When an Ornstein–Uhlenbeck Process Meets a Decreasing Random Boundary

A diffusion approximation for coldstart hitting times

Jonatha Anselmi
Inria – Ghost

University of Padua – Department of Mathematics

Padova – October 2026

The mathematical question

A coupled first-passage problem

Let X be an Ornstein–Uhlenbeck process and W a *decreasing random boundary* driven by X. Estimate

$$\mathbb{E}[\tau] \quad \text{with} \quad \tau = \inf\{t \geq 0 : X_t = W_t\}.$$

- The boundary is random but *not an independent diffusion*.
- The gap is Gaussian, yet non-Markovian and nondifferentiable.
- We will obtain an explicit, accurate approximation and a rigorous logarithmic asymptotic.

Related literature. The OU first-passage problem has been studied for a *fixed boundary*, e.g., [Alili et al., 2005; Lipton and Kaushansky 2018; Martin et al. 2019; Zeng 2020], and for a *deterministic exponentially decaying boundary*, via PDE methods [Tuckwell and Wan, 1984]. Here, W **is instead *random and coupled to* X.**

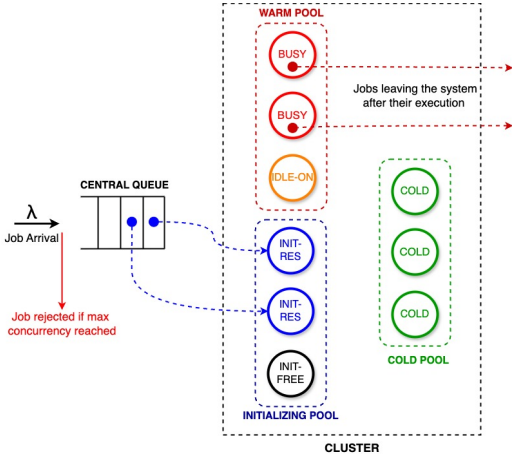
Road map

1. Motivation
2. Queueing model and diffusion convergence
3. Diffusion model and analysis

Road map

1. Motivation
2. Queueing model and diffusion convergence
3. Diffusion model and analysis

Autoscaling in one picture



A warm server

Ready to process an arriving job without startup delay.

A coldstart

An arrival finds *no idle warm server*; a new instance must be initialized.

Main question

Given the currently available spare capacity, when will the next coldstart occur?

Two autoscaling rules, one hitting time

1. On arrival, use an idle warm server; if none is available, trigger scale-up (a coldstart).
2. An idle warm server is switched off after an exponential idle timer of rate γ.

Pre-coldstart dynamics

$X_t^{(n)}$: number of **busy** servers; $W_t^{(n)}$: number of **warm** servers.

$$I_t^{(n)} = W_t^{(n)} - X_t^{(n)} \geq 0, \qquad \tau^{(n)} = \inf\{t \geq 0 : I_t^{(n)} = 0\}.$$

We study time *until* a coldstart, not the initialization latency incurred *after* one.



Supermarket checkout lines

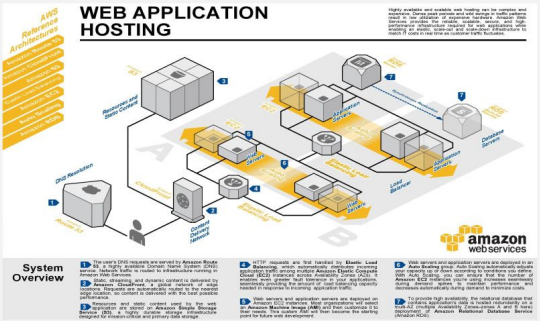


Call centers



Data centers

In Europe, data centers account for 2% of total electricity consumption [2025–2026 data, sources: IEA/EUDCA].



