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

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

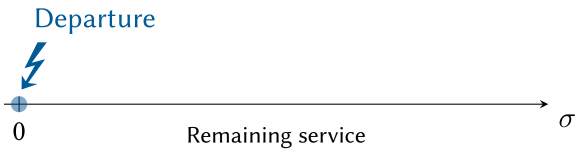
Consider the simple $M/G/\infty$ queue:

- 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 $S \sim g(\sigma)$.

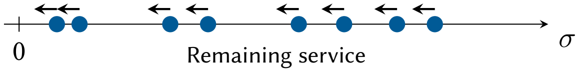


Measure valued stochastic processes

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

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

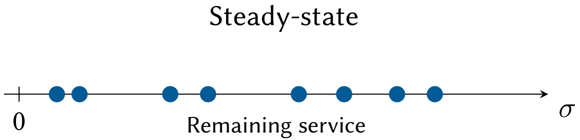
Steady-state



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:

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

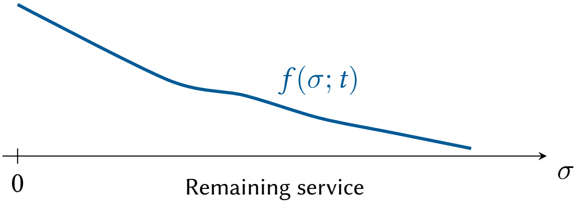
where $\bar{G}$ is the CCDF of S.

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.

$M/G/\infty$, 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:

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

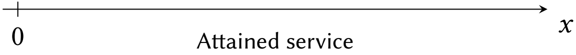
so the fluid approximation recovers the mean measure of Φ.

- 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

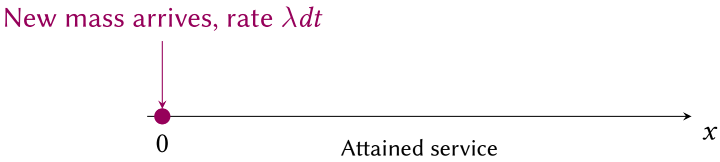
Here is another approach to model the same system [Kang and Ramanan, 2010]:



M/G/∞: take two

Attained service state descriptor

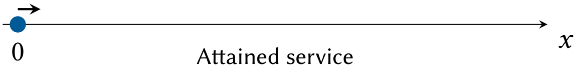
Here is another approach to model the same system [Kang and Ramanan, 2010]:



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

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

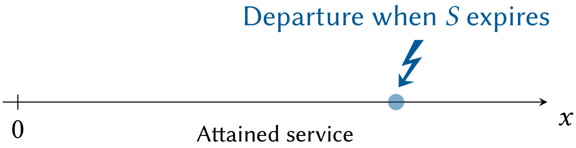
Drifts to the right ar rate 1



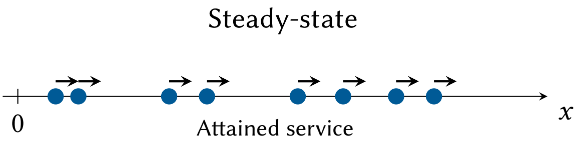
M/G/∞: take two

Attained service state descriptor

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



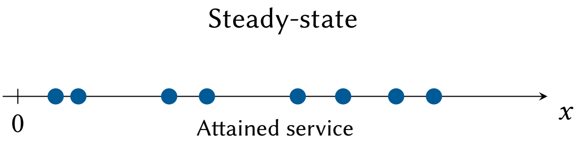
Here is another approach to model the same system [Kang and Ramanan, 2010]:



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:

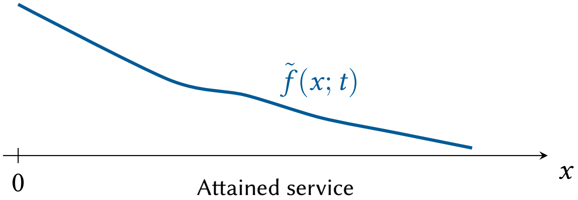
$$\tilde{\Phi} \sim \text{Poisson Process with mean measure } \nu(dx) = \lambda \bar{G}(x) dx$$

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.

