

On mean waiting time completeness and equivalence of EDD and HOL-PJ dynamic priority in 2-class M/G/1 queue

Manu K. Gupta^{*}
Industrial Engineering and
Operations Research,
IIT Bombay, Powai,
Mumbai 400076, India
manu.gupta@iitb.ac.in

N. Hemachandra[†]
Industrial Engineering and
Operations Research,
IIT Bombay, Powai,
Mumbai 400076, India
nh@iitb.ac.in

J. Venkateswaran
Industrial Engineering and
Operations Research,
IIT Bombay, Powai,
Mumbai 400076, India
jayendran@iitb.ac.in

ABSTRACT

This paper identifies two different parametrized dynamic priority queue disciplines, earliest due date (EDD) based and head of line priority jump (HOL-PJ), which are found to be mean waiting time complete in two class M/G/1 queue. An explicit one-to-one non linear transformation is obtained between earliest due date and delay dependent priority policy. Mean waiting time equivalence between these queue disciplines is established. Motivation behind the mean completeness and equivalence results is discussed from optimal control perspective. Notion of minmax fairness is introduced and it is argued that a simple global FCFS policy is the only solution for minmax fairness problem in two class by exploiting completeness in the structure of EDD based dynamic priority. Further, these completeness results are used to propose a simpler way for developing optimal control policy in celebrated c/ρ rule for two class M/G/1 queues.

Categories and Subject Descriptors

G.3 [Probability and Statistics]: Queueing theory; C.4 [Performance of Systems]: performance analysis; I.2.8 [Problem Solving, Control Methods, and Search]: Control theory, Scheduling

General Terms

Theory, Performance

Keywords

Parametrized dynamic priority, optimal control, multi-class queue, achievable region

^{*}Manu K. Gupta is a Ph.D. student at IE & OR, IIT Bombay.

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

VALUETOOLS 2014, December 09-11, Bratislava, Slovakia

Copyright © 2015 ICST 978-1-63190-057-0

DOI 10.4108/icst.valuetools.2014.258212

1. INTRODUCTION

Multi-class queues offer a flexible way of modelling a variety of complex dynamic real world problems where customers arrive over time and each class of customers have different quality of service requirement. Choice of queueing discipline plays a significant role in such scenarios. Different types of priority schemes are possible to schedule multiple classes of customers for service at a common resource. Suppose absolute or strict priority is given to one class of customers, then the lower priority class may starve for resource access for very long time.

Various types of parametrized dynamic priority rules are possible to overcome the starvation of lower priority classes. Kleinrock [15] proposed Delay Dependent Priority (DDP) scheme based on delay in queues. Other parametrized dynamic priority rules are Earliest Due Date (EDD) based dynamic priority (see [10]) and Head Of Line Priority Jump (HOL-PJ) [18]. Relative priority, recently proposed in [14], is yet another class of parametrized dynamic priority based on numbers in queue.

Another community of researchers focused on dynamic control of 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 [4], [5] and [17]). Optimal control policy for certain nonlinear optimization problems for two class work conserving queueing systems is derived in [13]. A finite step algorithm for optimally price server's surplus capacity in M/G/1 queue is proposed in [24] assuming that a certain conjecture is true. A complete class of parametrized (delay dependent) dynamic priority is used to schedule customers across classes in this pricing problem. Some further results on sensitivity of optimal operating parameters for the model described in [24] can be seen in [11]. Optimal control policy in two class polling system (non work conserving) for certain optimization problems using achievable region approach is recently developed in [22]. Note that in each of these, a suitable class of parametrized dynamic priority schemes are used; to ensure optimality, such classes have to be complete as discussed below.

For multi class single server priority queue, average waiting time vectors form a nice geometric structure (polytope)

driven by conservation laws under certain scheduling assumptions (see [8], [23]). This kind of structure also helps if one wants to optimize over all scheduling policies. Researchers in this field have come up with geometrical structure of achievable region in case of multiple servers and even for some networks (See [9], [6]). Unbounded achievable region of mean waiting time is recently identified in [22] for two class deterministic polling system.

Consider a service system for which demand consists of different types of jobs, for example computer systems, production facility, etc. Managers of such systems are often interested in selecting a scheduling strategy which will achieve a certain pre-specified performance measure for different job classes. A problem of more practical significance is to isolate the family of scheduling policies which have the property that any performance measure requirement that is realizable, can be obtained by a strategy from the family. A suitable policy from this isolated class achieves all possible performance vectors and this gives rise to the idea of completeness. In this paper, we deal with mean waiting time performance measure.

In a single server multi-class queue, a parametrized scheduling policy is called *complete* by Mitrani and Hine [20] if it achieves all possible vectors of mean waiting times by changing the parameter of scheduling policy. Discriminatory processor sharing (DPS) class of parametrized dynamic priority is identified as *complete* policy in the case of two class M/G/1 queue and used to find the optimal control policy in [13]. This idea of completeness is also useful in designing synthesis algorithms where service provider wants to design a system with certain service level (mean waiting time) for each class. Federgruen and Groenvelt [9] came up with synthesis algorithm using the completeness of mixed dynamic priority which is based on delay dependent priority proposed by Kleinrock [15].

In this paper, completeness via mean waiting time equivalence of parametrized dynamic priority rules (EDD and HOL-PJ) in two class M/G/1 queue is established. Completeness of relative priority is discussed in [12]. Some applications of these ideas are also presented.

This paper is organized as follows. Section 2 describes the idea of completeness and different types of parametrized dynamic priorities. Section 3 presents the results on mean completeness and mean equivalence between EDD and HOL-PJ dynamic priority. Section 4 discusses the applications of these completeness results where global FCFS scheduling is identified as minmax fair policy and the idea of alternate proof of c/ρ rule is given. Section 5 ends with conclusions and direction for future avenues.

