The last, the least and the urgent...

Fluid modeling and performance equivalence for scheduling policies in partial service queues with abandonment

Andres Ferragut

with Diego Goldsztajn y Fernando Paganini

IFIP Performance 2025

Outline

Outline

Introduction

Partial service queues

Deadline-oblivious policies

Simulations

Final remarks

Motivation

A bit of history...

- Several queueing systems have service and timing requirements.
- Examples:
  - Computing tasks with real-time constraints.
  - Item delivery problems in logistics.
  - Emergency response.
  - etc. etc. etc.
- This has led to a long and rich history of research about queues with abandonments [Barrer, 1957; Stanford, 1979; Baccelli et al., 1984].

Motivation

Recent developments...

One of the most used policies is Earliest-Deadline-First (EDF)

- Give priority to tasks with more urgent deadlines.

Motivation

Recent developments...

One of the most used policies is Earliest-Deadline-First (EDF)

- Give priority to tasks with more urgent deadlines.

Through fluid limits and diffusion approximations, establish performance:

- [Decreusefond and Moyal, 2008] establish EDF fluid limits in the single server case.
- [Kruk et al., 2011] provides diffusion approximations.
- [Moyal, 2013] establish some optimality properties of EDF.
- [Kang and Ramanan, 2010, 2012] analyze the many-server case.
- [Atar et al., 2018, 2023] establish asymptotic performance.

and many others...

Motivation

Common assumption

Customers renege *only* in the queue, and not during service.

Motivation

Common assumption

Customers renege *only* in the queue, and not during service.

We call this the *call-center scenario*:

- Akin to waiting for the customer-help line to pick your call while you listen to annoying music.
- The underlying idea is that when a task reaches service, it will stay until completion.

Key performance metric: number of satisfied tasks (or reneging probability).

Motivation

Partial service queues

In several queueing systems:

- Tasks may abandon during service.
- More importantly, all service provided may be useful.

We call this setting queues with partial service.

Motivation

Partial service queues

In several queueing systems:

- Tasks may abandon during service.
- More importantly, all service provided may be useful.

We call this setting queues with partial service.

Some examples:

- Electrical vehicle charging: customers leave the system with a *partial charge*.
- LLM inference: longer computation times lead to better answers, but these may be interrupted to deliver a quick response.
- File transfers over the Internet, that can be resumed later.

Key points of this talk

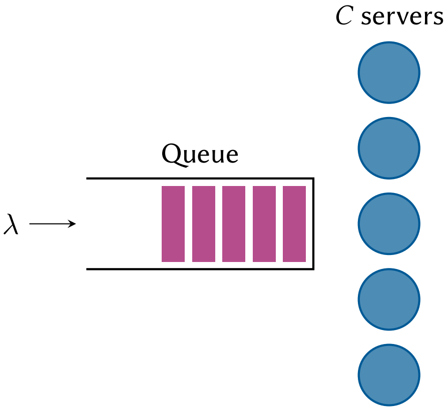
- Provide some suitable representation of the state space and dynamics of these partial service queues.
- Analyze several interesting policies under a suitable fluid model.
- Compute the main performance metric here: attained work.
- *Last but not least:* show that the simple LCFS policy exhibits the same performance than EDF in this setting, without using deadline information.

Outline

Setting

Consider an $M/G/C$ system where:

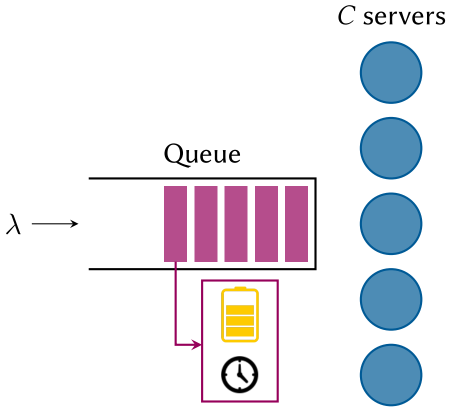
- Tasks arrive as a Poisson process of intensity λ.



