A Diffusion Approximation for the Coldstart Hitting Time in Autoscaling Systems

Jonatha Anselmi

***Abstract*—Existing cloud computing systems rely on auto-scaling technologies to dynamically adjust service capacity in response to workload fluctuations. Common scale-up rules activate new server instances whenever an incoming job finds no idle server available for immediate processing. Such jobs experience a *coldstart*, which slow them down. Our objective is to investigate the *coldstart hitting time*, i.e., the time until the next coldstart occurs, as a function of the current amount of spare (idle-on) servers. This is motivated by the fact that most autoscaling policies are highly reactive and adjust capacity basing on instantaneous measurements of the current system state.**

We develop a queueing model to define the coldstart hitting time and, under a diffusion scaling, show that it converges to the first-passage time (FPT) of a Gaussian process to zero. This connects our problem to a well-studied class of FPT problems, although the autoscaling setting makes the underlying Gaussian process fail to satisfy the classical assumptions. We then develop a heuristic for the expected FPT involving the Kac-Rice formula to obtain a simple and interpretable approximation. The key point of our heuristic is to replace the conditional upcrossing rate, which is not well defined in our case, by the conditional infinitesimal drift of the process. We prove that the proposed heuristic coincides with the exact expected FPT up to a constant term that does not depend on the fluctations of the idle server process. In the parameter regime of autoscaling, numerical simulations show that our approximation is remarkably accurate.

***Index Terms*—Auto-scaling, coldstarts, diffusion approximation, first-passage time, Gaussian process.**

I. Introduction

In cloud and serverless computing, performance and responsiveness strongly depend on how effectively the system can dynamically adjust its capacity to meet the fluctuating demand. This service elasticity is made possible by *autoscaling* technologies. In this context, a major problem is the occurrence of *coldstarts* [27], i.e., events in which unfortunate requests (jobs) find no ready-to-use (*idle-on*) servers upon arrival. In these cases, the jobs themselves trigger the activation of new server functions or virtual machines for their processing, and this increases user-perceived delay [31]. Frequent coldstarts can severely degrade user experience and, in latency-sensitive applications, translate into financial losses. It is well known that even a few extra milliseconds of latency can have a dramatic impact on user experience. Consequently, the development of coldstart mitigation techniques has become a central research challenge in the design of autoscaling mechanisms [25], [10]. Clearly, overprovisioning is not sustainable because it leads to substantial energy consumption and operational costs.

J. Anselmi is with Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, 38000 Grenoble, France. Email: jonatha.anselmi@inria.fr

A. Motivation and Context

The autoscaling principle currently employed by serverless computing platforms such as AWS Lambda, Google Cloud Functions, Microsoft Azure Functions, IBM Cloud Functions, and Apache OpenWhisk is as follows [31], [43]:

> *"Activate a new function instance (server) upon a job arrival only if no active idle instance exists, and deactivate a function instance once it has remained idle for a specified duration."*

Building on this principle, a wide range of strategies have been proposed over the past decade to mitigate coldstart occurrences [13]. This is a very active applied research area at several different levels. At the infrastructure level, techniques such as cold-container reuse and function pre-warming have demonstrated important reductions in latency under certain production workloads [24], [33]. At the application and workflow level, approaches like function fusion, minifying initialization code, and optimizing library loading have further pushed performance improvements [27], [10], [14]. More recently, machine-learning and forecasting methods have been introduced to predict invocation patterns and pre-allocate capacity, yielding promising results [38], [37], [5].

Despite the existing body of work, the scientific community is not entirely satisfied: many mitigation techniques remain heuristic, often tailored to specific platforms or workloads, and their trade-offs – between cost, resource utilization, and coldstart frequency – are only partially quantified [23]. At large scale, the influence of infrastructure configuration parameters – controlled by the system manager at the infrastructure level or by the application developer at the application level – on coldstart occurrences remains unclear. This may be due to the fact that the literature lacks a systematic theoretical characterization of the dominant dynamics. This gap between theory and reality motivates the need for a more rigorous and quantitative understanding of coldstart dynamics through analytical models [31], [37]. In this paper, we follow this perspective and base our analysis on a stochastic queueing-theoretic model.

B. Problem and Approach

We consider the following question: *how long does it take to observe the next coldstart given the current system state?* We refer to such performance metric simply as *coldstart hitting time*. Importantly, it should not be confused with the 'coldstart time/latency' metric, which in the computer systems literature refers to the time overhead of initializing container or virtual machine instances in serverless platforms [27]. Instead, the pro-

posed metric captures a state-dependent property, consequence of the stochastic dynamics of the servers.

Our goal is to investigate the coldstart hitting time and to this purpose, it is natural to consider a queueing model. Existing queueing models for autoscaling focus on stationary properties of the underlying system. In contrast, our approach focuses on the *transient* (non-stationary) dynamics and translates the problem to a *first-passage or hitting time* (FPT) problem; hence the name of the performance metric of interest. Although such problems have been broadly studied in probability and statistics, with applications across fields such as finance, neuroscience, and biology, the particular case of autoscaling deserves its attention because it presents its own distinctive features, see Section II-B.

