Atar, Rami; Shifrin, Mark
An asymptotic optimality result for the multiclass queue with finite buffers in heavy traffic.
(English) Zbl 1310.60124
Stoch. Syst. 4, No. 2, 556-603 (2014).

Summary: For a multiclass G/G/1 queue with finite buffers, admission and scheduling control, and holding and rejection costs, we construct a policy that is asymptotically optimal in the heavy traffic limit. The policy is specified in terms of a single parameter which constitutes the free boundary point from the Harrison-Taksar free boundary problem, but otherwise depends explicitly on the problem data. The $c\mu$ priority rule is also used by the policy, but in a way that is novel, and, in particular, different than that used in problems with infinite buffers. We also address an analogous problem where buffer constraints are replaced by throughput time constraints.

MSC:
60K25 Queueing theory (aspects of probability theory)
93E20 Optimal stochastic control
60F17 Functional limit theorems; invariance principles
60J60 Diffusion processes
90B22 Queues and service in operations research

Keywords:
multiclass G/G/1 queue; Brownian control problem; Harrison-Taksar free boundary problem; state dependent priorities; Reiman’s snapshot principle

Full Text: DOI arXiv Euclid

References:
[1] Ata, B., Dynamic control of a multiclass queue with thin arrival streams. Operations Research, 54(5):876-892, 2006. · Zbl 1167.90439 · doi:10.1287/opre.1060.0308
[2] Ata, B. and Kumar, S., Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. Ann. Appl. Probab., 15(1A):333-391, 2005. · Zbl 1071.60081 · doi:10.1214/105051604000000495
[3] Ata, B. and Olsen, T. L., Near-optimal dynamic lead-time quotation and scheduling under convex-concave customer delay costs. Operations Research, 57(3):753-768, 2009. · Zbl 1235.90177 · doi:10.1287/opre.1080.0608
[4] Atar, R., A diffusion regime with nondegenerate slowdown. Oper. Res., 60(2):490-500, 2012. · Zbl 1248.90038 · doi:10.1287/opre.1110.1030
[5] Atar, R., A diffusion regime with nondegenerate slowdown: Appendix. Electronic companion, downloadable at, 2012. · Zbl 1248.90038 · doi:10.1287/opre.1110.1030
[6] Atar, R. and Budhiraja, A., Singular control with state constraints on unbounded domain. Ann. Probab., 34(5):1864-1909, 2006. · Zbl 1106.93038 · doi:10.1214/009117906000000359
[7] Atar, R., Budhiraja, A., and Williams, R. J., HJB equations for certain singularly controlled diffusions. Ann. Appl. Probab., to appear, 2013. · Zbl 1320.60150 · doi:10.1214/13-AAP935
[8] Atar, R. and Gurvich, I., Scheduling parallel servers in the non-degenerate slowdown diffusion regime: Asymptotic optimality results. Ann. Appl. Probab., to appear, 2013. · Zbl 1320.60150 · doi:10.1214/13-AAP935
[9] Atar, R., Mandelbaum, A., and Reiman, M. I., Scheduling a multi class queue with many exponential servers: Asymptotic optimality in heavy traffic. Ann. Appl. Probab., 14(3):1084-1134, 2004. · Zbl 1049.60079 · doi:10.1214/105051604000000023
[10] Atar, R. and Solomon, N., Asymptotically optimal interruptible service policies for scheduling jobs in a diffusion regime with nondegenerate slowdown. Queueing Systems Theory Appl., 69(217-235), 2011. · Zbl 1244.90061 · doi:10.1007/s11134-011-9218-1
[11] Bell, S. L. and Williams, R. J., Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. Ann. Appl. Probab., 11(3):608-640, 2001. · Zbl 1015.60080 · doi:10.1214/aoap/1015345343
[12] Bennani, M. N. and Menascé, D. A., Resource allocation for autonomic data centers using analytic performance models. In Autonomic Computing, 2005. ICAC 2005. Proceedings. Second International Conference on, pages 229-240. IEEE, 2005.
[13] Billingsley, P., Convergence of Probability Measures. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. x+277 pp. A Wiley-Interscience Publication. · Zbl 0944.60003
Bharucha-Rezvani, Karim, Dynamic scheduling of a make-to-order, parallel-server system with cancellations. Oper. Res. 57(1):94-116, 2009.

Harrison, J. M. and Williams, R. J., Workload reduction of a generalized Brownian network. Ann. Appl. Probab., 15(4):2255-2295, 2005. · Zbl 1066.60036 · doi:10.1214/105051605000000458

Jacod, J. and Shiryaev, A. N., Limit Theorems for Stochastic Processes, volume 288 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1987. ISBN 3-540-17882-1. xviii+601 pp. · Zbl 0635.60021 · doi:10.1007/978-1-4613-8762-6

Budhiraja, Ajay, and Ghosh, A. P., Controlled stochastic networks in heavy traffic: Convergence of value functions. Ann. Appl. Probab., 22(2):734-791, 2012. · Zbl 1244.60091 · doi:10.1214/11-AAP784

Buyukkoc, Varaita, and Walrand, John, The (c, m) rule revisited. Adv. in Appl. Probab., 17(1):237-238, 1985. · Zbl 0557.60082 · doi:10.2307/1427064

Cox, D. R. and Smith, W. L., Queues. Methuen's Monographs on Statistical Subjects. Methuen & Co. Ltd., London; John Wiley & Sons Inc., New York, 1961. xii+180 pp.