Road map

1 Motivation

2 Queueing model and diffusion convergence

3 Diffusion model and analysis

Finite-system stochastic model

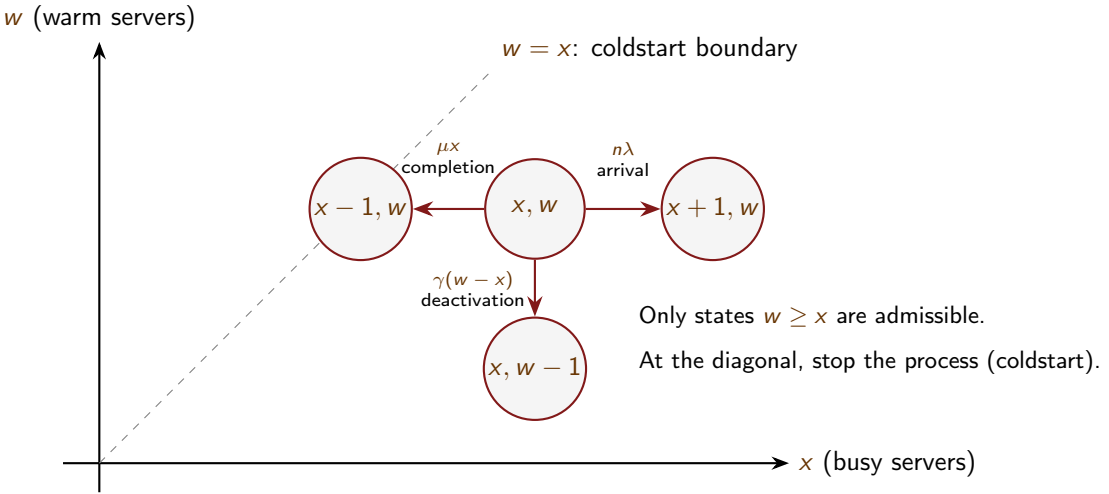
- System size n; renewal arrivals at rate $n\lambda$.
- Independent exponential service times: rate μ per busy server.
- Independent exponential idle timers: rate γ per idle warm server.

Before the first coldstart

Event	State change $(X^{(n)}, W^{(n)})$	Rate (Poisson-arrival case)
Arrival, if $X < W$	$(x, w) \to (x+1, w)$	$n\lambda$
Service completion	$(x, w) \to (x-1, w)$	μx
Idle timer expiry	$(x, w) \to (x, w-1)$	$\gamma(w-x)$

For general renewal arrivals, $(X^{(n)}, W^{(n)})$ need not itself be Markov: augment with arrival-age information. The displayed 2D CTMC applies to Poisson arrivals.

Two-dimensional jump geometry (Poisson arrivals)



Why fluid scaling is not enough

Set $\rho = \lambda/\mu$ and start with $X_0^{(n)} \approx n\rho$.

Fluid-scale picture

With $X^{(n)}/n \to x$, $W^{(n)}/n \to w$, and $x(0) = \rho$, the pre-coldstart equations are

$$x(t) = \rho, \qquad \dot{w}(t) = -\gamma(w(t) - \rho).$$

If $w(0) > \rho$, then $w(t) - x(t) = (w(0) - \rho)e^{-\gamma t} > 0$ at every finite t.

Consequence

The deterministic first-order description misses the first coldstart: fluctuations are essential.

Take an initial spare-capacity pool of order $\sqrt{n}$ and retain diffusion-scale fluctuations.

Diffusion scaling and limiting equations

$$\widehat{X}_t^{(n)} = \frac{X_t^{(n)} - n\rho}{\sqrt{n}}, \qquad \widehat{W}_t^{(n)} = \frac{W_t^{(n)} - n\rho}{\sqrt{n}}.$$

Candidate limit, before hitting

$$\begin{aligned} dX_t &= -\mu X_t\, dt + \sqrt{\theta}\, dB_t, \\ dW_t &= -\gamma(W_t - X_t)\, dt, \end{aligned} \qquad \theta = \lambda + \lambda^3 \nu^2,$$

where ν^2 is the limiting variance of the n-rescaled interarrival times.

- X: Ornstein–Uhlenbeck fluctuations of busy servers.
- W: random finite-variation boundary; no independent Brownian term.
- $W - X$ remains positive until the first hit.

Convergence of first-passage times

Theorem 1 [A. 2026]

Under the renewal-input assumptions, convergence of diffusion-scaled initial states, and uniformly bounded second moments,

$$\tau^{(n)} \Rightarrow \tau, \qquad \tau = \inf\{t \geq 0 : W_t = X_t\},$$

and the stopped pairs converge weakly in the Skorokhod space.