2. COMPLETENESS AND PARAMETRIZED DYNAMIC PRIORITY DESCRIPTION

In this section, we briefly describe the idea of completeness and different types of parametrized dynamic priorities in M/G/1 queue with N customer's classes.

Consider a single server system with N different classes of customers arriving in an independent Poisson streams each

with rate λ_i and mean service time be $1/\mu_i$ for class i , $i = 1, 2, \dots, N$. Let $\rho_i = \lambda_i/\mu_i$ and $\rho = \rho_1 + \rho_2 + \dots + \rho_N$. Assume that $\rho < 1$, i.e., system attains steady state. Let the service time variance for each class be finite, i.e., $\sigma_i^2 < \infty$. The performance of the system is measured by vector $\mathbf{W} = (w_1, w_2, \dots, w_N)$, where w_i is the expected waiting/delay time (time spent in queue not including service time) of class i jobs in steady state (see [28]). It is obvious that all performance vectors are not possible; for example $\mathbf{W} = \mathbf{0}$ is not possible (see [20]). We restrict our attention to system where the following conditions are satisfied.

1. Service discipline is non-preemptive.
2. Server is not idle when there are jobs in the system (work conserving).
3. Information about remaining processing time does not affect the system in any way (non anticipative).

Under the above mentioned conditions, Kleinrock's conservation law holds [16]:

$$\sum_{i=1}^N \rho_i w_i = \frac{\rho W_0}{1 - \rho} \quad (1)$$

where $W_0 = \sum_{i=1}^N \frac{\lambda_i}{2} \left(\sigma_i^2 + \frac{1}{\mu_i^2} \right)$.

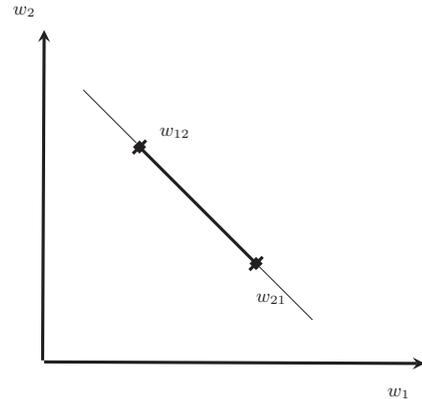


Figure 1: Achievable performance vectors in a two class M/G/1 queue [19]

In case of two classes, all achievable performance vectors $\mathbf{W} = (w_1, w_2)$ form the points lying on a *straight line segment* defined by conservation law. There are two special points on this line (refer Figure 1), \mathbf{w}_{12} and \mathbf{w}_{21} . These two points correspond to the expected waiting times when class 1 and class 2 are given strict priority, respectively. The priority policy (1,2) yields the lowest possible average waiting time for class 1 and the highest possible one for class 2; the situation is reversed with the policy (2,1). Thus, no point to the left of (1,2) or to the right of (2,1) can be achieved. Clearly, every point in the line segment is a convex combination of the extreme points \mathbf{w}_{12} and \mathbf{w}_{21} . Hence

$\alpha \mathbf{w}_{12} + (1 - \alpha) \mathbf{w}_{21}$ achieves all the points in line segment for $\alpha \in [0, 1]$.

Equation (1) defines a *hyperplane* in N -dimensional space of \mathbf{W} . In general, all achievable performance vectors lie in $(N-1)$ -dimensional *hyperplane* for N classes of customers. There are $(N)!$ extreme points, corresponding to $(N)!$ non-preemptive strict priority policies. Hence, the set of achievable performance vectors form a *polytope* with these *vertices*.

If for a given scheduling strategy S , the value of performance vector is W , we say that S achieves W . A family of scheduling strategy is called **complete** if it achieves the polytope described above (see [20]). The set of all scheduling strategies is trivially a complete family; thus one is interested in a subset of all strategies, parametrized suitably, but complete. In this paper, we identify two families of parametrized scheduling strategies which are complete. We now describe different types of parametrized dynamic priorities from literature.

2.1 Delay Dependent Priority

Delay dependent priorities (DDP) were first introduced by Kleinrock [15]. The logic of this discipline works as follows. Each customer class is assigned a queue discipline parameter, b_i , $i \in \{1, \dots, N\}$, $0 \leq b_1 \leq b_2 \leq \dots \leq b_N$. Higher the value of b_i , higher the priority for class i . The instantaneous dynamic priority for a customer of class i at time t , $q_i(t)$, is given by:

$$q_i(t) = (t - \tau) \times b_i, i = 1, 2, \dots, N. \quad (2)$$

where τ is the arrival time of the customer. After the current customer is served, the server will pick the customer with the highest instantaneous dynamic priority parameter $q_i(t)$ for service. Ties are broken using First-Come-First-Served rule. Hence, according to this discipline, the higher priority customers gain higher dynamic priority at higher rate.

Mean waiting time for p^{th} class of customers under this discipline is given by the following recursion [15]:

$$E(W_p) = \frac{W_0}{1 - \rho} + \frac{\sum_{i=1}^{p-1} \rho_i E(W_i) \left(1 - \frac{b_i}{b_p}\right)}{1 - \sum_{i=p+1}^N \rho_i \left(1 - \frac{b_p}{b_i}\right)} \quad (3)$$

where $\rho_i = \frac{\lambda_i}{\mu_i}$, $\rho = \sum_{p=1}^N \rho_i$ and $W_0 = \sum_{p=1}^N \frac{\lambda_p}{2} \left(\sigma_p^2 + \frac{1}{\mu_p^2}\right)$ and $0 \leq \rho < 1$.

Federgruen and Gruenevelt [9] proposed a synthesis algorithm by exploiting the completeness of mixed dynamic priority which is based on delay dependent priority. Mean waiting time expression in case of two classes of customers for DDP is given in Appendix A.