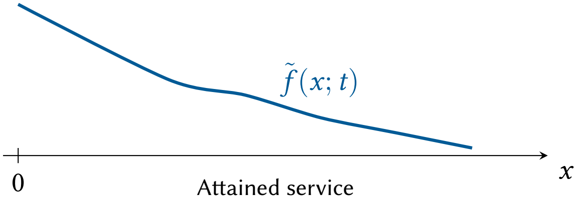
Suppose that we can replace $\tilde{\Phi}_t$ by a general measure ν_t with density $\tilde{f}(x;t)$.



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/∞: 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?

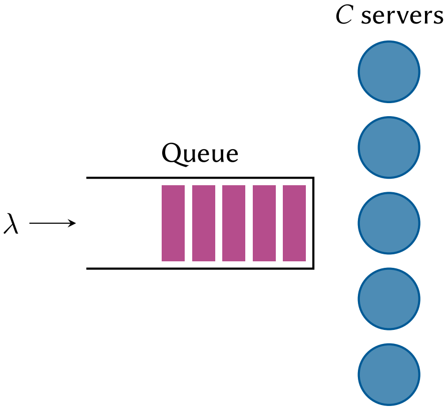
Outline

Partial service queues

Setting

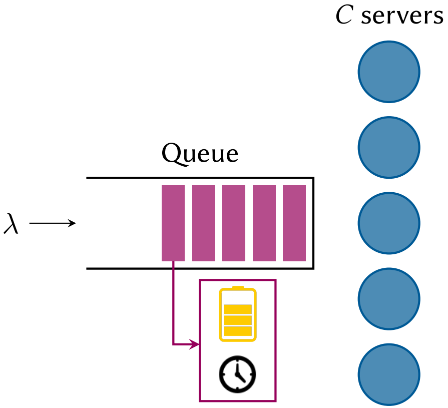
Consider an $M/G/C$ system where:

- Tasks arrive as a Poisson process of intensity λ.



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.

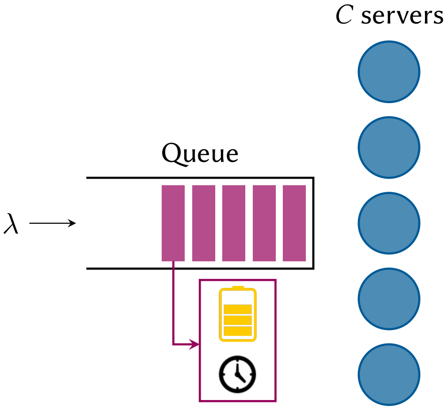


Partial service queues

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 laod:

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

System load

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

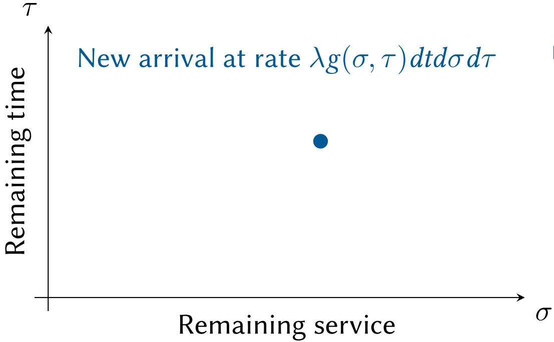
$$\rho := \lambda E[\min\{S, T\}].$$

- 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...

System evolution

Remaining service times

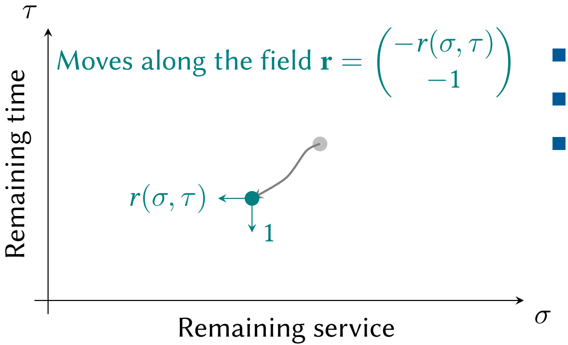
- Consider the remaining time space.



