The last, the least, and the urgent: fluid modeling and performance equivalence for scheduling policies in partial service queues with abandonment[★]

Andres Ferragut[a], Diego Goldsztajn[a], Fernando Paganini[a]

[a] *Universidad ORT Uruguay, Cuareim 1451, Montevideo, 11100, , Uruguay*

Abstract

In several queueing systems, arriving tasks or customers have both service and timing requirements, the latter expressed as a deadline for the task to be served. These systems with customer abandonment have a long and rich history in queueing theory, and have several applications in task scheduling in computer systems, operations research problems, etc. A common feature in all of these works is that they deal with customers reneging from the system only while in the queue, and not during service. However, in several applications, customers may also leave during service, and the partial work performed by the system during their stay is still useful.

In this paper we analyze these partial service queues with abandonment in a many-server setting, characterizing the equilibrium performance of several policies in terms of the amount of service attained by tasks. For this purpose, we develop fluid models with two-dimensional independent variables, corresponding to service and sojourn times, which take the form of partial differential equations expressed in weak form. These fluid models allow us to consider general and possibly correlated service and timing requirements, as well as a wide range of service disciplines. In particular, we focus on Earliest-Deadline-First, Least-Attained-Service and Last-Come-First-Served, and establish that all three policies have the same equilibrium performance, even though the latter two do not need any information about deadlines. This striking property means that designers may avoid the difficult job of estimating deadlines without incurring a performance penalty. The fluid model conclusions are validated by extensive numerical experiments.

Keywords: Deadline queues, Measure-valued processes, EDF, LCFS.

1. Introduction

In many queueing systems, it is necessary to consider not only the service time requested by the task or customer, but also the total amount of time that each given task is willing to stay in the system, i.e., its deadline. These systems with *customer reneging* or *abandonments* have a long and rich history in queueing theory, beginning with the seminal papers by Barrer [7], Stanford [29] and Baccelli et al. [6], and have several applications in task scheduling in computer systems, operations research problems (e.g., emergency assistance assignment), networking and processor scheduling environments, among many other examples.

One of the natural methods to incorporate deadlines is the *Earliest-Deadline-First (EDF)* scheduling policy, classical in operating systems task management, and thoroughly analyzed in [27] which first explored its optimality properties. This paper sparked continuously expanding research on this topic, leading to the development of stability results, fluid limits and diffusion approximations under heavy traffic conditions. For the single-server case, [8, 10, 11, 20, 21, 32] analyze this policy under increasingly broader assumptions on the service and deadline distributions, moving away from the basic exponential setting. A thorough review of this line of research can be found in [31]. See also [23] which uses a different approach, based on stochastic recurrences, still for single server systems.

The many-server case is more challenging; it has been analyzed in [17, 18], with focus on the case of orderly service, i.e., the First-Come-First-Served (FCFS) discipline. Nevertheless, the mathematical representation of the

[★] This work was partially supported by ANII Uruguay under grant PD_NAC_2024_1_182118 and AFOSR US under grant #FA9550-23-1-0350.

Email addresses: ferragut@ort.edu.uy (Andres Ferragut), goldsztajn@ort.edu.uy (Diego Goldsztajn), paganini@ort.edu.uy (Fernando Paganini)

system developed in these papers, as well as [19, 34] is a key stepping stone for the analysis of more complex policies; see in particular [1, 2] for recent results on the EDF policy.

The topic of reneging has also become important recently in the context of *matching queues*, frequent in operations research problems such as transportation and prompt order delivery. The main references in this regard are [3, 5, 28]. In these works the authors build upon the fluid models of [17, 19] to explore the tradeoffs between matching customers to less costly servers against the reneging costs, and develop optimal policies.

The mathematical tool that acts as the backbone of this line of research are *measure-valued processes*, which are needed to keep track of the system state, since for the problem to be interesting for applications, it is necessary to consider general service and deadline or sojourn time distributions (i.e., GI/GI/k-GI, in Barrer's [7] notation). In this regard, two approaches have evolved: the one in [17], that tracks *elapsed* times in the system, and the approach of [14, 22, 33], that tracks the *residual* times. Each approach has its advantages and both are required for the results of this paper. In both cases, some kind of *transport equation* is necessary to represent the dynamics.

While the aforementioned papers yield significant results for scheduling in reneging systems, almost all of them share a common assumption: *customers only abandon while waiting in the queue*, and if they reach service they stay until completion. We call this the *call-center scenario*, because it is more akin to how call center queues work. However, in many important applications tasks may abandon also *during service*, leaving the system with partially completed work. In some practical cases, this *partial service* received remains useful for the customer. We highlight some important examples: data transfers in networks, which may be canceled but resumed later, and therefore the transferred data is useful; here a processor-sharing discipline is a more adequate model, and is the focus of [9, 14]. A more closely related application to our analysis may be electrical vehicle charging [4], where users have energy requirements (service) and also patience time. Users may decide to leave with an incomplete charge, but the energy provided is still useful for them. Here, a many-server model is more adequate, due to power limitations, as explored in [33]. A final example is ML or AI inference in the cloud, where longer computation times (service) yield better results, but the system may decide to finish the job earlier in order to deliver a prompt response [30]. Moreover, generative AI chatbots have recently incorporated a functionality that allows users to interrupt the so-called thought process in order to obtain answers more quickly.

In this paper, we analyze these *partial service queues with abandonment* in the many-server setting. For this purpose, we develop fluid models based on partial differential equations (PDEs) in the two relevant coordinates of service and sojourn times. We admit general and possibly correlated service and patience requirements, which was identified in [24] as an important open problem. Using this framework, we analyze the equilibrium performance of different policies in terms of the service attained by tasks. Specifically, we find equilibrium measures for the PDEs and compute the rate at which tasks depart with a given amount of attained service, which characterizes the attained service distribution for a typical task in equilibrium.

We focus on the EDF policy and the deadline-oblivious policies *Least-Attained-Service* (LAS) and *Last-Come-First-Served* (LCFS). The former policy can be analyzed through a standard PDE that describes the distribution of the remaining service and time across the tasks in the system. For the other two policies, we consider the attained service and elapsed time of tasks, and we need PDEs written in weak form since the equilibrium measures are singular.

Our main result is that the deadline-oblivious policies LAS and LCFS match the equilibrium performance of EDF despite not using any information about deadlines. This striking property breaks away from the call-center scenario and has important practical implications, since estimating deadlines is considered a hard problem [15], and in some applications tasks are prone to under-reporting their deadlines to receive priority [13]. Our result implies that system designers may avoid the difficult job of estimating deadlines without incurring a performance penalty.

The rest of the paper is organized as follows. In Section 2 we define the load of the system and introduce two different descriptors for the state of the system, suitable for modeling different service policies. In Section 3 we motivate and define two fluid models for large-scale systems in equilibrium; each fluid model corresponds to a different state descriptor and consists of a PDE in weak form. In Section 4 we analyze the fluid models for systems in underload, and show that these systems behave as infinite-server queues. Section 5 provides a solution for the PDE associated with the EDF policy in overload, and use it to analyze the equilibrium performance in terms of the attained service. A similar analysis is performed in Section 6 for the LAS and LCFS policies. The relevant comparisons are made in Section 7, showing that the fluid models of all three policies have the same performance. Numerical experiments are discussed in Section 8; we consider a parametric setup where relevant quantities can be computed analytically, and a setup with correlated service and sojourn times. We conclude the paper in Section 9.

2. Basic notation, system load and state descriptors

In this section we introduce some basic notation, define the load of the system and provide two different state descriptors. Consider a service system where each arriving task i is associated with two random variables:

1. S_i is the amount of work required by task i, measured in units of time at a standardized service capacity,
2. T_i is the maximum amount of time that the task may stay in the system.

The random vectors (S_i, T_i) are independent and identically distributed with density $g(\sigma, \tau)$. In particular, this allows for possibly correlated service and patience requirements. In addition, we assume S_i and T_i have finite mean.

Suppose that tasks arrive as a Poisson process of intensity λ and the system has C servers, so that at most C tasks can be served simultaneously while the remaining tasks wait. There are two different ways in which task i may leave the system: if it has received S_i units of service time or if it has spent T_i units of time in the system; in the latter case, the task leaves the system even if it is currently in service. Unlike most of the literature on reneging queues, we consider that partial service is useful even if the task is not completed. The amount of service attained by task i by the time it leaves the system is denoted by $S_{a,i} \leqslant S_i$ and the amount of unfinished work is denoted by $S_{r,i} := S_i - S_{a,i}$. As an example, consider electrical vehicle charging: even if the user leaves while the battery is still not fully charged, the amount of energy received is valuable. The distribution of the attained service among the tasks and the total amount of unfinished work across the tasks characterize the performance of the system.

2.1. System load

Consider an auxiliary system with the same arriving tasks but infinitely many servers. In this situation, every task may be served at unit rate throughout its stay in the system, and thus it stays an amount of time equal to $\min\{S_i, T_i\}$. Let (S, T) be a random vector with density $g(\sigma, \tau)$. Then we have a classical $M/G/\infty$ queue where the service time distribution is $\min\{S, T\}$. In this case the average number of tasks in steady-state is the *offered load*:

$$\rho := \lambda E\left[\min\{S, T\}\right]. \tag{1}$$

In the original system, if the number of servers satisfies $C > \rho$, then we say that the system is in underload since the number of servers exceeds the average number of customers in the infinite-server system. On the other hand, we say that the system is overloaded if $C < \rho$ because in this case the server restriction will be binding, and users will be *curtailed* in their service, in the sense that they will not be able to achieve their full potential service $\min\{S, T\}$. How this curtailing is distributed across users will depend on the *service policy* and is one of the key points of this paper.

2.2. State descriptor based on residual times

Depending on which service policy is used to select the tasks being served, it may be convenient to describe the state of the system in different ways. For some service policies it is advantageous to describe the state in terms of the residual service times and residual patience of tasks. For this purpose, define:

1. $\sigma_i(t)$ as the remaining amount of service time at time t needed to complete task i,
2. $\tau_i(t)$ as the remaining patience of task i at time t.

The above information, for all the tasks currently in the system, can be encoded in a counting measure defined on the positive open orthant $\mathbb{R}^2_{++} = (0, \infty)^2$ as in [14]. Specifically, the state of the system is given by

$$\Phi_t = \sum_i \delta_{(\sigma_i(t), \tau_i(t))}, \tag{2}$$

where $\delta_{(\sigma,\tau)}$ is the Dirac measure at (σ, τ) and the sum is over all tasks present at time t.

Denote the set of finite point measures by $\mathcal{M}_P$. A *service policy* can be thought of as a mapping that takes the current state of the system $\Phi \in \mathcal{M}_P$ and returns a scalar field $r_\Phi(\sigma, \tau)$ specifying the service rate for tasks with remaining work σ and remaining patience τ. Naturally, we require that the following properties are satisfied by r_Φ:

1. For each task, the maximum (standardized) service rate is one, so $0 \leqslant r_\Phi \leqslant 1$ for all $\Phi \in \mathcal{M}_P$.

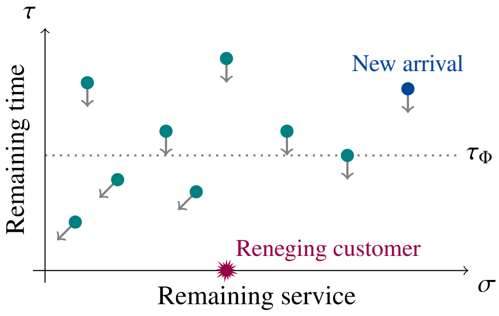

Figure 1: Illustration of the EDF policy for $C = 3$.

