Mathematical Methods in Operations Research 2005
Special issue in honor of *Arie Hordijk*

Routing to parallel homogeneous queues

Ger Koole

Vrije Universiteit Amsterdam

Received: December 2004 / Revised version: January 2005

Abstract We give an overview of results on routing to parallel homogeneous queues, some of which were up to now unavailable in the open literature.

1 Introduction

Arie Hordijk's work has been dominated by two research areas: queueing theory and stochastic dynamic programming. These were also the two advanced courses that Arie teached in Leiden while I was a Master student. It was at the intersection of these fields that I worked as a PhD student, and I still work on them, more than 15 years later. The first results that we obtained were on routing to parallel queues. The main result seems quite simple: just route to the shortest queue. In this short paper I'll show that there is more to that, by giving an overview of results concerning routing to parallel homogeneous queues. Crucial for a classification of results is the information that optimal policies are allowed to use: is the action a function of the queue lengths or even the workloads, or are these unknown? The more information, the better the decision is. In the next section we introduce the model.

2 Model formulation and results

We consider customers arriving according to some point process. All results can be generalized to any process, for simplicity we only consider Poisson processes. The service time of all customers has the same distribution. For a given realization of the process, we denote the service time of customer n with s_n. There are m queues with identical servers. At the arrival of the ith customer, queue j has x_j^n customers (including the one in service), and the total workload in the queue is w_j^n. In this paper we restrict to minimizing average workload. This is often equivalent to minimizing the average number of customers in the system, which in turn, according to Little's law, is equivalent to minimizing the average waiting time.

Bernoulli policies Consider the class of policies that does not depend on any of the variables s_n, w^n or x^n and for which we have the additional condition that the decision rules all have the same distribution on the actions (decision rules are allowed to randomize). The policy that selects each queue with the same probability m^{-1} is optimal in the sense that it minimizes the expected workload at any time. This follows readily from the next theorem that deals with a single queue with general arrival and service process, but arrivals are only admitted with a certain probability.

Theorem 1 *The expected workload in a single server queue is, at any time, a convex function of the admission probability.*

Proof We couple 4 systems, with admission probabilities λ, twice $\lambda + \delta$, and $\lambda + 2\delta$, such that $0 \leq \lambda \leq \lambda + 2\delta \leq 1$, with respective workloads $w_t^\lambda, w_t^{\lambda+\delta}, \tilde{w}_t^{\lambda+\delta}, w_t^{\lambda+2\delta}$ at time t. We take the arrival and service times the same for all 4 processes. The admission processes are coupled by the use of a uniform distribution for each arrival instant. For a realization u the admission is as follows. The arrival is admitted in all systems if $u \leq \lambda$, it is rejected in all systems if $u > \lambda + 2\delta$, it is admitted in the $w_t^{\lambda+\delta}$ and $w_t^{\lambda+2\delta}$-systems if $u \in (\lambda, \lambda + \delta]$ and it is admitted in the $\tilde{w}_t^{\lambda+\delta}$ and $w_t^{\lambda+2\delta}$-systems if $u \in (\lambda + \delta, \lambda + 2\delta]$. Now assume that at some t the following holds:

$$w_t^\lambda \leq w_t^{\lambda+\delta}, \tilde{w}_t^{\lambda+\delta} \leq w_t^{\lambda+2\delta}; \tag{1}$$

$$w_t^{\lambda+\delta} + \tilde{w}_t^{\lambda+\delta} \leq w_t^\lambda + w_t^{\lambda+2\delta}. \tag{2}$$

Note that they hold for an empty system. Now consider the system at $t+s$, before the next arrival. At that moment the workload is $(w_t^\bullet - s)^+$. By checking all possible values of s it is readily seen that (1)-(2) still hold. Now assume that an arrival occurs. Thanks to the way we coupled the admission it can be seen again that (1)-(2) remain valid. This shows that (1)-(2) hold for all t. The convexity follows from (2). QED

Other results on this type of *Bernoulli policies* can be found in [1,2,5].

Static policies We extend the class of policies in the following way: policies do not depend on any of the variables s_n, w^n or x^n, but the decision rules are allowed to differ at different times. For an initially empty system *cyclic* or *round-robin routing*, i.e., $(1, \ldots, m, 1, \ldots, m, \ldots)$, is optimal. For a proof see [9, Prop. 8.3.4], for exponential service times, or [8], for *ILR* service times.

