Exact Analysis

1. S.H. Xu and J.G. Shanthikumar, Optimal expulsion control - A dual approach to admission control of an ordered-entry system, Oper. Res. 41 (1993) 1137-1152.

2. Xu, S.H. A duality approach to admission and scheduling controls of queues. Queueing Syst 18, 273–300 (1994).

This paper studies theadmission andscheduling control problem in an queueing system with nonidentical processors. Admission control renders when a newly arrived job should be accepted, whereas scheduling control determines when an available processor should be utilized. The system received a rewardR when a job completes its service and pays a unit holding costC while a job is in the system. The main goal of the paper is to obtain the admission/scheduling policy that maximizes the expected discounted and long-run average profits (reward minus cost). We convert the system into its dual, a stochastically identical system subject toexpulsion/scheduling control, and prove that the individually optimal policy in the dual system is socially optimal in the original system. In contrast with the dynamic programming (DP) technique which considers the system as a whole, we adopt the viewpoint of an individual job and analyze the impact of its behavior on the social outcome. The key properties which simplify the analysis are that under the individually optimal policy the profit of a job under the preemptive last-come first-priority service discipline (LCFP-P) is independent of jobs arrived earlier than itself and that the system is insensitive to service discipline imposed. The former makes possible to bypass complex dynamic programming analyses and the latter serves as a vehicle in connecting the social and individual optimality. We also exploit system operational characteristics under LCFP-P to obtain simple and close approximations of the optimal thresholds.

3. Susan H. Xu and Y. Quennel Zhao (1996) Dynamic Routing and Jockeying Controls in a Two-Station Queueing System. Advances in Applied Probability 28(4): 1201-1226.

This paper studies optimal routing and jockeying policies in a two-station parallel queueing system. It is assumed that jobs arrive to the system in a Poisson stream with rate and are routed to one of two parallel stations. Each station has a single server and a buffer of infinite capacity. The service times are exponential with server-dependent rates, and . Jockeying between stations is permitted. The jockeying cost is cij when a job in station jockeys to station , . There is no cost when a new job joins either station. The holding cost in station is , , per job per unit time. We characterize the structure of the dynamic routing and jockeying policies that minimize the expected total (holding plus jockeying) cost, for both discounted and long-run average cost criteria. We show that the optimal routing and jockeying controls are described by three monotonically non-decreasing functions. We

study the properties of these control functions, their relationships, and their asymptotic behavior. We show that some well-known queueing control models, such as optimal routing to symmetric and asymmetric queues, preemptive or non-preemptive scheduling on homogeneous or heterogeneous servers, are special cases of our system.

4. Kim, J., Ahn, H., & Righter, R. (2011). Managing Queues With Heterogeneous Servers. Journal of Applied Probability, 48(2), 435-452.

5. Özkan, E., & Kharoufeh, J. (2014). Optimal Control of a Two-Server Queueing System with Failures. Probability in the Engineering and Informational Sciences, 28(4), 489-527.

We consider the problem of controlling a two-server Markovian queueing system with heterogeneous servers. The servers are differentiated by their service rates and reliability attributes (i.e., the slower server is perfectly reliable, whereas the faster server is subject to random failures). The aim is to dynamically route customers at arrival, service completion, server failure, and server repair epochs to minimize the long-run average number of customers in the system. Using a Markov decision process model, we prove that it is always optimal to route customers to the faster server when it is available, irrespective of its failure and repair rates, if the system is stable. For the slower server, there exists an optimal threshold policy that depends on the queue length and the state of the faster server. Additionally, we analyze a variant of the main model in which there are multiple unreliable servers with identical service rates, but distinct reliability characteristics. For that case it is always optimal to route customers to idle servers, and the optimal policy is insensitive to the servers' reliability characteristics.

Heavy Traffic Analysis

1. Kelly, F.P. & Laws, C.N. 1993. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. Queueing Systems 13: 47–86.

We present an introductory review of recent work on the control of open queueing networks. We assume that customers of different types arrive at a network and pass through the system via one of several possible routes; the set of routes available to a customer depends on its type. A route through the network is an ordered set of service stations: a customer queues for service at each station on its route and then leaves the system. The two methods of control we consider are the routing of customers through the network, and the sequencing of service at the stations, and our aim is to minimize the number of customers in the system. We concentrate especially on the insights which can be obtained from heavy traffic analysis, and in particular from Harrison's Brownian network models. Our main conclusion is that in many respects dynamic routing simplifies the behaviour of networks,

and that under good control policies it may well be possible to model the aggregate behaviour of a network quite straightforwardly.

