Probability in the Engineering and Informational Sciences, **18**, 2004, 329–337. Printed in the U.S.A.

ON A ROUTING PROBLEM

TORGNY LINDVALL
School of Mathematical Sciences
University of Göteborg
412 96 Göteborg, Sweden
E-mail: lindvall@math.chalmers.se

In a system with one queue and several service stations, it is a natural principle to route a customer to the idle station with the distributionwise shortest service time. For the case with exponentially distributed service times, we use a coupling to give strong support to that principle. We also treat another topic. A modified version of our methods brings support to the design principle: It is better with few but quick servers.

1. INTRODUCTION

Let $\mathcal{S}$ and $\mathcal{S}'$ be stochastic models of two queuing systems. Suppose that they may be constructed on the same probability space in such a way that

$$Y_t \leq Y_t' \quad \text{for all } t \geq 0 \tag{1}$$

when the initial conditions of $\mathcal{S}$ and $\mathcal{S}'$ are identical; here, Y_t and Y_t' are the number of customers at time $t \geq 0$ in $\mathcal{S}$ and $\mathcal{S}'$, respectively. If (1) holds, we are able to make an interesting comparison between $\mathcal{S}$ and $\mathcal{S}'$; we will use it for the following routing problem.

Consider the standard $M/M/k$ system: Poisson arrival process, one queue, and k service stations with independent and identically distributed (i.i.d.) exponentially distributed service times that are also independent of the arrival process. Now, retain all of these conditions except the sameness of the service-time distributions; we assume that the service times of station i are $\exp(\delta_i)$-distributed, where $\delta_1 \geq \cdots \geq \delta_k$ and the δ_i's are not all equal. The problem is as follows: If a customer arrives and there are two or more idle stations, to which of these should she be routed? A natural policy is the following:

© 2004 Cambridge University Press 0269-9648/04 $16.00

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

Always route to the idle station with the highest service rate. **(2)**

For convenience, we use the term "route" also for the trivial case when a customer arrives and there is exactly one station idle or when at least one customer waits and one station becomes idle. We abbreviate (2) to the HSR (highest service rate) policy.

In Section 2, we will prove that if the HSR is used for $\mathcal{S}$ and some other policy is used for $\mathcal{S}'$, then (1) holds. Actually, that turns out to be the case for an arbitrary arrival process that is independent of the service times.

Note that if g is a bounded, nondecreasing, and nonnegative function and α is a constant > 0, then (1) implies

$$\mathbb{E}[g(Y_t)] \le \mathbb{E}[g(Y_t')] \quad \text{for all } t \ge 0$$

and

$$\mathbb{E}\left[\int_0^\infty e^{-\alpha s} g(Y_s)\, ds\right] \le \mathbb{E}\left[\int_0^\infty e^{-\alpha s} g(Y_s')\, ds\right]. \tag{3}$$

In fact, any reasonable cost criterion based only on the number of customers decides in favor of the HSR policy.

Considerable attention has been paid to comparisons between queuing systems (cf. Asmussen [1, Sect. XI.1], Stoyan [5, Chaps. 5 and 6], and Wolff [7, Chap. 11] for accounts). The importance of couplings (i.e., constructions of random elements on the same probability space) has been apparent for a long time; for a recent example, see Lindvall [3].

When it comes to routing problems, interesting uses of couplings turn up in Sparaggis, Towsley, and Cassandras [4] and Winkler [6].

In Section 3, we modify our setup in order to study systems with different numbers of stations. It turns out that if there are more but slower (to be made precise) stations in $\mathcal{S}'$ than in $\mathcal{S}$, then $\mathcal{S}$ is superior to $\mathcal{S}'$ in the sense that (1) holds if the maximal capacities of the systems are the same.

2. THE HSR POLICY IS OPTIMAL

Recall the setup presented in (1) and (2). To begin with, assume that the arrival process is deterministic, with arrivals at time points $s_j, j \ge 1$, such that

$$0 < s_1 < s_2 < \cdots$$

and

$$s_j \to \infty \quad \text{and} \quad j \to \infty. \tag{4}$$

Set $s_0 = 0$. We will prove that the HSR policy implies (1) by induction, using the following idea: Suppose we have been able to construct $\mathcal{S}$ and $\mathcal{S}'$ on $[0, s_j]$ in such a way that (1) holds for $t \in [0, s_j]$ and the states of $\mathcal{S}$ and $\mathcal{S}'$ at $t = s_j$ are feasible, in a sense to be made precise later, then we will be able to extend that to the interval $[0, s_{j+1}]$, for $j = 1, 2, \ldots$

