The last, the least and the urgent...

A story of three policies

Andres Ferragut

Joint work with Diego Goldsztajn and Fernando Paganini
Universidad ORT Uruguay

INRIA MATHNET Seminar – June 2025

Outline

Introduction

A crash course on measure valued processes

Partial service queues and Earliest-Deadline-First

Deadline-oblivious policies

Simulations

Final remarks

Outline

Introduction

A crash course on measure valued processes

Partial service queues and Earliest-Deadline-First

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].

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 $S \sim g(\sigma)$.

$$\begin{array}{ccc} \vdash & & \longrightarrow \\ 0 & \text{Remaining service} & \sigma \end{array}$$

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 $S \sim g(\sigma)$.

New mass arrives, rate $\lambda g(\sigma)d\sigma\, dt$

0   Remaining service σ

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 $S \sim g(\sigma)$.

Drifts to the left ar rate 1

0   Remaining service σ

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 $S \sim g(\sigma)$.

Departure

0

Remaining service

σ

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 $S \sim g(\sigma)$.

Steady-state

Remaining service

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 $S \sim g(\sigma)$.

Steady-state

Remaining service

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 $\sim \text{Poisson}(\rho)$ 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 $f(\sigma; t)$.

$f(\sigma; t)$

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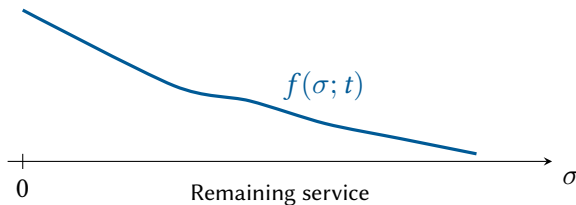
- Mass is transported to the left at rate 1.
- New mass arrives at σ with intensity $\lambda g(\sigma)\, d\sigma\, dt$.

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:

$$\frac{\partial f}{\partial \sigma} + \lambda g(\sigma) = 0 \implies f(\sigma) = \lambda \int_\sigma^\infty g(u) \, du = \lambda \bar{G}(\sigma),$$
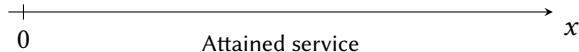
so the fluid approximation recovers the mean measure of Φ.

M/G/∞, fluid approximation.

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

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

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 ($\lambda \to \infty$).

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

$$\begin{array}{c} \vdash\!\!\longrightarrow\; x \\ 0 \qquad\quad \text{Attained service} \end{array}$$

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

New mass arrives, rate λdt

0   Attained service x

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

Drifts to the right ar rate 1

$\xrightarrow{\quad}$

●
0
Attained service
$\longrightarrow$ x

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

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

Steady-state

Attained service

x

0
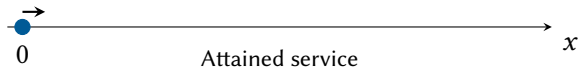
M/G/∞: take two
Attained service state descriptor

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

Steady-state

0   Attained service x
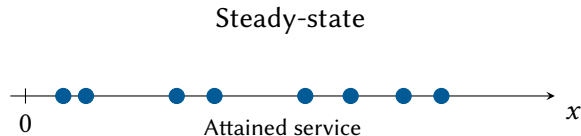
State-descriptor:

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

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

$\tilde{\Phi}_t$ 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:

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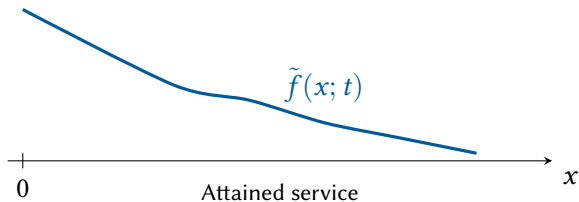
M/G/∞, take two

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.

