Scheduling for Efficiency and Fairness in Systems with Redundancy

Kristen Gardner*

Computer Science Department, Carnegie Mellon University

Mor Harchol-Balter

Computer Science Department, Carnegie Mellon University

Esa Hyytiä

Department of Computer Science, University of Iceland

Rhonda Righter

IEOR Department, UC Berkeley

Abstract

Server-side variability—the idea that the same job can take longer to run on one server than another due to server-dependent factors—is an increasingly important concern in many queueing systems. One strategy for overcoming server-side variability to achieve low response time is *redundancy*, under which jobs create copies of themselves and send these copies to multiple different servers, waiting for only one copy to complete service. Most of the existing theoretical work on redundancy has focused on developing bounds, approximations, and exact analysis to study the response time gains offered by redundancy. However, response time is not the only important metric in redundancy systems: in addition to providing low overall response time, the system should also be *fair* in the sense that no job class should have a worse mean response time in the system with redundancy than it did in the system before redundancy is allowed.

In this paper we use *scheduling* to address the simultaneous goals of (1) achieving low response time and (2) maintaining fairness across job classes. We develop new exact analysis for per-class response time under First-Come First-Served (FCFS) scheduling for a general type of system structure; our analysis shows that FCFS can be unfair in that it can hurt non-redundant jobs. We then introduce the Least Redundant First (LRF) scheduling policy, which we prove is optimal with respect to overall system response time, but which can be unfair in that it can hurt the jobs that become redundant. Finally, we introduce the Primaries First (PF) scheduling policy, which is provably fair and also achieves excellent overall mean response time.

Key words: queueing theory, redundancy, replication, scheduling, stochastic processes, resource allocation

*Corresponding author

Email addresses: ksgardne@cs.cmu.edu (Kristen Gardner), harchol@cs.cmu.edu (Mor Harchol-Balter), esa@hi.is (Esa Hyytiä), rrighter@berkeley.edu (Rhonda Righter)