2.2 Earliest Due Date Dynamic Priority

This type of parametrized 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 delay dependent priority. Each class i

has a constant urgency number u_i (weights) associated with it. Without loss of generality, classes are numbered so that $u_1 \leq u_2 \leq \dots \leq u_N$. When a customer from class i arrives at the system at time t_i , customer is assigned a real number $t_i + u_i$. The server chooses the next customer to go into service, from those present in queue, as the one with minimum value of $\{t_i + u_i\}$. Let W_r denote the waiting time of class r jobs. In steady state, $E(W_r)$ is given by [10]

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

for $r = 1, \dots, N$. Here $E(W) = \frac{W_0}{(1-\rho)}$ and ρ_i is the traffic due to class i . Note that above recursion is not a closed form equation; however, this helps in deriving mean completeness and equivalence results.

The formulation of the scheduling discipline in terms of urgency numbers facilitates various interpretations of the model. One primary interpretation of urgency numbers, u_i , correspond to the interval until the due date is reached. This model leads to a unified theory of scheduling with earliest due dates, which is an area of great practical importance (see [10]).

2.3 Head of Line Priority Jump (HOL-PJ)

This is another type of parametrized dynamic priority policy proposed in [18]. The fundamental principle of HOL-PJ is to give priority to the customers having the largest queueing delay in excess of its delay requirement. In HOL-PJ, an explicit priority is assigned to each class; the more stringent the delay requirement of the class, the higher the priority. From the server's point of view, HOL-PJ is the same as head of line (HOL) strict priority queue. Unlike HOL, the priorities of customers increase as their queueing delay increases relative to their delay requirements. This is performed by customer *priority jumping* (PJ) mechanism.

We briefly describe the practical significance of this model as pointed out in [18]. This model can be used in an integrated packet switching node serving multiple classes of delay sensitive traffic (eg. voice and video traffic). Implementation of this discipline is relatively simple and the processing overhead is minimal.

Consider a single server serving N classes of customers. Let D_j , $j = 1, 2, \dots, N$ be the delay requirement for class j customers where $0 < D_1 < D_2 < \dots < D_N \leq \infty$. Class 1 has the most stringent delay requirement and class N the least; class 1 has the highest priority and class N the least. T_j , $j = 2, 3, \dots, N$ is set to $D_j - D_{j-1}$. If a customer is still in queue after a period of time T_j , it jumps to the tail of queue $j - 1$. Figure 2 illustrates the operation of HOL-PJ. Excessive delay of a customer is defined as its queueing delay in excess of its original delay requirement. It is concluded in [18] that all the customers are queued according to largeness of their excessive delay. Mean waiting time for class k customers in HOL-PJ queueing discipline is derived

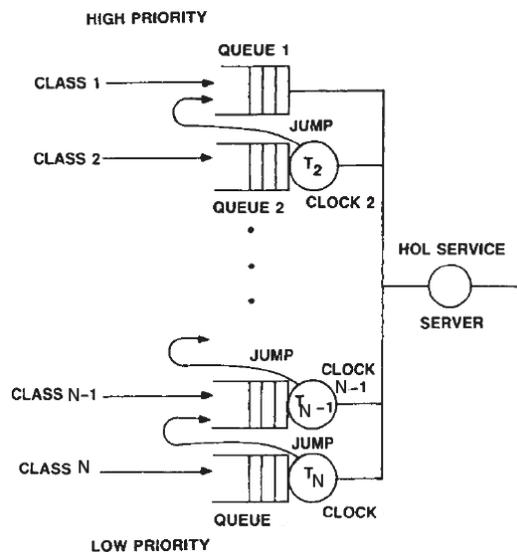


Figure 2: Head-of-line with priority jump [18]

in [18] as:

$$E(W_k) = E(W_{FIFO}) - \sum_{j=k+1}^N \rho_j \int_0^{\sum_{l=k+1}^j T_l} P(W_j > t) dt + \sum_{j=1}^{k-1} \rho_j \int_0^{\sum_{l=j+1}^k T_l} P(W_k > t) dt \quad (5)$$

Since $T_j = D_j - D_{j-1}$. This gives

$$E(W_k) = E(W_{FIFO}) + \sum_{j=1}^{k-1} \rho_j \int_0^{D_k - D_j} P(W_k > t) dt - \sum_{j=k+1}^N \rho_j \int_0^{D_j - D_k} P(W_j > t) dt \quad (6)$$

Here $E(W_{FIFO})$ is $\frac{W_0}{(1-\rho)}$. Note that above recursion is again not a closed form equation; however, this helps in deriving mean completeness and equivalence results.

3. MEAN EQUIVALENCE AND MEAN COMPLETENESS IN TWO CLASSES

In this section, we prove the mean completeness of EDD and HOL-PJ parametrized dynamic priorities for two class M/G/1 queue. We obtain the closed form expression for explicit one-to-one non-linear transformation from DDP class to EDD. DDP is known to be complete from [9]. Hence completeness of EDD follows. Completeness of HOL-PJ is argued by identifying the similarity in expressions of mean waiting time for EDD and HOL-PJ. Relative priority is another complete dynamic parametrized policy. More details about the completeness of relative priority can be seen in [12].

3.1 Motivation

These results on mean completeness and equivalence of dynamic priority are substantially useful in the theory of opti-

mal control. We discuss some of the advantages here.

The above discussed parametrized dynamic priority policies are shown to be mean complete and mean equivalent. Hence optimality of control policy will not be lost if it is posed over any of the complete parametrized dynamic priority discipline. One can suitably choose dynamic priority discipline according to application domain, on hand.

EDD dynamic priority often finds applications in scheduling for project management scenarios, where multiple jobs need to be completed before their respective deadlines, using shared resources. Due to parametrized (by urgency number) nature of EDD dynamic priority, appropriate urgency numbers can be designed for each type of job in a given project management problem.