Setting

Consider an $M/G/C$ system where:

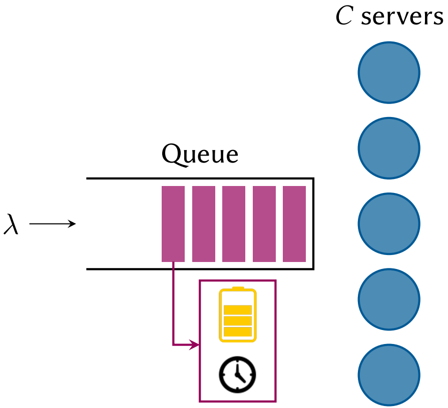
- Tasks arrive as a Poisson process of intensity λ.
- Each task i has two characteristics (marks):
  - S_i: service time (at rate 1).
  - T_i: sojourn time or deadline.



Setting

Consider an $M/G/C$ system where:

- Tasks arrive as a Poisson process of intensity λ.
- Each task i has two characteristics (marks):
  - S_i: service time (at rate 1).
  - T_i: sojourn time or deadline.
- (S_i, T_i) are independent across jobs.
- Follow a common distribution $G(\sigma, \tau)$, possibly correlated.



Partial service queues

Definition

Partial service queue

Customers depart whenever S_i is attained or the timer T_i expires.

Partial service queues

Definition

Partial service queue

Customers depart whenever S_i is attained or the timer T_i expires.

- In particular, they may leave during service.
- Key performance metrics:
  - S_a: amount of service attained.
  - Equivalently, $S_r := S - S_a$, amount of service reneged.

Partial service queue

Customers depart whenever S_i is attained or the timer T_i expires.

- In particular, they may leave during service.
- Key performance metrics:
  - S_a: amount of service attained.
  - Equivalently, $S_r := S - S_a$, amount of service reneged.
- Problem: we have to keep track of remaining service and deadlines simultaneously!

System load

- Before proceeding, it is useful to define the system load:

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

System load

- Before proceeding, it is useful to define the system load:

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

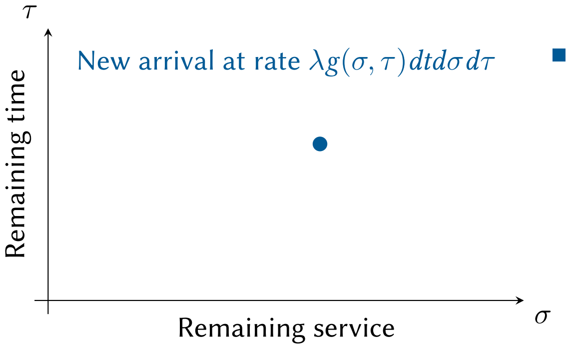
- Interpretation: the mean number of customers on a system with $C = \infty$.

- What we expect in a large scale fluid model:
  - If $\rho < C$ (underload), all tasks can be served, $S_a = \min\{S, T\}$.
  - If $\rho > C$ (overload), demand *curtailing* will occur. How? It depends on the policy...

System evolution

Remaining service times

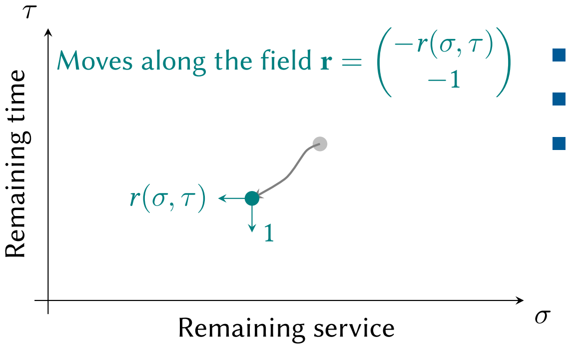
- Consider the remaining time space.



- Consider the remaining time space.

System evolution

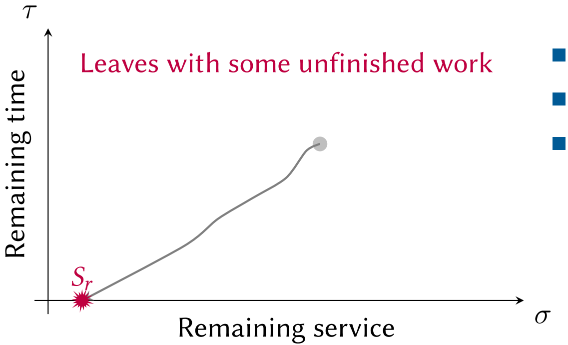
Remaining service times



- Consider the remaining time space.
- Policy defines how tasks are served.
- May depend on any combination of (σ, τ).

