Optimal Surplus Capacity Utilization in Polling Systems via Fluid Models

Ayush Rawal, Veeraruna Kavitha and Manu K. Gupta
Industrial Engineering and Operations Research, IIT Bombay, Powai, Mumbai - 400076, India
E-mail: ayush.rawal, vkavitha, manu.gupta @iitb.ac.in

***Abstract*—We discuss the idea of differential fairness in polling systems. One such example scenario is: primary customers demand certain Quality of Service (QoS) and the idea is to utilize the surplus server capacity to serve a secondary class of customers. We use achievable region approach for this. Towards this, we consider a two queue polling system and study its 'approximate achievable region' using a new class of delay priority kind of schedulers. We obtain this approximate region, via a limit polling system with fluid queues. The approximation is accurate in the limit when the arrival rates and the service rates converge towards infinity while maintaining the load factor and the ratio of arrival rates fixed. We show that the set of proposed schedulers and the exhaustive schedulers form a complete class: every point in the region is achieved by one of those schedulers. It is well known that exhaustive service policy optimizes system performances like unfinished work. In this paper, we show that it is also optimal from the perspective of individual queues.**

We further pose two constrained optimization problems: a) admission control, wherein the arrival rates of secondary customers is optimally designed; b) maximizing the revenue considering the losses when secondary arrival rate is fixed. We finally show that exhaustive service discipline at each queue turns out to be optimal.

***Index Terms*—Polling systems, achievable region, dynamic scheduling, fluid queues**

I. Introduction

Polling systems are special class of queueing systems where a single server visits a set of queues in some order and takes non-zero time to walk/switch between queues. This special class has acquired significant importance in queueing theory due to its wide range of applications in communication systems, production systems, traffic and transportation systems [1]. Analysis of polling systems started in early 1970's when cyclic polling systems was first used to model time sharing computer systems [2]. A rich literature in this field has evolved since then. Pseudo-conservation laws [3] are derived for various scheduling policies. Some dominance relations in polling systems can be seen in [4]. Extensive research is done in both discrete and continuous polling models (see [2], [5], [3], [1] etc., and references therein).

Many problems in wireless communications can be studied using polling systems. For example, consider a cognitive radio type scenario in which a network is providing service to primary customers. We consider a slightly modified approach, wherein, the primary customers are satisfied as long as their demands are fulfilled within a guaranteed average waiting time. The network now can utilize the surplus capacity (the spectrum, time slots etc.,) to serve a secondary set of users, while maintaining the QoS requirements of primary customers. One can model this scenario with polling systems, when it takes non-zero time to switch the services between the two classes of customers. Another example scenario is that of data and voice users utilizing the same wireless network. In this case, the network needs to maintain the drop probability of the impatient voice customers (who drop calls if not picked-up within a negligible waiting times) below an acceptable level. Alongside, it also needs to optimize the expected sojourn times of the data calls.

Polling systems can be categorized on the basis of different queue elements [11]. There are different types of switching disciplines possible: exhaustive, 1-exhaustive, gated, globally gated, absolute priority etc. Once server reaches queue there can be various queue disciplines: FCFS ((First Come First Serve), LCFS (Last Come First Serve), SJF (Shortest Job First), random order, processor sharing etc. Switching times can be random or deterministic and switching order can be cyclic or random or table based and so on.

Notion of fairness is well studied in literature (see [13], [14], [16], [17] etc.,). Proportional fairness idea is introduced by Kelly et. al., in [15] and a recent survey for various aspects of fairness (for work conserving queuing systems as well as for polling systems) can be seen in [13]. Here they mainly discuss the fairness based on the job-size of the customer. Raz et al., in [20], propose 'Resource Allocation Queuing Fairness Measure' to measure fairness in allocation of system resources. They argued that LCFS is the least 'fair' service discipline while processor sharing the most fair service discipline with respect to the measure proposed by them. Fairness by giving priority to low load customers was discussed in [17], in the context of internet users.

In [22], fairness is achieved for under privileged users (e.g., users far away from the 'wireless' service provider) via a constrained optimization (the users were guaranteed some minimum QoS) and is achieved using priority schedulers. The same idea is applicable even if a certain class of customers require a QoS larger than what they would have achieved in a social optimal solution. We refer this idea as *differential fairness*, wherein, in a system with N customer classes, n classes of customers demand/require a certain level of Quality of Service (QoS) while the performance of the remaining $N-n$ classes of customers need to be optimized. Differential fairness concept could be applied in various contexts: a) application driven, for e.g., voice calls need to be picked within negligible waiting periods while data calls can tolerate delays; b) price driven, for e.g., certain class of customers can pay higher money to get better QoS; c) market driven, for e.g., if the plant has to manufacture different varieties of items priority is given to the items whose demand is more etc. It is *not a completely new concept, and is proposed in several other contexts (e.g., wireless communications) but here we are trying to generalize this concept to polling systems.*

Achievable region approach is one of the popular techniques to solve optimization problems (see for e.g., [6], [7], [8], [10], [21] etc.,). A vast community of researchers have focused in characterizing the achievable region of mean waiting time for a *work conserving* queue. Coffman and Mitrani [6] were the first to identify such regions when they identified it for multiclass M/M/1 queue with pre-emptive priority discipline. Further structure of such achievable region is studied by many authors [7] [8]. It was identified for *work conserving* multi class single server priority queue that achievable region is bounded and forms a polytope (e.g., [6], [21]). Note that polling systems are non work conserving queues as server may not serve any customer while walking. Bertsimas et.al., [9] obtained the bounds on performance of an optimal policy and developed optimal or near-optimal policies for non-work conserving polling systems, multi-class queuing systems etc. Achievable region approach has also been used for optimal control of stochastic systems [10].

A. Main Contribution

We attempt to solve an example differential fairness problem (utilizing surplus capacity), via achievable region approach. A new class of parametrized schedulers are proposed, which determine the switching decision between queues, and with the motivation of finding the achievable region. These schedulers are inspired by delay priority schedulers proposed by Kleinrock [18], (see also [19]) for muti-class work conserving queuing systems. We extend these schedulers to non-work conserving, polling systems and parametrize them using $\beta = (\beta_1, \beta_2)$: β_i is the parameter used when switching decision has to be made while serving at queue i.

We consider well known fluid queues to study this problem (see for e.g., [10], [12] etc.,). We refer these as steady state fluid model (SSFM), as these models facilitate steady state analysis and are useful when the system behaves periodically as in the example case of polling systems. By SSFM we refer two queue polling system with continuous, deterministic arrival and departure flows. Departure flow at a queue is switched on only when the server is at that queue. We analyzed these fluid models, when switching policies are given by the proposed (β_1, β_2) schedulers. We obtained stability conditions and closed form expressions of stationary performance measures for the deterministic fluid model. We also *obtained the achievable region of waiting times which turned out to be unbounded.*

We conjecture that the performance analysis of a random system (one with discrete arrivals) converges to that of the SSFM, as the arrival and service rates converge to infinity while maintaining certain ratios constant. Some initial ideas are discussed in Appendix B. Thus we have the approximate achievable region for random systems with large arrival and departure rates and this fact is illustrated via numerical examples.

We consider two relevant differential fairness problems. In both the problems, exhaustive service discipline turns out to be the optimal policy, which both satisfies the QoS constraint at the primary queue and optimizes the performance at the other queue. *One of the important conclusions of this study is that for the systems with high arrival and departure rates the exhaustive policy turns out to be the optimal solution.* However this may not be the case for systems with moderate arrival rates, which are drastically different from fluid models. We already have some initial observations in this direction and this is the topic of future research.

B. Paper Organization

This paper is organized as follows. Sections II ,III describe model setting and (β_1, β_2) schedulers. Section V presents complete analysis of deterministic fluid queues. We also discuss stability conditions, stationary performance measure and achievable region of such system in this section. Some proofs are available in Appendix.

II. Problem Description: Differential fairness

Consider a queueing system offering service to a primary class of customers. The system is operating in stable regime, while maintaining certain demanded QoS defined in terms of average waiting time. Operating in stable conditions implies the system is under utilizing its capacity (the service rate has to be less than the arrival rate). The service provider wants to utilize the surplus capacity to derive additional income from secondary set of customers and this has to be done, while maintaining the QoS of the primary customers. This is a two class scenario in which differential fairness is to be implemented and we achieve this via the following optimization problem:

$$\text{Optimize } \mathcal{F}(\bar{w}_1, \bar{w}_2) \tag{1}$$

$$\text{Subject to: } \bar{w}_1 \leq \eta_1, \tag{2}$$

where $\bar{w}_1, \bar{w}_2$ are the mean waiting times of class 1 and class 2 customers respectively, $\mathcal{F}$ is an appropriate function of average waiting times and the allowed/admissible arrival rates of the the two classes and some other relevant factors. In this paper, we consider two examples of the above problem in Section VI. We assume that it takes non zero time to switch services from one class to the other and hence study this problem in the context of polling systems.

