Report 9?-027

Effects of Different Priority Policies on the Capacity Design for Multiclass Queues

E. L. Örmeci

Effects of Different Priority Policies on the Capacity Design for Multiclass Queues*

E. Lerzan Ormeci, Apostolos Burnetas and Hamilton Emmons

September 10, 1999

Abstract

We consider an $M/M/c$ queue with two classes of customers. We examine the effect of different priority policies on the optimal number of servers to minimize the long-run expected average server and holding costs of the system. Three different policies are considered; the non-idling non-preemptive and preemptive policies as well as the idling policy which minimizes the long-run expected average holding cost of the system in the class of non-preemptive policies. We show that their effect on the optimal number of servers is not significant for many real life systems. This conclusion allows us to consider the design problem under any of the above policies, which introduces significant computational simplification for the system under the idling policy.

1 Introduction

In a multiserver queueing system with multiple customer classes, several priority rules for scheduling service of customers can be employed. Depending on the system modeled, some rules may be more appropriate than others. For example, preemptive priority policies are not feasible in service systems such as restaurants, or in production systems where service involves a chemical process that cannot be interrupted. However they may be suitable in telecommunications (packet switching etc.).

In this paper we examine how different priority control policies affect the design of service capacity of an $M/M/c$ queue with infinite waiting room and two classes of customers. We consider three types of priority policies. Under the first policy, class-1 customers have preemptive priority. Under the second policy, class-1 customers have non-preemptive priority. Both policies are non-idling. The third type of policy, on the other hand, is a non-preemptive threshold policy which gives service to class-2 customers only if there are at least a certain number of class-2 customers in the system. Thus, it allows for some servers to remain idle while class-2 customers are present in the queue, in anticipation of future arrivals of high priority class-1 customers. Specifically, the idling policy is the non-preemptive policy which minimizes the long-run expected average cost of the system. Idling allows for more favorable treatment of the high priority class than the standard non-preemptive rule and can be used in situations where the preemptive rule is not feasible. We denote the three policy types described above by the letters N, P, I.

*This research was supported by a summer grant from the Research Committee of the Weatherhead School of Management, Case Western Reserve University.

For customers of class i, $i = 1, 2$, let λ_i denote the Poisson rate of arrivals, μ_i the exponential service rate and h_i a holding cost per unit of time for each customer. We assume $\mu_1 = \mu_2 = \mu$ and $h_1 \geq h_2$. In this case, class 1 should be the high priority class with respect to minimizing holding costs. Let $g^\pi(c)$ denote the expected average holding cost in a c-server system under priority policy π. We incorporate the priority policy into the problem of service capacity design as follows. Let B denote the cost per unit time for each server. Cost B will be referred to as the server cost and it may represent the purchase cost distributed over the lifetime of the machine in addition to maintenance and operating costs. Alternatively, it may be the annual cost over the machine's lifetime computed by straight-line depreciation, assuming that a server is repeatedly replaced by an identical one at the end of its lifetime. The objective of the capacity design problem under policy π is to determine the optimal number of servers to minimize $G^\pi(c) = g^\pi(c) + Bc$.

For policies N and P, convexity of $G^\pi(c)$ in c is a natural consequence of existing results, therefore efficient procedures can be employed to identify the optimal number of servers. However, showing the convexity of $G^I(c)$ requires comparison of the value functions for different numbers of servers, each of which is an output of an infinite state space MDP. We use $G^N(c)$ and $G^P(c)$ to draw conclusions for policy I, instead of analyzing $G^I(c)$ directly.

We analyze an upper bound, Δg^{NP}, on the difference of average costs due to different priority policies. We show that Δg^{NP} is increasing in the traffic load, but stays finite even under heavy traffic. It is also proved to be decreasing in the number of servers and $\Delta g^{NP} \to 0$ as the number of servers tends to ∞. We also derive an upper bound on the maximum difference within the optimal number of servers under P, N, I. These results in conjunction with extensive numerical experiments show that the difference due to different priority rules is not significant.

The following section examines the sensitivity of $G^\pi(c)$ in priority rules, P, N, and I. Section 3 provides an upper bound on the maximum difference within the optimal number of servers under policies P, N, and I. Both Section 2 and 3 include numerical examples that support their analysis and indicate the effect of different variables on the optimal design. In Section 4, we present the conclusion and possible future research areas.

2 Sensitivity of average purchase and holding costs in the priority rule

The system under policy P or N can be modeled as a continuous time Markov reward process (MRP) and under policy I as a Markov decision process (MDP). Xu, Righter & Shantikumar (1992) uses the MDP formulation for policy I to show that policy I is characterized by numbers l_i^c, $i = 0, ..., c-1$, such that whenever there are i busy servers out of c servers, the number of class-2 customers waiting in line must be greater than l_i^c in order to start the service of one class-2 customer. In addition, l_i^c is increasing in i, for all $i = 0, 1, ..., c-1$ and $l_c^c = \infty$. Policies P and N are static control policies since they are described independent of the current state of the system, whereas policy I is dynamic because of the thresholds which depend on the current state. Regarding the relations among these policies, it is easy to see intuitively that $g^P(c) \leq g^I(c) \leq g^N(c)$ for all c: Policy I is the optimal scheduling policy in the class of all non-preemptive policies, so it performs better than policy N when the number of servers is fixed. Thus, we conclude that $g^I(c) \leq g^N(c)$ for all c. Intuitively, it is also easy to see that the preemptive policy performs better than any other scheduling policy, and so better than policy I. Ormeci (1998) uses the MDP/MRP formulation to show this formally.

The objective of the design problem is to find the optimal number of servers, c_π^*, which minimizes total purchase and holding costs per unit time under policy π, $G^\pi(c) = g^\pi(c) + Bc$, i.e., to determine c_π^* such that

$$G^\pi(c_\pi^*) = \min_{c \geq c_{min}} \{g^\pi(c) + cB\}, \tag{1}$$

where $c_{min} = \min\{c : \lambda < c\mu\}$ and $\lambda = \lambda_1 + \lambda_2$. The constraint $\lambda < c\mu$ ensures that the system is stable so that $g^\pi(c) < \infty$, for $\pi = P, N, I$. In this section, we consider the effect of different priority rules on total costs (holding and purchase costs) per unit time. In particular, we are interested in the quantity $G^N(c_N^*) - G^P(c_P^*)$, the maximum difference in the minimum total costs due to different policies. We observe that