- Consider the remaining time space.

System evolution

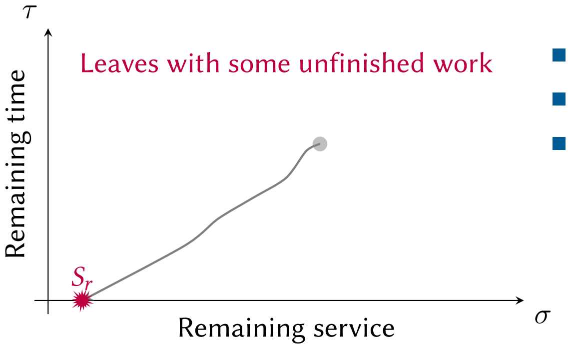
Remaining service times



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

System evolution

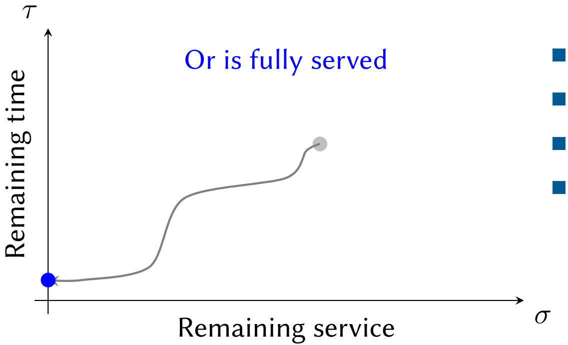
Remaining service times



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

System evolution

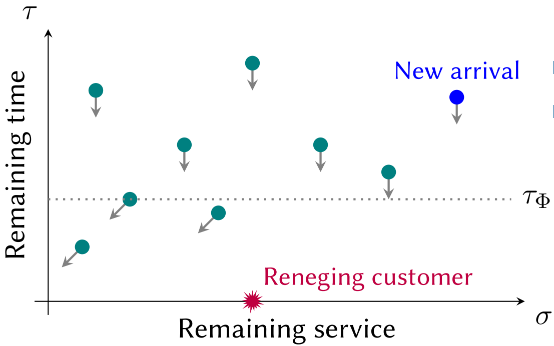
Remaining service times



- 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

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

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

$$
\begin{aligned}
\frac{\partial}{\partial t}\langle \varphi, \mu_t\rangle &= \lim_{dt\to 0}\iint \frac{1}{dt}\left[\varphi(\sigma - r_{\mu_t}dt, \tau - dt) - \varphi(\sigma,\tau)\right]\mu_t(d\sigma, d\tau) \\
&\quad + \lambda \iint \varphi(\sigma,\tau)g(\sigma,\tau)d\sigma\, d\tau \\
&= -\iint \left[r_{\mu_t}(\sigma,\tau)\varphi_\sigma(\sigma,\tau) + \varphi_\tau(\sigma,\tau)\right]\mu_t(d\sigma, d\tau) + \lambda \iint \varphi(\sigma,\tau)g(\sigma,\tau)d\sigma\, d\tau,
\end{aligned}
$$

Fluid model dynamics

Weak formulation

Equivalently:

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

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

Fluid model dynamics

Weak formulation

Equivalently:

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

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

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.

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.

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

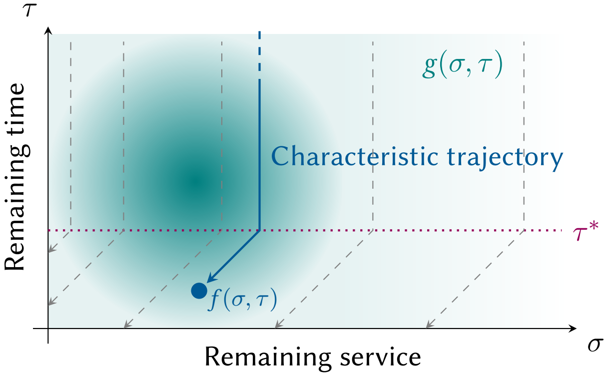
Imposing equilibrium we get:

- $\tau_{\mu^*} = \tau^*$ 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:

