Last but not least...

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

Andres Ferragut

Universidad ORT Uruguay

INRIA Paris Seminar – June 2025

Outline

Outline

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

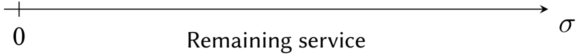
- Provide some suitable representation of the state space and dynamics of these partial service queues.
- Analyze several interesting policies under a suitable fluid model.
- Compute the main performance metric here: attained work.
- *Last but not least:* show that the simple LCFS policy exhibits the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

Outline

Measure valued stochastic processes

Consider the simple $M/G/\infty$ queue:

- Tasks arrive as a Poisson process of intensity λ.
- Each task has a service requirement $S \sim g(\sigma)$.



Measure valued stochastic processes

Consider the simple $M/G/\infty$ queue:

- Tasks arrive as a Poisson process of intensity λ.
- Each task has a service requirement $S \sim g(\sigma)$.

Measure valued stochastic processes

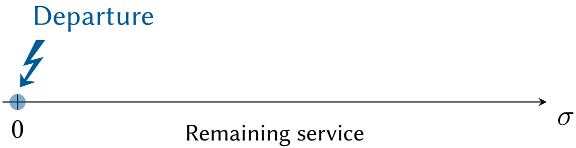
Consider the simple $M/G/\infty$ queue:

- Tasks arrive as a Poisson process of intensity λ.
- Each task has a service requirement $S \sim g(\sigma)$.

Measure valued stochastic processes

Consider the simple $M/G/\infty$ queue:

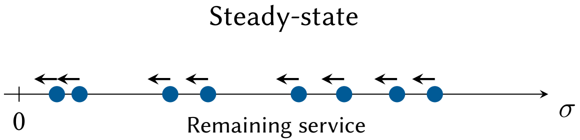
- Tasks arrive as a Poisson process of intensity λ.
- Each task has a service requirement $S \sim g(\sigma)$.



Measure valued stochastic processes

Consider the simple $M/G/\infty$ queue:

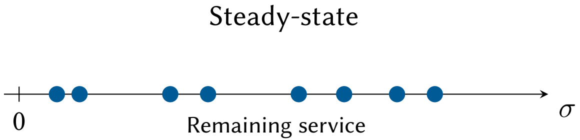
- Tasks arrive as a Poisson process of intensity λ.
- Each task has a service requirement $S \sim g(\sigma)$.



Measure valued stochastic processes

Consider the simple $M/G/\infty$ queue:

- Tasks arrive as a Poisson process of intensity λ.
- Each task has a service requirement $S \sim g(\sigma)$.



State-descriptor:

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

a Point-process on the positive half-line.

M/G/∞, steady state

- Φ_t is a measure-valued Markov process.
- Its dynamics can be characterized through its generator.
- In steady state:

$$\Phi \sim \text{Poisson Process with mean measure } \mu(d\sigma) = \lambda \bar{G}(\sigma) d\sigma$$

where $\bar{G}$ is the CCDF of S.

M/G/∞, steady state

- Φ_t is a measure-valued Markov process.
- Its dynamics can be characterized through its generator.
- In steady state:

$$\Phi \sim \text{Poisson Process with mean measure } \mu(d\sigma) = \lambda \bar{G}(\sigma) d\sigma$$

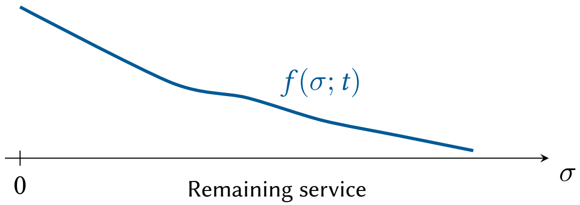
where $\bar{G}$ is the CCDF of S.

Interpretation:

- Write $\mu(d\sigma) = \rho \left[\frac{1}{E[S]}(1 - G(\sigma)) \right] d\sigma$, with $\rho = \lambda E[S]$.
- Then $\left[\frac{1}{E[S]}(1 - G(\sigma)) \right] d\sigma$ is the *residual service time distribution* associated to G.
- In steady-state, the total number of customers $\sim$ Poisson(ρ) and distributed in σ as the residual lifetime distribution.

M/G/∞, fluid approximation.

Suppose that we can replace Φ_t by a general measure μ_t with density $f(\sigma;t)$.