1. Known infinite-server diffusion limit gives $\widehat{X}^{(n)} \Rightarrow X$ [Whitt 1982].
2. Doob–Meyer: timer-expiration martingale vanishes after $\sqrt{n}$ normalization.
3. Gronwall/continuous mapping identifies the coupled limit for W.
4. Boundary regularity yields convergence of stopping times.

Proof sketch: why the timer noise disappears

Before $\tau^{(n)}$, write the scale-down equation

$$dW_t^{(n)} = -\gamma(W_t^{(n)} - X_t^{(n)})\,dt - dM_t^{(n)}.$$

The predictable quadratic variation is

$$\langle M^{(n)} \rangle_T = \gamma \int_0^T (W_s^{(n)} - X_s^{(n)})\,ds.$$

Key scale separation

The spare capacity is $O(\sqrt{n})$, so $\mathbb{E}\langle M^{(n)} \rangle_T = O(\sqrt{n})$; Doob's inequality gives

$$\sup_{t \le T} \frac{|M_t^{(n)}|}{\sqrt{n}} \xrightarrow{\mathbb{P}} 0.$$

Therefore, W inherits randomness from X, *not* from its own martingale.

Road map

1 Motivation

2 Queueing model and diffusion convergence

3 Diffusion model and analysis

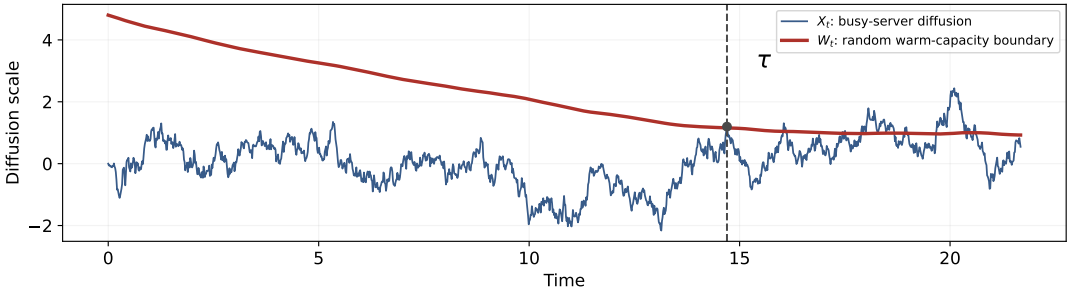
OU process and an endogenous moving boundary

OU process

$$dX_t = -\mu X_t dt + \sqrt{\theta}\, dB_t.$$

Random boundary

$$W_t = e^{-\gamma t} w_0 + \gamma \int_0^t e^{-\gamma(t-s)} X_s \, ds.$$

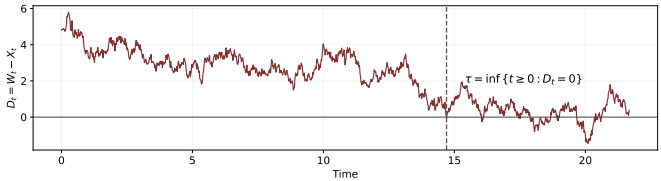


Illustrative simulated trajectory; the vertical line denotes the first meeting.

Equivalent Gaussian first-passage problem

Let $D_t = W_t - X_t$, $D_0 = d > 0$. Then

$$\begin{aligned} dX_t &= -\mu X_t dt + \sqrt{\theta}\, dB_t, \\ dD_t &= (\mu X_t - \gamma D_t)dt - \sqrt{\theta}\, dB_t, \\ \tau &= \inf\{t \geq 0 : D_t = 0\}. \end{aligned}$$



The joint pair (X, D) is Markov; its scalar coordinate D is generally not.

Explicit Gaussian representation

Assume $X_0 = 0$ and $\mu \neq \gamma$ (the equal-rate case follows by continuity). Then

$$D_t = de^{-\gamma t} + \frac{\sqrt{\theta}}{\mu - \gamma} \int_0^t \left(\gamma e^{-\gamma(t-s)} - \mu e^{-\mu(t-s)} \right) dB_s.$$

- $m_D(t) = de^{-\gamma t}$: an exponentially decaying mean.
- The Gaussian fluctuation has uniformly bounded variance:

$$\mathrm{Var}(D_t) \longrightarrow \frac{\theta}{2(\mu + \gamma)} \quad (t \to \infty).$$

- Continuous paths, but no mean-square derivative; infinitely many local crossings of a level after a hit.

