International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 5 Issue 2, January-February 2021 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Resources Allocation Queue-Fairness Model of Multi-Server Petroleum Products Distribution System

Israel J. Udoh[1], Alabi Oluwadamilare I[1], Abam Ayeni O[2]

[1]Applied Mathematics & Simulation Advanced Research Centre (AMSARC),
Sheda Science & Technology Complex (SHESTCO), Garki, Abuja FCT, Nigeria
[2]Department of Mathematics and Statistics, Federal University of Lafia, Nassarawa, Nigeria

ABSTRACT

Customer classification and service prioritization policy in a multi-server single queuing system is one major scheduling policy employed by most service-oriented institutions to provide preferential treatment to customers, as well as control services in the queuing system. Jobs seniority and service requirement differences are two fundamental principles of queue fairness governing such preferential services. In this study, we invoke the RAQFM analytical framework to investigate the dominance of these parameters in determining class discrimination and system unfairness in atypical multi-class multi-server single-queuing architecture under preemptive priority service policy. In a comparative variant, we evaluate and compare the relative fairness as well as class discrimination coefficients of an alternative setup -resource dedication system, i.e. dedicating a separate server or set of server(s) to each class of customers served under a single FCFS queuing policy. Tentatively, the result of the analysis shows that granting preemptive priority service to customers' classes based on some socio-economic or geo-political considerations order than the fundamental principles of queue fairness - service requirements and job seniority differences maybe not justify. As classes of short jobs from low priority class which have arrived the system early have to wait for eternity for the completion of many classes of long jobs from the high priority classes that arrive behind them; thus, to an unfair treatment by the system and a violation of the two fundamental principles of queue fairness. Such policy did not only breed high negative discrimination as well as general system unfairness to classes with lower service requirements and higher inter-arrival time but also violate the basic principle of RAQFM. To address such conflict in a multi-class multi-server system, the study recommended the dedication of separate server(s)or set of servers to each class of customer associated with a single FCFS queuing policy as a fairer alternative if preferential services must be granted to customers on other criteria than the two fundamental principles of queue fairness.

KEYWORDS: *RAQFM, classification, prioritization, discrimination, unfairness, FCFS, multi-server, preemptive priority, service requirement, jobs seniority, socio-economic and geo-political consideration*

How to cite this paper: Israel J. Udoh | Alabi Oluwadamilare I | Abam Ayeni O "Resources Allocation Queue-Fairness Model of Multi-Server Petroleum Products Distribution System" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-5 | Issue-2, February 2021, pp.770-788, URL: www.ijtsrd.com/papers/ijtsrd38497.pdf

IJTSRD38497

1. INTRODUCTION

Queuing models have long served key roles in a wide variety of fields and applications, including human service delivery systems such as supermarkets, airports, government offices, etc., as well as computer and telecommunication systems to control the services given by them. The classification of customers/jobs based on their service requirement and job seniority as well as the prioritization of such classes is a very common queuing policy used by most service-oriented institutions to provide preferential treatment to customers as well as control the services in the queuing systems. Everyday examples of classification and preferential treatment of in queuing systems include (i) classifications of passengers as domestic and international in airport customs queues, (ii) the gender classification in public toilet queuing facilities; (iii) the classification of short-jobs and long-jobs in a supermarkets, and (iv) the prioritization of jobs in computer systems. Classification of jobs is usually done based on many characteristics. For example in computerized call centers and internet web servers, customers can be classified and prioritized by IP address range, transaction type, etc. Most common situation especially in the supermarkets and groceries stores includes short-jobs (customers with a few items in hand) receiving preferential service through special servers (cashiers) dedicated to them over long-jobs (customers with much items in cart).

A major reason for prioritizing jobs as well as giving preferential service to customers is psychologically to maintain some degree of fairness in the queuing system.

That is, the desire to make the system operation "fair". Fairness among customers/jobs is a crucial and fundamental issue in queuing systems. Fairness or social justice has always been and still is a cardinal issue in all cultures and traditions. It is the fabric holding the human societies together. In any queuing situation, emotions and resentment may flare if unfairness is practiced or is perceived as being practiced, while courtesy and even camaraderie due to same experience-sharing may result when fairness in treatment is perceived. A recent experimental psychology study of attitude of people in queues shows that fairness in the queuing system is very important to people, perhaps not less than the wait itself [25, 26].

Despite its fundamental role in queuing systems, fairness in queues has hardly been studied and appraised empirically. In particular, the issue of how prioritization of jobs as well as granting preferential services to some class of customers/jobs affect fairness in queues has not been practically evaluated in a quantitative manner and thus not fully understood. Our focus in the present study is to quantitatively examine the effect of *preemptive priority* and *resource dedication* police son the fairness factors of a multi-class multi-servers single-queue system using the Resources Allocation Queuing Fairness (RAQF) model proposed in literature by Raz et al[28,29].In order to focus on the pure fairness and pure queuing properties, RAQFM analytical framework considered systems where job classification is based only on service characteristics. However, in practical systems customer/job classification have other attributes such as different socio-economic or geo-political values, which is the scope of this study.

As observed earlier, two common queuing architectures used to grant preferential services to customers/jobs are: (i) the class *prioritization*; in which some classes of customers/jobs are categorized as high or low priority class, and priorities such as preemptive or non-preemptive priority is given to customers belonging to higher priority classes over those belonging to lower priority classes, and (ii) *Resource dedication*; in which each class has a dedicated server or a set of servers, and a queue dedicated to it. Our objective is to evaluate the relative unfairness and class discrimination index of these two architectures in respect of a practical service oriented queuing system – the Petroleum Products and Pipelines Marketing Company (PPMC), Suleija, Niger state Nigeria. And by comparative analyses, determines the effect of preemptive priority policy on the multi-class multi-server single-queue system. Such analysis will provide measures of fairness for the system under study that can be used to quantitatively account for fairness when considering alternative designs. The quantitative approach would enhance the existing design approaches in which efficiency (e.g., utilization and delays) is accounted for quantitatively, while fairness is accounted for only in a qualitative way.

The PPMC, one of the subsidiary companies of the Nigerian National Petroleum Corporation (NNPC) was established in 1988 with the noble objective of providing excellent customer services by transporting crude oil to the refineries and moving white petroleum products to the existing and future markets efficiently and at low cost through a safe and well maintained network of pipelines and depots. It is also part of the PPMC's objectives to profitably and efficiently market refined petroleum products in the domestic as well as export markets especially in the ECOWAS sub-region, provide marine services and also maintain uninterrupted movement of refined petroleum products from the local refineries. In its determination and drive to deliver on its primary objective and thus meets the national level of petroleum products demand, stimulates the consumption of refined petroleum product to a significant level in the domestic energy mix, and possible optimize the sales of these products, as well as promotes their import substitution, the PPMC has strategically established and located about 29 of its depots around major cities of the nation's six geopolitical zone. Strategically, among the key service descriptions of the PPMC are:

- Marketing of refined petroleum and petrochemical products in the domestic as well as export markets, marine services and efficient evacuation of refined petroleum products from the local refineries.
- Construct, maintain, lease and hire pump stations, depots and pipelines for the storage and transportation of petroleum and petroleum products, liquids and gases; transport such liquids and gases by means of such pipelines and to utilize, sell and supply liquids and gases to others, to store the same in tanks or otherwise and to lay, buy, lease, sell and operate such pipelines, tanks and other storage facilities.
- Carry on the transportation by ships of crude oil, petroleum products, and petrochemical products and to perform other activities relating thereto.
- Collect and evacuate crude oil, refined oil, liquefied natural gas, liquefied petroleum gas or other allied products from various oil fields, oil terminals or refineries in the Federal Republic of Nigeria or elsewhere and to transport such to anywhere in Nigeria or elsewhere in the World.

Significantly, among the major petroleum products and capacity under the storage and marketing purviews of the PPMC depots are: (i) Premium Motor Spirit (PMS) - 994,500 cubic metres, (ii) Dual purpose Kerosene (DPK) - 430,700 cubic metres, (iii) Automotive Gas Oil (AGO) - 673,100 cubic metres, (iv) Aviation Turbine Kerosene (ATK) - 74,000 cubic metres, (v) Liquefied Petroleum Gas (LPG), (iv) Low Pour Fuel Oil (LPFO), and (vii) High Pour Fuel Oil (HPFO). Other specialized bye products also include (i) Paraffin Wax, (ii) Base oil (iii) Bitumen, (iv) Asphalt, and (v) Carbon Black [11,12].

1.1. Statement of the Problem

Over the years, especially during festive season, the management of PPMC has been meted with severe criticisms for lacking behind or failing from its primary responsibility as a result of the perceived unfair allocation and distribution of the petroleum products to its respective customers' classes; especially when unending queues of motorists and commuters spend days or weeks on major streets, highways, and immediate neighborhoods of petroleum filling stations waiting hopelessly to purchase the various petroleum products, yet to arrive from the various PPMC depots within the country. Even around the vicinities and extended neighborhoods of most PPMC depots, unending queues of long tanker vehicles spend days to weeks waiting to lift the respective

petroleum products to their respective customers' destinations.

Premised on this often perennial scarcity of the products occasioned by either delay in arrival or insufficient allocation of the products to the respective customers' classes, the present study seek to examine the viability of the queue architecture employed by the PPMC in its petroleum products distribution system, as well as the fairness characteristics associated with such system. Thus, in terms of equitable distribution and allocation of the premium motor spirit (PMS) sub-products to its catchments areas as well as to the queuing customers, ***"how fair is the customer classification and service prioritization policy of the PPMC system with respect to its multi-classes of customers?***

To carry out the analysis, the Section 2.0 briefly reviewed some related literature on the subject matter of queue fairness, customer/job classification and service prioritization policy, brief conceptual framework and policy justification for applying the RAQFM model to evaluate the performance measure of the PPMC queuing architecture, a briefly reviewed the RAQFM analytical framework and its basic properties as well as its applicability to multi-class multi-server single-queue system. A practical application of RAQFM analytical framework in a typical queuing system of a service oriented institution – the PPMC is carried out in Section 3.0. As an alternative to the perceived unfair PPMC system, RAQFM variant of resource dedication is also consider to comparing the discrimination and unfairness characteristics of the two systems. Finally, in Section 4.0, we present and discuss the result of the analysis, conclusion and some recommendations.

2. Relevant Academic Literature

Customer classification and service prioritization based on service requirement (job size) and seniority difference (arrival time) are common queuing policies used in many applications to provide preferential services to customers/jobs in the queuing system. For instance in computer systems, a handful of literatures abound on the use of single-server to serve multi-class jobs under priority service discipline - where the jobs' classes may differ from each other in their service requirements and arrival rates. The service policies of such systems are often time class-prioritization, namely, high priority jobs are always served ahead of lower priority jobs. However, analysis of such systems has always focused on evaluating the system performance based on the expected waiting time, or the mean waiting cost under linear cost parameters[2],[10],[32],[34]. Optimization of these systems with non-preemptive priorities based on these performance objectives shows that the optimal scheduling policy is to provide a higher priority to jobs with smaller mean service requirements [2],[10].Research have shown that such priority service may, however, result in long jobs (larger mean service requirements) waiting for the completion of many short jobs who arrive behind them, and thus, possibly, to unfair treatment by the system. Therefore system operation that accounts for both efficiency and fairness might have to resort to a different scheduling policy[7].

The fairness factors associated with waiting in queues has been recognized in many works and applications. For example Kirill[15] and *Michael and* Garrett*[21]*in their discussion papers on the psychology of waiting on the queue recognizes the central role played by "Social Justice", (another name for fairness), and its perception by customers. Rothkopf and Rech[30]had also addressed the concept of fairness in their paper discussing customers' perceptions in queues; where an impressive list of quantifiable variables shows that contrary to the common belief, combining queues may not be economically advantageous. At the end of their analysis, the authors conceded however, that all these quantifiable variables may not have sufficient weight to overcome the unfairness perceived by customers served in a separate queues structure.

Aspects of fairness in queues were also qualitatively appraised earlier by Alex Stone[3] who judged the annoyance caused by congestion as torture, PerryKuklin[24] who appraised the impact of waiting times on customer behavior, while Jenny and Refael[13] and Woo-SungKim and Dae-EunLim[35]addressed overtaking especially in queues. Scientific evidence of the importance of queue fairness was also provided by Rafaeli et al[25,26], where the reaction of humans waiting in queues with various service scheduling policies was studied using experimental psychology approach. The studies revealed that for humans waiting in queues, the issue of fairness is highly important, sometimes even more important than the duration of the wait (delay probability). For the case of multi-server single-queue versus a separate queue at each server (resource dedication), the authors observed that the former was perceived as more fair than the later. Probably for this may not be unconnected with the reason we find separate queues mostly in systems where a common queue is physically not possible, such as traffic toll booths and supermarkets. Recently, the issue of fairness was also discussed in the context of practical computer applications and web servers by Harchol-Balter et al[29]where a queuing policy was shown to have reduced response times, but at the expense of unfairness to large jobs.

Considering the apparent importance of queue fairness, there is very little published work providing reliable quantitative results on job fairness in priority queues. Therefore, the present study, a review of the RAQFM analytical model seek to practically appraise the unfairness and discrimination characteristics associated with multi-server single-queue architectures in the present of a multi-customer classes under preemptive priority service discipline. The choice of RAQFM for studying the PPMC queuing system may not be unconnected with the facts that RAQFM's analytical philosophy prioritize the basic principle of social justices, which demand that at any epoch *"equally needy members of a group should share equally the resources (the pie) available to the group"*[22],[23].Significantly, RAQFM analytical framework does not only tracks customers/jobs inter-relationship and the resulting unfairness throughout the queuing progress process, but is also sensitive to the fundamental principles of queue fairness - the principle of jobs seniority (arrival time) and service requirement (jobs size) differences inherent in every priority queue. Therefore, using RAQFM analytical frame-work to appraise the performance measure (discrimination and

fairness characteristics) of a multi-class multi-server queuing system is expedient and justified.