- Mass is transported to the left at rate 1.
- New mass arrives at σ with intensity $\lambda g(\sigma)d\sigma dt$.

We can combine this in the following transport equation:

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial \sigma} + \lambda g(\sigma).$$

M/G/∞, fluid approximation.

Imposing equilibrium and the boundary condition $f(\sigma) \to 0$ as $\sigma \to \infty$ we get:

$$\frac{\partial f}{\partial \sigma} + \lambda g(\sigma) = 0 \Longrightarrow f(\sigma) = \lambda \int_{\sigma}^{\infty} g(u)\, du = \lambda \bar{G}(\sigma),$$

so the fluid approximation recovers the mean measure of Φ.

M/G/∞, fluid approximation.

Imposing equilibrium and the boundary condition $f(\sigma) \to 0$ as $\sigma \to \infty$ we get:

$$\frac{\partial f}{\partial \sigma} + \lambda g(\sigma) = 0 \Longrightarrow f(\sigma) = \lambda \int_{\sigma}^{\infty} g(u)\, du = \lambda \bar{G}(\sigma),$$

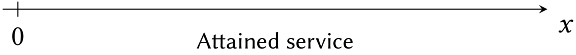
so the fluid approximation recovers the mean measure of Φ.

- This is a deterministic measure, with total mass ρ...
- ...distributed in the real line as the residual service distribution.
- Serves as an approximation of Φ in a large scale system ($\lambda \to \infty$).

M/G/∞: take two

Attained service state descriptor

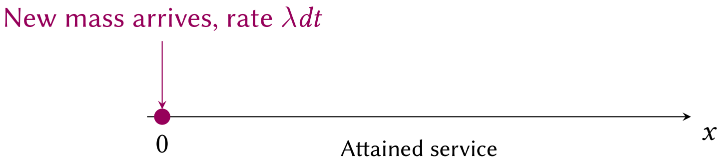
Here is another approach to model the same system [Kang and Ramanan, 2010]:



M/G/∞: take two

Attained service state descriptor

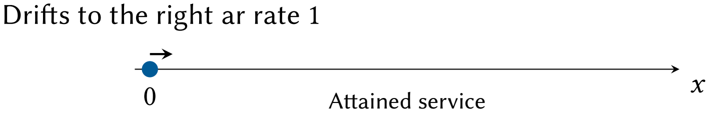
Here is another approach to model the same system [Kang and Ramanan, 2010]:



M/G/∞: take two

Attained service state descriptor

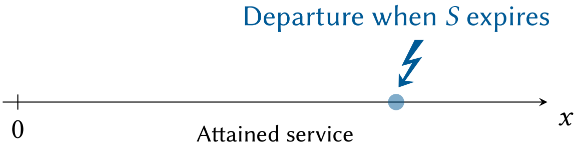
Here is another approach to model the same system [Kang and Ramanan, 2010]:



M/G/∞: take two

Attained service state descriptor

Here is another approach to model the same system [Kang and Ramanan, 2010]:

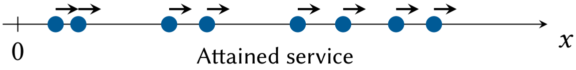


M/G/∞: take two

Attained service state descriptor

Here is another approach to model the same system [Kang and Ramanan, 2010]:

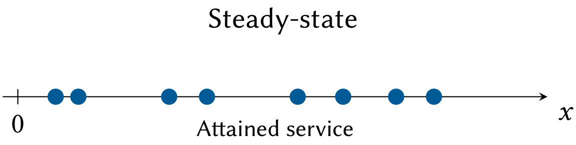
Steady-state



M/G/∞: take two

Attained service state descriptor

Here is another approach to model the same system [Kang and Ramanan, 2010]:



State-descriptor:

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

a Point-process on the positive half-line, where $x_i(t)$ is the elapsed time in the system

M/G/∞, take two

Steady-state

$\tilde{\Phi}_t$ is a measure-valued Markov process.

- Mass always arrive at 0 with rate λdt.
- Transports to the right at rate 1.
- Leaves the system at rate $h(x)$, the hazard rate function:

$$h(x) = \lim_{dt \to 0} P(S \in [x, x + dt] \mid S > x) = \frac{g(x)}{\bar{G}(x)} = -\frac{\partial}{\partial x} \log \bar{G}(x).$$

M/G/∞, take two
Steady-state

$\tilde{\Phi}_t$ is a measure-valued Markov process.