$$0 \leq G^N(c_N^*) - G^P(c_N^*) \leq G^N(c_N^*) - G^P(c_P^*) \leq G^N(c_P^*) - G^P(c_P^*) = g^N(c_P^*) - g^P(c_P^*) = \Delta g^{NP}(c_P^*),$$

where $\Delta g^{NP}(c) = g^N(c) - g^P(c)$. The inequalities follow either from the definition of c_π^* or from the fact that $g^P(c) \leq g^N(c)$ for all c. Thus, $G^N(c_N^*) - G^P(c_P^*)$ is non-negative and bounded from above by the difference $\Delta g^{NP}(c_P^*)$. Although the optimal number of servers under each policy depends on the machine cost rate, B, the sensitivity of total average cost can be measured by Δg^{NP}, which is independent of the machine cost rate, B. Δg^{NP} also provides the maximum difference in the average holding costs due to priority rules for fixed number of servers. Therefore, we devote this section to the analysis of Δg^{NP} with respect to the traffic load, $r = \frac{\lambda_1+\lambda_2}{\mu}$, as well as the number of servers. We show that Δg^{NP} is increasing in r, but has a finite limit when $r \to c$. Also Δg^{NP} is proved to be decreasing in the number of servers, c, and $\lim_{c\to\infty} \Delta g^{NP}(c) = 0$. The implications of these results are discussed in the end of the section through numerical examples.

To show these results, we use the work conservation law, which is a formalization of the intuitive property that one class of customers can be treated more favorably only at the expense of other classes (see e.g., Kleinrock (1976)). This law applies only to work-conserving queues. The system under policy I, which allows creation of work through idling, is not work conserving, whereas it is well-known that the nonpreemptive nonidling policy is work conserving. The nonidling preemptive policy also induces a work-conserving system because the service times for each class of customers follow an exponential distribution, although it is not work conserving under general service time distributions.

When the service times of all classes follow exponential distributions, the work conservation law reduces to the conservation of the total number of customers in the system under all work conserving policies, i.e., the expected numbers of total customers in all these systems are the same. Let $L_i^\pi(r, r_1, c)$ be the expected number of class-i customers in the system with c servers and total traffic load $r = \frac{\lambda_1+\lambda_2}{\mu}$, of which $r_1 = \frac{\lambda_1}{\mu}$ is due to class-1 customers, under policy π, for $i = 1, 2$ and $\pi = P, N$ and $L^F(r, c)$ the expected number of customers in an $M/M/c$ system under first come first served discipline. Then

$$L_1^N(r, r_1, c) + L_2^N(r, r_1, c) = L_1^P(r, r_1, c) + L_2^P(r, r_1, c) = L^F(r, c) \tag{2}$$

We note that L_i^π, and so g^π, depends not only on the number of servers, but also on the total traffic load, r, and the traffic load with respect to class-1 customers, r_1. For notational simplicity, only the dependence on one variable may be indicated in the sequel. For $\pi = P, N$, $g^\pi(c)$ can be expressed as:

$$g^\pi(c) = h_1 L_1^\pi(c) + h_2 L_2^\pi(c),$$

and from (2),

$$g^\pi(c) = h_1 L_1^\pi(c) + h_2(L^F(c) - L_1^\pi(c)) = (h_1 - h_2)L_1^\pi(c) + h_2 L^F(c). \tag{3}$$

Define $\Delta L_1^{NP}(c) = L_1^N(c) - L_1^P(c)$ as the difference in the average number of class-1 customers between the nonpreemptive and preemptive policies. From equation (3), we have:

$$\Delta g^{NP}(c) = g^N(c) - g^P(c) = (h_1 - h_2)\Delta L_1^{NP}(c). \tag{4}$$

We have the following expressions for L_1^π (see Buzen & Bondi (1983) for $\pi = P$ and Gross & Harris (1974) for $\pi = N$):

$$L_1^P(r, r_1, c) = r_1 + \frac{r_1}{c - r_1}\mathcal{D}(r_1, c) \tag{5}$$

$$L_1^N(r, r_1, c) = r_1 + \frac{r_1}{c - r_1}\mathcal{D}(r, c) \tag{6}$$

where

$$\mathcal{D}(r, c) = \frac{r^c}{(1 - \frac{r}{c})c!}p_0(r, c) \quad \text{and} \tag{7}$$

$$p_0(r, c) = \left(\sum_{n=0}^{c-1}\frac{r^n}{n!} + \frac{r^c}{c!(1 - \frac{r}{c})}\right)^{-1} \tag{8}$$

The quantity $\mathcal{D}(r, c)$, also known as the Erlang delay formula, is equal to the probability that all servers are busy in an $M/M/c$ system with load r and $p_0(r, c)$ is the probability of having no customers in the system with a traffic load of r and c servers. Thus, by equations (5) and (6):

$$\Delta L_1^{NP}(r, r_1, c) = \frac{r_1}{c - r_1}\left(\mathcal{D}(r, c) - \mathcal{D}(r_1, c)\right). \tag{9}$$

Consider the behavior of Δg^{NP} as a function of the traffic load r. The following theorem shows that Δg^{NP} is increasing in r, but has a finite limit when $r \to c$. In this theorem, as the load of the system increases, the load due to class-1 customers increases with the same rate, so that $r_1 = r\gamma$ with $0 < \gamma < 1$; therefore $\gamma 100\%$ of the arrivals are due to class-1 customers. However, the theorem remains valid if r_1 is fixed.

Theorem 1 *(i)* $\Delta g^{NP}(r, r\gamma, c)$ *is increasing in* r.

(ii) $\lim_{r \to c} \Delta g^{NP}(r, r\gamma, c) < \infty$.

Proof From equation (4), it is sufficient to consider ΔL_1^{NP}.

(i) Take the partial derivative of $\Delta L_1^{NP}(r, r\gamma, c)$ with respect to r. By equation (9):

$$\frac{\partial \Delta L_1^{NP}(r, r\gamma, c)}{\partial r} = \frac{\gamma c}{(c - r\gamma)^2}\left(\mathcal{D}(r, c) - \mathcal{D}(r\gamma, c)\right) + \frac{r\gamma}{c - r\gamma}\left(\mathcal{D}'(r, c) - \gamma\mathcal{D}'(r\gamma, c)\right).$$

In Lee & Cohen (1983), it is shown that $\mathcal{D}(r, c)$ is nondecreasing and convex in r. Thus $\mathcal{D}(r, c) \geq \mathcal{D}(r\gamma, c)$ by monotonicity of $\mathcal{D}(r, c)$ in r and $\mathcal{D}'(r, c) \geq \mathcal{D}'(r\gamma, c)$ by convexity of $\mathcal{D}(r, c)$ in r; since also $\gamma < 1$, these imply that $\frac{\partial \Delta L_1^{NP}}{\partial r} \geq 0$.