HOL-PJ is *computationally most efficient* among all class of dynamic priorities discussed here. No additional processing delay is involved with HOL-PJ compared to HOL. Implementation of the priority jump (PJ) mechanism would require at each priority queue a local clock, a list of the arrival times of the customers currently in queue, and circuitry for the timing mechanism to make the customer at the head of the queue jump (see [18]). It is suggested in [18] that this information can be provided without incurring too much system complexity and/or cost. Note that this dynamic priority will have *less switching rate* as compare to other dynamic priorities due to its mechanism being similar to HOL.

3.2 Equivalence and Completeness Results

In this section, we discuss the mean completeness and mean equivalence of EDD based and HOL-PJ dynamic priority.

3.2.1 EDD based dynamic priority

In case of two classes, the expected waiting time is [10, Theorem 2]:

$$E(W_h) = E(W) - \rho_l \int_0^u P(T_h[W] > y) dy \quad (7)$$

$$E(W_l) = E(W) + \rho_h \int_0^u P(T_h[W] > y) dy \quad (8)$$

Here index h and l are for higher and lower priority classes respectively. u_l and u_h are the weights associated with lower and higher classes, where $u = u_l - u_h \geq 0$. $T_h[W] = \lim_{t \rightarrow \infty} T_h[W(t)]$ and $T_h[W(t)]$ 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. $W(t) \rightarrow W$ as $t \rightarrow \infty$.

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

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

Consider the more general setting (in the view of completeness) with this type of priority where $u_1, u_2 \geq 0$ be the weights associated with class 1 and class 2. Let $\bar{u} = u_1 - u_2$. Thus \bar{u} can take value in interval $[-\infty, \infty]$. Class 1 will have higher or lower priority depending on \bar{u} being negative

or positive. By using equations (7) and (8), mean waiting time for this general setting in case of two classes can be written as:

$$E(W_1) = E(W) + \rho_2 \left[\int_0^{\bar{u}} P(T_2(W) > y) dy \mathbf{1}_{\{\bar{u} \geq 0\}} - \int_0^{-\bar{u}} P(T_1(W) > y) dy \mathbf{1}_{\{\bar{u} < 0\}} \right] \quad (9)$$

$$E(W_2) = E(W) + \rho_1 \left[\int_0^{\bar{u}} P(T_2(W) > y) dy \mathbf{1}_{\{\bar{u} \geq 0\}} - \int_0^{-\bar{u}} P(T_1(W) > y) dy \mathbf{1}_{\{\bar{u} < 0\}} \right] \quad (10)$$

Note that $\bar{u} = -\infty$ and $\bar{u} = \infty$ provide corresponding mean waiting times when strict higher priority is given to class 1 and class 2 respectively. DDP also achieves these strict priority mean waiting time at $\beta = 0$ and $\beta = \infty$ (see Appendix A). Hence we suspect a one-to-one transformation from DDP to EDD priority. We prove our intuition below and find the explicit nonlinear transformation.

LEMMA 3.1. *Delay dependent priority and earliest due date priority are mean equivalent in two classes and their priority parameters β and \bar{u} are related as:*

$$\beta = \frac{\mu - \lambda}{\lambda_2 + \frac{\rho_2}{\mu W_0}(\mu - \lambda)\lambda_1 \tilde{I}(\bar{u})} \left[\frac{\lambda_2}{\mu - \lambda} - \frac{\rho_2(\mu - \lambda_1)\tilde{I}(\bar{u})}{\mu W_0} \right] \\ \times \mathbf{1}_{\{-\infty \leq \bar{u} \leq 0\}} + \frac{\lambda_2 \left(\frac{\mu W_0}{\mu - \lambda} + \rho_2 I(\bar{u}) \right)}{\frac{\mu \lambda_2 W_0}{\mu - \lambda} - \rho_2(\mu - \lambda_2)I(\bar{u})} \mathbf{1}_{\{0 \leq \bar{u} \leq \infty\}}$$

where integrals $\tilde{I}(\bar{u}) = \int_0^{-\bar{u}} P(T_1(W) > y) dy$ and $I(\bar{u}) = \int_0^{\bar{u}} P(T_2(W) > y) dy$.

PROOF. See Appendix B. \square

Note that β is a monotone function of $\tilde{I}(\bar{u})$, and $\tilde{I}(\bar{u})$ is a monotone function of \bar{u} . Due to this monotonicity, there is a one-to-one transformation between \bar{u} and β . Since DDP is a mean complete dynamic priority discipline in case of two classes [9], EDD is also mean complete for two classes of customers. Thus we have following result:

THEOREM 3.2. *EDD with two classes of priorities is mean complete.*

A separate proof of above theorem is also given in Appendix B. This proof gives one-to-one correspondence between \bar{u} of EDD and α , convex combination parameter, of conservation law (as discussed in Section 2).

3.2.2 HOL-PJ dynamic priority

It can be observed from Equation (4) and (6) that the mean waiting time expression in HOL-PJ is same as that in EDD priority policy. Urgency number and overdue in EDD correspond to delay requirement and excessive delay in HOL-PJ.

Similar to EDD, we consider the more general setting in HOL-PJ where $D_1, D_2 \geq 0$ is the delay requirement associated with class 1 and class 2. Let $\bar{D} = D_1 - D_2$ be the parameter associated with HOL-PJ similar to \bar{u} of EDD. Note that the mean waiting time expression for HOL-PJ is same as EDD. Hence, we have following lemma from our previous result on mean equivalence and completeness of two class EDD dynamic priority.

THEOREM 3.3. *There is a one-to-one non-linear transformation for mean waiting time between HOL-PJ and DDP discipline, and hence HOL-PJ is mean complete in two class M/G/1 queues.* \blacksquare

4. APPLICATIONS

In this section, we consider global FCFS policy in case of two class parametrized queue. We discuss the notion of minmax fairness and argue that the only solution of minmax fairness problem is global FCFS policy by posing this problem as continuous semi-infinite program (see [3],[1]). We further illustrate the usefulness of these completeness results in finding the optimal scheduling policy for linear cost objective in two class M/G/1 queue.