III. System model and our approach

We consider two queue, $(\mathcal{Q}_1, \mathcal{Q}_2)$, single server polling system. Let λ_i and μ_i be the arrival and service rates respectively of $\mathcal{Q}_i$; $i = 1, 2$. Let S denote the random time required by server to switch from one queue to another. The sequence of consecutive switching times $\{S_k\}_k$ are assumed to be an independent and identically distributed (IID) sequence with mean s. Both the queues have infinite buffer capacity. FCFS and Non-preemptive queue discipline is used within a class. Customers leave system only when their service is completed. Once server reaches a queue, at least one customer is served. We consider those class of switching/scheduling policies, which are employed at every departure epoch, which can possibly depend upon the state of the system and which are time invariant/stationary. That is, these policies can depend for example upon the number of waiting customers in each queue, or the waiting time of the longest waiting customers etc., but do not depend upon the number of times the server has visited a queue. The system utilization factor $\rho = \rho_1 + \rho_2$ with $\rho_i = \lambda_i/\mu_i$ for each i.

We use achievable region based approach to solve (1). Achievable region for two queue polling system is defined as:

$$\mathcal{A} = \{(\bar{w}_1, \bar{w}_2) : \ \beta \textit{ is any stationary scheduling policy}\}.$$

And now the problem (1) is equivalent to

$$\text{Optimize}_{(\bar{w}_1,\bar{w}_2)\in\mathcal{A}} \mathcal{F}(\bar{w}_1, \bar{w}_2) \text{ s.t. } \bar{w}_1 \leq \eta_1.$$

If a *class $\mathcal{B}$ of schedulers is complete,* i.e., if every pair $(\bar{w}_1, \bar{w}_2) \in \mathcal{A}$ is achieved by a scheduler policy $\beta \in \mathcal{B}$, then the problem (1) is equivalent to

$$\text{Optimize}_{\beta\in\mathcal{B}} \mathcal{F}(\bar{w}_1(\beta), \bar{w}_2(\beta)) \text{ s.t. } \bar{w}_1(\beta) \leq \eta_1.$$

Motivated by Kleinrock's ([18]) delay priority queues, we introduce a class of priority type schedulers parametrized by parameter $\beta := (\beta_1, \beta_2)$, which along with exhaustive policy will be shown to be complete (in coming sections). Here β_i is delay priority parameter associated with $\mathcal{Q}_i$ and index $-i$ represents the label of other queue (when $i = 1$, $-i = 2$ and when $i = 2$, $-i = 1$). To include exhaustive policies, we let β_i take value in $\mathbb{R}^e = \mathbb{R} \cup \{ex\}$. Let $\tilde{w}_i$ and $\tilde{w}_{-i}$ be the waiting time of longest waiting customer in $\mathcal{Q}_i$ and $\mathcal{Q}_{-i}$ respectively. Switching decision is taken at every departure epoch and it depends upon the queue in which the server is currently working. When in $\mathcal{Q}_i$, switching rule $\beta = (\beta_1, \beta_2)$ implies the following scheduling decisions:

1) When $\beta_i = ex$ switch from $\mathcal{Q}_i$, when $\mathcal{Q}_i$ is empty.
2) When $\beta_i \neq ex$ switch from $\mathcal{Q}_i$, when $\tilde{w}_i\beta_i \leq \tilde{w}_{-i}$.

Note that (ex, β_1) is the scheduler where exhaustive policy is implemented at $\mathcal{Q}_1$ while priority type policy is implemented in $\mathcal{Q}_2$. Now define:

$$\mathcal{B}^P \equiv \{\beta = (\beta_1, \beta_2) : \beta_1, \beta_2 \in \mathbb{R}^e\}.$$

We will show that $\mathcal{B}^P$ is a complete class of schedulers for an 'approximate' achievable region of the two class polling system. We also obtain approximate performance (average waiting times $(\bar{w}_1, \bar{w}_2)$) for each scheduler of $\mathcal{B}^P$. We argue that these approximations are achieved asymptotically as the arrival and service rates increase towards infinity. Considering two problems of the type (1), that of admission control and revenue optimization, we obtain optimal scheduling policies using $\mathcal{B}^P$.

IV. Convergence of Random System

In this paper, we consider scenarios with high traffic and service rates. Some examples of such scenarios are, road traffic coming from different directions, manufacturing units for different types of items, packet level arrivals in a communication systems from higher layers. We obtain the asymptotic analysis in the limit $\mu \to \infty$ with ratios $\rho := (\lambda_1 + \lambda_2)/\mu$ and $\nu := \lambda_1/\lambda_2$ fixed. We will also require the switching times S to converge towards their mean s as $\mu \to \infty$. Our conjecture is that the performance measures of the random system converges towards that of a limit system which is deterministic. In this section, we discuss the supporting arguments for this conjecture.

To begin with, we make an observation: the random system converges towards a limiting and deterministic fluid queue system. Let $\{A^{\lambda_i}_{n,i}\}_n$ represent the successive inter-arrival times of customers of queue $\mathcal{Q}_i$. Let $\Lambda^{\lambda_i}(t)$ represent the number of arrivals in time $[0, t]$ when the arrival rate is λ_i:

$$\Lambda^{\lambda_i}(t) = \sup_k \left\{ \sum_{n=1}^{k} A^{\lambda_i}_{n,i} \leq t \right\}.$$

We assume inter arrival times $\{A^{\lambda_i}_{n,i}\}$ are IID with mean equal to $1/\lambda_i$. Let $\Gamma^\mu(t)$ represent the number of uninterrupted services in time $[0, t]$, i.e., the number of services completed if service was offered without a break, when the service rate is μ:

$$\Gamma^\mu(t) = \sup_k \left\{ \sum_{n=1}^{k} \xi_n \leq t \right\},$$

where $\{\xi_k\}_k$ are IID service times with mean $1/\mu$. We have the following:

Theorem 1. *With $\stackrel{d}{=}$ representing the stochastic equivalence, assume that $\{\Lambda^\lambda(t); t \geq 0\} \stackrel{d}{=} \{\Lambda^1(\lambda t); t \geq 0\}$. Then we as $\lambda \to \infty$ we have:*

$$\frac{\Lambda^\lambda(t)}{\lambda} \to t \textit{ a.s. for all } t. \qquad \blacksquare$$

If S^λ is a random variable independent of other processes, depending upon the parameter λ and satisfies the following additional assumption:

$$S^\lambda \to s, \textit{ with } s \textit{ a constant, almost surely as } \lambda \to \infty.$$

Then we have:

$$\frac{\Lambda^\lambda(S)}{S\lambda} \to 1 \ a.s.$$

We can have a similar theorem for uninterrupted service process $\{\Gamma^\mu(t); t \geq 0\}$. And the assumptions of this theorem are satisfied by many processes. One can easily see that Poisson process is one such example. Further we have the following theorem (proofs in Appendix).

Theorem 2. *Consider any IID sequence $\{A^1_n\}_n$ with unit mean, i.e., $E[A^1_n] = 1$ for all n. Then the number of arrivals*

$\{\Lambda^\lambda(t); t \geq 0\}$ *with inter-arrival process* $\{A_n^\lambda\}_k$ *given by:*

$$A_n^\lambda = \frac{1}{\lambda} A_n^1 \text{ for each } n,$$

satisfies the assumptions of Theorem 1. ■

Thus from Theorem 1 for large μ (which implies large λ_1 and λ_2 determined by μ because of ratios ρ, ν), random system is close to a deterministic fluid queue system. This system has continuous flows of arrivals at each queue as according to $\Lambda^{\lambda_i}(t) = \lambda_i t$, $i = 1, 2$ for all t, and a similar continuous departure flow whenever service is offered. We call these fluid queues as steady state fluid model (SSFM).

Our idea is to study first the performance and then the achievable region of the random system via the corresponding ones of the SSFM. But that the performances of random system converges towards that of the limit system requires an explicit proof. We are currently working towards this for general systems. In the Appendix B we present an example proof for a single queue with vacations. At the end of the next section after deriving the achievable region for SSFM, we provide a numerical illustration to show that the achievable region of the random system (even for moderately large values of μ) has similar structure as that of SSFM.

V. Steady state fluid model (SSFM)

As shown in the previous section with arrival/departure rates converging towards infinity, we have continuous flow of arrivals and departures (when service is offered), resulting in fluid queues. The SSFM consists of two storage tanks, two inlets and one outlet pipe. The fluid flows from inlets to the corresponding storage tanks at a constant rate. Outlet pipe is controlled by switch to move it from one storage tank to another. Consider the following notational analogy.