Policies based on queue length Now we go to a class of policies where information on the queue lengths x^n and the elapsed service times are available. The latter is relevant in the case of non-exponential service times; see [6] for a positive result for non-exponential service time, and [10] for some counterexamples. Here we concentrate on exponential service times, then the elapsed service time gives no additional information. Routing to the shortest queue is optimal, shown first in [11], but there is a remarkably simple dynamic programming (dp) proof. Note that the expected workload is simply the number of customers times the expected service time. The dp equation for the numbers of customers in the queues, after uniformization ([7]), with the usual notation and $\lambda + m\mu = 1$, is given by:

$$V_{n+1}(x) = \sum_{j=1}^{m} x_j + \lambda \min_{1\leq j\leq m} \Big\{ V_n(x+e_j) \Big\} + \mu \sum_{j=1}^{m} V_n((x-e_j)^+).$$

Using induction it can be shown that:

$$V_n(x+e_i) \leq V_n(x+e_j) \text{ for } x_i \leq x_j;$$

$$V_n(x) = V_n(x') \text{ with } x' \text{ a permutation of } x;$$

$$V_n(x) \leq V_n(x+e_1).$$

The (in)equalities can be shown to hold for all n by induction, the optimality of routing to the shortest queue follows from the first inequality. A complete proof can be found in our first joint publication [3].

Policies based on queue length and workload Our next set of policies routes customers on the basis of workloads and queue lengths. The queue lengths do not give additional information on the evolution and direct costs of the process, therefore it suffices to consider the workloads. Again we formulate the dp equation, now for general service times. With u_n we denote the nth interarrival time, numbered backwards, and with F the distribution function of the service time. Then the dp equation for the workloads reads:

$$V_{n+1}(w) = \sum_{j=1}^{m} w_j + \min_{1\leq j\leq m} \Big\{ \int_0^\infty V_n((w+se_j-u_n e)^+)dF(s) \Big\}.$$

Similar equations as for the case with queue lengths hold for the workload case:

$$\int V_n(w+se_i)dF(s) \leq \int V_n(w+se_j)dF(s) \text{ for } w_i \leq w_j; \qquad (3)$$

$$V_n(w) = V_n(w') \text{ for } w' \text{ a permutation of } w;$$

$$V_n(w) \leq V_n(w+se_1) \text{ for } s \geq 0.$$

Equation (3) shows that routing to the queue with the shortest workload is optimal. This equation should not be confused with $V_n(w+se_i) \leq V_n(w+$

se_j) for some or all s. Having the integral inside the minimization means that the routing decision is made *not knowing the service time of the arriving customer*. If this is the case we get the dp equation

$$V_{n+1}(w) = \int_0^\infty \min_{1\leq j\leq m} \Big\{ V_n((w + se_j - u_n e)^+) \Big\} dF(s).$$

It is easy to construct counterexamples against the optimality of the shortest workload policy for this case. This was to be expected, given the relation with the NP-complete deterministic machine scheduling problem ([4]).

Simulation We end the paper by reporting on simulation experiments executed for $m = 5$, exponential service times with unit mean, and varying load. By variance analysis we verified that all digits are significant. Bernoulli with thinning leads to 5 independent M/M/1 queues, shortest workload is equivalent to one M/M/5 queue. We verified the simulation for these policies using theoretical results. We see that the expected workload decreases as the amount of information increases.

load	Bernoulli	cyclic	shortest queue	shortest workload
50%	5.00	3.49	2.77	2.63
80%	20.0	12.6	7.0	6.2

Acknowlegdment I thank Floske Spieksma for some useful comments.

References

1. C.S. Chang. A new ordering for stochastic majorization: Theory and applications. *Advances in Applied Probability*, 24:604–634, 1992.
2. M.B. Combé and O.J. Boxma. Optimization of static traffic allocation policies. *Theoretical Computer Science*, 125:17–43, 1994.
3. A. Hordijk and G.M. Koole. On the optimality of the generalized shortest queue policy. *Probability in the Engineering and Informational Sciences*, 4:477–487, 1990.
4. L.C.M. Kallenberg and G.M. Koole. Oral communication during PhD defense, 1992.
5. G.M. Koole. On the pathwise optimal Bernoulli routing policy for homogeneous parallel servers. *Mathematics of Operations Research*, 21:469–476, 1996.
6. G.M. Koole, P.D. Sparaggis, and D. Towsley. Minimizing response times and queue lengths in systems of parallel queues. *Journal of Applied Probability*, 36:1185–1193, 1999.
7. S.A. Lippman. Applying a new device in the optimization of exponential queueing systems. *Operations Research*, 23:687–710, 1975.
8. Z. Liu and D. Towsley. Optimality of the round-robin routing policy. *Journal of Applied Probability*, 31:466–475, 1994.
9. J. Walrand. *An Introduction to Queueing Networks.* Prentice-Hall, 1988.
10. W. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34:55–62, 1986.
11. W.L. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14:181–189, 1977.