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

$\tilde{f}(x; t)$

0   Attained service x

Suppose that we can replace $\tilde{\Phi}_t$ by a general measure ν_t with density $\tilde{f}(x; t)$.
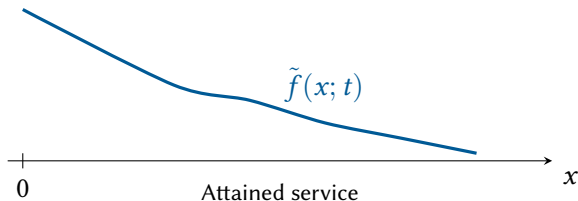
The corresponding transport equation is (informally):

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

M/G/∞: 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

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

- Tasks arrive as a Poisson process of intensity λ.

C servers

Queue

$\lambda \longrightarrow$

Partial service queues

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.

C servers

Queue

$\lambda \longrightarrow$

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.
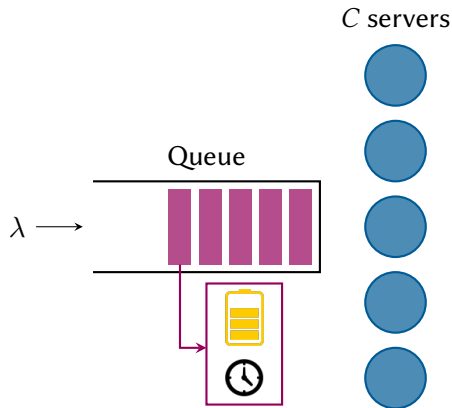
Partial service queue

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 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.

- Problem: we have to keep track of remaining service and deadlines simultaneously!

System load

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

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

System load

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

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

- 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...

- Consider the remaining time space.

New arrival at rate $\lambda g(\sigma, \tau) dt d\sigma d\tau$

- Consider the remaining time space.

Moves along the field $\mathbf{r} = \begin{pmatrix} -r(\sigma, \tau) \\ -1 \end{pmatrix}$

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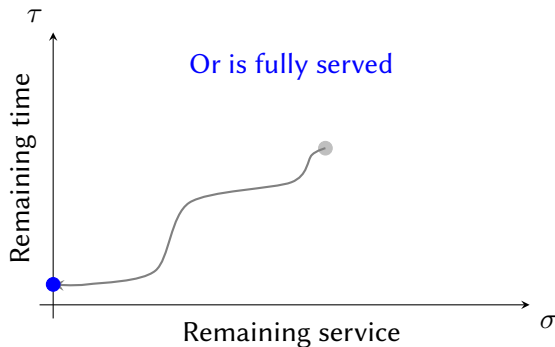- May depend on any combination of (σ, τ).
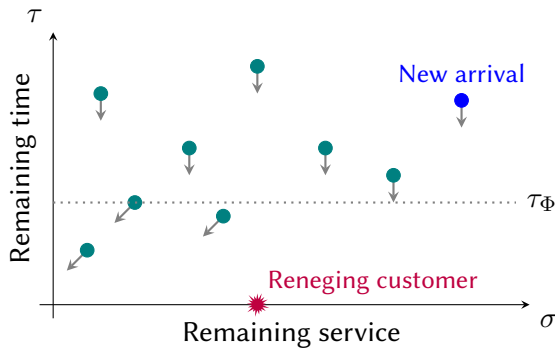
τ

Remaining time

$r(\sigma, \tau)$

1

Remaining service

σ

- 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

Weak formulation

We will describe these dynamics in terms of the projections

$$\langle \varphi, \mu \rangle := \iint \varphi(\sigma, \tau) \, \mu(d\sigma, d\tau)$$

of the state measure with respect to a test function $\varphi : \mathbb{R}_{++}^2 \to \mathbb{R}$, with continuous derivatives and compact support, i.e. $\varphi \in \mathcal{C}_c^1(\mathbb{R}_{++}^2)$.

