On 2-moment Completeness of Non Pre-emptive, Non Anticipative Work Conserving Scheduling Policies in Some Single Class Queues

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

***Abstract*— Completeness of some scheduling policies with mean waiting time performance measure is used quiet extensively in literature for dynamic control of multi-class queues due to its wide range of applications in computers, communication networks and manufacturing systems. For a single class queue, we introduce the idea of *2-moment completeness* of a parametrized class of policies that also have to be non pre-emptive, non anticipative and work conserving. Significance of this idea lies in the importance of variance (or second moment) of waiting time in any queuing system. Some parametrized classes of policies are identified and shown to be *2-moment complete* for $M/M/1$ queues. Some well known queue disciplines viz random order of service (ROS), Random Assigned Priority (RAP), etc., turn out to be 2-moment incomplete. We introduce a parametrized priority scheme that also turns out to be 2-moment incomplete. Further, few pre-emptive and anticipative scheduling disciplines are shown to have second moment beyond the achievable region of non pre-emptive, non anticipative and work conserving scheduling policies. Some optimal control problems are discussed to illustrate the possible applications of 2-moment complete parametrized set of policies.**

***Index Terms*— parametrized dynamic priority, scheduling disciplines, variance of waiting time, achievable region, optimal control of queues**

I. Introduction

The mean of stationary waiting time of customers in a single class single server queue among all non anticipative work conserving scheduling policies is a constant by Little's law ([1], [2] [3], [4] etc.); this is same as that when, say, 'First come first serve' (FCFS) policy is used. However, the distribution of waiting times among various policies of this class depend on the particular policy used. As a consequence, the variance of stationary waiting times of customer depends on the particular policy used. In particular, it is known that the variance of FCFS policy is least and that of Last Come First Serve (LCFS) is highest among all non anticipative work conserving policies ([5] and [6]). However, when we use a policy which allows the server to know the amount of work a job needs (such policies are called anticipative, in literature) the distribution of stationary waiting times can be quite different. For example, Longest Remaining Processing Time, (LRPT) which is an anticipative and work conserving policy has mean and variance of stationary waiting times higher than that of policies offered by non anticipative work conserving class [7]. Another example of anticipative policy could be the Shortest Processing Time (SPT) which has lower mean than that of FCFS (non anticipative work conserving) policy.

Mean waiting time in a single class single server queue among all non-anticipative work-conserving scheduling policies is a constant if (and only if) the service times are exponential, i.e., the M/M/1 queue. For a general M/G/1 queue, the mean waiting time is constant only if the scheduling policy is additionally required to be non-preemptive. We would like to explore the dependence of variance of waiting times on classes of policies that are parametrized by a single parameter. In fact, borrowing an idea from [8], we also explore the notion of completeness of such parametrized class of policies for second moment (equivalently, variance) of waiting time distributions. We are interested in investigating if there is a one-to-one correspondence between the range of the parameter describing each policy and all possible second moments of waiting times that are incurred when any non pre-emptive, non anticipative work conserving policy is used. Roughly speaking, this means that as the parameter 'sweeps' its domain, the second moment of waiting time also takes all possible values for the class of polices under consideration and hence such polices are good enough for optimization purposes.

Various types of queue disciplines are possible for scheduling in single class queue under the regime of non pre-emptive, non anticipative and work conserving scheduling disciplines. Some popular queue disciplines are FCFS, LCFS, random order of service (ROS), etc. These popular queue disciplines are found to be *2-moment incomplete* and achieve only some part of achievable region. We identify a parametrized priority scheme, 2-level priority that is 2-momemt incomplete. However, some parametrized scheduling classes do exist that are 2-moment complete; see Section 2.

Some pre-emptive queue disciplines are considered and it is shown that these queue disciplines (processor sharing and pre-emptive last in first out) achieve the second moment *outside* the 2-moment completeness range; this is one of the motivation for focusing on non pre-emptive priority scheduling queue discipline. We also consider an anticipative and pre-emptive discipline and the resulting second moment is beyond what is achieved by work conserving, non pre-emptive and non anticipative parametrized class of policies.

Queueing models and its optimal control have significant role in

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).

978-3-9018-8273-9/15/ ©2015 IFIP

computer communication systems and communication networks (See [9], [10], [11] and references therein). See [12] and [13] for textbook treatment of such important topics. Some recent surveys with applications focused on wireless communication can be seen in [14] and [15]. This idea of second moment completeness is important for the class of optimal control problems in queueing system where variance (or second moment) plays significant roll. To further illustrate the implication of the notion of 2-moment completeness, we develop optimal control policy by exploiting the 2-moment completeness structure of a parametrized queue discipline for certain optimal control problems motivated from various regime.

A. Related literature

We briefly describe related ideas from *multi-class* queues. Average waiting time for each class forms a nice geometric structure (polytope) driven by conservation laws under certain scheduling assumptions for multi-class single server priority queue (see [16], [17]). This kind of structure also helps if one wants to optimize a suitable objective 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 networks of queues ([18], [19]). Unbounded achievable region for mean waiting time in two class deterministic polling system (non work conserving) is recently identified in [20]. Note that achievable region described in literature so for is with respect to mean waiting time.

A parametrized scheduling policy is called *complete* in [8] if it achieves all possible vectors of mean waiting time in multi-class queue. This question of completeness is important in following aspect. A complete scheduling class can be used to find the optimal control policy over all scheduling disciplines. This idea is 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 Groenevelt [18] came up with a synthesis algorithm using the completeness of mixed dynamic priority which is based on delay dependent priority proposed by Kleinrock [21].