2. The total service rate must satisfy

$$\iint r_\Phi(\sigma,\tau)\Phi(d\sigma,d\tau) = \sum_i r_\Phi(\sigma_i,\tau_i) \leqslant \min\{\Phi(\mathbb{R}^2_{++}), C\} \quad \text{for all} \quad \Phi \in \mathcal{M}_P. \tag{3}$$

   We call a service policy r_Φ *efficient* if equality is attained in (3) for all Φ. For such policies, servers may only be idle if all tasks present are receiving maximum rate and their total number does not exceed capacity.

This formulation is rather general and allows to model any policy that depends on the residual service time and patience time of the tasks in the system. For example, the processor-sharing policy in [14] can be recovered by taking $C = 1$ and $r_\Phi(\sigma,\tau) \equiv 1/\Phi(\mathbb{R}^2_{++})$, and several other policies, such as the exact-scheduling policy in [25, 26], or the EV charging policies described in [33], can also be described using a scalar field r_Φ. An important policy to be analyzed in Section 5 is the (preemptive) Earliest-Deadline-First policy depicted in Figure 1. In this case

$$r_\Phi(\sigma,\tau) = \mathbf{1}_{\{\tau<\tau_\Phi\}} \quad \text{with} \quad \tau_\Phi := \sup\{\tau \geqslant 0 : \Phi(\mathbb{R}_{++}\times(0,\tau]) \leqslant C\}; \tag{4}$$

we let $\tau_\Phi = \infty$ if $\Phi(\mathbb{R}^2_{++}) < C$, i.e., if there are less than C tasks in the system. This definition means that the C tasks with more urgent deadlines are served at full rate, i.e., those with patience τ_i closer to zero. If $\Phi(\mathbb{R}^2_{++}) > C$, then the interpretation of τ_Φ is the remaining time in the system of the most urgent customer not in service.

Given a service policy, the measure-valued process Φ_t evolves over time as follows. Arrivals occur as a Poisson process of rate λ, creating atoms at coordinates (σ,τ) that are sampled from the density $g(\sigma,\tau)$ independently of everything else. Each atom moves over the orthant following the vector field $(-r_\Phi(\sigma,\tau),-1)$ until it reaches the $\sigma = 0$ or $\tau = 0$ axes. These events correspond to a task being fully served or a task having reached its deadline, respectively, and are immediately followed by the task departure, i.e., the atom disappears; see Figure 2.

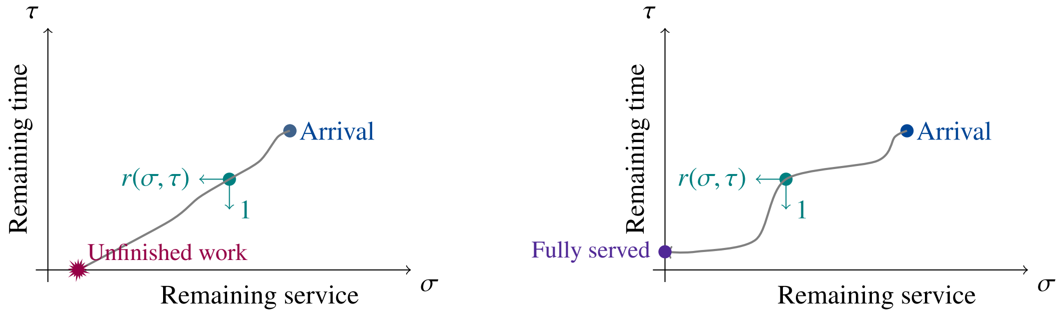

Figure 2: Dynamics in the remaining service and time descriptor. On the left a task leaves with partially finished work, on the right is fully served.

2.3. State descriptor based on elapsed times

The state descriptor defined above can be used to model several interesting policies, but it cannot be used to represent policies that rely on information about the arrival times of tasks, such as FCFS or Last-Come-First-Served (LCFS). It is also not possible to model service policies that use information about the amount of attained service time, such as Least-Attained-Service (LAS). Therefore, we now introduce a dual representation of the state of the system where we keep track of the times that tasks have spent in service and in the system. For this purpose, define:

1. $x_i(t)$ as the amount of service time that task i has attained by time t,
2. $y_i(t)$ as the amount of time that task i has spent in the system by time t.

For each task, the above information can be represented by a point in the closed[1] positive orthant $\mathbb{R}_+^2 = [0, \infty)^2$. Then the state of the system at time t is given by the following counting measure:

$$\tilde{\Phi}_t = \sum_i \delta_{(x_i(t), y_i(t))}, \tag{5}$$

where $\delta_{(x,y)}$ is the Dirac measure at the point (x, y) and the sum is over all tasks present at time t.

In this case a service policy is a scalar field $r_{\tilde{\Phi}}(x, y)$ that specifies at which rate tasks are served if they have already received x units of service time and have spent y units of time in the system; the constraints for choosing $r_{\tilde{\Phi}}$ are analogous to those stated in Section 2.2 for the residual times state descriptor.

As an example, consider the policies mentioned above. The scalar field $r_{\tilde{\Phi}}$ can be described in terms of a threshold that depends in feedback on the current state $\tilde{\Phi}$. In particular, for the FCFS policy, we have

$$r_{\tilde{\Phi}}(x, y) = \mathbf{1}_{\{y > u_{\tilde{\Phi}}\}} \quad \text{with} \quad u_{\tilde{\Phi}} := \inf\{y : \tilde{\Phi}(\mathbb{R}_+ \times [y, \infty)) \leqslant C\}, \tag{6}$$

where $u_{\tilde{\Phi}} = 0$ if $\tilde{\Phi}(\mathbb{R}_+^2) < C$; this means that the C tasks with the largest elapsed times in the system are served at maximum rate. The (preemptive) LCFS policy allocates full rate to the C jobs with least elapsed time in the system:

$$r_{\tilde{\Phi}}(x, y) = \mathbf{1}_{\{y < y_{\tilde{\Phi}}\}} \quad \text{with} \quad y_{\tilde{\Phi}} := \sup\{y : \tilde{\Phi}(\mathbb{R}_+ \times [0, y]) \leqslant C\}, \tag{7}$$

and $y_{\tilde{\Phi}} = \infty$ if $\tilde{\Phi}(\mathbb{R}_+^2) < C$. In addition, the (preemptive) LAS policy has

$$r_{\tilde{\Phi}}(x, y) = \mathbf{1}_{\{x < x_{\tilde{\Phi}}\}} \quad \text{with} \quad x_{\tilde{\Phi}} := \sup\{x : \tilde{\Phi}([0, x] \times \mathbb{R}_+) \leqslant C\}, \tag{8}$$

where $x_{\tilde{\Phi}} = \infty$ if $\tilde{\Phi}(\mathbb{R}_+^2) < C$; it serves the C tasks with the least attained service. A depiction of LAS and LCFS is given in Figure 3. If customers are considered in increasing order of their attained service, then $x_{\tilde{\Phi}}$ is the attained service of the first customer not in service under LAS, and if customers are considered in increasing order of the time spent in the system, then $y_{\tilde{\Phi}}$ is the time spent in the system by the first customer not in service under LCFS.

Given the service policy, the state of the system $\tilde{\Phi}_t$ evolves as follows. Tasks arrive as a Poisson process of rate λ and each task i generates an atom at the origin $(x_i, y_i) = (0, 0)$ of the orthant, which then moves following the vector field $(r_{\tilde{\Phi}}(x, y), 1)$ and departs when $x_i = S_i$ or $y_i = T_i$. In the former situation the task leaves because it has been completed, whereas in the latter situation it leaves because it has reached its maximum sojourn time; see Figure 4.

3. Fluid models for systems in equilibrium

In this section we introduce fluid models that describe large-scale systems for both state descriptors. For the residual time descriptor, a similar model has been proposed in [14] for a processor-sharing policy, and for the elapsed time descriptor, [17, 19] have derived fluid models for FCFS many-servers queues; without abandonment in [19] and with abandonment while not in service in [17]. In [10, 21] the authors analyzed EDF in a single-server scenario, and more recently, in [22] the authors analyzed a many-servers queue with random order of service and deadlines.

In all the above papers, the authors prove process-level convergence to the fluid model, performing a suitable scaling of the system parameters; in our setup, the scaling corresponds to setting $\lambda^{(n)} = n\lambda$, $C^{(n)} = nC$ and considering

[1] The reason why in this case we consider the closed orthant is that $(x_i, y_i) = (0, 0)$ if task i has just arrived.

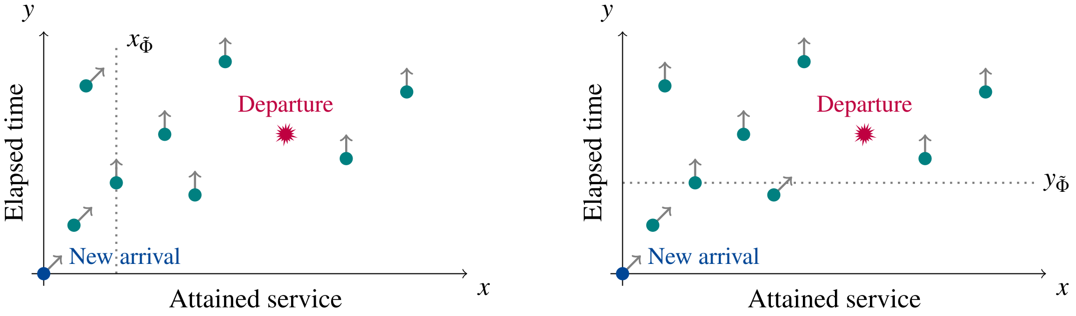

Figure 3: Illustration of the LAS and LCFS policies for $C = 3$.

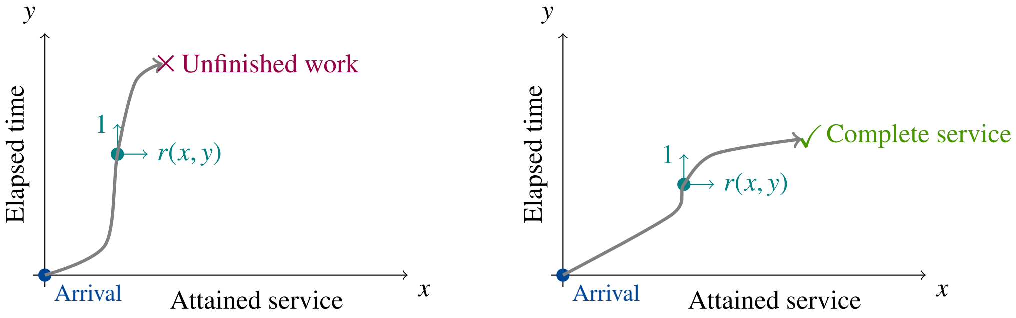

Figure 4: Schematic of the elapsed time dynamics. On the left, the task abandons with unfinished work, on the right, it is fully served.

the limit as $n \to \infty$ of the process $\Phi_t^{(n)}/n$ or $\tilde{\Phi}_t^{(n)}/n$. Here we will state the fluid model without proving such a limit; our main objective is to apply the fluid model to study the *performance* of large-scale systems in equilibrium. The analysis of process-level convergence is mathematically challenging and highly technical, we thus leave it for future research. Nevertheless, we validate the quality of the fluid approximation numerically in Section 8.