We have:

$$\langle \varphi, \mu_{t+dt} \rangle = \iint \varphi(\sigma - r_{\mu_t} dt, \tau - dt) \, \mu_t(d\sigma, d\tau) + \lambda dt \iint \varphi(\sigma, \tau) g(\sigma, \tau) \, d\sigma d\tau + o(dt).$$

Fluid model dynamics

Weak formulation

$$\frac{\partial}{\partial t} \langle \varphi, \mu_t \rangle = \lim_{dt \to 0} \iint \frac{1}{dt} \left[\varphi(\sigma - r_{\mu_t} dt, \tau - dt) - \varphi(\sigma, \tau) \right] \mu_t(d\sigma, d\tau)$$

$$+ \lambda \iint \varphi(\sigma, \tau) g(\sigma, \tau) d\sigma d\tau$$

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

Equivalently:

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \int_0^t \left[- \iint \left[r_{\mu_s}(\sigma, \tau) \varphi_\sigma(\sigma, \tau) + \varphi_\tau(\sigma, \tau) \right] \mu_t(d\sigma, d\tau) \right.$$
$$\left. + \lambda \iint \varphi(\sigma, \tau) g(\sigma, \tau) \, d\sigma \, d\tau \right] ds,$$

for any $\varphi \in \mathcal{C}_c^1(\mathbb{R}_{++}^2)$.

Fluid model dynamics

Weak formulation

Equivalently:

$$\langle \varphi, \mu_t \rangle = \langle \varphi, \mu_0 \rangle + \int_0^t \left[-\iint \left[r_{\mu_s}(\sigma, \tau) \varphi_\sigma(\sigma, \tau) + \varphi_\tau(\sigma, \tau) \right] \mu_t(d\sigma, d\tau) \right.$$
$$\left. + \lambda \iint \varphi(\sigma, \tau) g(\sigma, \tau) \, d\sigma \, d\tau \right] ds,$$

for any $\varphi \in \mathcal{C}_c^1(\mathbb{R}^2_{++})$.

Looks daunting, but is not that bad...

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

EDF in overload

Fluid model equilibrium
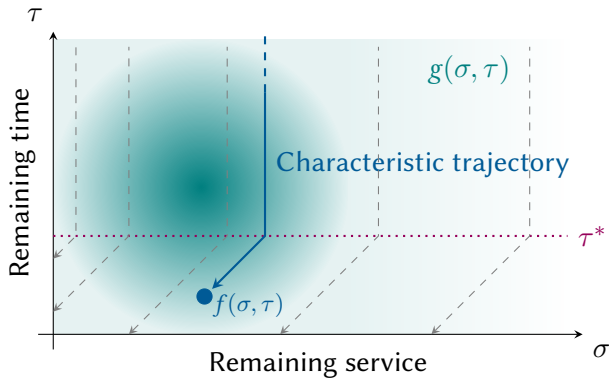
Theorem

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 $S_r < \sigma_0$.
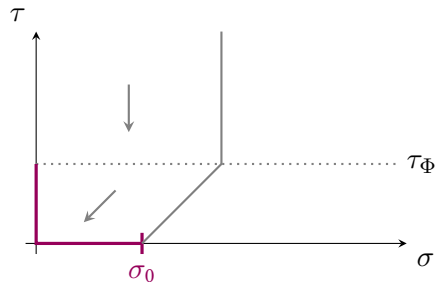
Proposition

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

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

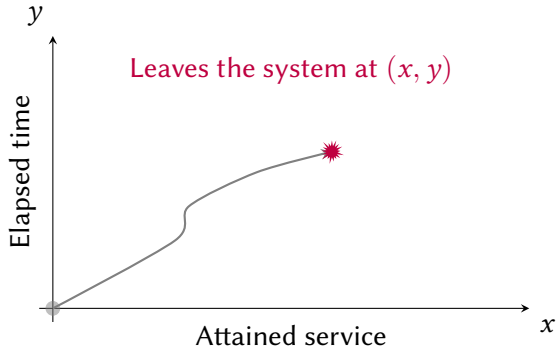- Consider the elapsed time space.