1) Switching time: s units be the time required to move outlet pipe from one storage tank to another.
2) Service rate: μ_1 and μ_2 are rate of outflow from tank 1 and tank 2 respectively. We will take $\mu_1 = \mu_2 = \mu$ whenever required in later sections.
3) Arrival rate: λ_1 and λ_2 are rate of inflow in tank 1 and tank 2 respectively.
4) Switching policy parameters: γ_1 and γ_2 are delay priority parameters*.

Let Q_1 and Q_2 be the fluid level in tank 1 and tank 2 respectively. Analogous switching rules will be as follows:

- When outlet pipe is in tank i: if $\gamma_i Q_i \geq Q_{-i}$ then stay at tank i else switch.

A. Switching Cycle and Stability Condition

Let k_1 and k_2 be the levels of fluid in tank 1 and tank 2 respectively in steady state, when service just starts. Assume that o_i is the level of fluid in tank $-i$ when service starts in tank i. Figure 1 illustrates the change in fluid level for one

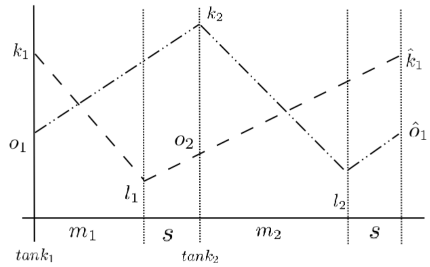

Fig. 1: Fluid level of deterministic system in one cycle

cycle. Consider the following steps for cycle completion:

Step 1: We start with tank 1. Let's assume that outlet pipe is moved after m_1 time units. Visit time m_1 satisfies:

$$\begin{aligned} \gamma_1(k_1 + m_1(\lambda_1 - \mu_1)) &= o_1 + m_1\lambda_2 \text{ and} \\ k_1 &= l_1 - (\lambda_1 - \mu_1)m_1, \end{aligned} \quad (3)$$

where $m_1\lambda_1$ and $m_1\lambda_2$ are amount of fluid added to tank 1 and tank 2 respectively, while $m_1\mu_1$ is amount of fluid removed from tank 1.

Step 2: During switching time s which is used to move outlet pipe from tank 1 to tank 2, $s\lambda_1$, $s\lambda_2$ amount of fluid gets added to tank 1 and tank 2 respectively. So the levels of fluid when outlet pipe reaches tank 2 are:

$$o_2 = k_1 + m_1(\lambda_1 - \mu_1) + s\lambda_1, \quad (4)$$
$$k_2 = o_1 + m_1\lambda_2 + s\lambda_2. \quad (5)$$

Step 3: Starting from tank 2, Let's assume that outlet pipe is moved after m_2 time units. Visit time m_2 satisfies:

$$\gamma_2(k_2 + m_2(\lambda_2 - \mu_2)) = o_2 + m_2\lambda_1 \quad (6)$$

where $m_2\lambda_1$ and $m_2\lambda_2$ are amount of fluid added to tank 1 and tank 2 respectively, while $m_2\mu_2$ is amount of fluid removed from tank 2.

Step 4: During switching time s which is used to move outlet pipe from tank 2 to tank 1; $s\lambda_1$ and $s\lambda_2$ be the amount of fluid gets added to tank 1 and tank 2 respectively. So the level of fluid in tank 1 and tank 2 when outlet pipe reaches tank 1 equals:

$$\hat{k}_1 = (k_1 + m_1(\lambda_1 - \mu_1) + s\lambda_1) + m_2\lambda_1 + s\lambda_1, \quad (7)$$
$$\hat{o}_1 = (o_1 + m_1\lambda_2 + s\lambda_2) + m_2(\lambda_2 - \mu_2) + s\lambda_2. \quad (8)$$

In steady state (if possible to reach) $\hat{k}_1 = k_1$ and $\hat{o}_1 = o_1$, i.e., level of fluid after each cycle is constant. This implies (using equations (7) and (8)):

$$\rho_1 = \frac{\lambda_1}{\mu_1} = \frac{m_1}{m_1 + m_2 + 2s} \text{ and } \rho_2 = \frac{\lambda_2}{\mu_2} = \frac{m_2}{m_1 + m_2 + 2s}. \quad (9)$$

Using the above equations, we obtain the following theorem whose proof is in Appendix.

Theorem 3. *SSFM with* (γ_1, γ_2) *scheduler is stable if* $\rho < 1$ *and* $\gamma_1\gamma_2 > 1$. ■

*γ_i and β_i are different from each other, as β_is were switching parameters when we considered waiting time of longest waiting customer, while γ_i are switching parameters defined on height of fluid in tank, or equivalently number of waiting customers. We also derive a relation between β_i and γ_i towards the end of section V.

B. Stationary Performance

Stationary visit times: On solving the linear equations of (9):

$$m_1^* = \frac{2s\rho_1}{1-\rho} \text{ and } m_2^* = \frac{2s\rho_2}{1-\rho}. \quad (10)$$

Average waiting time: Total fluid in a tank during one cycle can be calculated by determining area under the curve in Figure 1. Total fluid in tank 1 in one cycle equals:

$$\frac{1}{2}\left[(\lambda_1 - \mu_1)m_1^*\right] m_1^* + \frac{1}{2}\left[\lambda_1(m_2^* + 2s)\right](m_2^* + 2s)$$
$$+[k_1 - \lambda_1(m_2^* + 2s)](m_1^* + m_2^* + 2s).$$

Under stationarity ($k_1 = \hat{k}_1$), $(\lambda_1 - \mu_1)m_1^* = \lambda_1(m_2^* + 2s)$ holds and the above equation simplifies to:

$$\frac{1}{2}(2k_1 - \lambda_1(m_2^* + 2s))(m_1^* + m_2^* + 2s).$$

$$\text{Average fluid in tank } 1 = \frac{\text{total fluid in tank 1}}{\text{cycle time}}$$
$$= k_1 - \frac{1}{2}\lambda_1(m_2^* + 2s). \quad (11)$$

Using Little's law we obtain the average waiting time of fluid in tank 1,

$$\bar{w}_1 = \frac{k_1}{\lambda_1} - \frac{(m_2^* + 2s)}{2}.$$

Using equations (3) and (22) of Appendix, we further simplify:

$$\bar{w}_1 = \frac{1}{\lambda_1}\left(\frac{c_1 + \gamma_2 c_2}{\gamma_1\gamma_2 - 1}\right) + \varpi_1, \quad \varpi_1 := \frac{s(1-\rho_1)}{(1-\rho)}, \quad (12)$$
$$c_1 = \frac{s\lambda_1(1 - \rho_1 + \rho_2)}{1-\rho} \text{ and } c_2 = \frac{s\lambda_2(1 + \rho_1 - \rho_2)}{1-\rho}.$$

Similarly, average waiting time of fluid in tank 2

$$\boxed{\bar{w}_2 = \frac{1}{\lambda_2}\left(\frac{c_2 + \gamma_1 c_1}{\gamma_1\gamma_2 - 1}\right) + \varpi_2 \text{ with } \varpi_2 := \frac{s(1-\rho_2)}{(1-\rho)}.} \quad (13)$$

C. Achievable Region

Following theorem characterizes the achievable region with (γ_1, γ_2) schedulers (proof in Appendix).

Theorem 4. *Achievable region of performance for (γ_1, γ_2) scheduler is given by:*
$\mathcal{A}^P = \{(\bar{w}_1, \bar{w}_2) : \bar{w}_i > \varpi_i, i = 1, 2\}$ *with* $\varpi_i := \frac{s(1-\rho_i)}{1-\rho}$. ■

Recall that (γ_1, γ_2) scheduler defines the switching rule using the numbers in queues while (β_1, β_2) scheduler utilizes the waiting times of the longest waiting customers. Following theorem gives us the relation under which the two types of schedulers achieve the same performance. The proofs of Theorems 5, 6 are available in Appendix.

Theorem 5. *Stability criterion for (β_1, β_2) schedulers is given by $\rho < 1$ and $\beta_1\beta_2 > 1$. Further, the average waiting time performance of (γ_1, γ_2) as well as (β_1, β_2) schedulers is the same when they are related according to:*

$$\frac{\gamma_1}{\beta_1} = \frac{\beta_2}{\gamma_2} = \frac{\lambda_2}{\lambda_1}.$$ ■

Recall that the notation ex implies exhaustive service discipline. Following theorem characterizes the performance measure of policies of type (β, ex) or (ex, β).

Theorem 6. *When (ex, β) policy, with $\beta \in \mathbb{R}$, is implemented the average waiting time of customers at $\mathcal{Q}_1$ is ϖ_1 of Theorem 4. The average waiting time at $\mathcal{Q}_2$ takes any value in interval (ϖ_2, ∞) depending upon β.*

Following theorem helps in characterizing the completeness of achievable region by (β_1, β_2) scheduler along with exhaustive scheduler, i.e., completeness of $\mathcal{B}^P$. This says that to achieve the performance below the one given by exhaustive policy one has to made the other queue stable (proof in Appendix).

