Introduction

Special Issue: ACM SIGMETRICS 2015

Ayalvadi Ganesh¹ · Florian Simatos²

Published online: 18 July 2016

© Springer Science+Business Media New York 2016

The six papers in this special issue are based on a selection of top theoretical papers from the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, which took place in Portland, Oregon, in June 2015. This is the premier conference on the measurement and modelling of computer systems, and its technical program featured papers that covered both theory and applications from a wide variety of areas, including algorithms, communication networks, dynamics and control, optimization, performance analysis, resource allocation and scheduling, and stochastic modeling, among others. The program consisted of 40 papers selected from over 200 submissions from all over the world. We seek with this special issue to highlight some of the recent theoretical work on the mathematical analysis and modeling of computer systems.

The paper of Gardner et al. studies the use of redundant requests to reduce latency caused by queueing delays. It brings together a well-motivated problem, a simple algorithm and sophisticated analysis to make the argument for the benefits of replication, by presenting an exact analysis of a fairly broad class of such queueing models.

A related queueing model corresponds to splitting a task rather than replicating it; the chunks are submitted to different servers, and the last of these to complete (rather than the first) determines the service time of the task. This is the classical fork-join queueing model, and is studied in the paper of Rizk, Poloczek and Ciucu, who obtain stochastic bounds on delays in such a system.

The paper of Shah and de Veciana considers the performance analysis of queueing systems employing various fair scheduling policies, such as alpha-fairness or balanced

✉ Ayalvadi Ganesh
a.ganesh@bristol.ac.uk

¹ University of Bristol, Bristol, UK

² ISAE-SUPAERO, Toulouse, France

fairness. They also consider heterogeneity in jobs and servers. In a suitable asymptotic regime, they show that system performance is robust to the fairness criteria employed, and to a degree of heterogeneity.

The paper of Aalto, Lassila and Osti deals with optimal scheduling of jobs in a wireless downlink server, in which the channel states are time-varying. Extending previous work on non-anticipative policies (which assume job sizes are unknown), it considers the setting with known job sizes. Following the Whittle index approach, which involves a relaxation of the optimal stochastic control problem, it shows that the relaxed problem is indexable, i.e., that its optimal solution is a simple index policy.

Scheduling in the context of a wireless network rather than a single server is the topic of the paper of Li and Srikant. The objective here is not optimizing performance but maximizing the stability region. The Backpressure algorithm is a canonical example of such a scheduling policy. The paper of Li and Srikant presents a new algorithm for this problem that uses only per-link rather than per-destination queues at each node, and still achieves the full capacity region.

The paper of Gast and van Houdt presents an analysis of a class of cache replacement algorithms under the Independent Reference Model (IRM) for requests. The analysis covers both steady-state and transient performance, and both exact results and more easily computable bounds are presented. Based on the insights yielded by the analysis, they also show how to choose a good algorithm within this class, and demonstrate that it outperforms the well-known least recently used (LRU) algorithm.

We are grateful to the authors for amending and expanding their conference papers into journal versions tailored to the Queueing Systems community. The papers have gone through a regular Queueing Systems review cycle, and we are indebted to the anonymous referees not only for the quality of their work, but also for their assistance in working with our tight timeline.