Another community of researchers have exploited this notion of achievable region and completeness to find optimal control policy in multi-class queue (see [22], [23] and [24]). Optimal pricing and admission control problem for two classes is solved by exploiting the completeness structure of delay dependent priority queue (see [25] and [26]). Optimal control policy in two class polling system for certain optimization problems using achievable region approach is recently developed in [20].

This paper introduces the notion of achievable region for second moment of waiting time in *single* class queue for non pre-emptive, non anticipative and work conserving scheduling policies. The idea of 2-moment completeness for second moment of waiting time for $M/G/1$ and $M/M/1$ queues is discussed. This is done by identifying certain parametrized queue disciplines from literature and then showing them to be *2-moment complete*.

B. Paper organization

This paper is organised as follows: We introduce the notion of *2-moment completeness* in section II. Some parametrized policies are identified from literature and are shown to be *2-moment complete* in the same section. Various well known queue disciplines are discussed which are 2-moment incomplete in section III that puts in perspective the notion of 2-moment completeness. Few anticipative pre-emptive queue disciplines are discussed in section IV. Some optimal control problems are solved to illustrate the application of methodology in section V. This paper ends with a discussion on some future avenues in section VI.

II. 2-MOMENT COMPLETENESS

In single class queue, it is a well known fact that queue discipline does not affect the mean (first moment) waiting time. But the second moment, hence variance, depends heavily on queue discipline used. It has been proved that waiting time variance (or second moment) is minimum with FCFS (see [5]) and maximum with LCFS (see [6]) queue discipline under the assumption that busy period (time queue takes to empty) is finite almost surely or 'null state' of empty queue is recurrent, equivalently load factor, $\rho < 1$. Let l and u be the second moment of waiting time associated with FCFS and LCFS queue discipline respectively. The achievable region for second moment of waiting time is the interval $[l, u]$. Let $p \in I \subset \mathbb{R}$ and say class of these policies are denoted by $\{\mathcal{F}\}_{p\in I}$.

Definition 1: A set of parametrized queue discipline policies $\{\mathcal{F}\}_{p\in I}$ is called non pre-emptive, non anticipative, work conserving **2-moment complete** if these set of policies satisfy the following conditions.

1) Service is non pre-emptive.
2) Customers are selected for service in a manner that is independent of their subsequent service time.
3) If the service mechanism is ready to receive (serve) a customer at a time when the queue is non empty, then one of the customers present will be immediately served.
4) There exists a one-one mapping $V_{\mathcal{F}}(p) : I \to [l, u]$.

The first three conditions ensure the queue discipline to be non pre-emptive, non anticipative and work conserving. Condition 4 states that all possible second moments of waiting time are achieved by parametrized queue discipline policy. Note that this discussion with respect to second moment or variance of waiting time is equivalent as mean waiting time remains same for non pre-emptive, non anticipative and work conserving scheduling policies of $\{\mathcal{F}\}_{p\in I}$. We also use the term 2-moment complete in place of non pre-emptive, non anticipative work conserving 2-moment complete queueing discipline in further discussion.

Importance of such 2-moment complete parametrized policies lie in solving optimal control problems involving higher moments. Optimization over set of all non pre-emptive, non antic-

ipative work conserving scheduling policies can be performed by simply optimizing over a 2-moment complete policy. Some illustrative examples are discussed in Section V.

We discuss some policies in Section III that do not satisfy some of the above mentioned conditions for 2-moment completeness. Hence their second moment of waiting time lies outside the interval $[l, u]$. In fact, we observe that some popular policies have second moment more than u (achieved by LCFS) and less than l (achieved by FCFS). Note that such policies need extra information of the service time of the jobs waiting. We now identify some parametrized queue disciplines from literature which are 2-moment complete.

A. Impolite Customer class for M/G/1 queue

Consider the impolite arrival discipline introduced in [27] for a single class $M/G/1$ queue. This scheduling discipline is parametrized by p taking values in $[0, 1]$. An arriving customer joins the front of the queue with probability p and joins in the end of queue with probability $(1 - p)$ as shown in the Figure 1.

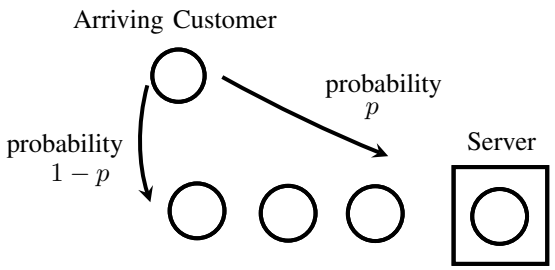

Fig. 1. Impolite arrival discipline

Theorem 1: Impolite customer class proposed in [27] is 2-moment complete.