Attained service state descriptor
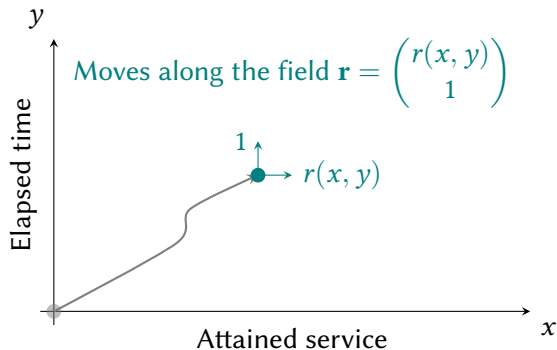
New arrival at rate λdt at $(0, 0)$

- Consider the elapsed time space.

y — Elapsed time

x — Attained service

Attained service state descriptor

Moves along the field $\mathbf{r} = \begin{pmatrix} r(x, y) \\ 1 \end{pmatrix}$

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

Leaves the system at (x, y)

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

Attained service state descriptor

Leaves the system at (x, y)

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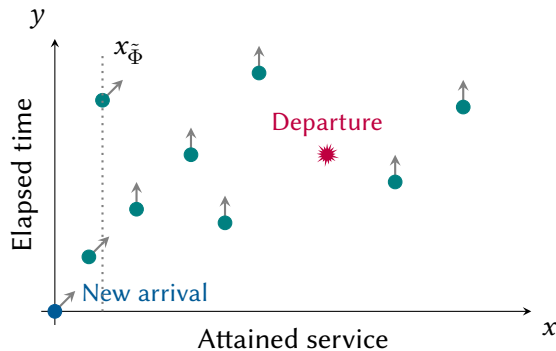
- Serve the C least-served tasks.
- Corresponds to taking:

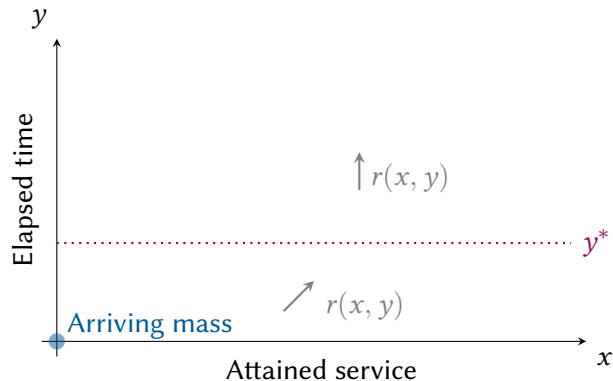$$r_{\tilde{\Phi}}(x, y) = \mathbf{1}_{\{x \leqslant x_{\tilde{\Phi}}\}}$$

with

$$x_{\tilde{\Phi}} := \sup\{x : \tilde{\Phi}([0, x] \times \mathbb{R}_+) \leqslant C\}.$$
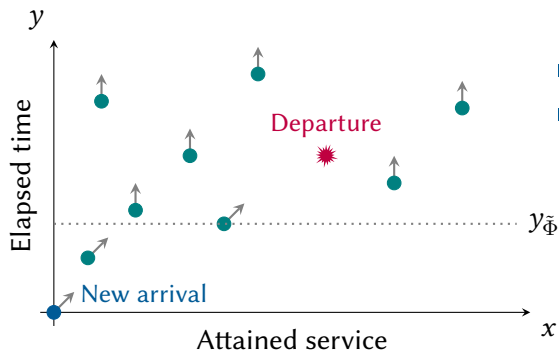
- Serve the C more recent tasks.
- Corresponds to taking:

