On a Conservation Law and the Achievable Region for Waiting Time Tail Probabilities in 2-class M/G/1 Queueing Systems

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

***Abstract*— Conservation laws and the related achievable region for mean waiting times are important concepts in multi-class queues. The nice geometric polytope structure of this region driven by the conservation law is exploited extensively for dynamic control of multi-class queues. Such control problems have wide range of applications in computers, communication networks and manufacturing systems. Tail probability of each class's waiting time is another important performance measure in any multi-class queue. This paper studies an approximate conservation law, the related achievable region and completeness of the tail probability of waiting time of each class in two class M/G/1 queues. We use completeness of the recently introduced relative priority scheme for mean waiting time vector as well as a suitable partition of the stability region of the queue to show that this approximate achievable region for tail probabilities is enclosed in a trapezium. We also study the tightness of bounds based on this decomposition of the stability region.**

***Index Terms*— Dynamic priority, tail probability, achievable region, conservation law, multi class queues, optimal control of queues, non convex optimization**

I. Introduction

Multi-class queueing systems are an important class of queueing systems with multiple types of customers which may differ in their arrival processes and service requirements. Such queues are used to model complex systems, and hence have several important applications in telecommunication, computer, transportation and job shop manufacturing systems. There is a huge literature in analysing their performance (see [1], [2], [3], [4], [5] and references therein).

Many problems in wireless communication can be studied using multi-class queueing systems. Examples include a system which serves both voice and data calls or a cognitive radio type of scenario with primary and secondary class of customers. Optimal control of such multi-class systems is important for efficient system design. Different optimal control problems can be posed depending on the application regime. Many optimal control problems involving mean waiting time performance measure are well studied in literature (See [6] and [7]). Some recent surveys with applications focused on wireless communication can be seen in [8] and [9]. See [10] and [11] for textbook treatment of such important topics.

Manu K. Gupta is a Ph.D. student at IE&OR, IIT Bombay and he is partially supported by a Teaching Assistantship offered by Government of India.

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

E-mail addresses: manu.gupta@iitb.ac.in (Manu K. Gupta), nh@iitb.ac.in (N. Hemachandra).

One of the useful tools to solve such optimal control problems is the characterization of the achievable region for performance measure of interest, because one can then use optimization methods to find the optimal control policy (see [12], [7]). Researchers in this field have extensively studied the achievable region for mean waiting time performance measure and related optimal control problems. Coffman and Mitrani [13] were the first to identify the achievable region for *mean waiting time* in multi-class M/M/1 queue with pre-emptive priority discipline. Further structure of such achievable region is studied by many authors under certain scheduling assumptions (See [12], [14]). Conservations laws proposed by Kleinrock [15] play an important role in analysing the geometric structure of achievable region. Achievable region for mean waiting time in two class non work-conserving queueing (polling) system is recently explored in [16].

Geometric structure for mean waiting time vector in a work conserving multi-class single server priority queue was identified as polytope ([14], [17]). A set of parametrized scheduling strategy is called *complete* [14] if it achieves each point in this polytope for suitable value of the parameter. So, identifying a complete parametrized class of policy is quiet useful in solving optimal control problems as discussed in [18]. Some admission control problems are solved in two class polling system by exploiting the unbounded structure of achievable region for mean waiting time (see [16]).

Despite being such an important topic, there seems to be no research in exploring conservation law or achievable region for performance measures other than mean waiting time. Tail probability is another important performance measure to model quality of service in communication network (see [19]).

In this paper, we explore the conservation laws and achievable region for waiting time tail probabilities in two class M/G/1 queue for a given tail value. Unlike Kleinrock's conservation law [20] for mean waiting time which is an affine function and results in a nice geometric structure (polytope) of achievable region, tail probability conservation laws are non linear in nature. For a given tail value, we study an approximate conservation law and discuss the idea of completeness with respect to tail probability performance measure. We introduce the notion of partially complete parametrized class for tail probability. Relative priority scheme turns out to be a *partially* complete class. We also study the associated approximate achievable region for probabilities that waiting times exceed the tail value.

978-3-9018-8273-9/15/ ©2015 IFIP

This approximate achievable region, given by a non linear curve, is bounded by a trapezium in two class queue (See Figure 2). We study the tightness of the approximation of this achievable region by the trapezium. For this performance measure, system has family of conservation laws and achievable regions parametrized by tail value x.

Waiting time tail probabilities are not known in multi-class queues with dynamic priority across classes. To the best of our knowledge, only mean waiting times are known (See [15], [21]). This motivates us to use the approximation proposed in [19] for such tail probabilities. Computational experiments given in [19] show that this approximation is fairly accurate. We also numerically note that these approximations are good. In view of space limitations, proofs and many details are given in technical report [22].

Such ideas will be helpful in solving optimal control problems related to tail probabilities. Many optimal control problems involving mean waiting time are solved in literature by exploiting the achievable region for mean waiting time (see [6], [7], [23]). Constraint on tail probability can be a measure for Quality of Service, QoS, in many wireless communication problems. These ideas will be helpful in solving such optimization problems.

A. Paper organisation

This paper is organised as follows. Section II presents the system setting and relative priority scheduling. Section III describes the approximate waiting time tail probability conservation law. Section IV discusses the approximate achievable region and solution of certain optimization problems to find the bound for tail probability conservation law. Section V describes the tightness of bounds. Section VI presents the approximation error via simulation. Section VII ends with discussion on various interesting future avenues.