4.1 Global FCFS as minmax fair policy

Global FCFS is the policy where customers are served according to the order of their arrivals irrespective of the priority classes. Global FCFS is achieved by DDP, EDD, and HOL-PJ based priority by keeping all b_i 's, u_i 's and D_i 's equal. The mean waiting time for each class is equal in global FCFS policy and is given by $\frac{W_0}{(1-\rho)}$. In the case of two class parametrized queueing system, global FCFS policy is realized by delay dependent priority with $\beta = 1$, by EDD with $\bar{u} = 0$ and by HOL-PJ dynamic priority with $\bar{D} = 0$.

In addition to the focus on performance metrics such as response time, queue length, throughput, etc., it is often important to ensure that the customers (jobs) across classes are *fairly* treated. A vast literature has evolved on the refinement of the notion of fairness (as in [26] [27] [2]). We define fairness for multi-class queues: *minimize the maximum dissatisfaction of the system*. Here dissatisfaction of a customer is quantified in terms of mean waiting time of that customer's class. Mathematically, it can be written as:

$$\min_{\alpha \in \mathcal{F}} \max_{i \in \mathcal{I}} E(W_\alpha^{(i)}) \quad (11)$$

where \mathcal{I} is set of classes and \mathcal{F} is given class of work conserving, non pre-emptive and non anticipative queueing discipline. $E(W_\alpha^{(i)})$ is the mean waiting time for class i customers when scheduling policy $\alpha \in \mathcal{F}$ is employed. Minmax problem can also be described as an optimization problem via lexicographic ordering (see [21], [25]). We solve this problem by rewriting it as *continuous semi-infinite program* [1]:

$$\min_{\alpha} \epsilon_\alpha$$

$$E(W_\alpha^{(i)}) \leq \epsilon_\alpha \quad \alpha \in \mathcal{F}, i \in \mathcal{I} \quad (12)$$

$$\epsilon_\alpha \geq 0, \quad (13)$$

We now consider two classes of customers. Since EDD is a *complete* parametrized dynamic priority discipline in case of

two classes, it can be re-written as

$$\min_{\bar{u}} \epsilon_{\bar{u}}$$

$$E(W_{\bar{u}}^{(i)}) \leq \epsilon_{\bar{u}} \quad \bar{u} \in [-\infty, \infty], i \in \mathcal{I} \quad (14)$$

$$\epsilon_{\bar{u}} \geq 0, \quad (15)$$

$$\sum_{i \in \mathcal{I}} \rho_i E(W_{\bar{u}}^{(i)}) = \frac{\rho W_0}{(1-\rho)} \quad (16)$$

Constraint (16) is necessary as parametrized policy should satisfy conservation law. Let $E(W_0^{(i)})$ be the mean waiting time of class i with scheduling policy $\bar{u} = 0$. It follows from Equation (9) and (10) that $E(W_0^{(i)}) = \frac{W_0}{(1-\rho)}$ for $i = 1, 2$. Note that $\bar{u} = 0$ is one of the feasible solution to the above optimization problem. We now argue the optimality of policy $\bar{u} = 0$. It is clear from Equations (9) and (10) that with a policy $\bar{u} \neq 0$ (say \bar{u}') the mean waiting time of one of the class (class 1 or class 2) will be more than $E(W_0^{(i)})$. Hence $\epsilon_{\bar{u}'}$ corresponding to that policy will be always more than $\epsilon_{\bar{u}=0}$ (due to Constraint (14)). Hence, the minima of the semi-infinite program will be given by the policy $\bar{u} = 0$. Similarly, the semi-infinite program can be solved by posing it over HOL-PJ parametrized discipline and $\bar{D} = 0$ will be the solution. $\bar{u} = 0$ and $\bar{D} = 0$ correspond to global FCFS. Note that the global FCFS is achieved by relative priority also if all p_i 's are equal for multi-class queues ($p_1 = p_2 = 1/2$ achieves mean waiting time for global FCFS for two classes of customers).

Now, we find the weights given to the extreme points of line segment of Figure 1 to achieve global FCFS in case of two classes. Consider weights $\alpha_1 = \frac{(1-\rho_1)}{(2-\rho_1-\rho_2)}$ to class 1 and $\alpha_2 = \frac{(1-\rho_2)}{(2-\rho_1-\rho_2)}$ to class 2. On simplifying, we have

$$[\alpha_1 \quad \alpha_2] \begin{bmatrix} E(W_{12}^{(1)}) & E(W_{12}^{(2)}) \\ E(W_{21}^{(1)}) & E(W_{21}^{(2)}) \end{bmatrix} = \begin{bmatrix} \frac{W_0}{1-\rho} & \frac{W_0}{1-\rho} \end{bmatrix} \quad (17)$$

where $E(W_{12}^{(i)})$ is the mean waiting time for class i , $i = 1, 2$, when class 1 has strict priority over class 2. Note that with two classes, we have exactly *one and unique* pair of weights given to extreme points to get the global FCFS point in the interior of the polytope (line segment).

4.2 Optimal Scheduling Policy

It is well known in literature (see [19, page 110], [29]) that linear cost objective function of mean waiting time, $C = \sum_{i=1}^K c_i E(W_i)$, is minimized by c/ρ rule if all work conserving non-preemptive scheduling policies are considered. Here c_i and $E(W_i)$ are the cost and mean waiting time associated with class i respectively. This rule states that the optimal scheduling discipline with respect to objective C is strict priority scheme where priority is given in the order of the ratios c_i/ρ_i .