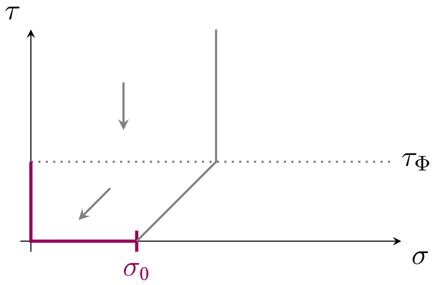
$$f(\sigma,\tau) = \lambda\left[\int_0^{(\tau^*-\tau)^+} g(\sigma+u,\tau+u)du + \int_{(\tau^*-\tau)^+}^{\infty} g\left(\sigma+(\tau^*-\tau)^+,\tau+u\right)du\right].$$

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

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

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!

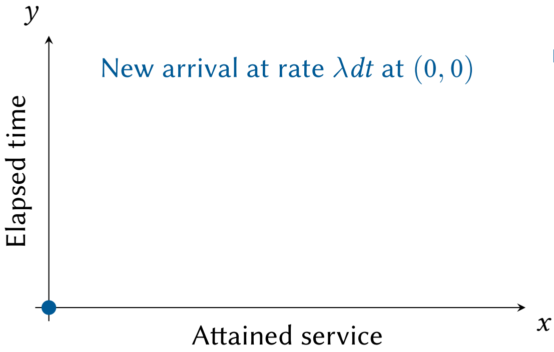
- What about deadline-oblivious policies?
  - Can we model them?
  - What is their performance?

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

Attained service state descriptor

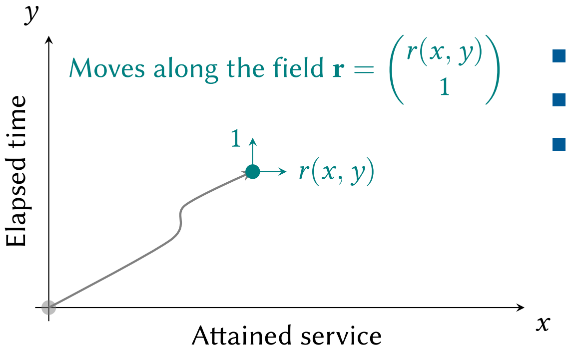
- Consider the elapsed time space.

Attained service state descriptor



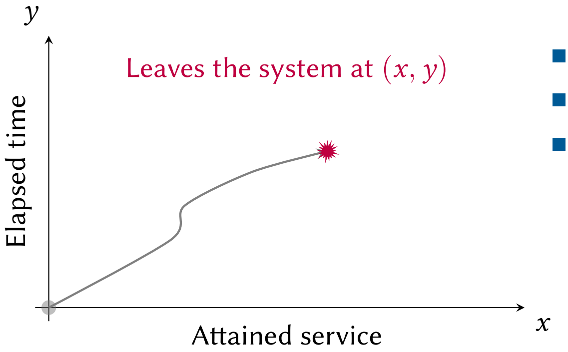
- Consider the elapsed time space.

Attained service state descriptor



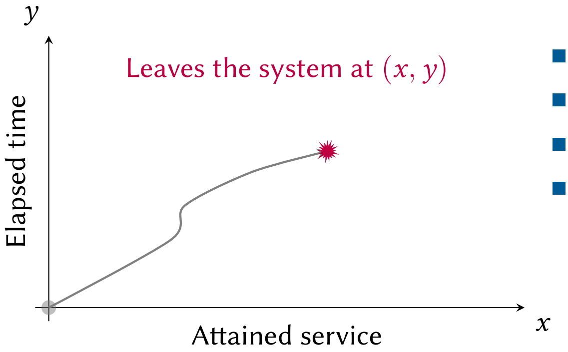
- Consider the elapsed time space.
- Policy again defines how tasks are served.
- May depend on any combination of (x, y).