II. System Description

Consider a multi-class queueing system with $i = 1, 2, ..., N$ number of classes and each class has independent Poisson arrival rate λ_i and general service time distribution with mean $1/\mu_i$. Let $\mathcal{F}$ be the set of all work conserving, non pre-emptive and non anticipative scheduling policies across classes in multi-class queueing system. Let π be a scheduling policy in $\mathcal{F}$ and $\bar{W}_i^\pi$ be the mean waiting time of class i under scheduling policy π, $i = 1, 2, ..., N$. Achievable region for mean waiting time can be mathematically written as following set $\mathcal{W}$.

$$\mathcal{W} = \{(\bar{W}_1^\pi, \bar{W}_2^\pi, \cdots, \bar{W}_N^\pi) : \pi \in \mathcal{F}\}$$

Kleinrock's conservation law [15] is given by

$$\sum_{i=1}^{N} \rho_i \bar{W}_i^\pi = \frac{\rho W_0}{1-\rho} \quad \text{(constant)} \tag{1}$$

where $\rho_i = \lambda_i/\mu_i$, $\rho = \sum_{i=1}^{N} \rho_i$ and $W_0 = \sum_{i=1}^{N} \frac{\lambda_i}{2}\left(\sigma_i^2 + \frac{1}{\mu_i^2}\right)$. Here σ_i^2 is the variance of service time for class i. Note that the right hand side of above equation is independent of scheduling policy. And this helps in characterizing the achievable region. Achievable region in two class, say $\mathcal{W}'$ as described below, forms a line segment in case of two classes and a polytope for more than two classes (see [14]).

$$\mathcal{W}' = \{(\bar{W}_1^\pi, \bar{W}_2^\pi) : \pi \in \mathcal{F}\}$$

Achievable region for mean waiting time vector with two classes is shown in Figure 1. In this figure, W_{12} and W_{21} are two extreme points of line segment corresponding to strict priorities given to class 1 and class 2 respectively.

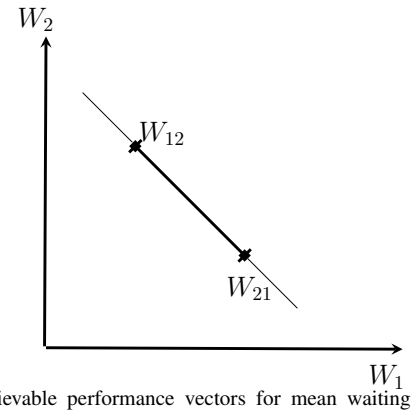

Fig. 1. Achievable performance vectors for mean waiting time driven by Kleinrock's conservation law in two class M/G/1 queue

Now, we look for conservation law and achievable region for waiting time tail probability in two class queue. Tail probability of waiting time of class i, W_i, is defined as $P(W_i > x)$ for a given tail value x, $i = 1, 2$. Mathematically, achievable region for tail probability can be represented by following space $\mathcal{T}_x$ for a given tail value x:

$$\mathcal{T}_x = \{(P(W_1^\pi > x), P(W_2^\pi > x)) : \pi \in \mathcal{F}\}$$

We study the approximate conservation law related to the above set. We also obtain the approximation for it in log transformed axis.

Consider non pre-emptive work conserving scheduling discipline in multi-class queue where customers within each queue are served in first come first serve order. Approximation of waiting time tail probability for class i is given by [19]:

$$P(W_i^\pi > x) \approx \rho e^{-\rho x/\bar{W}_i^\pi}, i = 1, 2 \tag{2}$$

We now briefly describe relative priority scheduling which will be helpful in solving certain optimization problems to obtain approximate achievable region for tail probability in two class M/G/1 queue.

A. Relative Priority

This is a dynamic priority scheduling across classes first proposed by Moshe Haviv and Van Der Wal [21]. In this multi-class priority system, a *positive* parameter p_i is associated with each class i. Let each class have independent Poisson arrival rate λ_i. 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:

$$\frac{n_i p_i}{\sum_{j=1}^{N} n_j p_j}, \quad 1 \leq i \leq N \tag{3}$$

Mean waiting time for class i, $\bar{W}_i$, in case of two classes is given by [21]:

$$\bar{W}_i = \frac{1 - \rho p_i}{(1-\rho_1-p_2\rho_2)(1-\rho_1-p_1\rho_1) - p_1p_2\rho_1\rho_2} W_0, \tag{4}$$

for $i = 1,\ 2$; here $\rho = \rho_1 + \rho_2$ and $W_0 = \sum_{i=1}^{2} \lambda_i \bar{x}_i^2$ where $\bar{x}_i^2$ is the second moment of service time of class i. Also $p_1 + p_2 = 1$. Relative priorities are modified in natural way in [24] so that it achieves all the mean waiting time vectors in the achievable region (line segment) shown in Figure 1.

Modified Relative priorities: We consider weights given to class i, $i = 1, 2$, as p_i to be non-negative instead of only positive and this happens for at most one p_i. Also when $p_i = 0$ and $n_i > 0 = n_j\ \forall\ j$ then class i customer is served, $i = 1, 2$. Remaining setting is same as *relative priority*.

Modified relative priority is proved to span entire achievable region (complete) in two class M/G/1 queue (see [24]).

III. Approximate conservation law

In this section, we describe approximate conservation law for tail probability similar to Kleinrock's mean waiting time conservation law in Equation (1). This is an approximate conservation law as we use an approximation for tail probability described in Equation (2).