- Mass always arrive at 0 with rate λdt.
- Transports to the right at rate 1.
- Leaves the system at rate $h(x)$, the hazard rate function:

$$h(x) = \lim_{dt \to 0} P(S \in [x, x + dt] \mid S > x) = \frac{g(x)}{\bar{G}(x)} = -\frac{\partial}{\partial x} \log \bar{G}(x).$$

Steady-state:

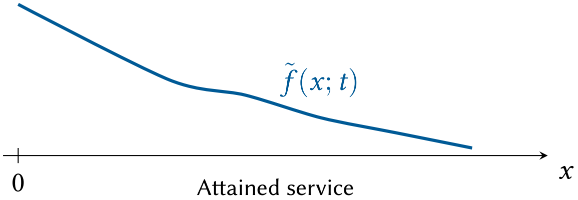
$$\tilde{\Phi} \sim \text{Poisson Process with mean measure } \nu(dx) = \lambda \bar{G}(x) dx$$

So the reversed representation has the same distribution, because in a random point in time the elapsed service and the remaining service have the same distribution.

M/G/∞: take two

Fluid approximation.

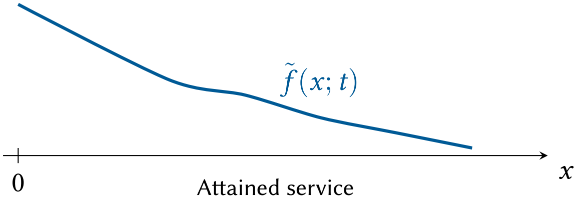
Suppose that we can replace $\tilde{\Phi}_t$ by a general measure ν_t with density $\tilde{f}(x;t)$.



M/G/∞: take two

Fluid approximation.

Suppose that we can replace $\tilde{\Phi}_t$ by a general measure ν_t with density $\tilde{f}(x;t)$.



The corresponding transport equation is (informally):

$$\frac{\partial \tilde{f}}{\partial t} = -\frac{\partial \tilde{f}}{\partial x} - h(x)\tilde{f} + \lambda\delta_0.$$

$M/G/\infty$: take two

Fluid equilibrium.

Imposing equilibrium we get:

$$\frac{\partial \tilde{f}}{\partial x} = -h(x)\tilde{f} + \lambda\delta_0.$$

Solving (in a distribution sense) with the boundary condition $\tilde{f}(\infty) = 0$ we get:

$$\tilde{f}(x) = \lambda e^{-\int_0^x h(u)du}.$$

But by definition $\int_0^x h(u)du = -\log \bar{G}(x)$, and thus:

$$\tilde{f}(x) = \lambda\bar{G}(x)$$

So the transport fluid equation recovers again the mean measure of the steady-state.

Lessons learned

- We can model M/G systems by using two state descriptors:
  - The remaining service Φ.
  - The attained service $\tilde{\Phi}$.
- Both admit reasonable fluid approximations, which correspond to transport equations.
- In fact this has been used in the literature to model abandonments (since they operate as $M/G/\infty$ systems in some sense).

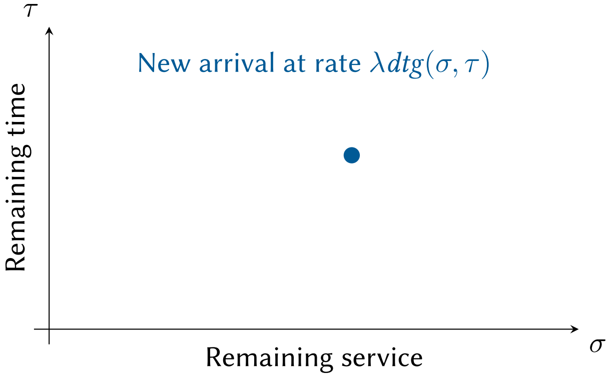
Lessons learned

- We can model M/G systems by using two state descriptors:
  - The remaining service Φ.
  - The attained service $\tilde{\Phi}$.
- Both admit reasonable fluid approximations, which correspond to transport equations.
- In fact this has been used in the literature to model abandonments (since they operate as $M/G/\infty$ systems in some sense).

