Last but not least...

Matching Earliest-Deadline-First performance through deadline-oblivious policies

Andres Ferragut

Universidad ORT Uruguay

INRIA Paris Seminar – June 2025

Outline

Introduction

A crash course on measure valued processes

Partial service queues

Performance analysis

Simulations

Future work

Outline

Introduction

A crash course on measure valued processes

Partial service queues

Performance analysis

Simulations

Future work

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)

- Serve the customers with more urgent deadlines in priority.

Motivation

Recent developments...

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

- Serve the customers with more urgent deadlines in priority.

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

Introduction

A crash course on measure valued processes

Partial service queues

Performance analysis

Simulations

Future work

Outline

Introduction

A crash course on measure valued processes

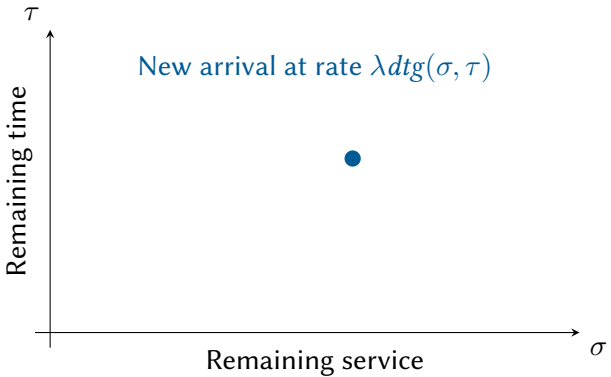