Proof: Note that this service discipline is non pre-emptive, non anticipative and work conserving. So first three conditions of 2-moment completeness are satisfied. Second moment of waiting time is given by ([27]):

$$E(W^2)|_{\text{imp}} = \frac{1}{1 - p\rho}\left(\frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2 (E(S^2))^2}{2(1-\rho)^2}\right)$$

where $E(S^n)$ denotes the nth moment of service time and $E(W^2)|_{\text{imp}}$ is the second moment of waiting time for impolite class of customers. Let $I = [0, 1]$ and $\{\mathcal{F}\}_I$ be the impolite parametrized class of policies. It is clear from queue mechanism itself that $p = 0$ and $p = 1$ correspond to FCFS and LCFS service disciplines. Hence, $p = 0$ and $p = 1$ achieve the end points of the achievable region $[l, u]$ respectively. The function $E(W^2)|_{\text{imp}}$ is proportional to reciprocal of an affine function of p. Hence $E(W^2)|_{\text{imp}}$ will have one to one mapping from $I \to [l, u]$. This implies that the impolite scheduling discipline (parametrized policy) class is *2-moment complete*. ■

B. A parametrized queue discipline for M/M/1 queue

Consider the queue discipline parametrization for $M/M/1$ setting as proposed in [28]. This queue discipline works as follows:

- Newly arriving customer joins the queue at its end.
- Whenever the server becomes free it picks first or last customer with probability δ and $1 - \delta$ respectively and $0 \leq \delta \leq 1$.

Theorem 2: Queue discipline parametrization proposed in [28] is 2-moment complete.

Proof: It is clear from the above scheduling mechanism of [28] that first three conditions for 2-moment completeness are satisfied. Second moment of waiting time is given by ([28]):

$$E(W^2)|_\delta = \frac{2\lambda}{(\mu - \lambda)^2(\mu - \lambda + \delta\lambda)} \tag{1}$$

Taking $I = [0, 1]$ and the above the parametrized class of policies as $\{\mathcal{F}\}_I$, note that $\delta = 1$ recovers FCFS and $\delta = 0$ recovers LCFS. Hence, $\delta = 1$ and $\delta = 0$ achieves the end points of achievable region $[l, u]$ respectively. Again, $E(W^2)|_\delta$ is proportional to reciprocal of an affine function of δ. Hence, $E(W^2)|_\delta$ will have one to one mapping from $I \to [l, u]$. This means that this parametrization is *2-moment complete*. ■

III. Some 2-moment incomplete classes

We now observe below that some queue disciplines that have been analysed for their waiting time distributions turn out to be 2-moment *in*complete. In a sense the analysis below brings out the importance of queueing disciplines of Section II as they are not only *2-moment complete*, but, are easy to describe and implement. First, we show that Random Order of Service and Random Insertion policies achieve, for any load, a single point $(1/2)$ of the $[0, 1]$ domain of δ. Then, we show that Random Assigned Priority policy achieves $(0, \frac{1}{2})$ of the $[0, 1]$ domain of δ. Finally, we introduce a simple policy that randomly segregates the arrivals into two artificial 'classes' and this policy also turns out to be 2-moment *incomplete*.

A. Random Order of Service

Random order of service (ROS) queue discipline works as follows. Whenever it is time for customer to enter service and there are already $n \geq 1$ customers in the queue, each customer will have equal probability $\frac{1}{n}$ of getting selected for service. Delay distribution for $M/G/1$ setting with ROS queue discipline was first calculated by [29]. A list of second moments for various queue disciplines is shown in [30]. Second moment of waiting time in queue with $M/M/1$ setting for random order of service discipline and 2-moment complete

parametrized queue discipline is given by (see [30] and [28]):

$$E(W^2)_{ROS} = \frac{1}{(\mu-\lambda)^2}\left(\frac{4\rho}{2-\rho}\right) \quad (2)$$

$$E(W^2)|_{\delta} = \frac{1}{(\mu-\lambda)^2}\left(\frac{2\lambda}{\mu-\lambda+\delta\lambda}\right) \quad (3)$$

where ρ is the load factor. $E(W^2)_{ROS}$ and $E(W^2)|_{\delta}$ represent the second moment of waiting time with random order of service and parametrized queue discipline respectively. On equating above two equations, we get $\delta|_{ROS} = 1/2$. Hence ROS queue discipline achieves a single point in interval $[l, u]$ corresponding to $\delta = 1/2$.

B. Random Insertion

Random insertion (RI) queue discipline was introduced in [31]. This works in the following manner. Customers in queue are ordered from right to left, i.e., right most customer will have the position 1 and so on. If there are n customers waiting in queue, a newly arrived customer will be inserted in any of the $(n+1)$ positions with probability $1/(n+1)$. At a service beginning epoch, customer in position 1 goes in service. It has been proved in [31] that RI has same waiting time distribution as ROS. Hence RI queue discipline will also achieve a single point in interval $[l, u]$ corresponding to $\delta|_{RI} = 1/2$.

C. Random Assigned Priority

Random assigned priority (RAP) queue discipline, also introduced in [31], works as follows. As each customer arrives at queue, customer is independently assigned a random value that is uniformly distributed over interval $[0, 1]$. Customers in queue are then served according to non pre-emptive priority based on their assigned values. Smaller values have priority over larger values. Second moment of waiting time for this discipline is given by [31]: $E(W^2)_{RAP} =$

$$\frac{\rho(1-\rho)(2-\rho)E(S)E(S^3) + \rho^2(3-\rho)[E(S^2)]^2}{6(1-\rho)^3[E(S)]^2} \quad (4)$$

where $E(S^n)$ is the nth moment of service time. Note that for standard $M/M/1$ case that we are considering, $E(S^n) = (n)!/\mu^n$. Using these values of $E(S^n)$ and solving for δ using Equation (3), we get:

$$\delta|_{RAP} = \frac{(\mu-\lambda)(3\mu-\lambda)}{3\mu(2\mu-\lambda)+\lambda^2} = \frac{(1-\rho)(3-\rho)}{3(2-\rho)+\rho^2} \quad (5)$$

It can be easily verified from the stability of queue ($\rho < 1$) that $0 < \delta|_{RAP} < 1/2$. Hence, second moment of RAP is greater than that of ROS. So, we have following result.