As an analytical model, the present study seek to provide a measure (metric) of fairness for the system under studied that can be used to account quantitatively for queue fairness when considering alternative configuration. Our quantitative appraisal will also enhance the existing queue design, in which efficiency (e.g. utilization and delay) is usually accounted for quantitatively, while fairness only attract qualitative appraisal. Given the veracity of our analysis, the study will provide the practitioners with useful tools by which they can evaluate the fairness of a variety of very common operational strategies. It will also serve as a useful reference and conceptual framework for all types of end users such as scholars, researchers, teachers, command levels, data processing programmers and functional support non-programming end users and for management and decision support system.

2.1. The Concept of Queuing Fairness

What is fairness in reality? Although almost every child, if asked, can tell what is fair and what isn't fair. However, to arrive at a commonly agreed definition for fairness is quite a demanding undertaking, much more so when it comes to defining a quantitative measure of the level of fairness. Therefore, to psychologically comprehend the concept of queue fairness, our approach in this work is to consider the queuing system as a microcosm social construct, whose fairness properties should therefore conform to the general cultural perception of social justice in a particular society [22],[23].The issue of fairness and social justice has always been, and is still a cardinal issue in all cultures, as it is the cement holding the society together. Thus issue of fairness or social justice has been a subject of debate by philosophers, prophets and spiritual leaders since the beginning of recorded history.

Perhaps one of the first formulations of fairness issue is Aristotle's idea in "Nicomachean Ethics" which states that *"justice consists, at least in part, in treating equal cases equally and unequal cases in proportional manner… also to implement equality between the persons and the shares, the ratio between the shares must be the same as that between the persons"*[4]. In modern time, the controversies surrounding fairness have attracted inputs from philosophers, economists, social and behavioral scientists, and as is to be expected, large volume of researches and publications of reviews and interpretation of justice as well as fairness doctrine abound in literature[1],[5],[7]. However, a most prominent and comprehensive publication on fairness issue date back toJ. Rawl's[27] "Theory of social justice", whose general conception stipulated that:*"All social primary goods such as liberty and opportunity, income and wealth, and the bases for self-respect, should be distributed equally unless an unequal distribution of any or all of these goods is to the advantage of the least favored".*

Central to the social justice debate is the concept of fair allocation of resources. That is fairness is conceivable when the resource (pie) is appropriately divided between the contending consumers or jobs. But what is the pie in the case of a queuing system, and how should it be divided fairly? A very similar concept in the area of computer network is that of Processor Sharing (PS), as embodied by the ideal PS policy, analyzed in[6],[14],[16,17],[31].The PS or egalitarian PS is a service policy where the customers, clients or jobs are all served simultaneously, each receiving an equal fraction of the service capacity available. The generalized PS is a multi-class adaptation of the policy which shares service capacity according to positive weight factors to all non-empty job classes at the node, irrespective of the number of jobs of each class present. Often it is assumed that the jobs within a class form a queue and that queue is served on a FCFS basis. The idea rooted in PS and of course any air service policy is that *"at every moment of time, the servers' service rate should be divided equally amongst the jobs or customers present in the system, and a violation of this policy connotes unfairness to the job or customers being serve".* This is perhaps, the idea behind RAQFMs' basic principle which states that *"at every epoch all jobs present in the system deserve an equal share of the system's service rate or servers' resources and a deviation from it create discriminations (positive or negative)",* and accounting for these discriminations with summary statistics give yields a measure of unfairness[28,29].

2.2. Fundamental Issues in Queue Fairness Measures:

According to the proponents of queue fairness measures [1],[5],[8],[18],[25],[28,29], two fundamental principles determines queuing fairness process and job scheduling policies. These are (i.) Job seniority differences - the arrival time of a customer/job and (ii.) the service requirement differences i.e. service times of the customer (job size). To analytically represent these concepts, consider a typical queuing setup consisting of a server$m$ and customers $C_i, (i = 1,2, \dots,)$ arriving at the system at arbitrary arrival epochs$a_i, (i - 1,2, \dots,)$ respectively, such that$a_i \leq a_{i+1}$. Suppose customer $C_i$requests some amount of service s_i at the server, $(s_i, measure\ in\ units\ of\ time)$ through some scheduling policy. Once $C_i$receives its full amount of service s_i(which does not necessarily have to be given continuously or at full rate) it departs the system at an epoch d_i. The duration $C_i$stays in the system,$t_i = d_i - a_i$ is called the system time. The duration $C_i$waits and does not get service $\omega_i = t_i - s_i = (d_i - a_i) - s_i$ is call the waiting time of C_i, except for PS disciplines, where this conventional definition of waiting time may not be applicable. These notations $a_i, s_i, d_i, t_i, \omega_i,$ denote the actual values attributed to $C_i$in a specific sample path of the queuing system.The seniority of C_i at epoch t is $t_i - a_i$, and the service requirement of C_i, at epoch t is s_i. It is natural to expect that a "fair" scheduling policy will give preferential service to highly senior customers (larger arrival time) and to customer with smaller service-requirement (small job size). Thus, the fundamental principles of job seniority and service requirements state:

2.2.1. Service-Requirement Preference Principle:

If all customers in the system have the same arrival time, then for customer $C_i$and $C_j$arriving at the same time and residing concurrently in the system, if $s_i < s_j$, then it will be more fair to complete service of $C_i$ahead of $C_j$than vice versa. The service-requirement preference principle is rooted in the belief that it is "less fair" to have short jobs wait for long ones.

2.2.2. Jobs Seniority Preference Principle:

If all jobs in the system have the same service times, then for customer $C_i$and $C_j$residing concurrently in the system,

if $a_i < a_j$, then it will be more fair to complete service of $C_i$ahead of $C_j$than vice versa. The jobs seniority preference principle is rooted in the common belief that jobs arriving at the system earlier "deserve" to leave it earlier. It should be noted that when $a_i < a_j$ but $s_i > s_j$ or $a_i = a_j$ and $s_i = s_j$ the two principles may conflict each other; and thus the relative fairness of the possible scheduling of customer $C_i$and $C_j$is likely to depends on the relative values of the parameters. These two preference principles can be considered as two axioms expressing one's basic belief in queue fairness, hence a fairness measure is said to follow either of the two preference principles if it associates higher fairness values with schedules that are more fair:

Axiom-2.0 - *Jobs Seniority Preference:* Consider customer $C_i$and $C_j$requiring equal service times and obeying $a_i < a_j$. Let πbe a scheduling policy where the service of $C_i$is completed before that of $C_j$and π' be identical to π, except for exchanging the service schedule of $C_i$and C_j.A fairness measure is said to adhere to the *seniority preference* principle if the fairness value it associates with π is higher than that it associates with π'.

xiom-2.1 - *Service-requirement Preference*: Consider jobs customer $C_i$and C_j, arriving the same time at the system and obeying $s_i < s_j$. Let πbe a scheduling policy where the service of $C_i$is completed before that of $C_j$and π'be identical to π, except for exchanging the service schedule of $C_i$and C_j..A fairness measure is said to adhere to the service-requirement preference principle if the fairness value it associates with π is higher than that it associates with π'

2.3. The Resources Allocation Queue Fairness Measure(RAQFM)

Divergence of opinions, thoughts and theories abound on the measurements or quantifications and analysis of queue fairness as well as the effect of customers' classification and service prioritization. While a host of these studies and measurements are predicated on the delay distribution perspective - the average time spent by a tagged customer in the queue system[1], [5], [8], [25], [16,17], a few are poised to consider queue fairness in terms of whether the actual services rendered to the customers commensurate with their delay probabilities. It is on such premise that the RAQFM's variant of class discrimination that accounts for the expected discrimination experienced by a tagged class-j customer was proposed by Raz et al[28,29].Thus by RAQFM'sprinciple *"at every epoch tat which there are N(t) customers/jobs present in the system, they all are entitled to an equal share of the server's time (system resources), and any deviation from this principle represent in a discrimination (positive or negative)"*

By this conceptual framework, RAQFM enjoyed not only the unique properties of being sensitivity to jobs seniority (arrival time), and service requirement (job size) differences, but also tracking the customers/jobs inter-relationship as well as the resulting unfairness throughout the queue progress process. These allow for understanding fairness at both the individual job perspective as well as the job classes' levels and not only at the level of job classes. This property also allows for the evaluation of the individual discrimination as well as the unfairness of specific scenario, at the one hand, and the overall unfairness of the system or policy, on the other hand. Thus, the RAQFM model allows accounting for individual job discrimination as well as system unfairness. This measure is therefore composed of two distinctive parts*(i)the job Discrimination*- each customer is given a single measure representing how well the customer was treated. A positive number means the customer was well treated, and a negative number means the customer was not treated well,*(ii)the system unfairness* - A summary measure taken over the discriminations. The non-negative result of this summary measure is the system's unfairness, thus, a low measure means a more fair system while a high measure connotes order wise.

2.3.1. RAQFM Analytical Model:

Consider a non-idling queuing system (i.e. a queuing system where if there are $n$customers in the system, and the system is composed of $m$independent servers, where $min\{m, n\}$ of these servers are operational) with n servers, $n = 1,2, \dots, m$. All servers have equal service rate; for simplicity, a rate of one (1) (unit of service time per unit time). The system is subject to the arrival of stream of customers $C_1, C_2, \dots$,who arrived at the system at this order. Leta_l andd_l denote the arrival and departure epochs of C_l respectively. LetS_l denote the service requirement (measured in time units) of C_l. RAQFM evaluates the unfairness in the system as follows: The basic fundamental assumption is that at each epoch, all customers present in the system deserve an equal share of the total service granted by the server at that epoch.Let $0 \le \omega(t) \le m$ denote the total service rate granted at epoch t, (which usually is an integer equaling the number of working servers at that epoch), and $N(t)$ denotes the number of customers in the system at epoch t, then the fair share; called the *momentary warranted* service rate of C_l given by

$$R_l(t) = \frac{\omega(t)}{N(t)} \quad (2.0.1)$$

Let $\sigma_l(t)$ be the momentary rate at which service is given toC_l at epoch t. This is called the momentary granted service rate of C_l. The *momentary discrimination* rate of C_l at epoch t, whenC_l is in service, denoted by $\delta_l(t)$ is, therefore, the difference between its granted service and the warranted service:

$$\delta_l(t) = \sigma_l(t) - R_l(t) = \sigma_l(t) - \frac{\omega(t)}{N(t)} = \frac{\sigma_l N - \omega}{N} \quad (2.0.2)$$

Equation (3.0.2) can be viewed as the rate at which customers' discrimination accumulates for C_l at epoch t.Let$\delta_l(t) \stackrel{\text{def}}{=} 0$, if $C_l$is not in the system at epoch t. However, as we are only interested in $\delta_i(t)$whenC_l is in the system, and for the omitted, the total discrimination ofC_l, denoted D_l is given by

$$D_l = \int_{a_l}^{d_l} \delta_l(t)dt \quad (2.0.3)$$

Definition2.0 - *Alternative Definition for Momentary Warranted Service and Discrimination*: The definition of the momentary warranted service and discrimination given above is based on the rulethat a customer deserves an equal share of the total*resources granted* $\omega(t)$by the server at that epoch and any deviation from it creates discrimination among the customers residing in the system. If some of the resources are not granted at epoch t, e.g., due to system idling, or due to the use of only part of the servers (as applicable to the system under study), it

may be considered as being *inefficient* but not as a discrimination and unfairness. In such situation one could consider an alternative concept by which at epoch t,a customer deserves an equal share of all the available m-servers' resources, thus the warranted service rate is given by

$$R_l(t) = \frac{m}{N(t)} \quad (2.1.1)$$

And the momentary discrimination$\delta_l(t)$will be replaced by,

$$\delta_l(t) = \sigma_l(t) - \frac{m}{N(t)} = \frac{\sigma_l N - m}{N} \quad (2.1.2)$$

The difference between equation (2.0.1) and (2.1.1) is conceptual and relates to situations where the system does not grant all of its resources. One such case is a multi-server system at epochs where the number of customers is smaller than the number of servers, $N(t) < m$. Another case is a system which allows server idling (when there are customers in the system). This issue and the tradeoff between the alternative is more pronounced in multi-server multi-queue systems. For the present work we choose to focus on the concept of fair division of the granted resources (equations2.0.1); this might be appealing since the cases in this work where the system does not grant all resources are limited to situations resulting from system operations constraints (system cannot serve a single customer by many servers), and thus may possibly be interpreted by customers as non-discriminatory. For work conserving systems (systems in which the total service given to a customer over time equals its service requirement:

$$\int_{a_l}^{\infty} \sigma_l(t)dt = s_l(t) \quad (2.1.3)$$

We have from equations (2.0.2) and (2.0.3), that

$$D_l = s_l(t) - \int_{a_l}^{d_l} \left(\frac{\omega(t)}{N(t)}\right) dt = \int_{a_l}^{\infty} \sigma_l(t)dt - \int_{a_l}^{d_l} \left(\frac{\omega(t)}{N(t)}\right) dt \quad (2.1.4)$$

A positive or negative value ofD_l means that a customer received better or worse treatment than it fairly deserves, and therefore it is positively or negatively discriminated. For a single server, work conserving and non-idling system, the expected value of discrimination always obey $E[D] = 0$, thus, the unfairness of the system can be taken as the second moment of discrimination, $E[D]^2$. The same property also holds in multi-server non-idling system.

