Mean Field Analysis of Join-Below-Threshold Load Balancing for Resource Sharing Servers

Illés Antal Horváth
MTA-BME Information Systems
Research Group
Budapest, Hungary
horvath.illes.antal@gmail.com

Ziv Scully
Carnegie Mellon University
Pittsburgh, PA, USA
zscully@andrew.cmu.edu

Benny Van Houdt
University of Antwerp
Antwerp, Belgium
benny.vanhoudt@uantwerpen.be

ABSTRACT

Load balancing plays a crucial role in many large scale computer systems. Much prior work has focused on systems with First-Come-First-Served (FCFS) servers. However, servers in practical systems are more complicated. They serve multiple jobs at once, and their service rate can depend on the number of jobs in service. Motivated by this, we study load balancing for systems using Limited-Processor-Sharing (LPS). Our model has heterogeneous servers, meaning the service rate curve and multiprogramming level (limit on the number of jobs sharing the processor) differs between servers. We focus on a specific load balancing policy: Join-Below-Threshold (JBT), which associates a threshold with each server and, whenever possible, dispatches to a server which has fewer jobs than its threshold. Given this setup, we ask: how should we configure the system to optimize objectives such as mean response time? Configuring the system means choosing both a load balancing threshold and a multiprogramming level for each server. To make this question tractable, we study the many-server mean field regime.

In this paper we provide a comprehensive study of JBT in the mean field regime. We begin by developing a mean field model for the case of exponentially distributed job sizes. The evolution of our model is described by a differential inclusion, which complicates its analysis. We prove that the sequence of stationary measures of the finite systems converges to the fixed point of the differential inclusion, provided a unique fixed point exists. We derive simple conditions on the service rate curves to guarantee the existence of a unique fixed point. We demonstrate that when these conditions are not satisfied, there may be multiple fixed points, meaning metastability may occur. Finally, we give a simple method for determining the optimal system configuration to minimize the mean response time and related metrics.

While our theoretical results are proven for the special case of exponentially distributed job sizes, we provide evidence from simulation that the system becomes insensitive to the job size distribution in the mean field regime, suggesting our results are more generally applicable.

CCS CONCEPTS

• **Mathematics of computing → Probability and statistics**; • **Networks → Network performance modeling**.

KEYWORDS

load balancing; mean field model; large scale computer systems; heterogeneous networks; differential inclusion; limited processor sharing

ACM Reference Format:
Illés Antal Horváth, Ziv Scully, and Benny Van Houdt. 2020. Mean Field Analysis of Join-Below-Threshold Load Balancing for Resource Sharing Servers. In *ACM SIGMETRICS / International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS '20 Abstracts), June 8–12, 2020, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3393691.3394199

1 INTRODUCTION

Most studies on large-scale load balancing systems consider servers that operate in a first-come-first-served (FCFS) manner and have a fixed service rate. In such a setting a simple policy such as the Join-Idle-Queue (JIQ) policy achieves asymptotic zero delay in both homogeneous and heterogeneous clusters [3]. However, servers in practical systems are more complex as they serve multiple jobs at once and need to share resources (such as caches).

Motivated by this we consider a cluster of so-called *resource sharing servers.* In a resource sharing server the throughput tends to increase as more jobs start sharing the server, but starts decreasing beyond some point due to contention and thrashing [1]. Hence, in such a server the total service rate $\mu(k)$ depends on the number of jobs k being processed simultaneously. In addition a limit, called the multi-programming level (MPL), is implemented that dictates how many jobs can equally share the server. When the number of jobs in a server exceeds its MPL, some jobs must wait before their service starts. In other words the server operates in a limited processor sharing (LPS) fashion, except that the total service rate also depends on the number of jobs in service. As stated before the rates $\mu(k)$ tend to increase up to some point $\kappa = \operatorname{argmax}_k \mu(k)$, and decreases afterward. In practice the MPL is often set equal to κ, though this may not be optimal in case of an isolated resource sharing server (see [1, 4]).

Our paper [2] studies the Join-Below-Threshold (JBT) dispatch policy in a heterogeneous cluster of resource sharing servers. In JBT a threshold is associated with every server and the server informs the dispatcher whenever its queue length either reaches or drops below its threshold. This allows the dispatcher to maintain a list of all the server IDs for which the server's queue length

SIGMETRICS '20 Abstracts, June 8–12, 2020, Boston, MA, USA
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7985-4/20/06.
https://doi.org/10.1145/3393691.3394199

is below its associated threshold. Incoming jobs are assigned by the dispatcher to a random server with a queue length below its threshold, unless no such server exists, in which case the job is assigned to a random server. If we set the threshold of a server equal to its MPL, this policy is a natural generalization of the JIQ policy for resource sharing servers as the dispatcher keeps track of all the servers where the incoming job can immediately start service. It also retains the advantage of JIQ that the communication overhead is upper bounded by one message per job.

2 CONTRIBUTIONS

The main contributions of our paper [2] are the following:

(1) We develop a mean field model to assess the performance of JBT in a heterogeneous cluster of resource sharing servers assuming exponential job sizes, the evolution of which is described by a differential inclusion (DI).
(2) We prove that as the number of servers tends to infinity, the sample paths of the stochastic systems converge towards the unique solution of the DI and the fixed point of the DI corresponds to the weak limit of the stationary measures of the stochastic systems provided that this fixed point is unique.
(3) Contrary to our initial expectations, we show that for arbitrary service rate curves multiple fixed points can exist (even in a homogeneous cluster), which implies that the system is metastable in such case. We identify simple conditions for the existence of a unique fixed point that we expect to hold for real service rate curves.
(4) We indicate how to determine the thresholds that minimize the mean response time (and related metrics) for a given arrival rate. This turns out to be surprisingly easy provided that we do not necessarily assign the same threshold to all the servers of the same type.
(5) We perform simulation experiments that support our belief that the queue length distribution becomes insensitive to the job size distribution as the system size tends to infinity. This indicates that our results are also relevant in systems with highly variable job sizes.

ACKNOWLEDGMENTS

Ziv Scully was supported by an ARCS Foundation scholarship and the NSF Graduate Research Fellowship Program under Grant Nos. DGE-1745016 and DGE-125222. Benny Van Houdt was supported by FWO contract number G033119N.

REFERENCES

[1] V. Gupta and M. Harchol-Balter. Self-adaptive admission control policies for resource-sharing systems. *SIGMETRICS Perform. Eval. Rev.*, 37(1):311–322, June 2009.
[2] Illés Antal Horváth, Ziv Scully, and Benny Van Houdt. Mean field analysis of join-below-threshold load balancing for resource sharing servers. *Proc. ACM Meas. Anal. Comput. Syst.*, 3(3), December 2019.
[3] A.L. Stolyar. Pull-based load distribution in large-scale heterogeneous service systems. *Queueing Systems*, 80(4):341–361, 2015.
[4] M. Telek and B. Van Houdt. Response time distribution of a class of limited processor sharing queues. *SIGMETRICS Perform. Eval. Rev.*, 45(3):143–155, March 2018.