System evolution

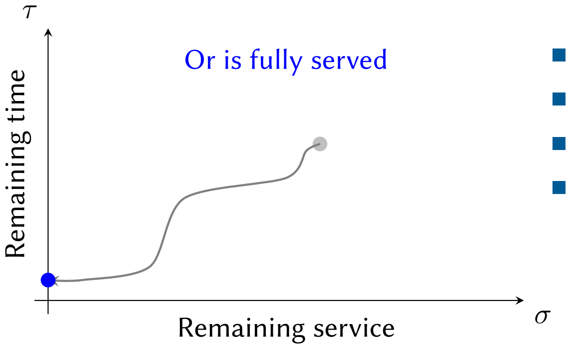
Remaining service times



- Consider the remaining time space.
- Policy defines how tasks are served.
- May depend on any combination of (σ, τ).

System evolution

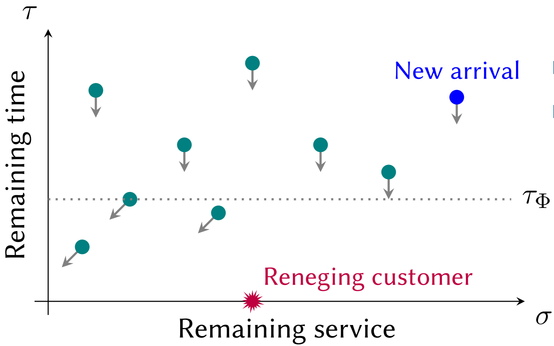
Remaining service times



- Consider the remaining time space.
- Policy defines how tasks are served.
- May depend on any combination of (σ, τ).
- State descriptor:

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

Example: Earliest-deadline-first



- Serve the C most urgent customers.
- Corresponds to taking:

$$r_\Phi(\sigma,\tau) = \mathbf{1}_{\{\tau \leqslant \tau_\Phi\}}$$

with

$$\tau_\Phi := \sup\{\tau \geqslant 0 : \Phi(\mathbb{R}_+ \times (0,\tau]) < C\}.$$

Fluid model dynamics

- Replace Φ_t by a (fluid) measure μ_t.
- Now mass drifts along the field:

$$\mathbf{r}_\mu(\sigma,\tau) = \begin{pmatrix} -r_\mu(\sigma,\tau) \\ -1 \end{pmatrix}$$

- With r_μ satisfying:

$$0 \leqslant r_\mu \leqslant 1$$
$$\iint r_\mu(\sigma,\tau)\mu(d\sigma,d\tau) \leqslant \min\{\mu(\mathbb{R}^2_{++}), C\}.$$

Fluid model dynamics
Transport PDE

If μ_t admits a density $f(\sigma, \tau; t)$ with respect to the Lebesgue measure, it corresponds to:

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

a transport equation.

If μ_t admits a density $f(\sigma, \tau; t)$ with respect to the Lebesgue measure, it corresponds to:

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

a transport equation.

Example: EDF

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial \sigma} \mathbf{1}_{\{\tau < \tau_{\mu_t}\}} + \frac{\partial f}{\partial \tau} + \lambda g$$

EDF Fluid model equilibrium

Imposing equilibrium we get:

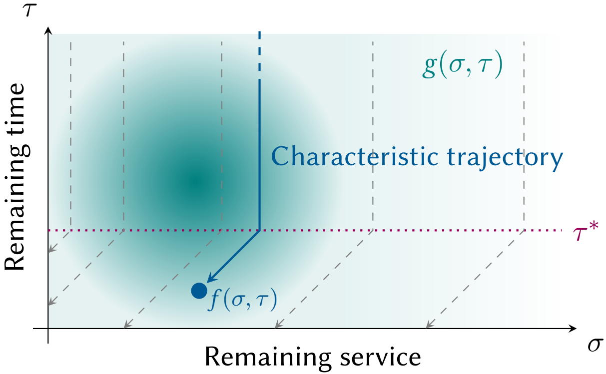
- $\tau_{\mu^*} = \tau^*$ becomes a constant.

- The measure μ^* must satisfy:

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

- Linear PDE that can be easily solved by the method of characteristics.