2.3.2. RAQFM Model of Multi-Server Single-Queue System:

Consider a work conserving non-idling system with m servers, (M/M/m), and u classes of customers; where class-j arrivals follow a Poisson process with rateλ_j, and their required service times are identically independent random variable $(i.i.d)$ exponentially distributed with mean$\mu_j^{-1}, j = 1,2, \dots, u$. The total arrival rate is given by

$$\lambda \stackrel{\text{def}}{=} \sum_{j=1}^{u} \lambda_j \quad (2.1.5)$$

And for stability, it is assumed that the traffic intensity is given by

$$\rho \stackrel{\text{def}}{=} \sum_{j=1}^{u} \frac{\lambda_j}{m\mu_j} < m \quad (2.1.6)$$

Due to the Markovian nature of the system, and for mathematical convenience, we consider the time as being slotted where the sequence of arrivals and departure formed the slot boundary. Let$T_i, i = 0,1,2 \dots$, be the duration of the i^{th} slot. Here the analysis is limit to systems where a service decision is made only on arrival and departure epoch. Thus the numbers of available servers, the number of servers actually giving service and the rate at which service is given to each customer are constant during each slot. Let$0 \le \omega_i \le m$ be the numbers of working servers in the i^{th} slot, $\sigma_{i,j}$ be the rate at which service is given to C_l at the i^{th}slot and N_i as the number of customers in the system during the i^{th} slot. Let $\delta_{i,j}$ be the momentary discrimination of C_l during the i^{th}slot, which is the rate at which customers discrimination accumulate for C_l at this slot.By equation (2.0.2):

$$\delta_{i,l} = \sigma_{i,j} - \frac{\omega_i}{N_i} = \frac{N_i\sigma_{i,j} - \omega_i}{N_i} \quad (2.1.7)$$

Let a_l, d_l denote the indexes of the arrival and departure slot of C_l respectively (C_l arrives at the beginning of the $a_{l^{th}}$ slot and departs at the end of the$d_{l^{th}}$).Then the total discrimination accumulated for C_l during thei^{th} slot is $C_{i,l}T_i$. Thus, the slotted version of equation (2.0.3)becomes:

$$D_l = \sum_{i=a_l}^{d_l} \delta_{i,l}T_i \quad (2.1.8)$$

And for work conserving systems the slotted version of equation (2.0.3) becomes:

$$D_l = s_l - \sum_{i=a_l}^{d_l} \frac{\omega_i T_i}{N_i} \quad (2.1.9)$$

2.3.3. Momentary Discrimination:

Let C be an arbitrary tagged customer of class-j, and let $a = 0, m, m+1, \dots$,denote the number of customers ahead of Cin the queue, including served customers and b, the numbers of customers behind C. If $C$is in service, then $a = 0$, and $b \in \mathbb{N}^0$includes also customers served by other servers. Note (unavoidable) jump occurring in the value of b when C enters service, and the fact that the values $1 \le a < m$ are invalid. Due to the Markovian and the non-idling nature of the system, the state (a, b)captures all that is needed to predict the future of C.The number of customers in the system at a slot where Cobserves the state $(a, b)$is $\mathbb{N}(a, b) = a + b + 1$. The rate of service given to Cat that slot is $1(a = 0)$and thetotal service rate is a constant $\omega(a, b) = \sigma(a, b)$, where $\sigma(a, b) = min(m, a + b + 1)$ is thenumber of active servers. The rates of arrival and departure are$\mu(a, b) = \sigma(a, b)\mu$and $\lambda(a, b) = \lambda$. The momentary discrimination at state(a, b), denoted by $\delta(a, b)$is given by

$$\delta(a, b) = 1(a = 0) - \frac{\sigma(a,b)}{(a+b+1)} \quad (2.2.0)$$

2.3.4. The Systems Unfairness:

Let $E[D^2|k]$for $k = 0,1,2, \dots$denote the expected value of the square of discrimination, given that customer Cencounters k customers on arrival (including the ones being served). LetP_k be the steady state probability that there are $k$customers in the system. Thus, the second moment of D (the unfairness) follows:

$$E[D^2] = \sum_{k=0}^{\infty} E[D^2|k]P_k \quad (2.2.1)$$

WhereP_k, the steady state probabilities for the single queue M/M/m system and given by:

$$P_k = \begin{cases} P_0 \frac{[m\rho]^k}{k!}; k < m \\ P_0 \frac{\rho^k m^m}{m!}; k \geq m \end{cases} \text{ and } P_0 = \left[\frac{[m\rho]^m}{m![1-\rho]} + \sum_{k=0}^{m-1} \frac{[m\rho]^k}{k!}\right]^{-1}$$

(2.2.2)

Where $m$denote the number of parallel server,$k$denotethe number of customer's class in the system and $p_0$denote the probability of an empty system. Let $D(a,b)$be a random variable denoting the discrimination experienced by a customer, through a walk starting at (a,b), and ending at its departure. Then

$$E[D^2|k] = \begin{cases} E[D^2(k,0)];\ k \geq m \\ E[D^2(0,k)];\ k \leq m \end{cases} \quad (2.2.3)$$

Let $d^{(1)}(a,b)$ and $d^{(2)}(a,b)$be the first two moments of $D(a,b)$. In the single queue M/M/m system, the slot lengths are exponentially distributed with parameter $\lambda + s\mu$ and the first two moments are:

$$\left.\begin{aligned} t^{(1)}(a,b) &= [\lambda + \sigma\mu]^{-1}; and \\ t^{(2)}(a,b) &= 2[\lambda + \sigma\mu]^{-2} = 2\left[t^{(1)}\right]^2 \end{aligned}\right\} \quad (2.2.4)$$

Let$\tilde{\lambda}(a,b)$denote the probability that the slot will end with a customer arrival event and$\tilde{\mu}(a,b)$, the probability that the slot will end with a customer departure from a specific active server. Then

$$\left.\begin{aligned} \tilde{\lambda}(a,b) &= \lambda[\lambda + \sigma(a,b)\mu]^{-1}; and \\ \tilde{\mu}(a,b) &= \mu[\lambda + \sigma(a,b)\mu]^{-1} \end{aligned}\right\} \quad (2.2.5)$$

Note that throughout this analysis $\tilde{\lambda}$refers to the probability of an arrival of *any* customer, while $\tilde{\mu}$refers to the probability of a departure of a customer from *one specific* queue. This seeming inconsistency is required for mathematical brevity. Assume C is in state (a,b)at the slot's end; the system will encounter one of the following events and $C's$ state will change accordingly:

A. A customer arrives to the system: The probability of this even is$\tilde{\lambda}(a,b) \stackrel{\text{def}}{=} \lambda[\lambda + \sigma(a,b)\mu]^{-1}$.$C's$ state changes to$(a, b+1)$.

B. For $a > 0$:A customer leaves the system. The probability of this event is $m\tilde{\mu}(a,b)$, where $\tilde{\mu}(a,b) = \mu[\lambda + \sigma(a,b)\mu]^{-1}$.If $a > m$; C's state changes to $(a-1, b)$. Otherwise $(a = m)$C's state changes to $(0, a+b-1)$.

C. For $a = 0, and\ b > 0$:A customer other than C leaves the system. The probability of this event is $(\sigma(a,b) - 1)\tilde{\mu}(a,b)$. C's state changes to $(a, b-1)$, and

D. For $a = 0$:C leaves the system. The probability of this even is$\tilde{\mu}(a,b)$.

2.4. Class Prioritization in Queuing System

According to Raz et al[28,29] and Hanoch et al[8], a common way to give preferential service to customers is to prioritized services to customers' classes based on their service characteristics. This can be done in various ways such as assigning each class of customers to a special queue or serving each class of customer ahead of the others in the queues. The quantitative analysis of RAQFM variant of class discrimination, accounts for the expected discrimination experienced by the customers of a certain class under the *Preemptive Priority* and the *Preemptive Resume* service policy. By *Preemptive Priority* class of scheduling policies, we mean a policy in which the server always serves the customer with the highest priority present in the system. If a higher priority customer arrives, and finds a lower priority customer in service, the served customer is displaced by the arriving customer. In the *Preemptive Resume* variant, the preempted customer returns to the head of the queue of its class, and resumes its service from the point it was interrupted, upon reentering service. The order of service within each class of customers is usually FCFS.

One practical way to implement the *preemptive resume* service policy is to assign each class of customers to its own service queue. The server always serves each nonempty queue using the FCFS service policy, until the queue is empty. Our interest in this work is to evaluate the *fairness factor of the preemptive resume service policy on a multi-server single-queue system, and compare it to the alternative of policy of assigning each class of customers to its own separate server with a FCFS queue policy.* As we wish to focus on the pure fairness issues inherent in the system, our discussion is limited to systems where customer/job classification is based only on service characteristics.

2.4.1. RAQFM Model of Class Prioritization:

Suppose the system is subject to the arrival of stream of customers $C_j, (j = 1, 2, \dots, u)$with eachbelongs to one of $u$classes. By equation (2.1.5), the arrival rate of class $u$customers: $\lambda = \sum_{u=1}^{U} \lambda_u$ .An order of priorities is assigned to theclasses, where lower class index means higher priority.In the *Preemptive Priority* class of scheduling policies, the order of service within each class of customers is FCFS, and preempted customers return to the head of the queue of their class. For a class $u$the discrimination Dexperienced by an arbitrary customer C, when thesystem is in steady state, is a random variable denoted $D_{(u)} \stackrel{\text{def}}{=} D|C \in u$. Our interest is the expected discrimination experienced by $u's$ customers, namely $E[D_{(u)}]$termed *Class Discrimination*. A second useful notion is that of *class discrimination rate*. The *instantaneous discrimination rate* of class $u$at time $t$is the sum of discriminations over all $u's$ customerspresent in the system at time t, and given by

$$\widetilde{D}_{(u)}(t) \stackrel{\text{def}}{=} \sum_{l \in u} \delta_l(t) \quad (2.2.6)$$

Therefore the *instantaneous discrimination rate* of class u when the system is in steady state is a random variable

$$\widetilde{D}_{(u)} \stackrel{\text{def}}{=} lim_{t \to \infty} \widetilde{D}_{(u)}(t) \quad (2.2.7)$$

Taking expectation of equation (2.2.7), we get the *class discrimination rate*, $E[D_{(u)}]$. The relationship between the variables $D_{(u)}$ and $\widetilde{D}_{(u)}$ is analogous tothe equilibrium relationship between the variables *customer delay*, i.e. the delay experienced by an arbitrary customer and *number of customers in the system*, i.e. the number of customers present at an arbitrary moment, in a stationary queuing system. While $D_{(u)}$ is more appropriate to describe the customer's perception, $\widetilde{D}_{(u)}$ might be moreappropriate to describe the system's state. We therefore choose to focus on $D_{(u)}$. However, we recall from J. Little's[19] Theorem, that $N = \lambda T$, where T, λ,and N denote expected customer delay, arrival rate and expected number of customers in the system, respectively. Thus, the class "discrimination version" of Little's Theorem becomes:

$$E\left[\tilde{D}_{(u)}\right] = \lambda_u E\left[D_{(u)}\right] \quad (2.2.8)$$

Applying this rule to the whole population, rather than to a certain class, would results in $E[D] = 0$.

Theorem 2.0: *The class discrimination of class u is bounded from above by the overall system unfairness as follows:*

$$\frac{\lambda_u}{\lambda}\left|E\left[D_{(u)}\right]\right| \le \left[E[D^2]\right]^{0.5} \quad (2.2.9)$$

Proof: Considering that

$$E\left[D_{(u)}{}^2\right] - \left[E\left[D_{(u)}\right]\right]^2 = 0,$$

We have that

$$\frac{\lambda_u}{\lambda}\left|E\left[D_{(u)}\right]\right| \le \frac{\lambda_u}{\lambda}\left[E\left[D_{(u)}{}^2\right]\right]^{0.5} \quad (2.3.0)$$

And

$$\frac{\lambda_u}{\lambda}\left[E\left[D_{(u)}{}^2\right]\right]^{0.5} \le \left[\frac{\lambda_u}{\lambda}E\left[D_{(u)}{}^2\right]\right]^{0.5} \le \left[\sum_{i=1}^{U}\frac{\lambda_i}{\lambda}E\left[D_{(i)}{}^2\right]\right]^{0.5} = \left[E[\mathrm{D}^2]\right]^{0.5} \blacksquare \quad (2.3.1)$$

Corollary 2.0: *Consider an arbitrary system with U customer classes. If the system unfairness obeys* $E[\mathrm{D}^2] = 0$ *then for every class* $1 \le u \le U$ *the class discrimination obeys* $E\left[D_{(u)}\right] = 0$. The proof is immediate from Theorem 2.1 below.

Theorem 2.1: *Consider a system with U classes. Assume that the class discrimination of each class u obeys* $E\left[D_{(u)}\right] = 0$. *Then the system unfairness* $E[D^2] = 0$ *can still be positive.*

Proof - ***By numerical example:*** Consider a system with two classes, A and B. Assume that the service requirement is one unit for all customers and the arrival process is in pairs, one customer of each type. Assume that the inter-arrival time is given by $x > 2$ and that for half the pairs the server serves A first and for the other half it serves B first. One can easily observe that half of the customers experience positive discrimination of 0.5 and half experience negative discrimination of -0.5; thus, $E[D^2] = 0.25$. Nonetheless the expected class discrimination is zero for both classes. The implications of these results are:

A. If one maintains very low system unfairness it guarantees that the class discrimination of large population classes (classes with relatively high arrival rates) will be very small, while the discrimination of a lightly populated class can still be very high, and
B. Maintaining low class discrimination to all classes does not guarantee a fair system, since there could be unfairness in treatment of customers within a class.

2.4.2. Effect of Class Prioritization on System Unfairness:

One of queue prioritization strategies is the common practice of serving customers' with smaller service requirement (Short jobs) ahead of those with larger service requirement (Long jobs). In this section we first show by theorem 2.2 below, that generally speaking, prioritizing short jobs is justified, since otherwise these jobs are negatively discriminated. We then show the effectiveness of class prioritization and that while prioritization can guarantee positive discrimination to the class with highest priority and negative discrimination to the class with lowest priority; it cannot guarantee monotonicity in discrimination.

Definition 2.1 - ***Stochastic Dominance between Random Variables***: Consider non negative random variables X_1, X_2, whose distributions are $F_{X_1}(t) = Pr\{X_1 \le t\}$, $F_{X_2}(t) = Pr\{X_2 \le t\}$. We say that X_1 *stochastically dominates* X_2, denoted $X_1 \succ X_2$, if $F_{X_1}(t) \le F_{X_2}(t), \forall t \ge 0$.

Theorem 2.2 - ***Justifying the Prioritization of Short Jobs:*** *Let* C_l *be a customer with service requirement* s_l. *Consider a* $G/G/m$ *system under non-preemptive service policy, where the service decision is independent of the service times. Let* $D_l^{(s_l)}$ *be a random variable denoting the discrimination of* C_l, *when it arrives at the system in steady state. Then* $D_l^{(s_l)}$ *is monotone non-decreasing in* s_l, *namely if* $s_l' > s_l$ *then* $D_l^{(s_l')} \succ D_l^{(s_l)}$.

