Extremal Properties of the Shortest Non-Full Queue and the Longest Non-Full Queue Policies in Finite Capacity Systems with State-Dependent Service Rates

P.D. Sparaggis, D. Towsley, C.G. Cassandras

COINS Technical Report 91-24
February 1991

EXTREMAL PROPERTIES OF THE SHORTEST NON-FULL QUEUE AND THE LONGEST NON-FULL QUEUE POLICIES IN FINITE CAPACITY SYSTEMS WITH STATE-DEPENDENT SERVICE RATES

P.D.SPARAGGIS,[1] D.TOWSLEY,[2] and C.G.CASSANDRAS[1]

ABSTRACT

We consider the problem of routing jobs to parallel queues with identical exponential servers and *unequal finite* buffer capacities. Service rates are state-dependent and non-decreasing with respect to queue lengths. We establish the extremal properties of the *Shortest Non-Full Queue* (SNQ) and the *Longest Non-Full Queue* (LNQ) policies, in systems with concave/convex service rates. Our analysis is based on the weak majorization of joint queue lengths which leads to stochastic orderings of critical performance indices. Moreover, we solve the buffer allocation problem, i.e., the problem of how to distribute a number of buffers among the queues. The two optimal allocation schemes are also 'extreme', in the sense of capacity balancing. Some extensions are also discussed.

Keywords: Optimal routing, buffer allocation, weak majorization, state-dependent service rates.

February, 1991

Submitted to *Journal of Applied Probability*

[1] Department of Electrical and Computer Engineering, University of Massachusetts, Amherst MA 01003. Partially supported by the Office of Naval Research under contract N00014-87-0304 and by the National Science Foundation under Grant ECS-8801912.

[2] Department of Computer and Information Science, University of Massachusetts, Amherst MA 01003. Partially supported by the Office of Naval Research under contract N00014-87-0304.

1 Introduction

A classical problem in the control of queues arises when routing decisions have to be taken for customers that arrive in front of a system which consists of a number of parallel queues with identical *exponential* servers. If the queue lengths are observed, then the intuitive '*Join the Shortest Queue*' (SQ) policy has been shown to be optimal with respect to various performance measures, such as throughput and delay, several times in the literature. The optimality of the SQ policy was first established by Winston in [14]. He proved that, in a purely Markovian system with infinite capacities, the SQ policy minimizes the discounted number of jobs that complete service by a certain time.

Weber [12], Ephremides et al. [4], and Walrand [11] extended Winston's results to systems with *general* interarrival time distributions. In particular, Walrand also provided a strong and comprehensive stochastic ordering framework which can treat consistently different versions of the problem. Menich [8] established the optimality of the SQ policy in systems with state-dependent service rates and Poisson arrivals by means of the standard uniformization method, which is used to convert a continuous time model into a discrete time one. A similar approach was used by Johri [6]. Whitt [13] called attention on the exponentiality assumption regarding the service times and presented counterexamples to demonstrate that there exist service time disrtibutions for which it is not always optimal to join the shortest queue. All the above authors considered systems consisting of queues with *infinite* buffer capacity. The optimality of the SQ policy was extended to finite capacity queueing systems in [5, 10]. Specifically, it was shown that the optimal policy always routes an incoming customer to the queue with the smallest queue length that is not at capacity. The approach in [5] was to use dynamic programming for optimization, whereas in [10] the analysis involved sample path comparisons of state and performance descriptors, based on the weak majorization of the joint queue lengths.

In this paper, we study systems, in which service times are exponentially distributed, but service rates can be state-dependent. This implies that servers may operate at different rates, depending on the number of customers in the queue. In particular, we assume that service rates are non-decreasing with respect to queue lengths and can be described by either concave or convex functions. Note that although concave functions are more frequently encountered in practice, convex functions are also reasonable to describe the behavior of finite capacity systems in which service rates are bounded from above by the rate that corresponds to the total system capacity. We establish the extremal properties of the *Shortest Non-full Queue* (SNQ) and the *Longest Non-full Queue* (LNQ) policies. Under the latter policy, a customer is routed to the queue that contains the most customers and is not at capacity. In particular, we show that the SNQ policy provides the best and worst performance in systems with unequal capacities and service rates that are respectively concave and convex functions of the queue lengths. On the other hand, the LNQ policy provides the best and worst performance when service rates are respectively convex and concave functions of the queue lengths, provided that all queue capacities are equal. Interestingly, the LNQ policy does not maintain these extremal properties when the queue capacities are unequal due to trade-offs that will be discussed later in the paper. The performance metrics we study include the number of jobs that are present in the system at any time t as well as the total number of jobs that are rejected by t.

Our analysis is based on the *weak majorization* of joint queue lengths under different policies for a single sample path. Specifically, we use *weak submajorization* and *weak supermajorization* which respectively lead to *weak Schur-convex* and *weak Schur-concave* orderings between the joint queue lengths under an extremal policy and any other dynamic policy. These orderings are defined via weak Schur-concave/convex functions. This paper completely departs from [10] and the previous literature in the way service completion events are handled. Our new arguments include the aggregation of service rates in individual systems, the coupling of service completion events between different systems by use of the max-operation on the individual aggregate service rates and the proper use of the preservation property of majorization under convex or concave functional operators. Moreover, in the case of the LNQ policy, we need to make use of the fact that all capacities should be equal in order to establish the policy's extremal properties. Some new arguments are introduced for this purpose.

Finally, we study the optimal buffer allocation problem. We show that when service rates are concave, then for a given total buffer capacity, the optimal allocation scheme is the one in which the difference between the maximum and minimum queue capacities is minimized, i.e., becomes either 0 or 1. On the other hand, when service rates are convex, it is optimal to allocate all available buffers to one queue, leaving only one buffer, that accommodates the customer in the server, to each of the remaining queues. Clearly, these two schemes are 'extremal' in the sense of balancing the number of buffers that are assigned to different queues.

The paper is organized as follows. In section 2 we define weak majorization and related orderings, and present some preliminary results. In section 3 we establish the extremal properties of SNQ and LNQ policies in systems with concave service rates. Section 4 contains systems with convex service rates and provides similar properties. The buffer allocation problem is treated in section 5. Finally, some extensions are discussed in section 6.

2 Weak majorizations and related orderings

In this section, we present the mathematical framework on which our analysis will be based and obtain some preliminary results. Some of the material used here can be found in [7]. Let $\mathbf{N}, \mathbf{M} \in I\!N^K$, $I\!N = \{0, 1, 2, ...\}$, be two arbitrary K-dimensional, integer-valued vectors. We introduce the notation $\hat{N}_k$ to denote the *k-th largest element* in vector $\mathbf{N}$ and define the following dominance relations.

Definition 1 *For any two vectors* $\mathbf{N}, \mathbf{M}$ *we say that* $\mathbf{N}$ *weakly submajorizes* $\mathbf{M}$ *(or,* $\mathbf{N}$ *weakly majorizes* $\mathbf{M}$ *from below) (written* $\mathbf{M} \prec_w \mathbf{N}$*) if*

$$\sum_{i=1}^{k} \hat{N}_i \geq \sum_{i=1}^{k} \hat{M}_i, \quad k = 1, \cdots, K.$$

For instance, if $\mathbf{N} = (5,0,3,3,1), \mathbf{M} = (4,1,2,2,3)$ we write $\mathbf{M} \prec_w \mathbf{N}$.

Remark. Whenever $\mathbf{N}$ and $\mathbf{M}$ further satisfy the relation

$$\sum_{i=1}^{K} \hat{N}_i = \sum_{i=1}^{K} \hat{M}_i$$

then $\mathbf{N}$ is said to *majorize* $\mathbf{M}$. In this case we write, $\mathbf{M} \prec \mathbf{N}$. The reader is referred to [7] for further details on this relation.

Definition 2 *For any two vectors* $\mathbf{N}, \mathbf{M}$ *we say that* $\mathbf{N}$ *weakly supermajorizes* $\mathbf{M}$ *(or,* $\mathbf{N}$ *weakly majorizes* $\mathbf{M}$ *from above) (written* $\mathbf{M} \prec^w \mathbf{N}$*) if*

$$\sum_{i=k}^{K} \hat{N}_i \leq \sum_{i=k}^{K} \hat{M}_i, \quad k = 1, \cdots, K.$$

Remark. The definitions of majorization and weak majorization are not restricted to vectors whose components are non-negative integers. However, we have presented them as such as our usage of these comparisons is restricted solely to such vectors.