Theorem 7. *Consider two class polling system where $\mathcal{Q}_1$ is stable and β is any time invariant policy. If $\bar{w}_1 < \varpi_1$, then $\mathcal{Q}_2$ is unstable.* ■

Using the Theorems 4, 5, 6 and 7 we obtain the completeness of $\mathcal{B}^P$ and summarize it as below:

Theorem 8. *Achievable region for SSFM is given by*

$$\mathcal{A} = \{[\varpi_1, \infty) \times [\varpi_2, \infty)\} \quad (14)$$

and $\mathcal{B}^P$ is a complete class of scheduling policies. Similarly, $\mathcal{B}^P_\gamma := \{(\gamma_1, \gamma_2); \gamma_i \in \mathbb{R}^e\}$ is another complete class. ■

Thus in contrast to the work conserving queuing systems, *we have an unbounded achievable region.* However we notice that the exhaustive service discipline (ex, ex) achieves the minimum waiting time $((\varpi_1, \varpi_2))$ at both the queues. It is well known that exhaustive is optimal for unfinished workload. Here we notice that exhaustive is optimal even from the perspective of individual classes and this is true for high arrival and departure rates.

Numerical illustration: Random systems with large μ

We build a Monte-Carlo simulator for two class polling system with β_1, β_2 scheduler. We also implemented exhaustive schedulers. We validated the simulator, for the case with switching time equal to 0, with the theoretical results of delay priority schedulers of [18].

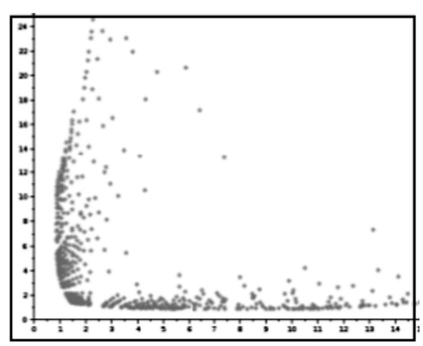

Fig. 2: Achievable region via simulations for large λ and μ

Simulations are conducted using Poisson arrivals and exponential service times while keeping switching time $s = 0.1$.

Figure 2 is plotted for rates, $\lambda_i = 4.5$ and $\mu_i = 10$ for all i. By Theorem 8, the achievable region for SSFM is unbounded and is rectangular in shape and we notice a similar shape in Figure 2. Thus the shape of the achievable region of the random system is close to that of the SSFM. These simulations were conducted at a load factor $\rho = 0.9$ and we have similar approximation result even for smaller load factors in Table I.

Results of Table I are obtained with $\lambda_1/\lambda_2 = 0.5$ and $\rho = 0.3$ and for varying values of μ for two different schedulers β. We notice that the average waiting times converge towards that of the SSFM. The performance is considerably close to that of SSFM, for values of μ greater than 200.

			Simulation		SSFM	
μ	β_1	β_2	$\bar{w}_1$	$\bar{w}_2$	$\bar{w}_1$	$\bar{w}_2$
100	2	4	0.1848	0.1446	0.2245	0.1775
200	2	4	0.2048	0.1615	0.2245	0.1775
500	2	4	0.2169	0.1712	0.2245	0.1775
100	7	17	0.1420	0.1167	0.1484	0.1246
200	7	17	0.1464	0.1228	0.1484	0.1246
500	7	17	0.1478	0.1242	0.1484	0.1246

TABLE I: Performance of Random System and SSFM

VI. OPTIMIZATION

We consider two constraint optimization problems: admission control and revenue maximization considering the losses. We consider $\mu_1 = \mu_2 = \mu$ to simplify the discussions.

A. Admission Control

Consider a two class polling system with primary and secondary customers, with respective arrival rates λ_1 and λ_2. Let $P(\bar{w}_2)$ be the price paid by secondary class customers which depends on its QoS, $\bar{w}_2$. Assume that price function *$P(\bar{w}_2)$ is monotonically decreasing in $\bar{w}_2$*. The revenue can be increased by increasing the arrival rate λ_2 of secondary customers, however this has to be done while maintaining the QoS at $\mathcal{Q}_1$. Thus we are interested in the following admission control problem

$$\textbf{P1:} \quad \max_{\lambda_2, \beta \in \mathcal{B}^P} \lambda_2 P(\bar{w}_2) \quad \text{Subject to: } \bar{w}_1 \le \eta_1.$$

By virtue of Theorem 8, we consider $\mathcal{B}^P$ in **P1**. Clearly, for any λ_2 if there exists a $\beta \in \mathcal{B}^P$ such that $\bar{w}_1 \le \eta_1$ then $\varpi_1(\lambda_2) \le \eta_1$[†]. By Theorem 8 , ϖ_1 is the best achievable mean waiting time for class 1 (achieved when $\beta_1 = ex$) and hence the above is true. Thus the solution for problem **P1** is also obtained by maximizing over the following alternate domain:

$$\mathcal{Y} = \left\{ (\lambda_2, \beta) : \lambda_2 > 0, \frac{s(1-\rho_1)}{(1-\rho)} \le \eta_1, \beta = (ex, \beta_2) \in \mathcal{B}^P \right\}.$$

Note that for every fixed λ_2, mean waiting time in secondary class, $\bar{w}_2$, is minimized (i.e., $P(\bar{w}_2)$ is maximized) by exhaustive scheduling policy at class 2, i.e., $\beta_2 = ex$. Hence optimality will not be lost if we optimize over following set:

$$\mathcal{Z} = \left\{ (\lambda_2, \beta) : \lambda_2 > 0, \frac{s(1-\rho_1)}{(1-\rho)} \le \eta_1, \beta = (ex, ex) \right\}.$$

[†]Dependency of ϖ_1 on λ_2 is explicitly mentioned hence forth.

Substituting the performance at (ex, ex) **P1** simplifies to:

$$\max_{\lambda_2} \lambda_2 P\left(\frac{s(\mu - \lambda_2)}{\mu - \lambda_1 - \lambda_2} \right) \text{ such that } \frac{s(\mu - \lambda_1)}{\mu - \lambda_1 - \lambda_2} \le \eta_1.$$

By monotonicity of functions involved, the optimizer of the above problem λ_2^* satisfies the constraint with equality, i.e.,

$$\frac{s(\mu - \lambda_1)}{(\mu - \lambda_1 - \lambda_2^*)} = \eta_1. \text{ Simplifying, } \lambda_2^* = \frac{(\eta_1 - s)(1-\rho_1)\mu}{\eta_1}.$$

Thus when one has to maintain the average waiting time of a queue below a level, however low the level could be, exhaustive policy (ex, ex) turns out to be the optimal policy as long as the QoS demanded is achievable ($\eta_1 < \varpi_1(\lambda_2)$ for some λ_2). This is not surprising given the rectangular shape of the achievable region (14) of Theorem 8.

B. Revenue maximization with losses

We now consider a scenario in which the arrival rate λ_2 is fixed. When λ_2 is such that, QoS of the primary customer can't be maintained, i.e., when $\varpi_1(\lambda_2) < \eta_1$, then the service provider can provide service only for a fraction of the secondary customers. Towards this we assume there is a limit, B, on the buffer size at $\mathcal{Q}_2$ and choose the buffer size B optimally. Let f be the fraction of customers lost in $\mathcal{Q}_2$ with buffer size B. Then only $\lambda_2(1-f)$ fraction of the customers are served and hence profit is obtained only from them. Thus with $\eta_1 < \varpi_1(\lambda_2)$, it is appropriate to consider the following revenue optimization problem[‡]:

$$\textbf{P2:} \quad \max_{B, \beta \in \mathcal{B}^P} \lambda_2(1-f)P(\bar{w}_2) \quad \text{Subject to: } \bar{w}_1 \le \eta_1.$$

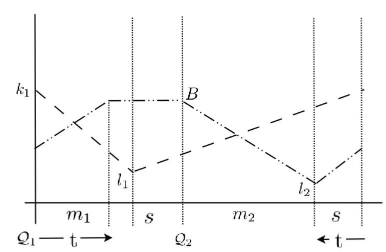

Fig. 3: Flow of fluid with buffer constraint in $\mathcal{Q}_2$

Fraction of losses: Consider SSFM and the notation t for time at which tank level reaches B after switching from $\mathcal{Q}_2$. Note that

$$B = l_2 + t\lambda_2. \tag{15}$$

Flow balance equation at $\mathcal{Q}_1$ and $\mathcal{Q}_2$ gives us (see Figure 3):

$$t\lambda_2 = m_2(\mu_2 - \lambda_2), \quad m_1(\mu_1 - \lambda_1) = (m_2 + 2s)\lambda_1. \tag{16}$$

In **P2** the constraint is $\bar{w}_1 \le \eta_1$ and we have $\varpi_1(\lambda_2) > \eta_1$. One can achieve a smaller average waiting time at $\mathcal{Q}_1$ only when there are losses at $\mathcal{Q}_2$. This implies $t < m_1 + 2s$ for

[‡]We again need to establish the completeness of $\mathcal{B}^P$, and we are working towards it.

all B satisfying the constraint. For such B, the fraction of customers lost equals:

$$f = \frac{(m_1 + 2s - t)\lambda_2}{(m_1 + m_2 + 2s)\lambda_2}.$$

Using equation (16), and upon further simplification using (15)

$$f = 1 - \frac{m_2\mu_2(\mu_1 - \lambda_1)}{(m_2 + 2s)\lambda_2\mu_1} = \frac{2s(1-\rho_2)}{(B - l_2)\rho_2 + 2s(1-\rho_2)}. \quad (17)$$

Note that the losses increases with decrease in B, i.e., with $B \downarrow$, $f \uparrow$.