We give the idea for the proof of this result in two class M/G/1 queue using completeness arguments discussed in this paper. Consider the objective

$$\mathbf{P1} \quad \min_{\alpha \in \mathcal{F}} c_1 E(W_{\alpha}^{(1)}) + c_2 E(W_{\alpha}^{(2)})$$

where \mathcal{F} is set of all work conserving, non pre-emptive and non anticipative scheduling policies. Note that optimizing over \mathcal{F} is same as optimizing over set of EDD priority as this class of priority is mean complete by Theorem 3.2. Hence problem **P1** is equivalent to **P2** defined below:

$$\mathbf{P2} \quad \min_{\bar{u} \in [-\infty, \infty]} c_1 E(W_{\bar{u}}^{(1)}) + c_2 E(W_{\bar{u}}^{(2)})$$

Note that above optimization problem **P2** can be easily solved to yield the optimal c/ρ rule. Relative priority is also identified as a complete class and optimal c/ρ rule is obtained using relative priority in [12].

5. CONCLUSIONS AND FUTURE WORK

The notion of completeness is discussed for work conserving queueing systems. Certain parametrized dynamic priorities (EDD and HOL-PJ) are shown to be mean complete in two class M/G/1 queue. Mean equivalence between EDD, DDP and HOL-PJ is established. An explicit one-to-one nonlinear transformation is given between EDD and DDP. Significance of these results in optimal control of queueing system is also discussed. The notion of minmax fairness in queues is introduced and solution of this minmax fairness is obtained by exploiting the mean completeness of EDD policy. Importance of global FCFS policy is discussed and it is shown as minmax fair policy in two class M/G/1 queue. It will be interesting to extend these ideas in higher dimensions (N class queue). Another fascinating future avenue will be to come up with synthesis algorithm with different parametrized dynamic priority. Other parametrized dynamic policy classes can also be explored.

Acknowledgement

We thank the reviewers for their comments that helped us to improve the presentation. Manu K. Gupta was partially supported by a Teaching Assistantship offered by MHRD, Government of India.

6. REFERENCES

- [1] ANDERSON, E. J., AND NASH, P. *Linear Programming in Infinite Dimensional Spaces: Theory and Applications*. Academic Press, 1980.
- [2] AVI-ITZHAK, B., AND LEVY, H. On measuring fairness in queues. *Adv. of Applied Probability* 36, 3 (2004), 919–936.
- [3] BERTSEKAS, D. P. *Nonlinear programming*. Athena Scientific, 1999.
- [4] BERTSIMAS, D. The achievable region method in the optimal control of queueing systems; formulations, bounds and policies. *Queueing systems* 21, 3-4 (1995), 337–389.
- [5] BERTSIMAS, D., AND NIÑO-MORA, J. Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Mathematics of Operations Research* 21, 2 (1996), 257–306.
- [6] BERTSIMAS, D., PASCHALIDIS, I., AND TISTSIKLIS, J. N. Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *The Annals of Applied Probability* 4 (1994), 43–75.

- [7] COBHAM, A. Priority assignment in waiting-line problems. *Operations Research* 9 (1954), 383–387.
- [8] COFFMAN JR, E., AND MITRANI, I. A characterization of waiting time performance realizable by single-server queues. *Operations Research* 28, 3-part-ii (1980), 810–821.
- [9] FEDERGRUEN, A., AND GROENEVELT, H. M/G/c queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules. *Management Science* 9 (1988), 1121–1138.
- [10] GOLDBERG, H. M. Analysis of earliest due date scheduling rule in queueing systems. *Mathematics of Operations Research* 2 (1977), 145–154.
- [11] GUPTA, M. K., HEMACHANDRA, N., RAGHAV, B. S., AND VENKATESWARAN, J. On a conjecture and performance of a two class delay dependent priority queue arising from pricing surplus server capacity for mean waiting time sensitive customers. Tech. rep., IIT Bombay (submitted), 2014. <http://www.ieor.iitb.ac.in/files/ConjectureTR.pdf>.
- [12] GUPTA, M. K., HEMACHANDRA, N., AND VENKATESWARAN, J. On completeness and equivalence of some dynamic priority schemes. Tech. rep., IIT Bombay, 2014. <http://www.ieor.iitb.ac.in/files/completeTR.pdf>.
- [13] HASSIN, R., PUERTO, J., AND FERNÁNDEZ, F. R. The use of relative priorities in optimizing the performance of a queueing system. *European Journal of Operational Research* 193, 2 (2009), 476–483.
- [14] HAVIV, M., AND VAN DER WAL, J. Waiting times in queues with relative priorities. *Operations Research Letters* 35 (2007), 591–594.
- [15] KLEINROCK, L. A delay dependent queue discipline. *Naval Research Logistics Quarterly* 11 (September–December 1964), 329–341.
- [16] KLEINROCK, L. A conservation law for wide class of queue disciplines. *Naval Research Logistics Quarterly* 12 (June–September 1965), 118–192.
- [17] LI, C.-P., AND NEELY, M. J. Delay and rate-optimal control in a multi-class priority queue with adjustable service rates. In *INFOCOM, Proceedings IEEE* (2012), pp. 2976–2980.
- [18] LIM, Y., AND KOBZA, J. E. Analysis of delay dependent priority discipline in an integrated multiclass traffic fast packet switch. *IEEE Transactions on Communications* 38 (5) (1990), 351–358.
- [19] MITRANI, I. *Probabilistic Modelling*. Cambridge University Press, 2004.
- [20] MITRANI, I., AND HINE, J. Complete parametrized families of job scheduling strategies. *Acta Informatica* 8 (1977), 61–73.
- [21] OSBORNE, M. J., AND RUBINSTEIN, A. *A course in game theory*. MIT press, 1994.
- [22] RAWAL, A., KAVITHA, V., AND GUPTA, M. K. Optimal surplus capacity utilization in polling systems via fluid models. In *WiOpt, Proceedings IEEE* (2014), pp. 381–388.
- [23] SHANTHIKUMAR, J. G., AND YAO, D. D. Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research* 40, 3-supplement-2 (1992), S293–S299.
- [24] SINHA, S. K., RANGARAJ, N., AND HEMACHANDRA, N. Pricing surplus server capacity for mean waiting time sensitive customers. *European Journal of Operational Research* 205, 1 (2010), 159–171.
- [25] VANAM, K. C., AND HEMACHANDRA, N. Some excess-based solutions for cooperative games with transferable utility. *International Game Theory Review* 15, 04 (2013).
- [26] WIERMAN, A. Fairness and classifications. *Performance Evaluation Review* 34, 4 (2007), 4–12.
- [27] WIERMAN, A. *Scheduling for today’s computer systems: Bridging theory and practice*. PhD thesis, Carnegie Mellon University, 2007.
- [28] WOLFF, R. W. *Stochastic modelling and the theory of queues*. Englewood Cliffs, NJ, 1989.
- [29] YAO, D. D. Dynamic scheduling via polymatroid optimization. In *Performance Evaluation of Complex Systems: Techniques and Tools*. Springer, 2002, pp. 89–113.