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

It is due time for the details. Because the service times are exponentially distributed, the states of $\mathcal{S}$ and $\mathcal{S}'$ are completely described by the number of customers and the configurations of busy stations. Let

$$B = \{0,1\}^k \tag{5}$$

and for $\mathbf{b} = (b_1, \ldots, b_k) \in B$,

$$C(\mathbf{b}) = \{j; b_j = 0\},$$
$$D(\mathbf{b}) = \{j; b_j = 1\},$$
$$E = \{(i, \mathbf{b}); i \in \mathbb{Z}_+, \mathbf{b} \in B, \#D(\mathbf{b}) = i \wedge k\}.$$

That a system is in state $(i, \mathbf{b}) \in E$ will mean that there are i customers in it, with the configuration of busy stations given by $\mathbf{b}$; $b_j = 1$ indicates that station j is busy.

For the partial ordering on E that will be defined now, let $D(\mathbf{b}) \neq \emptyset$ to start. We define $\preccurlyeq$ on E as follows: For $i \geq 1$,

$$(i, \mathbf{b}) \preccurlyeq (i', \mathbf{b}') \tag{6}$$

exactly when the following hold:

(i) $i \leq i'$

(ii) There exists a strictly increasing mapping $\varphi(\mathbf{b}, \mathbf{b}', \cdot) : D(\mathbf{b}) \to \{1, \ldots, k\}$ such that $\varphi(\mathbf{b}, \mathbf{b}', j) \geq j$ for all $j \in D(\mathbf{b})$, and $\varphi(\mathbf{b}, \mathbf{b}', D(\mathbf{b})) \subseteq D(\mathbf{b}')$.

To complete the definition, let $(0, (0, \ldots, 0)) \preccurlyeq (i', \mathbf{b}')$ for all $(i', \mathbf{b}') \in E$.

The meaning of (6(ii)) is that we can match each busy station in $\mathcal{S}$ with one busy station in $\mathcal{S}'$ with a smaller or equal service rate.

We will use $X_t = (Y_t, B_t)$, $t \geq 0$, to denote queuing process when the HSR policy is used and use $X'_t = (Y'_t, B'_t)$, $t \geq 0$, to denote the competing process with some other routing policy; the meanings of B_t and B'_t, $t \geq 0$, are the obvious ones. We will produce a coupling such that $X_t \preccurlyeq X'_t$ for all $t \geq 0$, under the assumption that the initial distributions are the same.

We follow the convention about right-continuity of paths of the type of processes we are dealing with here. That means that we will produce (X_t, X'_t), $t \geq 0$, as a time-homogeneous jump Markov process on each of the intervals $[s_j, s_{j+1})$, $j = 0,1,\ldots$, and then pay special attention to the jump at s_{j+1}. We are now able to be precise about what the proof idea requires:

$$\text{for each } T > 0, \text{ with no arrival in } [0, T) \text{ and an arrival at } T, \tag{7}$$

(i) produce a coupling such that (X_t, X'_t), $t \geq 0$, is a jump Markov process on $[0, T)$ satisfying $X_t \preccurlyeq X'_t$ for all $t \in [0, T)$ if $X_0 \preccurlyeq X'_0$

(ii) see to it that $X_T \preccurlyeq X'_T$.

We used the word "feasible" in the presentation of the proof idea; now (7(ii)) makes that precise. We will apply (7) to the intervals $[s_j, s_{j+1}]$, $j = 0,1,\ldots$, successively, with $X_0 = X'_0$ for the application to the interval $[0, s_1]$. That yields

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

$$X_t \preccurlyeq X_t' \quad \text{for all } t \geq 0 \tag{8}$$

if $X_0 = X_0'$, which is actually stronger than (1).

For (7(i)), recall the properties of the mappings $\varphi(\mathbf{b}, \mathbf{b}', \cdot)$ in (6(ii)) and that the stations have intensities $\delta_1, \ldots, \delta_k$ satisfying $\delta_1 \geq \cdots \geq \delta_k$.

Let $e_j = (0, \ldots, 1, \ldots 0)$ be the jth unit vector. For convenience, we will occasionally abbreviate $\varphi(\mathbf{b}, \mathbf{b}', j)$ to j'.

We will produce (X_t, X_t'), $t \geq 0$, in terms of transition intensities. Fix two elements in E such that $(i, \mathbf{b}) \preccurlyeq (i', \mathbf{b}')$, to be thought of as states of X_0 and X_0', respectively. We sort out the presentation of the intensities into three cases.