A key difference between our fluid models and those in previous work is that we need to keep track of both the residual/elapsed service time and patience of each task *simultaneously*, in a two-dimensional space. This is due to the fact that tasks may leave the system *at any time* either due to service completion or deadline expiration. This is different to the setting in [10, 17, 21, 22, 28], where tasks do not leave the system after service starts. In that case, the dynamics correspond to two systems in tandem: one representing the queue, where abandonment may occur, and the other representing service stations. Hence, the fluid dynamics can always be written in terms of a system of coupled PDEs, in each of which the "spatial" dimension is one dimensional. In contrast, we need to consider a PDE with "spatial" variables in the two-dimensional orthant.

3.1. Fluid model for the residual times state descriptor

Before stating the fluid model, let us provide an informal motivation for it. For this purpose, we replace the counting measure $\Phi_t \in \mathcal{M}_P(\mathbb{R}^2_{++})$ considered in Section 2.2 by a general measure $\mu_t \in \mathcal{M}(\mathbb{R}^2_{++})$, the space of finite nonnegative measures defined on the orthant $\mathbb{R}^2_{++}$. We further assume that the service policy is given by a scalar field $r_\mu(\sigma, \tau)$ that is well-defined as a function of measures $\mu \in \mathcal{M}(\mathbb{R}^2_{++})$. As before, we require that:

$$0 \leqslant r_\mu \leqslant 1, \tag{9a}$$

$$\iint r_\mu(\sigma, \tau)\mu(d\sigma, d\tau) \leqslant \min\left\{\mu(\mathbb{R}^2_{++}), C\right\}. \tag{9b}$$

In addition, we call the policy *efficient* if equality is attained in (9b).

The measure μ_t is fully determined by the projections

$$\langle \varphi, \mu_t \rangle := \iint \varphi(\sigma, \tau)\, \mu_t(d\sigma, d\tau) \tag{10}$$

with respect to test functions $\varphi : \mathbb{R}^2_{++} \to \mathbb{R}$ having continuous derivatives and compact support, i.e., $\varphi \in C^1_c(\mathbb{R}^2_{++})$. In our model, the derivative of the above expression with respect to t is given by:

$$\frac{d}{dt} \langle \varphi, \mu_t \rangle = \iint -\left[r_{\mu_t}(\sigma, \tau)\varphi_\sigma(\sigma, \tau) + \varphi_\tau(\sigma, \tau) \right] \mu_t(d\sigma, d\tau) + \lambda \iint \varphi(\sigma, \tau) g(\sigma, \tau)\, d\sigma d\tau, \tag{11}$$

where $\varphi_\sigma = \partial\varphi/\partial\sigma$ and $\varphi_\tau = \partial\varphi/\partial\tau$ denote the partial derivatives of φ. The first term on the right-hand side accounts for the movement of tasks over the orthant following the vector field $(-r_\mu(\sigma, \tau), -1)$. Indeed, for a task at (σ, τ), the instantaneous change in $\varphi(\sigma, \tau)$ when the task moves along the latter vector field is exactly the integrand in the first term of the right-hand side; the integral over μ_t gives the overall variation due to such transport motion. The second term accounts for the arrival of new tasks, which appear at rate λ and are placed at the point (σ, τ) with infinitesimal probability $g(\sigma, \tau) d\sigma d\tau$.

Equation (11) is a PDE in weak form, which is often called *transport equation* or *advection equation*. In fact, if μ_t has a density $f(\sigma, \tau; t)$ with respect to the Lebesgue measure, then after substitution and integration by parts we have:

$$\frac{\partial f}{\partial t} + \nabla \cdot \left[\mathbf{r}_{\mu_t} f \right] = \lambda g, \tag{12}$$

where $\nabla\cdot$ is the divergence operator and $\mathbf{r}_{\mu_t}$ is the vector field $\mathbf{r}_{\mu_t} := [-r_{\mu_t}(\sigma, \tau), -1]^\intercal$, which indicates the velocity at which mass is moving along the orthant. The reason for using a weak formulation is to allow the equilibrium to be more general (possibly without a density with respect to the Lebesgue measure).

We are interested in equilibrium solutions of (11), which must satisfy (11) when its left-hand side is zero. Formally, our fluid model for large-scale systems in equilibrium is given in the following definition.

Definition 1. We say that a finite measure $\mu \in \mathcal{M}(\mathbb{R}^2_{++})$ is a fluid equilibrium for the policy $r_\mu(\sigma, \tau)$ in the residual times state descriptor if and only if the following equation holds for all $\varphi \in C^1_c(\mathbb{R}^2_{++})$:

$$-\iint \left[r_\mu(\sigma, \tau)\varphi_\sigma(\sigma, \tau) + \varphi_\tau(\sigma, \tau) \right] \mu(d\sigma, d\tau) + \lambda \iint \varphi(\sigma, \tau) g(\sigma, \tau)\, d\sigma d\tau = 0. \tag{13}$$

In Section 5, we will consider (13) for the EDF policy, and we will find and analyze a fluid equilibrium.

3.2. *Fluid model for the elapsed times state descriptor*

As in the previous section, we will provide an informal motivation before stating the fluid model. For this purpose, replace the counting measure $\tilde{\Phi}_t \in \mathcal{M}_P(\mathbb{R}^2_+)$ considered in Section 2.3 by a general finite and nonnegative measure $\nu_t \in \mathcal{M}(\mathbb{R}^2_+)$. In addition, assume that the service policy is given by a scalar field $r_\nu(x, y)$ that is well-defined as a function of measures $\nu \in \mathcal{M}(\mathbb{R}^2_+)$ and satisfies the natural constraints, as in (9).

The measure ν_t is determined by the projections (10) against test functions in $C^1_c(\mathbb{R}^2_+)$. The dynamic model is:

$$\frac{d}{dt} \langle \varphi, \nu_t \rangle = \lambda\varphi(0, 0) + \iint \left[r_{\nu_t}(x, y)\varphi_x(x, y) + \varphi_y(x, y) \right] \nu_t(dx, dy) - \iint \varphi(x, y)\eta_{\nu_t}(x, y)\, \nu_t(dx, dy). \tag{14}$$

In this equation, the first term on the right-hand side corresponds to tasks appearing in the origin at rate λ, and the second term accounts for tasks moving over the orthant along the vector field $(r_\nu(x, y), 1)$, analogously to (11). The third term corresponds to tasks leaving the system; here $\eta_{\nu_t}(x, y)$ represents the departure rate of tasks that have received x units of service time and have spent y units of time in the system. Next we give an expression for $\eta_{\nu_t}(x, y)$ incorporating concepts from survival analysis.

Recall that for a positive random variable X with density $f(x)$, the hazard rate function is defined as:

$$h(x) := \lim_{h \downarrow 0} \frac{1}{h} P\left(X \in (x, x + h) \mid X > x \right) = \frac{f(x)}{\bar{F}(x)} = -\frac{d}{dx} \log \bar{F}(x),$$

where the complementary CDF $\bar{F}(x) := P(X > x)$ is also called the survival function.

In [16], the authors extend this hazard rate concept to the multivariate case, which is appropriate for our two-dimensional setup in which there are two possibly correlated random variables, the service time S and the sojourn time T, determining the "survival" of tasks in the system.

Definition 2 (Hazard rate vector field, [16])**.** Given a pair of positive random variables (S, T) with joint density $g(x, y)$, consider their joint survival function $\bar{G}(x, y) = P(S > x, T > y)$ and define the *hazard rate vector field* as:

$$\mathbf{h}(x, y) = \begin{pmatrix} h^x(x, y) \\ h^y(x, y) \end{pmatrix} := -\nabla \log \bar{G}(x, y) = -\frac{1}{\bar{G}(x, y)} \nabla \bar{G}(x, y). \tag{15}$$

As an example, in the case of independent and exponentially distributed S and T, with rates α and β, the hazard rate vector field is $\mathbf{h}(x, y) \equiv (\alpha, \beta)^\intercal$, generalizing the notion of constant hazard rate of the one–dimensional case.

We apply this definition to evaluate the departure rate $\eta_{\nu_t}(x, y)$. Suppose that at time t a task has received x units of service time and has spent y units of time in the system. The following calculation assumes that the service rate $r = r_{\nu_t}(x, y)$ remains constant in $(t, t + h)$; the probability that the task leaves the system before time $t + h$ is then:

$$\begin{aligned} P\left([\{S \in (x, x + hr)\} \cup \{T \in (y, y + h)\}] \cap \{S > x, T > y\}\right) = &\int_x^{x+hr} \left[\int_y^\infty g(u, v)\, dv\right] du + \int_x^\infty \left[\int_y^{y+h} g(u, v)\, dv\right] du \\ &- \int_x^{x+hr} \left[\int_y^{y+h} g(u, v)\, dv\right] du. \end{aligned}$$

For the integral in the second line, the integration domain has Lebesgue measure rh^2. Thus, under mild assumptions on the density $g(u, v)$ (e.g., local boundedness), it will disappear when computing the departure rate, as follows:

$$\begin{aligned} &\lim_{h \downarrow 0} \frac{1}{h} P\left([\{S \in (x, x + hr)\} \cup \{T \in (y, y + h)\}] \cap \{S > x, T > y\}\right) \\ &\quad = \lim_{h \downarrow 0} \frac{1}{h} \int_x^{x+hr} \left[\int_y^\infty g(u, v)\, dv\right] du + \lim_{h \downarrow 0} \frac{1}{h} \int_y^{y+h} \left[\int_x^\infty g(u, v)\, du\right] dv \\ &\quad = r \int_y^\infty g(x, v)\, dv + \int_x^\infty g(u, y)\, du \\ &\quad = -r \frac{\partial \bar{G}}{\partial x}(x, y) - \frac{\partial \bar{G}}{\partial y}(x, y) \\ &\quad = \left[r h^x(x, y) + h^y(x, y)\right] \bar{G}(x, y), \end{aligned}$$

where the second equality holds outside a set of Lebesgue measure zero, by Lebesgue's differentiation theorem. Then

$$\lim_{h \downarrow 0} \frac{1}{h} P(\{S \in (x, x + hr)\} \cup \{T \in (y, y + h)\} \mid S > x, T > y) = r h^x(x, y) + h^y(x, y).$$

Using the above expression, we set $\eta_{\nu_t}(x, y) = r_{\nu_t}(x, y) h^x(x, y) + h^y(x, y)$ and rewrite (14) as follows:

$$\begin{aligned} \frac{d}{dt} \langle \varphi, \nu_t \rangle = \lambda \varphi(0, 0) &+ \iint \left[r_{\nu_t}(x, y) \varphi_x(x, y) + \varphi_y(x, y)\right] \nu_t(dx, dy) \\ &- \iint \varphi(x, y) \left[r_{\nu_t}(x, y) h^x(x, y) + h^y(x, y)\right] \nu_t(dx, dy). \end{aligned} \tag{16}$$

Equation (16) is again a transport or advection equation written in weak form. If ν_t admits a density $f(x, y; t)$ with respect to the Lebesgue measure on $\mathbb{R}_+^2$, then (16) can be informally stated as:

$$\frac{\partial f}{\partial t} + \nabla \cdot [\mathbf{r}_{\nu_t} f] + [\mathbf{r}_{\nu_t} \cdot \mathbf{h}] f = \lambda \delta_{(0,0)}.$$

However, since mass arrives only at $(0, 0)$, we have an impulse function at the origin as the forcing field, and thus it is better to work with (16). Also, we will see in Section 6 that solutions without a density are relevant.