APPENDIX

A. DELAY DEPENDENT PRIORITY (DDP)

It is clear from Equation (3) that average waiting time expressions in delay dependent priority (DDP) depends on ratios b_i only. Define $\beta := b_2/b_1$. Average waiting time expression for class 1, $E(W_\beta^{(1)})$, and for class 2, $E(W_\beta^{(2)})$, in two customer’s classes DDP can be derived using Equation (3) as follows:

$$E(W_\beta^{(1)}) = \frac{\lambda\psi(\mu - \lambda(1 - \beta))}{\mu(\mu - \lambda)(\mu - \lambda_1(1 - \beta))} \mathbf{1}_{\{\beta \leq 1\}} + \frac{\lambda\psi}{(\mu - \lambda)(\mu - \lambda_2(1 - \frac{1}{\beta}))} \mathbf{1}_{\{\beta > 1\}} \quad (18)$$

$$E(W_\beta^{(2)}) = \frac{\lambda\psi}{(\mu - \lambda)(\mu - \lambda_1(1 - \beta))} \mathbf{1}_{\{\beta \leq 1\}} + \frac{\lambda\psi(\mu - \lambda(1 - \frac{1}{\beta}))}{\mu(\mu - \lambda)(\mu - \lambda_2(1 - \frac{1}{\beta}))} \mathbf{1}_{\{\beta > 1\}} \quad (19)$$

where $\lambda = \lambda_1 + \lambda_2$, $\psi = (1 + \sigma^2\mu^2)/2$ and $\mathbf{1}_{\{\cdot\}}$ is indicator function. Here μ is the common service rate for both the classes and σ^2 is the variance of service time. Note that $\beta = 0$ and $\beta = \infty$ give the corresponding mean waiting time when strict higher priority is given to class 1 and class 2 respectively. Also $\beta = 1$ achieves mean waiting time for global FCFS scheduling policy.

B. PROOFS OF LEMMA AND CLAIMS

Proof of Lemma 3.1: Global FCFS, strict priorities are achieved by $\bar{u} = 0, -\infty, \infty$ in EDD and $\beta = 1, 0, \infty$ in DDP respectively. This can be verified using the expressions (Equations (9), (10), (18) and (19)) of mean waiting times. On considering following two cases, we have

1. $-\infty \leq \bar{u} \leq 0$ and $0 \leq \beta \leq 1$: On equating the mean waiting time for class 1 under these two dynamic priorities, we have (using Equations (18) and (9)):

$$E(W) - \rho_2 \int_0^{-\bar{u}} P(T_1(W) > y) dy = \frac{\lambda\psi(\mu - \lambda(1 - \beta))}{\mu(\mu - \lambda)(\mu - \lambda_1(1 - \beta))} \quad (20)$$

On simplifying the above equation for β , we have

$$\begin{aligned} \beta &= \frac{\mu - \lambda}{\lambda_2 + \frac{\rho_2}{\mu W_0}(\mu - \lambda)\lambda_1 \tilde{I}(\bar{u})} \times \\ &\quad \left[\frac{\mu - \lambda_1}{\mu(1 - \rho)} - \frac{\rho_2(\mu - \lambda_1)\tilde{I}(\bar{u})}{\mu W_0} - 1 \right] \\ &= \frac{\mu - \lambda}{\lambda_2 + \frac{\rho_2}{\mu W_0}(\mu - \lambda)\lambda_1 \tilde{I}(\bar{u})} \times \\ &\quad \left[\frac{\lambda_2}{\mu - \lambda} - \frac{\rho_2(\mu - \lambda_1)\tilde{I}(\bar{u})}{\mu W_0} \right] \quad (21) \end{aligned}$$

as $\bar{u} \rightarrow 0, \tilde{I}(\bar{u}) \rightarrow 0$ so $\beta \rightarrow 1$ from above equation. Similarly, as $\bar{u} \rightarrow -\infty, \tilde{I}(\bar{u}) \rightarrow E(T_1(W))$ or $\frac{W_0}{(1 - \rho)(1 - \rho_1)}$ (mean waiting time of class 1 when class 1 has strict priority over class 2). Hence $\beta \rightarrow 0$. So

$$-\infty \leq \bar{u} \leq 0 \Leftrightarrow 0 \leq \beta \leq 1$$

Above relation follows from Equation (21) and by the fact that β is monotonically increasing with \bar{u} as $\tilde{I}(\bar{u})$ is monotonically decreasing.