(ii) Equations (9) and (7) imply:

$$\Delta L_1^{NP}(r, r\gamma, c) = \frac{r\gamma}{(c - r\gamma)c!}\left(\frac{r^c}{1 - \frac{r}{c}}p_0(r, c) - \frac{(r\gamma)^c}{1 - \frac{r\gamma}{c}}p_0(r\gamma, c)\right). \tag{10}$$

First consider $\lim_{r\to c} h(r)$, where $h(r) = \frac{p_0(r,c)}{1-\frac{r}{c}}$. Because both $p_0(r,c)$ and $1-\frac{r}{c}$ converge to 0 as $r \to c$, we can use L'Hospital's rule:

$$\lim_{r\to c} h(r) = \lim_{r\to c} \frac{p_0'(r,c)}{-\frac{1}{c}} = -c \lim_{r\to c} p_0'(r,c)$$

where $p_0'(r,c) = \partial p_0(r,c)/\partial r$. Now consider $-p_0'(r,c)$:

$$\begin{aligned}
\lim_{r\to c}[-p_0'(r,c)] &= \lim_{r\to c} \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!(1-\frac{r}{c})}\right)^{-2} \left[\sum_{n=0}^{c-2} \frac{r^n}{n!} + \frac{r^c}{c!(1-\frac{r}{c})}\frac{c}{r} + \frac{r^c}{c!(1-\frac{r}{c})^2 c}\right] \\
&= \lim_{r\to c} p_0(r,c) + \lim_{r\to c} \frac{r^c}{c!(1-\frac{r}{c})^2 c} \left(\frac{c!(1-\frac{r}{c})\sum_{n=0}^{c-1}\frac{r^n}{n!} + r^c}{c!(1-\frac{r}{c})}\right)^{-2} \\
&= \lim_{r\to c} \frac{r^c (c-1)!}{\left(c!(1-\frac{r}{c})\sum_{n=0}^{c-1}\frac{r^n}{n!} + r^c\right)^2} = \frac{(c-1)!}{c^c},
\end{aligned}$$

because $\lim_{r\to c} p_0(r,c) = 0$ and $\lim_{r\to c} c!(1-\frac{r}{c})\sum_{n=0}^{c-1}\frac{r^n}{n!} = 0$. Thus:

$$\lim_{r\to c} h(r) = \frac{c!}{c^c}.$$

Therefore, from (10),

$$\lim_{r\to c} \Delta L_1^{NP}(r, r\gamma, c) = \frac{\gamma}{1-\gamma}\left(1 - \frac{(c\gamma)^c}{c!(1-\gamma)} p_0(c\gamma, c)\right) < \infty. \tag{11}$$

□

We want to establish the monotonicity and asymptotic behavior of Δg^{NP} with respect to the number of servers, c. First, we need to consider the behavior of $\mathcal{D}(r,c) - \mathcal{D}(r_1,c)$ with respect to the number of servers: c:

Lemma 1 *$\mathcal{D}(r,c) - \mathcal{D}(r_1,c)$ is nonincreasing in c.*

Proof See the Appendix. □

Now, we present Theorem 2, which examines Δg^{NP} as the number of servers, c, varies.

Theorem 2 *(i) $\Delta g^{NP}(r, r_1, c)$ is nonincreasing in c.*

(ii) $\lim_{c\to\infty} \Delta g^{NP}(r, r_1, c) = 0$.

Proof By equation (4), it suffices to consider ΔL_1^{NP} only.

(i) From equation (9):

$$\begin{aligned}
\Delta L_1^{NP}(r, r_1, c+1) &= \frac{r_1}{c+1-r_1}\left[\mathcal{D}(r, c+1) - \mathcal{D}(r_1, c+1)\right] \\
&\le \frac{r_1}{c-r_1}\left[\mathcal{D}(r, c) - \mathcal{D}(r_1, c)\right] \\
&= \Delta L_1^{NP}(r, r_1, c),
\end{aligned}$$

where the inequality follows from Lemma 1. Thus, $\Delta L_1^{NP}(r, r_1, c)$ is nonincreasing in c.

(ii) Recall that ΔL_1^{NP} is, by definition, equal to the difference in the number of class-1 customers between the nonpreemptive and preemptive policies. When $c \to \infty$, the behavior of class-1 customers under both policies approaches that of an $M/M/\infty$ system with traffic intensity r_1, since all incoming customers find at least one idle server. Thus $\lim_{c\to\infty} L_1^N(r, r_1, c) = \lim_{c\to\infty} L_1^P(r, r_1, c) = r_1$, which proves (ii). □

These results describe the behavior of Δg^{NP} with respect to traffic load, r, and the number of servers, c: Δg^{NP} is finite even under heavy traffic although $g \to \infty$, and $\Delta g^{NP} \to 0$ as $c \to \infty$. We next consider a new quantity $\Delta g^{NP}/g^P$, which can be regarded as an upper bound on the relative error due to different priority policies. Thus, it is a fair measure of the effect of priority policies for fixed c. We first present an approximation of $\Delta g^{NP}/g^P$:

$$\begin{aligned}
\frac{\Delta g^{NP}}{g^P} &= \frac{(h_1 - h_2)\frac{r_1}{c-r_1}(\mathcal{D}(r,c) - \mathcal{D}(r_1,c))}{h_2\left(r + \frac{r}{c-r}\mathcal{D}(r,c)\right) + (h_1 - h_2)\left(r_1 + \frac{r_1}{c-r_1}\mathcal{D}(r_1,c)\right)} \\
&\approx (h_1 - h_2)\frac{r_1}{c-r_1} \times \\
&\quad \frac{\left[2r + (c-r)^2 - (c-r)\sqrt{5r + (c-r)^2}\right]^+ - \left[2r_1 + (c-r_1)^2 - (c-r_1)\sqrt{5r_1 + (c-r_1)^2}\right]^+}{h_1 r_1 + h_2(r - r_1) + h_2\frac{r}{c-r}\left[2r + (c-r)^2 - (c-r)\sqrt{5r + (c-r)^2}\right]^+}
\end{aligned}$$

where $\mathcal{D}(r,c)$ can be approximated as $EB2 = \left[2r + (c-r)^2 - (c-r)\sqrt{5r + (c-r)^2}\right]^+$, using results from Harel (1988). Under low traffic, i.e., when r/c is low, $EB2 \approx 0$, and so $\Delta g^{NP}/g^P \approx 0$. Thus, we need to consider the behavior of $\Delta g^{NP}/g^P$ only under high traffic. This quantity is neither monotone nor convex or concave in r or c. Therefore, instead of analyzing this quantity analytically, we compute $\Delta g^{NP}/g^P$ for 840 different examples to see the effect of different parameters. We set $h = h_1/h_2$ with $h_2 = 1$ and $\mu = 1$ so that $r_1 = \lambda_1$ and $r = \lambda_1 + \lambda_2$. We consider seven different values of h, $h = 2, 3, 5, 7, 10, 20$ and as $h \to \infty$, and six different values of r, $r = 3, 7, 11, 19, 49, 99$. For each fixed r and h, four different values of r_1 and five values of c are considered; $r_1/r = 2/9, 4/9, 6/9, 8/9$ and $c = r + 1, ..., r + 5$.