Approximation given by Equation (2) for two class queueing system results in following expression:

$$P(W_1^\pi > x)^{\rho_2} P(W_2^\pi > x)^{\rho_1} = \rho^{\rho_1+\rho_2} e^{-\rho x \left(\frac{\rho_1 \bar{W}_1^\pi + \rho_2 \bar{W}_2^\pi}{\bar{W}_1^\pi \bar{W}_2^\pi} \right)} \tag{5}$$

On using Kleinrock's conservation law for two classes from Equation (1), we have

$$P(W_1^\pi > x)^{\rho_2} P(W_2^\pi > x)^{\rho_1} = \rho^{\rho_1+\rho_2} e^{-\rho x \left(\frac{\rho W_0}{\bar{W}_1^\pi \bar{W}_2^\pi (1-\rho)} \right)} \tag{6}$$

Theorem 1: Approximate waiting time tail probability conservation law is given by:

$$\rho_2 \log P(W_1^\pi > x) + \rho_1 \log P(W_2^\pi > x)$$

$$+ \frac{\rho^2 x W_0}{(1-\rho)\int_0^\infty P(W_1^\pi > y)dy \int_0^\infty P(W_2^\pi > y)dy} = \rho \log \rho \tag{7}$$

Proof: Follows by taking natural logarithm of Equation (6) and using the fact that mean waiting time, $\bar{W}_i = \int_0^\infty P(W_i^\pi > y)dy$ for $i = 1,\ 2$. ∎

Recall from Equation (1), Kleinrock's conservation law for mean waiting time, that right hand side is a constant and independent of scheduling policy. Equation (7) has similar interpretation as RHS is a constant and it is independent of scheduling policy. Hence Equation (7) forms approximate tail probability conservation law. Interestingly, this constant $(\rho \log \rho)$ does not depend on tail value x also.

Completeness of (modified) relative priority for mean waiting time implies that given a scheduling policy $\pi \in \mathcal{F}$, under which the mean waiting time vector is $(\bar{W}_1^\pi, \bar{W}_2^\pi)$, there is a (modified) relative priority scheduling policy π_{rp} for which the mean waiting time vector is exactly $(\bar{W}_1^\pi, \bar{W}_2^\pi)$. We refer this as *mean waiting time completeness*.

Similarly, one can consider completeness of *tail probabilities of waiting times*. It follows from mean completeness of relative priority that the term $\rho \log \rho - \frac{\rho^2 x W_0}{\bar{W}_1^\pi \bar{W}_2^\pi (1-\rho)}$ can be recovered by appropriately choosing the relative priority parameter. However, the tail probabilities for an arbitrary policy π, $(P(W_1^\pi > x), P(W_2^\pi > x))$ may not be same as that with chosen relative priority scheme. Hence, we introduce the notion of partially tail probability complete class below.

For given tail value x and policy π, approximate conservation law is a non-linear curve. A series of such non linear curves can be obtained by changing scheduling policy π for a fixed tail value x. We discuss the notion of *partially* complete class of scheduling policy for space $\mathcal{T}_x = \{(P(W_1^\pi > x), P(W_2^\pi > x)) : \pi \in \mathcal{F}\}$ in the context of these non linear curves. A parametrized policy forms a *partially* complete class if it achieves all possible curves of tail probability vector driven by conservation law defined in Theorem 1 for given value of x.

For an arbitrary policy π, a *partially* complete parametrized class achieves the same non linear curve by choosing appropriate parameter. However, exact tail probability vector with arbitrary policy π, $(P(W_1^\pi > x), P(W_2^\pi > x))$ may not be achieved by the appropriately chosen parameter from a partially complete class. Hence, a *partially* complete parametrized class may not result in the same tail probability vector even if non linear curve is same.

By using a (modified) relative scheduling policy in Equation (7), we have that the approximate tail probability curve (defined in Theorem 1) under an arbitrary scheduling policy $\pi \in \mathcal{F}$ is achieved by using (modified) the relative probability scheduling policy π_{rp}. Thus, we have

Theorem 2: For a given tail value $x > 0$, the (modified) relative priority class is *partially* complete class for the approximate tail probability curves in the space $\mathcal{T}_x = \{(P(W_1^\pi > x), P(W_2^\pi > x)) : \pi \in \mathcal{F}\}$.

Remark 1: The above approximate conservation law traces a continuous non linear curve in $P(W_1^\pi > x)$ and $P(W_2^\pi > x)$ plane as $\bar{W}_1^\pi \bar{W}_2^\pi$ is continuous in p.

IV. Approximate achievable region

In this section, we study the approximate achievable region for tail probability motivated by line segment for mean waiting time vectors driven by Kleinrock's conservation law.

The achievable region for waiting time tail probability, τ_x, is a subset in the unit square, $[0,1] \times [0,1]$ with a non linear boundary. We are not aware of any explicit expression for tail probabilities of waiting times in two class queues under (pure) dynamic priority policy. We now study the approximate achievable region by a log transformation of the approximate conservation law (Equation (6)) which turns out to be included in a semi open trapezium (See figure 2).

Rewriting Equation (7) as

$$\rho_2 \log P(W_1^\pi > x) + \rho_1 \log P(W_2^\pi > x) = \rho \log \rho - \frac{\rho^2 x W_0}{\bar{W}_1^\pi \bar{W}_2^\pi (1-\rho)} \quad (8)$$

For a given x, the RHS of Equation (8) depends on scheduling

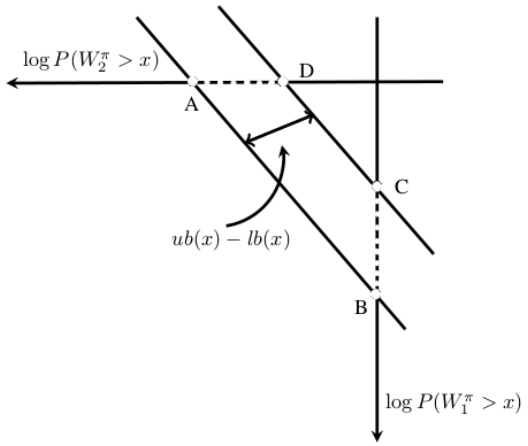

Fig. 2. Approximate achievable performance vectors for tail probability of waiting time driven by bounds from Equation (9) in two class M/G/1 queue (for a given tail value x), τ_x

policy by mean waiting time expression. A uniform bound, independent of scheduling policy, can be obtained by solving following optimization problems **P1** and **P2** for a given tail value x.

