilization," *IEEE transactions on vehicular technology*, vol. 62, no. 5, pp. 2315–2326, 2013.

[3] L. Lei, Z. Zhong, C. Lin, and X. Shen, "Operator controlled device-to-device communications in lte-advanced networks," *IEEE Wireless Communications*, vol. 19, no. 3, p. 96, 2012.

[4] N. Golrezaei, A. G. Dimakis, and A. F. Molisch, "Device-to-device collaboration through distributed storage," in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 2397–2402.

[5] N. Golrezaei, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," in *Communications (ICC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 7077–7081.

[6] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," 2013.

[7] S. Mumtaz and J. Rodriguez, *Smart Device to Smart Device Communication*. Springer, 2014.

[8] L. Kleinrock, "A delay dependent queue discipline," *Naval Research Logistics Quarterly*, vol. 11, pp. 329–341, September-December 1964.

[9] M. Haviv and J. van der Wal, "Waiting times in queues with relative priorities," *Operations Research Letters*, vol. 35, pp. 591 – 594, 2007.

[10] H. M. Goldberg, "Analysis of the earliest due date scheduling rule in queueing systems," *Mathematics of Operations Research*, vol. 2(2), pp. 145–154, 1977.

[11] Y. Jiang, C.-K. Tham, and C.-C. Ko, "Delay analysis of a probabilistic priority discipline," *European transactions on telecommunications*, vol. 13, no. 6, pp. 563–577, 2002.

[12] Y. Lim and J. E. Kobza, "Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch," *IEEE Transactions on Communications*, vol. 38 (5), pp. 351–358, 1990.

[13] T. Maertens, J. Walraevens, M. Moeneclaey, and H. Bruneel, "Performance analysis of a discrete-time queueing system with priority jumps," *AEU-International Journal of Electronics and Communications*, vol. 63, no. 10, pp. 853–858, 2009.

[14] R. Jain, *The art of computer systems performance analysis*. John Wiley & Sons, 2008.

[15] S. K. Bose, *An introduction to queueing systems*. The Rosen Publishing Group, 2002.

[16] S. Asmussen, *Applied probability and queues*. Springer, 2003, vol. 51.

[17] C.-p. Li and M. J. Neely, "Delay and rate-optimal control in a multi-class priority queue with adjustable service rates," in *INFOCOM, Proceedings IEEE*, 2012, pp. 2976–2980.

[18] D. Bertsimas and J. Niño-Mora, "Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems," *Mathematics of Operations Research*, vol. 21, no. 2, pp. 257–306, 1996.

[19] A. Rawal, V. Kavitha, and M. K. Gupta, "Optimal surplus capacity utilization in polling systems via fluid models," in *WiOpt, Proceedings IEEE*, 2014, pp. 381–388.

[20] R. Hassin, J. Puerto, and F. R. Fernández, "The use of relative priorities in optimizing the performance of a queueing system," *European Journal of Operational Research*, vol. 193, no. 2, pp. 476–483, 2009.

[21] P. J. Burke, "The output of a queuing system," *Operations research*, vol. 4, no. 6, pp. 699–704, 1956.

[22] D. A. Stanford, "Interdeparture-time distributions in the non-preemptive priority $\sum M_i/G_i/1$ queue," *Performance Evaluation*, vol. 12, no. 1, pp. 43–60, 1991.

[23] P. Mayekar, J. Venkatswaran, M. K. Gupta, and N. Hemachandra, "Simulation and analysis of various dynamic priority schemes in multi-class queues," IIT Bombay, Tech. Rep., 2014, http://www.ieor.iitb.ac.in/files/faculty/nh/simulationTR.pdf.

[24] N. Matloff, "Introduction to discrete-event simulation and the simpy language," *Davis, CA. Dept of Computer Science. University of California at Davis. Retrieved on August*, vol. 2, p. 2009, 2008.

[25] M. K. Gupta, N. Hemachandra, and J. Venkateswaran, "On completeness and equivalence of some dynamic priority schemes," IIT Bombay, Tech. Rep., 2014, http://www.ieor.iitb.ac.in/files/faculty/nh/completeTR.pdf.

[26] D. D. Yao, "Dynamic scheduling via polymatroid optimization," in *Performance Evaluation of Complex Systems: Techniques and Tools*. Springer, 2002, pp. 89–113.