Queueing Systems Vol. 100
Special Issue: 100 views on queues

Routing with Too Much Information?

Esa Hyytiä · Peter Jacko · Rhonda Righter

the date of receipt and acceptance should be inserted later

1 Introduction

An important problem with many applications is routing jobs to parallel processors with dedicated queues to minimize job response times or delays. Often, job sizes are known or well-estimated upon arrival, and the router can use this information to minimize delays. If job sizes, as well as routing history, are known, then the router also knows the current states at the servers without message passing. Somewhat surprisingly, the problem with job-size information available is much harder than the problem without this information, and there are very few theoretical results to provide general guidelines. Counter-examples show that many "intuitively obvious" results are in fact not true.

2 Routing to homogeneous FCFS queues

Optimal routing to homogeneous FCFS queues has been extensively investigated for a variety of scenarios with varying amounts of partial information. When job sizes are unknown, depending on the state information, variants of "Join the Smallest Expected Work" policies, such as Round-Robin (RR) [6,16,17], Join the Shortest Queue (JSQ) [18,19], and Least Work Left (LWL), [1,4,8,15], are often optimal. On the other hand, when we know arriving job sizes, but assume the server states are unknown, Size Interval Task Assignment (SITA), in which jobs of similar size are routed to the same server, has been shown to be optimal [7]. Multi-level SITA and RR policies, that also ignore state information, have been explored [2,13]. We focus on FCFS queues, but note that

Esa Hyytiä
Department of Computer Science, University of Iceland. E-mail: esa@hi.is

Peter Jacko
Department of Management Science, Lancanster University Management School, and UK & Berry Consultants, UK. E-mail: p.jacko@lancaster.ac.uk

Rhonda Righter
Department of Ind. Eng. and Opns. Res., UC Berkeley, USA. E-mail: rrighter@berkeley.edu

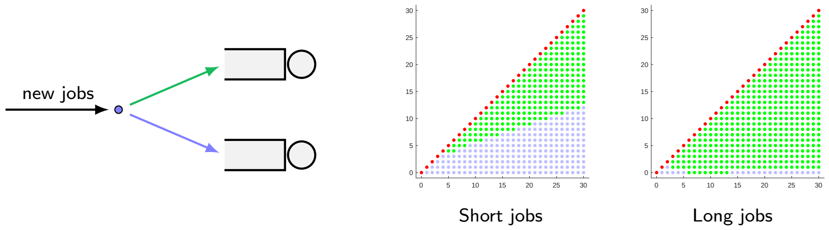

Fig. 1 Optimal routing for short and long jobs to two servers (green, blue, and red correspond to routing to queue 1, queue 2, and either queue). The axes represent the work of the two queues, with $(0,0)$ in the bottom-left corner.

optimal routing to queues following shortest remaining processing time (SRPT) is also an open problem. (See [5] and [9] for preliminary results.)

In sum, we have nice optimality results for routing to FCFS queues when job sizes are known, but we ignore server state information, or when job sizes are unknown but we have server state information. What can we say about the optimal policy when we know job sizes and use them to keep track of server state information? Heuristics based on first policy iteration have been investigated [10,12,14], but there is no theory to evaluate them.

3 What does the optimal routing policy look like and how good are heuristics?

We have investigated properties of the optimal policy for the special case of two servers (so the only actions in a given state are LWL and MWL - Most Work Left) and a two-point job-size distribution in discrete time. That is, in each time slot a job of size i arrives with probability p_i, $i = 0, 1, m$. Under FCFS, the state information is $s = (u_S, u_L, x)$, where $x \in \{1, m\}$ is the current job size, and $u_S \leq u_L$ (for Smallest and Largest) are the ordered current backlogs (work), at the servers. A two-point job size distribution provides a starting point for understanding more general size distributions.

Through counter-examples, we know that LWL is not optimal, even for the "short" (size-1) jobs. For "long" (size-m) jobs, MWL is often optimal, even when the "short" queue is idle. Indeed, "unbalancing" the workload with MWL for long jobs, and sometimes even for short jobs, may benefit enough later arriving short jobs to be optimal.

Structural results for the optimal policy seem to be equally elusive. For example, it seems that if LWL is optimal for a short job in state $(u_S, u_L, 1)$, then this would also be the case in state $(u_S, u'_L, 1)$, for $u'_L \geq u_L$, but this does not necessarily hold. That is, there is not a single threshold in u_L, for fixed u_S and x, that determines the optimality of LWL, nor is LWL necessarily optimal when $u_L \to \infty$. We also considered the amount of workload imbalance, $\Delta = u_L - u_S$. Again, reasonable conjectures for structural results based on this measure do not necessarily hold. For example, holding Δ and x fixed, we do not have monotonicity in terms of u_s in terms of when to follow LWL.