P1: $\max_{\mathcal{F}} \ \bar{W}_1^\pi \bar{W}_2^\pi$ Subject to $\rho_1 \bar{W}_1^\pi + \rho_2 \bar{W}_2^\pi = \frac{\rho W_0}{1-\rho}$

P2: $\min_{\mathcal{F}} \ \bar{W}_1^\pi \bar{W}_2^\pi$ Subject to $\rho_1 \bar{W}_1^\pi + \rho_2 \bar{W}_2^\pi = \frac{\rho W_0}{1-\rho}$

Let u^* and l^* be the optimal objective for optimization problem P1 and P2 respectively. Lower bound, $lb(x)$, and upper bound, $ub(x)$, for Equation (8) can be obtained as a function of l^* and u^* irrespective of scheduling policy. Hence, we have

$$lb(x) \leq \rho_2 \log P(W_1^\pi > x) + \rho_1 \log P(W_2^\pi > x) \leq ub(x) \quad (9)$$

where

$$lb(x) = \rho \log \rho - \frac{\rho^2 x W_0}{(1-\rho) l^*} \text{ and} \quad (10)$$

$$ub(x) = \rho \log \rho - \frac{\rho^2 x W_0}{(1-\rho) u^*} \quad (11)$$

Geometric interpretation of mean waiting time conservation law is that all mean waiting time vectors lie in the hyperplane defined by conservation law (See Figure 1). Similar geometric interpretation for tail probability conservation law for a given x can be given. Equation (9) gives a bounded approximate achievable region for tail probability in log transformed axis (see figure 2). ABCD is the approximate achievable region for tail probability in figure 2 for a given x. Note that line segment BC and AD are not achievable as corresponding logarithm value is 0 and hence tail probability has to be 1.

Now, we look for the solution of optimization problems P1 and P2 to explicitly calculate lower bound, $lb(x)$ and upper bound, $ub(x)$ for given value of x using modified relative priority scheduling discipline.

A. Solution of optimization problems P1 and P2

In this subsection, we discuss the solution of optimization problem P1 and P2. We also explicitly compute the bounds on tail probability conservation law using Equation (9). We first look for solution of optimization problem P1.

Modified relative priorities are proved to be complete in case of two classes (see [24]). This implies that optimizing over set of all work conserving, non pre-emptive and non anticipative policies is equivalent to optimize over the range of relative priority parameter. Hence, optimization problem P1 is equivalent to following transformed problem T1.

T1: $\max_{0 \leq p \leq 1} \ \bar{W}_1^p \bar{W}_2^p$

Subject to

$$\rho_1 \bar{W}_1^p + \rho_2 \bar{W}_2^p = \frac{\rho W_0}{1-\rho} \quad (12)$$

where $\bar{W}_i^p$ is the mean waiting time of class i with p as relative priority scheduling parameter across classes. Note that p and $1-p$ are the parameters associated with class 1 and class 2 respectively in relative priority scheduling mechanism. On using the expression of mean waiting time in relative priority [21], it is noted that conservation law (Equation (12)) is trivially satisfied. Hence the transformed optimization problem T1 reduces to

$$\max_{0 \leq p \leq 1} \frac{(1-\rho p)(1-\rho(1-p)) W_0^2}{((1-\rho_1-(1-p)\rho_2)(1-\rho_2-p\rho_1)-p(1-p)\rho_1\rho_2)^2}$$

The objective of above unconstrained optimization problem is a function of p. Let us denote it by $f(p)$. Its derivative is calculated using Mathematica and further simplified as follows.

$$\frac{df(p)}{dp} = \frac{C_1 + pC_2}{[g(p)]^3} \quad (13)$$

where

$$C_1 = (1-\rho)[\rho^2(1-\rho_2) + 2(\rho_1(1-\rho_1) - \rho_2(1-\rho_2))] \quad (14)$$

$$C_2 = \rho^2[\rho_1(1-\rho_1) - \rho_2(1-\rho_2) - 2(1-\rho_2)(1-\rho)] \quad (15)$$

and $g(p) = [(-1+p)p\rho_1\rho_2 + (1-p\rho_1-\rho_2)(1-\rho_1-(1-p)\rho_2)]$. The derivative in Equation (13) is zero if $p = -\frac{C_1}{C_2}$ and $g(p) \neq$

0. On simplifying $g(p) \neq 0$, we get

$$p \neq \frac{(1-\rho_2)(1-\rho_1-\rho_2)}{\rho_1(1-\rho_1)-\rho_2(1-\rho_2)} \quad (16)$$

Theorem 3: Pure dynamic policy will be the optimal solution to problem P1 with $p^* = -C_1/C_2$ if $\lambda_1,\ \lambda_2$ and μ are in following stability region D:

$$D \equiv \{\lambda_1,\ \lambda_2,\ \mu : \beta_1 < Y < \beta_2\}$$

where $Y = \rho_1(1-\rho_1) - \rho_2(1-\rho_2)$, $\beta_1 = -\rho^2(1-\rho_2)/2$ and $\beta_2 = \dfrac{\rho^2(1-\rho_2)(1-\rho)}{\rho^2+2(1-\rho)}$. And objective function is concave in nature.