The approach above is motivated by the extremely reactive nature of current autoscaling mechanisms, which continuously adapt based on real-time system observations. Serverless platforms allow the system manager to implement scale-up policies that can activate a large number of servers instantaneously, potentially at any order of magnitude [1]. Such abrupt state changes somewhat affect stationary assumptions. In this context, up to our knowledge, our work is the first to provide a state-dependent analytical characterization of the coldstart hitting time within a stochastic dynamic model.

Since the proposed stochastic model is intractable, we analyze dynamics on asymptotic regimes that are representative of large-scale production systems. In this respect, a natural first step consists in investigating the fluid or mean-field limit, which gives a deterministic, first-order description of the system dynamics. However, as we demonstrate in Section IV, this approach is not precise enough to capture the stochastic fluctuations that drive coldstart events. To overcome this limitation, we consider the system in a *diffusion scaling*. This scaling appears to be appropriate, as it yields non-trivial but tractable approximations of the mean coldstart hitting time while providing insights into how system parameters influence it.

C. Contribution

We consider a queueing model for the dynamics induced by the autoscaling principle described above. Our model for the number of busy servers will be a GI/M/∞ queue, denoted by $(X_t^{(n)})$ at time $t \geq 0$. The process representing the number of active (warm) servers, denoted by $(W_t^{(n)})$, will be then coupled to $(X_t^{(n)})$ and satisfy $W_t^{(n)} \geq X_t^{(n)}$. Each *idle* warm server deactivates itself after a timer with mean duration $1/\gamma$ expires. Here, n is a parameter related to the system size, which we will let grow to infinity. In this framework, the coldstart problem translates into an FPT problem: we are interested in the first time the trajectories of $W_t^{(n)}$ and $X_t^{(n)}$ meet, $\tau^{(n)} := \min\{t \geq 0 : X_t^{(n)} = W_t^{(n)}\}$, as this represents the first time where no warm server is idle, or ready to serve an incoming job. An exact analysis of $\mathbb{E}[\tau^{(n)}]$ appears analytically intractable. This motivates the study of $(X_t^{(n)}, W_t^{(n)})$ under a diffusion scaling.

Within this scaling, we first establish in Theorem 1 a process-level (weak) convergence to a two-dimensional diffusion process and the weak convergence of $\tau^{(n)}$ to τ, interpreted as the FPT of a Gaussian process to zero. Since in practice n is "large", this justifies the first approximation

$$\mathbb{E}[\tau^{(n)}] \approx \mathbb{E}[\tau].$$

The starting point of our proof is the seminal work developed by Whitt in [45]. Our argument here follows a common approach – Doob–Meyer decomposition of W_t, martingale control and continuous mapping theorem – but its application is technical due to the state-dependent coupling between X_t and W_t. This establishes a connection between the problem of interest and a well-known FPT problem.

Although FPT problems related to Gaussian processes have been extensively studied over the past century, they remain analytically intractable, as existing approaches are typically based either on numerical computations or on asymptotic analyses. Moreover, the specific FPT problem for autoscaling considered here has the following inconvenient characteristics on the underlying Gaussian process (defined in (7)): it is non-stationary, non-centered, it has no C^1 paths, nor mean-square differentiability. In addition, the parameter regime of interest is such that the mean passage time is neither asymptotically small nor large, which implies that the boundary is not far away. These conditions place our problem somewhat beyond the direct applicability of existing theoretical results.

Given the analytical intractability of our FPT problem, we develop an approximation for $\mathbb{E}[\tau]$ that relies on an heuristic argument involving the celebrated Kac-Rice formula for the expected number of barrier crossings. This formula cannot be literally applied because our Gaussian process does not enjoy the usual assumptions of the classical theory, in particular sample path differentiability. However, we apply it anyway *by replacing the conditional crossing rate through the barrier (Rice prefactor), which is not well defined in our case, by the conditional infinitesimal drift of the process*; see Section V-A. This is the key step of our heuristic, which leads to the following approximation for the coldstart hitting time: provided that $\gamma/\mu + \operatorname{erfc}(d\sqrt{(\mu+\gamma)/\theta}) < 1$,

$$\mathbb{E}[\tau] \approx \frac{1}{\gamma}\left(\log d + \frac{1}{2}\log\frac{\mu+\gamma}{\theta} - \log z(\gamma, d)\right) \qquad (1)$$

with

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

where γ is the deactivation rate of idle warm servers, d the current (normalized by $\sqrt{n}$) number of active idle servers, μ the job service rate, $\theta \stackrel{\text{def}}{=} \lambda + \lambda^3\nu^2$ with λ^{-1} and ν the mean and standard deviation of job interarrival times, and finally, erfc denotes the complementary error function $\operatorname{erfc}(x) := \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt$, for all $x \in \mathbb{R}$.

Our main theoretical result on $\mathbb{E}[\tau]$ is given in Theorem 2, where we show that the RHS of (1) approximates $\mathbb{E}[\tau]$ up to an additive constant that does not depend on d. We prove this result using a mix of probabilistic arguments mainly involving a Foster–Lyapunov contraction argument and concentration inequalities such as Borell-TIS [2]; see Appendix B. The magnitude of this additive constant is nevertheless significant

in the parameter regime of autoscaling, and neglecting it would lead to inaccuracies as we show numerically.