Proof: Consider service times s_l, s_l'; $s_l' > s_l$, and observe a customer C_l. Under any non-preemptive service policy, C_l waits until epoch q_l when it enters service, and stays in service until its departure. Equation (2.0.3) can thus be written as

$$D_l = \int_{a_l}^{q_l} \delta_l(t)dt + \int_{q_l}^{d_l} \delta_l(t)dt \quad (2.3.2)$$

The first term in this sum is independent of the service requirement. The second term is an integral of $\delta_l(t)$, over the interval $(q_l, d_l) \Rightarrow d_l - q_l = s_l$, which is the service time of C_l. To prove the monotonicity we consider a specific sample path π and compare the values of $D_l^{(s_l)}$ and $D_l^{(s_l')}$ for this path, denoted by $D_{l,\pi}^{(s_l)}$ and $D_{l,\pi}^{(s_l')}$. From equation (2.3.2) we have

$$D_{l,\pi}^{(s_l')} - D_{l,\pi}^{(s_l)} = \int_{q_l}^{q_{l+s'}} \delta_l(t)dt - \int_{q_l}^{q_{l+s}} \delta_l(t)dt = \int_{q_{l+s}}^{q_{l+s'}} \delta_l(t)dt \ge 0 \quad (2.3.3)$$

Where the last inequality is due to $\delta_l(t)dt \ge 0$, which is obvious from equation (2.1.2). Since equation (2.3.3) holds for every sample path π, the proof follows. ■

The theorem 2.2, stated in terms of deterministic service requirements can also be stated using stochastic service requirements. That is, if the customer's service requirements are stochastic variables S_l and S_l', and $S_l' \succ S_l$, then $D^{S_l} \succ D^{S_l'}$ and clearly $E[D^{S_l'}] \ge E[D^{S_l}]$. Similarly, using class notation, if S_u is the service requirement distribution of class u customers, then, $S_u \succ S_{u'} \Rightarrow E[D_{(u)}] \ge E[D_{u'}]$. In conclusion, service policies that do not give preferential service to shorter jobs actually discriminate against those jobs. This provides one more justification for prioritizing shorter jobs.

Remark 2.0: Using the same arguments, it can be shown that theorem 2.2 also holds in the case of a preemptive system, provided that the preemption of a customer with service $(s' > s)$ during the period at which it receives the first s units of service is unchanged, i.e. preemptions are not determined by the customer's service requirement.

2.4.3. Effect of Prioritization on Class Discrimination:

How does class prioritization (i.e. serving classes with smaller service requirements ahead of classes with larger

service requirements) affects class discrimination? The theorem 2.3 below addresses this question.

Theorem 2.3: *In a G/G/m system with U classes, if the services scheduling policy belongs to the class of preemptive priority scheduling policies, then* $E[D_{(1)}] \geq 0$, and $E[D_U \leq 0]$.

Proof: Let $N_u(t)$ be the number of class $u$customers in the system at epoch t. If $N_1(t) \leq m$, then all $N_1(t)$ customers will be served at epoch t; otherwise if $N_1(t) > m$, then only $N_1(t) = m$will be served. Thus

$$\widetilde{D}_1(t) = \begin{cases} N_1(t) - \frac{\omega(t)N_1(t)}{N(t)} = N_1(t)\left[1 - \frac{\omega(t)}{N(t)}\right]; & N_1(t) \leq m \\ m - \frac{mN_1(t)}{N(t)} = m\left[1 - \frac{N_1(t)}{N(t)}\right]; & N_1(t) > m \end{cases} \quad (2.3.4)$$

Equation (2.3.4) is greater or equal to zero since $\omega(t) \leq N(t))$ and $N_1(t) \leq N(t)$. Thus

$$\widetilde{D}_1(t) \geq 0 \Rightarrow \widetilde{D}_1 \geq 0 \Rightarrow E[\widetilde{D}_{(1)}] \geq 0 \quad (2.3.5)$$

And from equation (2.2.8), $E[D_{(1)}] \geq 0$. Note that equation (2.3.4) also provides the only epochs in which$\widetilde{D}_1(t) = 0$, namely when either $N_1(t) = N(t)$, (all the customers in the system are of class-1), or $N(t) < m$, (there are less than m customers in the system), or $N_1(t) = 0$. In fact, for every class u, $\widetilde{D}_{(u)}(t) = 0$ when either $N_u(t) = N(t)$ or $N(t) < m$, or$N_u(t) = 0$. And$\widetilde{D}_U(t) = 0$, when either $N_U(t) = N(t)$ or $N(t) < m$, or $N_U(t) = 0$. Otherwise there are two cases, either $N(t) - N_U(t) \geq m$ or $N(t) - N_U(t) < m$. In the first case there are more than $m$customers of higher priority in the system, and thus no class$U$customers are being served. Therefore,

$$\widetilde{D}_U(t) = -N_U(t)m[N(t)]^{-1} \leq 0 \quad (2.3.6)$$

In the second case there are some class U customers being served. In this case let $\omega_U(t)$ be the number of class U customers served at epoch t; using this notation

$$\widetilde{D}_U(t) = \omega_U(t) - \frac{N_U(t)m}{N(t)} = \frac{\omega_U(t)N(t) - N_U(t)m}{N(t)} \leq 0 \quad (2.3.7)$$

To prove that equation (2.4.7), let $N'(t) = N(t) - m$denote the number of customers waiting at epoch t, all of whom must be of class U. We can write$N(t) = m + N'(t)$ and $N_U(t) = \omega_U(t) + N'(t)$, and substituting these expression into equation (3.8.2) yields

$$\widetilde{D}_U(t) = \frac{\omega_U(t)(m+N'(t)) - (\omega_U(t)+N'(t))m}{N(t)} = \frac{(\omega_U(t) - m)N'(t)}{N(t)} \leq 0 \quad (2.3.8)$$

Since $\omega_U(t) \leq m$;thus $\widetilde{D}_U(t) \leq 0 \Rightarrow \widetilde{D}_U \leq 0 \Rightarrow E[\widetilde{D}_U] \leq 0$ and from equation (3.9.4), $E[D_U] \leq 0$ ■

The significant of theorem 2.3 is that, the most prioritized class has greater or zero discrimination, even if the customers are extremely small. This means that at least for the first priority class, certain discrimination can be guaranteed. Having shown that the discrimination of the most prioritized class is always non-negative, and that the discrimination of the least prioritized class is always non-positive, however the discrimination is not monotonic with respect to the class priority, because class prioritization is limited in its effect in multiple class systems.

2.5. Resource Dedication to Classes (Multi- M/M/1 Systems)

Perhaps the most common alternative approach for dedicating resources to customer is by assigning each customer's class a set of one or more servers and associating each class with a single FCFS queue policy, thus, transforming M/M/m single queue architecture to multi-M/M/1single queue systems. This architecture is very common in human-service facilities, including airport passport control systems which are customarily divided to alien and non-alien classes, and public restrooms. Since in some of these systems, customers are classified based on personal properties, such as gender or nationality, or economic values (e.g. VIP), etc., fairness aspects of these systems are highly important. An operational question of interest in this system is whether to allocate equal amount of resources to the different classes or to grant more resources to the class with the larger service requirement. The answer to this question is not immediate since one of the basic principles of the RAQFM fairness measure is that short jobs should get preference over long jobs.

3. ANALYSIS OF THE PPMC PREEMPTIVE PRIORITY QUEUING SYSTEM

To practically demonstrate the effect of service prioritization on a multi-class multi-server single queue system, we take a case study of the PPMC Petroleum products distribution queuing system. By personal facility tour of the organization's queuing system reveals that there are eight electronically operated parallel servers, running for eight hours per day for five working days per week. However, two hours per day is set aside for daily routine maintenance and repairs of any malfunctioned servers and documentation. Of the eight (8) parallel servers, four are designated for PMS (petrol); two for AGO (diesel) and two for DPK (kerosene), each serving at an average rate of 11,000 litres per minutes. To guide against the eventuality of sudden breakdown of any server during operations, two of the four PMS servers are sometimes kept on reserve when there are less work load in the depot.

On the service delivery policy, the PPMC service records indicate that services are prioritized to two major economy and geo-political classes of customers respectively, at different pre-emptive priority levels. The major economic class consist of customers from the major petroleum products marketing companies (major marketers) owned by multi-national oil companies (e.g. Mobil, Oando, Total, Texaco, AP etc), and customers from licensed independent petroleum products marketing companies owned by private individuals (independent marketers). The geo-political classes consist of marketers operating within the central area district of Abuja FCT and its environs, and customers from outside the FCT and the neighboring states (e.g. Kaduna, Nassarawa, Kogi, Niger, etc.).Tentatively the PPMC priority classes are:

- Class-1: Major marketer from within the Abuja geopolitical zone,
- Class-2: Independent marketer from within Abuja geopolitical zone,
- Class-3: Major marketers from outside the Abuja geo-political zone, and
- Class-4: Independent marketers from outside the Abuja geo-political zone

An order of priority is assigned to the classes, where lower class index means higher priority, and a *preemptive* service policy is assigned, where the order of service within each class is FCFS, and preempted customers return to the head of the queue of their class. These customers are queued up within an imaginary single queue finite waiting space provided by the organization. Predicated on the assumption that all jobs arriving at the system have equal service requirement (capacity of 33,000 liters/truck), and that the present system performance can be predicted from the records of past operations, we proceed to analyzed specifically the performance measure of the queuing system dedicated to the distribution of PMS from 2000-2006.

3.1. The PPMC 4-Class M/M/4Priority System

As shown diagrammatically in figure 1.0, the PPMC queuing configuration designated to PMS distribution is an M/M/4 system - consists specifically of four parallel servers and four customers classes served under a single FCFS queue. The system is subject to the arrival of a stream of 4 classes of customers, $C_j, (j = 1,2, \dots ,4)$, respectively whose arrival rate,$\lambda_j, (j = 1,2, \dots ,4)$ follow a Poisson process and their required service rate are identically independent variable (i.i.d) with mean,$\mu_j, (j = 1,2, \dots ,4)$. An order of priority is assigned to the classes, where lower class index mean higher priority. To ascertain the validity of the observed data, a chi-square goodness of fit test was invoke to determine if the customers' arrival rates are Poisson distributed and the service rates are exponentially distributed. Thus, the mean arrival rate of a Class-j customer: $\lambda_1 = 6.5 trucks/hr$; $\lambda_2 = 5.57 trucks/hr$; $\lambda_3 = 4.43 trucks/hr$; $\lambda_4 = 2.83 trucks/hr$,giving the system total arrival rate: $\lambda = 19.33 trucks/hr$. Thus, at every one hour of the day, a total of 19.33trucks must arrived the system for service, while the service rate of each of the parallel servers is exponentially distributed with the parameter: $\mu = 11{,}000 litres/\min \Rightarrow 20 trucks/hr$. Thus at every one hour of the day, an average of 20 trucks must depart the system after service.

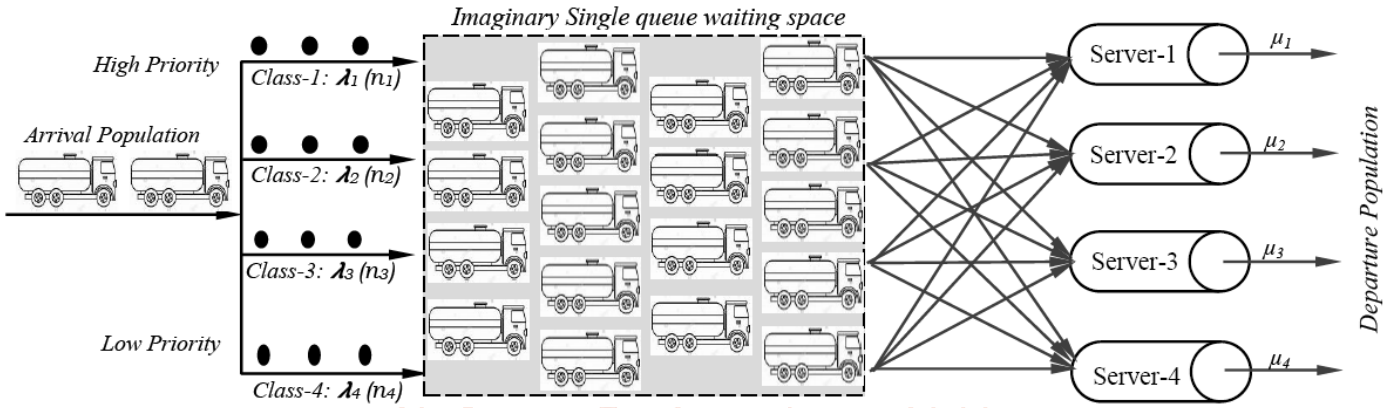

Figure 1.0: The PPMC Class Prioritization Queuing System

3.1.1. Operating Characteristics of the PPMC system:

The traffic intensity of class-1 customers: $\rho_1 = 0.0813$, this gives an average of 8.13% class-1 customers per busy server per hour.If a class-1 customer enters the server, we assume that it must be served fully before departure since he is in the high priority class. However, if the class-1 customer's service is preempted due to system failure, the additional time required to complete a class-1 services follows an exponential distribution with mean$(\rho_1^{-1} = 12.3)$. This means that a class-1customer spent an average time of 12.3minutesfrom entering the system to departure. The system utilization rate$(1 - \rho_1 = 0.9187)$, implies that about 991.87% of class-1 customers are on queue per hour.

The traffic intensity of class-2 customers: $\rho_2 = 0.0696$, this gives an average of 6.96% of class-2 customers per busy server per hour. If a class-2 customer enters the server, we assume that it must be served fully before departure or preempted by a higher priority customer. Therefore, the additional time required to complete a class-2 services follows an exponential distribution with mean$(\rho_2^{-1} = 14.37)$. This means that a class-2 customer spent an average time of 14.37 minutes from entering the system to departure. The system utilization rate$(1 - \rho_2 = 0.9304)$, implies that about 93.04% of class-2 customers are on queue per hour.