Solving the EDF transport equation



Theorem

Assume that $\rho > C$ and the equation

$$\lambda E[\min\{S, T, \tau^*\}] = C$$

has a unique solution $\tau^* > 0$. Consider the measure μ^* given by the following density:

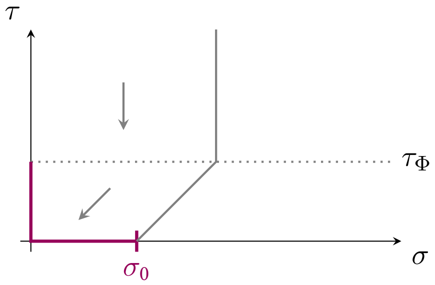
$$f(\sigma, \tau) = \lambda \left[\int_0^{(\tau^* - \tau)^+} g(\sigma + u, \tau + u) du + \int_{(\tau^* - \tau)^+}^{\infty} g\left(\sigma + (\tau^* - \tau)^+, \tau + u\right) du \right].$$

This measure is a fluid equilibrium for the EDF policy, and

$$\tau^* = \sup\left\{\tau \geq 0 : \mu^*(\mathbb{R}_{++} \times (0, \tau]) \leq C\right\}.$$

EDF performance in equilibrium

- Let us compute the rate at which work is reneged.
- Compute the rate at which mass exits with $S_r < \sigma_0$.



Proposition

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

i.e. $S_a = S - S_r = \min\{S, T, \tau^*\}$.

Outline

What if we do not know the deadlines?

- Deadlines are often hard to estimate in practice.
- Moreover, tasks may under-report their deadline to get priority!
- What about deadline-oblivious policies?
  - Can we model them?
  - What is their performance?

What if we do not know the deadlines?

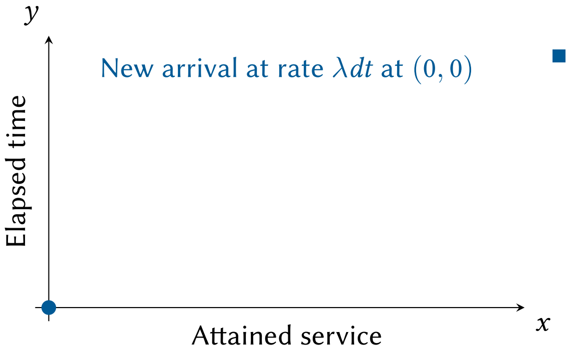
- Deadlines are often hard to estimate in practice.
- Moreover, tasks may under-report their deadline to get priority!
- What about deadline-oblivious policies?
  - Can we model them?
  - What is their performance?

Problem: we need a new state-space...

Attained service state descriptor

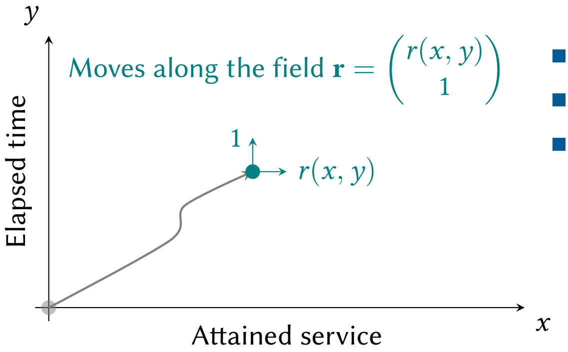
- Consider the elapsed time space.

Attained service state descriptor



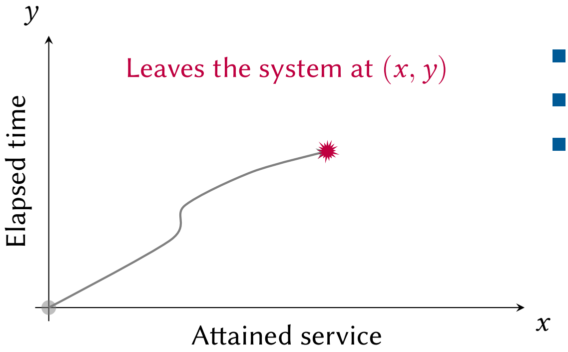
- Consider the elapsed time space.

Attained service state descriptor



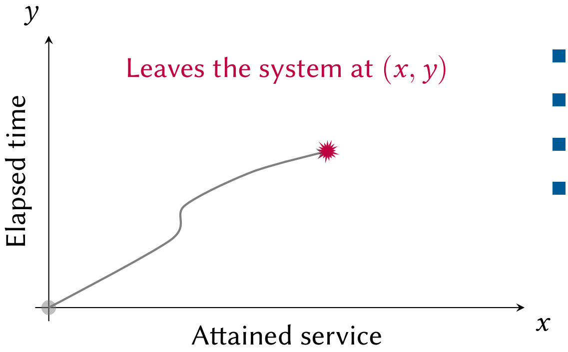
- Consider the elapsed time space.
- Policy again defines how tasks are served.
- May depend on any combination of (x, y).