We observe that $\Delta g^{NP}/g^P$ tends to be higher for low r and high r/c in general. Under this condition, the number of servers is low. Then the service level for class-1 customers under policy N is substantially lower than that under policy P, since the servers do not get idle frequently. Thus the effect of the control policy in this case is more significant. For example, $\Delta g^{NP}/g^P$ has its maximum at 16.8% for the system with $r = 3$, $\gamma = 4/9$ and $c = 4$ as $h \to \infty$, although for more reasonable values of h, $\Delta g^{NP}/g^P$ is also more reasonable; e.g. see Figure 1 which shows the effect of the load, r, on $\Delta g^{NP}/g^P$ for different values of r_1/r with $h = 5$ and $c = r + 1$. An additional server in the systems with low r decreases the traffic (r/c) decreases substantially, which, in turn, causes a considerable decrease in $\Delta g^{NP}/g^P$. For example, when $r = 3$ and $c = 5$, a maximum of 6.1% is observed as $h \to \infty$, which is almost one third of 16.8%, the maximum for $c = 4$. Thus, even with only one more server, the effect of control policy almost vanishes. When r is high, different control policies do not affect the total cost that significantly, since the servers become available more often. For example, when $r \geq 19$ and $h \leq 20$, $\Delta g^{NP}/g^P$ has a maximum of 10%.

We next consider the effect of $h = h_1/h_2$ (see Figure 2): Naturally, as h increases, the service level of class-1 customers affect the total cost more; which increases the effect of control policies. Although $\Delta g^{NP}/g^P$ can be high as $h \to \infty$, it is always less than 9.5% for more realistic values of h, i.e., when $h \leq 5$. Thus, unless h is high, the control policy does not influence the total cost very significantly.

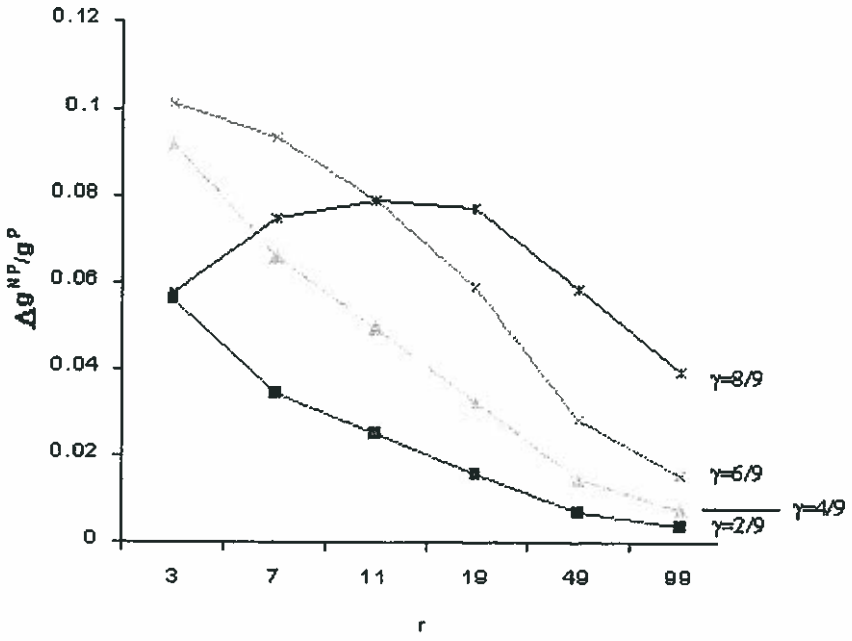

Figure 1: Behavior of $\Delta g^{NP}/g^P$ with respect to r, where $h = 5$ and $c = r + 1$

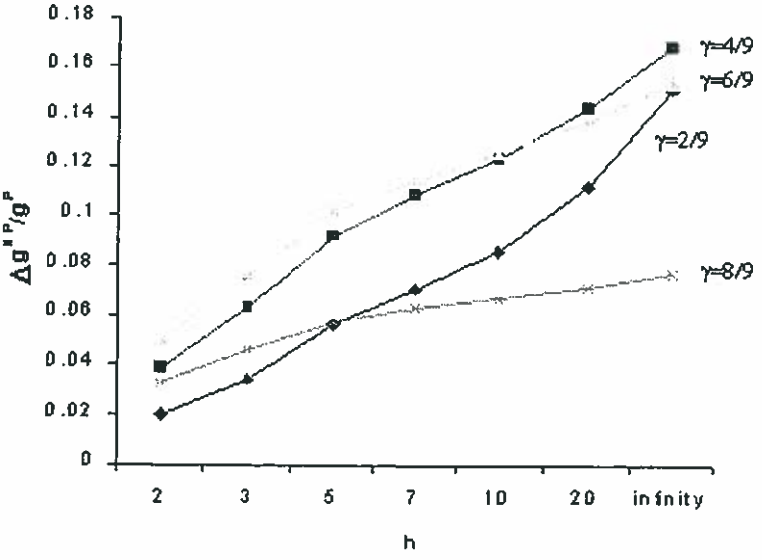

Figure 2: Behavior of $\Delta g^{NP}/g^P$ with respect to h, where $r = 3$ and $c = 4$

