Reservation Strategies for Stochastic Jobs

Guillaume Aupy* , Ana Gainaru[1], Valentin Honoré , Padma Raghavan , Yves Robert[2], and Hongyang Sun[3]

[1]Mellanox – United States

[2]ROMA (ENS Lyon / CNRS / Inria Grenoble Rhône-Alpes) – CNRS : UMR5668, Laboratoire d'informatique du Parallélisme, École Normale Supérieure (ENS) - Lyon, INRIA – Laboratoire de l'Informatique du Parallélisme 46 Allée d'Italie 69364 Lyon, France

[3]Laboratoire de l'Informatique du Parallélisme (LIP) – École Normale Supérieure (ENS) - Lyon, INRIA, CNRS : UMR5668, PRES Université de Lyon, Université Claude Bernard - Lyon I (UCBL) – 46 Allée d'Italie 69364 LYON CEDEX 07, France

Abstract

We are interested in scheduling stochastic jobs on a reservation-based platform. Specifically, we consider jobs whose execution time follows a known probability distribution. The platform is reservation-based, meaning that the user has to request fixed-length time slots. The cost depends on both the request duration and the actual execution time of the job. A reservation strategy is a sequence of increasing-length reservations, which are paid for until one of them allows the job to successfully complete. The goal is to minimize the total expected cost of the strategy. I will present different scheduling strategies and properties of an optimal solution.

Keywords: stochastic applications, hpc, runtime system, resource manager

*Speaker