Attained service state descriptor



- Consider the elapsed time space.
- Policy again defines how tasks are served.
- May depend on any combination of (x, y).

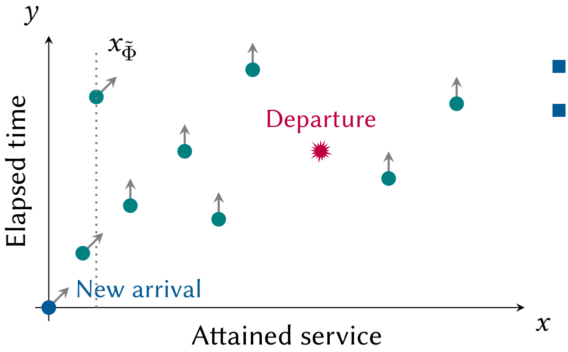
Attained service state descriptor



- Consider the elapsed time space.
- Policy again defines how tasks are served.
- May depend on any combination of (x, y).
- State descriptor:

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

Example: Least-Attained-Service policy



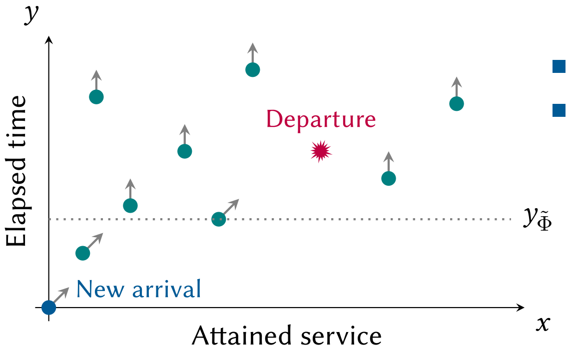
- Serve the C least-served tasks.
- Corresponds to taking:

$$r_{\tilde{\Phi}}(x, y) = \mathbf{1}_{\{x \leqslant x_{\tilde{\Phi}}\}}$$

with

$$x_{\tilde{\Phi}} := \sup\{x : \tilde{\Phi}([0, x] \times \mathbb{R}_+) \leqslant C\}.$$

Example: Last-Come-First-Served policy



- Serve the C more recent tasks.
- Corresponds to taking:

$$r_{\tilde{\Phi}}(x, y) = \mathbf{1}_{\{y \leqslant y_{\tilde{\Phi}}\}}$$

with

$$y_{\tilde{\Phi}} := \sup\{y : \tilde{\Phi}(\mathbb{R}_+ \times [0, y]) \leqslant C\}$$

The hazard rate field

We have a new problem: what is the rate at which users leave the system?

The hazard rate field

We have a new problem: what is the rate at which users leave the system?

Let $\bar{G}(x,y) = P(S > x, T > y)$ and define:

Definition (Hazard rate field)

$$\mathbf{h}(x,y) = -\nabla \log \bar{G}(x,y) \quad \text{i.e.}$$

- $h^x(x,y) = P(S \in [x, x+dx], T > S \mid S > x, T > y)$
- $h^y(x,y) = P(T \in [y, y+dy], S > T \mid S > x, T > y)$

Interpretation: $\mathbf{h}$ stores the rate at which $\min\{S,T\}$ is attained due to S or T expiring.

Fluid model dynamics

- Replace $\tilde{\Phi}_t$ by a (fluid) measure ν_t.
- Now mass arrives at $(0,0)$ at rate λ.
- Drifts along the field:

$$\mathbf{r}_\nu(x,y) = \begin{pmatrix} r_\nu(x,y) \\ 1 \end{pmatrix}$$

- With r_ν satisfying:

$$0 \leqslant r_\nu \leqslant 1$$

$$\iint r_\nu(x,y)\nu(dx,dy) \leqslant \min\{\nu(\mathbb{R}^2_+), C\}.$$

Attained service transport equation

- We now have all ingredients to formulate the dynamics of the system.
- The transport equation in the elapsed service space is (informally):

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

where $\tilde{f}$ is the density of ν_t.

Attained service transport equation

- We now have all ingredients to formulate the dynamics of the system.
- The transport equation in the elapsed service space is (informally):

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

where $\tilde{f}$ is the density of ν_t.

- The above equation must be treated in weak form:
  - To account for the impulse mass at $(0,0)$ driving the system.
  - To allow solutions without a density as we shall see.