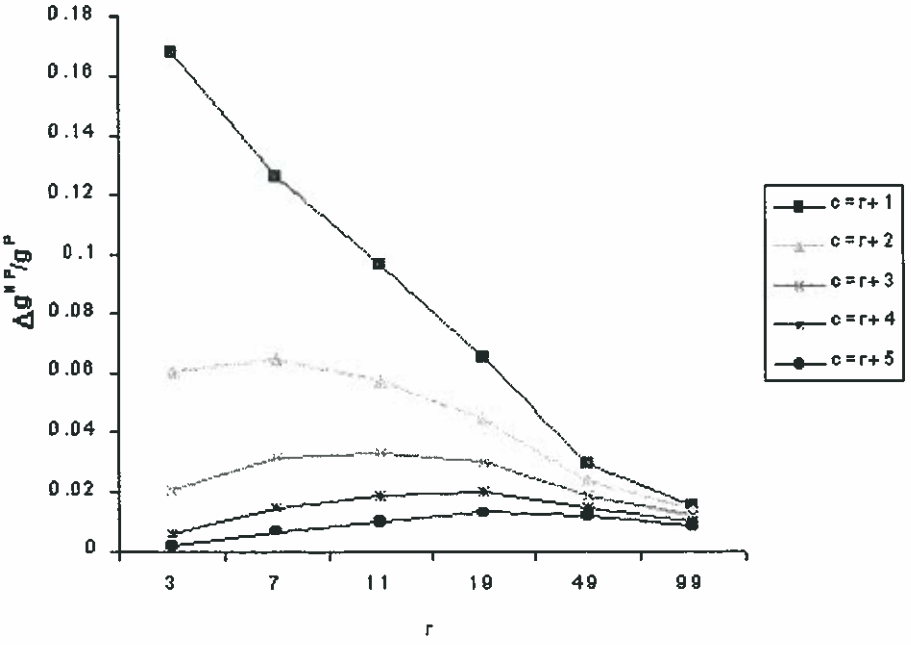

Figure 3: Behavior of $\Delta g^{NP}/g^P$ with respect to r, where $r_1/r = 4/9$ and $h \to \infty$

We observe that the effect of r_1 depends on the value of r: For low r, $\Delta g^{NP}/g^P$ is high when r_1/r is close to 50%, whereas for high r, it is high if r_1/r is also high. In real life situations, r_1/r can generally be expected to be low since class-1 customers are more expensive. When $r_1 = 2/9$ and $h \leq 5$, $\Delta g^{NP}/g^P \leq 5.6\%$, which is really low.

The number of servers is the most effective parameter of all, since we observe that although $\Delta g^{NP}/g^P$ can be high for $c = c_{\min}$, it is less than 5.5% for all r and r_1 when $c \geq c_{\min} + 1$ even for $h \to \infty$ (see Figure 3). Thus, we can conclude that the effect of an additional server is much more substantial than the choice of a control policy. When the load r is low, this can be explained by the drastic decrease in r/c. For low r ($r = 3, 7$), the decrease in $\Delta g^{NP}/g^P$ is usually over 50%; whereas for moderate r ($r = 11, 19$), the decrease is around 30-40%. When r is high ($r = 49, 99$), $\Delta g^{NP}/g^P$ is low even when $c = c_{\min}$, and although the decrease can be around 20% in some cases, it is not that significant in general.

To summarize, we observe that for more realistic values of different parameters, $\Delta g^{NP}/g^P$ is low. The results of this section also support this observation. Under the extreme cases, $\Delta g^{NP}/g^P \to 0$; in particular, when $r \to c$, $\Delta g^{NP} < \infty$ $g^P \to \infty$ and when $c \to \infty$, $\Delta g^{NP} \to 0$ and $g^P > 0$. Hence, we can conclude that for many real life situations the effect of different priority policies is not significant.

3 Optimal number of servers under different control policies

In determining the optimal number of servers under policy π, the convexity of G^π in c is an important property which facilitates the solution process of the design problem. By definition of G^π, it is enough to consider the convexity of g^π. From equation (3), g^π is convex in c if $L^F(c)$ and $L_1^\pi(c)$ are convex. Dyer & Proll (1977) have shown that the average waiting time of a customer in a single class $M/M/c$ system is nonincreasing and convex in c, thus by Little's law, $L^F(c)$ is also nonincreasing and convex in c. For specific π, the following are known in the literature: Under preemptive priority, the behavior

of class-1 customers is not affected by class-2 customers, so that the probability distribution of the number of class-1 customers is the same as that of the total number of customers in an $M/M/c$ system with arrival rate λ_1 and service rate μ (Buzen & Bondi 1983). Therefore, $g^P(c)$ is nonincreasing and convex in c. As for g^N, we first note that Harel (1990) has shown that $\mathcal{D}(r,c)$ is nonincreasing and convex in c. Taking the appropriate differences with respect to c in equation (6), it is easy to show that g^N is also convex in c.

Using convexity of P and N, the optimal number of servers under policy P and N can be found by:

$$c_\pi^* = \min\{c : g^\pi(c) - g^\pi(c+1) \leq B\} = \min\{c : \Delta g^\pi(c) \leq B\}, \tag{12}$$

where $\Delta g^\pi(c)$ is equal to the savings in $g^\pi(c)$ achieved by adding a server to a system with c servers. Note that if the total operating cost of c servers can be represented by an increasing and convex function, f (rather than a linear function), we can still use a similar formula to find c_π^*:

$$c_\pi^* = \min\{c : \Delta g^\pi(c) \leq f(c+1) - f(c)\}.$$

Hence we stop adding servers when the savings in holding costs introduced by an additional server are less than its marginal cost.

The above analysis applies to policies P and N, because the average holding costs under these policies are convex. The situation is more complex for policy I, since it is determined by a set of thresholds, which requires to solve an MDP with an infinite state space. Proving convexity of $g^I(c)$ requires properties on the behavior of threshold levels as the number of servers varies. Although we have not established that $g^I(c)$ is convex in c, we have not seen any counterexamples in our numerical computations. Because of the lack of a convexity proof, we cannot employ efficient procedures which use the first differences to check the global optimality of a current solution, as we could for policies P and N. The design problem in this case is further complicated by the fact that for any c, $g^I(c)$ can only be computed by solving an infinite state space MDP.

In this section, instead of solving the design problem under each priority rule, we concentrate on the difference of optimal number of servers due to different priority rules and derive bounds on this difference. For this, we first establish upper and lower bounds on the optimal number of servers for all policies. This also facilitates the design problem under policy I, since it provides an upper bound on the number of MDPs we need to solve. In fact, it further simplifies the whole problem, since these bounds are very tight for many systems. This, with the conclusions of the previous section, suggests that solving the problem under any of these policies performs good enough in terms of total cost rate. Then, we can solve the problem for policy P or N, which has convex cost structure and closed form solution for g^π, instead of considering the system under policy I.