Only weak submajorization was used in the context of routing problems before. Specifically, it was used by Walrand [11] and Menich [8] for showing the optimality of the shortest queue (SQ) policy in an *infinite capacity system.* It has also been used in [9] to show the optimality of certain classes of longest queue policies in the context of a network flow control problem. Finally, it was used in [10] to show the optimality of the *Shortest Non-full Queue* (SNQ) policy in a finite capacity system. In this paper, we study both weak submajorization and supermajorization; we analyze preservation properties that are important from a queueing perspective, i.e., the preservation of majorization under arrival or departure operators, and introduce the associated stochastic orderings. We begin by defining the following operators.

Operator A: Let $A_k\mathbf{N}$ denote the vector that adds a unit quantity to the k-th largest element of $\mathbf{N}$.

Operator D: Let $D_k\mathbf{N}$ denote the vector that results by subtracting a unit quantity from the k-th largest element of $\mathbf{N}$; if $\hat{N}_k = 0$ then we set $D_k\mathbf{N} = \mathbf{N}$.

We denote the l-th largest element in $A_k\mathbf{N}$, $D_k\mathbf{N}$ by $(\widehat{A_k\mathbf{N}})_l$, $(\widehat{D_k\mathbf{N}})_l$ respectively. Note that it is not necessarily true that $(\widehat{D_k\mathbf{N}})_k = (\hat{N}_k - 1, 0)^+$, or, $(\widehat{A_k\mathbf{N}})_k = \hat{N}_k + 1$. For instance, if $\mathbf{N} = (3,3,3)$ then $D_2\mathbf{N} = (3,3,2)$, where $(\widehat{D_2\mathbf{N}})_2 = \hat{N}_2$ rather than $(\widehat{D_2\mathbf{N}})_2 = \hat{N}_2 - 1$. When the vector $\mathbf{N}$ represents queue lengths, the two operators, A and D, correspond to arrivals and service completions respectively.

We now state some conditions under which the relations '$\prec_w$' and '$\prec^w$' are preserved with respect to either A or D, in the following lemmas.

Lemma 1 *For any two vectors* $\mathbf{N}, \mathbf{M} \in I\!N^K$ *such that* $\mathbf{M} \prec_w \mathbf{N}$ *it follows,*

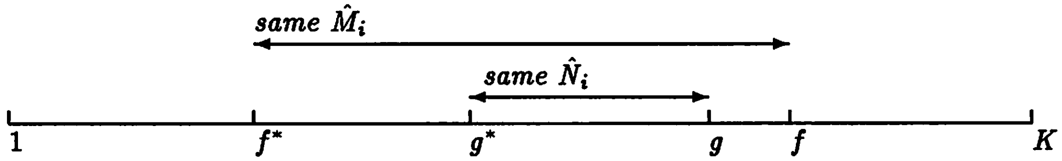

Figure 1: Index ordering for Lemma 2.

1. $A_f\mathbf{M} \prec_w A_g\mathbf{N}, \;\; g \leq f, \;\; g, f \in \{1, \ldots, K\}$
2. $D_f\mathbf{M} \prec_w D_g\mathbf{N}, \;\; g \geq f, \;\; g, f \in \{1, \ldots, K\}$
3. $\mathbf{M} \prec_w D_g\mathbf{N}$, *for* $\sum_{i=1}^{l} \hat{M}_i < \sum_{i=1}^{l} \hat{N}_i, \forall l \geq g, \;\; g, l \in \{1, \ldots, K\}$
4. $D_f\mathbf{M} \prec_w \mathbf{N}, \;\; f \in \{1, \ldots, K\}$

Proof: Properties 1,2 follows from Lemma 1 in [10]. Properties 3,4 follow easily from the definition of '$\prec_w$'. ∎

Lemma 2 *For any two vectors* $\mathbf{N}, \mathbf{M} \in I\!N^K$ *such that* $\mathbf{M} \prec^w \mathbf{N}$ *it follows,*

1. $A_f\mathbf{M} \prec^w A_g\mathbf{N}, \;\; g \leq f, \;\; g, f \in \{1, \ldots, K\}$
2. $D_f\mathbf{M} \prec^w D_g\mathbf{N}, \;\; g \geq f, \;\; g, f \in \{1, \ldots, K\}$
3. $D_f\mathbf{M} \prec^w \mathbf{N}$, *for* $\sum_{i=l}^{K} \hat{M}_i > \sum_{i=l}^{K} \hat{N}_i, \forall l \leq f, \;\; f, l \in \{1, \ldots, K\}$
4. $\mathbf{M} \prec^w D_g\mathbf{N}, \;\; g \in \{1, \ldots, K\}$

Proof: We prove part 1. Then part 2 can be proved in a similar way. Let $g^* = \min\{i : \hat{N}_i = \hat{N}_g, i \leq g\}$; likewise, let $f^* = \min\{i : \hat{M}_i = \hat{M}_f, i \leq f\}$. It follows,

$$\sum_{i=l}^{K} (\widehat{A_f\mathbf{M}})_i = \sum_{i=l}^{K} \hat{M}_i + 1, \; l \leq f^*; \quad \sum_{i=l}^{K} (\widehat{A_f\mathbf{M}})_i = \sum_{i=l}^{K} \hat{M}_i, \; l > f^*.$$

Likewise,

$$\sum_{i=l}^{K} (\widehat{A_g\mathbf{N}})_i = \sum_{i=l}^{K} \hat{N}_i + 1, \; l \leq g^*; \quad \sum_{i=l}^{K} (\widehat{A_g\mathbf{N}})_i = \sum_{i=l}^{K} \hat{N}_i, \; l > g^*.$$

Due to the above equations if $g^* \leq f^*$ it immediately follows that $A_f\mathbf{M} \prec^w A_g\mathbf{N}$. Now assume, $g^* > f^*$ (see Figure 1). Clearly, it suffices to prove that,

$$\sum_{i=l}^{K} \hat{M}_i > \sum_{i=l}^{K} \hat{N}_i, \;\; l = f^* + 1, \ldots, g^* \tag{1}$$

Proceed by contradiction assuming that

$$\sum_{i=l}^{K} \hat{M}_i = \sum_{i=l}^{K} \hat{N}_i. \tag{2}$$

for some $l \in \{f^* + 1, \ldots, g^*\}$. Since $\sum_{i=l-1}^{K} \hat{M}_i \geq \sum_{i=l-1}^{K} \hat{N}_i$, it is implied by (2) that $\hat{M}_{l-1} \geq \hat{N}_{l-1}$. Since $\hat{N}_i \leq \hat{N}_{l-1} \leq \hat{M}_{l-1} = \hat{M}_i$ for $i = l - 1, \ldots, g^* - 1$ it follows that $\hat{N}_i \leq \hat{M}_i$ for $i = l - 1, \ldots, g^* - 1$. In particular, $\hat{N}_{g^*} < \hat{N}_{g^*-1} \leq \hat{N}_{l-1} \leq \hat{M}_{l-1} = \hat{M}_{g^*}$, which implies $\hat{N}_{g^*} < \hat{M}_{g^*}$. Hence, $\sum_{i=l}^{g^*} \hat{M}_i > \sum_{i=l}^{g^*} \hat{N}_i$, which due to (2) implies, $\sum_{g^*+1}^{K} \hat{M}_i < \sum_{g^*+1}^{K} \hat{N}_i$ if $g^* + 1 < K$. This contradicts our hypothesis $\mathbf{M} \prec^w \mathbf{N}$. If $g^*(= g) = K$ then $\hat{N}_i \leq \hat{M}_i$ for $i = l - 1, \ldots, K - 1$ and $\hat{N}_K < \hat{M}_K$. Thus, $\sum_{i=l}^{K} \hat{N}_i < \sum_{i=l}^{K} \hat{M}_i$, which contradicts (2) and the proof is complete.

