Redundancy in Processor Sharing servers

E. Anton[1,3], U. Ayesta[1,2,3,4], M. Jonckheere[5] and I.M. Verloop[1,3]

[1] CNRS, IRIT, 2 rue Charles Camichel, 31071 Toulouse, France
[2] IKERBASQUE - Basque Foundation for Science, 48011 Bilbao, Spain
[3] Université de Toulouse, INP, 31071 Toulouse, France
[4] UPV/EHU, University of the Basque Country, 20018 Donostia, Spain
[5] Instituto de Cálculo - Conicet, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (1428) Pabellón II, Ciudad Universitaria Buenos Aires, Argentina

Keywords : *redundancy, heterogeneous servers, processor sharing.*

1 Introduction

The main motivation to investigate redundancy models comes from empirical evidence suggesting that redundancy can help improve the performance of real-world applications, see for instance [2]. Under redundancy-d, a job that arrives to this system is dispatched to d servers uniformly chosen at random in order to benefit from the variability of the queue length of these servers. As soon as one of the copies finishes service, the job (and its copies) is removed from the system. As a consequence, a job's delay is given by the minimum delay among the servers its copies are sent to.

We consider a parallel-server system where Processor Sharing (PS) is employed in each of the servers. Jobs have exponentially distributed service times and all the copies of a job have the same size, i.e., identical copies. Under these assumptions, Anton et al. [1] prove that when servers have homogeneous capacities, i.e., all have the same computation capacity, the stability region dramatically degrades as the number of copies d increases. In this extended abstract, we characterize the stability condition when servers have heterogeneous capacities. The main conclusion from our work is that when server capacities are sufficiently heterogeneous, redundancy significantly improves the stability region of the system. Moreover, we obtain sufficient conditions under which redundancy-d improves the stability region of the system compared to the system where there is no redundancy, $d = 1$.

2 Model description

We consider a K parallel-server system. Servers have capacities μ_k, $k = 1, \ldots, K$, where w.l.o.g. $\mu_1 < \ldots < \mu_K$. Each server implements PS, and has its own queue. Jobs arrive to the system according to a Poison process of rate λ. Upon arrival, each job chooses d servers out of K uniformly at random and sends a copy to each of these servers. We consider that all the copies of a job are exact replicas, that is, all the copies have the same size. The jobs are exponentially distributed with unit mean. The job departs the system as soon as one of the d copies has completed service.

We denote by $S = \{1, \ldots, K\}$ the set of all servers. Each job is labelled with a type c that represents the subset of d servers to which the copies of the jobs are sent, i.e., $c = \{s_1, \ldots s_d\}$, where $s_1, \ldots, s_d \in S$ and $s_i \neq s_j$, for all $i \neq j$. We denote by $\mathcal{C}$ the set of all types, i.e, $\mathcal{C} = \{c = \{s_1, \ldots, s_d\} \subset S : s_i \neq s_j, i \neq j\}$ and $|\mathcal{C}| = \binom{K}{d}$.

We denote by $N_c(t)$ the number of type-c jobs at time t. For the i-th type-c job, let b_{ci} denote the realization of the service requirement of this job, $i = 1, \ldots, N_c(t)$, $s \in c$. Since copies are identical, there is a departure of a job from the server where it has attained most

service so far. We let $a_{cis}(t)$ denote the attained service in server s of the i-th type-c job at time t and by $A_c(t) = (a_{cis}(t))_{is}$ a matrix on $\mathbb{R}_+$. Hence, the Markovian descriptor of the system is $\{N_c(t), A_c(t), c \in \mathcal{C}\}_{t\geq 0}$.

3 Stability condition

The stability condition of the system is that of the system where each type of job is dispatched to the server with maximum capacity that servers this type of job. For instance, let us consider $i \in \{d, \dots, K\}$. The number of different job types that have server i as the server with maximum capacity that severs this type of job is $\binom{i-1}{d-1}$. Thus, in the latter system, server i has arrival rate $\lambda\binom{i-1}{d-1}/\binom{K}{d}$ and capacity μ_i. Hence, server i has a strictly negative drift if $\lambda\binom{i-1}{d-1}/\binom{K}{d} < \mu_i$.

In the following proposition we characterize the stability condition of this system. The proof is based on fluid-scaling techniques. We first prove that the fluid limit of the number of copies per server in the system is Harris recurrent. Then, we conclude that the stochastic system is as well stable.

Proposition 1 *The system is stable if* $\lambda\frac{\binom{i-1}{d-1}}{\binom{K}{d}} < \mu_i$, *for* $i = d, \dots, K$. *The system is unstable if there exists* $i \in \{d, \dots, K\}$ *such that* $\lambda\frac{\binom{i-1}{d-1}}{\binom{K}{d}} > \mu_i$.

Remark 1 *In the case where servers have homogeneous capacities,* $\mu_i = \mu\ \forall i$, *the stability condition of the system is given by* $\lambda < \mu\frac{K}{d}$, *see [1].*

In the following, we consider the system with heterogeneous servers and obtain sufficient conditions for which the stability region is improved under redundancy-d, compared to the system where there is no redundancy, $d = 1$. We note that in the case where there is no redundancy and jobs are uniformly routed over the K servers, the stability condition is $\lambda/K < \mu_1$. The set of conditions in Proposition 1 is equivalent to $\lambda < \min_{i=d}^{K}\{\mu_i\binom{K}{d}/\binom{i-1}{d-1}\}$.

Corollary 1 *The system under redundancy-d has larger stability region than the system with no redundancy if* $\mu_1 d < \mu_d$.

Hence, if there exists d such that $\mu_1 d < \mu_K$, then adding d redundant copies to the system improves its stability region.

We further analyze the scenario where redundancy degrades the stability region of the system.

Corollary 2 *The system under no redundancy has largest stability region if* $\min_{d=2}^{K}\left\{\mu_d\binom{K}{d}\right\} < \mu_1 K$.

We note that when $\mu_K/\mu_1 \approx 1$, redundancy degrades the stability region of the system. This condition states that even if servers have heterogeneous capacities, these are not heterogeneous enough in order to ensure that redundancy improves the stability region of the system. This result is in agreement with the result obtained in [1] (see Remark 1), where the authors show that when servers are homogeneous the stability condition dramatically degrades when adding redundant copies.

References

[1] Elene Anton, Urtzi Ayesta, Matthieu Jonckheere and Ina Maria Verloop. *On the stability of redundancy models.* arXiv:1903.04414, 2019.

[2] Ashish Vulimiri, Philip Brighten Godfrey, Radhika Mittal, Justine Sherry, Sylvia Ratnasamy and Scott Shenker, *Low latency via redundancy. Proceedings of the ACM conference on Emerging networking experiments and technologies.* 283-294 (ACM).