Scheduling in Multiserver Systems: Approaches and Open Problems (Part I)

Mor Harchol-Balter
Computer Science Department, Carnegie Mellon University

Server farms are ubiquitous in applications ranging from Web server farms to high-performance supercomputing systems to call centers. The popularity of the server farm architecture is understandable, as it allows for increased performance, while being cost-effective and easily scalable.

Given the prevalence of server farms, it is surprising that even at this late date so little is understood regarding their performance as compared with their single-server counterpart, particularly with respect to scheduling. Part of the problem is that there are at least three disjoint communities studying scheduling in server farms, including the SIGMETRICS community, the INFORMS community, and the SPAA/STOC/FOCS community, all of which have different approaches and goals. One of our goals in this tutorial is to make researchers aware of results in these different communities.

Our primary focus is the evaluation of different routing/dispatching policies in server farms. The emphasis will be on *intuition*, so that the talk is accessible to newcomers as well as old-timers. In surveying the newest results, we will also present some practical open problems.

Since server farms are composed of many individual servers, each operating under some scheduling policy, Part I of this tutorial will begin by briefly examining single-server systems, and the effect of scheduling therein. Here we will pay particular attention to the effect of heavy-tailed job size distributions witnessed in computer system environments [6, 1, 11], in determining which scheduling policies are most effective in practice. We will point out several counter-intuitive results, such as the fact that scheduling policies that favor short jobs may actually help long jobs as well [8, 12, 2], and the fact that scheduling results in closed system models can be very different from those in open system models [10].

We will then move on to studying server farm models representative of those used in supercomputing and manufacturing. These involve non-preemptive, First-Come-First-Serve (FCFS) scheduling at the individual servers. We will see that the mean response time of such FCFS server farms can vary by orders of magnitude depending on the routing/dispatching policy used for assigning jobs to servers [5]. We will question common wisdoms, like whether load should be balanced among identical servers [9, 3]. We will also discuss the benefits of cycle stealing in such models [7], and what one can do when the size of jobs isn't known a priori [4].

References

[1] P. Barford and M. E. Crovella. Generating representative Web workloads for network and server performance evaluation. In *ACM SIGMETRICS Conference*, pages 151–160, July 1998.

[2] P. Brown. Comparing FB and PS Scheduling Policies. In *Eighth Workkshop on Mathematical Performance Modeling and Analysis (MAMA 2006)*, June 2006.

[3] P. Glynn, M. Harchol-Balter, and K. Ramanan. Optimal Cutoffs for Size-Based Task Assignment in Heavy Traffic. Work in progress 2006.

[4] M. Harchol-Balter. Task Assignment with Unknown Duration. *Journal of the ACM*, 49(2):260–288, 2002.

[5] M. Harchol-Balter, M. Crovella, and C. Murta. On Choosing a Task Assignment Policy for a Distributed Server System. *IEEE Journal of Parallel and Distributed Computing*, 59:204 – 228, 1999.

[6] M. Harchol-Balter and A. Downey. Exploiting process lifetime distributions for dynamic load balancing. *ACM Transactions on Computer Systems*, 15(3), 1997.

[7] M. Harchol-Balter, C. Li, T. Osogami, A. Scheller-Wolf, and M. Squillante. Cycle Stealing under Immediate Dispatch Task Assignment. In *Fifteenth ACM Annual Symposium on Parallel Algorithms and Architectures (SPAA 03)*, pages 274–285, June 2003.

[8] M. Harchol-Balter, B. Schroeder, N. Bansal, and M. Agrawal. Size-based Scheduling to Improve Web Performance. *Transactions of Computer Systems*, 21(2):207–233, May 2003.

[9] M. Harchol-Balter and R. Vesilo. Optimal Cutoffs in Size-Based Task Allocation Systems. Work in progress 2006.

[10] B. Schroeder, A. Wierman, and M. Harchol-Balter. Closed versus Open System Models: a Cautionary Tale. In *Proceedings of Networked Systems Design and Implementation (NSDI 2006)*, pages 239–252, May 2006.

[11] A. Shaikh, J. Rexford, and K. G. Shin. Load-sensitive routing of long-lived ip flows. In *Proceedings of SIGCOMM*, September 1999.

[12] A. Wierman and M. Harchol-Balter. Classifying Scheduling Policies with respect to Unfairness in an M/GI/1. In *Proceedings of the ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)*, pages 238–249, June 2003.