Challenge

Classical smooth-Gaussian Kac–Rice formulas do *not* count the crossings of D.

The desired object: a mean first-passage time

Mathematical question

$$\mathbb{E}_d[\tau], \qquad \tau = \inf\{t \geq 0 : D_t \leq 0\}, \qquad D_0 = d > 0.$$

Even though D_t is Gaussian, does not enjoy the usual assumptions, e.g., differentiability.

Our route

Construct a *smoothed, Rice-inspired crossing proxy*:

- approximate the *expected* first-passage time by the point in time where the *expected* number of downcrossings of the barrier at level zero reaches one
- adapt the Kac–Rice formula

Then,

- solve the resulting equation explicitly
- prove an error bound and the logarithmic scaling

Why the classical Rice argument breaks

For a *smooth* Gaussian process Z, the Kac–Rice crossing formula reads

$$\mathbb{E}[N([0,T],u)] = \int_0^T \mathbb{E}[|\dot{Z}_t| \mid Z_t = u] f_{Z_t}(u)\, dt.$$

For our D,

- $\dot{D}_t$ does not exist as a pathwise derivative;
- the raw number of zero crossings is not finite;
- replacing first passage by the time of the first *expected* crossing is itself heuristic.

Important distinction

What follows is **not** a rigorous application of Kac–Rice: it is a proposal for an effective, smoothed flux.

Core idea: replace crossing speed by conditional drift

The finite-variation part of the gap has drift $b(X_t, D_t) = \mu X_t - \gamma D_t$.

Conditional infinitesimal drift at the boundary

With $f(z) = z$, Dynkin's formula gives

$$\mathcal{L}_t f(0) := \lim_{h \downarrow 0} \frac{\mathbb{E}[D_{t+h} - D_t \mid D_t = 0]}{h} = \mu \mathbb{E}[X_t \mid D_t = 0].$$

By Gaussian regression (when $X_0 = 0$),

$$\mathcal{L}_t f(0) = -\mu \, de^{-\gamma t} \frac{\mathrm{Cov}(X_t, D_t)}{\mathrm{Var}(D_t)}.$$

For large t, $\mathcal{L}_t f(0) \sim \mu d e^{-\gamma t}$.

Interpretation

A surrogate *mean velocity*, not a pathwise crossing intensity.

From a surrogate flux to an explicit equation

Set $w(t) = de^{-\gamma t}$ and $\sigma^2 = \theta/[2(\mu+\gamma)]$. For times at which the variance has approximately equilibrated,

$$f_{D_t}(0) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{w(t)^2}{2\sigma^2}\right).$$

Using $|\mathcal{L}_t f(0)| \approx \mu w(t)$, propose

$$N_{\mathrm{heu}}(T) = \int_0^T \frac{\mu w(t)}{\sqrt{\pi\theta/(\mu+\gamma)}} \exp\left(-\frac{w(t)^2}{\theta/(\mu+\gamma)}\right) dt.$$

Rice-inspired closure (heuristic)

$$\boxed{N_{\mathrm{heu}}(T^*) = \tfrac{1}{2}.}$$

The factor $1/2$ converts a two-sided crossing proxy to a downcrossing proxy.

Closed-form mean hitting-time approximation

Let $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-u^2}\, du$ and

$$a = d\sqrt{\frac{\mu+\gamma}{\theta}}, \qquad z = \operatorname{erfc}^{-1}\left(\frac{\gamma}{\mu} + \operatorname{erfc}(a)\right).$$

Proposition 2 (explicit solution)

If $0 < \gamma/\mu + \operatorname{erfc}(a) < 1$, the heuristic equation admits a unique finite solution

$$\boxed{T^*(\gamma, d) = \frac{1}{\gamma}\left[\log d + \frac{1}{2}\log\frac{\mu+\gamma}{\theta} - \log z\right].}$$

- Decreasing γ (slower shutdown) strongly postpones the coldstart.
- Increasing initial spare capacity d gives primarily logarithmic gains.