Attained service state descriptor



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

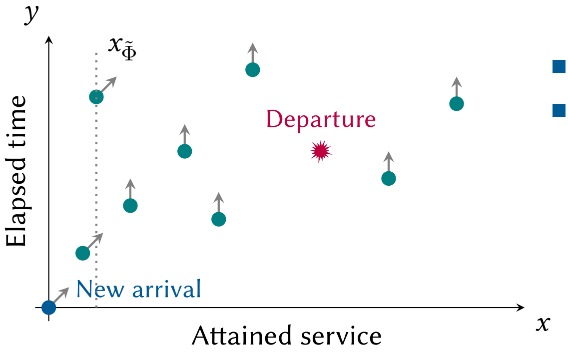
Attained service state descriptor



- 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



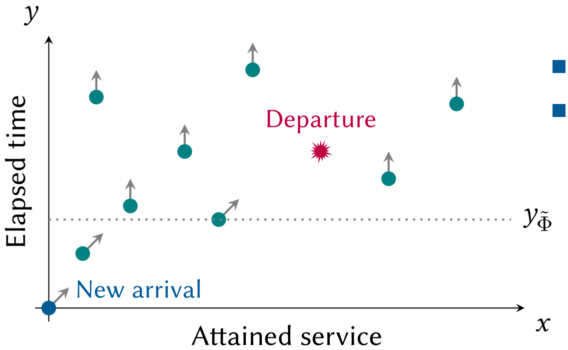
- Serve the C least-served tasks.
- Corresponds to taking:

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

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)

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

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

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:

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

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

Departure rate

Now we have to compute the departure rate $\eta_n u(x, y)$:

$$\eta_\nu(x, y) := \lim_{dt \to 0} \frac{1}{dt} P\left(\left\{S \in \left(x, x + r_{\tilde{\Phi}} dt\right)\right\} \cup \{T \in (y, y + dt)\} \mid S > x, T > y\right)$$

By the chain rule and some computations:

$$\eta_\nu(x, y) = \frac{1}{\bar{G}(x, y)} \left[-\frac{\partial}{\partial x} \bar{G}(x, y) r_{\tilde{\Phi}}(x, y) - \frac{\partial}{\partial y} \bar{G}(x, y) \right]$$

Departure rate

Now we have to compute the departure rate $\eta_n u(x,y)$:

$$\eta_\nu(x,y) := \lim_{dt\to 0} \frac{1}{dt} P\left(\left\{S \in \left(x, x + r_{\tilde{\Phi}} dt\right)\right\} \cup \left\{T \in (y, y+dt)\right\} \mid S > x, T > y\right)$$

By the chain rule and some computations:

$$\eta_\nu(x,y) = \frac{1}{\bar{G}(x,y)} \left[-\frac{\partial}{\partial x}\bar{G}(x,y) r_{\tilde{\Phi}}(x,y) - \frac{\partial}{\partial y}\bar{G}(x,y) \right]$$

Therefore:

$$\eta_\nu(x,y) = h^x(x,y) r_\nu(x,y) + h^y(x,y) = \mathbf{r}_\nu(x,y) \cdot \mathbf{h}(x,y).$$

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):

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

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

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

Recall that LCFS can be modeled by:

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

with

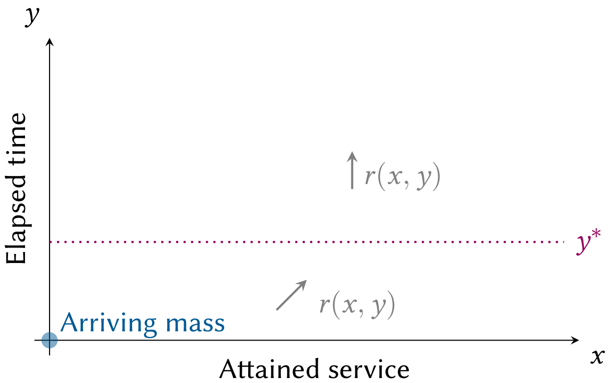
$$y^* = \sup\left\{y \geq 0 : \nu^*(\mathbb{R}_+ \times [0,y]) \leqslant C\right\}.$$