Expressions for $\bar{w}_1$ and $\bar{w}_2$: Let z be the level of fluid in $\mathcal{Q}_2$ after time m_1. Note that

$$z = \min\{B, (m_1 + s)\lambda_2 + l_2\}.$$

Switching condition gives us

$$l_1 = \frac{z}{\beta_1} \text{ and } l_2 = \frac{l_1 + (m_2 + s)\lambda_1}{\beta_2}.$$

Further solving for l_2, we get

$$l_2 = \frac{(\mu_2 - \lambda_2)l_1 + (B + s(\mu_2 - \lambda_2))\lambda_1}{\beta_2(\mu_2 - \lambda_2) + \lambda_1}. \quad (18)$$

Following the similar analysis as earlier, we obtain mean waiting times as

$$\begin{aligned} \bar{w}_1 &= \frac{m_2 + 2s}{2} + \frac{l_1}{\lambda_1} = \frac{(B - l_2) + 2s(\mu_2 - \lambda_2)}{2(\mu_2 - \lambda_2)} + \frac{l_1}{\lambda_1}, \\ \bar{w}_2 &= \frac{B - l_2}{2\lambda_2(1-f)} + \frac{l_2}{\lambda_2(1-f)} = \frac{B + l_2}{2\lambda_2(1-f)}. \quad (19) \end{aligned}$$

Note in the above that when Little's law has to be applied to get $\bar{w}_2$ we need to divide by $\lambda_2(1-f)$ as this is the effective rate at which arrivals enter $\mathcal{Q}_2$.

Step 1: *Optimizer for any fixed B and β_2:*

With $\beta_1 \to \infty$, we have following limits which are smaller than the corresponding values with β_1 finite (note $z \le B$),

$$l_1 \to 0 \text{ and } l_2 \to \frac{(B + s(\mu_2 - \lambda_2))\lambda_1}{\beta_2(\mu_2 - \lambda_2) + \lambda_1}.$$

Clearly from equations (17), (19), the fraction lost f as well as both the average waiting times $\bar{w}_1$, $\bar{w}_2$ are simultaneously minimized with $\beta_1 = ex$ (or equivalently letting $\beta_1 \to \infty$).

Thus the optimization problem **P2** simplifies to:

$$\mathbf{P2}' : \max_{B,\beta_2} \lambda_2(1-f)P(\bar{w}_2(ex, \beta_2)) \quad \text{s.t. } \bar{w}_1(ex, \beta_2) \le \eta_1.$$

Step 2: *With* $\beta_1 = ex$ the required functions become:

$$\begin{aligned} \bar{w}_1 &= \frac{(B - l_2) + 2s(\mu_2 - \lambda_2)}{2(\mu_2 - \lambda_2)}, \; l_2 = \frac{(B + s(\mu_2 - \lambda_2))\lambda_1}{\beta_2(\mu_2 - \lambda_2) + \lambda_1}, \\ \bar{w}_2 &= \frac{B + l_2}{2\lambda_2(1-f)}, \; f = \frac{2s(1-\rho_2)}{(B - l_2)\rho_2 + 2s(1-\rho_2)}. \quad (20) \end{aligned}$$

Consider any B, β_2 satisfying the constraint ($\bar{w}_1 \le \eta_1$):

$$B - l_2(B, \beta 2) \le \eta' := \eta_1 2(\mu_2 - \lambda_2)(1 - 2s).$$

Then there exists a pair (B', l_2'), with $B' := B - l_2(B, \beta 2)$ and $l_2' := 0$ (or equivalently $\beta_2 = ex$), which still satisfies the constraint and

$$f(B', l_2') = f(B, l_2) \text{ and } \bar{w}_2(B', l_2') < \bar{w}_2(B, l_2).$$

Since the price function $P(.)$ increases with decrease in $\bar{w}_2$ we can conclude (via contradiction) that **P2** is equivalent to:

$$\mathbf{P2}'' : \max_{B} \lambda_2(1-f)P(\bar{w}_2(ex, ex)) \text{ s.t., } \bar{w}_1(ex, ex) \le \eta_1.$$

To conclude we have

Theorem 9. *The revenue optimization problem **P2** is optimized by exhaustive schedulers $\beta = (ex, ex)$ and the optimal level B^* is obtained by solving the simplified optimization problem:*

$$\max_B \frac{B\rho_2\lambda_2}{B\rho_2 + 2s(1-\rho_2)} P\left(\frac{B\rho_2 + 2s(1-\rho_2)}{2\rho_2\lambda_2^2}\right) \text{ s.t., } B \le \eta'.$$

■

Further optimization of the problem depends upon the price function $P(.)$. However we again notice that exhaustive (ex, ex) is the optimal policy.

VII. Conclusions

We investigated the idea of differential fairness in polling systems. Here different classes of customers were given different levels of fairness depending either upon the requirement or upon the price a class pays for the service. We used achievable region approach for solving two relevant problems. Towards this, we first proposed and proved that a class of delay priority kind of schedulers along with exhaustive policy forms a complete class for two queue polling systems. We obtained the performance measures and then the achievable region for a limit system with fluid queues, when it uses the proposed priority schedulers. Using Monte Carlo simulations we showed that the performance of the random systems (ones with discrete arrivals) converges towards that of the analyzed limit system with fluid queues. We conclude this study with the following observations for systems with large arrival and departure rates: a) the achievable region is unbounded; b) exhaustive service discipline is optimal even from the perspective of the individual classes. However exhaustive policy may not be optimal in above sense, for systems with moderate arrival rates. We have some initial observations in this direction and this is a topic of future research.

References

[1] M. Boon, R. Van der Mei, and E. Winands, "Applications of polling systems," *Surveys in Operations Research and Management Science*, vol. 16, no. 2, pp. 67–82, 2011.

[2] H. Levy and M. Sidi, "Polling systems: Applications, modeling, and optimization," *Communications, IEEE Transactions on*, 1990.

[3] O. Boxma and W. Groenendijk, "Pseudo-conservation laws in cyclic-service systems," *Journal of Applied Probability*, pp. 949–964, 1987.

[4] H. Levy, M. Sidi, and O. J. Boxma, "Dominance relations in polling systems," *Queueing systems*, vol. 6, no. 1, pp. 155–171, 1990.

[5] V. Kavitha and E. Altman, "Continuous polling models and application to ferry assisted wlan," *Annals of Operations Research*, vol. 198, no. 1, pp. 185–218, 2012.