Derivation: a one-line change of variables

Write $a = d\sqrt{(\mu+\gamma)/\theta}$ and $y(t) = ae^{-\gamma t}$.

$$N_{\text{heu}}(T) = \frac{\mu}{\gamma\sqrt{\pi}} \int_{ae^{-\gamma T}}^{a} e^{-y^2}\, dy$$
$$= \frac{\mu}{2\gamma} \left[\operatorname{erfc}(ae^{-\gamma T}) - \operatorname{erfc}(a)\right].$$

Imposing $N_{\text{heu}}(T^*) = \frac{1}{2}$ yields

$$\operatorname{erfc}(ae^{-\gamma T^*}) = \frac{\gamma}{\mu} + \operatorname{erfc}(a).$$

Why a closed form exists

Exponential decay of the boundary's mean and a Gaussian flux combine into a complementary error function.

Error bound behind the heuristic

Theorem 2 [A. 2026]

For fixed $\mu, \gamma, \theta > 0$, $X_0 = 0$, and $d \to \infty$,

$$\mathbb{E}_d[\tau] = T^*(\gamma, d) + O(1) = \frac{\log d}{\gamma} + O(1).$$

The bounded additive term is independent of the *initial gap* d (but may depend on μ, γ, θ).

- This is a **leading-order / bounded-error result**, not a small absolute-error guarantee.
- The relative error is $O(1/\log d)$ as $d \to \infty$ at fixed parameters.
- The proof uses contraction and Gaussian concentration, *not* a rigorous Rice formula.

Proof sketch I: center time and control the upper tail

Decompose

$$D_t = de^{-\gamma t} + Z_t, \qquad \sup_{t\geq 0} \mathbb{E} Z_t^2 < \infty; \qquad t_d := \frac{\log d}{\gamma}.$$

At time t_d, $D_{t_d} = 1 + Z_{t_d}$ has uniformly bounded second moment.

Upper bound by Lyapunov contraction

For the two-dimensional Markov process (X, D), take

$$V(x, d) = x^2 + cd^2, \qquad \mathcal{L}V \leq -\kappa V + b.$$

The drift bound plus a compact-set hitting/return argument gives

$$\mathbb{E}_{x,d}\tau \leq C(1 + V(x, d)).$$

Consequently $\sup_{d\geq 1} \mathbb{E}[(\tau - t_d)^+] < \infty$.

$$\therefore \quad \mathbb{E}\tau \leq t_d + O(1).$$

Proof sketch II: early hits require Gaussian extremes

An early crossing at $t \le t_d - u$ requires

$$Z_t \le -de^{-\gamma t} \le -e^{\gamma u}.$$

Borell–TIS on backward unit-time blocks

Uniform variance and expected-supremum estimates give, for large j,

$$\mathbb{P}\left(\inf_{t \in [t_d - j - 1, t_d - j]} Z_t \le -e^{\gamma j} \right) \le 2\exp(-ce^{2\gamma j}).$$

A union bound over blocks, followed by integration over u, yields

$$\sup_{d \ge 1} \mathbb{E}[(t_d - \tau)^+] < \infty.$$

$$\therefore \quad \mathbb{E}\tau \ge t_d - O(1), \quad \text{completing the logarithmic law.}$$

What the theorem does *not* say

- It does not control the additive constant uniformly as $\gamma \to 0$.
- It does not prove that T^* has a uniformly small *absolute* error.
- It does not follow from high-barrier asymptotics: the effective crossing level