Theorem 3: $\delta|_{RAP} < 1/2$ and $E(W^2)_{RAP} > E(W^2)_{ROS}$.

Remark 1: Also note that in heavy traffic, i.e., $\rho \to 1 \Rightarrow \delta \to 0$ hence queue discipline behaves as LCFS and in low traffic, i.e., $\rho \to 0 \Rightarrow \delta \to 1/2$ hence queue discipline behaves as ROS from variance of waiting time perspective.

Remark 2: If one uses RAP class of policies for optimizing over set of all non pre-emptive, non anticipative work conserving scheduling policy (for example problem P1 discussed in Section V), we will get suboptimal solution.

These are illustrated in Figure 2.

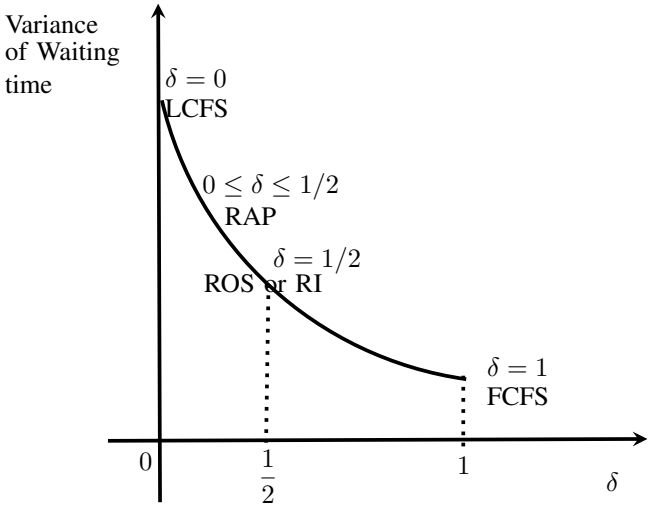

Fig. 2. Illustration of variance (or second moment) of waiting time for different queue disciplines vs parameter δ

D. Two Level Priority

This queue discipline is defined as follows: Arriving customers are divided in higher and lower priority class with probability p and $(1-p)$ respectively. Higher priority class (class 1) will have strict static priority. Queue discipline is FCFS within a class. Note that this is not a multi-class queue. Class discrimination is just a way of scheduling customers. Second moment of waiting time for each class is calculated in [32], [33]. Second moment and mean of waiting time of an arbitrary customer in such system can be obtained by conditioning on class as follows.

$$E(W) = E(W|\text{Class 1})P(\text{Class 1}) + E(W|\text{Class 2})P(\text{Class 2})$$

On calculating the mean waiting time for exponential arrivals and service by using above expression, we get mean waiting time same as in $M/M/1$ queue. Hence variance will differ in terms of second moments only. We calculate the second moment of system described above by conditioning:

$$E(W^2)|_p = \frac{2\lambda p}{\mu(\mu-\lambda p)^2} + \frac{2\lambda(\mu^2-\lambda^2 p)(1-p)}{(\mu-\lambda)^2(\mu-\lambda p)^3} \quad (6)$$

where $E(W^2)|_p$ is the second moment of waiting time for system with two level priority. Equating this expression with that of (3), we have that, $p = 1 \Rightarrow \delta = 1$ and $p = 0 \Rightarrow \delta = 1$. This matches with intuition as $\delta = 1$ corresponds to FCFS queue discipline and $p = 1$ or $p = 0$ implies that entire traffic goes in only one queue. Hence it will again give FCFS queue discipline. It is clear from scheduling mechanism of two level priority that this policy can never achieve LCFS queue discipline irrespective of value of p and hence, it is 2-moment incomplete. On simplifying for δ by equating second

moment from Equation (6) with parametrized second moment from Equation (3), we get $\delta|_{2lp}$ as

$$\frac{(1-p\rho)^3-(1-\rho)(p(1-\rho)^2(1-\rho p)+(1-p)(1-\rho^2 p))}{\rho(1-\rho)^2(1-p\rho)+\rho(1-p)(1-\rho^2 p)} \tag{7}$$

We argue that above equation does not achieve $\delta|_{2lp}=0$ for feasible range of p and ρ. Note that denominator is always positive for $0<p<1$ and $0<\rho<1$.

Parameter $\delta_{2lp}=0$ iff numerator is zero. Numerator simplifies to following cubic in p.

$$g(p)=\rho^2p^3+(\rho^3-4\rho^2+\rho-1)p^2+(1+2\rho)p-1$$

$g(0)=-1<0$ and $g(1)=(\rho-1)^3<0$ for feasible range of p and ρ. Also note that second derivative $g''(p)=2(3p\rho^2+(\rho-1)^2(\rho+1))>0$. This implies $g(p)$ is convex in p over $(0,1)$ for any given ρ. Hence numerator $g(p)<0$ for $0<p<1$. Thus, $\delta_{2lp}\neq 0$ for any given ρ and hence the range of δ_{2lp} is a strict subset of $[0,1]$. Combining this with Theorem 2, we have:

Theorem 4: The two level priority scheme is 2-moment incomplete.