Dai, J. G. and Dai, W., A heavy traffic limit theorem for a class of open queueing networks with finite buffers. Queueing Systems Theory Appl., 32(1-3):5-40, 1999. · Zbl 0997.60038 · doi:10.1023/A:1009217903658

Dellacherie, C. and Meyer, P.-A., Probabilities and Potential , volume 29 of North-Holland Mathematics Studies. North-Holland Publishing Co., Amsterdam-New York, 1978. ISBN 0-7204-0701-X. viii+189 pp. · Zbl 0494.60001

Durrett, Richard, Probability: Theory and Examples. Duxbury Press, Belmont, CA, second edition, 1996. ISBN 0-534-24318-5.

Kruk, L. Lehoczky, J., Ramanan, K., and Shreve, S., An explicit formula for the Skorokhod map on [0,a]. Ann. Probab., 22(4):1712-1737, 1994. · Zbl 0911.90162 · doi:10.1214/aop/1024769131

Whitt, William, Weak convergence theorems for priority queues: Preemptive-resume discipline. J. Appl. Probability, 8:74-94, 1971. · Zbl 0215.33801 · doi:10.2307/321839

Williams, R. J., Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. Queueing Systems Theory Appl., 30(1-2):27-88, 1998. · Zbl 0911.90171 · doi:10.1023/A:1019108819713

Rubino, M. and Ata, B., Dynamic control of a make-to-order, parallel-server system with cancellations. Oper. Res., 57(1):94-108, 2009. · Zbl 1181.60138 · doi:10.1287/opre.1080.0532

Shihrin, M., Arai, R., and Cidon, I., Optimal scheduling in the hybrid-cloud. In IFIP/IEEE International Symposium on Integrated Network Management . IEEE, 2013.

Smith, W. E., Various optimizers for single-stage production. Naval Res. Logist. Quart., 3:59-66, 1956. · doi:10.1002/nav.3800030106

van Mieghem, J. M. and Williams, R. J., Brownian models of open queueing networks with heterogeneous customer populations. In Stochastic Differential Systems, Stochastic Control Theory and Applications (Minneapolis, Minn., 1986), volume 10 ofIMA Vol. Math. Appl., pages 147-186. Springer, New York, 1988. · Zbl 0691.60006 · doi:10.1007/978-1-4613-8762-6

Harrison, J. M., Brownian models of open processing networks: canonical representation of workload. Ann. Appl. Probab., 10(1):75-100, 2000. · Zbl 1031.60036 · doi:10.1214/aop/1019178265

Harrison, J. M., A broader view of Brownian networks. Ann. Appl. Probab., 13(3):1119-1150, 2003. · Zbl 1060.90020 · doi:10.1214/aoap/1050516057

Harrison, J. M. and Taksar, M. I., Instantaneous control of Brownian motion. Math. Oper. Res., 8(3):439-453, 1983. · Zbl 0523.90068 · doi:10.1287/moor.8.3.439

Harrison, J. M. and Van Mieghem, J. A., Dynamic control of Brownian networks: space state collapse and equivalent workload formulations. Ann. Appl. Probab., 7(3):747-777, 1997. · Zbl 0885.60080 · doi:10.1214/aoap/1055426355

Harrison, J. M. and Williams, R. J., Workload reduction of a generalized Brownian network. Ann. Appl. Probab., 15(4):2255-2295, 2005. · Zbl 1096.60036 · doi:10.1214/105051605000000458

Reiman, M. I., The heavy traffic diffusion approximation for sojourn times in Jackson networks. In Applied Probability-Computer Science: The Interface, Vol. II (Boca Raton, Fla., 1981), volume 3 ofProgr. Comput. Sci. and Systems, Stochastic Control Theory and Applications (Minneapolis, Minn., 1986), volume 10 ofIMA Vol. Math. Appl., pages 147-186. Springer, New York, 1988. · Zbl 0691.60006 · doi:10.1007/978-1-4613-8762-6

Whitt, William, Weak convergence theorems for priority queues: Preemptive-resume discipline. J. Appl. Probability, 8:74-94, 1971. · Zbl 0215.33801 · doi:10.2307/321839

Williams, R. J., Diffusion approximations for open multiclass queueing networks: Sufficient conditions involving state space collapse. Queueing Systems Theory Appl., 30(1-2):27-88, 1998. · Zbl 0911.90171 · doi:10.1023/A:1019108819713

Rubino, M. and Ata, B., Dynamic control of a make-to-order, parallel-server system with cancellations. Oper. Res., 57(1):94-108, 2009. · Zbl 1181.60138 · doi:10.1287/opre.1080.0532

Shihrin, M., Arai, R., and Cidon, I., Optimal scheduling in the hybrid-cloud. In IFIP/IEEE International Symposium on Integrated Network Management . IEEE, 2013.

Smith, W. E., Various optimizers for single-stage production. Naval Res. Logist. Quart., 3:59-66, 1956. · doi:10.1002/nav.3800030106

van Mieghem, J. A., Dynamic scheduling with convex delay costs: the generalized (c,m) rule. Ann. Appl. Probab., 5(3):809-833, 1995. · Zbl 0843.90047 · doi:10.1214/aop/1177004706

Ward, A. R. and Kumar, S., Asymptotically optimal admission control of a queue with impatient customers. Math. Oper. Res., 33(1):167-202, 2008. · Zbl 1159.90017 · doi:10.1287/moor.1070.0287
This reference list is based on information provided by the publisher or from digital mathematics libraries. Its items are heuristically matched to zbMATH identifiers and may contain data conversion errors. It attempts to reflect the references listed in the original paper as accurately as possible without claiming the completeness or perfect precision of the matching.