Optimal Speed of a DVFS Processor under Soft Deadlines

Jonatha Anselmi, Bruno Gaujal, Louis-Sébastien Rebuffi

Optimal Speed of a DVFS Processor under Soft Deadlines

Jonatha Anselmi[1], Bruno Gaujal[1], Louis-Sébastien Rebuffi[1]

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France.

jonatha.anselmi.inria.fr,bruno.gaujal@inria.fr,

louis-sebastien.rebuffi@univ-grenoble-alpes.fr

1 Introduction

We consider a Dynamic Voltage and Frequency Scaling (DVFS) processor executing jobs with obsolescence deadlines. Forcing hard deadlines requires to consider the worst cases with only a finite number of jobs [2]. In this paper we use *soft* real-time constraints instead, i.e., jobs may miss their deadlines, at some cost. If they do, they are immediately removed from the system. The objective is to design a dynamic speed policy for the processor that minimizes its average energy consumption plus an obsolescence cost per deadline miss.

Under Poisson arrivals and exponentially distributed deadlines and job sizes, we show that this problem can be modeled as a continuous time Markov decision process (MDP) with unbounded state space and unbounded rates. Inspired by the scaling method introduced recently by Blok and Spieksma, we show the existence of an optimal speed profile that is increasing in the number of jobs in the system and upper bounded by some constant, that does not depend on the deadlines and arrival rates. In addition, it yields a simple approximation for the optimal policy and several numerical tests show that such approximation is accurate in heavy-traffic conditions.

2 Methodology and Main Result

We consider a DVFS processor whose speed can continuously vary in the interval $[0, S_{\max}]$. We consider that speed changes are immediate and induce no energy cost. When the processor works at speed s, it processes s units of work per second while its power dissipation is $w(s)$ watts. We require that $w(s)$ is continuous, increasing and strictly convex in the speed s.
Jobs join the system following a Poisson process with rate λ. Deadlines and job sizes are i.i.d. exponentially distributed random variables with with rates δ and μ respectively. Without loss of generality, we assume that $\mu = 1$. At any point in time t, the processor chooses its speed $s(t)$ and executes one of the jobs in its backlog queue.

2.1 Markov Decision Process

We formulate the problem of interest as an MDP whose state space is $\mathbb{N}$ and where the states represent the number of jobs in the system. The action space is $[0, S_{\max}]$, i.e., the set of available speeds for the DVFS processor. Let $\sigma = (\sigma_i)_{i\in\mathbb{N}}$ denote a stationary and deterministic speed policy adopted by the processor, i.e., $\sigma_i \in [0, S_{\max}]$ is the speed used in state i. It is well known that focusing on stationary and deterministic policies can be done with no loss of optimality in our case [3, Theorem 5.9].

The Markov chain X^σ is ergodic under all policies σ, therefore the long-run expected cost exists. Letting $\mathbb{E}^\sigma$ denote the expectation given a speed policy σ, we have

$$J(\sigma) := \limsup_{T\to\infty} \frac{1}{T} \int_0^T \mathbb{E}^\sigma c(X^\sigma(t), \sigma)\, \mathrm{d}t.$$

In this equation, the function $c(\cdot,\cdot)$ is the expected cost incurred by the system at time t: $c(i,\sigma) := Ci\delta + w(\sigma_i)$.

Stationary policies that minimize $J(\sigma)$ are optimal speed policies for the model. Also, our MDP satisfies all the conditions given in [3, Theorem 5.9] to assert the existence of an optimal stationary deterministic policy σ^* and an optimality equation of the form

$$J^* = J(\sigma^*) = \min_{s\in[0,S_{\max}]} c(i,s) + \sum_j h^*(j) q_{i,j}(s),\ \forall i \in \mathbb{N} \tag{1}$$

where h^* is a real function defined on $\mathbb{N}$, usually referred to as *bias* of the optimal policy.

2.2 Structural properties of the optimal policies

The following theorem is our main result.

Theorem 1 *There exists a deterministic optimal policy $\sigma^* = (\sigma_i^*)_{i\in\mathbb{N}}$ that is increasing in i and upper bounded by B, where:*

$$B := \operatorname*{arg\,min}_{s\in\mathbb{R}^+\cup\{+\infty\}} (w(s) + C(\lambda - \mu s)). \tag{2}$$

The optimal speed policy of the processor is always bounded by a finite constant, namely $\min(B, S_{\max})$. We remark that B is *independent of the arrival and deadline rate*. Indeed, if B is finite, one can set *a priori* the maximal speed of the processor to $S_{\max} := B$, so that no cost reduction would be possible by using a more powerful processor. If these parameters and the power dissipation w were related to units of work instead of units of time, B would also be independent of μ and therefore of all parameters.

Underlying the proof of our main result, there are some technical challenges that we now discuss. The proposed MDP satisfies the regularity assumptions (stability, unichain) needed to establish an optimality equation as described in [3]. However, this is not enough to show structural properties of the optimal policy. In fact, the classical approach to do this is to uniformize the MDP and to investigate the properties of the corresponding discrete time value iteration operator. Unfortunately, this is not possible in our case because the transition rates are unbounded. To uniformize the MDP, a typical approach consists of truncating the state space. However, a naive truncation will not help here because the truncation barrier has a strong impact on the structure of the optimal policy in the sense that it would not preserve any monotonicity property that it may have without truncation. Instead, we use the technique proposed by Blok and Spieksma in [1] on discounted costs, which smoothly scales down the upward rates of the truncated system as a function of the size of its state space. Here, we use the same truncation technique but we apply it to the average cost. In our specific case, the convergence to the infinite system will be guaranteed by the monotone convergence theorem.

References

[1] H. Blok and F. M. Spieksma. Countable state Markov decision processes with unbounded jump rates and discounted cost: optimality equation and approximations. *Advances in Applied Probability*, 47(4):1088 – 1107, 2015.

[2] B. Gaujal, A. Girault, and S. Plassart. Dynamic Speed Scaling Minimizing Expected Energy Consumption for Real-Time Tasks. *Journal of Scheduling*, pages 1–25, July 2020.

[3] X. Guo and O. Hernandez-Lerma. *Continuous-time Markov decision processes. Theory and applications*, volume 62. 01 2009.