Note that Equation (7) is a highly non linear function of p and it is difficult to find the exact range of $\delta|_{2lp}$ for $p\in[0,1]$. However, for $p=1/2$, we get

$$\delta|_{2lp}=\frac{4\left(1-\frac{\rho}{2}\right)^3-(1-\rho)^4-(1-\rho)(3-2\rho)}{\rho((1-\rho)^3+(3-2\rho))} \tag{8}$$

Remark 3: $\rho\to 1\Rightarrow\delta|_{2lp}\to 1/2$. Hence this simple two level priority system in heavy traffic can be good approximation for complex ROS or RI system for second moment or variance of waiting time.

IV. Some pre-emptive anticipative work conserving queue disciplines

In this section, we shift our attention to a couple of queue disciplines where variance is beyond the range of 2-moment complete policies, due to either anticipative, pre-emptive or non work conserving nature of queue discipline. A certain range of load factor can be found for processor sharing scheduling discipline where variance is within 2-moment complete range, while variance is beyond 2-moment complete range for the other range of load factor. Variance of longest remaining processing time (LRPT) and pre-emptive last in first out (PLIFO) scheduling policy are found beyond 2-moment complete range for any load factor. This means that the conditions in definition of 2-moment completeness are indeed necessary, if scheduler only uses the information of number in the system.

A. Processor Sharing

Processor sharing (PS) queue discipline is often used in computer systems for scheduling of processors. Note that this is a pre-emptive service queue discipline. Waiting time distribution with $M/M/1$ setting was first derived in [34] and that with $M/G/1$ setting in [35]. Conditional (conditioned on service time τ) mean waiting time (total time), $E[T|\tau]$, and conditional variance, $Var[T|\tau]$, in $M/M/1/PS$ queue is given by [36]:

$$E[T|\tau] = \frac{\tau}{(1-\rho)} \tag{9}$$

$$Var[T|\tau] = \frac{2\rho\tau}{\mu(1-\rho)^3}-\frac{2\rho}{\mu^2(1-\rho)^4}\left[1-e^{-\mu\tau(1-\rho),}\right] \tag{10}$$

Now, we use the following expressions to derive unconditional variance of waiting time:

$$Var(T)=E(Var(T|\tau))+Var(E(T|\tau)) \tag{11}$$

$$E(Var(T|\tau))=\int_\tau Var(T|\tau)f(\tau)d\tau$$

where $f(\tau)$ is service time density. For exponential service density, we have

$$\begin{aligned} E(Var(T|\tau)) &= \int_0^\infty Var(T|\tau)\mu e^{-\mu\tau}d\tau \\ &= \frac{2\rho}{\mu^2(1-\rho)^2(2-\rho)} \\ Var(E(T|\tau)) &= \frac{1}{\mu^2(1-\rho)^2} \end{aligned}$$

On simplifying, we get the following unconditional variance for mean waiting time when processor sharing is used as scheduling policy.

$$Var(T)_{PS} = \frac{1}{\mu^2(1-\rho)^2}\frac{2+\rho}{2-\rho} \tag{12}$$

$$Var(T)_\delta = \frac{1}{(\mu-\lambda)^2}\left[\frac{2\lambda}{\mu-\lambda+\delta\lambda}-\rho^2\right]+\frac{1}{\mu^2} \tag{13}$$

where $Var(T)_\delta$ denotes variance for parametrized queue discipline discussed in Section II-B. On equating the above variances and solving for δ, we have

$$\delta=1-\frac{\mu^2}{\lambda(3\mu-\lambda)}=1-\frac{1}{\rho(3-\rho)} \tag{14}$$

Note that $\delta<1$ is trivially true. In heavy traffic, $\rho\to 1\Rightarrow\delta\to 1/2$. Hence processor sharing behaves like ROS in high traffic from variance view point.

Note that on simplifying $\delta\geq 0$, we get the quadratic $\rho^2-3\rho+1\leq 0$ or we have

$$\left(\rho-\frac{3+\sqrt{5}}{2}\right)\left(\rho-\frac{3-\sqrt{5}}{2}\right)\leq 0. \tag{15}$$

This implies $\delta\geq 0$ for $\rho\in[\frac{3-\sqrt{5}}{2},1]$ (see Figure 3). As $\rho\downarrow\frac{3-\sqrt{5}}{2}\Rightarrow\delta\downarrow 0$. Hence processor sharing queue behaves as LCFS for $\rho\approx\frac{3-\sqrt{5}}{2}$.

For $\rho\in(0,\frac{3-\sqrt{5}}{2})$, there is no δ in range $[0,1]$. Hence variance of PS is beyond 2-moment complete parametrized queue discipline range. In fact, on simplifying the expression,

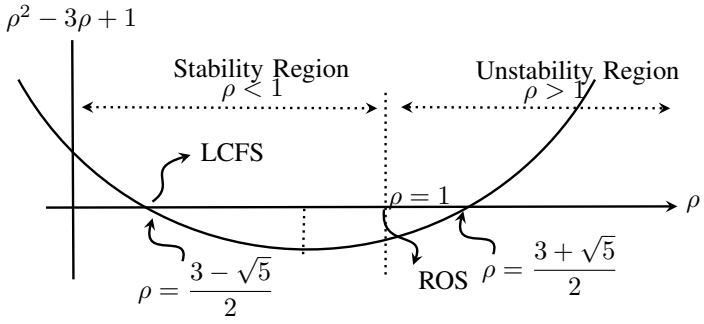

Fig. 3. Change in sign of quadratic $\rho^2 - 3\rho + 1$ w.r.t. ρ

$Var(W)_{PS} > Var(W)_{LCFS}$ holds iff $\rho^2 - 3\rho + 1 > 0$ which is true for $\rho \in (0, \frac{3-\sqrt{5}}{2})$. So there is a range of load factor for which variance in processor sharing happens to be beyond 2-moment complete range, i.e., more than that of LCFS.