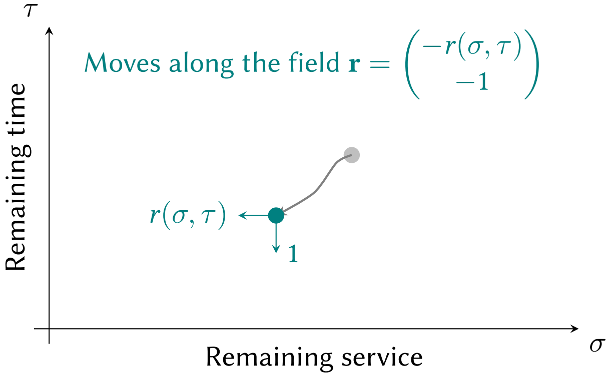
Question: can we do more using this machinery of measure-valued processes?

Outline

Remaining service state descriptor

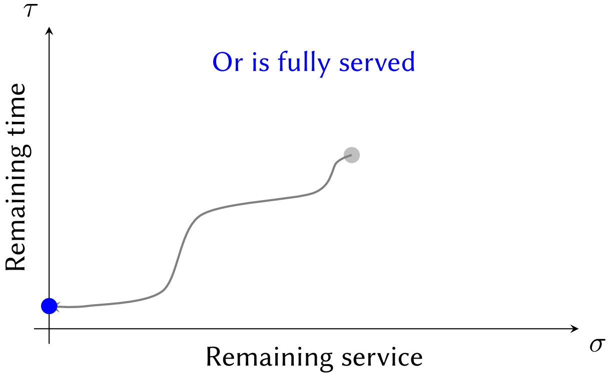


Remaining service state descriptor



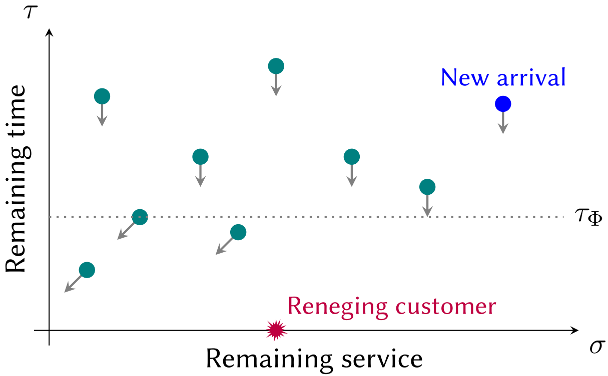
Remaining service state descriptor

Remaining service state descriptor

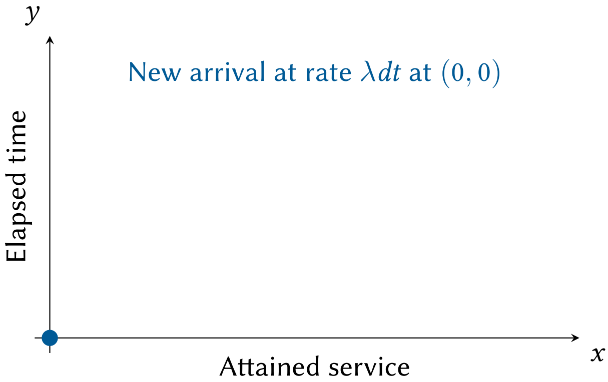


Example

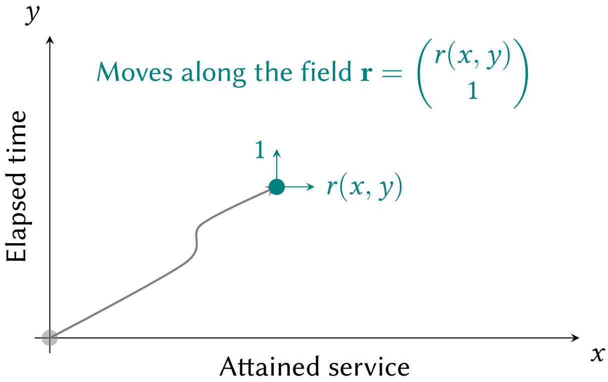
Earliest-deadline-first



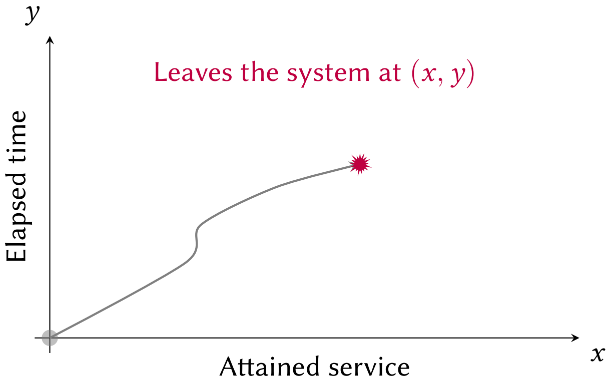
Attained service state descriptor