Since we are interested in equilibrium solutions, we replace the left-hand side of (16) by zero. Formally, our fluid model for equilibrium in the elapsed time state descriptor is given in the following definition.

Definition 3. We say that a finite measure $\nu \in \mathcal{M}(\mathbb{R}_+^2)$ is a fluid equilibrium for the policy $r_\nu(x,y)$ in the elapsed times state descriptor if and only if the following equation holds for all $\varphi \in C_c^1(\mathbb{R}_+^2)$:

$$\lambda\varphi(0,0) + \iint \left[r_\nu(x,y)\varphi_x(x,y) + \varphi_y(x,y)\right] \nu(dx,dy) - \iint \varphi(x,y)\left[r_\nu(x,y)h^x(x,y) + h^y(x,y)\right] \nu(dx,dy) = 0. \quad (17)$$

In Section 6 we will use (17) to find and analyze fluid equilibria for LCFS and LAS.

4. Equilibrium behavior in underload

First, we analyze fluid equilibria for systems in underload, i.e., with $\rho < C$. The following proposition establishes that the fluid equilibrium behaves as in the infinite server case, with all tasks served at full rate.

Proposition 1. *Assume that $\rho < C$. Then $r_\mu \equiv 1$ is an efficient policy that satisfies the conditions in* (9) *and admits the following fluid equilibrium $\mu^* \in \mathcal{M}(\mathbb{R}_{++}^2)$ in the sense of Definition 1. The measure μ^* has density*

$$f(\sigma,\tau) = \lambda \int_0^\infty g(\sigma+u,\tau+u)du \quad \textit{for all} \quad \sigma,\tau > 0. \quad (18)$$

In addition, $r_\nu \equiv 1$ is an efficient policy and admits a fluid equilibrium $\nu^* \in \mathcal{M}(\mathbb{R}_+^2)$ in the sense of Definition 3, where

$$\int_{\mathbb{R}_+^2} \varphi(x,y)\nu^*(dx,dy) = \lambda \int_0^\infty \varphi(u,u)\bar{G}(u,u)du \quad \textit{for all} \quad \varphi \in C_c(\mathbb{R}_+^2). \quad (19)$$

Proof. First we show that μ^* satisfies Definition 1 for $r_\mu \equiv 1$. Note that in this case, imposing equilibrium in (12), we end up with a linear PDE than can be solved by the method of characteristic curves [12]. In this case, since the driving field is $\mathbf{r} = [-1,-1]^\intercal$, these curves are simply $\{(\sigma+u,\tau+u) : u > 0\}$. Integrating along these curves yields the expression (18). We verify that this is a solution:

$$\frac{\partial f}{\partial \sigma}(\sigma,\tau) + \frac{\partial f}{\partial \tau}(\sigma,\tau) = \lambda \int_0^\infty \left[g_\sigma(\sigma+u,\tau+u) + g_\tau(\sigma+u,\tau+u)\right] du = \lambda g(\sigma+u,\tau+u)\Big|_0^\infty = -\lambda g(\sigma,\tau).$$

Therefore, f satisfies the stationary version of (12), and thus μ^* solves (13) as required. Furthermore,

$$\begin{aligned}\mu^*\left(\mathbb{R}_{++}^2\right) = \int_0^\infty \int_0^\infty f(\sigma,\tau)d\sigma d\tau &= \lambda \int_0^\infty \int_0^\infty \int_0^\infty g(\sigma+u,\tau+u)dud\sigma d\tau \\ &= \lambda \int_0^\infty P\left(\min\{S,T\} > u\right) du = \lambda E\left[\min\{S,T\}\right] = \rho < C,\end{aligned}$$

which shows that μ^* is finite and (9b) holds with equality. In particular, r_μ is an efficient policy.

The elapsed time case is slightly more involved: since tasks receive service immediately upon arrival, mass concentrates on the diagonal $\{(u,u) : u \geqslant 0\}$; this is captured by (19), which is written (inevitably) in weak form.

We verify that this is a solution. Consider $r_\nu \equiv 1$ in (17), fix any $\varphi \in C_c^1(\mathbb{R}_+^2)$ and observe that:

$$\begin{aligned}\int_{\mathbb{R}_+^2} \left[\varphi_x(x,y) + \varphi_y(x,y)\right] \nu^*(dx,dy) &= \lambda \int_0^\infty \left[\varphi_x(u,u) + \varphi_y(u,u)\right] \bar{G}(u,u)du \\ &= -\lambda\varphi(0,0) - \lambda \int_0^\infty \varphi(u,u)\left[\bar{G}_x(u,u) + \bar{G}_y(u,u)\right] du \\ &= -\lambda\varphi(0,0) + \lambda \int_0^\infty \varphi(u,u)h^x(u,u)\bar{G}(u,u)du + \lambda \int_0^\infty \varphi(u,u)h^y(u,u)\bar{G}(u,u)du \\ &= -\lambda\varphi(0,0) + \int_{\mathbb{R}_+^2} \varphi(x,y)\left[h^x(x,y) + h^y(x,y)\right] \nu^*(dx,dy),\end{aligned}$$

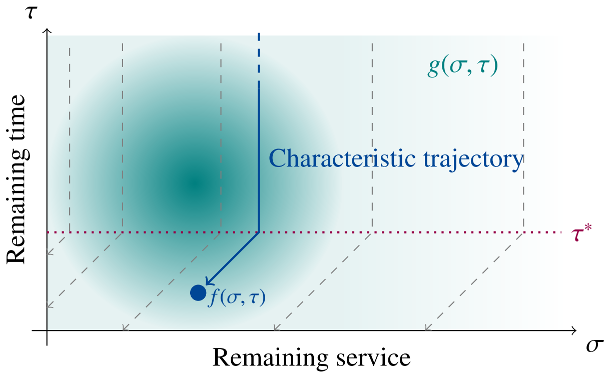

Figure 5: Characteristic curves of the EDF partial differential equation (21).

where the second line follows from integration by parts and the fact that φ has compact support. This shows that ν^* satisfies (17). In addition, we have

$$\nu^*\left(\mathbb{R}^2_+\right) = \lambda \int_0^\infty \bar{G}(u,u)du = \lambda \int_0^\infty P\left(\min\left\{S,T\right\} > u\right) du = \lambda E\left[\min\left\{S,T\right\}\right] = \rho < C.$$

This proves that ν^* is finite and efficient. □

The proposition shows that the policy $r \equiv 1$, which serves each task at full rate, admits an equilibrium measure as in Definitions 1 and 3 under the underload condition $\rho < C$. In particular, the equilibrium measure has total mass smaller than the number of servers C. Moreover, the threshold policies mentioned in Sections 2.2 and 2.3 admit the same equilibrium measure since the total mass is less than C, and thus the thresholds are trivial, i.e., equal to infinity or zero. Note that in the elapsed times state descriptor all mass concentrates on the diagonal, and the density along the diagonal is the well-known residual lifetime distribution of $\min\{S,T\}$, as in the fluid limit of the $M/G/\infty$ queue.

We will now turn to the overload case $\rho > C$ where curtailing of service will occur, thresholds will be non-trivial and differences may appear between the policies. We begin by analyzing the EDF policy.

5. Earliest-Deadline-First in overload

We begin by defining the fluid counterpart of the EDF policy defined in (4). Let μ be a finite measure in $\mathbb{R}^2_{++}$, then the EDF policy is the following threshold policy:

$$r_\mu(\sigma,\tau) = \mathbf{1}_{\{\tau<\tau_\mu\}} \quad \text{with} \quad \tau_\mu := \sup\{\tau \geqslant 0 : \mu(\mathbb{R}_{++} \times (0,\tau]) \leqslant C\}. \tag{20}$$

A fluid equilibrium μ^* must satisfy (13) for any $\varphi \in C_c^1(\mathbb{R}^2_{++})$ with $\tau^* := \tau_{\mu^*}$ a fixed value. Assuming that μ^* has a density f and imposing the equilibrium condition in (12), we obtain:

$$\frac{\partial f}{\partial \sigma}\mathbf{1}_{\{\tau<\tau^*\}} + \frac{\partial f}{\partial \tau} + \lambda g = 0. \tag{21}$$

Again, this is a linear PDE that can be solved by following the characteristic trajectories, which are represented in the diagram of Figure 5. We are now ready to state the main result of this section.

Theorem 1. *Assume that $\rho > C$. Then there exists a unique $\tau^* \geqslant 0$ such that*

$$\lambda E[\min\{S,T,\tau^*\}] = \lambda \int_0^{\tau^*} P\left(\min\{S,T\} > u\right) du = C. \tag{22}$$

Consider the measure μ^* given by the following density:

$$f(\sigma,\tau) = \lambda\left[\int_0^{(\tau^*-\tau)^+} g(\sigma+u,\tau+u)du + \int_{(\tau^*-\tau)^+}^{\infty} g\left(\sigma+(\tau^*-\tau)^+,\tau+u\right)du\right] \quad \textit{for all} \quad \sigma,\tau>0. \tag{23}$$

This measure is a fluid equilibrium per Definition 1 for the policy (20)*, and* $\tau^* = \sup\{\tau \geqslant 0 : \mu^*(\mathbb{R}_{++}\times[0,\tau]) \leqslant C\}$.

Proof. We prove that μ^* satisfies (13) by checking that the density f satisfies (21). For this purpose, note that

$$\frac{\partial f}{\partial\sigma}(\sigma,\tau)\mathbf{1}_{\{\tau<\tau^*\}} = \lambda\left[\int_0^{(\tau^*-\tau)^+} g_\sigma(\sigma+u,\tau+u)du + \int_{\tau^*-\tau}^{\infty} g_\sigma(\sigma+\tau^*-\tau,\tau+u)du\right]\mathbf{1}_{\{\tau<\tau^*\}}.$$

Computing the partial derivative with respect to τ requires more care since the integration limits in (23) depend on τ when $\tau<\tau^*$. However, using Leibniz's integral rule, we get:

$$\begin{aligned}
\frac{\partial f}{\partial\tau}(\sigma,\tau) &= -\lambda g(\sigma+\tau^*-\tau,\tau^*)\mathbf{1}_{\{\tau<\tau^*\}} + \lambda\int_0^{(\tau^*-\tau)^+} g_\tau(\sigma+u,\tau+u)du \\
&\quad + \lambda g(\sigma+\tau^*-\tau,\tau^*)\mathbf{1}_{\{\tau<\tau^*\}} + \lambda\int_{(\tau^*-\tau)^+}^{\infty}\left[-g_\sigma(\sigma+\tau^*-\tau,\tau+u)\mathbf{1}_{\{\tau<\tau^*\}} + g_\tau(\sigma+(\tau^*-\tau)^+,\tau+u)\right]du \\
&= \lambda\left[\int_0^{(\tau^*-\tau)^+} g_\tau(\sigma+u,\tau+u)du + \int_{(\tau^*-\tau)^+}^{\infty} g_\tau(\sigma+(\tau^*-\tau)^+,\tau+u) - \int_{\tau^*-\tau}^{\infty} g_\sigma(\sigma+\tau^*-\tau,\tau+u)du\mathbf{1}_{\{\tau<\tau^*\}}\right].
\end{aligned}$$