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

with

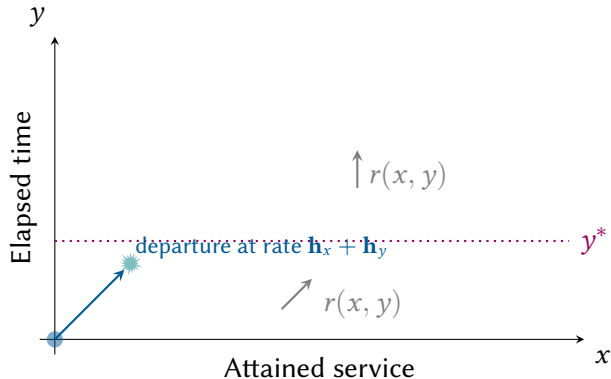$$y_{\tilde{\Phi}} := \sup\{y : \tilde{\Phi}(\mathbb{R}_+ \times [0, y]) \leqslant C\}$$

The figure shows axes labeled y (Elapsed time) and x (Attained service), with points marked, a "New arrival" at origin, a "Departure" marked, and a horizontal dotted line at $y_{\tilde{\Phi}}$.
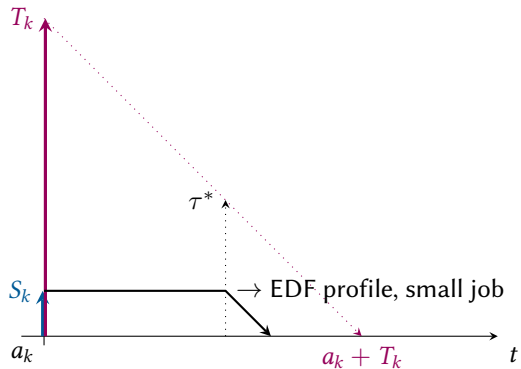
The hazard rate field

Solving the transport equation
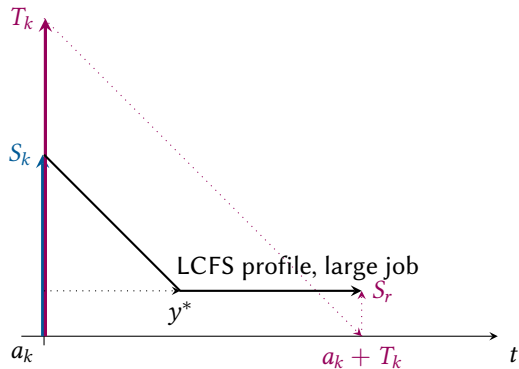
Attained service case
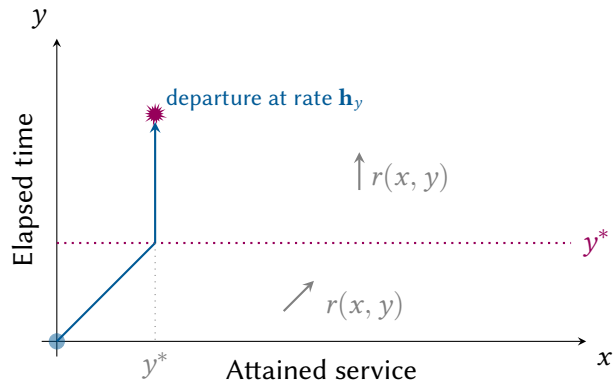
Perceived performance in EDF
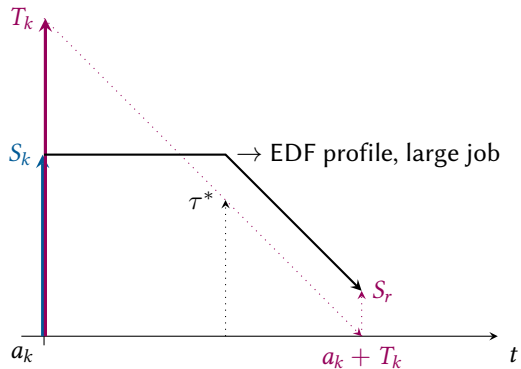
$\rightarrow$ EDF profile, small job

Perceived performance in EDF

T_k

S_k

$\rightarrow$ LCFS profile, small job

a_k

$a_k + T_k$

t

LCFS profile, large job

T_k

S_k

y^*

S_r

a_k

$a_k + T_k$

t

Outline

Outline

Messages from the talk

Future work

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.