Attained service state descriptor

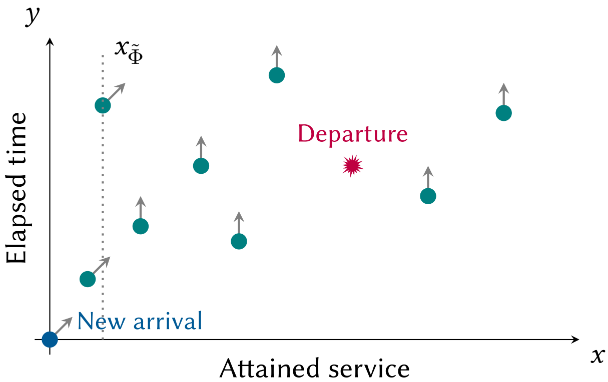


Attained service state descriptor



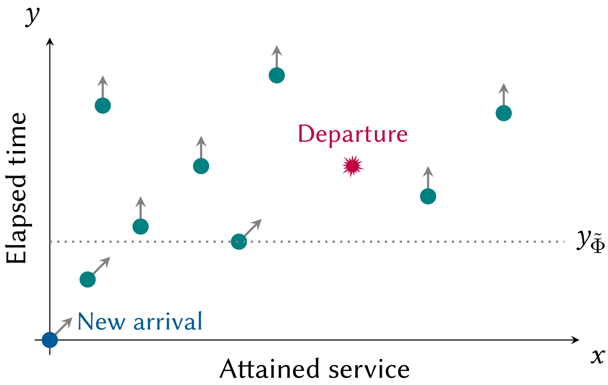
Example

Least-Attained-Service policy



Example

Last-Come-First-Served policy

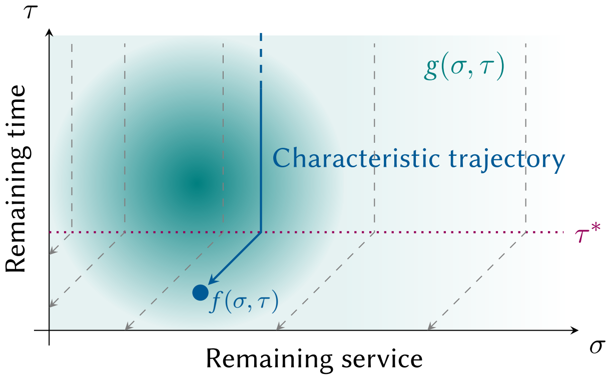


The hazard rate field

Outline

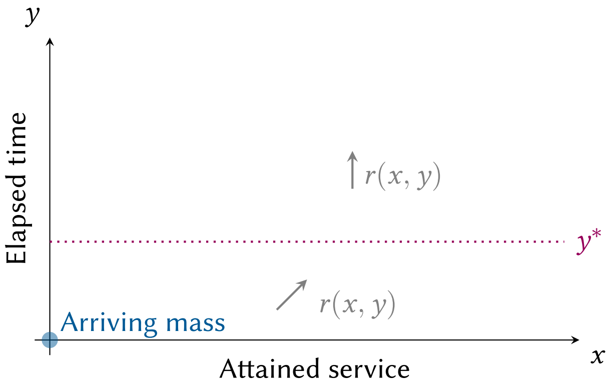
Solving the transport equation

Remaining service case



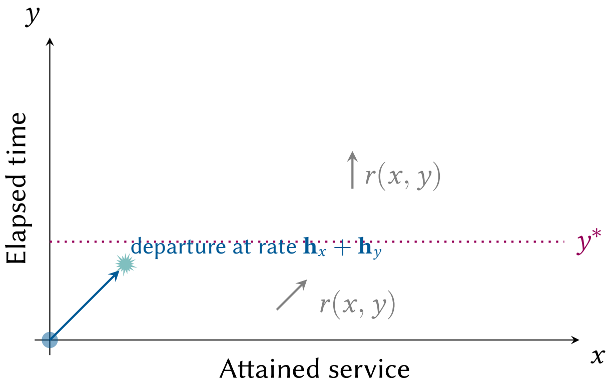
Solving the transport equation

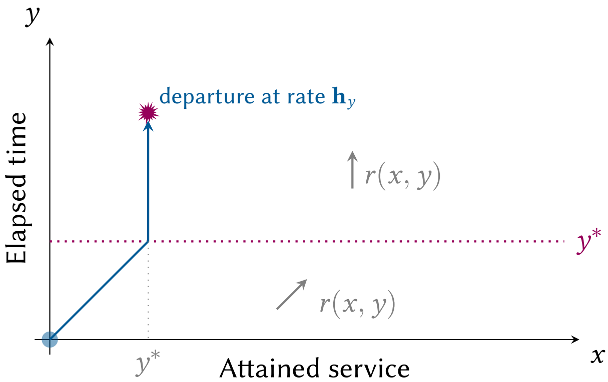
Attained service case



