A Fluid model based Heuristic for Optimal Speed-scaling in Bandwidth-sharing Networks

Olivier Brun, Henda Ben Cheikh, Balakrishna Prabhu

▶ **To cite this version:**

Olivier Brun, Henda Ben Cheikh, Balakrishna Prabhu. A Fluid model based Heuristic for Optimal Speed-scaling in Bandwidth-sharing Networks. IFIP WG 7.3 Performance 2015 - The 33rd International Symposium on Computer Performance, Modeling, Measurements and Evaluation 2015, IFIP, Oct 2015, Sydney, Australia. hal-01094506v3

HAL Id: hal-01094506
https://hal.science/hal-01094506v3

Submitted on 13 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Fluid model based Heuristic for Optimal Speed-scaling of Multi-class Single Server Queues

O. Brun[†,⋆], H. Ben Cheikh[†,⋆], B.J. Prabhu[†,⋆]

[†] CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France.
[⋆] Université de Toulouse, LAAS, F-31400 Toulouse, France.

Abstract

We investigate the energy-delay tradeoff in multi-class queues in which the server can regulate its speed according to the load of the system. Assuming that the queue is initially congested, we investigate the rate allocation to the classes that drains out the queue with minimum total energy and delay cost. We propose to solve this stochastic problem using a deterministic fluid approximation. We show that the optimal-fluid solution follows the well-known $c\mu$ rule and obtain an explicit expression for the optimal speed. Numerical results show the utility and the applicability of the fluid-optimal policy.

1 Introduction

The main question that we investigate is the following: assuming that a single server queue is initially in a congested state (e.g., due to flash-crowd), how to share the available bandwidth between multiple traffic classes in order to drain out the congestion with minimum total cost which comprises of the mean response times experienced by traffic classes as well as the energy consumption. Although the optimal speed-scaling policy taking into account the energy-delay trade-off is known for a single server and single class of traffic (see, e.g., [2]), there is no corresponding result for the multi-class problem studied in this paper.

2 Stochastic and fluid control problems

2.1 Stochastic control problem

We consider a single server queue shared by a set $\mathcal{F}$ of N classes of jobs. Class-i jobs arrive according to a Poisson process at rate λ_i and have exponentially distributed sizes of mean $1/\mu_i$, $i \in \mathcal{F}$. In the following, we define $\boldsymbol{X}(t) = (X_1(t), X_2(t), \ldots, X_N(t))$ as the state of the system, where $X_i(t)$ is the number of class-i jobs in the system at time t.

For simplicity, we assume that the server can be operated at any speed in the interval $[0, \infty)$. We let $\boldsymbol{u}_i(\boldsymbol{x})$ be the capacity allocated to class i when the system is in state $\boldsymbol{x}$, and denote by $\boldsymbol{u}(\boldsymbol{x})$ the vector $(\boldsymbol{u}_1(\boldsymbol{x}), \boldsymbol{u}_2(\boldsymbol{x}), \ldots, \boldsymbol{u}_N(\boldsymbol{x}))$. We further assume that all class-i jobs share the capacity allocated to the class according to the PS discipline. The power required to operate the server at rate $\sum_i u_i(\boldsymbol{x})$ is assumed to be proportional to $(\sum_i u_i(\boldsymbol{x}))^\gamma$ where $\gamma > 1$ [2].

We note that for any given stationary policy $\boldsymbol{u}$, the queue state $\boldsymbol{X}(t)$ is a multi-dimensional birth-and-death process. We shall assume that at time 0 the queue finds itself in a congestion state $\boldsymbol{x}(0) \gg \mathbf{0}$. The goal is to find the capacity allocation policy that will bring the queue to the state where all classes have zero jobs while minimizing the total cost incurred. Formally, we aim to find $\boldsymbol{u}^* : \mathbb{N}^N \to \mathbb{R}_+^N$ solving the following problem:

$$\text{Minimize } \mathbb{E}_{\boldsymbol{x}(0)} \left\{ \int_0^T f\left(\boldsymbol{X}(t), \boldsymbol{u}(\boldsymbol{X}(t))\right) dt \right\}, \quad (1)$$

where T is the first time the queue is empty, i.e., $\boldsymbol{X}(T) = \mathbf{0}$, and $f(\boldsymbol{x}, \boldsymbol{u})$ represents the cost rate in state $\boldsymbol{x}$ when a control $\boldsymbol{u}$ is applied, that is,

$$f(\boldsymbol{x}, \boldsymbol{u}) = \sum_{i \in \mathcal{F}} c_i x_i + \kappa \left(\sum_{i \in \mathcal{F}} u_i \right)^\gamma. \quad (2)$$

In (2), κ is a parameter controlling the relative weights of energy consumption and delay, whereas $\mathbf{c}$ is a vector giving the relative weights of the delays of the classes. Note that the cost comprises of two conflicting components: one for the holding cost of the jobs and the other for the energy consumption of the server. Intuitively, to lower the holding cost of jobs, one has to increase the speed of the server which then increases the energy consumption, and vice versa.