Proof: Stability region D is obtained by looking into range of $\lambda_1,\ \lambda_2$ and μ such that p^* lies in $(0,1)$. Nature of objective function is obtained by exploring the sign of second derivative of objective function. More details can be seen in technical report [22]. ■

We now look for the solution of optimization problem P2 in stability region D. Due to concave nature of objective, minima will lie at either $p = 0$ or $p = 1$. Further decomposition of stability region D is discussed below based on the solution of optimization problem P2.

Decomposition of stability region: We calculate the difference in the value of objective function, $\bar{W}_1\bar{W}_2$, at $p = 0$ and $p = 1$ to identify global minima. We have

$$\bar{W}_1\bar{W}_2|_{p=0} - \bar{W}_1\bar{W}_2|_{p=1} = \frac{(\rho_2-\rho_1)(2-\rho_1-\rho_2)W_0^2}{(1-\rho)(1-\rho_1)^2(1-\rho_2)^2} \quad (17)$$

Sign of above term is decided by $(\rho_2 - \rho_1)$. So is the minima and solution of optimization problem P2.

Stability condition for queue is $\rho < 1$. Let S be the stability region defined as $S \equiv \{\lambda_1,\ \lambda_2,\ \mu : \lambda_1 + \lambda_2 < \mu\}$. It follows from Equation (17) and Theorem 3 that solution of optimization P2 depends on relative values of ρ_1 and ρ_2. Hence, this stability region D is further decomposed in two parts D_1 and D_2 as shown below:

$$D_1 \equiv \{\lambda_1,\ \lambda_2,\ \mu :\ \beta_1 < Y < \beta_2 \text{ and } \rho_2 > \rho_1\} \quad (18)$$
$$D_2 \equiv \{\lambda_1,\ \lambda_2,\ \mu :\ \beta_1 < Y < \beta_2 \text{ and } \rho_2 < \rho_1\} \quad (19)$$

The difference in Equation (17) is positive for region D_1. Hence global minima will be given at $p^* = 1$, i.e., class 1 should be given strict priority over class 2. Similarly, this difference will be negative for region D_2, so $p^* = 0$ will be the global minimizer. Note that when $\rho_1 = \rho_2$, difference in Equation (17) is 0 and derivative $\dfrac{df(p)}{dp} = 0$ at $p = 1/2$. Hence global maxima will be at $p^* = 1/2$ and minima will be at $p^* = 0$ or $p^* = 1$. A schematic of function, $f(p)$, is shown in Figure 3.

It follows from Theorem 3 that $p^* = -C_1/C_2 \notin (0,1)$ for the range beyond stability region D. Hence objective function will be monotone for $p \in [0,1]$ and strict priority will be optimal for stability region D^c. Consider further decomposition of stability region in S_1 and S_2 depending on relative value of ρ_1 and ρ_2

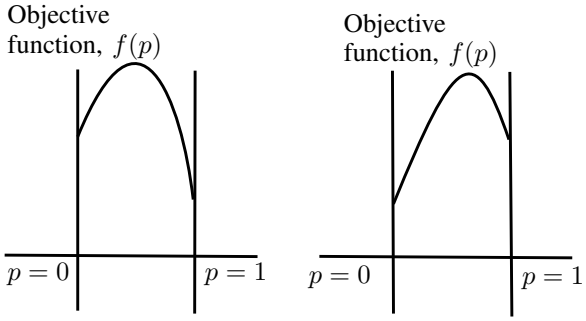

Fig. 3. Left figure captures the behaviour of objective function in stability region D_1 and right one captures the same in D_2

as follows.

$$S_1 \equiv \{\lambda_1,\ \lambda_2,\ \mu : Y \in (\beta_1,\beta_2)^c \text{ and } \rho_2 > \rho_1\}$$
$$S_2 \equiv \{\lambda_1,\ \lambda_2,\ \mu : Y \in (\beta_1,\beta_2)^c \text{ and } \rho_2 < \rho_1\}$$

By using definition of $\beta_1,\ \beta_2$ and Y, sets S_1 and S_2 can be rewritten as:

$$S_1 \equiv \{\lambda_1,\ \lambda_2,\ \mu : Y \in (-\infty,\beta_1] \text{ and } \rho_2 > \rho_1\} \quad (20)$$
$$S_2 \equiv \{\lambda_1,\ \lambda_2,\ \mu : Y \in [\beta_2,\infty) \text{ and } \rho_2 < \rho_1\} \quad (21)$$

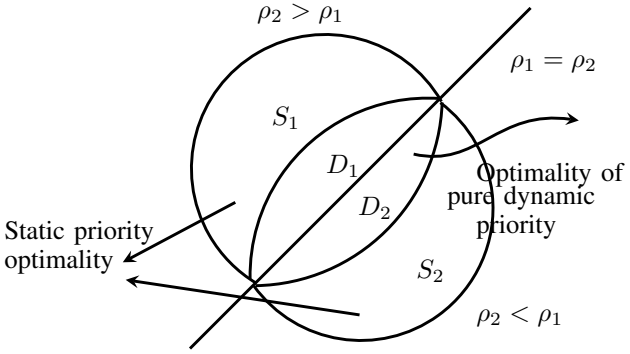

Fig. 4. Decomposition of stability region

Decomposition of stability region is shown in Figure 4. Theorem 4 below discusses the nature of objective function in stability region S_1 and S_2.

Theorem 4: Objective function of optimization problem P1 is monotonically decreasing in stability region S_1 while it is monotonically increasing in stability region S_2.

Proof: Such a nature of objective function is obtained by exploring the nature of first derivative of objective function in stability region S_1 and S_2. See technical report [22]. ■

It is clear from Figure 5 that static priority will be optimal for both optimization problems P1 and P2 in stability regions S_1 and S_2.