We use two different approaches to derive an upper bound, one of which also provides a lower bound. Both approaches make use of the convexity of g^N and g^P. The first method uses the difference $\Delta g^{NP}(c) = g^N(c) - g^P(c)$ to derive an upper bound, $\bar{c}^1$, on c_π^*. The second method uses only the convexity of g^N and g^P and the fact $g^N(c) \leq g^I(c) \leq g^P(c)$ to provide a second upper bound $\bar{c}^2$ as well as a lower bound $\underline{c}$.

3.1 Method 1

Recall that $\Delta g^\pi(c)$ is equal to the savings in $g^\pi(c)$ introduced by an additional server, i.e., $\Delta g^\pi(c) = g^\pi(c) - g^\pi(c+1)$. From equation (1), adding a server to a system with c servers results in an immediate

net benefit only when $g^\pi(c) - g^\pi(c+1) > B$. Using the relation $g^P(c) \le g^I(c) \le g^N(c)$ for the c- and $(c+1)$-server systems, we have:

$$\Delta g^\pi(c) \le \bar{\Delta} g(c), \tag{13}$$

where we set $\bar{\Delta} g(c) = g^N(c) - g^P(c+1)$. To determine an upper bound on c_π^*, it suffices to show that $\bar{\Delta} g(c)$ is nonincreasing in c and that there exists a $\hat{c}$ such that $\bar{\Delta} g(\hat{c}) < B$. Indeed these imply that for all $c \ge \hat{c}$, $\Delta g^\pi(c) < B$, therefore $c_\pi^* \le \hat{c}$. To show the two claims, we rewrite $\bar{\Delta} g(c)$ as:

$$\begin{aligned} \bar{\Delta} g(c) &= g^N(c) - g^P(c+1) \\ &= g^N(c) - g^N(c+1) + g^N(c+1) - g^P(c+1) \\ &= \Delta g^N(c) + \Delta g^{NP}(c+1). \end{aligned} \tag{14}$$

$\Delta g^N(c)$ is nonincreasing in c since $g^N(c)$ is convex in c. Also notice that $\Delta g^N(c) \to 0$ as $c \to \infty$. By Theorem 2, $\Delta g^{NP}(c+1)$ is also nonincreasing in c and $\Delta g^{NP}(c+1) \to 0$ as $c \to \infty$. Then, we have the following result:

Corollary 1 *(i) $\bar{\Delta} g(c)$ is nonincreasing in c.*

(ii) $\lim_{c\to\infty} \bar{\Delta} g(c) = 0$.

Part *(ii)* of the corollary guarantees that there exists a $\hat{c}$ with $\bar{\Delta} g(\hat{c}) \le B$, while part *(i)* ensures that $\bar{\Delta} g(c) < B$ for all $c \ge \hat{c}$. Let

$$\bar{c}^1 = \min\{c : \bar{\Delta} g(c) \le B\}.$$

By Corollary 1, $\bar{c}^1$ is well-defined and finite. Inequality (13) implies that $\Delta g^\pi(c) < B$ for all $c \ge \bar{c}^1$, thus $c_\pi^* \le \bar{c}^1$ for $\pi = P, N, I$.

3.2 Method 2

For an alternative method of bounding c_π^*, recall that for $\pi = P, N$, $G^\pi(c)$ is convex in c and c_π^* is well-defined. Define $\bar{c}^2$ and $\underline{c}$ as follows (see Figure 4):

$$\bar{c}^2 = \min\{c \ge c_N^* : G^N(c_N^*) \le G^P(c)\} - 1, \tag{15}$$

$$\underline{c}^2 = \max\{c \le c_N^* : G^N(c_N^*) \le G^P(c)\} + 1, \quad \text{and}$$

$$\underline{c} = \max\left\{\underline{c}^2, c_{min}\right\} \tag{16}$$

where we set $\max\emptyset = -\infty$. Recall that $c_{min} = \min\{c : \lambda < c\mu\}$. Since $G^N(c)$ and $G^P(c)$ are convex in c, $g^\pi(c) - g^\pi(c+1) \to 0$ as $c \to \infty$, $B > 0$ and $G^N(c) \ge G^P(c)$ for all c, quantities $\bar{c}^2$ and $\underline{c}$ are well-defined. We now show that $\bar{c}^2$ and $\underline{c}$ represent bounds for c_π^*, for $\pi = P, N, I$.

Proposition 1 *Let $\bar{c}^2$ and $\underline{c}$ be defined as in (15) and (16). Then*

$$\underline{c} \le c_\pi^* \le \bar{c}^2 \qquad \text{for } \pi = P, N, I. \tag{17}$$

Proof Consider the first inequality in (17). If $\underline{c} = c_{min}$, then the statement is true by definition of c_{min}. Thus, assume that $\underline{c} > c_{min}$. Then, we have:

$$G^P(c_N^*) \le G^I(c_N^*) \le G^N(c_N^*) \le G^P(c) \le G^I(c) \le G^P(c)$$

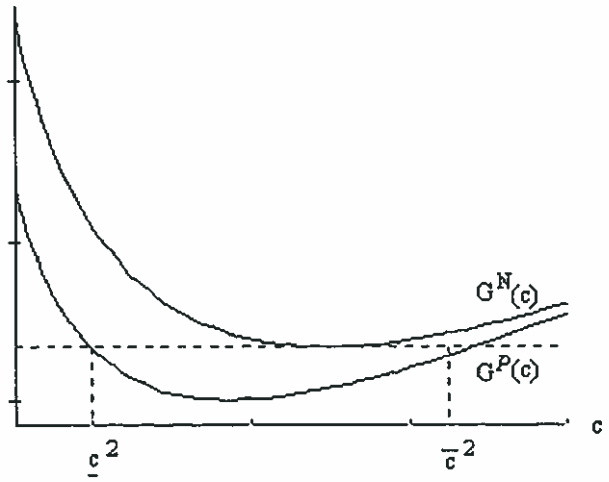

Figure 4: Upper and lower bound on the number of servers using Method 2

where $c < \underline{c}$. The first two and the last two inequalities are true since $G^P(c) \leq G^I(c) \leq G^N(c)$ for all c, the third inequality follows by the definition of $\underline{c}$ and because $c < \underline{c}$. Thus c_N^* performs better than all $c < \underline{c}$ under all policies, therefore the optimal number of servers cannot be smaller than $\underline{c}$ under any of these policies; i.e., $\underline{c} \leq c_\pi^*$, for $\pi = P, N, I$.