B. Pre-emptive Last In First Out (PLIFO)

Under this service discipline, server is always working on most recent arrival to the system. Job at the server is pre-empted on an arrival and may only resume service once the system is empty of newer arrivals. Thus, PLCFS acts as a stack where new jobs are placed on the top of the stack and the server is always working on the job at the top. Variance of waiting time for PLCFS is given by [7]:

$$Var(T)|_{PLCFS} = \frac{Var(S)}{(1-\rho)^3} + \lambda \left(\frac{E(S)}{1-\rho}\right)^3. \quad (16)$$

Variance of total time in non preemptive LCFS queue discipline is known. On simplifying the expressions for exponential service, we have

$$Var(T)_{PLCFS} - Var(T)_{LCFS} = \frac{2\mu\lambda}{\mu^2(\mu-\lambda)^2} > 0 \quad (17)$$

Hence, variance in pre-emptive LCFS is more than that of LCFS irrespective of value of load factor.

C. Longest Remaining Processing Time (LRPT)

In this queue discipline, the job in the system with the longest remaining size is given pre-emptive priority. Hence no job can finish service before the end of a busy period. LRPT finishes every job at the last moment possible under any work conserving policy. Note that this is an anticipative queue discipline. Conditional (conditioned on service time) mean and variance in LRPT are given by [7]:

$$E(T|\tau)_{LRPT} = \frac{\tau}{1-\rho} + \frac{\lambda E[S^2]}{2(1-\rho)^2}$$

$$Var(T|\tau)_{LRPT} = \frac{\lambda\tau E[S^2]}{(1-\rho)^3} + \frac{\lambda E[S^3]}{(1-\rho)^3} + \frac{3}{4}\left(\frac{\lambda E[S^2]}{(1-\rho)^2}\right)^2$$

On unconditioning the above mean and variance for exponential service time similar to processor sharing case (see Equation (11)), we have

$$E(T)_{LRPT} = \frac{\mu}{(\mu-\lambda)^2} > E(T)_{FCFS} \quad (18)$$

$$Var(T)_{LRPT} = \frac{\mu^2 - 4\lambda^2 + 6\mu\lambda}{(\mu-\lambda)^4} \quad (19)$$

On calculating the difference, we have

$$Var(T)_{LRPT} - Var(T)_{LCFS} = \frac{2\lambda^3 + \mu^2\lambda + 7\mu\lambda(\mu-\lambda)}{\mu(\mu-\lambda)^4} > 0 \quad (20)$$

Variance of this anticipative and pre-emptive queue discipline is more than that of LCFS queue discipline for any load factor.

We summarize the above discussion as:

Theorem 5: Variance of waiting time with Pre-emptive Last in First Out (PLIFO) or Longest remaining processing time (LRPT) is more than that with LCFS scheduling policy for any load factor while variance with processor sharing is beyond 2-moment complete range for load factor, $\rho \in (0, \frac{3-\sqrt{5}}{2})$.

Remark 4: Variance of waiting time can be beyond 2-moment complete range if scheduling policy violates any of the condition on queue discipline being non pre-emptive, non anticipative and work conserving as described in definition 1 of Section II.

V. Some Applications

In this section, we illustrate the implication of the idea of *2-moment completeness* with the help of some optimal control problems.

A. Illustrative example 1

First, we consider the problem, **P1**, of minimizing variance of waiting time subject to constraint on lower bound on it over set of all non pre-emptive, non anticipative and work conserving scheduling policies in M/M/1 queue. Mathematically, we have

$$\textbf{P1:} \quad \min_{\mathcal{F}} \quad Var(W)$$

Subject to

$$Var(W) \geq \gamma$$

for a given γ, where $\mathcal{F}$ is the set of all non pre-emptive, non anticipative and work conserving scheduling policies for this $M/M/1$ queue. Since, parametrized queue discipline discussed in section II-B is shown to be *2-moment complete*, the optimization problem P1 is equivalent to transformed problem T1

$$\textbf{T1:} \quad \min_{0\leq\delta\leq1} \quad Var(W)$$

Subject to

$$Var(W) \geq \gamma$$

Clearly, if $\gamma > Var(W)|_{LCFS}$, the problem is infeasible. When $\gamma < Var(W)|_{FCFS}$, trivial solution for the above problem will be FCFS scheduling policy ($\delta = 1$). For the range

$Var(W)|_{LCFS} \geq \gamma \geq Var(W)|_{FCFS}$, optimal scheduling policy can be easily obtained by exploiting monotonic nature of variance function for *2-moment complete* range and hence by solving $Var(W) = \gamma$. Optimal scheduling policy is pure dynamic for $\gamma \in (Var(W)|_{FCFS}, Var(W)|_{LCFS})$.