Part 3 is proved as follows. Define $f' = \max\{i : \hat{M}_i = \hat{M}_f, i \geq f\}$. It follows,

$$\sum_{i=l}^{K} (\widehat{D_f \mathbf{M}})_i = \sum_{i=l}^{K} \hat{M}_i, \; l < f'; \quad \sum_{i=l}^{K} (\widehat{D_f \mathbf{M}})_i = \sum_{i=l}^{K} \hat{M}_i - 1, \; l \geq f'.$$

Due to the above and our hypothesis $\sum_{i=l}^{K} \hat{M}_i \geq \sum_{i=l}^{K} \hat{N}_i, \forall l \leq f$ it suffices to show that

$$\sum_{i=l}^{K} \hat{M}_i > \sum_{i=l}^{K} \hat{N}_i, \; l = f + 1, \ldots, f', \;\; \text{if } f' > f.$$

Proceeding by contradiction assume that

$$\sum_{i=l}^{K} \hat{M}_i = \sum_{i=l}^{K} \hat{N}_i \tag{3}$$

for some $l \in \{f + 1, \ldots, f'\}$. In particular, let l be the minimum such integer in $\{f + 1, \ldots, f'\}$, i.e., $\sum_{i=l-1}^{K} \hat{M}_i > \sum_{i=l-1}^{K} \hat{N}_i$. This implies that $\hat{M}_{l-1} > \hat{N}_{l-1}$. Since $\hat{M}_l = \hat{M}_{l-1} > \hat{N}_{l-1} \geq \hat{N}_l$, (3) implies that $\sum_{i=l+1}^{K} \hat{M}_i < \sum_{i=l+1}^{K} \hat{N}_i$ if $l < K$ which contradicts our hypothesis $\mathbf{M} \prec^w \mathbf{N}$. If $l = K$ then

$$\hat{M}_K = \hat{M}_{K-1} = \hat{M}_{l-1} > \hat{N}_{l-1} = \hat{N}_{K-1} \geq \hat{N}_K,$$

which contradicts (3).

Finally, part 4 follows easily from the definition of '$\prec^w$'.

∎

We now define functions associated with weak majorizations.

Definition 3 *A function $\phi : \mathbb{N}^K \to \mathbb{R}$ is said to be a weak Schur-convex function iff*

$$\mathbf{M} \prec_w \mathbf{N} \;\Rightarrow\; \phi(\mathbf{M}) \leq \phi(\mathbf{N}),\ \forall \mathbf{M}, \mathbf{N} \in \mathbb{N}^K.$$

Examples of weak Schur-convex functions include $\sum_{k=1}^{K} f(N_k)$, for all convex f (e.g., $\sum_{k=1}^{K} N_k$) and $\max_k N_k$.

Definition 4 *A function $\psi : \mathbb{N}^K \to \mathbb{R}$ is said to be a weak Schur-concave function iff*

$$\mathbf{M} \prec^w \mathbf{N} \;\Rightarrow\; \psi(\mathbf{M}) \geq \psi(\mathbf{N}),\ \forall \mathbf{M}, \mathbf{N} \in \mathbb{N}^K.$$

The above introduced functions are related to the classes of *Shur-convex* and *Shur-concave* functions, that have been already studied in the literature (see [7] for definitions and references). These relations are described in the following lemma.

Lemma 3 *Consider a function $\phi : \mathbb{N}^K \to \mathbb{R}$.*

1. *ϕ is weak Schur-convex iff ϕ is non-decreasing and Shur-convex.*
2. *ϕ is weak Schur-concave iff ϕ is non-decreasing and Shur-concave.*

Proof: 1 follows from the first part of Theorem 3.A.8. in [7]. 2 follows from the second part of the same theorem and the fact that ψ is Shur-concave iff $-\psi$ is Shur-convex. ∎

Next, we define stochastic orderings among random vectors that are related to weak majorizations.

Definition 5 *If* $\mathbf{N}$ *and* $\mathbf{M}$ *are random vectors of dimension K, we say that* $\mathbf{N}$ *is larger than* $\mathbf{M}$ *in the sense of weak Schur-convex order (written* $\mathbf{M} \leq_{wscx} \mathbf{N}$*) iff*

$$E[\phi(\mathbf{M})] \leq E[\phi(\mathbf{N})], \quad \textit{for all weak Schur-convex functions } \phi$$

Remark. In the case that $K = 1$, this reduces to the standard stochastic ordering among real-valued random variables (r.v.'s), i.e., for r.v.'s X and Y we write $Y \leq_{st} X$ iff $\Pr(X \leq x) \leq \Pr(Y \leq x)$, $x \in \mathbb{R}$.

Definition 6 *If* $\mathbf{N}$ *and* $\mathbf{M}$ *are random vectors of dimension K, we say that* $\mathbf{N}$ *is larger than* $\mathbf{M}$ *in the sense of weak Schur-concave order (written* $\mathbf{M} \leq_{wscv} \mathbf{N}$*) iff*

$$E[\psi(\mathbf{M})] \geq E[\psi(\mathbf{N})], \quad \textit{for all weak Schur-concave functions } \psi$$

In concluding this section, we give sufficient conditions that involve the preservation of majorizations. These are due to Theorem 5.A.1. in [7].

Lemma 4 *Consider a function $f : I\!R \rightarrow I\!R$.*

1. *If f is convex, then*

$$\mathbf{M} \prec \mathbf{N} \Rightarrow (f(M_1), \ldots, f(M_K)) \prec_w (f(N_1), \ldots, f(N_K))$$

.

2. *If f is concave, then*

$$\mathbf{M} \prec \mathbf{N} \Rightarrow (f(M_1), \ldots, f(M_K)) \prec^w (f(N_1), \ldots, f(N_K))$$

.

3 Extremal properties in systems with non-decreasing concave service rates

We consider a system of K queues, each with its own server, labelled $k = 1, 2, \cdots, K$, which are fed by a single arrival stream. We assume that queues may have *unequal, finite* capacities. Let $0 < a_1 < \cdots < a_n < \cdots$ be the sequence of arrival times, i.e., the n-th job arrives at time a_n, and let $\{\tau_n\}_{n=1}^{\infty}$ denote the interarrival times, $\tau_n = a_n - a_{n-1}$, $n = 1, 2, \cdots$, $a_0 = 0$. The customers arrive at a controller which routes them to the different queues. We assume that the service times at each queue are i.i.d. exponential r.v.'s independent of the arrival times as well as the decisions made by the controller. Furthermore, we also assume that service rates in different queues are all equal. However, service rates in each queue can be state-dependent. We assume that service rates are non-decreasing and concave with respect to the number of customers in the queue.

We consider a class of routing policies, Σ, that have instantaneous queue length information available to them and that are required to route jobs to some queue that has available space, if one exists. Define SNQ to be the policy that always routes a job to the *non-full* queue with the least number of jobs. In case of a tie, any rule can be used to choose the destination queue. Note, however, that since queues with equal lengths may have unequal residual capacities, the controller may have to choose among a set of *distinct* queues in case of a tie. As a result of this, we define Σ_{SNQ} to be the class of all feasible SNQ policies. Clearly, $\Sigma_{SNQ} \subset \Sigma$.

Let $\mathbf{N}^{\pi}(t) = (N_1^{\pi}(t), \cdots, N_K^{\pi}(t))$ denote the joint queue lengths at time $t > 0$ under policy $\pi \in \Sigma$. Let $L^{\pi}(t)$ denote the number of jobs lost due to buffer overflow under policy π by time t. Further, let $\mu(\hat{N}_i^{\pi}(t))$ denote the service rate in the i-th largest queue at time t, $i = 1, \cdots, K$. In this section, we assume that μ is a non-decreasing and concave function, $\mu : I\!N \rightarrow I\!R$. Note that $\mu(\hat{N}_i^{\pi}(t)) = \mu(\hat{N}_j^{\pi}(t))$ whenever $\hat{N}_i^{\pi}(t) = \hat{N}_j^{\pi}(t)$, $i \neq j$. Finally, let $r^{\pi}(t) = \sum_{i=1}^{K} \mu(\hat{N}_i^{\pi}(t))$.

For any two policies $\pi \in \Sigma, \gamma \in \Sigma_{SNQ}$ let A^π and A^γ be arrival operators under π, γ respectively. For instance, if an arrival occurs at time t and the system's state is $\mathbf{N}^\pi(t)$ under policy π, then state $A^\pi \mathbf{N}^\pi(t)$ will be reached after the customer is routed to its appropriate queue, as induced by π. The following result is a consequence of Theorem 2 in [10].