For the second integral above, note that the first argument of the integrand is independent of τ if $\tau \geqslant \tau^*$. Combining the above derivations, we conclude that f indeed satisfies (21), because:

$$\begin{aligned}
\frac{\partial f}{\partial\sigma}(\sigma,\tau)\mathbf{1}_{\{\tau<\tau^*\}} + \frac{\partial f}{\partial\tau}(\sigma,\tau) &= \lambda\left[\int_0^{(\tau^*-\tau)^+}\left[g_\sigma(\sigma+u,\tau+u)+g_\tau(\sigma+u,\tau+u)\right]du + \int_{(\tau^*-\tau)^+}^{\infty} g_\tau(\sigma+(\tau^*-\tau)^+,\tau+u)du\right] \\
&= \lambda\left[g(\sigma+u,\tau+u)\Big|_0^{(\tau^*-\tau)^+} - g(\sigma+(\tau^*-\tau)^+,\tau+(\tau^*-\tau)^+)\right] = -\lambda g(\sigma,\tau).
\end{aligned}$$

It remains to check that μ^* is finite, (9b) holds and $\tau^* = \sup\{\tau\geqslant 0 : \mu^*(\mathbb{R}_{++}\times[0,\tau])\leqslant C\}$. Note that:

$$\begin{aligned}
\int_0^{\tau^*}\left[\int_0^{\infty} f(\sigma,\tau)d\sigma\right]d\tau =&\lambda\int_0^{\tau^*}\left[\int_0^{\infty}\left[\int_0^{\tau^*-\tau} g(\sigma+u,\tau+u)du\right]d\sigma\right]d\tau \\
&+\lambda\int_0^{\tau^*}\left[\int_0^{\infty}\left[\int_{\tau^*-\tau}^{\infty} g(\sigma+\tau^*-\tau,\tau+u)du\right]d\sigma\right]d\tau \\
=&\lambda\int_0^{\tau^*}\left[\int_0^{\tau^*-u}\left[\int_0^{\infty} g(\sigma+u,\tau+u)d\sigma\right]d\tau\right]du \\
&+\lambda\int_0^{\tau^*}\left[\int_{\tau^*-\tau}^{\infty}\left[\int_0^{\infty} g(\sigma+\tau^*-\tau,\tau+u)d\sigma\right]du\right]d\tau \\
=&\lambda\int_0^{\tau^*}\left[\int_z^{\tau^*}\left[\int_z^{\infty} g(x,y)dx\right]dy\right]dz + \lambda\int_0^{\tau^*}\left[\int_{\tau^*}^{\infty}\left[\int_{\tau^*-z}^{\infty} g(x,y)dx\right]dy\right]dz \\
=&\lambda\int_0^{\tau^*}\left[\int_z^{\tau^*}\left[\int_z^{\infty} g(x,y)dx\right]dy\right]dz + \lambda\int_0^{\tau^*}\left[\int_{\tau^*}^{\infty}\left[\int_z^{\infty} g(x,y)dx\right]dy\right]dz \\
=&\lambda\int_0^{\tau^*} P(S>z,T>z)\,dz = \lambda E\left[\min\{S,T,\tau^*\}\right] = C,
\end{aligned}$$

which implies that (9b) holds and $\tau^* = \sup\{\tau\geqslant 0 : \mu^*(\mathbb{R}_{++}\times[0,\tau])\leqslant C\}$. Moreover, μ^* is a finite measure since

$$\begin{aligned}
\int_{\tau^*}^{\infty}\left[\int_0^{\infty} f(\sigma,\tau)d\sigma\right]d\tau &= \lambda\int_{\tau^*}^{\infty}\left[\int_0^{\infty}\left[\int_0^{\infty} g(\sigma,\tau+u)du\right]d\sigma\right]d\tau \\
&= \lambda\int_{\tau^*}^{\infty}\left[\int_z^{\infty}\left[\int_0^{\infty} g(x,y)dx\right]dy\right]dz = \lambda\int_{\tau^*}^{\infty} P(T>z)\,dz \leqslant \lambda E[T] < \infty.
\end{aligned}$$

This completes the proof. □

We now turn to analyzing the *performance* of the EDF policy, in regard to the service curtailing that it imposes in equilibrium. In particular, we wish to characterize the distribution of unfinished work across tasks. For this purpose, consider the one-dimensional integral

$$\int_0^{\tau^*} f(0,\tau)d\tau$$

of the population density along the boundary $\{0\} \times [0,\tau^*]$, where *completed* tasks leave the orthant. This integral has units of *rate*[2] and represents the speed at which tasks depart without any unfinished work. Similarly, the integral

$$\int_0^{\sigma_0} f(\sigma,0)d\sigma$$

represents the rate at which tasks depart with a nonzero amount of remaining work that is smaller than σ_0. Hence, the sum of these two integrals represents the total rate at which tasks leave the system with an amount of unfinished work smaller than σ_0. The following proposition computes this sum.

Proposition 2. *Assume that* $\rho > C$ *and* τ^* *is the unique solution of* (22). *Then*

$$\int_0^{\tau^*} f(0,\tau)d\tau + \int_0^{\sigma_0} f(\sigma,0)d\sigma = \lambda P(S - \min\{S,T,\tau^*\} < \sigma_0)\,. \tag{24}$$

Proof. First, observe that

$$\begin{aligned}\int_0^{\tau^*} f(0,\tau)d\tau &= \lambda\int_0^{\tau^*}\left[\int_0^{\tau^*-\tau} g(u,\tau+u)du\right]d\tau + \lambda\int_0^{\tau^*}\left[\int_{\tau^*-\tau}^{\infty} g(\tau^*-\tau,\tau+u)du\right]d\tau\\ &= \lambda\int_0^{\tau^*}\left[\int_0^{y} g(x,y)dx\right]dy + \lambda\int_{\tau^*}^{\infty}\left[\int_0^{\tau^*} g(x,y)dx\right]dy\\ &= \lambda P(S < T < \tau^*) + \lambda P(0 < S < \tau^*, T > \tau^*) = \lambda P(S < \tau^*, T > S)\,.\end{aligned}$$

For the second integral on the left-hand side of (24), we obtain:

$$\begin{aligned}\int_0^{\sigma_0} f(\sigma,0)d\sigma &= \lambda\int_0^{\sigma_0}\left[\int_0^{\tau^*} g(\sigma+u,u)du\right]d\sigma + \lambda\int_0^{\sigma_0}\left[\int_{\tau^*}^{\infty} g(\sigma+\tau^*,u)du\right]d\sigma\\ &= \lambda\int_0^{\tau^*}\left[\int_y^{y+\sigma_0} g(x,y)dx\right]dy + \lambda\int_{\tau^*}^{\infty}\left[\int_{\tau^*}^{\tau^*+\sigma_0} g(x,y)dx\right]dy\\ &= \lambda P(T < S < T+\sigma_0, T < \tau^*) + \lambda P(\tau^* < S < \tau^*+\sigma_0, T > \tau^*)\,.\end{aligned}$$

Putting everything together, we conclude that

$$\begin{aligned}\int_0^{\tau^*} f(0,\tau)d\tau + \int_0^{\sigma_0} f(\sigma,0)d\sigma &= \lambda P(S < \tau^*, T > S) + \lambda P(T < S < T+\sigma_0, T < \tau^*)\\ &\quad + \lambda P(\tau^* < S < \tau^*+\sigma_0, T > \tau^*)\\ &= \lambda P(S-\tau^* < 0, S-T < 0) + \lambda P(0 < S-T < \sigma_0, T < \tau^*)\\ &\quad + \lambda P(0 < S-\tau^* < \sigma_0, T > \tau^*)\,.\end{aligned}$$

Since (S,T) is absolutely continuous, the three probabilities in the last expression add up to the probability of

$$\begin{aligned}\{S-T < \sigma_0, S-\tau^* < \sigma_0\} &= \{S-T < \sigma_0, T < \tau^*, S-T > 0\} \cup \{T < \tau^*, S-T \leqslant 0\}\\ &\quad \cup \{S-\tau^* < \sigma_0, T \geqslant \tau^*, S-\tau^* > 0\} \cup \{T \geqslant \tau^*, S-\tau^* \leqslant 0\}\\ &= \{S-T \leqslant 0, S-\tau^* \leqslant 0\} \cup \{0 < S-T < \sigma_0, T < \tau^*\} \cup \{0 < S-\tau^* < \sigma_0, T \geqslant \tau^*\}.\end{aligned}$$

[2] The density $f(\sigma,\tau)$ has units of number of tasks divided units of time squared, so that its double integral over the residual service and patience times has units of number of tasks. Thus, simple integrals have units of the number of tasks per unit of time, i.e., units of rate.

Therefore, we conclude that

$$\int_0^{\tau^*} f(0,\tau)d\tau + \int_0^{\sigma_0} f(\sigma,0)d\sigma = \lambda P\left(\max\{0, S-T, S-\tau^*\} < \sigma_0\right) = \lambda P\left(S - \min\{S,T,\tau^*\} < \sigma_0\right),$$

where the last step follows from $-\min\{S,T,\tau^*\} = \max\{-S,-T,-\tau^*\}$. □

Since in equilibrium the overall departure rate must be λ, we have shown that the fraction of tasks which depart with unfinished work smaller than σ_0 (i.e., the cumulative distribution of the unfinished work S_r) coincides with the distribution of the random variable $S - \min\{S,T,\tau^*\}$. Since $S - S_r = S_a$, the attained work, this is consistent with an attained work distributed as $\min\{S,T,\tau^*\}$. A further discussion is deferred to Section 7.

6. The deadline-oblivious policies LAS and LCFS in overload

An important disadvantage of EDF, and any other deadline-aware policy, is that it requires knowing the deadlines of tasks beforehand. This is problematic when deadlines are uncertain, or in settings with strategic customers that may report lower values than their real deadlines, in order to obtain priority. In this section we derive equilibrium measures for two *deadline-oblivious* policies that will lead to the same overall performance.

Namely, we consider the LAS and LCFS policies defined in Section 2. Recall that the former serves the C users with the least attained service, whereas the latter serves the C users with the least time in the system. In particular, they do not depend on task deadlines or any other exogenous information.

Note that, according to their definition, these policies depend in feedback on a threshold value that determines either the attained service or the elapsed time of the $(C+1)$-th customer. In order to satisfy equilibrium conditions as in Definition 3, these thresholds must be fixed, and thus the service rates for LAS and LCFS become:

$$r_{LAS}(x,y) = \mathbf{1}_{\{x<x^*\}} \quad \text{and} \quad x^* = \sup\{x \geqslant 0 : \nu^*_{LAS}([0,x]\times\mathbb{R}_+) \leqslant C\}, \tag{25}$$

$$r_{LCFS}(x,y) = \mathbf{1}_{\{y<y^*\}} \quad \text{and} \quad y^* = \sup\{y \geqslant 0 : \nu^*_{LCFS}(\mathbb{R}_+\times[0,y]) \leqslant C\}, \tag{26}$$

where ν^*_{LAS} and ν^*_{LCFS} are equilibrium measures. The following theorem shows that both policies admit the same fluid equilibrium in overload. Unlike the fluid equilibrium of EDF, this equilibrium is not absolutely continuous since the measure is concentrated in a one-dimensional set of the form $\{(u,u) : 0 \leqslant u \leqslant z^*\} \cup \{(z^*,u) : u \geqslant z^*\}$.