The traffic intensity of class-3 customers: $\rho_3 = 0.0554$, this gives an average of 5.54% of class-3 customers per busy server per hour. If a class-3 customer enters the server, we assume that it must be served fully before departure or preempted by higher priority customers. Therefore, the additional time required to complete a class-2 services follows an exponential distribution with mean$(\rho_3^{-1} = 18.05)$. This means that a class-3 customer spentan average time of 18.05 minutes from entering the system to departure. The system utilization rate$(1 - \rho_3 = 0.9446)$, implies that about 94.46% of class-3 customers are on queue per hour.

The traffic intensity of class-4 customers: $\rho_4 = 0.0354$, this gives an average of 3.54% of class-4 customers per busy server per hour. If a class-4 customer enters the server, we assume that it must be served fully before departure or preempted thrice by higher priority customers. Therefore, the additional time required to complete a class-4 services follows an exponential distribution with mean$(\rho_4^{-1} = 28.25)$. This means that a class-4 customer spent an average time of 28.25 minutes from entering the system to departure. The system utilization rate$(1 - \rho_4 = 0.9646)$, implies that about 96.46% of class-4 customers are on queue per hour.

By equation (2.1.6) the traffic intensity of the M/M/4 system, $\rho = 0.2417$, this gives an average of 24.17% of class-j customers per busy server per hour. If a class-j customer enters the server, we assume that it must be served fully before

departure if in the highest priority class, or partially if displaced by a high priority customer. Therefore, the additional time required to complete a class-j services if preempted by a higher priority customer follows an exponential distribution with mean($\rho^{-1} = 4.14$). This gives an average time of 4.14 minutes from entering the system to departure. The system utilization rate($1 - \ \rho = 0.7584$), implies that about 75.84% of the customers are on queue per hour

Consequentially, the throughput($\gamma = m\rho\mu$)or the mean number of class-1requests serviced per a time unit:$\gamma_1 = 6.504$, thus an average 6.504 class-1 services are granted per hour; also class-2 throughput:$\gamma_2 = 5.568$ –an average of 5.568 class-2 customers' service are granted per hour. And the class-3 throughput: $\gamma_3 = 4.432$ –anaverage of 4.432 class-3 customers' services are granted per hour, while the clas-4 throughput$\gamma_4 = 2.832$, thus, an average ofonly 2.832 class-4 customers' services are granted. This give a total system throughput: $\gamma = 19.328 trucks/hr$, thus an average of 19.328 customers' services are granted per hour.

By equation (2.2.2), the probability that the system is empty, i.e. there is no class-j customer in the system,$P_0 = 0.3807$.This implies that the server is idle at 38.07% of the time, while its get busy at 61.93% of the time. Similarly, the probability that there is no class-1 customer in the system, $P_{0_1} = 0.7224$ (72.24%), thus the servers only get busy with class-1 customers only at 27.76% of the time. Also the probability that there is no class-2 customer in the system,$P_{0_2} = 0.757$ (75.7%), thus the servers only get busy with class-2 customers only at 24.3% of the time. The probability that there is no class-3 customer in the system, $P_{0_3} = 0.8012$ (80.12%), thus the servers only get busy with class-3 customers only at 19.88% of the time. Finally, the probability that there is no class-4 customer in the system, $P_{0_4} = 0.867$ (86.7%), thus the servers only get busy with class-4 customers only at 13.3% of the time.Let $\overline{W_j}$ denotes the mean waiting time of a class-j customer in the system, by Sztrik (2012),

$$\overline{W_j} = \frac{1}{\mu} + \overline{W_q} = \frac{1}{\mu} + \frac{\left[\frac{m!}{\rho!(m-\rho)!}\right]^{\frac{1}{\mu}}}{m\left[1-\left(\frac{1}{\mu}\sum_{i=1}^{j-1}\lambda_i\right)/m\right]\left[1-\left(\frac{1}{\mu}\sum_{i=1}^{j}\lambda_i\right)/m\right]}, j = 1,2,\dots,4 \quad (3.0.1)$$

So $\overline{W}_{q_1} = 0.02141hrs$; $\overline{W}_{q_2} = 0.02522hrs$; $\overline{W}_{q_3} = 0.02919hrs$; $\overline{W}_{q_4} = 0.03268$hrs, giving an average waiting time in the queue:$\overline{W}_q = 0.02713$hrs per customers' class.Therefore, $\overline{W_1} = 0.0714hrs$; $\overline{W_2} = 0.0752hrs$; $\overline{W_3} = 0.0792hrs$; and $\overline{W_4} = 0.0827hrs$;giving an average system waiting time:$\overline{W} = 0.0771hrs/truck$.

Similarly, the mean number of class-j customers waiting on the queue (queue length) at any epoch:$\bar{Q}_j = \lambda_j \overline{W}_j$; thus, $\bar{Q}_1 = 0.464trucks$; $\bar{Q}_2 = 0.419trucks$; $\bar{Q}_3 = 0.351trucks$; $\bar{Q}_4 = 0.234trucks$,giving a total number ofqueuelengths:$Q = 1.468trucks/hr$ or 8.808trucks/day.

Therefore, the mean response time of a class-j customer; $\overline{T_j} = \overline{W_j} + \frac{1}{\mu}$, thus,$\bar{T}_1 = 0.1214hrs$; $\bar{T}_2 = 0.1252hrs$; $\bar{T}_3 = 0.1292hrs$; $\bar{T}_4 = 0.1327hrs$; giving an averagesystem response time:$\bar{T} = 0.5085$hrs. While the mean number of class-j customers in the system (waiting and service) any epoch: $\bar{N} = \lambda\bar{T}$, thus, $\bar{N}_1 = 0.789trucks$; $\bar{N}_2 = 0.697trucks$; $\bar{N}_3 = 0.572trucks$; $\bar{N}_4 = 0.376trucks$;giving a totalnumber of customers in the system:$N = 2.434\ tucks/hrs\ or\ 14.604trucks/day$. Finally, by Erlang loss formula[20], the delay probability of class-j customer is given by:

$$D = C(m,\rho) = P_m = \frac{\rho(m-1-\rho)\cdot C(m-1,\rho)}{(m-1)(m-\rho) - \rho C(m-1,\rho)} \quad (3.0.2)$$

Thus,$D_1 = 0.0814$, (8.14%), i.e. at 8.14% of the time an arriving class-1 customer has to wait on the queue. For class-2 customers,$D_2 = 0.0856$ (8.56%), implies at 8.56% of the time an arriving class-2 customer has to wait on the queue. While for class-3 customers,$D_3 = 0.0903$ (9.03%), implies that at 9.03% of the time an arriving class-3 customer has to wait on the queue. Finally, for class-4,$D_4 = 0.0963$, (9.63%), implies that at 9.63% of the time an arriving class-4 customer has to wait on the queue. The system delay probability$D = 0.0884$ (8.84%) implies that on the average, an arriving customer has to wait for 8.84% of the time on the queue.

The figure 1.1 below shows the variability of customer's response/sojourn time with their arrival rates. Considering the preemptive priority service policy of the PPMC system, with the average system response time of 0.1271 hours, the red curve shows that classes with high arrival rates (larger service requirements) enjoyed smaller response time while classes with low arrival rates (smaller service requirement) enjoyed larger response time. Similarly, with the average system sojourn time of 0.069 hours, the blue curve shows that classes with low arrival rates (smaller service requirement) sojourned more or spent longer time in the system than the jobs with high arrival rates (high service requirement).

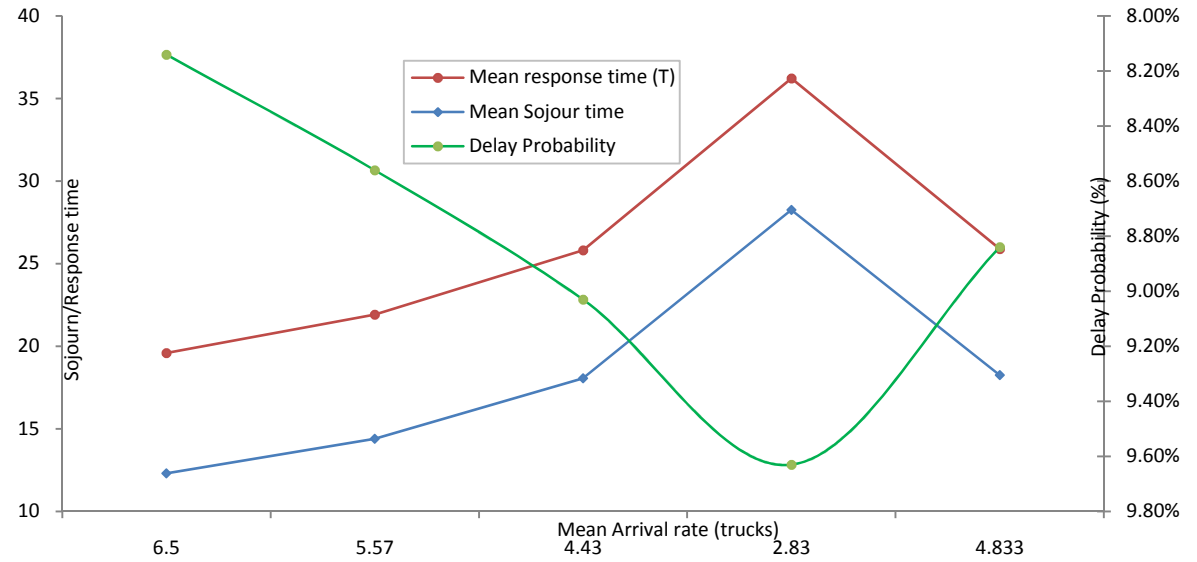

Figure 1.1: Variability of System Response/Sojourn time with mean arrival rate

With the mean delay probability of 8.84%, the green curve of figure 1.1 above, also shows that classes with high arrival rates (high service requirement) enjoyed lower delay probability, while classes with low arrival rates (low service requirement) are highly delayed in the system. Therefore, by these characteristics, the preemptive priority service policy of the PPMC queuing system violate the service requirement preference principle (Axioms 3.1), hence an unfair system.

3.1.2. The Discrimination Coefficient:

From the PPMC records, each of the four servers runs for five working days per a week at 6 hours per day, while 2 hours is dedicated to maintenance services on the servers. Thus, the total resources granted by the 4-servers per day, $\omega(t) = 4(6)(60)(11{,}000)litres/day = 15840000litres/day$ =480trucks/day.Byequation (2.0.1), each class of customer in the system at that epoch deserved a momentary warranted service rate:$R(t) = 32.8677trucks/day$. However, from table-1.0 the momentary granted rates for a class-j customer are: $\sigma_1(t) = 33.6trucks/day$; $\sigma_2(t) = 28trucks/day$; $\sigma_3(t) = 21.2trucks/day$; and$\sigma_4(t) = 14.8trucks/day$. Givingthe total grantedrate: $\sigma(t) = 97trucks/day$. Therefore, by equation (2.0.2), the momentary class discrimination of class-j customer:$\delta_1(t) = 0.7323trucks/day > 0$; $\delta_2(t) = -4.8677trucks/day < 0$; $\delta_3(t) = -11.6677trucks/day < 0$; and $\delta_4(t) = -18.0677trucks/day < 0$. Givingtotal momentary discrimination,$\delta(t) = -33.8708$trucks/day < 0.Therefore, by equation (2.0.3), the accumulative discrimination experienced by a class-j customer over the 5 working days:$D_1(t) = 3.6615trucks/week < 0$; $D_2(t) = -24.3385trucks/week < 0$; $D_3(t) = -58.3385trucks/week < 0$; $D_4(t) = -90.3385trucks/week < 0$.This giving the totalClass discrimination rate,$D = -169.354trucks/week < 0$, thus, yielding a deficit supply of 169.354trucks of PMS per week.

The figure 1.2 below shows the variability of momentary/accumulative discrimination of a class-j customer with mean arrival into the system. Considering the preemptive priority service policy of the PPMC, the blue and the green curves both shows that classes with high daily arrival rates (larger service requirements) are less negatively discriminated, while classes with low daily arrival rates (smaller service requirements) are more negatively discriminated momentarily and accumulatively. Similarly, with the average class discriminative index of -70.97, the red curve shows that the low priority classes (smaller service requirements) are more negatively discriminated while the high priority classes (large service requirements) as less negatively discriminated. Therefore, with these characteristics, the preemptive priority service policy of the PPMC queuing system violates the service requirement preference principle as well as the RAQFM measure, hence a negatively discriminative and an unfair system.