Numerical examples for illustration of nature of objective function in different decomposed stability regions can be seen

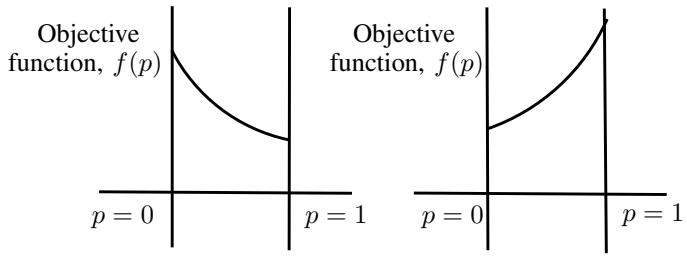

Fig. 5. Left figure captures the behaviour of objective function in stability region S_1 and right one captures the same in S_2

in technical report [22]. Same phenomenon as observed by Theorem 3 and 4 is noted in numerical examples (see [22]).

B. Calculation of bounds, $lb(x)$ and $ub(x)$

In this subsection, we explicitly calculate the bounds $lb(x)$ and $ub(x)$ for various stability region and this completes the approximation for achievable region shown in Figure 2. Following Achievable Region Bounds, ARB, algorithm determines the $lb(x)$ and $ub(x)$ for given input parameter setting.

These $lb(x)$ and $ub(x)$ give the bounds on approximate achievable region of tail probability for a given tail value x as shown in Figure 2.

We summarize the above in following theorem.

Theorem 5: For a given $x > 0$, approximate achievable region for tail probability, $(P(W_1 > x), P(W_2 > x))$, is a semi open trapezium in 3rd orthand of $\mathbb{R}^2$ bounded by $lb(x) \leq \rho_2 \log P(W_1 > x) + \rho_1 \log P(W_2 > x) \leq ub(x)$ where $lb(x)$ and $ub(x)$ are calculated by ARB algorithm.

V. Tightness of bounds

In this section, we discuss the tightness of the bounds obtained in previous section. We explicitly calculate the tightness as the difference between upper and lower bound on tail probability conservation law in different stability regions. Mathematically,

$$t(x) := ub(x) - lb(x)$$

By using the definition of $ub(x)$ and $lb(x)$ from Equation (11) and (10), we have

$$t(x) = \frac{\rho^2 x W_0 (u^* - l^*)}{(1-\rho) l^* u^*} \tag{22}$$

Remark 2: $t(x)$ is linear in tail value x.

We obtain closed form expression for $t(x)$ when input parameters are in stability region S_1, S_2 and when load factors of both classes are equal. We discuss the tightness of bounds with the help of numerical examples for region D_1 and D_2.

A. Region S_1 and S_2

It follows from Theorem 4 that lower and upper bounds are given by $p = 0$ and $p = 1$ respectively for stability region S_1 while lower and upper bounds are given by $p = 1$ and $p = 0$ respectively for stability region S_2. Due to this symmetry and Equation (22), tightness $t(x)$ will be same for both stability regions S_1 and S_2.

By using mean waiting time expressions from [21] at $p = 0$ and $p = 1$, we get l^* and u^* for stability region S_1 as follows.

$$l^* = \bar{W}_1^{\pi} \bar{W}_2^{\pi}|_{p=0} = \frac{W_0^2}{(1-\rho)(1-\rho_2)^2} \text{ and}$$

$$u^* = \bar{W}_1^{\pi} \bar{W}_2^{\pi}|_{p=1} = \frac{W_0^2}{(1-\rho)(1-\rho_1)^2}$$

Algorithm 1 ARB algorithm to compute $lb(x)$ and $ub(x)$

Inputs: $\lambda_1,\ \lambda_2,\ \mu,\ x$

Determine the stability region among $S_1,\ S_2,\ D_1,\ D_2$ for given input parameters using Equations (18), (19), (20), (21).

if $\lambda_1,\ \lambda_2,\ \mu \in S_1$ **then**

$l^* = \bar{W}_1 \bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1 \bar{W}_2|_{p=0}$

else if $\lambda_1,\ \lambda_2,\ \mu \in S_2$ **then**

$l^* = \bar{W}_1 \bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1 \bar{W}_2|_{p=0}$

else if $\lambda_1,\ \lambda_2,\ \mu \in D_1$ **then**

$l^* = \bar{W}_1 \bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1 \bar{W}_2|_{p=-C_1/C_2}$

else if $\lambda_1,\ \lambda_2,\ \mu \in D_2$ **then**

$l^* = \bar{W}_1 \bar{W}_2|_{p=0}$ and $u^* = \bar{W}_1 \bar{W}_2|_{p=-C_1/C_2}$

else if $\rho_1 = \rho_2$ **then**

$l^* = \bar{W}_1 \bar{W}_2|_{p=0}$ or $\bar{W}_1 \bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1 \bar{W}_2|_{p=1/2}$

Compute $\bar{W}_1 \bar{W}_2$ as below to calculate l^* and u^*

$$\bar{W}_1^{\pi} \bar{W}_2^{\pi}|_{p=0} = \frac{W_0^2}{(1-\rho)(1-\rho_2)^2},\ \bar{W}_1^{\pi} \bar{W}_2^{\pi}|_{p=1} = \frac{W_0^2}{(1-\rho)(1-\rho_1)^2} \text{ and } \bar{W}_1^{\pi} \bar{W}_2^{\pi}|_{p=\frac{1}{2}} = \frac{W_0^2}{(1-2\rho_1)^2}$$

To compute u^* for region D_1 or D_2, calculate $p^* = -C_1/C_2$ using Equation (14) and (15) and then use mean waiting time Equation (4).