2. $0 \leq \bar{u} \leq \infty$ and $1 \leq \beta \leq \infty$: Again on equating the mean waiting time for class 1 under these two dynamic priorities by using equations (18) and (9):

$$E(W) + \rho_2 \int_0^{\bar{u}} P(T_2(W) > y) dy = \frac{\lambda\psi}{(\mu - \lambda)(\mu - \lambda_2(1 - \frac{1}{\beta}))} \quad (22)$$

On simplifying the above equation for β , we have

$$\beta = \frac{\lambda_2 \left(\frac{\mu W_0}{\mu - \lambda} + \rho_2 I(\bar{u}) \right)}{\frac{\mu \lambda_2 W_0}{\mu - \lambda} - \rho_2(\mu - \lambda_2)I(\bar{u})} \quad (23)$$

as $\bar{u} \rightarrow 0, I(\bar{u}) \rightarrow 0$ so $\beta \rightarrow 1$ from above equation. Similarly, as $\bar{u} \rightarrow \infty, I(\bar{u}) \rightarrow E(T_2(W))$ or $\frac{W_0}{(1 - \rho)(1 - \rho_2)}$ (mean waiting time of class 1 when class 2 has strict priority over class 1). Hence $\beta \rightarrow \infty$. So

$$0 \leq \bar{u} \leq \infty \Leftrightarrow 1 \leq \beta \leq \infty$$

Above relation follow from Equation (23) and by the fact that β is monotonically increasing with \bar{u} as $\tilde{I}(\bar{u})$ is monotonically increasing.

Note that above analysis considered mean waiting time for class 1. Similar analysis can be done with the mean waiting time of class 2. Hence lemma follows. ■

Proof of Theorem 3.2: Consider the notation $E(W_{12}^{(i)})$ to be the mean waiting time for class $i, i = 1, 2$, when class 1 has strict priority over class 2. Mean waiting time are [7]:

$$E(W_{12}^{(1)}) = \frac{W_0}{1 - \rho_1} \text{ and } E(W_{12}^{(2)}) = \frac{1}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} W_0 \quad (24)$$

Point $W_{12} = (E(W_{12}^{(1)}), E(W_{12}^{(2)}))$ is shown in Figure 1. Similarly, when class 2 has strict priority over class 1. We have

$$E(W_{21}^{(1)}) = \frac{W_0}{(1 - \rho_2)(1 - \rho_1 - \rho_2)} \text{ and } E(W_{21}^{(2)}) = \frac{1}{(1 - \rho_2)} W_0 \quad (25)$$

$W_{21} = (E(W_{21}^{(1)}), E(W_{21}^{(2)}))$ is other extreme point shown in Figure 1. Consider the notation $W^{\alpha, i}$ for class i as $W^{\alpha, i} = \alpha E(W_{12}^{(i)}) + (1 - \alpha)E(W_{21}^{(i)})$. On using Equation (24) and (25), we have

$$W^{\alpha, 1} = \frac{\alpha \rho_2(\rho_1 + \rho_2 - 2) + 1 - \rho_1}{(1 - \rho_1)(1 - \rho_2)(1 - \rho_1 - \rho_2)} W_0 \quad (26)$$

Average waiting time in EDD priority depends on \bar{u} being positive or negative (see equation (9) and (10)), consider the following two cases:

1. $0 \leq \bar{u} \leq \infty$: Expected waiting time for class 1 is given by (using equation (9))

$$\begin{aligned} E(W_1) &= E(W) + \rho_2 \int_0^{\bar{u}} P(T_2(w) > y) dy \\ &= E(W) + \rho_2 I(\bar{u}) \end{aligned}$$

On equating $E(W_1)$ with $W^{\alpha, 1}$ and solving for α , we have

$$\alpha = \frac{1 - \rho_1}{2 - \rho_1 - \rho_2} - \frac{I(\bar{u})(1 - \rho_1)(1 - \rho_2)(1 - \rho)}{W_0(2 - \rho_1 - \rho_2)} \quad (27)$$

$\bar{u} = 0 \Rightarrow I(\bar{u}) = 0$ so $\alpha = \frac{1 - \rho_1}{2 - \rho_1 - \rho_2}$ and $\bar{u} = \infty \Rightarrow I(\bar{u}) = \int_0^{\infty} P(T_2(w) > y) dy = E(T_2(W)) = \frac{W_0}{(1 - \rho)(1 - \rho_2)}$ (See [10, page 152]). On putting back this value of $I(\bar{u})$ in (27), we get $\alpha = 0$. Since $I(\bar{u})$ is monotone increasing in \bar{u} , we have

$$0 \leq \bar{u} \leq \infty \Leftrightarrow 0 \leq \alpha \leq \frac{1 - \rho_1}{2 - \rho_1 - \rho_2} \quad (28)$$

2. $-\infty \leq \bar{u} \leq 0$: Expected waiting time for class 1 is given by (using Equation (9))

$$\begin{aligned} E(W_1) &= E(W) - \rho_2 \int_0^{-\bar{u}} P(T_1(W) > y) dy \\ &= E(W) - \tilde{I}(\bar{u}) \end{aligned}$$

On equating $E(W_1)$ with $W^{\alpha, 1}$ and solving for α , we have

$$\alpha = \frac{(1 - \rho_1)(1 - \rho_2)(1 - \rho)\tilde{I}(\bar{u})}{(2 - \rho_1 - \rho_2)W_0} + \frac{(1 - \rho_1)}{(2 - \rho_1 - \rho_2)} \quad (29)$$

$\bar{u} = 0 \Rightarrow \tilde{I}(\bar{u}) = 0$ so $\alpha = \frac{1 - \rho_1}{2 - \rho_1 - \rho_2}$ and $\bar{u} = -\infty \Rightarrow \tilde{I}(\bar{u}) = \int_0^{\infty} P(T_1(w) > y) dy = E(T_1(W)) = \frac{W_0}{(1 - \rho)(1 - \rho_2)}$ (See [10, page 152]). On putting back this value of $I(\bar{u})$ in (29), we get $\alpha = 1$. Since $\tilde{I}(\bar{u})$ is monotone decreasing in \bar{u} , we have

$$-\infty \leq \bar{u} \leq 0 \Leftrightarrow \frac{1 - \rho_1}{2 - \rho_1 - \rho_2} \leq \alpha \leq 1 \quad (30)$$

So, entire range of α is achieved by unique value of \bar{u} . Similar arguments will work if mean waiting time of other class is considered. Hence result follows. ■