Recall that LCFS can be modeled by:

$$r_\nu(x, y) = \mathbf{1}_{\{y < y_\nu\}}$$

with

$$y_\nu = \sup\left\{y \geq 0 : \nu(\mathbb{R}_+ \times [0, y]) \leqslant C\right\}.$$

Last come first served

Fluid equilibrium

Recall that LCFS can be modeled by:

$$r_\nu(x, y) = \mathbf{1}_{\{y < y_\nu\}}$$

with

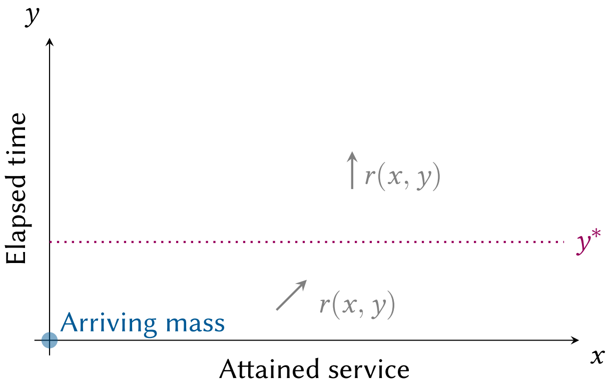
$$y_\nu = \sup\left\{y \geq 0 : \nu(\mathbb{R}_+ \times [0, y]) \leqslant C\right\}.$$

Imposing equilibrium, ν^*, y^* fixed, we have to solve:

$$\nabla \cdot \left[\mathbf{r}_{\nu^*} \bar{f}\right] + [\mathbf{r}_{\nu^*} \cdot \mathbf{h}]\bar{f} = \lambda \delta_{(0,0)}.$$

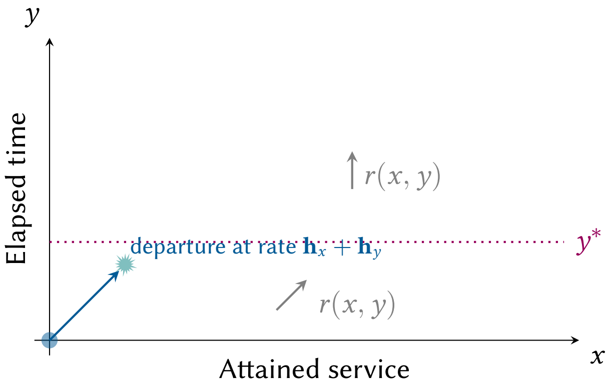
Solving the transport equation

Last come first served case



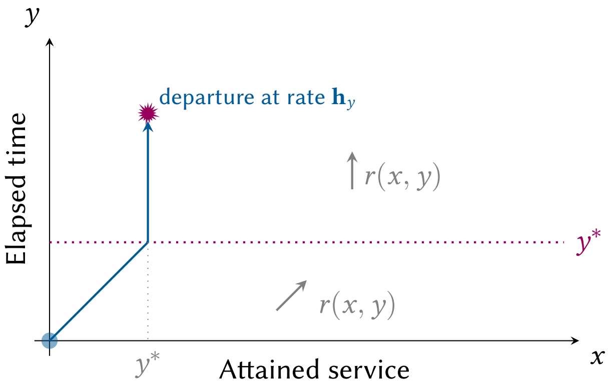
Solving the transport equation

Last come first served case



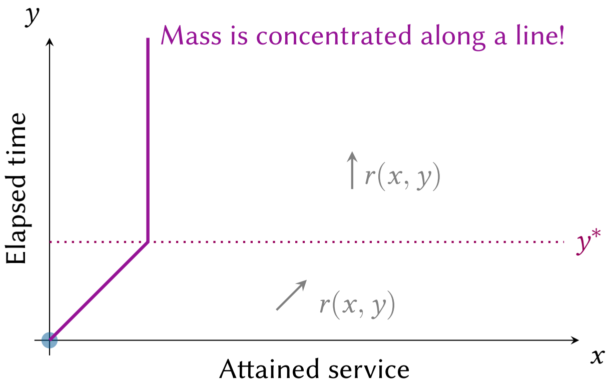
Solving the transport equation

Last come first served case



Solving the transport equation

Last come first served case



Deadline-oblivious policies in overload

Theorem

Assume that $\rho > C$ *and the equation*

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

has a unique solution $z^* > 0$*. Consider the measure* ν^* *given by:*

$$\langle \varphi, \nu^* \rangle = \lambda \left[\int_0^{z^*} \varphi(u,u)\bar{G}(u,u)du + \int_{z^*}^{\infty} \varphi\left(z^*, u\right) \bar{G}(z^*, u) du \right],$$

for all $\varphi \in C_c(\mathbb{R}_+^2)$*. Then this measure is the equilibrium measure for both the Least-Attained-Service and Last-Come-First-Served policies.*