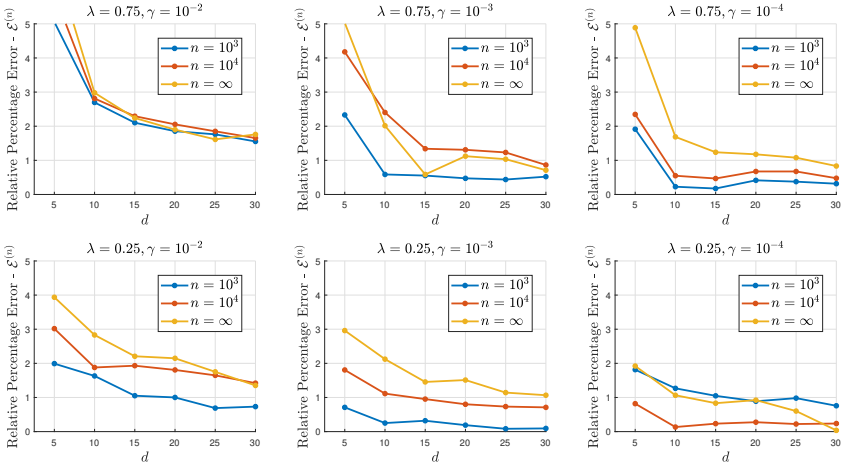
$$w(T^*) = z\sqrt{\frac{\theta}{\mu+\gamma}}$$

stays bounded as $d \to \infty$.

Separation of evidence

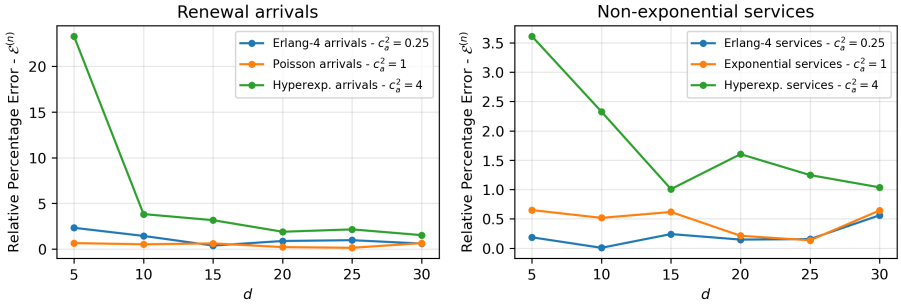
Rigorous: logarithmic growth and bounded additive discrepancy.
Empirical: strikingly high accuracy in the autoscaling parameter range.

Numerical validation I: approximation accuracy



Relative error $\mathcal{E}^{(n)} = 100|T^* - \tau_{\text{sim}}^{(n)}|/\tau_{\text{sim}}^{(n)}$. Accuracy generally improves for larger d and smaller γ; $n = 10^3, 10^4$ are comparable to the diffusion benchmark $n = \infty$.

Numerical validation II: robustness



Renewal arrivals

Squared coefficient of variation $c_a^2 \in \{0.25, 1, 4\}$. The approximation remains accurate. The diffusion limit of the busy-server process remains a scalar OU process, with adjusted variance parameter $\theta = \lambda(1 + c_a^2)$.

Non-exponential service times

Squared coefficient of variation $c_s^2 \in \{0.25, 1, 4\}$. The approximation remains accurate, even though the diffusion limit of the busy-server process is no longer a scalar OU process in this case.

Interpretation: spare capacity versus shutdown policy

$$T^*(\gamma, d) = \frac{\log d}{\gamma} + O_{\mu,\gamma,\theta}(1), \qquad d \to \infty.$$

Qualitative conclusions

1. Adding more initially idle warm servers yields *diminishing logarithmic returns*.
2. Increasing idle-server lifetime ($\gamma \downarrow$) can have a much stronger effect.
3. The formula offers a state-dependent proxy for triggering proactive scale-up.

The precise engineering optimum requires energy, delay, and capacity costs not incorporated in this first-passage analysis.

Conclusions and open mathematical directions

Main takeaway

An applied coldstart problem leads to an *OU process hitting an endogenous, decreasing random boundary*. A Rice-inspired smoothed flux yields a closed-form mean approximation, supported by a logarithmic-growth theorem.

- Can the $O(1)$ discrepancy be sharpened, especially as $\gamma \downarrow 0$?
- Thresholds $D_t = b > 0$ and deterministic idle timers.
- Higher-dimensional OU limits for general service times.
- First passage with finite-capacity or heterogeneous-server constraints.

Thank you!

Questions and discussion

Jonatha Anselmi
jonatha.anselmi@inria.fr

Based on
J. Anselmi *A Diffusion Approximation for the Coldstart Hitting Time in Autoscaling Systems*, IEEE/ACM Transactions on Networking, 2026