Partial service queues

Performance analysis

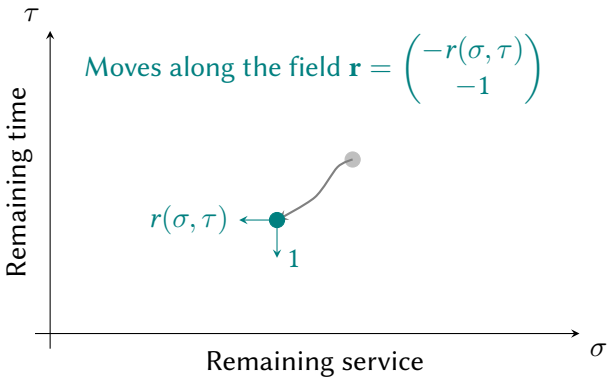
Simulations

Future work

Remaining service state descriptor

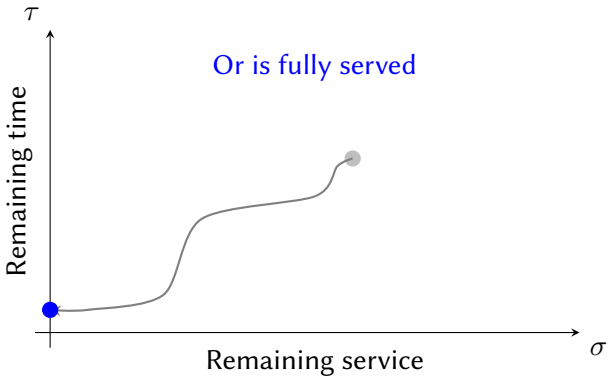


Remaining service state descriptor



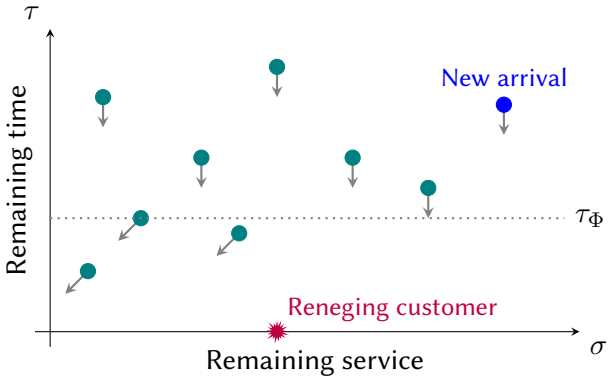
Remaining service state descriptor

Remaining service state descriptor

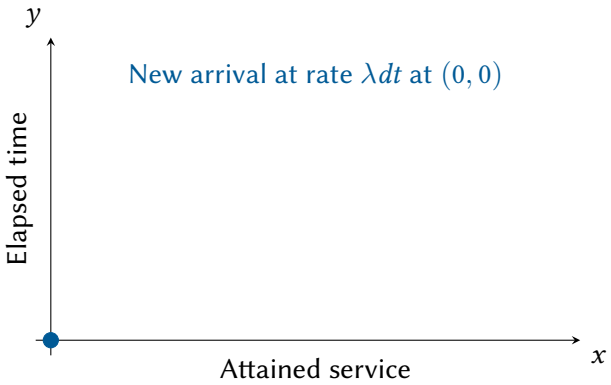


Example

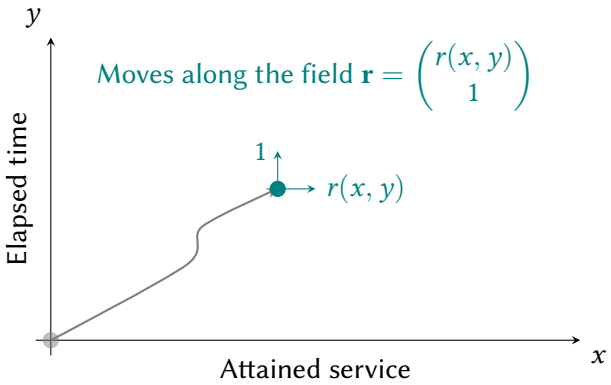
Earliest-deadline-first



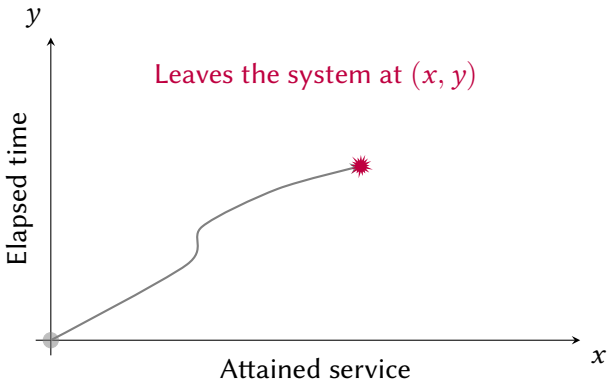
Attained service state descriptor



Attained service state descriptor

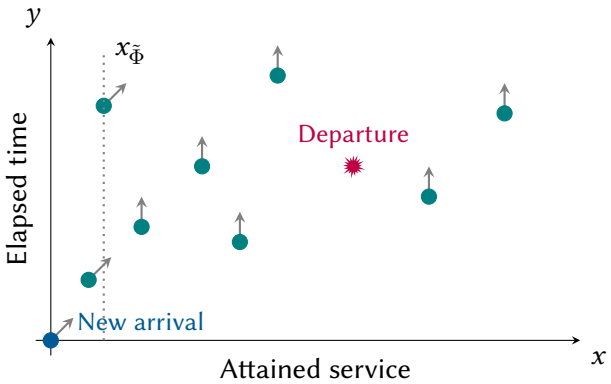


Attained service state descriptor



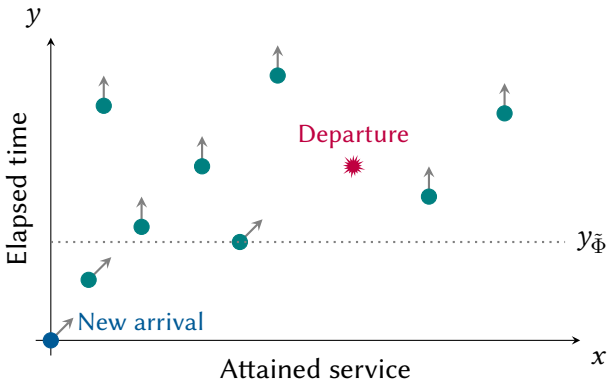
Example

Least-Attained-Service policy



Example

Last-Come-First-Served policy



The hazard rate field

Outline

Introduction

A crash course on measure valued processes

Partial service queues

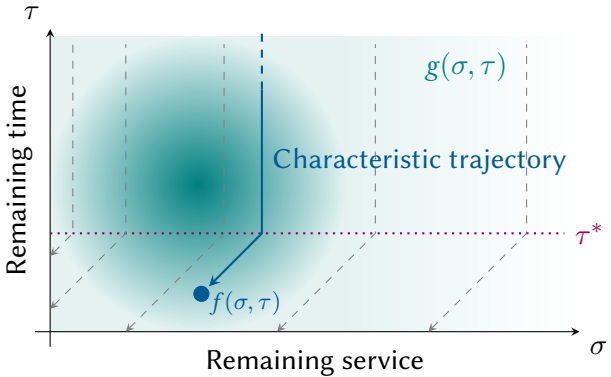
Performance analysis

Simulations

Future work