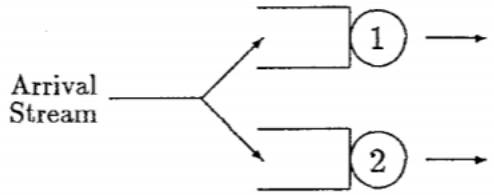
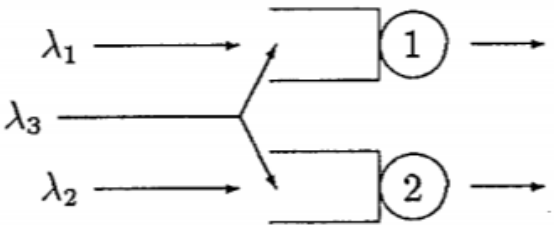


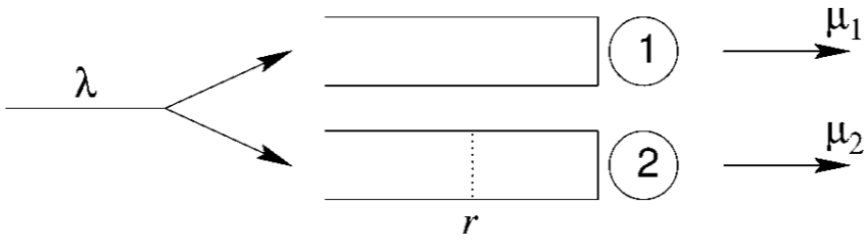
Fig. 1. Two parallel queues.

Summary

> 0. Asymptotic analysis
> 1. Symmetric queue:
>
> * JSQ is optimal under Poisson arrival and exponential service, by Winston; extended to arbitrary arrival and service times with non-decreasing failure rate by Weber.
>
> * However, Whitt showed that JSQ is not optimal for general service times even with Poisson arrival process. **Intuition**: there exists a service time distribution such that when the difference between the numbers in the two queues is small, the longer queue is likely to have a sudden series of departures and hence be a better choice.
>
> * **Other interesting results**: In heavy traffic, the two-server system behaves like a pooled system.
> 2. SED rule: join the queue with the shorter expected delay. Whitt showed that it is not optimal in general even for Poisson arrival and exponential service times. However, Houck showed via simulation that it is close to being optimal in many cases.
> 3. Reiman, heavy traffic analysis
>
> 
>
>
> Fig. 7. Parallel queues.

2. Teh, Y., Ward, A.R. Critical Thresholds for Dynamic Routing in Queueing Networks. Queueing Systems 42, 297–316 (2002).

This paper studies dynamic routing in a parallel server queueing network with a single Poisson arrival process and two servers with exponential processing times of different rates. Each customer must be routed at the time of arrival to one of the two queues in the network. We establish that this system operating under a threshold policy can be well approximated by a one-dimensional reflected Brownian motion when the arrival rate to the network is close to the processing capacity of the two servers. As the heavy traffic limit is approached, thresholds which grow at a logarithmic rate are critical in determining the behavior of the limiting system. We provide necessary and sufficient conditions on the growth rate of the threshold for (i) approximation of the network by a reflected Brownian motion (ii) positive recurrence of the limiting Brownian diffusion and (iii) asymptotic optimality of the threshold policy.



Two queues in parallel with a single discretionary arrival stream and unequal service rates $\mu_1 > \mu_2$. Arrivals are routed to queue 1 unless queue 2 is below the threshold r.

Summary

> 0. Asymptotic analysis
> 1. Poisson arrival, homogeneous customers, exponential processing times with rates and .
> 2. Policies of interest: A simple threshold policy routes customers to queue 1 unless queue 2 is shorter than an associated threshold r, in which case arrivals are routed to queue 2.

3. Stolyar, A. (2005). OPTIMAL ROUTING IN OUTPUT-QUEUED FLEXIBLE SERVER SYSTEMS. Probability in the Engineering and Informational Sciences, 19(2), 141-189.

We consider a queuing system with multitype customers and nonhomogeneous flexible servers, in the heavy traffic asymptotic regime and under a complete resource pooling (CRP) condition. For the input-queued (IQ) version of such a system (with customers being queued at the system “entrance,” one queue per each type), it was shown in the work of Mandelbaum and Stolyar that a simple parsimonious scheduling rule is optimal in that it asymptotically minimizes the system customer workload and some strictly convex queuing costs. In this article, we consider a different—output-queued (OQ)—version of the model, where each arriving customer must be assigned to one of the servers immediately upon arrival. We introduce the MinDrift routing rule for OQ systems (which is as

simple and parsimonious as Gcμ) and show that this rule, in conjunction with arbitrary work-conserving disciplines at the servers, has asymptotic optimality properties analogous to those Gcμ rule has for IQ systems. A key element of the analysis is the notion of system server workload, which, in particular, majorizes customer workload. We show that (1) the MinDrift rule asymptotically minimizes server workload process among all OQ-system disciplines and (2) this minimal process matches the minimal possible customer workload process in the corresponding IQ system. As a corollary, MinDrift asymptotically minimizes customer workload among all disciplines in either the OQ or IQ system.