Our numerical examples show that generally, if MWL is optimal in state $(u_S, u_L, 1)$, then it is also optimal in state (u_S, u_L, m), but, again, this is not always true. On the other hand, all of our examples support the conjecture of a threshold in u_S, for fixed x and u_L, such that MWL is more likely to be optimal for larger values of u_S. Also, our counter-examples to simple policies being optimal are generally for "rare" states, and the differences in the value functions for the two actions in anomalous states are generally small, suggesting that simple heuristics may be "nearly optimal."

We were able to show that the heuristic of LWL (MWL) for short (long) jobs in all states is asymptotically optimal in heavy traffic. Can this be extended to more general service time distributions, or more than two servers? Note that heuristics such as LWL are sub-optimal in this limit even for exponential jobs. For a given heuristic, can we determine which performance metrics it optimizes, and under what conditions? How do we mathematically evaluate heuristics, such as first-policy-iteration, that seem to perform well in practice? And how do we determine the value of job size information, in terms of comparing good policies that use it to policies like JSQ or LWL?

References

1. Akgun, O., Righter, R., and Wolff, R.: Partial flexibility in routing and scheduling. Adv. Appl. Prob. **45**, 673–691 (2013)
2. Anselmi, J.: Combining size-based load balancing with round-robin for scalable low latency. IEEE Trans. on Parallel and Distributed Syst. **31**, 886–896 (2020)
3. Bansal, N. and Harchol-Balter, M.: Analysis of SRPT scheduling: Investigating unfairness. Proc. ACM Sigmetrics '01 (2001)
4. Daley, D.J.: Certain optimality properties of the first-come first-served discipline for G/G/s queues. Stoch. Proc. and Appl. **25**, 301–308 (1987)
5. Down, D. and Wu, R.: Multi-layered round robin routing for parallel servers. Queueing Syst., **53**, 177–188 (2006)
6. Ephremides, A., Varaiya, P., and Walrand, J.: A simple dynamic routing problem, IEEE Trans. Aut. Cont. **25**(4), 690–693 (1980)
7. Feng, H., Misra, V. and Rubenstein, D.: Optimal state-free, size-aware dispatching for heterogeneous M/G/-type systems. Perf. Eval. **62**(1-4), 475–492 (2005)
8. Foss, S.G.: Approximation of multichannel queueing systems. Sib. Math. J. **21**, 851–857 (1980)
9. Grosof, I., Scully, Z., Harchol-Balter, M.: Load balancing guardrails: keeping your heavy traffic on the road to low response times Proc. ACM Meas. Anal. Comput. Syst., **3**(2) 42:1-31 (2019)
10. Hyytiä, E.: Lookahead actions in dispatching to parallel servers. Perf. Eval. **70**, 859–872 (2013)
11. Hyytiä, E., Penttinen, A., and Aalto, S. Size- and state-aware dispatching problem with queue-specific job sizes. European J. Oper. Res. **217**(2), 357–370 (2012)
12. Hyytiä, E. and Righter, R.: Routing jobs with deadlines to heterogeneous parallel servers. Opns. Res. Lett. **44**, 507–513 (2016)
13. Hyytiä, E. and Righter, R.: STAR and RATS, Multi-level dispatching policies. Int. Teletraffic Conf., **ITC 32**, 81–89 (2020)
14. Hyytiä, E., Righter, R., and Aalto, S.: Task assignment in a heterogeneous server farm with switching delays and general energy-aware cost structure. Perf. Eval. **75-76**, 17-35 (2014)
15. Koole, G.: On the optimality of FCFS for networks of multi-server queues. CWI, Amsterdam, TR BS-R9235, 1992.
16. Liu, Z. and Towsley, D.: Optimality of the round-robin routing policy. J. Appl. Prob. **31**(2), 466–475 (1994)
17. Liu, Z. and Righter, R.: Optimal load balancing on distributed homogeneous unreliable processors. Opns. Res. **46**(4), 563–573 (1998)
18. Weber, R.R.: On the optimal assignment of customers to parallel servers. J. Appl. Prob. **15**(2), 406–413 (1978)
19. Winston, W.: Optimality of the shortest line discipline. J. Appl. Prob. **14**, 181–189 (1977)