B. Illustrative example 2

Consider another unconstrained optimal control problem, **P2**, of minimizing the total cost where cost is associated with variance and unfairness in standard M/M/1 queue. Mathematically, we have

$$\textbf{P2:} \quad \min_{\mathcal{F}} \quad c_1 Var(W) + c_2 f(W)$$

where c_1 and c_2 are the costs associated with variance and unfairness respectively. $f(W)$ represents unfairness of a job and unfairness is quantified according to [37]. Following ordering in scheduling policies is identified under this fairness index: FCFS $>$ ROS $>$ LCFS. FCFS scheduling policy also has minimum variance in $\mathcal{F}$ (from 2-moment completeness). Hence the optimal scheduling policy for problem **P2** will be FCFS.

C. Illustrative example 3

Now, we illustrate another application of 2-moment completeness in higher moment optimal control problems. Study of higher moments is quiet popular in queueing systems as well as other application areas (see [38], [39], [40]).

This optimal control problem below, **P3**, is motivated from Markovitz mean-variance (MV) model (see [41]) where variance of a portfolio of assets is minimized subject to constraint on first moment of returns of the portfolio. We consider an extension of this model and minimize the third moment of waiting time subject to constraint on second moment over all non pre-emptive, non anticipative and work conserving scheduling policies which span entire feasible space. Some similar problems of minimizing skewness (related to 3rd moment) under mean and variance constraints are solved in finance literature (See [39]).

$$\textbf{P3:} \quad \min_{\mathcal{F}'} \quad E(W^3)$$

Subject to

$$E(W^2) \leq \beta$$

for a given β, where $\mathcal{F}'$ is the set of all non pre-emptive, non anticipative and work conserving scheduling policies for this $M/M/1$ queue which span the feasible space. Since, parametrized queue discipline discussed in section II-B is shown to be *2-moment complete*, the optimization problem P3 is equivalent to transformed problem T3

$$\textbf{T3:} \quad \min_{0 \leq \delta \leq 1} \quad E(W^3)$$

Subject to

$$E(W^2) \leq \beta$$

By the above result, the parametrized queue discipline discussed in section II-B is 2-moment complete class and hence spans the entire feasible space of this optimization problem.

Note that we are not aware if there is a parametrized class of schedulers that are 3rd moment complete or even if there is a such a notion. A 3rd moment complete class (if exists) may contain 2nd moment complete class also. Hence it might be possible that optimizing over another such 2-moment complete class may give a minima different from optimal objective of T3. We assume that the 2-moment complete class identified in this paper gives the best solution for problem P3 and solve the problem as below.

We obtain third moment of waiting time for the above optimization problem using characteristic function for parametrized queue discipline. Expression for characteristic function of waiting time distribution is given by (see [28]):

$$\phi_W(s) = 1 + \frac{is\rho\alpha}{\lambda(1-\alpha)} \text{ where } \alpha = \frac{\mu + \lambda - \delta\lambda - is - \{(\mu + \lambda - \delta\lambda - is)^2 - 4\lambda\mu(1-\delta)\}^{1/2}}{2\mu(1-\delta)}.$$

Since, the third moment of waiting time is

$$E(W^3) = (-i)^3 \phi_W^{(3)}(0)$$

we get the following expression after some simplifications:

$$E(W^3) = \frac{6\lambda(\mu(\mu-\lambda) + \lambda(\mu - \lambda(1-\delta)))}{(\mu-\lambda)^3(\mu-\lambda(1-\delta))^3}$$

Hence optimization problem T3 can be rewritten as

$$\textbf{T3:} \quad \min_{0 \leq \delta \leq 1} \quad \frac{6\lambda(\mu(\mu-\lambda) + \lambda(\mu - \lambda(1-\delta)))}{(\mu-\lambda)^3(\mu-\lambda(1-\delta))^3} := f(\delta)$$

Subject to

$$E(W^2) = \frac{2\lambda}{(\mu-\lambda)^2(\mu-\lambda+\delta\lambda)} \leq \beta \tag{21}$$

The above constraint can be rewritten as

$$\delta \geq \frac{2}{(\mu-\lambda)^2\beta} - \frac{\mu-\lambda}{\lambda} := \delta'$$

Note that if $\beta < E(W^2)|_{\text{FCFS}}$ no feasible solution will exist to problem P3 and if $\beta > E(W^2)|_{\text{LCFS}}$, constraint (21) will become redundant. Derivative of objective $f(\delta)$ simplifies to

$$f'(\delta) = \frac{6\lambda^2}{(\mu-\lambda)^3(\mu-\lambda+\delta\lambda)^6}[2\lambda^2 + \mu(\lambda - 3\mu) - 2\delta\lambda^2]$$

It can be noted that $f'(\delta) < 0$ for $\delta \in (0,1)$ and hence $f(\delta)$ will be decreasing for $\delta \in (0,1)$.

Based on above, we have the following lemma describing the solution of optimization problem P3.

Lemma 1: Under the above assumption, optimal scheduling policy for problem P3 is given by FCFS queue discipline as long as P3 is feasible.

Remark 5: If second moment of waiting time is constrained in $[\beta_1, \beta_2]$ for some suitable and given $\beta_1 > 0$ and β_2, optimal

queue discipline will be pure dynamic with $\delta^* \in (0, 1)$.