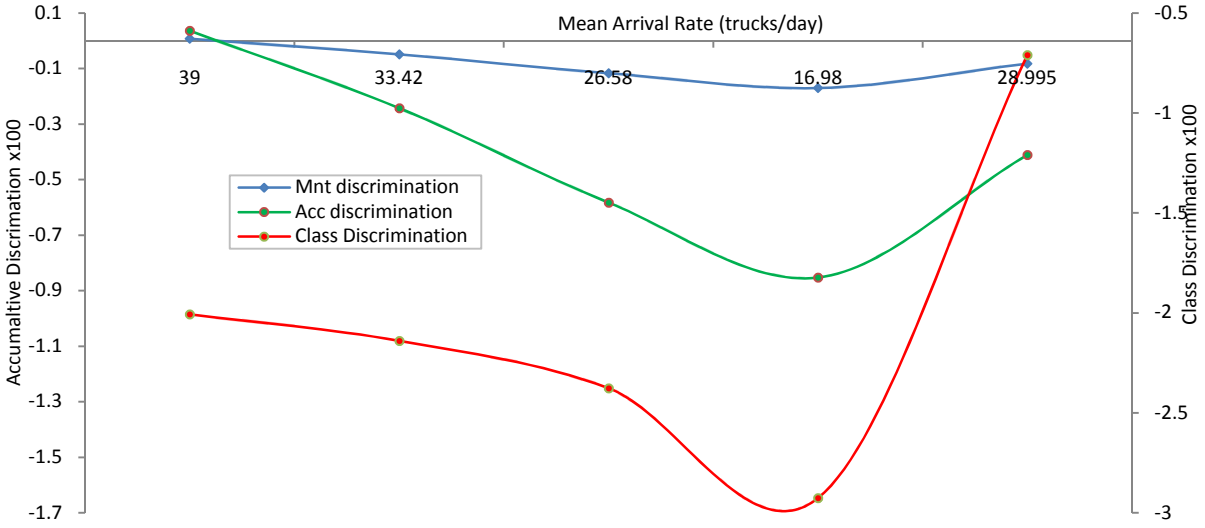

Figure 1.2: Variability of Discrimination Coefficient with Mean Arrival rate

3.1.3. The System Unfairness/Class Discrimination:

By equation (2.2.2), the probability that a customer meets $k = 0.789$ class-1 trucks/hr in the system:$P_{k_1} = 0.3207$ (32.07%). The probability that a customer meets $k = 0.697$ class-2 trucks/hr in the system: $P_{k_2} = 0.3419$ (34.19%). Similarly, the probability that a customer meets $k = 0.572$ class-3 trucks/hr in the system:$P_{k_3} = 0.3799$ (37.99%). Also the probability that a customer meets $k = 0.376$ class-4 trucks/hr in the system:$P_{k_4} = 0.4677$ (46.77%). And the probability that a customer meets a total of $k = 2.434$ trucks/hr in the system:$P_{k_n} = 0.1134$ (11.34%). Therefore, by equation (2.2.1)and (2.2.8) the system unfairnessand class discrimination given that a class-j customer meets at least$k$ −class-j trucks/hr in the system:$E[D^2]$ and $E\left[\tilde{D}_{(u)}\right]$ respectively, can be represented by table 1.1and 1.2 below.

Table1.0: Class Unfairness With Respect To System Unfairness

No of Customer:	Zero Trucks	$k_1 = 0.789$	$k_2 = 0.697$	$k_3 = 0.572$	$k_4 = 0.376$	$k_n = 2.434$
Prob. of k-Trucks:	0.3807	0.3207	0.3419	0.3799	0.4677	0.1134
$E\left[D_{1,k}^2\right] = D_1^2 P_k$:	3.3572	2.8281	3.015	3.3501	4.1244	1.00
$E\left[D_{2,k}^2\right] = D_2^2 P_k$:	3.3572	2.8281	3.015	3.3501	4.1244	1.00
$E\left[D_{3,k}^2\right] = D_3^2 P_k$:	3.3572	2.8281	3.015	3.3501	4.1244	1.00
$E\left[D_{4,k}^2\right] = D_4^2 P_k$:	3.3572	2.8281	3.015	3.3501	4.1244	1.00
$E[D^2]$:	**13.4288**	**11.3124**	**12.06**	**13.4004**	**16.4976**	**4.00**

From table 1.0 above, the system unfairness given that there are no arriving customer but there are k numbers of class-jtrucks in the system: $E[D^2] = E[D^2(0,k)] = 17.6748$. Similarly, from table 1.2 below, the system class discrimination given that there are no arriving customers but there are k numbers of class-j trucks in the system: $E\left[\tilde{D}_{(u)}\right] = E\left[\tilde{D}_{(u)}(0,k)\right]$, are $E\left[\tilde{D}_{(1)}\right] = 17.6748; E\left[\tilde{D}_{(2)}\right] = E\left[\tilde{D}_{(3)}\right] = E\left[\tilde{D}_{(4)}\right] = -17.6748 < 0$.

Table 1.1: Class Discrimination with respect to System Discrimination

No of Customer:	Zero trucks	$k_1 = 0.789$	$k_2 = 0.697$	$k_3 = 0.572$	$k_4 = 0.376$	$k_n = 2.434$
Prob. of k-Trucks:	0.3807	0.3207	0.3419	0.3799	0.4677	0.1134
$E\left[\tilde{D}_{(1,k)}\right] = \lambda_1[D_1 P_k]$	3.3572	2.8281	3.015	3.3501	4.1244	1.00
$E\left[\tilde{D}_{(2,k)}\right] = \lambda_2[D_2 P_k]$	-3.3572	-2.8281	-3.015	-3.3501	-4.1244	-1.00
$E\left[\tilde{D}_{(3,k)}\right] = \lambda_3[D_3 P_k]$	-3.3572	-2.8281	-3.015	-3.3501	-4.1244	-1.00
$E\left[\tilde{D}_{(4,k)}\right] = \lambda_4[D_4 P_k]$	-3.3572	-2.8281	-3.015	-3.3501	-4.1244	-1.00
$E\left[\tilde{D}_{(u)}\right]$	**-6.7144**	**-5.6562**	**-6.03**	**-6.7002**	**-8.2488**	**-2.00**

Figure 1.3 below represents the variability of class discrimination and the system unfairness coefficients with respect to the number of k class-j trucks a lass-j customer meeting in the system. Considering the preemptive priority service policy of the PPMC; the blue curves shows that with the mean system discrimination index of -2, classes with large number of customers in the system (large service requirement) are less negatively discriminated, while classes with smaller number of customers in the system (small service requirement) are more negatively discriminated. Similarly, with the mean system unfairness index of +4, the red curve shows that the PPM system is less unfair to classes with large number of customers in the system (high priority), while the system is more unfair to classes with smaller number of customers in the system. This is a violation of both the service requirement preference principle (Theorem 2.1) and the RAQFM measure.

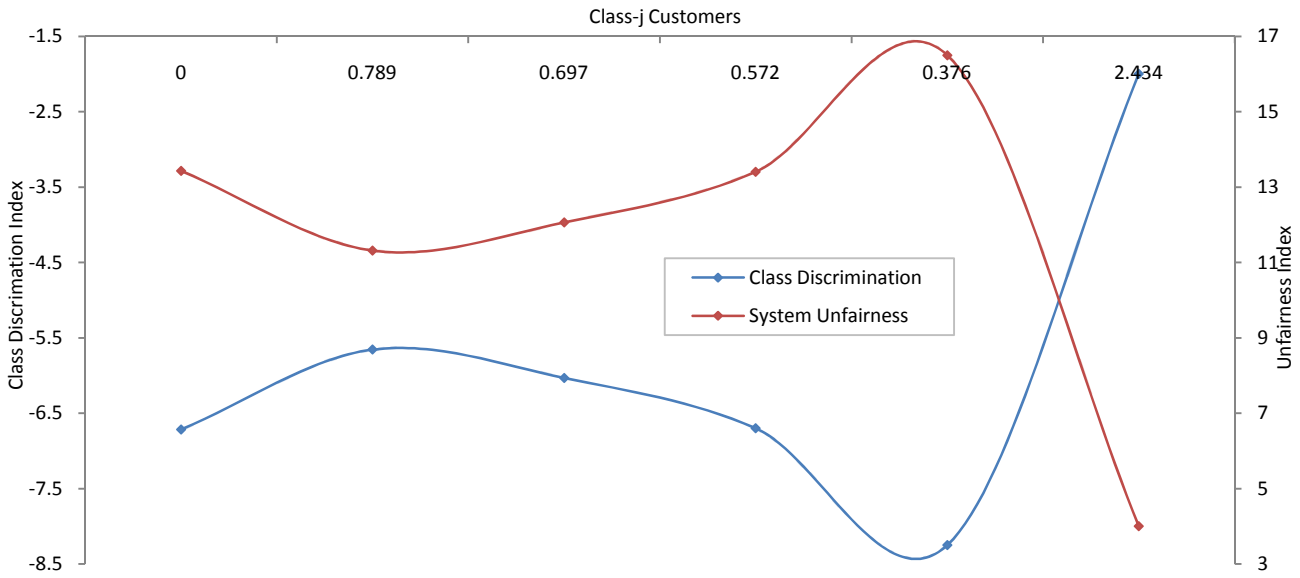

Figure 1.3: Variability of System Discrimination/Unfairness with No of Customers in system

3.2. The Resource Dedication in PPMC Queuing System (Alternative Four-M/M/1 Systems)

As an alternative configuration for guaranteeing queuing fairness in the PPMC, we considered a queuing system where each customer's class is assigned a dedicated server or set of servers associated with a single FCFS queuing policy, thus converting the 4-classes M/M/4 single queue system to four M/M/1 system. In the context of dedicating resource to classes of customers, several questions need to be addressed: (i) how should servers be assigned to each class so as to engender maximal fair scheduling, (ii) how fair is the resource dedication design, and (iii) what is the class-discrimination experienced under each M/M/1 system? To address these questions, the figure 3.0 below illustrates the system configuration of the four M/M/1 queuing system.

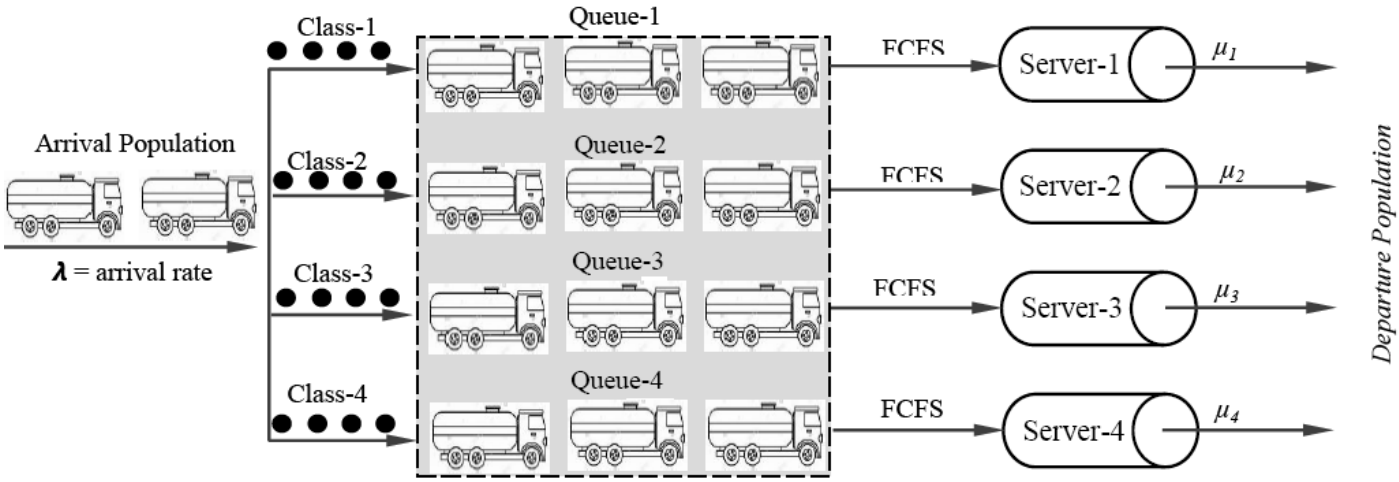

Figure 1.4: Proposed Resource Dedication Queuing System

3.2.1. Operating Characteristics of the PPMC Alternative System:

Considering that each class of customer is assigned a separate server associated with a FCFS service policy, and there are a total of 580 trucks that arrived the system weekly; the respective mean arrival rates of a class-j customer:$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 19.33 trucks/hr$and mean service rate: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 20 trucks/hr$. The traffic intensity of the each of the M/M/1systems,$\rho_1 = \rho_2 = \rho_3 = \rho_4 = 0.9665$. This gives an average of 96.65%class-jcustomers at each busy server per hour. If a customer enters the server, we assume that it must be served fully before departure or partially if there is system breakdown. Therefore, the additional time required to complete a customer's services follows an exponential distribution with mean: $\rho_1^{-1} = \rho_2^{-1} = \rho_3^{-1} = \rho_4^{-1} = 1.0347$. This gives an average time of 1.0347 minutes from entering the system to departure. The system utilization rate for each of the four systems:$1 - \rho_1 = 1 - \rho_2 = 1 - \rho_3 = 1 - \rho_4 = 0.0335$, implies that only3.35%of customers are on each queue per hour.

By Mathew [20]the probability that the system is empty: $P_0 = 1 - \rho \Rightarrow P_{0_1} = P_{0_2} = P_{0_3} = P_{0_4} = 0.0335$. This implies that each of theservers isidle only at 3.35% of the time, while theyare all busy at 96.65% of the time. By Mathew [20]the mean number of customers on each of the four queues at any epoch:

$$\bar{Q}_1 = \bar{Q}_2 = \bar{Q}_3 = \bar{Q}_4 = \frac{\rho^2}{1-\rho} = \frac{\rho^2}{P_0} = 27.884\ trucks/hrs \quad (3.0.3)$$

Giving the total number of customers on the four queues: $Q = 111.536 trucks/hrs$.Similarly, the average number of customers in each of the systems (waiting and service) at any epoch:

$$\bar{N}_1 = \bar{N}_2 = \bar{N}_3 = \bar{N}_4 = \frac{\rho}{1-\rho} = \frac{\rho}{P_0} = 28.8507 trucks \quad (3.0.4)$$

Giving the total number of customers in the four systems: $N = 115.4028 trucks/hr$. Also from Mathew [20] the mean waiting (response) time of a customer in each of four the systems:

$$\bar{T} = \bar{T}_1 = \bar{T}_2 = \bar{T}_3 = \bar{T}_4 = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu P_0} = 1.4925 mins \quad (3.0.5)$$

Similarly, the mean waiting time of a customer on each of the four queues:

$$\bar{W}_q = \bar{W}_{q_1} = \bar{W}_{q_2} = \bar{W}_{q_3} = \bar{W}_{q_4} = \frac{\rho}{\mu(1-\rho)} = \frac{\rho}{\mu P_0} = 1.4425 mins \quad (3.0.6)$$

Also, the probability that an arriving customer in system must wait (will be delayed) on the queue: $D = \rho \Rightarrow D_1 = D_2 = D_3 = D_4 = 0.9665$ (96.65%); thus, at 96.65%of the time an arriving customer has to wait on the queue.