LAS/LCFS performance in equilibrium

Compute the rate at which mass leaves the system with less than x_0 attained service:

$$\iint_{[0,x_0]\times\mathbb{R}_+} \eta_{\nu^*}(x,y)\nu^*(dx,dy).$$

LAS/LCFS performance in equilibrium

Compute the rate at which mass leaves the system with less than x_0 attained service:

$$\iint_{[0,x_0]\times\mathbb{R}_+} \eta_{\nu^*}(x,y)\nu^*(dx,dy).$$

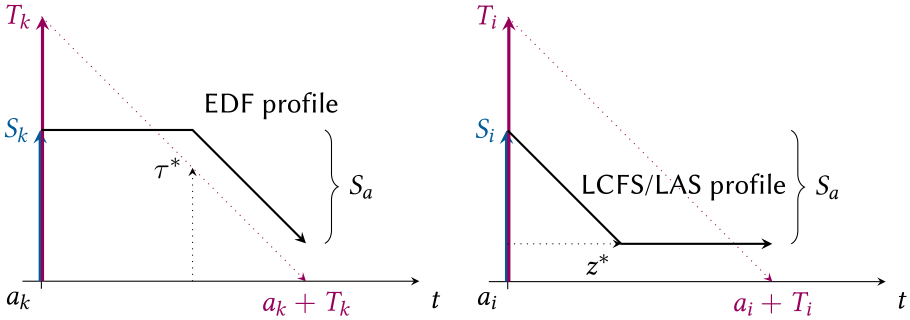
Proposition

Assume that $\rho > C$. Then

$$\int_{[0,x_0]\times\mathbb{R}_+} \left[h^x(x,y)\mathbf{1}_{\{y<z^*\}} + h^y(x,y)\right]\nu^*(dx,dy) = \lambda P\left(\min\{S,T,z^*\} \leqslant x_0\right).$$

So again the attained work is $S_a = \min\{S,T,z^*\}$!!

Graphical explanation



Since $\tau^* = x^* = y^* = z^*$, performance is the same in all three policies!!!

Outline

Introduction

Partial service queues

Deadline-oblivious policies

Simulations

Final remarks

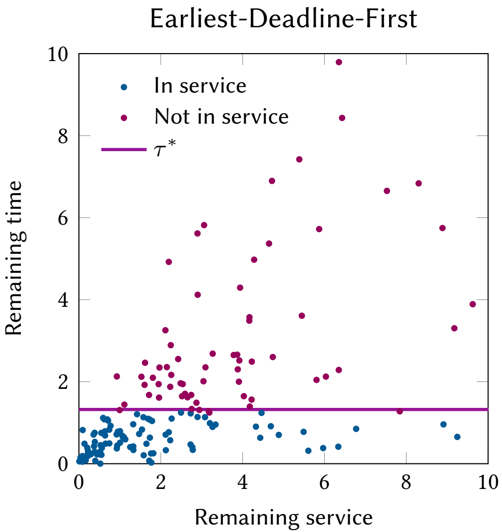
Simulations with correlated S and T

- We finally validate our fluid approximation by stochastic simulations
- In order to account for correlations, we take:

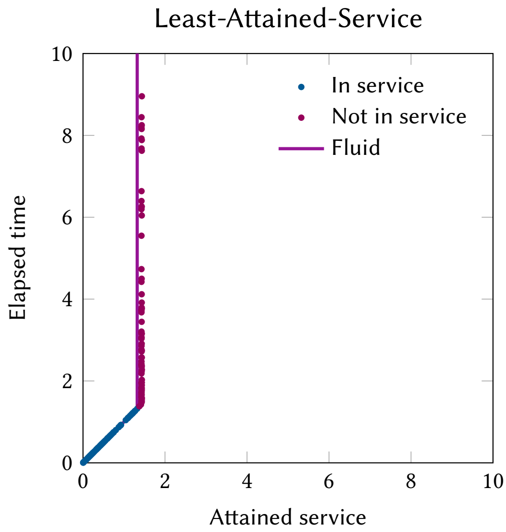
$$S = e^U \quad \text{and} \quad T = e^V \quad \text{with} \quad (U, V) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right).$$

- In particular, the random variables U and V are correlated with normal distributions, and therefore S and T are correlated with log-normal distributions.
- In this case, $E[\min\{S, T\}] \approx 1.36$ can only be numerically estimated.
- We choose $\lambda = 120$ and $C = 100$, then $\rho \approx 160$ and $z^* \approx 1.322$.

