Contribution ID: **120** Type: **not specified**

Online Demand Scheduling with Failovers

Friday, July 14, 2023 10:55 AM (20 minutes)

Konstantina Mellou, Marco Molinaro and Rudy Zhou

Abstract: Motivated by cloud computing applications, we study the problem of how to optimally deploy new hardware subject to both power and robustness constraints. To model the situation observed in large-scale data centers, we introduce the \emph{Online Demand Scheduling with Failover} problem. There are m identical devices with capacity constraints. Demands come one-by-one and, to be robust against a device failure, need to be assigned to a pair of devices. When a device fails (in a failover scenario), each demand assigned to it is rerouted to its paired device (which may now run at increased capacity). The goal is to assign demands to the devices to maximize the total utilization subject to both the normal capacity constraints as well as these novel failover constraints. These latter constraints introduce new decision tradeoffs not present in classic assignment problems such as the Multiple Knapsack problem and AdWords.

In the worst-case model, we design a deterministic $\approx \frac{1}{2}$-competitive algorithm, and show this is essentially tight. To circumvent this constant-factor loss, which in the context of big cloud providers represents substantial capital losses, we consider the stochastic arrival model, where all demands come i.i.d. from an unknown distribution. In this model we design an algorithm that achieves a sub-linear additive regret (i.e. as opt or m increases, the multiplicative competitive ratio goes to 1). This requires a combination of different techniques, including a configuration LP with a non-trivial post-processing step and an online monotone matching procedure introduced by Rhee and Talagrand.

Presenter: ZHOU, Rudy

Session Classification: Track A-2