Lemma 5 *If* $\mathbf{N}^\gamma(t) \prec_w \mathbf{N}^\pi(t)$ *and an arrival occurs at time t both under* γ *and* π*,then*

$$A^\gamma \mathbf{N}^\gamma(t) \prec_w A^\pi \mathbf{N}^\pi(t) \tag{4}$$

for all $\pi \in \Sigma$, $\gamma \in \Sigma_{SNQ}$, $t \geq 0$.

Inherent in the proof of the above lemma is the difficulty of making sample path comparisons between systems that employ different routing policies when the queue capacities in each system are unequal. Lemma 5 is used in the theorem that follows, since the presence of state-dependent service rates does not change the behavior of a policy γ in Σ_{SNQ} at the occurrence of an arrival event. This theorem proves that under any policy γ in Σ_{SNQ} the number of jobs that are rejected by any time t is minimized (in a stochastic sense). Moreover, the vector $\mathbf{N}^\pi(t)$ is shown to be larger than $\mathbf{N}^\gamma(t)$ in the sense of weak Schur-convex order, for any $\pi \in \Sigma$ and all times t. Based on this last result, one can immediately conclude that the total number of jobs present in the system at any time t is minimized under the SNQ policy.

Theorem 1

$$\begin{aligned} L^\gamma(t) &\leq_{st} L^\pi(t), & (5)\\ \mathbf{N}^\gamma(t) &\leq_{wscx} \mathbf{N}^\pi(t). & (6) \end{aligned}$$

for all $\pi \in \Sigma$, $\gamma \in \Sigma_{SNQ}$, $t > 0$ *provided that* $\mathbf{N}^\pi(0) =_{st} \mathbf{N}^\gamma(0)$.

Proof. We condition on the arrival times, service times, and initial queue lengths. The proof is by induction on event times (i.e., arrival times or departure times), $t_0 = 0, t_1, t_2, \cdots$. Specifically, we will show that

$$\begin{aligned} L^\gamma(t) &\leq L^\pi(t), & (7)\\ \mathbf{N}^\gamma(t) &\prec_w \mathbf{N}^\pi(t). & (8) \end{aligned}$$

on the given sample path. We consider $L^\pi(0) = L^\gamma(0) = 0$. Further, we can take initial queue lengths such that $\mathbf{N}^\pi(0) = \mathbf{N}^\gamma(0)$. Although capital letters are usually reserved to denote random variables within the proof of the theorem, as well as within all proofs to follow, they also indicate the values of the variables at specific time instants on *a single sample path.*

To carry on a forward induction, we couple arrival and service times in the two systems. In particular, we couple service completion events as follows. After an event has occurred, say at time t, we randomly draw a real number ϕ_t from the interval $(0, \max(r^\gamma(t), r^\pi(t))]$ and also schedule a service completion event to occur after some time s_t, according to an underlying exponential distribution with parameter $\max(r^{\gamma}(t), r^{\pi}(t))$. We let $t+s_t$ be the time of the next service completion event at the l-th largest queue under γ, if

$$\sum_{i=1}^{l-1} \mu(\hat{N}_i^{\gamma}(t)) < \phi_t \le \sum_{i=1}^{l} \mu(\hat{N}_i^{\gamma}(t)), \tag{9}$$

for some $l \in \{1, \cdots, K\}$. Likewise, we let $t+s_t$ be the time of the next service completion event at the l-th largest queue under π, if

$$\sum_{i=1}^{l-1} \mu(\hat{N}_i^{\pi}(t)) < \phi_t \le \sum_{i=1}^{l} \mu(\hat{N}_i^{\pi}(t)), \tag{10}$$

for some $l \in \{1, \cdots, K\}$. By convention we set $\sum_{i=1}^{0} \mu(\hat{N}_i^{h}(t)) = 0$, $h \in \{\gamma, \pi\}$. This coupling on the service completion epochs is allowed since

1. we can aggregate service rates in each system due to the fact that individual service times in the queues are exponentially distributed, and
2. exponential aggregate rates enable us to have the next service completion event in the system with the smaller aggregate rate (i.e., $\min(r^{\gamma}(t), r^{\pi}(t))$) occur at the same time at which the next service completion event occurs in the system with the larger rate, with probability $\frac{\min(r^{\gamma}(t), r^{\pi}(t))}{\max(r^{\gamma}(t), r^{\pi}(t))}$. This is equivalent to the above procedure.

Basis step. By the statement of the theorem, the relations hold for $t = t_0$.

Inductive step. Assume that the relations hold up through $t = t_n$. Clearly they hold for $t_n < t < t_{n+1}$. For $t = t_{n+1}$ we consider the following two cases.

Case 1. Arrival. Suppose that the next event on the observed sample path is an arrival of a customer. Clearly, the inductive hypothesis $\sum_{i=1}^{K} \hat{N}_i^{\pi}(t_n) \ge \sum_{i=1}^{K} \hat{N}_i^{\gamma}(t_n)$ guarantees that $L^{\pi}(t_{n+1}) \ge L^{\pi}(t_{n+1})$. Hence, (7) holds at t_{n+1}. As to equation (8), it follows at time t_{n+1} by the inductive hypothesis and Lemma 5.

Case 2. Service completion. Suppose that the next event on the given sample path is a completion under both or exactly one of the two policies. Equation (7) follows easily at t_{n+1}, since,

$$L^{\pi}(t_{n+1}) = L^{\pi}(t_n) \ge L^{\gamma}(t_n) = L^{\gamma}(t_{n+1}).$$

As to equation (8), we consider the following cases.

2.1. Service completion under γ only.
In this case (8) follows easily at t_{n+1}, due to the inductive hypothesis and part 4 of Lemma 1.

2.2. Service completion under π only.
Suppose a service completion occurs at the k-th largest queue under π only. Then for all $l \ge k$ it follows,

$$\sum_{i=1}^{l} \mu(\hat{N}_i^{\pi}(t_n)) \ge \sum_{i=1}^{k} \mu(\hat{N}_i^{\pi}(t_n)) \ge \phi_{t_n} > \sum_{i=1}^{K} \mu(\hat{N}_i^{\gamma}(t_n)) \ge \sum_{i=1}^{l} \mu(\hat{N}_i^{\gamma}(t_n)), \tag{11}$$

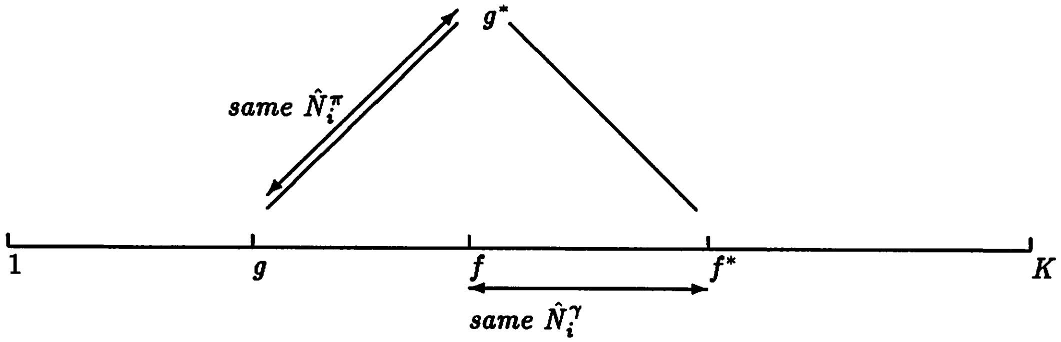

Figure 2: Index ordering for Theorem 2.

where the middle inequalities are induced by the way we schedule service completions (see (9) and (10)). Therefore,

$$\sum_{i=1}^{l} \hat{N}_i^\pi(t_n) > \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n), \ \ \forall l \geq k, \tag{12}$$

because, if instead $\sum_{i=1}^{l} \hat{N}_i^\pi(t_n) = \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n)$, then $(\hat{N}_1^\gamma(t_n), \ldots, \hat{N}_l^\gamma(t_n)) \prec (\hat{N}_1^\pi(t_n), \ldots, \hat{N}_l^\pi(t_n))$, which along with part 2 of Lemma 4 implies $\sum_{i=1}^{l} \mu(\hat{N}_i^\pi(t_n)) \leq \sum_{i=1}^{l} \mu(\hat{N}_i^\gamma(t_n))$. This contradicts (11). Therefore, (12) holds and along with part 3 of Lemma 1 implies (8) at t_{n+1}.

