Parallel-Server Stochastic Systems with Dynamic Affinity Scheduling and Load Balancing

Mark S. Squillante*
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, NY 10598, USA
mss@watson.ibm.com

We consider a parallel-server stochastic system in which a scheduling policy directs customers to the server where they inherently can be served most efficiently (so-called affinity scheduling [3]) and in which a threshold-based scheduling policy (both sender-initiated, where overloaded servers transfer arriving customers to underloaded servers, and receiver-initiated, where underloaded servers transfer waiting customers from overloaded servers; refer to [4]) manages the fundamental tradeoff between balancing the workload among the servers and serving customers where they can be served most efficiently (so-called load balancing [4], as well as resource pooling [1]). This particular (symmetric) stochastic system arises in many applicationareas, such as parallel computing environments, multi-tiered Internet server environments, and various business applications.

We first establish the optimality of our general dynamic threshold-based scheduling policy within the context of fluid and diffusion limits. We then derive a matrix-analytic analysis and fixed-point solution of this parallel-server system under fairly general stochastic processes as input and across all traffic intensities, obtaining explicit results in several cases. This solution is shown to be asymptotically exact (in terms of the number of servers) under certain conditions. The results of this analysis also provide key insights into the fundamental probabilistic behavior of dynamic threshold-based policies in large-scale parallel-server stochastic systems. This includes numerically determining the optimal threshold values as a function of the workload, and demonstrating the potential for some unstable behavior under dynamic threshold-based scheduling policies if the thresholds are selected improperly.

Lastly, based on this analysis we consider the corresponding stochastic optimization problem of dynamically determining the threshold values that minimize expected sojourn time as a function of time-varying workloads.

References

[1] S. L. Bell and R. J. Williams. Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Annals of Applied Probability*, 11:608–649, 2001.

[2] R. D. Nelson and M. S. Squillante. Parallel-server stochastic systems with dynamic affinity scheduling and load balancing. Preprint, 2005.

[3] M. S. Squillante and E. D. Lazowska. Using processor-cache affinity information in shared-memory multiprocessor scheduling. *IEEE Transactions on Parallel and Distributed Systems*, 4(2):131–143, February 1993.

[4] M. S. Squillante and R. D. Nelson. Analysis of task migration in shared-memory multiprocessors. In *Proceedings of ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems*, pages 143–155, May 1991.

*Based mostly on joint work with Randolph D. Nelson [2], OTA Limited Partnership, One Manhattanville Road, Purchase, NY 10377, USA.