VI. Discussion

The idea of 2-moment completeness introduced in this paper is useful in solving the optimal control problems which involve second moment (or variance) of waiting time for single class queues and optimization needs to be done in the class of non pre-emptive, non anticipative and work conserving scheduling disciplines. A parametrized policy (2-level priority) is found to be 2-moment incomplete. Thus, optimization over such a two moment incomplete policy will give a sub-optimal solution. This brings out the importance of identifying 2-moment complete policies as in Theorem 1 and Theorem 2. It will be interesting to explore the applicability of this idea in solving various useful optimal control problems motivated from wireless communication and computer networks. Few such problems are discussed in this paper. Extending this idea of 2-moment completeness to multi class queues and queueing networks will be another fascinating future avenue.

References

[1] J. D. Little, "A proof of queueing formula: L=λ W," *Operations Research*, vol. 9, pp. 383–387, 1961.

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

[3] W. Whitt, "A review of L= λW and extensions," *Queueing Systems*, vol. 9, no. 3, pp. 235–268, 1991.

[4] ——, "Correction note on L= λW," *Queueing Systems*, pp. 431–432, 1992.

[5] J. F. C. Kingman, "The effect of queue discipline on waiting time variance," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, pp. 163–164, 1962.

[6] P. Eschenfeldt, B. Gross, and N. Pippenger, "A bound on the variance of the waiting time in a queueing system," June 2011, arXiv:1106.0074v1.

[7] A. Wierman, "Scheduling for today's computer systems: Bridging theory and practice," Ph.D. dissertation, Carnegie Mellon University, 2007.

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

[9] H. Kobayashi and A. G. Konheim, "Queueing models for computer communications system analysis," *Communications, IEEE Transactions on*, vol. 25, no. 1, pp. 2–29, 1977.

[10] S. Stidham Jr, "Optimal control of admission to a queueing system," *Automatic Control, IEEE Transactions on*, vol. 30, no. 8, pp. 705–713, 1985.

[11] C.-H. Ng and S. Boon-Hee, *Queueing modelling fundamentals: With applications in communication networks*. John Wiley & Sons, 2008.

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

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

[14] 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.

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

[16] 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.

[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] A. Federgruen and H. Groenevelt, "M/G/c queueing systems with multiple customer classes: Characterization and control of achievable performance under nonpreemptive priority rules," *Management Science*, vol. 9, pp. 1121– 1138, 1988.

[19] 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.

[20] 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.

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

[22] D. Bertsimas, "The achievable region method in the optimal control of queueing systems; formulations, bounds and policies," *Queueing systems*, vol. 21, no. 3-4, pp. 337–389, 1995.

[23] 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.

[24] 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.

[25] 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.

[26] ——, "A model for service level based pricing of shared resources at container depots," in *Proceedings of the international conference on transportation system studies*, 2008.

[27] S. Ozekici, J. Li, and F. S. Chou, "Waiting time in M/G/1 queues with impolite arrival disciplines," *Probability in the Engineering and Informational Sciences*, vol. 9, pp. 255–267, 1995.

[28] V. Dufkova and F. Zitek, "On a class of queue disciplines," *Aplikace Mathematiky*, vol. 20, pp. 345–357, 1974.

[29] L. Takacs, "Delay distributions for one line with Poisson input, general holding times, and various orders of service," *Bell System Technical Journal*, vol. 42, pp. 487–503, 1963.

[30] M. Scholl and L. Kleinrock, "On the M/G/1 queue with rest periods and certain service-independent queueing disciplines," *Operations Research*, vol. 31, pp. 705–719, 1983.

[31] S. W. Fuhrmann and I. Iliadis, "A comparison of three random discipline," *Queueing Systems*, vol. 18, pp. 249–271, 1994.

[32] H. Kesten, J. T. Runnenburg, and D. van Dantzig, *Priority in waiting line problems*. Koninklijke Nederlandse Akademie van Wetenschappen, 1957.

[33] L. Durr, "A single-server priority queuing system with general holding times, Poisson input, and reverse-order-of-arrival queuing discipline," *Operations Research*, vol. 17 (2), pp. 351–358, 1969.

[34] E. G. Coffman, R. R. Muntz, and H. Trotter, "Waiting time distributions for processor-sharing systems," *Journal of the ACM*, vol. 17, pp. 123–130, 1970.

[35] T. J. Ott, "The sojourn-time distribution in the M/G/1 queue with processor sharing," *Journal of Applied Probability*, vol. 21, pp. 360–378, 1984.

[36] S. F. Yashkov, "Processor sharing queues: Some progress in analysis," *Queueing Systems*, vol. 2, pp. 1–17, 1987.

[37] D. Raz, H. Levy, and B. Avi-Itzhak, "RAQFM: A resource allocation queueing fairness measure," in *Proceedings of ACM SIGMETRICS Conference, New York, NY*, 2004.

[38] N. Gülpınar, U. Harder, P. Harrison, T. Field, B. Rustem, and L.-F. Pau, "Mean-variance performance optimization of response time in a tandem router network with batch arrivals," *Cluster Computing*, vol. 10, no. 2, pp. 203–216, 2007.

[39] M. Mhiri and J.-L. Prigent, "International portfolio optimization with higher moments," *International Journal of Economics and Finance*, vol. 2, no. 5, pp. 157–169, 2010.

[40] S. X. Liao, "Image analysis by moments," Ph.D. dissertation, University of Manitoba, 1993.

[41] H. M. Markowitz, *Portfolio selection: efficient diversification of investments*. Yale university press, 1968, vol. 16.