2.3. Service completion under both policies.
Suppose that the next service completion occurs at the f-th largest queue under γ and the g-th largest queue under π. If $g \geq f$ then (8) follows easily at time t_{n+1} due to the inductive hypothesis and part 2 of Lemma 1. Next, assume $g < f$. Let, $f^* = \max\{j \geq f : \hat{N}_j^\gamma(t_n) = \hat{N}_f^\gamma(t_n)\}$ and likewise $g^* = \max\{j \geq g : \hat{N}_j^\pi(t_n) = \hat{N}_g^\pi(t_n)\}$. It is seen that,

$$\sum_{i=1}^{l} \hat{N}_i^\gamma(t_{n+1}) = \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n), l < f^*; \ \ \sum_{i=1}^{l} \hat{N}_i^\gamma(t_{n+1}) = \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n) - 1, l \geq f^*. \tag{13}$$

Likewise,

$$\sum_{i=1}^{l} \hat{N}_i^\pi(t_{n+1}) = \sum_{i=1}^{l} \hat{N}_i^\pi(t_n), l < g^*; \ \ \sum_{i=1}^{l} \hat{N}_i^\pi(t_{n+1}) = \sum_{i=1}^{l} \hat{N}_i^\pi(t_n) - 1, l \geq g^*. \tag{14}$$

Due to the above two equations, if $f^* \leq g^*$ it immediately follows that $\mathrm{N}^\gamma(t_{n+1}) \prec_w \mathrm{N}^\pi(t_{n+1})$. We now consider the case $f^* > g^*$ (see Figure 2). Clearly, due to (13) and (14) it suffices to prove the following strict inequality,

$$\sum_{i=1}^{l} \hat{N}_i^\pi(t_n) > \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n), \ \ l = g^*, \cdots, f^* - 1. \tag{15}$$

First, we prove that,

$$\sum_{i=1}^{l} \hat{N}_i^\pi(t_n) > \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n), \ \ l = g, \cdots, f - 1. \tag{16}$$

We proceed by contradiction. Assume that $\sum_{i=1}^{l} \hat{N}_i^\pi(t_n) = \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n)$, for some $l \in \{g, \cdots, f-1\}$. Then, by using Lemma 4 as in case 2.2 it follows,

$$\sum_{i=1}^{l} \mu(\hat{N}_i^\pi(t_n)) \leq \sum_{i=1}^{l} \mu(\hat{N}_i^\gamma(t_n)). \tag{17}$$

Moreover, by the fact that the service completion occurs at the g-th and f-th largest queue under π and γ respectively, and $g \leq l < f$, we get, $\sum_{i=1}^{l} \mu(\hat{N}_i^\pi(t_n)) \geq \phi_{t_n} > \sum_{i=1}^{l} \mu(\hat{N}_i^\gamma(t_n))$ which contadicts (17). Thus, (16) holds for $l \in \{g, \cdots, f-1\}$.

Note that in case $f = f^*$, (16) yields (15) immediately. Next, we assume that $f^* > f$ and prove (15) for $l = f, \cdots, f^* - 1$. Again, assume that,

$$\sum_{i=1}^{l} \hat{N}_i^\pi(t_n) = \sum_{i=1}^{l} \hat{N}_i^\gamma(t_n), \tag{18}$$

for some $l \in \{f, \cdots, f^* - 1\}$. In particular, consider l to be the minimum such integer in $\{f, \cdots, f^* - 1\}$. Thus, $\sum_{i=1}^{l-1} \hat{N}_i^\pi(t_n) > \sum_{i=1}^{l-1} \hat{N}_i^\gamma(t_n)$, which combined with (18) implies, $\hat{N}_l^\pi(t_n) < \hat{N}_l^\gamma(t_n)$. Since, $\hat{N}_l^\pi(t_n) \geq \hat{N}_{l+1}^\pi(t_n)$ and $\hat{N}_l^\gamma(t_n) = \hat{N}_{l+1}^\gamma(t_n)$, we get, $\hat{N}_{l+1}^\pi(t_n) < \hat{N}_{l+1}^\gamma(t_n)$. This, combined with (18), implies, $\sum_{i=1}^{l+1} \hat{N}_i^\pi(t_n) < \sum_{i=1}^{l+1} \hat{N}_i^\gamma(t_n)$, which contradicts our induction hypothesis. Thus, (15) holds for $l = f, \cdots, f^*$.

Removal of the conditioning on arrival times and service times completes the theorem. ∎

It is noteworthy that the SNQ policy minimizes the expected number of jobs that are present in the system at any time instant t, while minimizing the total number of jobs that are rejected by t. Moreover, the SNQ policy is both individually and socially optimal (e.g., [1]). Specifically, the SNQ policy is optimal for any incoming customer, in the sense of minimizing its individual expected waiting time; and at the same time, it is socially optimal, in the sense of minimizing a total cost function (see below) to which not all customers contribute the same.

Next, define a cost function of the form

$$\begin{aligned} V_\alpha^\pi(\mathbf{n}) \;=\; & E\left[\int_0^\infty e^{-\alpha t} \phi(\mathbf{N}^\pi(t)) dt | \mathbf{N}(0) = \mathbf{n}\right] \\ & + E\left[\int_0^\infty e^{-\beta t} (L^\pi(t) - L^\pi(t^-)) dt | \mathbf{N}(0) = \mathbf{n}\right] \end{aligned} \tag{19}$$

for any weak Schur-convex function ϕ, $\alpha, \beta > 0$, $\mathbf{n} \in \{0, \cdots, B\}^K$, and $\pi \in \Sigma$. Note that $V_\alpha^\pi(\mathbf{n})$ is a Shur-convex function in $\mathbf{n}$. Here, the first term accounts for α-discounted holding costs for jobs that are buffered in the system, whereas the second term accounts for β-discounted loss penalties for jobs that are rejected. Holding costs are appropriate to express both throughput and delay in systems with infinite queues. However, in finite capacity systems it is possible that a policy which minimizes holding costs will, in fact, maximize the mean delay if it sufficiently decreases the throughput (i.e., the number of customers that are not rejected). Interestingly, the

SNQ policy minimizes both holding and blocking costs. The discounting factors $e^{-\alpha t}, e^{-\beta t}$ above guarantee that the cost function is well defined over an *infinite* horizon (see [2] for example). We assume that the sequence $\{n\}_{n=1}^{\infty}$ is *non-explosive* (see [3], chapter 2). The optimality of the SNQ policy is established in the following corollary.

Corollary 1 *Any policy $\gamma \in \Sigma_{SNQ}$ minimizes the cost function in (19) over all policies in Σ.*

Proof. The proof follows from the definition of $\leq_{st}$, $\leq_{wscx}$, and Theorem 1. ∎

In the remainder of this section, we restrict attention to systems in which all queues have equal capacities. Let C be the capacity of a single queue. Define the *Longest Non-full Queue* (LNQ) policy to be the policy that always routes a job to the non-full queue that contains the most jobs. We prove that LNQ provides the worst performance over all policies in Σ. Note, however, that in systems with unequal capacities, the LNQ policy does not preserve this extremal property. Specifically, in that case there exists a trade-off between routing to the queue with the longest queue length and routing to some other queue with fewer jobs but larger capacity. This latter policy will increase the chance of filling up the queue with the larger capacity at a future time; therefore, it provides the potential for performing worse than the LNQ policy. Note that since all queues have equal capacities, there is no need to define a class of LNQ policies. In the case of SNQ policies, in systems with unequal capacities, this need was necessitated by the fact that queues with the same queue length may have unequal residual capacities.

Theorem 2

$$L^{\pi}(t) \quad \leq_{st} \quad L^{LNQ}(t), \tag{20}$$

$$\mathbf{N}^{\pi}(t) \quad \leq_{wscx} \quad \mathbf{N}^{LNQ}(t), \tag{21}$$

for all $\pi \in \Sigma$, $t > 0$, provided that all queue capacities are equal, and $\mathbf{N}^{\pi}(0) =_{st} \mathbf{N}^{LNQ}(0)$.

