A Better Model for Task Assignment in Server Farms: How Replication can Help

Mor Harchol-Balter
Carnegie Mellon University, Pittsburgh, PA, USA
harchol@cs.cmu.edu

ABSTRACT

An age-old problem in the design of server farms is the choice of the task assignment policy. This is the algorithm that determines how to assign incoming jobs to servers. Popular policies include Round-Robin assignment, Join-the-Shortest-Queue, Join-Queue-with-Least-Work, and so on.

While much research has studied assignment policies, little has taken into account server-side variability -- the fact that the server we choose might be temporarily and unpredictably slow. We show that when server-side variability dominates runtime, replication of jobs can be very beneficial. We introduce the Replication-d algorithm that replicates each arrival to d servers chosen at random, where the job is considered “done” as soon as the first replica completes. We provide an exact closed-form analysis of Replication-d.

We next introduce a much more general model, one which takes both the inherent job size distribution and the server-side variability into account. This is a departure from traditional queueing models which only allow for one “size” distribution. We propose and analyze a new task assignment policy, Replicate-Idle-Queue (RIQ), which is designed to perform well given these dual sources of variability.

CCS Concepts

- **Mathematics of computing** → Queueing theory
- **Mathematics of computing** → Stochastic processes
- *Mathematics of computing* → Probabilistic algorithms
- **Networks** → Network performance evaluation
- **Theory of computation** → Online algorithms
- *Theory of computation* → Parallel algorithms

Keywords Resource allocation; replication; power-of- d ; queueing; load balancing

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author(s). Copyright is held by the owner/author(s).

SIGMETRICS'16, June 14–18, 2016, Antibes Juan-Les-Pins, France.

ACM ISBN 978-1-4503-4266-7/16/06.

DOI: <http://dx.doi.org/10.1145/2896377.2901900>

Short Bio

Mor Harchol-Balter is a Professor of Computer Science at Carnegie Mellon. She received her Ph.D. from U.C. Berkeley in Computer Science, completed her postdoc at MIT, and joined CMU in 1999. From 2008-2011, she served as the Associate Department Head for Computer Science. Mor is a recipient of the McCandless Junior Chair, the NSF CAREER award, and several teaching awards, including the Herbert A. Simon Award.

Mor's work focuses on designing new resource allocation policies, including load balancing policies, power management policies, and scheduling policies, for distributed systems. She is known for both her new techniques in queueing-theoretic analysis and Markov chains, as well as her work in computer systems implementation. Mor has co-authored over 100 publications in top journals and conferences, including a textbook, *Performance Analysis and Design of Computer Systems*, published by Cambridge University Press, 2013.

Mor is heavily involved in the ACM SIGMETRICS / PERFORMANCE research community, where she has published over two dozen papers, and received multiple best paper awards, and where she served as Technical Program Chair in 2007 and as General Chair in 2013. She is best known for her enthusiastic talks and her many PhD students, most of whom are professors in top academic institutions.