Case 1: $i > k$. From $((i, \mathbf{b}), (i', \mathbf{b}'))$, we make a jump

to state	with intensity
$((i-1, \mathbf{b}), (i'-1, \mathbf{b}'))$	$\sum_{j=1}^{k} \delta_j$

Case 2: $i \leq k, i' \leq k$. From $((i, \mathbf{b}), (i', \mathbf{b}'))$, we make jumps

to state	with intensity
$((i-1, \mathbf{b} - e_j), (i'-1, \mathbf{b}' - e_{j'}))$	$\delta_{j'}$
$((i-1, \mathbf{b} - e_j), (i', \mathbf{b}'))$	$\delta_j - \delta_{j'}$

for $j \in D(\mathbf{b})$ and

to state	with intensity
$((\mathrm{i}, \mathbf{b}), (i'-1, \mathbf{b}' - e_j))$	δ_j

for $j \in D(\mathbf{b}') \backslash \varphi(D(\mathbf{b}))$ if that set is not equal to $\varnothing$, which implies $i < i'$.

Case 3: $i \leq k, i' > k$. From $((i, \mathbf{b}), (i', \mathbf{b}'))$, we make jumps

to state	with intensity
$((i-1, \mathbf{b} - e_j), (i'-1, \mathbf{b}'))$	$\delta_{j'}$
$((i-1, \mathbf{b} - e_j), (i', \mathbf{b}'))$	$\delta_j - \delta_{j'}$

for $j \in D(\mathbf{b})$ and

to state	with intensity
$((i, \mathbf{b}), (i'-1, \mathbf{b}'))$	$\sum_{j \in D(\mathbf{b}') \backslash \varphi(D(\mathbf{b}))} \delta_j$

Clearly, all possible jumps go to states $\in E \times E$ with the inequality $\preccurlyeq$ preserved.

We turn to (7(ii)). It is obvious how the HSR policy works. However, we have not defined what is meant by a routing policy in general: It is a rule that decides about an arrival customer, given the state of the system at the arrival. If there is at

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

most one station idle, the rule is the trivial one; there is no choice. The rule could mean a randomization over idle stations; we omit the formalism involving a Markov kernel on $E \times E$ and think of $\mathcal{S}'$ using a policy deviating from the HSR policy.

Now, let $(i, \mathbf{b})$ and $(i', \mathbf{b}')$ be the states of X_{T-} and X'_{T-}, respectively, and let $(i+1, \mathbf{b}^*)$ and $(i'+1, \mathbf{b}^{*\prime})$ denote the states at X_T and X'_T. We have to find a function $\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, \cdot)$ satisfying (6(ii)).

If $i' \geq k$, the arriving customer to $\mathcal{S}'$ at time T finds all stations busy. That means that we may let

$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, j) = j \quad \text{for } j \in D(\mathbf{b}^*).$$

Therefore, suppose that $i' < k$. Clearly, due to the HSR policy,

$$\mathbf{b}^* = \mathbf{b} + e_m,$$

where

$$m = \min\{j; j \in C(\mathbf{b})\}.$$

We observe that $\{1, \ldots, m-1\} \subseteq D(\mathbf{b})$.

Suppose that the policy for $\mathcal{S}'$ routes the customer arriving at time T to station ℓ. To start, suppose that $\ell < m$. Let

$$L = \{j \in D(\mathbf{b}); \varphi(\mathbf{b}, \mathbf{b}', j) < \ell\}.$$

Since $\varphi(\mathbf{b}, \mathbf{b}', \cdot)$ is strictly increasing on $D(\mathbf{b})$ and $\{1, \ldots, m-1\} \subseteq D(\mathbf{b})$, L is either empty or a set $\{1, \ldots, r\} \subseteq \{1, \ldots, m-1\}$ for some $r \geq 1$.

If $L = \varnothing$, let

$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, 1) = \ell,$$
$$\varphi(\mathbf{b}, \mathbf{b}^{*\prime}, j) = \varphi(\mathbf{b}, \mathbf{b}', j-1) \quad \text{for } j \in \{2, \ldots, m\},$$

and

$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, j) = \varphi(\mathbf{b}, \mathbf{b}', j) \quad \text{for } j > m.$$

If $L \neq \varnothing$, let

$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, j) = \varphi(\mathbf{b}, \mathbf{b}', j) \quad \text{for } j \in \{1, \ldots, r\},$$
$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, r+1) = \ell,$$
$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, j) = \varphi(\mathbf{b}, \mathbf{b}', j-1) \quad \text{for } j \in \{r+2, \ldots, m\},$$

and

$$\varphi(\mathbf{b}^*, \mathbf{b}^{*\prime}, j) = \varphi(\mathbf{b}, \mathbf{b}', j) \quad \text{for } j > m, j \in D(\mathbf{b}).$$

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

A similar proof applies to the case $\ell \geq m$; let r be the integer $\in \{0, \ldots, k-1\}$ such that $L = \{1, \ldots, r\} \cap D(\mathbf{b})$ ($= \varnothing$ if $r = 0$) and note that special attention has to be paid to the case $r \geq m$.