3.2.2. Comparison of System Response/Sojourn Times of the PPMC with the Alternative Systems:

The Figure 1.5 below shows the variability of system response and sojourn time with customers'' classes in the two systems. Comparatively, the thick red curve shows that the mean sojourn time for jobs in the PPMC system are higher for low priority classes (smaller service requirements)and lower for high priority classes (high service requirements). In the alternative system, the broken red curve shows that all classes of customers or jobs enjoyed the same sojourn time of 1.0347 minutes irrespective of service requirements or service policy. Similarly, the broken blue curve shows that the mean response time for jobs in the PPMC system are higher for low priority classes (smaller service requirements) and lower for high priority classes (high service requirements), while in the alternative system, the thick blue curve shows that

all classes of customers or jobs enjoyed the same mean response time of 1.4925minutes irrespective of service requirements or service policy. By these characteristics, the resources dedication alternative i.e. assigning each customer's class to a server or a set of servers associated with a single FCFS queue policy obeyed the fundamental principle of queue fairness (service requirement preference principle). Thus, guaranteeing quick response as well as limited sojourn time to job classes irrespective of service requirement, hence a fairer alternative.

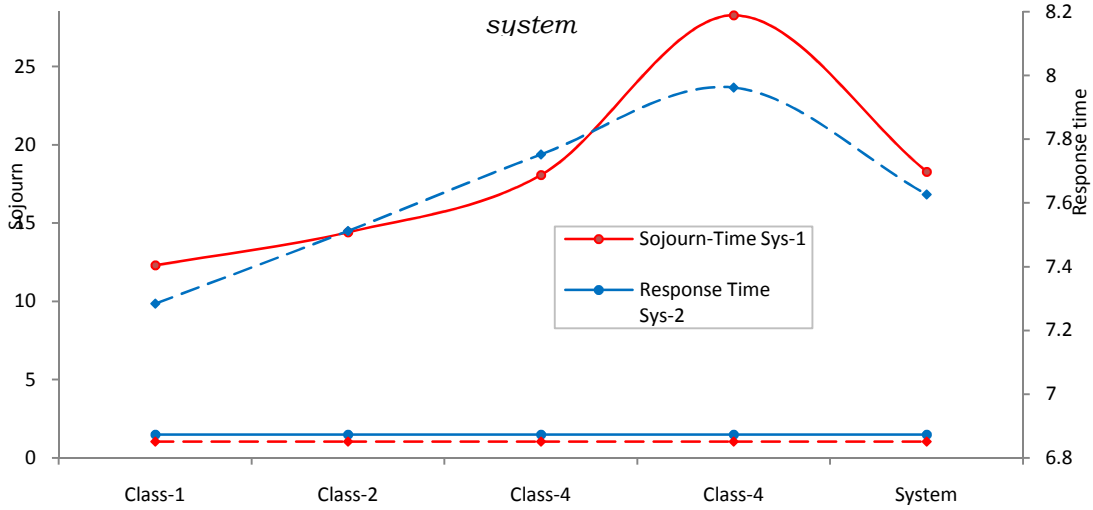

Figure1.5: Comparison of system Response/Sojourn time of the PPMC and Alternative

3.2.3. The System Discrimination:

Considering that the total granted resources of each server at any epoch:$\omega_1(t) = \omega_2(t) = \omega_3(t) = \omega_4(t) = 20trucks/hrs$, and there are$\bar{N}_1 = \bar{N}_2 = \bar{N}_3 = \bar{N}_4 = 28.8507trucks/hr$in each the system.By equation (2.0.1) the momentary warranted service of a class-j customer:$R_1(t) = R_2(t) = R_3(t) = R_4(t) = 0.6932trucks/hr$, but from table-1.0, the momentary granted rates of a class-j customer: $\sigma_1(t) = 5.6trucks/hr$; $\sigma_2(t) = 4.667trucks/hr$; $\sigma_3(t) = 3.5333trucks/hr$; $\sigma_4(t) = 2.4667trucks/hr$. Giving a total granted rate: $\sigma(t) = 16.267trucks/hr$. By equation (2.0.2), the momentary class discrimination of a class-j customer: $\delta_1(t) = 4.9068trucks\ or\ 29.4408trucks/day > 0$; $\delta_2(t) = 3.9738trucks\ or\ 23.8428trucks/day > 0$; $\delta_3(t) = 2.8401trucksor\ 17.0406trcks/day$and $\delta_4(t) = 1.7735trucks/hr\ or\ 10.641trucks/day > 0$. Giving a total momentary class discrimination:$\delta(t) = 80.9652$trucks/day. Therefore, by equation (2.0.3), the accumulative discrimination experienced by a class-j customer over the five working days: $D_1(t) = 147.204trucks/week > 0$; $D_2(t) = 119.214trucks/week > 0$; $D_3(t) = 85.203trucks/week > 0$; $D_4(t) = 53.205trucks/week > 0$. Thus, giving a total Class discrimination rate, $D = 404.826trucks/week > 0$, thus yielding anexcess supply of 13,356,858litres of PMS per week.

3.2.4. Comparison of System Response/Sojourn Times of the PPMC with the Alternative Systems:

The Figure 1.6 below shows the variability of momentary and accumulative discrimination index with customers 'classes in the two systems. Comparatively, the blue curves shows that in both momentary and accumulative variants, classes with low priority service (smaller service requirements) in the PPMC system are more negatively discriminated than classes with high priority service (high service requirements). In the alternative system, the red curves shows that both momentary and accumulative discrimination of a class-j customer increases more positively for classes with high priority service (larger service requirements)are than classes with low priority service (smaller service requirements). By these characteristics, the alternative system i.e. assigning each customer's class to a server or a set of servers associated with a single FCFS queue policy is more positively discriminative than the PPMC system. Thus, with high positive discrimination index, the alternative architecture guarantees a fairer petroleum products distribution to the respective customers' classes irrespective of service requirement, hence a fairer alternative.

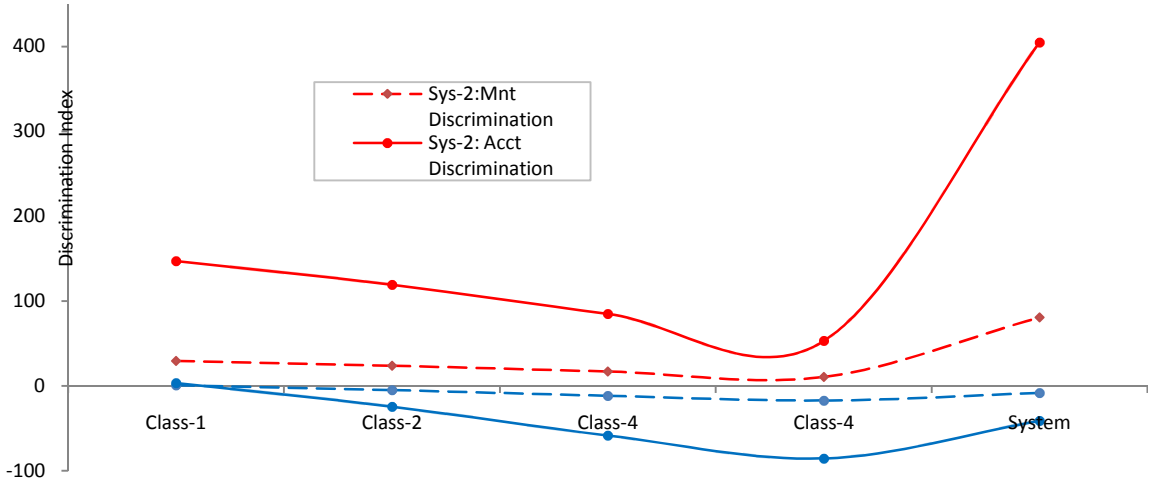

Figure1.6: Comparison of Momentary/Accumulative Discrimination of PPMC & Alternative system

3.2.5. The System Unfairness Coefficient:

By equation (2.2.2), the probability that there are $k = 28.8507$ trucks in each of the system:$P_{k_1} = P_{k_2} = P_{k_3} = P_{k_4} = 0.01254$ (1.254%), and the probability that there are a total of 115.4028trucks in the four systems: $P_{k_n} = 0.007$ (0.7%).Therefore, by equation (2.21) and (2.2.8) the system unfairness and class discrimination given that a class-j customer meets at least k class-j trucks in the system: $E[D^2]$ and $E[\tilde{D}_{(u)}]$ respectively, can be represented by table 1.5and 1.6 below.

Table 1.2: Class Unfairness With Respect To System Unfairness

No of Customer:	Zero Trucks	$k_1 = 28.851$	$k_2 = 28.851$	$k_3 = 28.851$	$k_4 = 28.851$	$k_n = 115.403$
Prob. of k-Trucks:	0.0335	0.0125	0.0125	0.0125	0.0125	0.007
$E[D_{1,k}^2] = D_1^2 P_k$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[D_{2,k}^2] = D_2^2 P_k$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[D_{3,k}^2] = D_3^2 P_k$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[D_{4,k}^2] = D_4^2 P_k$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[D^2]$	**2.5312**	**7.1428**	**7.1428**	**7.1428**	**7.1428**	**4.00**

From table 1.2 above, the total system unfairness given that there are no arriving customer but there are k numbers of class-j trucks in the system: $E[D_1^2] = E[D_2^2] = E[D_3^2] = E[D_4^2] = 3.4185$. Similarly, from table 1.6 below, the system class discrimination given that there are no arriving customers but there are k numbers of class-j trucks in the system: $E[\tilde{D}_{(1)}] = E[\tilde{D}_{(2)}] = E[\tilde{D}_{(3)}] = E[\tilde{D}_{(4)}] = 3.4185$.

Table 1.3: Class Discrimination with respect to System Discrimination

No of Customer:	Zero Trucks	$k_1 = 28.851$	$k_2 = 28.851$	$k_3 = 28.851$	$k_4 = 28.851$	$k_n = 115.403$
Prob. of k-Trucks:	0.0335	0.0125	0.0125	0.0125	0.0125	0.007
$E[\tilde{D}_{(1,k)}] = \lambda_1[D_1 P_k]$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[\tilde{D}_{(2,k)}] = \lambda_2[D_2 P_k]$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[\tilde{D}_{(3,k)}] = \lambda_3[D_3 P_k]$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[\tilde{D}_{(4,k)}] = \lambda_4[D_4 P_k]$	0.6328	1.7857	1.7857	1.7857	1.7857	1.00
$E[\tilde{D}_{(u)}]$	**2.5312**	**7.1428**	**7.1428**	**7.1428**	**7.1428**	**4.00**

3.2.6. Comparison of System Unfairness of the PPMC System with the Alternative System:

The Figure 1.7 below shows the variability of system unfairness index with customers' classes in the two systems. Comparatively, the blue curve shows that the PPMC service policy is more positively unfair to classes with low priority service (smaller service requirements) than classes with high priority service (high service requirements). While in the alternative system, the red curve shows that the system unfairness are marginally distributed across all classes of customers in the system irrespective of service requirements. By these characteristics, the alternative system i.e. assigning each customer's class to a server or a set of servers associated with a single FCFS queue policy is fairer than the PPMC system. Thus, with the mean system unfairness index of +4, the alternative architecture guarantees a fairer petroleum products distribution to the respective customers' classes irrespective of service requirements.

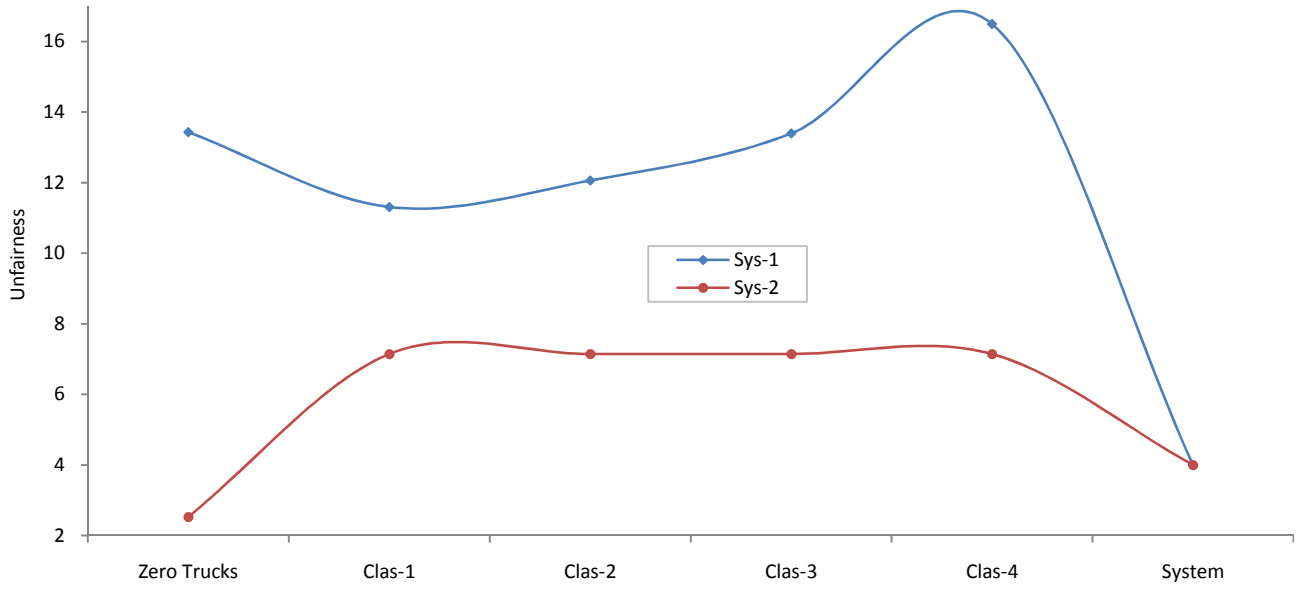

Figure1.7: Comparison of System Unfairness of PPMC and Alternative systems

3.2.7. Comparison of Class Discrimination of the PPMC System with the Alternative System:

The Figure 1.8 below shows the variability of class discrimination index with customers' classes in the two systems. Comparatively, the red curve shows that the PPMC service policy is more negatively discriminative to classes with low

priority service (smaller service requirements) than classes with high priority service (high service requirements). While in the alternative system, the blue curve shows that the discrimination coefficients are marginally distributed across all classes of customers in the system irrespective of service requirements.