Theorem 2. *Assume that $\rho > C$ and let $z^* \geqslant 0$ be the unique solution of*

$$\lambda E[\min\{S,T,z^*\}] = C; \tag{27}$$

this is the same equation as (22). *Consider the measure ν^* given by:*

$$\int_{\mathbb{R}^2_+} \varphi(x,y)\nu^*(dx,dy) = \lambda\left[\int_0^{z^*} \varphi(u,u)\bar{G}(u,u)du + \int_{z^*}^{\infty} \varphi(z^*,u)\,\bar{G}(z^*,u)du\right], \tag{28}$$

for all $\varphi \in C_c(\mathbb{R}^2_+)$. This measure is a fluid equilibrium, as in Definition 3, both for the LAS policy (25) *and the LCFS policy* (26)*. In the former case we have $x^* = z^*$ and in the latter case $y^* = z^*$.*

Proof. We only prove the above properties for LCFS; analogous arguments may be used for LAS. First, note that

$$\nu^*\left(\mathbb{R}_+\times[0,z^*]\right) = \lambda\int_0^{z^*} \bar{G}(u,u)du = \lambda\int_0^{\infty} P\left(\min\{S,T,z^*\} > u\right)du = \lambda E\left[\min\{S,T,z^*\}\right] = C,$$

which implies efficiency and that (26) holds. Moreover, ν^* is finite since

$$\nu^*\left(\mathbb{R}_+\times[z^*,\infty)\right) = \lambda\int_{z^*}^{\infty} \bar{G}(z^*,u)du \leqslant \lambda\int_{z^*}^{\infty} P\left(T > u\right)du \leqslant \lambda E[T] < \infty.$$

In order to check that ν^* satisfies the equilibrium condition (17), observe that:

$$\begin{aligned}\int_a^b \left[\varphi_x(u,u) + \varphi_y(u,u)\right] \bar{G}(u,u)du &= \varphi(u,u)\bar{G}(u,u)\Big|_a^b - \int_a^b \varphi(u,u)\left[\bar{G}_x(u,u) + \bar{G}_y(u,u)\right] du \\ &= \varphi(u,u)\bar{G}(u,u)\Big|_a^b + \int_a^b \varphi(u,u)\left[h^x(u,u)\bar{G}(u,u) + h^y(u,u)\bar{G}(u,u)\right] du\end{aligned}$$

for all $0 \leqslant a < b$ and $\varphi \in C_c^1(\mathbb{R}_+^2)$. We further have:

$$\int_a^b \varphi_y(z^*,u)\,\bar{G}(z^*,u)du = \varphi(z^*,u)\,\bar{G}(z^*,u)\Big|_a^b + \int_a^b \varphi(z^*,u)\,h^y(z^*,u)\bar{G}(z^*,u)du.$$

If $\varphi \in C_c^1(\mathbb{R}_+^2)$, then the above properties imply that:

$$\begin{aligned}\int_{\mathbb{R}_+^2} \left[\varphi_x(x,y)\mathbf{1}_{\{y<y^*\}} + \varphi_y(x,y)\right] \nu^*(dx,dy) =&\lambda \int_0^{z^*} \left[\varphi_x(u,u) + \varphi_y(u,u)\right] \bar{G}(u,u)du + \lambda \int_{z^*}^{\infty} \varphi_y(z^*,u)\,\bar{G}(z^*,u)du \\ =&\lambda\varphi(z^*,z^*)\,\bar{G}(z^*,z^*) - \lambda\varphi(0,0) \\ &+ \underbrace{\lambda \int_0^{z^*} \varphi(u,u)h^x(u,u)\bar{G}(u,u)du}_{(a)} + \underbrace{\lambda \int_0^{z^*} \varphi(u,u)h^y(u,u)\bar{G}(u,u)du}_{(b)} \\ &- \lambda\varphi(z^*,z^*)\,\bar{G}(z^*,z^*) + \underbrace{\lambda \int_{z^*}^{\infty} \varphi(z^*,u)\,h^y(z^*,u)\bar{G}(z^*,u)du}_{(c)} \\ =& - \lambda\varphi(0,0) + \underbrace{\int_{\mathbb{R}_+^2} \varphi(x,y)h^x(x,y)\mathbf{1}_{\{y<z^*\}}\nu^*(dx,dy)}_{(d)} \\ &+ \underbrace{\int_{\mathbb{R}_+^2} \varphi(x,y)h^y(x,y)\nu^*(dx,dy)}_{(e)}.\end{aligned}$$

Indeed, observe that (a) is equal to (d) and (b) plus (c) is equal to (e) by definition of ν^*. We conclude that ν^* is a solution of (17), which completes the proof for LCFS. □

Having established the common equilibrium distribution of LAS and LCFS, we proceed now to examine their performance, in terms of the attained service distribution obtained by the tasks visiting the system.

We will look at the rate at which tasks leave the system with attained service smaller than some given x_0, for the equilibrium distribution. This rate can be expressed as:

$$\begin{aligned}&\int_{\mathbb{R}_+^2} \left[h^x(x,y)\mathbf{1}_{\{x<z^*\}} + h^y(x,y)\right] \mathbf{1}_{\{x\leqslant x_0\}}\nu^*(dx,dy) \quad \text{for} \quad \text{LAS}, \\ &\int_{\mathbb{R}_+^2} \left[h^x(x,y)\mathbf{1}_{\{y<z^*\}} + h^y(x,y)\right] \mathbf{1}_{\{x\leqslant x_0\}}\nu^*(dx,dy) \quad \text{for} \quad \text{LCFS}.\end{aligned}$$

The above expressions correspond to integrating the instantaneous departure rate $\eta_{\nu^*}(x,y)$ over the set of tasks with attained service less than x_0, with respect to the equilibrium measure ν^*. Both integrals are equal since ν^* is supported on the set $\{(u,u) : 0 \leqslant u \leqslant z^*\} \cup \{(z^*,u) : u \geqslant z^*\}$. Hence, we only compute the latter integral.

Proposition 3. *Assume that $\rho > C$ and* (27) *holds. Then*

$$\int_{\mathbb{R}_+^2} \left[h^x(x,y)\mathbf{1}_{\{y<z^*\}} + h^y(x,y)\right] \mathbf{1}_{\{x\leqslant x_0\}}\nu^*(dx,dy) = \lambda P\left(\min\{S,T,z^*\} \leqslant x_0\right). \tag{29}$$

Proof. By definition of ν^*, we have:

$$\begin{aligned}\int_{\mathbb{R}_+^2} \left[h^x(x,y)\mathbf{1}_{\{y<z^*\}} + h^y(x,y)\right]\mathbf{1}_{\{x\leqslant x_0\}}\nu^*(dx,dy) =&\lambda \int_0^{x_0\wedge z^*} \left[h^x(u,u) + h^y(u,u)\right]\bar{G}(u,u)du\\ &+ \mathbf{1}_{\{x_0\geqslant z^*\}}\lambda\int_{z^*}^{\infty} h^y(z^*,u)\bar{G}(z^*,u)du\\ =&-\lambda\int_0^{x_0\wedge z^*}\left[\bar{G}_x(u,u)+\bar{G}_y(u,u)\right]du - \mathbf{1}_{\{x_0\geqslant z^*\}}\lambda\int_{z^*}^{\infty}\bar{G}_y(z^*,u)du\\ =&-\lambda\bar{G}(u,u)\Big|_0^{x_0\wedge z^*} - \mathbf{1}_{\{x_0\geqslant z^*\}}\lambda\bar{G}(z^*,u)\Big|_{z^*}^{\infty}\\ =&\lambda\left[1-P\left(\min\{S,T\}>x_0\wedge z^*\right)\right] + \mathbf{1}_{\{x_0\geqslant z^*\}}\lambda P\left(\min\{S,T\}>z^*\right)\\ =&\lambda\left[1-\mathbf{1}_{\{x_0<z^*\}}P\left(\min\{S,T\}>x_0\right)\right]\\ =&\lambda P\left(\min\{S,T,z^*\}\leqslant x_0\right).\end{aligned}$$

This completes the proof. □

Again, since the total departure rate is λ in equilibrium, the above proposition implies that the distribution of attained service S_a for departing tasks is the same as that of $\min\{S,T,z^*\}$.

7. Comparison between deadline-aware and deadline oblivious policies

We have analyzed fluid models for EDF, LAS and LCFS. We now interpret these fluid models and draw conclusions about the performance of the latter policies in terms of the amount of service attained by tasks in equilibrium.

In Section 5 we obtained the departure rate of tasks with unfinished work S_r smaller than σ_0 for EDF. Since this rate is given by $\lambda P(S-\min\{S,T,\tau^*\}\leqslant\sigma_0)$ and the arrival rate of tasks is λ, we interpret $P(S-\min\{S,T,\tau^*\}\leqslant\sigma_0)$ as the fraction of tasks that leave the system with an amount of unfinished work smaller than σ_0. Similarly, in Section 6 we considered LAS and LCFS, and obtained that the departure rate of tasks with attained service $S_a = S - S_r$ smaller than an arbitrary x_0 is $\lambda P(\min\{S,T,z^*\}\leqslant x_0)$. As for EDF, the total arrival rate is λ, and thus we interpret $P(\min\{S,T,z^*\}\leqslant x_0)$ as the fraction of tasks that leave the system with attained work smaller than x_0.

Recall that the thresholds τ^* and z^* solve the same equation $\lambda E[\min\{S,T,x\}] = C$. This equation has a unique solution when $\rho > C$, so $\tau^* = z^*$ and thus the above interpretations of the unfinished work and attained service in equilibrium mean that all three policies have the same performance in a large-scale regime.

We further note that the equilibrium measures for EDF, LAS and LCFS have threshold structures that can be interpreted similarly as the departure rates discussed above. Specifically, each task i receives service from the EDF policy (at unit rate) if and only if its remaining patience is below the threshold τ^*. As a result, the attained service is $\min\{S_i,T_i,\tau^*\}$, with (S_i,T_i) the service and patience requirements for task i. Analogously, for the LAS and LCFS policies, each task i receives service if and only if its attained service or elapsed time, respectively, is below the threshold z^*. In these cases the attained service is $\min\{S_i,T_i,z^*\}$ as well. Figure 6 illustrates the above interpretations. EDF postpones serving tasks until they are urgent enough, i.e., their patience reaches τ^*, and then serves tasks at full rate. In LAS and LCFS, the situation is the opposite: tasks receive service upon arrival and stop receiving it when the attained service/elapsed time reaches z^*.

As a concrete example, in the case where S and T are independent and exponentially distributed with rates α and β, $\min\{S,T\}$ is exponentially distributed with rate $\alpha+\beta$. In this case, the solution of (27) can be explicitly computed:

$$z^* = -\frac{1}{\alpha+\beta}\log\left(1-\frac{C}{\rho}\right),$$

for $\rho > C$ and the attained service of customers is distributed as $S_a \sim \min\{Z,z^*\}$ with $Z\sim\exp(\alpha+\beta)$. Further analytical expressions can be obtained by evaluating the fluid equilibrium conditions (13) or (17) for exponential $g(\sigma,\tau)$ or constant hazard rates.