Summary

> 0. Asymptotic analysis
> 1. types of customers; nonhomogeneous flexible servers; service rate , depends on **both customer type and server**
> 2. Cost function: strictly convex
> 3. Main results: optimality of the MinDrift Rule

Heuristics

1. Argon, N. T., Ding, L., Glazebrook, K. D., & Ziya, S. (2009). Dynamic routing of customers with general delay costs in a multiserver queuing system. Probability in the Engineering and Informational Sciences, 23(2), 175-203.

We consider a network of parallel service stations each modeled as a single-server queue. Each station serves its own dedicated customers as well as generic customers who are routed from a central controller. We suppose that the cost incurred by a customer is an increasing function of her time spent in the system. In a significant advance on most previous work, we do not require waiting costs to be convex, still less linear. With the objective of minimizing the long-run average waiting cost, we develop two heuristic routing policies, one of which is based on dynamic programming policy improvement and the other on Lagrangian relaxation. In developing the latter policy, we show that each station is “indexable” under mild conditions for customers’ waiting costs and also prove some structural results on the admission control problem that naturally arises as a result of the Lagrangian relaxation. We then test the performance of our heuristics in an extensive numerical study and show that the Lagrangian heuristic demonstrates a strong level of performance in a range of traffic conditions. In particular, it clearly outperforms both a greedy heuristic, which is a standard proposal in complex routing problems, and a recent proposal from the heavy traffic literature.

Summary

> 1. Multiple parallel M/M/1 queues with genetic and dedicated customers

2. General holding cost function, dependent on the server and the number of customers in the queue, **NOT on the customer's type**.
3. Objective: minimize long-run average waiting cost by optimally routing an incoming genetic customer to one of the parallel queues
4. Contribution: two heuristics, single-step policy improvement and Lagrangian relaxation
5. Future direction: Another direction for future research would be to consider multiple types of generic customers each differing in their waiting costs. Such a generalization would be of particular interest to call centers that serve a heterogeneous group of users and seek ways of providing a better service for their more “valuable” customers. It would also be of interest in health care operations, for which patients have significantly different waiting cost structures depending on their health conditions.

2. K. D. Glazebrook, C. Kirkbride, J. Ouenniche. (2009) Index Policies for the Admission Control and Routing of Impatient Customers to Heterogeneous Service Stations. Operations Research 57(4) pp. 975-989.

We propose a general Markovian model for the optimal control of admissions and subsequent routing of customers for service provided by a collection of heterogeneous stations. Queue-length information is available to inform all decisions. Admitted customers will abandon the system if required to wait too long for service. The optimisation goal is the maximisation of reward rate earned from service completions, net of the penalties paid whenever admission is denied, and the costs incurred upon every customer loss through impatience. We show that the system is indexable under mild conditions on model parameters and give an explicit construction of an index policy for admission control and routing founded on a proposal of Whittle for restless bandits. We are able to gain insights regarding the strength of performance of the index policy from the nature of solutions to the Lagrangian relaxation used to develop the indices. These insights are strengthened by the development of performance bounds. Although we are able to assert the optimality of the index heuristic in a range of asymptotic regimes, the performance bounds are also able to identify instances where its performance is relatively weak. Numerical studies are used to illustrate and support the theoretical analyses.

Summary

1. SMDP model: admission control and routing impatient customers for service
2. Each new customer must be either refused admission or routed to one of the stations; **customers are homogeneous**.
3. Station : service rate , loss rate , both depend on the number of customers .
4. A service completion at station generates a reward , a loss incurs a penalty , and a rejection incurs a penalty .
5. Objective: maximize the average net reward per unit of time over an infinite horizon.
6. Index heuristic and performance bound.

3. Ding, L., Glazebrook, K.D. Dynamic routing in distinguishable parallel queues: an application of product returns for remanufacturing. OR Spectrum 35, 585–608 (2013).

This paper deals with the dynamic routing of product returns in distinguishable parallel queues. Several vendors alongside an original equipment manufacturer are available in provision of remanufacturing service. Each has its own queue. The stream of the product returns follow a stochastic process. A central controller is employed to decide to which vendor an incoming product is sent to avoid excessive queues in front of some vendors and idle servers in the others. We develop models and index-based heuristics to support the dynamic routing decisions so as to minimize the overall recovering costs. The product concerned exemplifies a short-life cycle due to, for example, technology advance. Long delay during the remanufacturing process will render a substantial deterioration of reselling prices. Hence, in the paper we contend that the cost incurred for remanufacturing a product should take explicit account of the impact of long delays in the lead time. Both theoretical and simulation studies demonstrate the effectiveness of the Restless Bandit approach deployed to the dynamic routing of product returns among multiple vendors.

Summary

> This paper studies an application of the index policies; heuristic plus simulation.