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

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

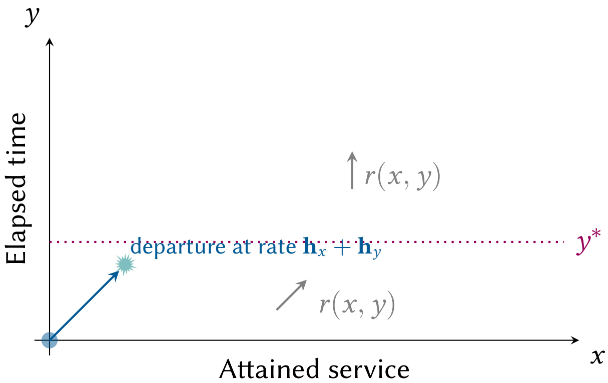
Solving the transport equation

Last come first served case



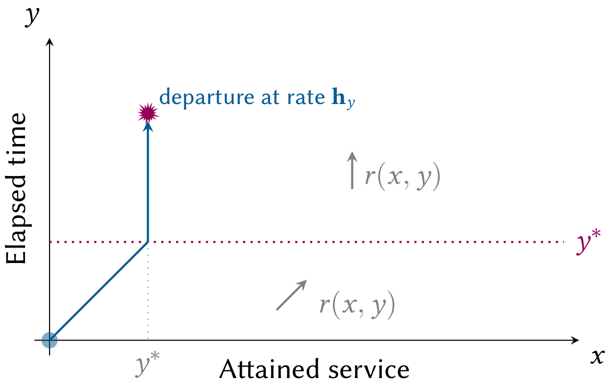
Solving the transport equation

Last come first served case



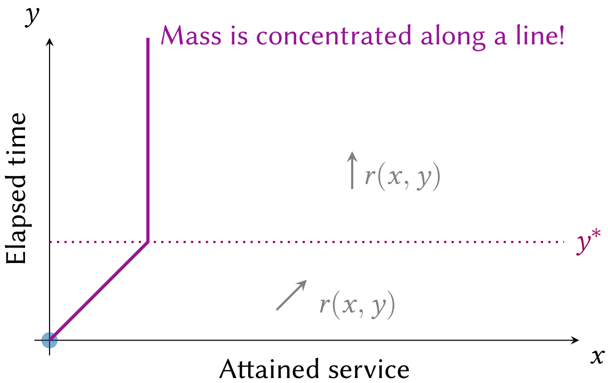
Solving the transport equation

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:

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

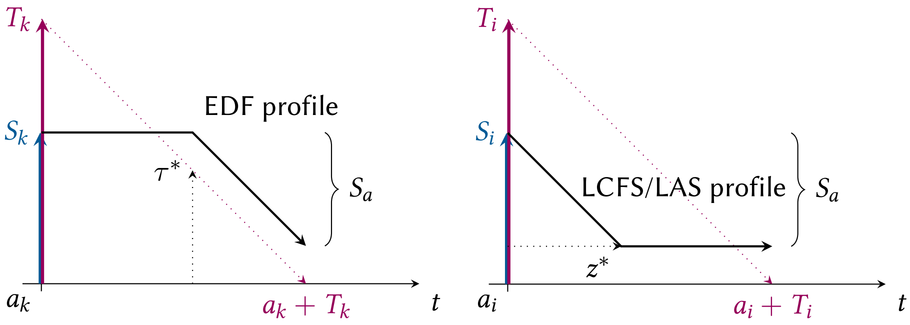
Proposition

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!!!