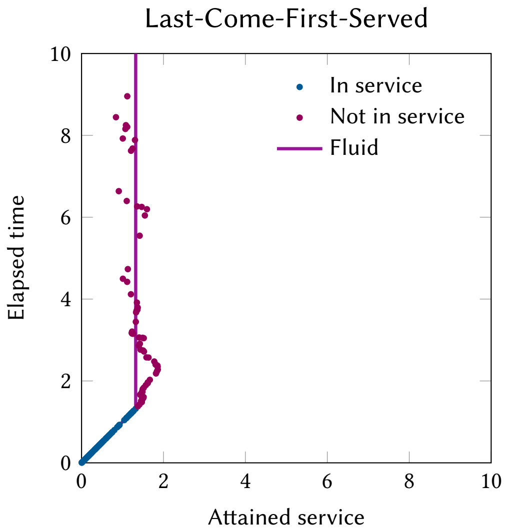
State space snapshots



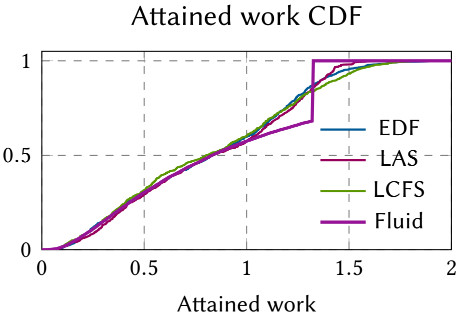
State space snapshots



State space snapshots



Attained work empirical CDF



Even in the pre-limit system, performance is similar!

Outline

Messages from the talk

- Measure-valued processes are a powerful tool to model general service queues.
- Partial service queues require two-dimensional measures.
- Our proposed dynamics for fluid models are tractable and approximate the real system.
- Last-but-not-least: in this setting, deadline-oblivious policies can be used without performance penalty!

Future work

- Analyze further policies using these tools (FCFS is easy for instance).
- Establish process-level convergence to the fluid models (almost done!)
- Devise new policies and/or analyze different settings:
  - Tasks stay until service completion, but we want to measure the average *tardiness*, i.e. how late they depart.

Thank you!

Andres Ferragut
ferragut@ort.edu.uy
https://aferragu.github.io

References I

R. Atar, A. Biswas, and H. Kaspi. Law of large numbers for the many-server earliest-deadline-first queue. *Stochastic Processes and their Applications*, 128(7):2270–2296, 2018.

R. Atar, W. Kang, H. Kaspi, and K. Ramanan. Long-time limit of nonlinearly coupled measure-valued equations that model many-server queues with reneging. *SIAM Journal on Mathematical Analysis*, 55(6):7189–7239, 2023.

F. Baccelli, P. Boyer, and G. Hebuterne. Single-server queues with impatient customers. *Advances in Applied Probability*, 16(4):887–905, 1984.

D. Barrer. Queuing with impatient customers and ordered service. *Operations Research*, 5(5): 650–656, 1957.

L. Decreusefond and P. Moyal. Fluid limit of a heavily loaded EDF queue with impatient customers. *Markov Processes and Related Fields*, 14(1):131–158, 2008.

W. Kang and K. Ramanan. Fluid limits of many-server queues with reneging. *Annals of Applied Probabability*, 20(6):2204–2260, Dec. 2010.

References II

W. Kang and K. Ramanan. Asymptotic approximations for stationary distributions of many-server queues with abandonment. *Annals of Applied Probabability*, 22(2):477–521, Apr. 2012.

Ł. Kruk, J. Lehoczky, K. Ramanan, and S. Shreve. Heavy traffic analysis for EDF queues with reneging. *Annals of Applied Probability*, 21(2):484–545, 2011.

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

R. E. Stanford. Reneging phenomena in single channel queues. *Mathematics of Operations Research*, 4(2):162–178, 1979.