Fair workload distribution for multi-server systems with pulling strategies

Andrea Marin
Università Ca' Foscari Venezia, Italy
marin@dais.unive.it

Sabina Rossi
Università Ca' Foscari Venezia, Italy
srossi@dais.unive.it

ABSTRACT

The problem of parallel job scheduling has been widely studied in the literature with the aim of improving some performance indices, such as the throughput, the response time, the fairness or a combination of these indices (see, e.g., [3]). One can distinguish two basic approaches to the problem depending on the phase at which the dispatcher is placed. The two approaches are: *dispatching before the queues* and *dispatching after the queue*. In the first approach, the dispatcher decides how to assign a job to a server according to some scheduling discipline which takes into account the queueing state and other information about the servers. In contrast, the second approach consists in storing the jobs in a shared queue and these are assigned by the dispatcher to a server as soon as it becomes available. The scheduling policy can rely on a pushing strategy, i.e., the dispatcher takes the initiative to send a job to a specific server or on a pulling strategy, i.e., the servers decide autonomously to fetch a job from the shared queue.

In this paper we study the second approach relying on a pulling strategy. In contrast to prior works, we focus on the balance of the total number of jobs served by a set of K identical servers. We propose a stateless scheduling discipline implemented by the servers so that the difference between the number of jobs served by each unity is finite in steady-state. Our scheduling discipline is based on a server rate-adaptation algorithm. Informally, each server maintains a variable to store the difference between the total number of jobs served by itself and a neighbour. Given just this piece of information, the server may decide to slow down its maximum service speed in order to reduce this difference. As soon as the server finishes its job, it fetches a new one from the queue. We study two rate-adaptation strategies. The first one, named *bimodal* strategy, uses only two distinct service rates: the highest is used when a server has served less jobs than its neighbour, while the slowest is used otherwise. The second rate-adaptation policy, named *proportional* strategy, requires a server to reduce a fixed maximum service rate in proportion to the number of extra jobs it has served with respect to its neighbour. We propose a Markovian model for such a scheduling discipline and for both the rate-adaptation strategies described above. We show that, despite the little knowledge that each server has about the state of the system, we can derive a necessary and sufficient condition for the job balance index to have finite expectation in the bimodal strategy whereas in the proportional strategy the job balance is unconditionally finite. We derive the exact expressions for two relevant performance indices: the system's throughput and the balance index. The latter measures the differences among the total number of jobs processed by each server, hence low values imply a well balanced system. Although the Markov process underlying the models has an infinite state space, these expressions involve finite sums derived from the evaluation of hypergeometric functions.

Our findings show that maintaining reasonable low values for the balance index reduces the throughput at around 70% of the maximum. More interestingly, numerical evidences show that this value scales slowly with the number of servers, which means that the rate adaptation policy scales well with the system's size. From a theoretical point of view, to analyse our model we resort to the notion of *ρ-reversibility* [2]. Indeed, we show that although the model is in general *not* reversible [1], it satisfies the Kolmogorov's criteria for ρ-reversibility allowing us to derive an analytical product-form expression for the invariant measure.

DOI: 10.1145/1235

CCS Concepts

•**Computing methodologies** → *Modeling and simulation;*

Keywords

Fork-join queueing systems, Markov models, load balancing, rate adaptation

1. REFERENCES

[1] F. Kelly. *Reversibility and stochastic networks.* Wiley, New York, 1979.

[2] A. Marin and S. Rossi. On the relations between Markov chain lumpability and reversibility. *Acta Informatica*, 2016. Available online.

[3] I. Tsimashenka, W. Knottenbelt, and P. Harrison. Controlling variability in split-merge systems and its impact on performance. *Annals of Oper. Res.*, 239:569–588, 2016.