[6] E. G. Coffman and I. Mitrani, “A characterization of waiting time performance realizable by single server queues,” *Operations Research*, vol. 28, pp. 810 – 821, 1979.
[7] I. Mitrani and J. Hine, “Complete parametrized families of job scheduling strategies,” *Acta Informatica*, vol. 8, pp. 61– 73, 1977.
[8] 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.
[9] D. Bertsimas, “The achievable region method in the optimal control of queueing systems; formulations, bounds and policies,” *Queueing Systems 21 (1995) 337-389*, vol. 21, pp. 337–389, 1995.
[10] O. Czerniak and U. Yechiali, “Fluid polling systems,” *Queueing Systems*, vol. 63, no. 1-4, pp. 401–435, 2009.
[11] H. Takagi, *Analysis of Polling Systems*. MIT Press, 1986.
[12] L. Rogers, “Fluid models in queueing theory and wiener-hopf factorization of markov chains,” *The Annals of Applied Probability*, pp. 390–413, 1994.
[13] A. Wierman, “Fairness and scheduling in single server queues,” *Surveys in Operations Research and Management Science*, vol. 16, no. 1, pp. 39–48, 2011.
[14] H. J. Kushner and P. A. Whiting, “Convergence of proportional-fair sharing algorithms under general conditions,” *Wireless Communications, IEEE Transactions on*, vol. 3, no. 4, pp. 1250–1259, 2004.
[15] F. P. Kelly, A. K. Maulloo, and D. K. Tan, “Rate control for communication networks: shadow prices, proportional fairness and stability,” *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.
[16] B. Avi-Itzhak, E. Brosh, and H. Levy, “Sqf: A slowdown queueing fairness measure,” *Performance Evaluation*, vol. 64, no. 9, pp. 1121–1136, 2007.
[17] T.-C. Lin, Y. S. Sun, S.-C. Chang, S.-I. Chu, Y.-T. Chou, and M.-W. Li, “Priority-based internet access control for fairness improvement and abuse reduction,” in *Quality of Service in Multiservice IP Networks*. Springer, 2003, pp. 659–671.
[18] L. Kleinrock, “A delay dependent queue discipline,” *Naval Research Logistics Quarterly*, vol. 11, pp. 329–341, September-December 1964.
[19] L. Kleinrock and R. P. Finkelstein, “Time dependent priority queues,” *Operations Research*, vol. 15, no. 1, pp. 104–116, 1967.
[20] D. Raz, H. Levy, and B. Avi-Itzhak, “A resource-allocation queueing fairness measure.” In Proc. of ACM Sigmetrics-Performance, 2004.
[21] J.G. Shanthikumar, and D. D. Yao. ”Multiclass queueing systems: polymatroidal structure and optimal scheduling control,” Oper. Res. 40, pp. 293-299, 1992.
[22] V. Kavitha, N. Hemachandra and D. Das, “Fairness via Priority Scheduling” , Allerton 2013.

APPENDIX

Proof of Theorem 1: To prove the result, we need to show that $P(\omega : \frac{\Lambda^\lambda(t)}{\lambda} \to t) = 1$. Equivalently, we will show that $P(\omega : \frac{\Lambda^\lambda(t)}{\lambda t} \to 1) = 1$. Note that from supposition of theorem it follows that

$$P\left(\omega : \frac{\Lambda^\lambda(t)}{\lambda t} \to 1\right) = P\left(\omega : \frac{\Lambda^1(\lambda t)}{\lambda t} \to 1\right).$$

Note that $\Lambda^1(.)$ is a renewal process with unit mean inter arrival times. By elementary renewal theorem applied to process $\{\Lambda^1(.)\}$, when $\lambda \to \infty$, we get:

$$P\left(\omega : \frac{\Lambda^1(\lambda t)}{\lambda t} \to 1\right) = 1.$$

So $\frac{\Lambda^\lambda(t)}{\lambda} \to t \ a.s.$ for all t. ∎

Proof of Theorem 2: To prove

$$\{\Lambda^\lambda(t); t \geq 0\} \stackrel{d}{=} \{\Lambda^1(\lambda t); t \geq 0\}.$$

That is, we need to show for any t and x

$$P(\omega : \Lambda^\lambda(t) \leq x) = P(\omega : \Lambda^1(\lambda t) \leq x).$$

Note that

$$\Lambda^\lambda(t) = \sup_k \left\{ \sum_{i=1}^k A_i^\lambda \leq t \right\} = \sup_k \left\{ \sum_{i=1}^k \frac{A_i^1}{\lambda} \leq t \right\}$$
$$= \sup_k \left\{ \sum_{i=1}^k A_i^1 \leq \lambda t \right\} = \Lambda^1(\lambda t).$$

So it follows that $P(\omega : \Lambda^\lambda(t) \leq x) = P(\omega : \Lambda^1(\lambda t) \leq x)$. ∎

Proof of Theorem 3: On solving the system of linear equations (9):

$$m_1 = \frac{2s\rho_1}{1-\rho} \text{ and } m_2 = \frac{2s\rho_2}{1-\rho}. \tag{21}$$

Note that m_1 and m_2 must be positive and finite for stability. Hence we get $\rho < 1$ as one of the stability conditions. Apart from this we will need that the level reached at switching time must be positive. That is, $l_1 = k_1 - m_1(\mu_1 - \lambda_1) \geq 0$ and $l_2 = k_2 - m_2(\mu_2 - \lambda_2) \geq 0$. We clearly have

$$\gamma_1 l_1 = l_2 + s\lambda_2 + m_1\lambda_2 \text{ and } \gamma_2 l_2 = l_1 + s\lambda_1 + m_2\lambda_1.$$

On solving above set of equations for l_1 and l_2, we get

$$l_1 = \frac{c_1 + \gamma_2 c_2}{\gamma_1\gamma_2 - 1} \text{ and } l_2 = \frac{c_2 + \gamma_1 c_1}{\gamma_1\gamma_2 - 1}. \tag{22}$$

Note that $c_1 = (s + m_2)\lambda_1$ and $c_2 = (s + m_1)\lambda_2$. So for l_1 and l_2 to be non negative, $\gamma_1\gamma_2 > 1$. ∎

Proof of Theorem 4: Let average waiting time of fluid in tank 1, $\bar{w}_1$, be fixed at $l + \varpi_1$ (with some $l > 0$), we will prove that one can achieve any value of $\bar{w}_2$ in (ϖ_2, ∞) by varying γ_1 and γ_2 appropriately and keeping $\bar{w}_1$ at level $l + \varpi_1$. Similar things can be proved for $\bar{w}_2$ and this will prove the result.

Now, we look at the value of γ_1 and γ_2 together that achieve $\bar{w}_1 = l + \varpi_1$. Using equation (12),

$$l = \frac{1}{\lambda_1}\left(\frac{c_1 + \gamma_2 c_2}{\gamma_1\gamma_2 - 1}\right) \tag{23}$$

On further simplifying, we get

$$\gamma_1 = \frac{c_1}{\lambda_1 l \gamma_2} + \frac{1}{\gamma_2} + \frac{c_2}{\lambda_1 l}. \tag{24}$$

From the above equation, $(\gamma_1\gamma_2 - 1) = (c_1 + \gamma_2 c_2)/(\lambda_1 l)$. Now from (13) we have:

$$\bar{w}_2 = \frac{c_2 + \gamma_1 c_1}{c_1 + \gamma_2 c_2}\frac{\lambda_1 l}{\lambda_2} + \varpi_2. \tag{25}$$

From (24) when γ_2 is decreased, γ_1 is increased to maintain $\bar{w}_1$ at $l + \varpi_1$ and then from (25), $\bar{w}_2$ increases. Large values of $\bar{w}_2$ are achieved as $\gamma_2 \to 0$ and in this limit $\gamma_1 \to \infty$. Thus we have, $\bar{w}_2 \uparrow \infty$ when $\gamma_2 \downarrow 0$.

Again from (24) when γ_2 is increased, γ_1 is decreased to maintain $\bar{w}_1 = l + \varpi_1$. Then, from (25), $\bar{w}_2$ decreases. Minimum value of $\bar{w}_2$ is achieved when $\gamma_2 \to \infty$ and then $\gamma_1 \to \frac{c_2}{\lambda_1 l}$ (from equation (24)) and then $\bar{w}_2 \downarrow \varpi_2$. ∎

Proof of Theorem 5: Let t_i be the maximum time fluid waits before completion of m_i where fluid arrived but not removed from the system in current cycles's m_i interval. Note that t_i satisfies

$$\lambda_1 t_1 = l_1 \text{ and } \lambda_2 t_2 = l_2 \tag{26}$$

Switching conditions are satisfied by a (β_1, β_2) scheduler at the same instance with (γ_1, γ_2) scheduler, if following conditions are satisfied:

$$\beta_1 t_1 = t_2 + s + m_1$$

where $s + m_1$ is the extra delay which fluid in tank 2 suffers, while fluid in tank 1 has t_1 delay. Similarly

$$\beta_2 t_2 = t_1 + s + m_2$$

On using equation (26), we get

$$\beta_1 \frac{l_1}{\lambda_1} = \frac{l_2}{\lambda_2} + s + m_1 \text{ and } \beta_2 \frac{l_2}{\lambda_2} = \frac{l_1}{\lambda_1} + s + m_2 \tag{27}$$

On substituting the value of l_1 and l_2 from equation (22), we get

$$\beta_1 = \frac{\lambda_1}{\lambda_2}\left(\frac{c_2 + \gamma_1 c_1}{c_1 + \gamma_2 c_2}\right) + \left(\frac{\gamma_1\gamma_2 - 1}{c_1 + \gamma_2 c_2}\right)(m_1 + s)\lambda_1 \tag{28}$$

$$\beta_2 = \frac{\lambda_2}{\lambda_1}\left(\frac{c_1 + \gamma_2 c_2}{c_2 + \gamma_1 c_1}\right) + \left(\frac{\gamma_1\gamma_2 - 1}{c_2 + \gamma_1 c_1}\right)(m_2 + s)\lambda_2 \tag{29}$$

Note that (β_1, β_2) and (γ_1, γ_2) schedulers will give same performance if they are related by above set of equation. Further by substitution, we observe that when scheduling parameters are related via following relation, then above set of equations are satisfied

$$\frac{\gamma_1}{\beta_1} = \frac{\beta_2}{\gamma_2} = \frac{\lambda_2}{\lambda_1}.$$

Hence statement of theorem follows. ■

Proof of Theorem 6: Note that exhaustive policy is implemented at $\mathcal{Q}_1$, i.e., server switches to $\mathcal{Q}_2$ only when $\mathcal{Q}_1$ is empty (See figure 4).