Proof. We condition on the arrival times, service times, and initial queue lengths, and couple the event times as in Theorem 1. The proof is by induction on event times (i.e., arrival times or departure times), $t_0 = 0, t_1, t_2, \cdots$. Specifically, we show that

$$L^{\pi}(t) \quad \leq \quad L^{LNQ}(t), \tag{22}$$

$$\mathbf{N}^{\pi}(t) \quad \prec_w \quad \mathbf{N}^{LNQ}(t). \tag{23}$$

on the given sample path. We consider $L^{\pi}(0) = L^{LNQ}(0) = 0$. Further, we can take initial queue lengths such that $\mathbf{N}^{\pi}(0) = \mathbf{N}^{\gamma}(0)$.

Basis step. By the statement of the theorem, the relations hold for $t = t_0$.

Inductive step. Assume that the relations hold up through $t = t_n$. Clearly they hold for $t_n < t < t_{n+1}$. For $t = t_{n+1}$ we consider the following two cases.

Case 1. Arrival. Suppose that the next event is an arrival of a customer at the f-th largest and g-th largest queue under π and LNQ respectively. Clearly, the inductive hypothesis $\sum_{i=1}^{K} \hat{N}_i^{\pi}(t_n) \leq \sum_{i=1}^{K} \hat{N}_i^{LNQ}(t_n)$ guarantees that $L^{\pi}(t_{n+1}) \leq L^{LNQ}(t_{n+1})$. Hence, (22) holds at t_{n+1}. As to equation (23), it follows immediately at time t_{n+1} by part 1 of Lemma 1 if $g \leq f$. If $g > f$ then define $f^* = \min\{i \leq f : \hat{N}_i^{\pi}(t_n) = \hat{N}_f^{\pi}(t_n)\}$ and $g^* = \min\{i \leq g : \hat{N}_i^{LNQ}(t_n) = \hat{N}_g^{LNQ}(t_n)\}$. It is seen that,

$$\sum_{i=1}^{l} \hat{N}_i^{\pi}(t_{n+1}) = \sum_{i=1}^{l} \hat{N}^{\pi}(t_n),\ l < f^*; \quad \sum_{i=1}^{l} \hat{N}_i^{\pi}(t_{n+1}) = \sum_{i=1}^{l} \hat{N}^{\pi}(t_n) + 1,\ l \geq f^*.$$

Likewise,

$$\sum_{i=1}^{l} \hat{N}_i^{LNQ}(t_{n+1}) = \sum_{i=1}^{l} \hat{N}^{LNQ}(t_n),\ l < g^*; \quad \sum_{i=1}^{l} \hat{N}_i^{LNQ}(t_{n+1}) = \sum_{i=1}^{l} \hat{N}^{LNQ}(t_n) + 1,\ l \geq g^*.$$

Clearly, if $f^* \geq g^*$ (23) holds at t_{n+1}. If now $g^* > f^*$ it suffices to prove that

$$\sum_{i=1}^{l} \hat{N}_i^{\pi}(t_n) < \sum_{i=1}^{l} \hat{N}_i^{LNQ}(t_n), \quad l = f^*, \ldots, g^* - 1.$$

This follows easily since $\sum_{i=1}^{l} \hat{N}_i^{\pi}(t_n) < lC$ $(\hat{N}_{f^*}^{\pi}(t_n) < C)$ and $\sum_{i=1}^{l} \hat{N}_i^{LNQ}(t_n) = lC$ for all $l \in \{f^*, \cdots, g^* - 1\}$.

Case 2. Service completion. The case of a service completion is handled exactly as in Theorem 1.

Removal of the conditioning on arrival times and service times completes the theorem. ∎

Thus, the LNQ policy provides the worst performance in the following sense.

Corollary 2 *The LNQ policy maximizes the cost function in (19) over all policies in Σ when all queue capacities are equal.*

4 Extremal properties in systems with non-decreasing convex service rates

In this section, we consider systems with non-decreasing *convex* state-dependent service rates. Again, we restrict our attention to systems in which all queues have equal capacities, given by C. We let Σ^* denote the class of feasible routing policies in this system. The main result of this section states that the LNQ policy provides the best performance over Σ^*, whereas the SNQ policy provides the worst. Note, however, that in systems with unequal capacities the LNQ policy does not maintain this optimality property. Specifically, in that case there exists a trade-off between routing to the queue with the most jobs and routing to some other queue with fewer jobs but larger capacity. This latter policy will increase the chance of filling up the queue with the larger capacity at a future time; therefore, it may perform better than the LNQ policy.

Theorem 3

$$\begin{aligned} L^{\pi}(t) &\ \ge_{st}\ L^{LNQ}(t), & (24)\\ \mathbf{N}^{\pi}(t) &\ \le_{wscv}\ \mathbf{N}^{LNQ}(t). & (25)\end{aligned}$$

for all $\pi \in \Sigma^*$, $t > 0$ *provided that all queues have equal capacities and* $\mathbf{N}^{\pi}(0) =_{st} \mathbf{N}^{LNQ}(0)$.

Proof. We use the same sample path approach as in Theorems 1 and 2 and prove by induction that on a single sample path the following relations hold.

$$\begin{aligned} L^{\pi}(t) &\ \ge\ L^{LNQ}(t), & (26)\\ \mathbf{N}^{\pi}(t) &\ \prec^{w}\ \mathbf{N}^{LNQ}(t). & (27)\end{aligned}$$

As before, we establish the above relations at time t_{n+1} assuming they hold at time t_n, where t_n, t_{n+1} are two consecutive event times on the considered sample path. In particular, after an event has occurred at time t, we randomly draw a real number ψ_t from the interval $(0, \max(r^{\pi}(t), r^{LNQ}(t))]$ and also schedule a service completion event to occur after some time s_t, which is exponentially distributed with a parameter equal to $\max(r^{\pi}(t), r^{LNQ}(t))$. Moreover, we let the next service completion event occur at the l-th largest queue under policy $h = \pi, LNQ$, if $\sum_{i=l}^{K} \mu(\hat{N}_i^h(t)) \ge \psi_t > \sum_{i=l+1}^{K} \mu(\hat{N}_i^h(t))$, for some $l \in \{1, \cdots, K\}$. By convention we set $\sum_{i=K+1}^{K} \mu(\hat{N}_i^h(t)) = 0$, $h = \pi, LNQ$.

We consider the following two cases.

Case 1. Arrival. Suppose that the next event on the observed sample path is an arrival of a customer at the f-th largest and g-th largest queue under π and LNQ respectively. Equation (26) holds at t_{n+1} as in case 1 of Theorems 1, 2. As to equation (27), it follows immediately at time t_{n+1} by part 1 of Lemma 2 if $g \le f$. If $g > f$ then as in Lemma 2 it suffices to prove that

$$\sum_{i=l}^{K} \hat{N}_i^{\pi}(t_n) > \sum_{i=l}^{K} \hat{N}_i^{LNQ}(t_n), \ \ l = f^* + 1, \ldots, g^*, \ \ f^* < g^*, \tag{28}$$

where f^*, g^* are defined as in Lemma 2. Note that

$$\sum_{i=1}^{l} \hat{N}_i^{\pi}(t_n) < lC = \sum_{i=1}^{l} \hat{N}_i^{LNQ}(t_n), \ \ l = f^*, \cdots, g^* - 1, \tag{29}$$

since $\hat{N}_{f^*}^{\pi}(t_n) < C$. Then by the inductive hypothesis $\sum_{i=1}^{K} \hat{N}_i^{\pi}(t_n) \ge \sum_{i=1}^{K} \hat{N}_i^{LNQ}(t_n)$ and (29), (28) follows.

Case 2. Service completion. Suppose that the next event on the given sample path is a completion under both or exactly one of the two policies. Equation (26) follows easily at t_{n+1} as in case 2 of Theorem 1. As to equation (27) we consider the following cases.

2.1. Service completion under LNQ only.
In this case (27) follows easily at t_{n+1} due to the inductive hypothesis and part 4 of Lemma 2.