The second inequality in (17) is proven similarly. □

Let $\bar{c} = \min\{\bar{c}^1, \bar{c}^2\}$. Now $\bar{\Delta}c = \bar{c} - \underline{c}$ is an upper bound on the difference of optimal number of servers due to different control policies. Analytic expressions on the average or worst case behavior of $\bar{\Delta}c$ are very difficult to establish, due to the complexity of the formulas. Instead, we again present sets of examples. We assume $h = h_1/h_2$, $h_2 = 1$ and $\mu = 1$. As a point of reference, we consider a system with parameters $h = 5$, $B = 10$, $r = 19$ and $r_1/r = 4/9$. In each of the tables below, we present $\bar{c}^1, \bar{c}^2, \underline{c}, \bar{\Delta}c, c_P^*$ and c_N^*, as each of the parameters vary, keeping others constant. The rationale for selecting these particular values are the following: As the holding cost of class-1 customers increases, the gap between different priority rules will also increase. We take $h = 5$, which is both sufficiently large to induce a "large" gap and still realistic. In real systems the machine cost rate is often a higher order magnitude than the holding cost rates. However, a very high machine cost rate, B, would force the design problem under any priority policy to use a "minimal" number of machines. We set $B = 10$ as a compromise. If the traffic load is too high relative to the service rate, then many servers will be needed to serve all the customers, which means that machines will frequently become available and subsequently make the difference between preemptive and non-preemptive policies immaterial. However, under low traffic the effect of one additional server is high since it reduces the traffic intensity of the system to a great extent, so that the difference induced by the priority policies is low when compared with the effect of an additional server (see Table 2). Thus, we consider a moderate size problem with $r = 19$. In any real world system, more expensive jobs (class 1) arrive less often than less expensive ones (class 2). However, if class-1 customers are very rare, then the effect of priority policies cannot be observed at all. Therefore, we assume $r_1/r = 4/9$.

We first consider the effect of the machine cost rate, B. Table 1 presents c_P^*, c_N^*, $\underline{c}$, $\bar{c}^1$, $\bar{c}^2$ and $\bar{\Delta}c$ when $h = 5$, $r = 19$ and $r_1 = 8.44$. When B is high, the machine cost dominates the total cost, and so the optimal number of servers does not vary with the priority policy: when $B \geq 20$, $c_\pi^* = c_{min} = 20$ for all P, N, I, i.e., the system uses the minimal number of machines to serve all the customers. However, even when B is small, the difference on the optimal number of servers is small, e.g., $\bar{\Delta}c = 2$ for $B = 1$.

B	$\bar{c}^1$	$\bar{c}^2$	$\underline{c}$	$\bar{\Delta}c$	c_P^*	c_N^*
1	24	24	22	2	23	23
2	23	23	21	2	22	22
5	21	21	21	0	21	21
10	21	21	20	1	20	20
20	21	20	20	0	20	20
50	21	20	20	0	20	20
100	21	20	20	0	20	20

Table 1: Effect of B when $h = 5$, $r = 19$ and $r_1 = 8.44$.

In Table 2, we consider the effect of r with $h = 5$, $B = 10$ and $r_1 = 4r/9$. Whenever r is relatively small or large, the difference, $\bar{\Delta}c$, is 0. In fact, $\bar{\Delta}c = 1$ only when $r = 19$. Thus, under any traffic load, the bounds on the optimal number of servers are very close.

r	$\bar{c}^1$	$\bar{c}^2$	$\underline{c}$	$\bar{\Delta}c$	c_P^*	c_N^*
3	4	4	4	0	4	4
7	8	8	8	0	8	8
11	12	12	12	0	12	12
19	21	21	20	1	20	20
49	51	51	51	0	51	51
99	102	102	102	0	102	102

Table 2: Effect of r when $h = 5$, $B = 10$ and $r_1 = 4r/9$.

In Table 3, we consider the effect of r_1 with $h = 5$, $B = 10$ and $r = 19$. As the arrival rate of class-1 customers to all customers, r_1, increases, $\bar{\Delta}c$ also increases. However, $\bar{\Delta}c \leq 1$ in all cases.

The effect of h with $r_1 = 8.44$, $B = 10$ and $r = 19$ is given in Table 4, which, again, shows that $\bar{\Delta}c \leq 1$ in all cases. We observe that $\bar{\Delta}c$ increases as h increases; however, it is always small.

The upper and lower bounds on the number of optimal servers are pretty tight. Thus, solving the design problem under either policy N or P, which is relatively easier, is good enough for design purposes. This also eliminates the problem of finding the optimal number of servers under policy I.

γ	$\bar{c}^1$	$\bar{c}^2$	$\underline{c}$	Δc	c_P^*	c_N^*
4.222	20	20	20	0	20	20
8.444	21	21	20	1	20	20
12.667	21	21	20	1	20	21
16.889	21	21	20	1	21	21

Table 3: Effect of r_1 when $h = 5$, $B = 10$ and $r = 19$.

h	$\bar{c}^1$	$\bar{c}^2$	$\underline{c}$	Δc	c_P^*	c_N^*
2	20	20	20	0	20	20
5	21	21	20	1	20	20
7	21	21	20	1	20	21
10	21	21	20	1	20	21
20	21	21	20	1	20	21

Table 4: Effect of h when $r_1 = 8.44$, $B = 10$ and $r = 19$.

4 Conclusion and further research

We consider an $M/M/c$ queue with infinite waiting room and two classes of customers and examine the effect of different priority policies, namely P, N, I, on the design of this system. The capacity designs under policy P and N are relatively easy, since we could show g^N and g^P are convex in the number of servers. The cost g^I and the thresholds for policy I, on the other hand, can be obtained only by solving a corresponding MDP with an infinite state space, so the analysis of policy I is carried through the analysis of policies P and N. We define measures for the effect of different priority policies on the average holding costs, Δg^{NP} and $\Delta g^{NP}/g^P$. We analyze the monotonicity and asymptotic behavior of Δg^{NP} with respect to the number of servers and traffic load of the system, and we present extensive numerical examples for $\Delta g^{NP}/g^P$. We also derive upper and lower bounds on the difference within the optimal number of servers under policies P, I and N. The numerical examples show that these bounds are very tight.

Both the analytical and numerical results suggest that the effect of priority policies on the design problem is not significant. In other words, optimizing the number of servers can be decomposed from the problem of priority. For systems with different service requirements for different classes, the priority policies may affect the average holding costs more heavily. A step in this direction is to consider a system with $\mu_1 \neq \mu_2$, or a system with two types of machines in two stations one of which is allocated to only class-1 customers and the other station is shared by both classes as in Xu et al. (1992). The effect of idling and preemption may be more significant in these cases.