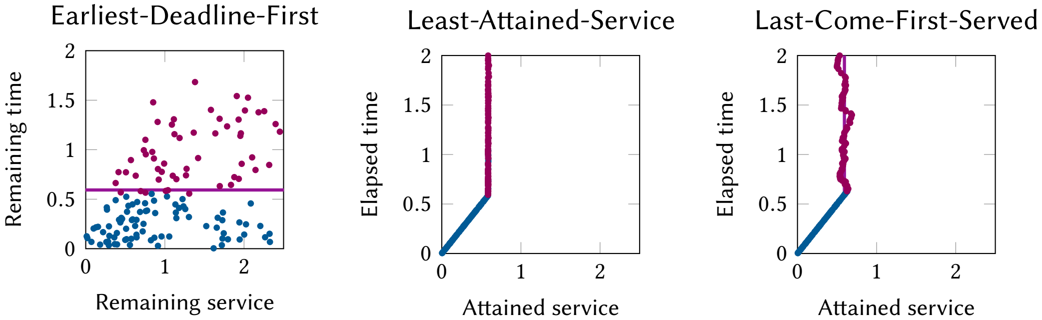
Outline

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

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

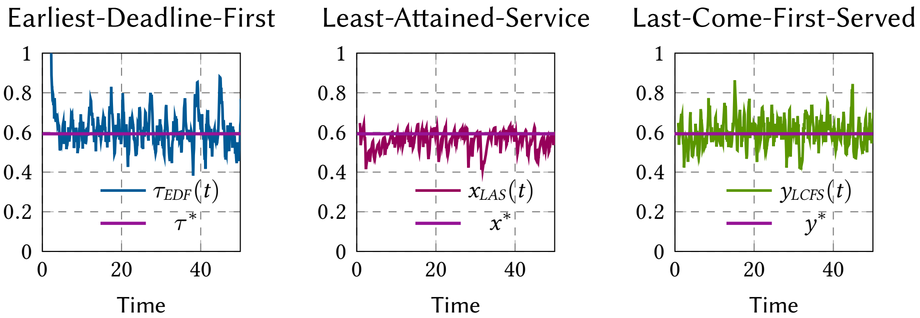
- In particular, the random variables U and V are correlated with normal distributions, and therefore S and T are correlated with log-normal distributions.
- In this case, $E[\min\{S, T\}] \approx 1.37$ can only be numerically estimated.
- We choose $\lambda = 200$ and $C = 100$, then $z^* \approx 0.593$.

State space snapshots

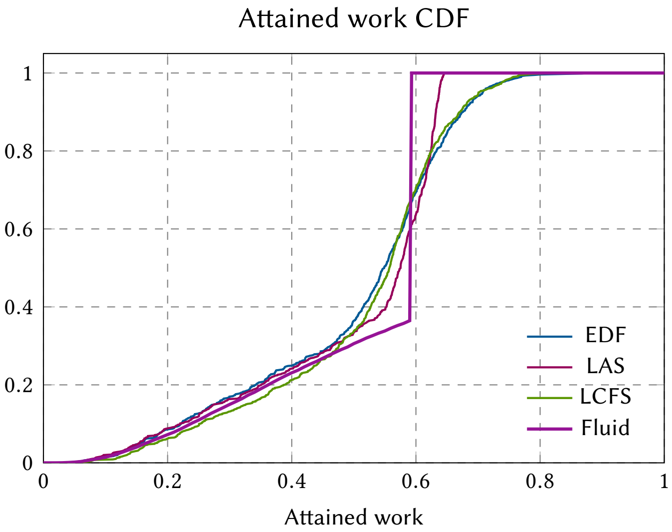


Blue dots are in service, red dots are not in service.

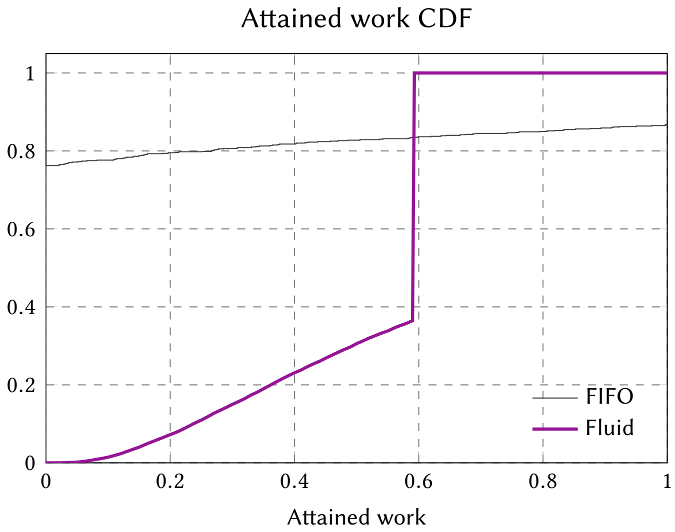
Stochastic threshold evolution



Attained work empirical CDF



Comparison with FIFO



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 (long work...help needed...)
- 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.

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.