2.2. Service completion under π only.
Suppose a service completion occurs at the k-th largest queue under π only. Then $\forall l \leq k$ it follows,

$$\sum_{i=l}^{K} \mu(\hat{N}_i^{\pi}(t_n)) \geq \sum_{i=k}^{K} \mu(\hat{N}_i^{\pi}(t_n)) \geq \psi_{t_n} > \sum_{i=1}^{K} \mu(\hat{N}_i^{LNQ}(t_n)) \geq \sum_{i=l}^{K} \mu(\hat{N}_i^{LNQ}(t_n)), \qquad (30)$$

where the middle inequalities are induced by the way we schedule service completions, given ψ_{t_n}. Therefore,

$$\sum_{i=l}^{K} \hat{N}_i^{\pi}(t_n) > \sum_{i=l}^{K} \hat{N}_i^{LNQ}(t_n), \ \ \forall l \leq k, \qquad (31)$$

because, if instead of (31) $\sum_{i=l}^{K} \hat{N}_i^{\pi}(t_n) = \sum_{i=l}^{K} \hat{N}_i^{LNQ}(t_n)$, then $(\hat{N}_l^{\pi}(t_n), \ldots, \hat{N}_K(t_n)^{\pi}) \prec (\hat{N}_l^{LNQ}(t_n), \ldots, \hat{N}_K(t_n)^{LNQ})$, which along with part 1 of Lemma 4 implies $\sum_{i=l}^{K} \mu(\hat{N}_i^{\pi}(t_n)) \leq \sum_{i=l}^{K} \mu(\hat{N}_i^{LNQ}(t_n))$. This contradicts (30). Thus, (31) holds and along with part 3 of Lemma 2 implies (27) at t_{n+1}. Note that (30) and (31) are the dual of (11) and (12) in Theorem 1.

2.3. Service completion under both policies.
Suppose that the next service completion occurs at the f-th largest queue under π and the g-th largest queue under LNQ. If $g \geq f$ then (27) follows easily at time t_{n+1} due to the inductive hypothesis and part 2 of Lemma 2. Next, assume $g < f$. Arguments similar to those in case 2.3 of Theorem 1 suffice to prove

$$\sum_{i=l}^{K} \hat{N}_i^{\pi}(t_n) > \sum_{i=l}^{K} \hat{N}_i^{LNQ}(t_n), \ \ l = g^* + 1, \cdots, f^*, \qquad (32)$$

where f^*, g^* are defined as in case 2.3, Theorem 1. This is the dual of (15). The rest of the proof involves arguments similar to those in case 2.3 of Theorem 1.

Removal of the conditioning on arrival times and service times completes the theorem. ∎

Now, define a cost function of the form

$$\begin{aligned} J_{\alpha}^{\pi}(\mathbf{n}) &= E\left[\int_0^{\infty} e^{-\alpha t}\psi(\mathbf{N}^{\pi}(t))dt | \mathbf{N}(0) = \mathbf{n}\right] \\ &\qquad + E\left[\int_0^{\infty} e^{-\beta t}(L^{\pi}(t) - L^{\pi}(t^{-}))dt | \mathbf{N}(0) = \mathbf{n}\right] \qquad (33) \end{aligned}$$

for any weak Schur-concave function ψ, $\alpha, \beta > 0$, $\mathbf{n} \in \{0, \cdots, B\}^K$, and $\pi \in \Sigma^*$.

The optimality of the LNQ policy is shown in the following corollary.

Corollary 3 *The LNQ policy minimizes the cost function in (33) over all policies in Σ^* when all queue capacities are equal.*

Similarly to Theorem 2, we can show that any policy $\gamma \in \Sigma^*_{SNQ}$ provides the worst performance. We define Σ^*_{SNQ} to represent the class of SNQ policies in systems with non-decreasing, convex service rates. The proofs of the following Theorem and corollary are omitted since they use arguments very similar to those described in detail before.

Theorem 4

$$L^{\pi}(t) \leq_{st} L^{\gamma}(t), \tag{34}$$

$$\mathbf{N}^{\gamma}(t) \leq_{wscv} \mathbf{N}^{\pi}(t). \tag{35}$$

for all $\pi \in \Sigma^*$, $\gamma \in \Sigma^*_{SNQ}$, $t > 0$ provided that all queues have equal capacities, and $\mathbf{N}^{\pi}(0) =_{st} \mathbf{N}^{\gamma}(0)$.

Corollary 4 *Any policy $\gamma \in \Sigma^*_{SNQ}$ maximizes the cost function in (33) over all policies in Σ^* when all queue capacities are equal.*

It can be shown that, in fact, any policy $\gamma \in \Sigma^*_{SNQ}$ provides the worst performance, even if queue capacities are not equal. This involves some additional algebraic arguments similar to those used in Lemma 3 and Theorem 2 in [10].

5 The optimal buffer allocation problem

A related problem is to determine the optimal allocation of B buffers to K parallel queues ($B \geq K$). Any feasible allocation scheme defines a system in which there exist a number of different routing policies that can be employed. Thus, our objective becomes to

1. specify the optimal routing policy in the buffer allocation scheme which we expect to perform optimally, and
2. show that under this policy the system defined by the optimal scheme does, indeed, outperform any other system that may be defined by a different scheme and may employ a different policy.

In case of systems with concave service rates, we can take advantage of the fact that the optimal routing policy (SNQ) has been already determined, in order to specify a *unique* allocation scheme, which is optimal in the sense of minimizing (19). The problem becomes more complicated when service rates are convex. In this case the optimal policy is not known, except when all capacities are equal. We show later in this section that when all buffers are assigned to a single queue, the LNQ policy is optimal and this allocation scheme provides the best performance.

Let $\boldsymbol{B} = (B_1, \cdots, B_K)$ be an allocation scheme such that $\sum_{i=1}^{K} B_i = B$, $B_i \geq 1$ for all $i = 1, .., K$. Let

$$\mathcal{B} = \{\boldsymbol{B} = (B_1, \cdots, B_K) \ : \ \sum_{k=1}^{K} B_k = B, B_i \geq B_{i+1} \geq 1, i = 1, \cdots, K-1\} \tag{36}$$

denote the class of all feasible allocation schemes.

Define the scheme $\boldsymbol{B}^o = (B_1^o, \cdots, B_K^o)$ such that

$$B_i^o = \begin{cases} \lfloor B/K \rfloor + 1, & B \bmod K \neq 0, i = 1, .., B \bmod K, \\ \lfloor B/K \rfloor, & \text{otherwise,} \end{cases} \tag{37}$$

i.e., the B_i's can differ by one at most. We show that $\boldsymbol{B}^o$ is the optimal allocation scheme provided that service rates are non-decreasing and concave. We begin the analysis in this section with a preliminary lemma.

Lemma 6

$$\boldsymbol{B}^o \prec \boldsymbol{B}, \quad \forall \boldsymbol{B} \in \mathcal{B}$$

Proof. Follows from the definition of $\boldsymbol{B}^o$ and "$\prec$". ∎

Next, we only consider systems that employ *optimal* policies. We modify our earlier notation so that whenever we are interested in the behavior of a system under the optimal policy, when the buffer allocation is determined by some scheme $\boldsymbol{B}^b \in \mathcal{B}$, we will use the superscript of $\boldsymbol{B}$ and write $L^b(t)$, $\mathbf{N}^b(t)$. The following result is a consequence of Lemma 5 in [10].

Lemma 7 *For any two allocation schemes $\boldsymbol{B}^1, \boldsymbol{B}^2 \in \mathcal{B}$, with $\boldsymbol{B}^2 \prec \boldsymbol{B}^1$, it follows*

$$\mathbf{N}^2(t) \prec_w \mathbf{N}^1(t) \ \Rightarrow A^{SNQ}\mathbf{N}^2(t) \prec_w A^{SNQ}\mathbf{N}^1(t)$$

$t \geq 0$.

Now the next result follows easily.

Theorem 5 *If $\boldsymbol{B}^2 \prec \boldsymbol{B}^1$, then*

$$L^2(t) \quad \leq_{st} \quad L^1(t) \tag{38}$$

$$\mathbf{N}^2(t) \quad \leq_{wscx} \quad \mathbf{N}^1(t) \tag{39}$$

for all $\boldsymbol{B}^1, \boldsymbol{B}^2 \in \mathcal{B}$, $t \geq 0$ provided that service rates are non-decreasing and concave, and $\mathbf{N}^2(0) =_{st} \mathbf{N}^1(0)$.