2.2 Fluid control problem

Problem (1) can be cast as a Markov decision process which proves to be both analytically and computationally challenging. Our approach is to analyze an associated fluid model which can be interpreted as a deterministic approximation of the stochastic problem. Let $x_i(t)$ represent the quantity of fluid associated to class i at time t and $u_i(t)$ be the rate allocated to this class at that time. The fluid control problem is then to find the rate allocation that drains out the queue with minimum total cost. Formally, the problem can be stated as

$$\text{minimize } J(\boldsymbol{u};\boldsymbol{x}_0) = \int_0^T f(\boldsymbol{x}(t),\boldsymbol{u}(t))dt \qquad \text{(OPT)}$$

subject to

$$\dot{x}_i(t) = \lambda_i - \mu_i u_i(t),\ i \in \mathcal{F}, \tag{3}$$

$$-\boldsymbol{x}(t) \leq \mathbf{0},, \tag{4}$$

$$-\boldsymbol{u}(t) \leq \mathbf{0}, \tag{5}$$

$$\boldsymbol{x}(0) = \boldsymbol{x}_0,\ \boldsymbol{x}(T) = \mathbf{0}, \tag{6}$$

where T denote the first time when the total amount of fluid in the network reaches 0. Note that the horizon T is not fixed, and it is also a part of the solution.

Using Pontryagin's Maximum Principle, we can show the following result.

Theorem 2.1. *At any time $t > 0$, $u_i(t) > 0$ if and only if either $x_i(t) = 0$ or $i = \arg\max\{c_j\mu_j : j \text{such that } x_j(t) > 0\}$. If $x_i(t) = 0$, then $u_i(t) = \rho_i$.*

The result says that amongst the classes with non-zero fluid, it is optimal to only serve the class with the largest value of $c_j\mu_j$. This is the same as the $c\mu$ rule when energy costs are not taken into account [4]. A direct consequence is that there exist $\tau_1 = 0 < \tau_2 < \ldots < \tau_{N+1} = T$ such that, in the optimal policy, class k receives a non-zero service rate in the interval $[\tau_k, \tau_{k+1})$ and a service rate of ρ_k in the interval $[\tau_{k+1}, T)$. The Pontryagin's Maximum Principle can be used to solve for T and to establish the speed at which class k is served in the interval $[\tau_k, \tau_{k+1})$. Due to the lack of space, we are not able to describe the details of the solution. We however mention that in the optimal policy the total server speed is decreasing as a function of time and scales as $(T-t)^{\frac{1}{\gamma-1}}$.

3 Numerical results

In order to illustrate the utility and the applicability of the fluid-optimal policy, we compare below the optimal stochastic policy obtained by solving a stochastic shortest path problem with the fluid policy obtained analytically, as well as with Bocop, an open source toolbox for optimal control problems [1]. In order to convert the fluid-optimal policy to the stochastic setting, we use a version of the Discrete-review method proposed in [3]. The idea is to periodically apply the fluid-optimal policy with the initial state as the state of the stochastic system sampled at review instants. In the scenario we consider, we assume that κ=1 and γ=3. The other parameters are as follows: ρ_0=0.2, ρ_1=0.1, ρ_2=0.3, $\mu_0 c_0$=5, $\mu_1 c_1$=6 and $\mu_2 c_2$=4. In Figure 1, we compare the optimal fluid rates allocated to each class with those that would be obtained under the optimal stochastic policy when the initial state is $\boldsymbol{x}(0) = (10, 11, 12)$. The fluid rates of each traffic class closely matches its stochastic rates and follows the $c\mu$ rule.

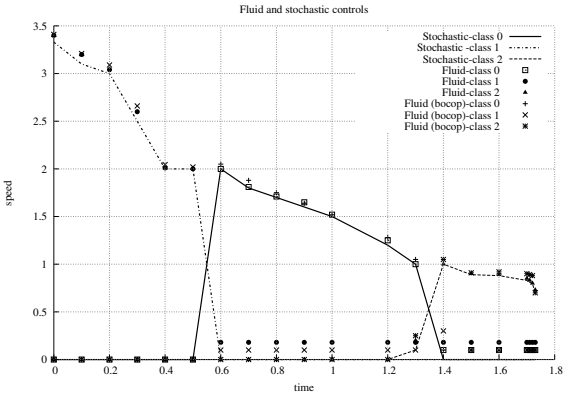

Figure 1: Optimal fluid speed vs. optimal stochastic speed for $x_0(0)$=10, $x_1(0)$=11 and $x_2(0)$=12.

References

[1] Bocop - the optimal control solver, http://bocop.saclay.inria.fr.

[2] L.L.H. Andrew, A. Wierman, and A. Tang. Optimal speed scaling under arbitrary power functions. *SIGMETRICS Perform. Eval. Rev.*, 37(2):39–41, October 2009.

[3] C. Maglaras. Discrete-review policies for scheduling stochastic networks: trajectory tracking and fluid-scale asymptotic optimality. *The Annals of Applied Probability*, 10(3):897–929, 2000.

[4] P. Nain and D. Towsley. Optimal scheduling in a machine with stochastic varying processing rate. *IEEE Transactions on Automatic Control*, 39:1853–1855, 1994.