The stability regions of preemptive and non-preemptive scheduling in randomly modulated queues

Nahuel Soprano-Loto[1*], Urtzi Ayesta[2], Matthieu Jonckheere[1], Ina Maria Verloop[2]

[1] CNRS, LAAS, Toulouse, France
[2] CNRS, IRIT, Toulouse, France

Key-words: *random modulations, scheduling, fluid analysis, preemptive*

1 Introduction

We consider a single-server multi-class system, where the service for each class is modulated by a dedicated dynamic ON-OFF environment. We analyze the role of timing by considering two types of scheduling: preemptive scheduling, where decisions can be made at any transition, and non-preemptive scheduling, where decisions can be made only at departure times. By explicitly describing the maximum stability regions in each case, we show that a significant impact on the stability occurs as a byproduct of these timing constraints. This contrasts with scenarios lacking modulations, where the well-established equivalence in terms of stability between preemptive and non-preemptive settings holds.

The preemptive case has been already studied in [1]. We introduce a novel fluid description that significantly simplifies the previous provided analysis, and that allows us to analyse also the non-preemptive case. In both cases, we show that serving the longest queue among the ones that are ON is a maximum stable policy.

2 Model description, main result

The model. Consider a single server system comprising K classes or queues, where the classes have their distinct arrival rates for incoming jobs, denoted by $\lambda_1, \ldots, \lambda_K$, and their distinct service rates, denoted by $\mu_1, \ldots, \mu_K$. Additionally, each queue operates within a two-state random environment characterized by ON and OFF states. When a queue's environment is in the OFF state, the queue is unable to receive service. The randomness governing the evolution of these environments may vary according to the class, and correlations between environments linked to different classes are allowed. We make the sole assumption that the overall environment, regarded as a collection of multiple environments across all queues, is ergodic. This assumption enables the establishment of a stationary distribution ν on the set of environments $\{(e_1, \ldots, e_K) : e_i \in \{\text{ON}, \text{OFF}\}\}$.

Preemptive vs non-preemptive. By simplicity, we restrict to policies that take decisions based solely on the current state of the system. The state contains the information about the environment or, in other words, observability of the environment is assumed.

In the preemptive case, for a fixed distribution of the environments, the maximum stability region is defined as

$$\text{MSR}_{\text{P}} = \{(\lambda_1, \ldots, \lambda_K, \mu_1, \ldots, \mu_K) : \exists \text{ a policy that stabilizes the system}\}. \tag{1}$$

Once a policy is fixed, we have a Markov process with countable state space, and stability is defined as positive recurrence.

In the non-preemptive case, the maximum stability region is called MSR_{NP}, and is defined as in (1) but running the set over non-preemptive policies.

*Main contributor: nahuel.soprano-loto@laas.fr

Serve the longest connected (SLC) policy. We define the Serve the longest connected (SLC) policy (see [2]) as the policy that serves the longest queue (the one with the highest number of waiting jobs) among the queues that are ON.

Main result. The following is our main result.

Theorem 1 *In the preemptive case,* $\mathrm{MSR_P}$ *is characterized by*

$$\sum_{i\in J}\rho_i < 1-\nu(e_i=\mathrm{OFF}\ \forall i\in J)\quad \forall J\subseteq\{1,\dots,K\}, \tag{2}$$

where $\rho_i=\frac{\lambda_i}{\mu_i}$ *stands for the load of the i-th queue. In the non-preemptive case,* $\mathrm{MSR_{NP}}$ *is characterized by*

$$\sum_{i=1}^{K}\frac{\rho_i}{\nu(e_i=\mathrm{ON})}<1. \tag{3}$$

In both cases, the maximum stability regions are attained by the SLC policy.

Proof idea: fluid limits. A simple rate-stability argument shows that, both in the preemptive and non-preemptive cases, the corresponding conditions are necessary to have stable policies. In the other direction, we need to prove that the SLC policy is stable under the parameter requirements, which is addressed by a fluid limit approach. More precisely, we prove that for any fluid limit, the derivative of the maximum fluid queue size remains below a negative constant. We describe how to do so only in the preemptive case, as the non-preemptive one is similar.

Let $U=(U_i)_i:[0,\infty)\to\mathbb{R}^K$ be a fluid limit. Assuming that $0\le t_1<t_2$ are such that

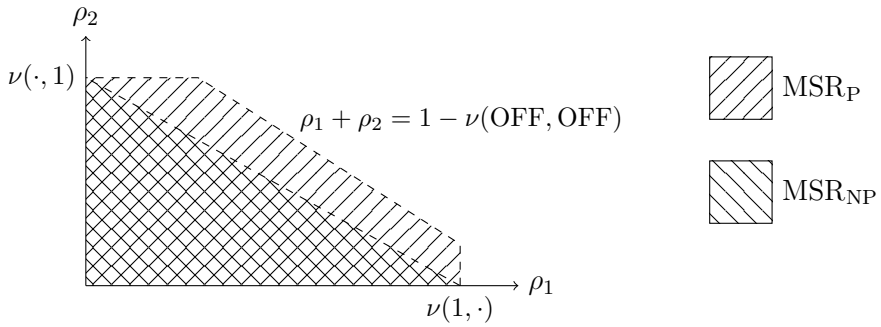
$$J:=\operatorname*{argmax}_{1\le i\le K}U_i(t_1)=\operatorname*{argmax}_{1\le i\le K}U_i(t_2)\quad\text{and}\quad\min_{i\in J}U_i(t)>\max_{i\notin J}U_i(t)\quad\forall t\in[t_1,t_2], \tag{4}$$

we are able to prove the following identity concerning the slope of the maximum:

$$\frac{\max_{i\in J}U_i(t_2)-\max_{i\in J}U_i(t_2)}{t_2-t_1}\sum_{i\in J}\frac{1}{\mu_i}=\sum_{i\in J}\rho_i-[1-\nu(e_i=\mathrm{OFF}\ \forall i\in J)]. \tag{5}$$

The attainment of this equality was possible thanks to the fact that, under hypotheses (4), we can precisely control the effective dedicated time of service to the queues in J. The aforementioned control over the derivative of the maximum now follows in conjunction with a series of formal additional steps, together with the fact that, due to (2), the r.h.s. of (5) is negative.

Graphical descriptions. To finish, we present a graphical representation of $\mathrm{MSR_P}$ and $\mathrm{MSR_{NP}}$ in the simple case $K=2$.



References

[1] N. Bambos and G. Michailidis. Queueing networks of random link topology: stationary dynamics of maximal throughput schedules. *Queueing Syst.*, vol. 50, no. 1, pp. 5–52, 2005.

[2] A. Ganti, E. Modiano and J. N. Tsitsiklis. Scheduling in Symmetric Communication Models With Intermittent Connectivity. *IEEE Transactions on Information Theory*, vol. 53, no. 3, pp. 998-1008, 2007.