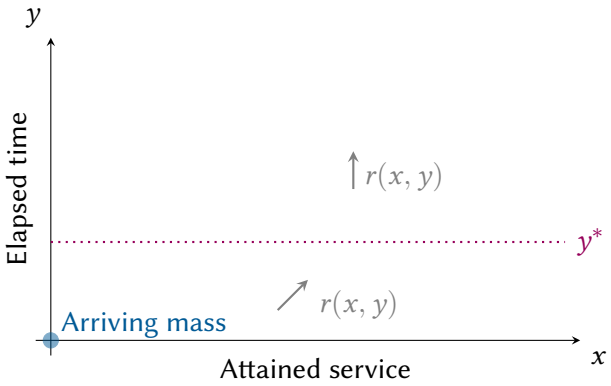
Solving the transport equation

Remaining service case



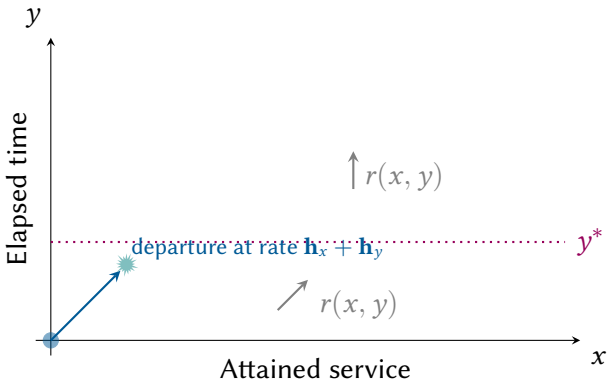
Solving the transport equation

Attained service case



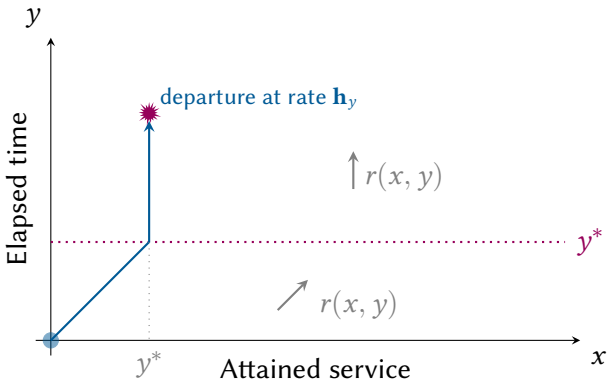
Solving the transport equation

Attained service case

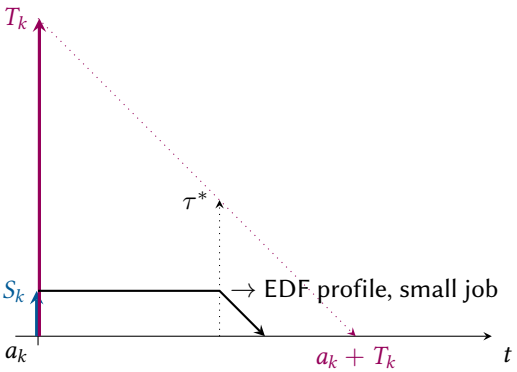


Solving the transport equation

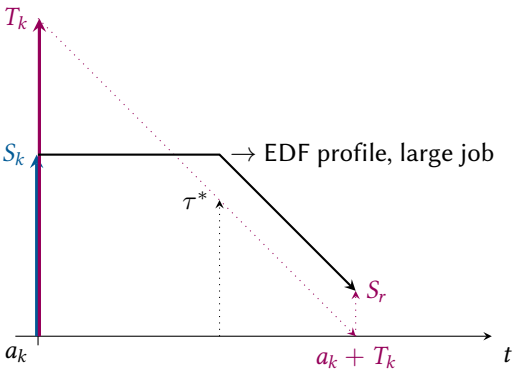
Attained service case



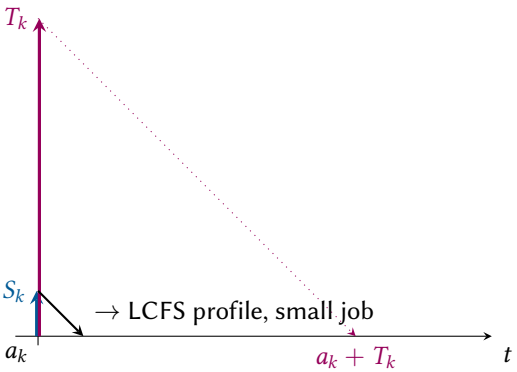
Perceived performance in EDF



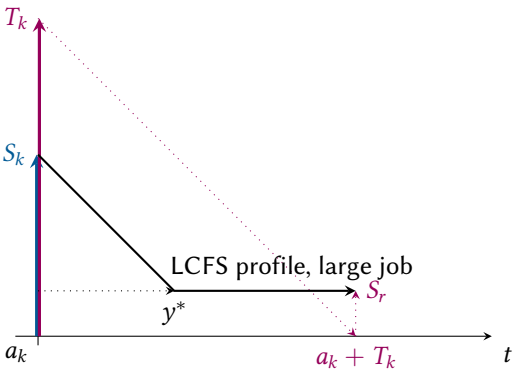
Perceived performance in EDF



Perceived performance in LAS and LCFS



Perceived performance in LAS and LCFS



Outline

Introduction

A crash course on measure valued processes

Partial service queues

Performance analysis

Simulations

Future work

Outline

Introduction

A crash course on measure valued processes

Partial service queues

Performance analysis

Simulations

Future work

Final remarks

Merci beaucoup!

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.