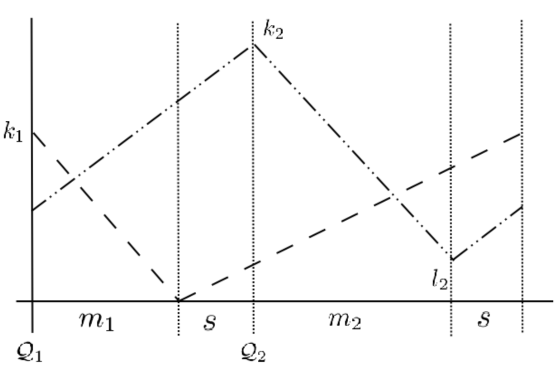

Fig. 4: Fluid level of deterministic system when $\mathcal{Q}_1$ is exhaustive

$$m_1(\mu_1 - \lambda_1) = k_1$$

Total fluid in Tank1 in one cycle

$$= \frac{1}{2} m_1(\mu_1 - \lambda_1) m_1 + \frac{1}{2}(m_2 + 2s)\lambda_1(m_2 + 2s)$$

Average fluid in tank 1, i.e., $\bar{w}_1 = \dfrac{\text{total fluid in tank 1}}{\text{cycle time}}$

$$= \frac{\frac{1}{2} m_1(\mu_1 - \lambda_1) m_1 + \frac{1}{2}(m_2 + 2s)\lambda_1(m_2 + 2s)}{(m_1 + m_2 + 2s)}$$

Under stationarity $m_1(\mu_1 - \lambda_1) = (m_2 + 2s)\lambda_1$. On using this in above equation we get $\bar{w}_1 = \varpi_1$. Let l_2 be the level of fluid in $\mathcal{Q}_2$ when service ends at $\mathcal{Q}_2$. From equation (13) and (22), we get $\bar{w}_2 = \dfrac{l_2}{\lambda_2} + \varpi_2$. Switching condition gives us:

$$\gamma_2 l_2 = k_1 - s\lambda_1 = m_2 + s\lambda_1$$

$$l_2 = \frac{m_2 + s\lambda_1}{\gamma_2}$$

clearly, on varying γ_2 (or equivalently varying β_2 from previous theorem), any value of $\bar{w}_2$ from (ϖ_2, ∞) can be achieved. Similar arguments can be made when $\mathcal{Q}_2$ is exhaustive. ■

Proof of Theorem 7: Consider any time invariant scheduling policy which renders $\mathcal{Q}_1$ stable. Number in $\mathcal{Q}_1$ for such policies should follow the pattern as in figure 1. That is, when server reaches $\mathcal{Q}_1$ it's level is k_1 and when it leaves $\mathcal{Q}_1$ it's level is l_1, irrespective of the number of visit.

Let m_1 and m_2 be the times spent respectively in $\mathcal{Q}_1$ and $\mathcal{Q}_2$. When $\mathcal{Q}_1$ is stable, these visit times will be stationary (i.e., do not depend upon the number of visit). From flow balance in $\mathcal{Q}_1$, we have $l_1 = k_1 - \lambda_1(m_2 + 2s)$.

Now the average number in $\mathcal{Q}_1$ equals $(m_2 + 2s)\lambda_1/2 + l_1$ and by Little's law, we get

$$\bar{w}_1 = \frac{m_2 + 2s}{2} + \frac{l_1}{\lambda_1}. \tag{30}$$

If $\mathcal{Q}_2$ were also stable then balance equations are also satisfied at $\mathcal{Q}_2$ and then the visit times m_1^*, m_2^* are given by (10). Recall,

$$m_2^* = \frac{2s\rho_2}{(1-\rho)} \text{ and } m_1^* = \frac{2s\rho_1}{(1-\rho)}.$$

Note that $\bar{w}_1 < \varpi_1$ implies $m_2 < m_2^*$, as otherwise we have a contradiction: from (30) $\bar{w}_1 > (m_2^* + 2s)/2 = \varpi_1$.

Let l_k be the level of fluid in $\mathcal{Q}_2$ at the time of switching in k-th cycle. Clearly, we have

$$l_{k+1} = l_k - m_2\mu_2 + (m_1 + m_2 + 2s)\lambda_2. \tag{31}$$

From flow balance in $\mathcal{Q}_1$, we have

$$m_1(\mu_1 - \lambda_1) = (m_2 + 2s)\lambda_2.$$

On using above expression in equation (31), we have

$$l_{k+1} - l_k = \frac{m_2(\mu_1\lambda_2 - \mu_1\mu_2 + \lambda_1\mu_2) + 2s\mu_1\lambda_2}{\mu_1 - \lambda_1} = \zeta.$$

Note that $\mu_1\lambda_2 - \mu_1\mu_2 + \lambda_1\mu_2 = \mu_1\mu_2(\rho - 1) < 0$ and so

$$l_{k+1} - l_k > \frac{m_2^*(\mu_1\lambda_2 - \mu_1\mu_2 + \lambda_1\mu_2) + 2s\mu_1\lambda_2}{\mu_1 - \lambda_1}.$$

On simplifying the above equation using $m_2^* = \frac{2s\rho_2}{1-\rho}$ we get RHS equal to 0. Hence

$$l_{k+1} - l_k > 0 \text{ if } m_2 < m_2^*.$$

Thus level of $\mathcal{Q}_2$ will always be increasing under above mentioned condition, hence $\mathcal{Q}_2$ will be unstable. Also note that the growth rate is linear in k as $l_k = k\zeta + l_0$ where l_0 is the initial level and that ζ is independent of k. ∎

Appendix B: Single queue with vacation

Consider a single queue in which the server takes vacation, when it finds the queue empty (exhaustive service policy). And consider the arrival rate λ and service rates μ which are large, converge towards infinity while their ratio $\rho = \lambda/\mu$ is fixed. We assume IID vacation/switching times $\{S_i\}$ with mean $s = E[S_1]$.

It is well known that this system is stable when $\rho < 1$ and we assume the same. The server leaves the queue only when it is empty. It returns to the queue after a random vacation time S and would find $\Lambda^\lambda(S)$ number of customers waiting for service. Let M^μ represents the stationary visit time or the time spent in the queue in any cycle. Then,

$$\begin{aligned} M^\mu &= \inf_{t\geq 0}\{Q^\mu(t) \leq 0\} \text{ where} && (32)\\ Q^\mu(t) &:= \Lambda^\lambda(S) + \Lambda^\lambda(t) - \Gamma^\mu(t). && (33)\end{aligned}$$

In the above $Q^\mu(t)$ is stochastically equivalent to the number of customers in the system (under stationarity) at t units of time, after the server has reached the queue. We will represent the number of customers at time t by $N^\mu(t)$. So, under stationarity,

$$Q^\mu(t) \stackrel{d}{=} N^\mu(\tau_k^\mu + t) \quad \text{for all } k,$$

where τ_k^μ represents the visit epoch of the server to the queue during the k-th cycle.