Output: $lb(x) = \rho \log \rho - \dfrac{\rho^2 x W_0}{l^*(1-\rho)}$ and $ub(x) = \rho \log \rho - \dfrac{\rho^2 x W_0}{u^*(1-\rho)}$

Using Equation (22) and above expressions, we get tightness

$$t(x) = (\rho_2 - \rho_1)(\rho_1 + \rho_2 - 2)\frac{\rho^2 x}{W_0}$$

B. Equal load factors

It follows from analysis in Section IV-A that upper bound is given by $p = 1/2$ and lower bound is given by either $p = 0$ or $p = 1$ as objective function value $f(p)$ is same. By using mean waiting time expressions from [21], we obtain

$$l^* = \bar{W}_1^{\pi}\bar{W}_2^{\pi}|_{p=0} = \bar{W}_1^{\pi}\bar{W}_2^{\pi}|_{p=1} = \frac{W_0^2}{(1-\rho_1)^2(1-2\rho_1)}$$

$$u^* = \bar{W}_1^{\pi}\bar{W}_2^{\pi}|_{p=1/2} = \frac{W_0^2}{(1-2\rho_1)^2}$$

Using above expressions, tightness of bounds for equal load factor is given by:

$$t(x) = \rho_1^2\frac{\rho^2 x}{W_0}$$

C. Region D_1 and D_2

It follows from Theorem 3 and analysis in Section IV-A that lower bound, l^*, is given at $p = 1$ and $p = 0$ for stability region D_1 and D_2 respectively while upper bound, u^*, is given at $p = -\frac{C_1}{C_2}$. Explicit calculation of upper bound is theoretically intractable due to highly non linear nature of expressions C_1 and C_2. So is the tightness of bounds in this stability region. However, $lb(x)$ can be obtained using Equation (10) for stability region D_1 and D_2.

$$\begin{aligned} lb(x)|_{D_1} &= \rho\log\rho - \frac{\rho^2 x}{W_0}(1-\rho_1)^2 \\ lb(x)|_{D_2} &= \rho\log\rho - \frac{\rho^2 x}{W_0}(1-\rho_2)^2 \end{aligned}$$

1) Examples: We compute the tightness of bounds, $t(x)$, numerically for stability region D_1 and D_2. As $lb(x)$ has a closed form expression as above, $ub(x)$ is computed in these stability regions by evaluating objective $f(p)$ at $p^* = -C_1/C_2$ while C_1 and C_2 are obtained from Equation (14) and (15). We consider some instances of λ_1, λ_2 and μ such that conditions for stability region D_1 and D_2 are satisfied. We report here the results with two set of service time distributions.

Example 1 Let the service time distribution be deterministic with rate μ. Hence variance will be 0 and second moment will be $1/\mu^2$.

Example 2 We consider a two sided service time distribution as follows.

$$S = \begin{cases} 1, & \text{with probability } 0.3 \\ 3, & \text{with probability } 0.7 \end{cases}$$

The tightness of bounds for different parameter setting with tail value $x = 0.1$ is tabulated in Table I and II for above examples in stability region D_1.

It is noted that the tightness is larger in heavy traffic. In all instances, difference is not more than 0.54. Hence bounds are pretty tight. Similar results are noted for other experiments with exponential and uniform service time distribution (see technical report [22]). Same tightness in bounds, $t(x)$, is observed for stability region D_2 due to symmetry and $p^*_{D_2}$ turns out to be $1 - p^*_{D_1}$ (See [22]).

TABLE I

ILLUSTRATION OF $t(x)$ FOR DETERMINISTIC SERVICE TIME DISTRIBUTION

λ_1	λ_2	μ	$p^*_{D_1}$	Tightness of bounds, $t(x)$ $x = 0.1$
1.2	3.5	5	0.2241	0.5389
1	1.5	5	0.2	0.3080
0.8	1	8	0.0617	0.2877

TABLE II

ILLUSTRATION OF $t(x)$ FOR TWO SIDED SERVICE TIME DISTRIBUTION

λ_1	λ_2	μ	$p^*_{D_1}$	Tightness of bounds $x = 0.1$
0.15	0.2	0.4167	0.4014	0.0140
0.1	0.15	0.4167	0.2667	0.0114
0.07	0.1	0.4167	0.1557	0.0110

VI. ERROR IN APPROXIMATION

In this section, we study the error in tail probability approximation of [19] which is used to obtain approximate tail probability conservation law and approximate achievable region. Tail probability approximation is given by [19]:

$$P(W_i > x) \approx \rho e^{-\rho x/\bar{W}_i} \quad i = 1, 2 \tag{23}$$

Note that tail probability and mean waiting time in standard M/M/1 queue is:

$$P(W > x) = \rho e^{-\mu(1-\rho)x} \text{ and } \bar{W} = \frac{\rho/\mu}{1-\rho}$$

It can be easily verified using above equations that approximation (23) is exact for single class standard M/M/1 queue. Different scheduling policies, strict priority, weighted fair queueing (WFQ) and weighted round robin (WRR) disciplines are simulated with four classes of queues in [19]. Numerical results for these disciplines are reported in [19]. These results show that approximation (23) is quiet adequate.

We use simulation to check the error in approximation for two classes. We build a simulator for two class queue where relative priority scheduling is implemented across classes as this scheduling policy is used for obtaining bounds on approximate achievable region in Section IV. This simulator is build in SymPy and simulation results are validated from theoretical known results (mean waiting time from [21]). See technical report [22] for details. Total run time is 100000 mins. Five replications with each setting are simulated and average value is reported in tables.