The equivalence of the attained service distribution for all three policies is a striking property: the main performance metric of the system can be completely emulated, in the fluid models, by deadline-oblivious policies. In

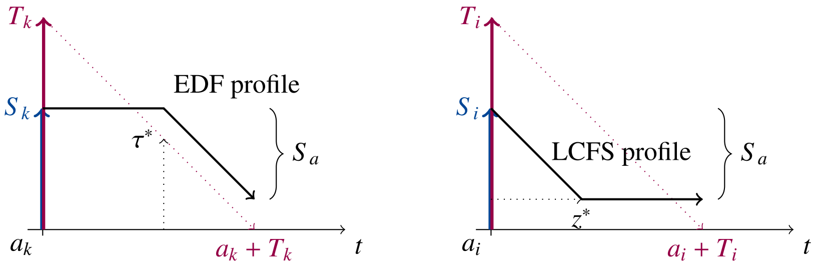

Figure 6: Service profile for the EDF policy (left) and the LAS and LCFS policies (right) for a given task.

particular, LCFS does not need more information than the order of arrival. We now provide simulations showing that this property holds approximately in the stochastic setting as well.

8. Simulation experiments

In this section we compare our fluid models to large-scale stochastic systems with general service and sojourn time distributions. We compare the equilibrium measure obtained in Section 5 with the distribution of the remaining service and time in stationarity, and similarly for the equilibrium measures derived in Section 6. In particular, we focus on the behavior of the thresholds used by these policies, which are shown to oscillate around the corresponding equilibrium values. We also estimate the attained service distribution for each policy, and compare with the distribution obtained from the fluid models. First, we consider a setup in which the equilibrium thresholds admit closed-form expressions. Then we discuss a setup where the service and patience requirements are correlated.

8.1. Parametric setup: Independent Pareto service and sojourn times

First, consider a system with Poisson arrivals at rate λ where S and T are independent and follow a Pareto distribution with common shape parameter $\alpha > 1$, i.e.,

$$P(S > \sigma) = \left(\frac{\theta_S}{\theta_S + \sigma}\right)^{\alpha}, \quad P(T > \tau) = \left(\frac{\theta_T}{\theta_T + \tau}\right)^{\alpha}.$$

In this case, by simple integration, $E[S] = \theta_S/(\alpha - 1)$ and $E[T] = \theta_T/(\alpha - 1)$.

From the above equations and the independence assumption, it follows that:

$$\bar{G}(\sigma, \tau) = \left(\frac{\theta_S \theta_T}{(\theta_S + \sigma)(\theta_T + \tau)}\right)^{\alpha}, \text{and } P(\min\{S, T\} > z) = \bar{G}(z, z) = \left(\frac{\theta_S \theta_T}{(\theta_S + z)(\theta_T + z)}\right)^{\alpha}.$$

From this equation one can compute $E[\min\{S, T\}]$ and $E[\min\{S, T, x\}]$ by integration. We focus on a case where (22) can be solved exactly: with $\theta_S = \theta_T = \theta$, i.e., S and T are also identically distributed. In this special case, the minimum follows a Pareto distribution with the same scale parameter θ and shape parameter 2α. We then have:

$$E[\min\{S, T\}] = \int_0^{\infty} \left(\frac{\theta}{\theta + z}\right)^{2\alpha} dz = \frac{\theta}{2\alpha - 1},$$

and we also have:

$$E[\min\{S, T, x\}] = \int_0^{x} \left(\frac{\theta}{\theta + z}\right)^{2\alpha} dz = \frac{\theta}{2\alpha - 1}\left[1 - \left(\frac{\theta}{\theta + x}\right)^{2\alpha - 1}\right] = E[\min\{S, T\}]\left[1 - \left(\frac{\theta}{\theta + x}\right)^{2\alpha - 1}\right].$$

It is now possible to solve eq. (27):

$$\lambda E[\min\{S, T, z^*\}] = C \Longleftrightarrow \lambda E[\min\{S, T\}]\left[1 - \left(\frac{\theta}{\theta + z^*}\right)^{2\alpha - 1}\right] = C.$$

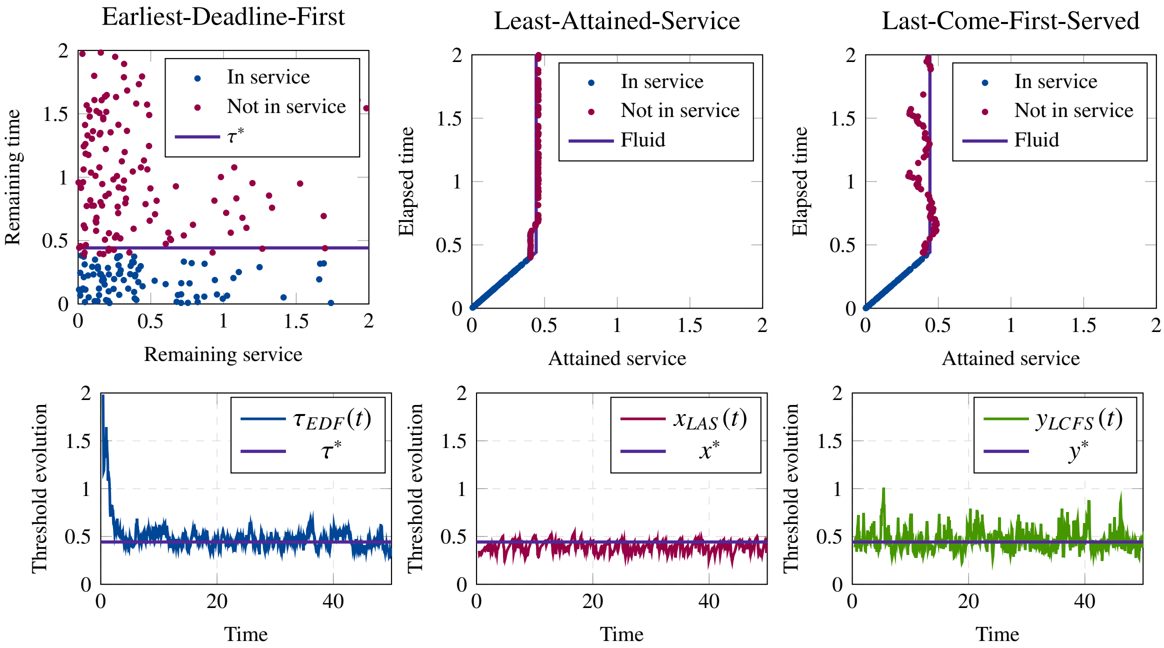

Figure 7: Steady-state snapshot of the state space and stochastic threshold evolution for the three analyzed policies under Pareto traffic.

Recognizing $\lambda E[\min\{S, T\}] = \rho$, the offered load, and solving for z^* we get:

$$1 - \left(\frac{\theta}{\theta + z^*}\right)^{2\alpha-1} = \frac{C}{\rho} \iff z^* = \theta\left[\left(1 - \frac{C}{\rho}\right)^{-\frac{1}{2\alpha-1}} - 1\right], \tag{30}$$

which is the unique solution provided that $\rho > C$, i.e., the system is in the overload condition. All three thresholds $\tau^* = x^* = y^*$ equal z^* and equations (23) and (28) give the equilibrium measures for EDF, LAS and LCFS respectively.

We ran a stochastic simulation of the partial service system, with arrival rate $\lambda = 450$ and $C = 100$ servers, with Pareto distributions for S and T as before, with $\theta = 1$ and $\alpha = 2$. In this case, the system load satisfies:

$$\rho = \lambda\frac{\theta}{3} = 150 > 100 = C,$$

so the system is in overload, and from (30):

$$\tau^* = x^* = y^* \approx 0.442.$$

Figure 7 shows a snapshot of the system for the three policies, in the corresponding state-space representation. The blue dots represent the tasks currently in service, and the red dots the tasks that are not being served. The threshold τ^* is drawn for EDF and the support of the equilibrium measure is depicted for LAS and LCFS. In all three cases, the fluid model captures the system state accurately. Figure 7 also shows how the stochastic thresholds $\tau_{EDF}(t)$, $x_{LAS}(t)$ and $y_{LCFS}(t)$ evolve over time. These thresholds oscillate around the values obtained through the fluid models.

Finally, according to the fluid models, tasks should leave the system with the same amount of attained service $\min\{S, T, z^*\}$ for all three policies. Figure 8 shows the empirical CDFs of the attained service and compares with the distribution obtained from the fluid models, which is just $\bar{G}(z, z)$ truncated at the threshold z^*. The CDFs of the discrete systems are similar to the distribution obtained from the fluid model except at z^*, where the discrete systems have a softer transition. Moreover, all three policies perform similarly, which is consistent with the assertion that in partial service queues deadline-oblivious policies can match the performance of EDF.

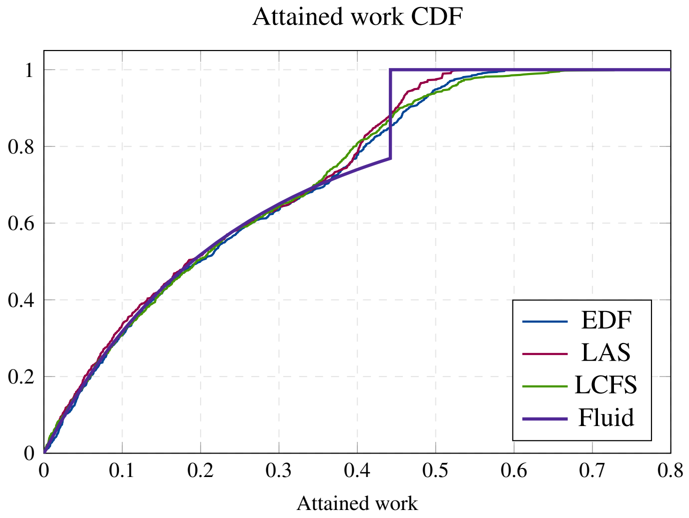

Figure 8: Attained service comparison for the three analyzed policies under Pareto traffic, and the fluid limit approximation.

8.2. *Correlated service and sojourn times*

We now consider a setup where S and T are correlated. Specifically, we set

$$S = e^U \quad \text{and} \quad T = e^V \quad \text{with} \quad (U, V) \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}\right).$$

In particular, the random variables U and V are correlated with normal distributions, and therefore S and T are correlated with log-normal distributions. For all three policies we will consider the same stream of incoming tasks, i.e., with the same arrival, service and patience time for each task and all three policies.

In this case, $E[\min\{S, T\}] \approx 1.36$ can only be numerically estimated. In order to satisfy the overload condition, We choose $\lambda = 120$ and $C = 100$, so the system is in a similar regime as in the previous case, $\rho \approx 160 > 100 = C$. The threshold z^* solving eq. (27) can be numerically estimated to be $z^* \approx 1.322$. Figure 9 shows a snapshot of the state space in steady state, and how the stochastic threshold evolves over time. As before, we observe that tasks remain close to the support of the equilibrium measure of LAS and LCFS. Moreover, the stochastic thresholds oscillate around the values obtained from the fluid models.

Finally, Figure 10 shows the empirical distribution of the attained service and compares with the distribution obtained from the fluid models, i.e., the distribution of $\min\{S, T, z^*\}$. The empirical CDFs are similar to the distribution obtained from the fluid models, and the three policies perform similarly for the discrete systems.

As a final comparison, we plot in Figure 10 the attained service obtained by following the simple FCFS policy, which is also deadline-oblivious. However, this policy is clearly a bad idea: most of the tasks abandon the system while waiting in the queue, and approximately 50% of the tasks leave the system before even reaching service.