That completes the proof of (8), for an arrival process satisfying (4). However, the condition in (4) excluding multiple arrivals was made to avoid some trite arguments; with multiple arrivals, we extend (7) to cover the case: $T = 0$ and (7(i)) dropped. No obstacles appear.

For a random point process ξ of arrivals, let $S = (S_i)_1^\infty$ be its points of occurrence in increasing order. If ξ is independent of the service times involved, then (8) is seen to hold for the random arrival case, after a conditioning on S.

That completes the proof of the optimality of the HSR policy; we refrain from a statement in terms of a theorem.

Remarks:

1. Our analysis works just as well when there are infinitely many stations. In fact, it becomes easier: We may use B as the state space, and the distinction among Cases 1–3 is unnecessary. The same comment applies to loss systems (i.e., when no waiting is allowed).
2. If the arrival process is stationary and with sufficiently low intensity (the expected number of arrivals per time unit), then $(Y_t)_0^\infty$ and $(Y_t')_0^\infty$ can be stationary too and (1) can be used to justify the HSR policy when performing measures under stationarity are used.
3. It is natural to believe that (1) holds when $\rho_1 \geq \cdots \geq \rho_k$, where ρ_i is the failure rate function for the distribution of the service times at station i. However, even an effort under the extra condition that the ρ_i functions are all increasing, using some of the ideas in Lindvall [3] in order to keep track of the attained service times, was in vain.
4. For this remark, suppose that the HSR policy is used for both $\mathcal{S}$ and $\mathcal{S}'$ and that the arrivals appear according to a time-homogeneous Poisson process (which renders X_t, $t \geq 0$, a Markov process). It is clear from the proof of (8) that if $\mathcal{S}$ is idle at time 0, then $X_t \preccurlyeq X_t'$ for all $t \geq 0$, regardless of the initial configuration of $\mathcal{S}'$. That is the condition of Lindvall [2, Sect. IV.3, Thm. (9.3)] and we can conclude from that theorem that X_t is stochastically increasing in t if $\mathcal{S}$ is idle at time 0; the same monotonicity for Y_t follows.

3. USE FEW BUT QUICK STATIONS

In Lindvall [3, Sect. 6], it was proved that if we compare two $M/M/\cdot$ systems, with k and k' stations and service rates δ and δ', respectively, where $k < k'$ and $k\delta = k'\delta'$, then (1) holds if $Y_0 \leq Y_0'$, for a certain natural and simple coupling. That speaks in favor of the design principle: Use few but quick stations.

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

We will now generalize that result to the case where the service rates $\delta_1,\dots,\delta_k$ and $\delta'_1,\dots,\delta'_{k'}$ of the stations in $\mathcal{S}$ and $\mathcal{S}'$, respectively, satisfy

$$\begin{aligned}&\delta_1 \geq \cdots \geq \delta_k,\\ &\delta'_1 \geq \cdots \geq \delta'_{k'},\\ &\delta_i \geq \delta'_i \quad \text{for } 1 \leq i \leq k,\end{aligned} \tag{9}$$

and

$$\sum_1^k \delta_i \geq \sum_1^{k'} \delta'_i,$$

and the arrival process is arbitrary but independent of the service times, as in Section 2. It is assumed that the HSR policy is used for both $\mathcal{S}$ and $\mathcal{S}'$, but it suffices that $\mathcal{S}$ uses that policy. We will be as brief as possible, concentrating on the necessary modification of the setup in Section 2 and a new coupling problem that has to be solved.

Some new notation is needed. Let

$$B' = \{0,1\}^{k'} \tag{10}$$

and

$$E' = \{(i',\mathbf{b}');i' \in \mathbb{Z}_+,\mathbf{b}' \in B',\#D(\mathbf{b}') = i \wedge k'\};$$

we let $C(\mathbf{b}')$ and $D(\mathbf{b}')$ have their obvious meanings for $\mathbf{b}' \in B'$. For the matching in (6(ii)), the mapping $\varphi(\mathbf{b},\mathbf{b}',\cdot)$ now has $\{1,\dots,k'\}$ as its range. That means that (6(i)) and (6(ii)) no longer define a partial ordering on E, but rather a relation on $E \times E'$; that is, however, sufficient for our purposes. We use $\preccurlyeq^\circ$ to denote that relation.