By Theorem 1, we have:

$$\frac{\Lambda^\lambda(t)}{\lambda t} \to 1\, a.s., \text{ and } \frac{\Gamma^\mu(t)}{\mu t} \to 1\, a.s.$$

Using conditional expectation we have

$$\begin{aligned} Prob(\Lambda^\lambda(S) = k) &= \int Prob(\Lambda^\lambda(s) = k) dP_S(s)\\ &= \int Prob(\Lambda^1(\lambda s) = k) dP_S(s)\\ &= Prob(\Lambda^1(\lambda S) = k) \quad \text{for all } k.\end{aligned}$$

Hence, proceeding as in Theorem 1 and additionally using bounded convergence theorem we will have:

$$\frac{\Lambda^\lambda(S)}{\lambda S} \to 1\, a.s.,$$

Thus from (33) we have the following for all t:

$$\begin{aligned} \frac{Q^\mu(t)}{\mu} &= \rho S \frac{\Lambda^\lambda(S)}{\lambda S} + \rho \frac{\Lambda^\lambda(t)}{\lambda} - \frac{\Gamma^\mu(t)}{\mu}\\ &\to q(t) := \rho s + \rho t - t \text{ a.s., as } \mu \to \infty,\end{aligned}$$

if *the switching times converge towards a deterministic switching time in the following manner:*

$$S^\mu \to s \textit{ a.s., as } \mu \to \infty.$$

We have the following theorem under these assumptions:

Lemma 10. *Assume $Prob(A_i^1 > 0) = 1$ for all i. Then the visit time*

$$\begin{aligned} M^\mu &= \int_0^\infty 1_{\{g^\mu(t)=0\}} dt \textit{ with} && (34)\\ g^\mu(t) &:= \int_0^t 1_{\{Q^\mu(v)\leq 0\}} dv.\end{aligned}$$

Proof: Clearly g^μ is a monotone function. And so the RHS of equation (34) equals $\tau_g := \inf_t\{g^\mu(t) > 0\}$ and it suffices to prove that this infimum equals the infimum of equation (32), i.e., that $\tau_g = M^\mu$. Let

$$B = \{\omega : A_i^1(\omega) > 0, \text{ for all } i\}.$$

By hypothesis $Prob(B) = 1$. It is clear to see that M^μ equals the first time that the queue length Q^μ becomes zero due to a departure. For any $\omega \in B$, there exists a small interval $(M^\mu, M^\mu + \epsilon]$ during which there is no arrival. Hence, $Q^\mu(t) \leq 0$ for all $t \in (M^\mu, M^\mu + \epsilon]$ and so, $g^\mu(t) > 0$ for all $t > M^\mu$. Thus $\tau_g = M^\mu$. ∎

By our previous arguments, $Q^\mu(v) \to q(v)$ for all v and almost surely. So, by bounded convergence theorem (applied w.r.t. to Lebesgue measure on $[0,t]$ interval) we have the following convergence, for any t, as $\mu \to \infty$

$$g^\mu(t) = \int_0^t 1_{\{Q^\mu(v)\leq 0\}} dv \to \int_0^t 1_{\{q(v)\leq 0\}} dv =: g(t).$$

By Lemma 10, we have for any $T > 0$

$$M^\mu = \sum_{k=0}^\infty \int_{kT}^{(k+1)T} 1_{\{g^\mu(t)=0\}} dt.$$

Every term inside the summation converges by bounded convergence theorem and then the series converges by Monotone convergence theorem to the following limit:

$$\begin{aligned} M^\mu &\to \sum_{k=0}^\infty \int_{kT}^{(k+1)T} 1_{\{g(t)=0\}} dt\\ &= \int_0^\infty 1_{\{g(t)=0\}} dt = \frac{s\rho}{1-\rho} := \bar{m}\end{aligned}$$

The last equation is obtained using the definition of $g(.)$.

Recall $N^\mu(t)$ represents the (random) number of waiting customers in the system at time t. Then the time average number of customers in the system equals

$$\bar{n}^\mu = \lim_{t\to\infty} \frac{1}{t}\int_0^t N^\mu(t) dt.$$

By renewal reward theorem (the queue exist times of the server represent the renewal epochs) this limit is a constant almost surely. Recall τ_k^μ represent the k-th time the server has reached the queue and let M_k^μ represents the time spent in queue in k-th cycle. $\tau_k^\mu + M_k^\mu$ represent the time epoch at which it leaves the queue after the k-th visit. When the system reaches stability, it is easy to see that each M_k^μ is stochastically equivalent to M^μ and that the trajectory of the number of customers N^μ during the time interval $[\tau_k^\mu, \tau_k^\mu + M_k^\mu]$ is stochastically equivalent to that of Q^μ during the interval $[0, M^\mu]$. And this repeats periodically. One can apply similar arguments for the vacation period, $[\tau_k^\mu + M_k^\mu, \tau_{k+1}^\mu]$. Thus by renewal reward theorem:

$$\begin{aligned}\bar{n}^\mu &= \frac{1}{E[M^\mu + S]}\left(E\left[\int_0^{M^\mu} Q^\mu(t)dt\right]\right.\\ &\quad \left. + E\left[\int_{M^\mu}^{M^\mu+S}\left(A^\lambda(M^\mu+t) - A^\lambda(M^\mu)\right)dt\right]\right)\\ &= \frac{1}{E[M^\mu]+s}\left(E\left[\int_0^\infty 1_{\{t\le M^\mu\}}Q^\mu(t)dt\right]\right.\\ &\quad \left. + E\left[\int_0^\infty 1_{\{0\le t\le S\}}A^\lambda(t)dt\right]\right).\end{aligned}$$

In all the expectations given above, we have already shown almost sure convergence, as $\mu \to \infty$, of the integrands, for example $Q^\mu(t)/\mu \to q(t)$ a.s., for all t. Now the integrals inside the expectations converge to that corresponding to the limit trajectories using similar arguments as above (using bounded convergence theorem and monotone convergence theroem), i.e., for example:

$$\begin{aligned}\int_0^\infty 1_{\{t\le M^\mu\}}\frac{Q^\mu(t)}{\mu}dt \;\to\; \int_0^\infty 1_{\{t\le \bar{m}\}}q(t)dt = \int_0^{\bar{m}} q(t)dt\\ = \rho s\bar{m} - \frac{(1-\rho)\bar{m}^2}{2} = \frac{s^2\rho^2}{2(1-\rho)}\end{aligned}$$

Thus as $\mu \to \infty$

$$\frac{\bar{n}^\mu}{\mu} \to \frac{s\rho^2}{2} + \rho\frac{s(1-\rho)}{2} = \frac{\rho s}{2}.$$

Thus by Little's law the average waiting time of the random system

$$\bar{w}^\mu = \frac{\bar{n}^\mu}{\lambda} = \frac{\bar{n}^\mu}{\rho\mu} \to \frac{s}{2} = \bar{w}^{SSFM}, \tag{35}$$

which equals the average waiting time of the single fluid queue with vacation, which is derived below as (36). ■

A. Steady state fluid model for single queue with vacation

Consider steady state fluid model for single server queue with vacation where server goes to vacation after serving all the customers waiting in queue (exhaustive policy). Here s is the deterministic vacation time. SSFM is as shown in figure 5. In steady state, level of fluid in the beginning of cycle will be same as the level of fluid at the end of cycle, i.e., $k_1 = k_2 = \lambda s$. Total volume of fluid in one cycle is $\frac{k_1 m_1}{2} + \frac{k_2 s}{2}$. Note that $k_1 = k_2 = \lambda s$ and cycle length is $m_1 + s$.

$$\text{Average level of fluid} = \frac{\lambda s}{2}.$$

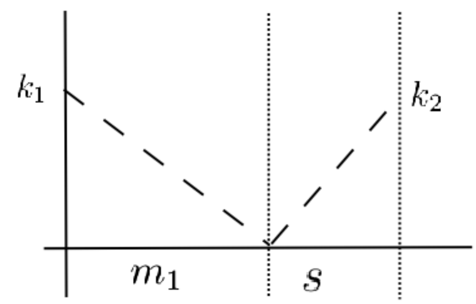

Fig. 5: Fluid level of deterministic system in one cycle for single class polling system

On using Little's law, we have

$$\text{Average waiting time} = \bar{w}^{SSFM} = \frac{s}{2}. \tag{36}$$

B. Single queue with vacations and with discrete arrivals

We now obtain the performance of single queue polling system with Poisson arrivals. This result is well known and we are reproducing the same to illustrate via an example the convergence given by (35).

Consider the pseudo conservation law for N-queue polling system as derived in [3]:

$$\begin{aligned}\sum_{i=1}^N \rho_i E(W_i) = \rho\frac{\sum_{i=1}^N \lambda_i\beta_i^{(2)}}{2(1-\rho)} + \rho\frac{s^{(2)}}{2s} + \frac{s}{2(1-\rho)}\left[\rho^2 - \sum_{i=1}^N \rho_i^2\right]\\ + \sum_{j=1}^N E(M_j^{(1)}).\end{aligned} \tag{37}$$

Here $\beta_i^{(2)}$ is the second moment of service time of class i. s and $s^{(2)}$ are first and second moment of switching time. $E(M_i^{(1)})$ is the mean amount of work in queue i at the departure epoch of server from queue i. When the switching times $S^\mu \to s$ a.s., $s^{(2)} \to s^2$ and $E[S] \to s$ (under the assumption that S^μ, for all μ, are bounded above by an integrable function). Note that in case of single queue with vacation $N = 1$, $E(M_i^{(1)}) = 0$, $\beta_i^{(2)} = 1/\mu^2$. On using equation (37) for single queue with vacation, we get

$$E(W) = \bar{w}^\mu = \frac{\rho}{2\mu(1-\rho)} + \frac{s^2}{2s}.$$

When $\mu \to \infty$ while keeping ρ fixed, $E(W) \to s/2$ as derived in equation (35) of Appendix B.