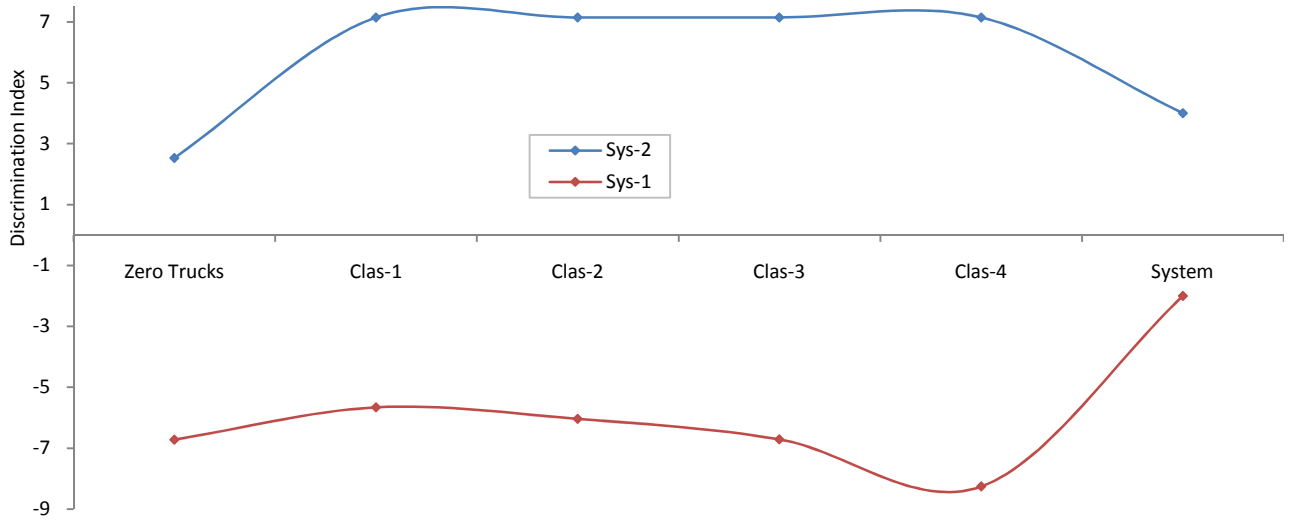

Figure1.8: Comparison of Class Discrimination of PPMC and Alternative systems

By the above characteristics, the alternative system i.e. assigning each customer's class to a server or a set of servers associated with a single FCFS queue policy is fairer than the PPMC system. Thus, with the mean system discrimination coefficient of +4, the alternative architecture guarantees a fairer petroleum products distribution to the respective customers' classes irrespective of service requirements.

4. RESULT OF ANALYSIS

Given the policy of prioritizing services in its multi-server single-queue system to four major socio-economic and geo-political classes of customers within its catchment areas, and the consequential perennial scarcity of these petroleum products during festive season, we invoke the RAQFM analytical framework to study and examine the viability of the PPMC queuing system. Thereby, addressing the all-important question *"how fair is the customer classification and service prioritization policy of the PPMC system in term of the quality and quantity of services rendered to its teaming customers?.*

Tentatively, from the individual class discrimination and the unfairness of specific scenario, to the overall system or policy unfairness, the result of the analysis shows that the PPMC's services prioritization base on socio-economic and geo-political considerations, grossly violate the fundamental principle of queue fairness both in terms of quality and quantity of services render to its teaming customers. This is evident, as the system isn't only highly unfair and negatively discriminative to customers' classes with smaller service requirements (low priority classes), but also marginally discriminate against even the higher priority classes (See figure 1.2&1.3).

A summary statistics also shows that the system does not only discriminate customers' classes to the tune of 169.354trucks deficit per week, but also aggravate the mean sojourn and response times of low priority classes (smaller service requirements)much longer than customers' class with larger service requirements (high priority)(see figure 1.1). Hence, the PPMC prioritization of service based on socio-economic and geo-political considerations rather than job seniority or service requirement difference is unfair, discriminative and not justify; as this policy contradicts both the RAQFM as well as the fundamental principle of queue fairness - service requirement preference principles.

Perhaps, this may have accounted for the perennial scarcity and large deficits of petroleum products supply and allocations to major and private independent marketers outside the Abuja zones and the neighboring states of Kaduna, Nassarawa, Kogi, Niger, etc. during festive seasons. Thus, the recurrent long queues of commuters often associated with this highly discriminative and unfair queuing system of the PPMC. To mitigate against this perennial drawback, the alternative setup proposed in this study may not only help to guarantee positive class discrimination and system fairness to all classes of customers irrespective of service requirements or job seniority differences (see figure 1.6-1.8) but also reduces customers' sojourn and response times to the barest minimum as well as guaranteeing surplus petroleum products allocation to marketers irrespective of their socio-economic or geo-political classes. Therefore, in a multi-class multi-server system, if the total granted service rate to the customers is the same across all classes irrespective of their job seniority or service requirement differences, then resources dedication or dedicating separate server(s) or set of server(s) to each class of customer/job served under a single FCFS queue policy is fairer and justified, otherwise the class prioritization must base on job seniority or service requirement preference principles.

Conclusion

To study the viability of the PPMC queuing system in terms of the quality and quantity of services rendered to its teaming customers, the study presents a strategic RAQFM analytical framework for assessing queue fairness in multi-class multi-server queuing system; where service prioritization is based on service requirement and job seniority differences. As a practical application of RAQFM,

the present study analyzes4-class M/M/4 single-queue architecture of the PPMCSuleijain comparison to the resource dedication service policies alternative. Tentatively, the result of the analyses shows that the expected positive discrimination (system unfairness) of the highest priority classes of customers decreases in its service length, while those of the low priority classes increases in its service length. That is classes with lager service requirements (high priority)are always marginally discriminated, while those with smaller service requirement(lower priority)are always highly discriminated from the individual discrimination, the unfairness of specific scenario to the overall system or policy unfairness.

Hence, prioritization of jobs classes base on socio-economic and geo-political considerations is not justify, and contradict the fundamental principles of queue fairness as well as the RAQFM principle, as classes of short jobs which have arrived the system early may have to wait for eternity for the completion of many classes of long jobs that arrive behind them. This is possibly an unfair treatment by the system and a violation of both jobs seniority and service requirement preference principles. Therefore, since all classes of jobs arriving at the system require equal or proportional service time or the servers' resources, then dedicating separate server or set of server(s) to each class of customer/job served under a single FCFS queue policy is justified and fair, otherwise the service prioritization must based on job seniority (arrival time) or service requirement differences.

Furthermore, addressing the issue of class discrimination that often arise in the management of a multi-class multi-server single-queue architectures, our analytical result shows that: (i.) the (weighted) value of class discrimination is always bounded by the system unfairness, that is, a class cannot be highly discriminated if the overall system unfairness is low, (ii) in the preemptive priority variant, the highest priority class may not always enjoy positive discrimination if service prioritization is bases on service requirements difference, except consideration is also given to customers' arrival time (seniority difference).

To address the conflict of class discrimination and the general system or policy unfairness characteristics that often arise in system whose prioritization is not based on any of the fundamental principles of queue fairness, the study recommended resource dedication policy as a fairer alternative. Comparatively, the alternative setup may not only guarantee positive class discrimination as well as zero unfairness index but also ensure that all customer classes share equal sojourn and response time irrespective of their socio-economic and geo-political class. The results derived in this work can serve for two purposes. First, the simpler results, which might sound intuitive to many researchers, can be used to build confidence in the RAQFM model, queue fairness and class discrimination metrics. These concepts of course are very new to the queuing theory world and require examination and trust building. Second, once the confidence is built, 'the RAQFM model can be used to evaluate and study systems where the results may not be explicit.

References

[1] Adam Wierman (2011): Fairness and scheduling in single server queues. Surveys in Operations Research and Management Science, Elsevier B.V. Science Direct, Vol 16, Issue 1, January 2011, pp 39-48.

[2] AlexandreBrandwajn, Thomas Begin (2017): Multi-server preemptive priority queue with general arrivals and service times. Performance Evaluation, Elsevier, Sept, 2017, 10.1016/j.peva.2017.08.003. hal01581118.

[3] Alex Stone (2012): Why Waiting Is Torture. Aug. 19, 2012, Section SR, Page 12 of the New York Time, https://www.nytimes.com/2012/08/19/opinion/sunday/why-waiting-in-line-is-torture.html

[4] Aristotle (1962): *Nicomachean Ethics*. Liberal Arts Press, Indianapolis, 1962. M. Oswald, Trans. 23.

[5] Benjamin Avi-Itzhak and Hanoch Levy (2016): On measuring fairness in queues. Published online by Cambridge University Press: **01 July 2016**

[6] Dimitrios Serpanos, Tilman Wolf (2011): Quality of service and security. *Architecture of Network Systems, 2011.* https://www.sciencedirect.com/science/article/pii/B9780123744944000104

[7] Dudin, S.; Dudina, O.; Samouylov, K.; Dudin, A. (2020): Improvement of fairness of non-preemptive priorities in transmission of heterogeneous traffic. Mathematics 2020, 8, 929.

[8] Hanoch Levy, Benjamin Avi-Itzhak, and David Raz (2011): Principles of Fairness Quantification in Queueing Systems. Network Performance Engineering. D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 284–300, 2011. Springer-Verlag Berlin Heidelberg 2011

[9] Harchol-Balter, M. Schroeder, B. Bansal, N., and Agrawal, M. (2003): Size-based scheduling to improve web performance. *ACM Transactions on Computer Systems*, 21(2):207-233, May 2003. 2.

[10] Hideaki Takagi (2014): Waiting Time in the M/M/M FCFS Non-preemptive Priority Queue With Impatient Customers. International Journal of Pure and Applied Mathematics Volume 97 No. 3 2014, 311-344. http://www.ijpam.eu

[11] https://ppmc.nnpcgroup.com/AboutUs/Pages/Default.aspx

[12] https://bpe.gov.ng/pipelines-and-products-marketing-company-limited-ppmc/

[13] Jenny Erlichman and RefaelHassin (2011): Strategic overtaking in a monopolistic M/M/1 queue. School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel,

[14] Karl Sigman (2013): Processor sharing queues. http://www.columbia.edu/~ks20/4404-Sigman/4404-Notes-PS.pdf

[15] *Kirill Tšernov* (2020): The Psychology of Queuing As a Key to Reducing Wait Time.

https://www.qminder.com/queue-psychology-reduce-time/

[16] Kleinrock, L. (1975): Queuing Systems, Volume 1: Theory, ed. John Wiley & Sons, New York, 1975,

[17] Kleinrock, L. (1976): Queuing Systems, Volume 2: Computer Applications. Wiley, 1976.

[18] Klimenok, V.; Dudin, A.; Dudina, O.; Kochetkova, I. (2020): Queuing System with Two Types of Customers and Dynamic Change of a Priority. Mathematics 2020, 8, 824.

[19] Little, J. D. C. (1961): A proof of the queuing formula $l = \lambda w$. *Operations Research*, 9: 380-387, 1961. 131.

[20] Mathew, T. V. (2014): Queuing Analysis. Transportation Systems Engineering, IIT Bombay, February 19, 2014.

[21] *Michael Burke and* Dr. Garrett Bomba *(2019):* The Psychology of Waiting and How It Affects Urgent Care Patients. https://www.experityhealth.com/resources/the-psychology-of-waiting-and-how-it-affects-urgent-care-patients/

[22] Mona Khechen (2013): Social Justice: Concepts, Principles, Tools And Challenges. Economic and Social Commission for Western Asia (ESCWA), Distr. LIMITED E/ESCWA/SDD/2013/Technical Paper.9 16 December 2013

[23] Olga Nikolić and Igor Cvejić (2017): Social Justice and the Formal Principle of Freedom. UDK: 316.34 https://doi.org/10.2298/FID1702270N Original scientific article Received: 15.5.2017 — Accepted: May 2017

[24] Perry Kuklin (2019): How Wait Times Impact Customer Behavior and Queue Management. Last update May 2019. https://www.lavi.com/en/resources-detail/customer-behavior-queue-management

[25] Rafaeli A., Kedmi E., Vashdi D., & Barron C. (2005): Queues and Fairness: A Multiple Study Investigation. Technical Report of the Faculty of Industrial Engineering and Management Technology, Haifa, Israel, 2005.

[26] Rafaeli A., Barron G., and K. Haber (2002): The Effects of Queue Structure on Attitudes. Journal of Service Research, 5(2):125-139, 2002.

[27] Rawls, J. (1971): *A Theory of Justice*. Harvard University Press, Cambridge MA, 1971. 24.

[28] Raz D., Avi-ltzhak B., & Levy H. (2004): Classes, Priorities and Fairness in Queuing systems. Technical Report RRR-1 1-2005, RUTCOR, Rutgers University, June 2004.

[29] Raz D., Levy H. and Avi-Itzhak, (2007): A resource allocation queuing fairness measure: properties and bounds.https://link.springer.com/article/10.1007/s11134-007-9025-x, Published: 13 June 2007

[30] Rothkopf, M. H., and Rech, P. (1987): Perspectives on queues: Combining queues is not always beneficial. *Operations Research*, 35:906-909, 1987. 2, 98.

[31] Sarah L. Harris, David Money Harris (2016): Microarchitecture*., Digital Design and Computer Architecture, 2016.* https://www.sciencedirect.com/science/article/pii/B9780128000564000078

[32] S.D. Ayuningtyas and N. Binatari (2017): An analysis on Kendall Lee queuing system with non-preemptive priority at BRI Ahmad Dahlan Yogyakarta. : Journal of Physics: Conf. Series 943 (2017) 012018.

[33] Sztrik, J. (2012): Basic Queuing Theory. University of Debrecen, Faculty of Informatic.http://irh.inf.unideb.hu/user/jsztrik/education/05/3f.html.

[34] ValentinaKlimenok, Alexander Dudin and Vladimir Vishnevsky (2020): Priority Multi-Server Queuing System with Heterogeneous Customers. Mathematics 2020, 8, 1501; doi:10.3390/math8091501 www.mdpi.com/journal/mathematics

[35] Woo-SungKim and Dae-EunLim (2016): Analysis of overtaking in M/M/c queues. Elsevier B.V. Science Direct, Vol. 101, November 2016, Pages 177-183