We will produce (X_t, X'_t), $t \geq 0$, such that

$$X_t \preccurlyeq^\circ X'_t \quad \text{for all } t \geq 0 \tag{11}$$

when $X_0 \preccurlyeq^\circ X'_0$. Clearly, that implies (1).

We will pay no attention to the analog of (7(ii)), since the proof of that is virtually the same as for the routing problem. For the analog of (7(i)), let $(i,\mathbf{b}) \in E$ and $(i',\mathbf{b}') \in E'$ be such that $(i,\mathbf{b}) \preccurlyeq^\circ (i',\mathbf{b}')$; we will produce couplings such that "inequality" is preserved when services are finished. Note that $i \leq i'$.

Case 1°: $i \leq k$. Use the obvious variants of Cases 2 and 3 in Section 2.

For the case $i > k$, we need couplings that are more subtle than those above. To that end, let

$$K_m = \left[\sum_1^{m-1}(\delta_l - \delta'_l), \sum_1^{m}(\delta_l - \delta'_l)\right] \quad \text{for } m = 1,2,\dots,k, \tag{12}$$

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

and

$$L_n = \left[\sum_{k+1}^{n-1} \delta_l', \sum_{k+1}^{n} \delta_l' \right] \quad \text{for } n = k+1, \dots, k'.$$

Further, let

$$\gamma_{mn} = \ell(K_m \cap L_n) \quad \text{for } m = 1, \dots, k,\ n = k+1, \dots, k'. \tag{13}$$

Here, $\ell(I)$ denotes the length of an interval I.

Case 2°: $i > k, i' = k'$. Let $\varphi(\mathbf{b}, \mathbf{b}', j) = j$ for $j = 1, \dots, k$. From $((i, \mathbf{b}), (i', \mathbf{b}'))$, we make jumps

to state	with intensity
$((i-1, \mathbf{b}), (i'-1, \mathbf{b}' - e_j'))$	δ_j'

for $j = 1, \dots, k$,

to state	with intensity
$((i-1, \mathbf{b}), (i', \mathbf{b}'))$	$\sum_1^k \delta_j - \sum_1^k \delta_j'$

and

to state	with intensity
$((i-1, \mathbf{b}), (i'-1, \mathbf{b}' - e_j'))$	δ_j'

for $j = k+1, \dots, k'$.

To understand the last intensity, consider γ_{mn} as the intensity for service finished both at station m in $\mathcal{S}$ and station n in $\mathcal{S}'$. We have $\delta_j' = \sum_{m=1}^k \gamma_{mj}$ due to the condition $\sum_1^k \delta_j \geq \sum_1^{k'} \delta_j'$. Because there is a customer waiting in $\mathcal{S}$, it does not become visible which station in $\mathcal{S}$ finishes service at the same time as station j in $\mathcal{S}'$.

We used e_j' to denote the jth k'-dimensional unit vector.

Case 3°: $i > k, i' > k'$. This case is easy, after understanding Case 2°.

Case 4°: $i > k, i' < k'$. Let $n_1', \dots, n_m'$ be the elements of $D(\mathbf{b}')$ enumerated in increasing order and let

$$\varphi(\mathbf{b}, \mathbf{b}', j) = n_j' \quad \text{for } 1 \leq j \leq k.$$

Now use an obvious variant of Case 2°.

For Cases 1°–4°, we find that the relation $\preccurlyeq^\circ$ is preserved after jumps; hence (11) is proved.

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press

References

1. Asmussen, S. (1987). *Applied probability and queues*. Chichester: Wiley.
2. Lindvall, T. (2002). *Lectures on the coupling method*. Mineola, NY: Dover.
3. Lindvall, T. (2002). Comparisons between certain queuing systems. *Probability in the Engineering and Informational Sciences* 16: 1–17.
4. Sparaggis, P.D., Towsley, D., & Cassandras, C.G. (1993). Extremal properties of the shortest/longest non-full queue policies in finite-capacity systems with state-dependent service rates. *Journal of Applied Probability* 30: 223–236.
5. Stoyan, D. (1983). *Comparison methods for queues and other stochastic models*. Chichester: Wiley.
6. Winkler, P. (2001). Optimality and greed in dynamic allocation. *Journal of Algorithms* 41(2): 244–261.
7. Wolff, R.W. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs, NJ: Prentice-Hall.

https://doi.org/10.1017/S0269964804183046 Published online by Cambridge University Press