References

Buzen, J. P. & Bondi, A. B. (1983), 'The response times of priority classes under preemptive resume in $M/M/m$ queue', *Operations Research* **31**, 456–465.

Dyer, M. E. & Proll, L. G. (1977), 'On the validity of marginal analysis for allocating servers in $M/M/c$ queues', *Manage. Sci.* **23**, 1019–1022.

Gross, D. & Harris, C. M. (1974), *Fundamentals of Queueing Theory*, John Wiley and Sons, New York.

Harel, A. (1988), 'Sharp bounds and simple approximations for the erlang delay and loss formula', *Management Science* **34**, 959–972.

Harel, A. (1990), The convexity of the erlang delay formula with respect to the number of servers, GSM Working Paper 90-24 The Graduate School of Management, Rutgers University.

Harel, A. & Zipkin, P. (1987), 'Strong convexity results for queueing systems', *Operations Research* **35**, 405–418.

Kleinrock, L. (1976), *Queueing Systems II*, John Wiley and Sons, New York.

Lee, H. L. & Cohen, M. A. (1983), 'A note on the convexity of performance measures of $M/M/c$ queueing systems', *Journal of Applied Probability* **20**, 920–923.

Ormeci, E. L. (1998), *Idling Rules for Queues with Preferred Customers*, Ph.D. Thesis in Operations Research, CWRU, Cleveland.

Xu, S. H., Righter, R. & Shantikumar, J. G. (1992), 'Optimal dynamic assignment of customers to heterogeneous servers in parallel', *Operations Research* **40**, 1126–1138.

5 Appendix

For the Erlang delay formula $\mathcal{D}(r,c)$, we show that $\mathcal{D}(r,c) - \mathcal{D}(r_1,c)$ is nonincreasing in c.

Lemma 1 *$\mathcal{D}(r,c) - \mathcal{D}(r_1,c)$ is nonincreasing in c.*

Proof Since $r_1 < r$, it suffices to prove that:

$$\frac{\partial f}{\partial r} \geq 0$$

where $f(r,c) = \mathcal{D}(r,c) - \mathcal{D}(r,c+1)$. It is shown in Harel (1990) that:

$$\mathcal{D}(r,c+1) = \frac{r(c-r)\mathcal{D}(r,c)}{c(c-r)+c-r\mathcal{D}(r,c)}$$

We substitute this expression in $f(r,c)$ and we will denote $\mathcal{D}(r,c)$ as $\mathcal{D}$ in the rest of the proof. Hence:

$$\frac{\partial f}{\partial r} = \mathcal{D}' - \frac{\left[(c-r-r)\,\mathcal{D} + \mathcal{D}' r(c-r)\right]\left[c(c-r)+c-r\mathcal{D}\right] - \left[-c-\mathcal{D}-r\mathcal{D}'\right] r(c-r)\mathcal{D}}{\left(c(c-r)+c-r\mathcal{D}\right)^2}$$

To show $\frac{\partial f}{\partial r} \geq 0$ it is enough to prove $A \geq 0$, where

$$\begin{aligned} A &= \mathcal{D}'\left(c(c-r)+c-r\mathcal{D}\right)^2 - \left[(c-r)\mathcal{D}-r\mathcal{D}+r(c-r)\mathcal{D}'\right]\left[c(c-r)+c-r\mathcal{D}\right] \\ &\quad + \left[c+\mathcal{D}+r\mathcal{D}'\right] r(c-r)\mathcal{D} \\ &= \mathcal{D}'\left[c(c-r)^3+(c-r\mathcal{D})^2+c(c-r)(c-r\mathcal{D})+(c-r)^2(c-r\mathcal{D})+r^2(c-r)\mathcal{D}\right] \\ &\quad +\mathcal{D}\left[-c(c-r)^2-(c-r)(c-r\mathcal{D})+2rc(c-r)+r(c-r\mathcal{D})\right]+r(c-r)\mathcal{D}^2. \end{aligned} \tag{18}$$

From Harel & Zipkin (1987):

$$\mathcal{D}' = \frac{\mathcal{D}\left[(1-\mathcal{D})\frac{r}{c}+c(1-\frac{r}{c})^2\right]}{r\left(1-\frac{r}{c}\right)}.$$

Substituting $\mathcal{D}'$ in equation(18), it follows after some algebra that:

$$\begin{aligned} \frac{A}{D} &= c(c-r)^2(1-\mathcal{D})+\frac{c}{r}(c-r)^4+\frac{1-\mathcal{D}}{c-r}(c-r\mathcal{D})^2+\frac{c-r}{r}(c-r\mathcal{D})^2+c(c-r\mathcal{D})(1-\mathcal{D}) \\ &\quad +\frac{c}{r}(c-r)^2(c-r\mathcal{D})+(c-r)(c-r\mathcal{D})(1-\mathcal{D})+\frac{(c-r)^3}{r}(c-r\mathcal{D})+r^2(1-\mathcal{D})\mathcal{D} \\ &\quad +r(c-r)^2\mathcal{D}-c(c-r)^2-(c-r)(c-r\mathcal{D})+2rc(c-r)+r(c-r\mathcal{D})+r(c-r)\mathcal{D} \\ &= (c-r)^2(\frac{c^2}{r}-2c\mathcal{D}+r\mathcal{D})+(c-r)(r\mathcal{D}^2+2rc+r\mathcal{D}-c\mathcal{D}) \\ &\quad +\frac{c}{r}(c-r)^4+\frac{1-\mathcal{D}}{c-r}(c-r\mathcal{D})^2+\frac{c-r}{r}(c-r\mathcal{D})^2+c(c-r\mathcal{D})(1-\mathcal{D}) \\ &\quad +\frac{(c-r)^3}{r}(c-r\mathcal{D})+r^2(1-\mathcal{D})\mathcal{D}+r(c-r\mathcal{D}). \end{aligned}$$

The first term is nonnegative because

$$\frac{c^2}{r}-2c\mathcal{D}+r\mathcal{D} = \frac{c^2-2rc\mathcal{D}+r^2\mathcal{D}}{r} \geq \frac{(c-r\mathcal{D})^2}{r} \geq 0.$$

The second term is nonnegative because $c\mathcal{D} \leq r$ (see e.g. Harel & Zipkin (1987)) and therefore

$$2rc+r\mathcal{D}^2+r\mathcal{D}-c\mathcal{D} \geq 2c^2\mathcal{D}-c\mathcal{D} = c\mathcal{D}(2c-1) \geq 0.$$

Thus, all the terms are nonnegative and the lemma holds. □