Solving the transport equation

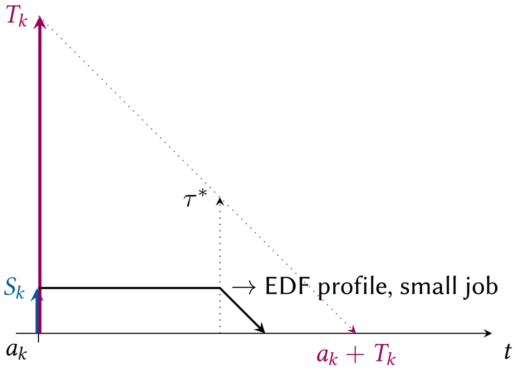
Attained service case



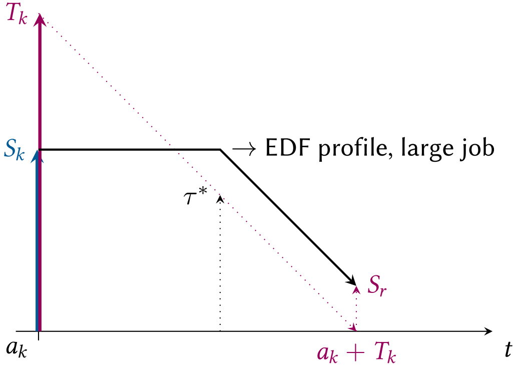


y
departure at rate $\mathbf{h}_y$
Elapsed time
$r(x, y)$
y^*
$r(x, y)$
y^*
Attained service
x

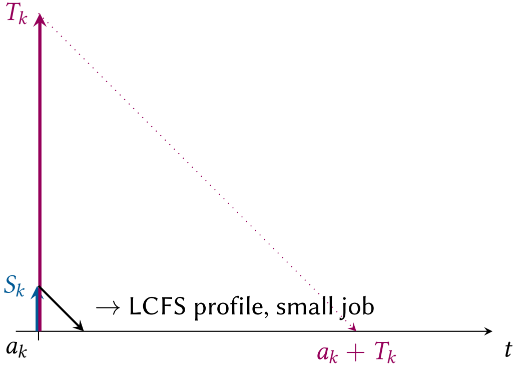
Perceived performance in EDF



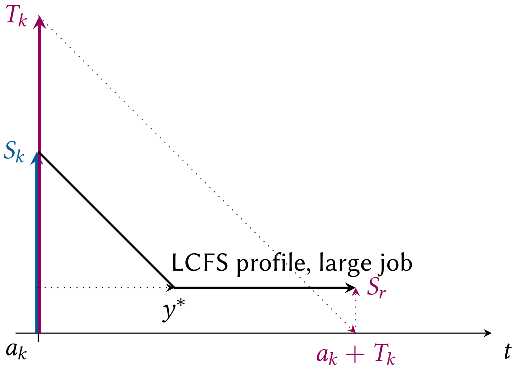
Perceived performance in EDF



Perceived performance in LAS and LCFS



Perceived performance in LAS and LCFS



Outline

Outline

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.