9. Conclusions and Future Work

We have analyzed multi–server queueing systems where arriving tasks have two separate characteristics: required service time S and available sojourn time T, with a general joint distribution; abandonments may occur during service and the partial attained service remains valuable, it is our main performance metric. The state representation under these conditions is naturally two-dimensional; we developed a fluid dynamic model in the form of a (weak) partial differential equation, in two versions, depending on whether residual or elapsed times are used. Each is natural for the modeling of different classes of service policies.

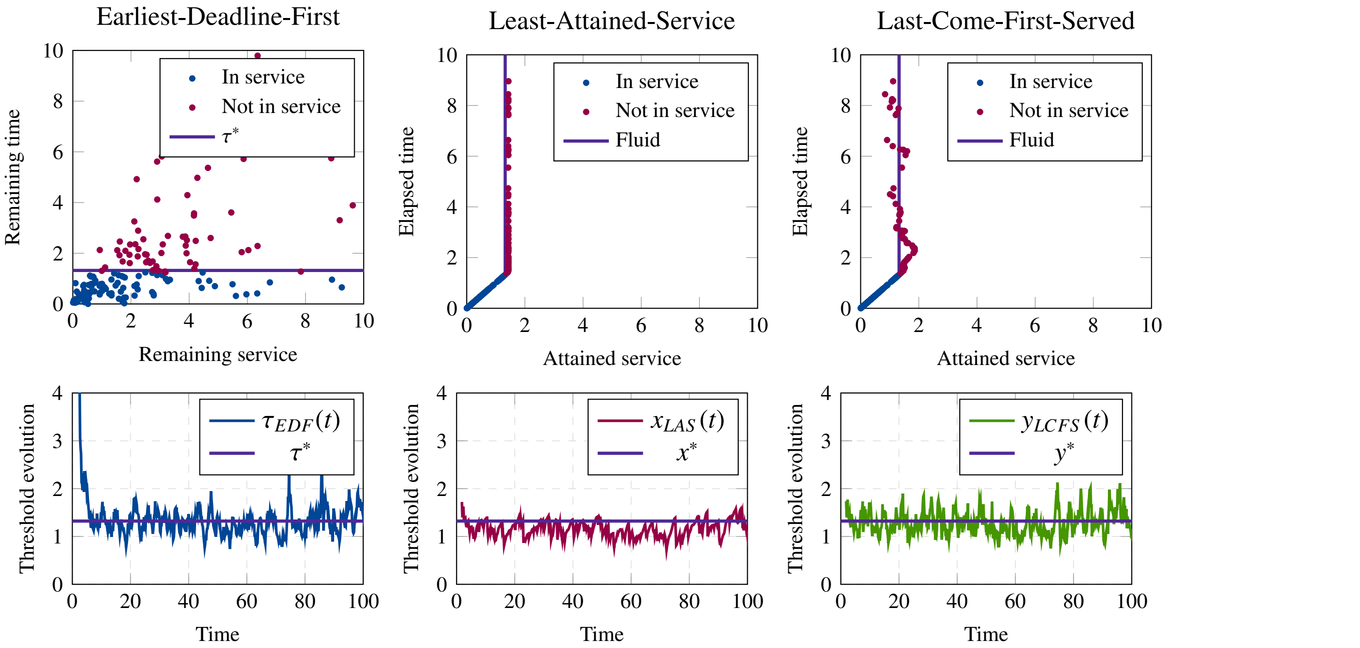

Figure 9: Steady state snapshot of the state space and stochastic threshold evolution for the three analyzed policies under correlated traffic.

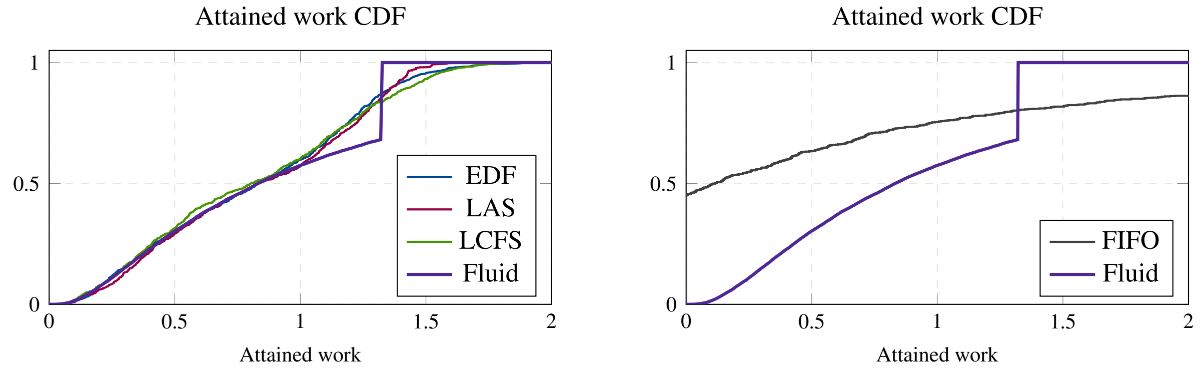

Figure 10: On the left, attained work comparison for the three analyzed policies under correlated traffic, and the (numerically computed) fluid model. On the right, a simulation of the attained work of the FCFS policy, showing that around 50% of the jobs receive no service and abandon while waiting.

In particular, we analyzed task scheduling based on EDF, LAS and LCFS, threshold policies with different information requirements. We characterized the equilibrium measure in each case, and the corresponding distribution of attained service achieved by these policies, which is found to be the *same* in the fluid model. Discrete, stochastic simulations with representative assumptions are used to validate the approximate accuracy of this main conclusion.

A natural continuation of this work is to extend the methodology to other popular scheduling policies. A mathematical question that remains open for future research is the formulation and proof of limit theorems that connect stochastic Markov models with a discrete measure state, with the corresponding fluid representations.

References

[1] Atar, R., Biswas, A., Kaspi, H., 2018. Law of large numbers for the many-server earliest-deadline-first queue. Stochastic Processes and their Applications 128, 2270–2296.

[2] Atar, R., Kang, W., Kaspi, H., Ramanan, K., 2023. Long-time limit of nonlinearly coupled measure-valued equations that model many-server queues with reneging. SIAM Journal on Mathematical Analysis 55, 7189–7239.

[3] Aveklouris, A., DeValve, L., Stock, M., Ward, A., 2024a. Matching impatient and heterogeneous demand and supply. Operations Research In advance (online).

[4] Aveklouris, A., Nakahira, Y., Vlasiou, M., Zwart, B., 2017. Electric vehicle charging: a queueing approach. ACM SIGMETRICS Performance Evaluation Review 45, 33–35.

[5] Aveklouris, A., Puha, A.L., Ward, A.R., 2024b. A fluid approximation for a matching model with general reneging distributions. Queueing Systems 106, 199–238.

[6] Baccelli, F., Boyer, P., Hebuterne, G., 1984. Single-server queues with impatient customers. Advances in Applied Probability 16, 887–905.

[7] Barrer, D., 1957. Queuing with impatient customers and ordered service. Operations Research 5, 650–656.

[8] Bramson, M., 2001. Stability of earliest-due-date, first-served queueing networks. Queueing systems 39, 79–102.

[9] Coffman Jr, E.G., Puhalskii, A.A., Reiman, M.I., Wright, P.E., 1994. Processor-shared buffers with reneging. Performance Evaluation 19, 25–46.

[10] Decreusefond, L., Moyal, P., 2008. Fluid limit of a heavily loaded EDF queue with impatient customers. Markov Processes and Related Fields 14, 131–158.

[11] Doytchinov, B., Lehoczky, J., Shreve, S., 2001. Real-time queues in heavy traffic with earliest-deadline-first queue discipline. Annals of Applied Probability , 332–378.

[12] Evans, L.C., 1998. Partial Differential Equations. AMS.

[13] Ferragut, A., Narbondo, L., Paganini, F., 2022. Edf vehicle charging under deadline uncertainty. ACM SIGMETRICS Performance Evaluation Review 49, 27–29.

[14] Gromoll, H.C., Robert, P., Zwart, B., 2008. Fluid limits for processor-sharing queues with impatience. Mathematics of Operations Research 33, 375–402.

[15] Inoue, Y., Ravner, L., Mandjes, M., 2023. Estimating customer impatience in a service system with unobserved balking. Stochastic Systems 13, 181–210.

[16] Johnson, N.L., Kotz, S., 1975. A vector multivariate hazard rate. Journal of Multivariate Analysis 5, 53–66.

[17] Kang, W., Ramanan, K., 2010. Fluid limits of many-server queues with reneging. Annals of Applied Probabability 20, 2204–2260.

[18] Kang, W., Ramanan, K., 2012. Asymptotic approximations for stationary distributions of many-server queues with abandonment. Annals of Applied Probabability 22, 477–521.

[19] Kaspi, H., Ramanan, K., 2011. Law of large numbers limits for many-server queues. Annals of Applied Probability 21, 33–114.

[20] Kruk, Ł., Lehoczky, J., Ramanan, K., Shreve, S., 2008. Double Skorokhod map and reneging real-time queues, in: Markov Processes and Related Topics: A Festschrift for Thomas G. Kurtz. Institute of Mathematical Statistics, pp. 169–193.

[21] Kruk, Ł., Lehoczky, J., Ramanan, K., Shreve, S., 2011. Heavy traffic analysis for EDF queues with reneging. Annals of Applied Probability 21, 484–545.

[22] Loeser, E.H., Williams, R.J., 2025. Fluid limit for a multi-server, multiclass random order of service queue with reneging and tracking of residual patience times. Queueing Systems 109, 12.

[23] Moyal, P., 2013. On queues with impatience: stability, and the optimality of earliest deadline first. Queueing Systems 75, 211–242.

[24] Moyal, P., Perry, O., 2022. Many-server limits for service systems with dependent service and patience times. Queueing Systems 100, 337–339.

[25] Nakahira, Y., Ferragut, A., Wierman, A., 2019. Minimal-variance distributed deadline scheduling in a stationary environment. SIGMETRICS Perform. Eval. Rev. 46, 56—-61.

[26] Nakahira, Y., Ferragut, A., Wierman, A., 2023. Generalized exact scheduling: A minimal-variance distributed deadline scheduler. Operations Research 71, 433–470.

[27] Panwar, S.S., Towsley, D., Wolf, J.K., 1988. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. Journal of the ACM 35, 832–844.

[28] Puha, A.L., Ward, A.R., 2022. Fluid limits for multiclass many-server queues with general reneging distributions and head-of-the-line scheduling. Mathematics of Operations Research 47, 1192–1228.

[29] Stanford, R.E., 1979. Reneging phenomena in single channel queues. Mathematics of Operations Research 4, 162–178.

[30] Trihinas, D., Michael, P., Symeonides, M., 2024. Towards low-cost and energy-aware inference for edgeai services via model swapping, in: 2024 IEEE International Conference on Cloud Engineering, pp. 168–177.

[31] Ward, A.R., 2012. Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. Surveys in Operations Research and Management Science 17, 1–14.

[32] Ward, A.R., Glynn, P.W., 2005. A diffusion approximation for a GI/GI/1 queue with balking or reneging. Queueing Systems 50, 371–400.

[33] Zeballos, M., Ferragut, A., Paganini, F., 2019. Proportional fairness for EV charging in overload. IEEE Transactions on Smart Grid 10, 6792–6801.

[34] Zuñiga, A.W., 2014. Fluid limits of many-server queues with abandonments, general service and continuous patience time distributions. Stochastic Processes and their Applications 124, 1436–1468.