Proof. The proof is similar to that of Theorem 1. In particular, arrival events can be treated easily due to Lemma 7. ∎

As a result of Lemma 6 and Theorem 5, we conclude that the buffer allocation scheme $\boldsymbol{B}^o$ provides the optimum performance in the following sense.

Corollary 5 *$\boldsymbol{B}^o$ minimizes the cost function in (19) over the class $\mathcal{B}$ when service rates are non-decreasing and concave.*

Now, define the scheme $\boldsymbol{B}^O = (B_1^O, \cdots, B_K^O)$ such that

$$B_i^O = \begin{cases} B - K + 1, & i = 1, \\ 1 & \text{otherwise.} \end{cases} \tag{40}$$

Implicit in the above definition is the assumption that at least one buffer has to be allocated to each server. In other words, we assume that servers must occupy one buffer. We show that the above scheme B^O provides the optimal performance when service rates are non-decreasing and convex and the LNQ policy is employed.

Theorem 6 *For any allocation scheme $\boldsymbol{B}^s \in \mathcal{B}$ it follows,*

$$L^s(t) \quad \geq_{st} \quad L^O(t) \tag{41}$$

$$\mathbf{N}^s(t) \quad \leq_{wscv} \quad \mathbf{N}^O(t) \tag{42}$$

provided that service rates are non-decreasing and convex, and $\mathbf{N}^s(0) =_{st} \mathbf{N}^O(0)$.

Proof: Let π be the policy employed in $\boldsymbol{B}^s$. We allow π to be arbitrary. This overcomes the obstacle of the optimal policy in $\boldsymbol{B}^s$ being unknown. The proof follows exactly that of Theorem 3, by substituting π by s and LNQ by O anywhere in the notation. The only difference is in equation (29) which should now read,

$$\sum_{i=1}^{l} \hat{N}_i^s(t_n) < (B - K + 1) + (l - 1) = \sum_{i=1}^{l} \hat{N}_i^O(t_n),\ l = f^*, \ldots, g^* - 1.$$

In particular, the first strict inequality in the above relation follows from the fact that for any allocation scheme $\boldsymbol{B}^s \in \mathcal{B}$, $\max(\sum_{i=1}^{l} \hat{N}_i^s(t)) \leq (B - K + 1) + (l - 1)$ at all times $t \geq 0$ (this maximum value is achieved when all queues are at capacity) and the queue $\hat{N}_{f^*}^s(t_n)$ is not full. ∎

As a result, we conclude that the buffer allocation scheme $\boldsymbol{B}^O$ provides the optimum performance in the following sense.

Corollary 6 *$\boldsymbol{B}^O$ minimizes the cost function in (19) over the class $\mathcal{B}$ when service rates are non-decreasing and convex.*

Remark: Note that when service rates are convex and servers can be removed out of the system, it is preferable to leave only one server in the system, thus allocating the $(K-1)$ buffer positions occupied by the $(K-1)$ servers to the single remaining server.

6 Extensions

Some extensions are possible. First, consider the case of monotonically increasing Poisson arrival rates. For instance, consider two systems S_1, S_2 with arrival rates λ_1, λ_2 respectively such that $\lambda_1 \leq \lambda_2$. Then, Theorem 1 should be extended to read,

$$L_1^\gamma(t) \quad \leq_{st} \quad L_2^\pi(t), \tag{43}$$

$$\mathbf{N}_1^\gamma(t) \quad \leq_{wscx} \quad \mathbf{N}_2^\pi(t) \tag{44}$$

where the subscripts 1,2 refer to S_1, S_2 respectively. The proof involves only one additional step: couple the arrival times in the two systems so that an arrival occurs in S_1 only if it occurs in S_2. This is allowed by our assumption of Poisson arrival rates. Note that arrivals only in S_2 will strengthen (43), (44) above. Similarly, all other results can be extended.

It is also possible to assume that when the system is full, the arrival process is ***shut-off*** so that no customer is ever lost. This model is more appropriate to use in systems that employ mechanisms for overflow avoidance (e.g., window protocols in communication networks). For instance, in view of Theorem 1, $\mathbf{N}^\gamma(t) \prec_w \mathbf{N}^\pi(t)$ implies that an arrival may occur either in both systems, or only under γ. In this latter case, we can reassure that the majorization is preserved since all queues under π are at capacity.

Another possible extension regards families of state-dependent service rates. Specifically, all results still hold if service rates are of the form $\mu(N_j^\pi(t)/\sum_{i=1}^K N_i^\pi(t))$ instead of $\mu(N_j^\pi(t))$ and the assumption of μ being either concave or convex is maintained where appropriate. This follows by a single extension of Lemma 4:

$$\mathbf{M} \prec \mathbf{N} \Rightarrow (\frac{M_1}{\sum_{i=1}^K M_i}, \ldots, \frac{M_K}{\sum_{i=1}^K M_i}) \prec (\frac{N_1}{\sum_{i=1}^K N_i}, \ldots, \frac{N_K}{\sum_{i=1}^K N_i}),$$

since $\sum_{i=1}^K N_i = \sum_{i=1}^K M_i$ by the definition of '$\prec$'. These families of state-dependent service rates can be used to model processor sharing systems in which the processing capacity is distributed among the different stations according to their load. The case of concave processing sharing service rates was also discussed in [8].

Last, another extension is to consider a system in which buffer space is available at the controller. This implies that whenever a customer arrives at the system, it may be queued at the controller before it is routed to one of the parallel service stations. In this case, it can be shown that if service rates are concave, the optimal policy always delays making routing decisions, i.e., holds customers at the controller as long as none of the parallel queues is empty. Then, when the controller becomes full, an incoming customer is routed to the station with the shortest queue length. This is an extension of a result in [10]. On the other hand, when service rates are convex, the controller should always route customers to the queue with the largest queue length and never delay routing of a customer.

References

[1] C.E.Bell and S.Stidham, "Individual and social optimization in the allocation of customers to alternative servers," *Management Science*, vol. 29, pp. 831-839, 1983.

[2] D.P.Bertsekas, *Dynamic Programming*, Prentice Hall, chapter 5, 1987.

[3] P.Bremaud, *Point processes and queues*, John Wiley and Sons, 1981.

[4] A.Ephremides, P.Varaiya and J.Walrand, "A simple dynamic routing problem," *IEEE Trans. on Aut. Control*, vol. AC-25, 1980.

[5] A.Hordijk and G.Koole, "On the optimality of the generalized shortest queue policy," *Probability in the Eng. and Info. Sciences*, vol. 4, pp. 477-487, 1990.

[6] P.K.Johri, "Optimality of the shortest line discipline with state-dependent service times," *European J. of Operational Research*, vol. 41, pp. 157-161, 1989.

[7] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, 1979.

[8] R.Menich, "Optimality of the Shortest Queue routing for dependent service stations," *Proc. of the 26th IEEE Conf. on Decision and Control*, L.A., California, December 1987.

[9] D. Towsley, S. Fdida, H. Santoso, "Design and evaluation of flow control protocols for Metropolitan Area Networks," *Architecture and performance issues of high-capacity local and metropolitan area networks*, ed. G.Pujolle, Springer Verlag, to appear.

[10] D.Towsley, P.D.Sparaggis and C.G.Cassandras, "Stochastic ordering and optimal routing control for a class of finite capacity queueing systems," *Proc. of the 29th IEEE Conf. on Decision and Control*, pp. 658-663, Honolulu, Hawaii, December 1990.

[11] J.Walrand, *An introduction to queueing networks*, Prentice Hall, 1988.

[12] R.R.Weber, "On the optimal assignment of customers to parallel queue," *J. of Applied Prob.*, vol. 15, pp. 406-413, 1978.

[13] W.Whitt, "Deciding which queue to join," *Oper. Res.*, vol. 34, pp. 55-62, 1986.

[14] W.Winston, "Optimality of the shortest line discipline," *J. of Applied Prob.*, vol. 14, pp. 181-189, 1977.