Now, we compute tail probability using relative priority simulator and check the error in approximation with Equation (23).

TABLE III
ERROR CALCULATION IN TAIL PROBABILITY VIA SIMULATION FOR $x = 0.5$

Settings ($\mu = 10$)	Relative Priority	Simulation		Approximation		Absolute Difference	
		P($W_1 > 0.5$)	P($W_2 > 0.5$)	P($W_1 > 0.5$)	P($W_2 > 0.5$)	Class 1	Class 2
$\lambda_1 = 1.5$ $\lambda_2 = 0.5$	$p = 0.1$	0.00463	0.00217	0.00431	0.00204	0.00063	0.00053
	$p = 0.4$	0.00406	0.00355	0.00382	0.00319	0.00049	0.00077
	$p = 0.8$	0.00353	0.00642	0.00317	0.00532	0.00065	0.00144
$\lambda_1 = 2$ $\lambda_2 = 4$	$p = 0.1$	0.14259	0.03869	0.16041	0.04049	0.01782	0.00200
	$p = 0.4$	0.09530	0.06851	0.10022	0.07165	0.00492	0.00314
	$p = 0.8$	0.03021	0.09542	0.03226	0.10666	0.00204	0.01124
$\lambda_1 = 6$ $\lambda_2 = 3$	$p = 0.1$	0.55027	0.14350	0.62282	0.15435	0.07254	0.01084
	$p = 0.4$	0.51010	0.42166	0.57207	0.47912	0.06196	0.05746
	$p = 0.8$	0.36855	0.58024	0.39582	0.67987	0.02726	0.09963

Total run time for simulation is 100000 mins. Five replications with each setting are performed and average value is reported. Table III contains simulation results with tail value 0.5. Note that the maximum error encountered in above experiments is 0.01 except for heavy traffic. Hence approximations are quiet accurate for low and moderate traffic and so is the bounds on tail probability conservation law and approximate achievable region. Experiments with tail value $x = 0.1$ are also performed and similar results are noted (See technical report [22]).

VII. DISCUSSION

A conservation law for waiting time tail probabilities in two class queues is explored in this paper. Notion of partially complete class is discussed and relative priority is identified as partially complete scheduling policy. Tail probability completeness of relative priority is yet to be discovered. Bounds on the related approximate achievable region for these probabilities are obtained by solving certain optimization problems. Tightness of these bounds is discussed. Error in approximation of tail probability is explored via simulation. It will be interesting to explore optimal control problems where quality of service is in terms of tail probabilities by exploiting this approximate achievable region. Further, expanding these ideas in multi-class queue and queueing networks will be another fascinating future avenue.

REFERENCES

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

[2] L. Ponomarenko, C. S. Kim, and A. Melikov, *Performance analysis and optimization of multi-traffic on communication networks*. Springer, 2010, vol. 208.

[3] R. W. Wolff, "Stochastic modelling and the theory of queues," *Englewood Cliffs, NJ*, 1989.

[4] L. Kleinrock, "Queueing systems, volume ii: computer applications," 1976.

[5] M. Haviv, *Queues: A Course in Queueing Theory*. Springer, 2013, vol. 191.

[6] 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.

[7] 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.

[8] A. Asadi and V. Mancuso, "A survey on opportunistic scheduling in wireless communications," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 4, pp. 1671–1688, 2013.

[9] R. Adams, "Active queue management: a survey," *Communications Surveys & Tutorials, IEEE*, vol. 15, no. 3, pp. 1425–1476, 2013.

[10] A. Kumar, D. Manjunath, and J. Kuri, *Communication networking: an analytical approach*. Elsevier, 2004.

[11] ——, *Wireless networking*. Morgan Kaufmann, 2008.

[12] D. Bertsimas, I. Paschalidis, and J. N. Tistsiklis, "Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance," *The Annals of Applied Probability*, vol. 4, pp. 43–75, 1994.

[13] E. Coffman Jr and I. Mitrani, "A characterization of waiting time performance realizable by single-server queues," *Operations Research*, vol. 28, no. 3-part-ii, pp. 810–821, 1980.

[14] I. Mitrani and J. Hine, "Complete parametrized families of job scheduling strategies," *Acta Informatica*, vol. 8, pp. 61– 73, 1977.

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

[16] 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.

[17] J. G. Shanthikumar and D. D. Yao, "Multiclass queueing systems: Polymatroidal structure and optimal scheduling control," *Operations Research*, vol. 40, no. 3-supplement-2, pp. S293–S299, 1992.

[18] M. K. Gupta, N. Hemachandra, and J. Venkateswaran, "On mean waiting time completeness and equivalence of EDD and HOL-PJ dynamic priority in 2-class M/G/1 queue," in *8th international conference on performance methodology and tools (Valuetools)*, 2014.

[19] Y. Jiang, C.-K. Tham, and C.-C. Ko, "An approximation for waiting time tail probabilities in multiclass systems," *IEEE Communications letters*, vol. 5, no. 4, pp. 175–177, 2001.

[20] L. Kleinrock, "A conservation law for wide class of queue disciplines," *Naval Research Logistics Quarterly*, vol. 12, pp. 118–192, 1965.

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

[22] M. K. Gupta and N. Hemachandra, "On a conservation law and the achievable region for waiting time tail probabilities in 2-class M/G/1 queueing systems," IIT Bombay, Tech. Rep., 2014. [Online]. Available: http://www.ieor.iitb.ac.in/files/faculty/nh/tailprobTR.pdf

[23] S. K. Sinha, N. Rangaraj, and N. Hemachandra, "Pricing surplus server capacity for mean waiting time sensitive customers," *European Journal of Operational Research*, vol. 205, pp. 159–171, August 2010.

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