When compared with estimates of the mean coldstart hitting time $\mathbb{E}[\tau^{(n)}]$ obtained from Monte Carlo simulations of the stochastic system $(X^{(n)}, W^{(n)})$, we find that the proposed approximation (1) provides a remarkable accuracy. Across a broad range of parameters that are typical of autoscaling, the percentage relative error of our approximation is less than 1.5%, significantly outperforming other approximation methods.

D. Applicability of the Approximation

Let us comment on the applicability of (1) to the case of autoscaling. First, the requirement $\frac{\gamma}{\mu}+\mathrm{erfc}(d\sqrt{(\mu+\gamma)/\theta}) < 1$, which is necessary to invert the erfc function in the right branch, is naturally satisfied in practice. This because idle-server timers are much longer than the job service times, $\gamma \ll \mu$. Also, since d represents the number of spare warm servers currently available for immediate processing normalized by $\sqrt{n}$, i.e., a second-order fluctuation, d take moderate to large values so that $\mathrm{erfc}(d\sqrt{(\mu+\gamma)/\theta}) \approx 0$.

In serverless and cloud systems, both γ and d are under the control of the system manager. The first is a platform configuration parameter, while the second can be used to define a scale-up policy with an handle on the coldstart hitting time. Here, our analysis and Formula (1) provide a simple and interpretable approximation of the mean coldstart hitting time as a function of these parameters. This yields the following theoretical and practical insights for designing more responsive and energy-efficient autoscaling policies.

Remark 1. *Provisioning too many warm servers at once (increasing d) yields logarithmic diminishing returns. Allowing idle servers to remain active longer (decreasing γ) may be more convenient because of the stronger $1/\gamma$ scaling. The precise trade-off between these two parameters depends on the specific cost function used to balance coldstart occurrences and energy consumption.*

Remark 2. *Platform users can trigger scale-up actions based on the current load, potentially activating a bulk of servers instantaneously [1]. In our model, this translates into a quick change of the initial condition, i.e., a change on d. A number of autoscaling policies aim to maintain a "safety" pool of idle-on servers of constant size (a.k.a. pre-warmed, pre-provisioned pool) [38], [5] – large enough to prevent coldstarts, yet small enough to avoid unnecessary energy consumption. In this setting, the message of our result is:*

> "If the idle-on pool has size $d\sqrt{n}$, the coldstart hitting time one gets is as in (1)."

The approximation can also be used as a building block for proactive autoscaling. The model does not prescribe that the controller should wait until a coldstart occurs. Rather, it quantifies the expected time to the next coldstart if no additional scale-up action is taken from the current state. Therefore, a controller may use $T^\star(\gamma, d)$ following this rule: "*whenever $T^\star(\gamma, d) \le T_{\mathrm{th}}$, for some target safety threshold T_{th}, the controller activates additional warm servers*". If k servers are activated in a system of size n, the diffusion-scaled idle-server gap is updated from d to $d + k/\sqrt{n}$, and the estimate $T^\star(\gamma, d + k/\sqrt{n})$ can then be recomputed. In this sense, our formula provides a state-dependent warning signal for triggering proactive scale-up actions before a coldstart is observed.

E. Organization

The remainder of the paper is organized as follows. Section II reviews the relevant literature. Section III introduces the stochastic framework for autoscaling. Section IV defines a diffusion model and establishes its relationship with the queueing model. Section V develops the proposed heuristic and gives the error bound result. Section VI numerically validates the quality of our approximation and, finally, Section VII draws the conclusions discussing how certain assumptions of our model can be removed.

II. Relevant Literature

Since our analysis ultimately reduces the problem of interest to a classical first-passage problem, we review two distinct lines of related literature. The first concerns queueing models for autoscaling systems, and the second relates to first-passage-time (FPT) problems.

A. Autoscaling and Queueing

The problem of mitigating coldstarts in autoscaling systems has been extensively investigated. As outlined in the introduction, it has been addressed from multiple perspectives [13], [33] Here we focus on the line of work that relies on queueing models, which is most closely related to ours.

Since the foundational M/G/1 queue with vacation model [22], systems in which servers can be turned off or on at the same timescale as job dynamics, with or without setup (initialization) times, have been extensively studied in queueing theory. A major milestone in this line of work is the Fuhrmann–Cooper decomposition [15], a structural result showing that the steady-state queue length can be expressed as the sum of two independent components. Similar decomposition results are often found in queueing systems [44]; see also [17].

Over the last decade, the previous classical models have inspired new variants designed to capture the specific features of modern computing technologies [17], [46], [20], [30], [6], [7]. The scale-down rule "turn off a server after a certain idle time", which is a de facto standard in serverless computing and also adopted in our model, has been analyzed in several studies [4], [16]. A recent contribution that closely reflects the autoscaling mechanisms used in platforms such as Amazon's AWS Lambda is the Markov model proposed in [31], which is the closest to ours. In that framework, active and busy servers are modeled as M/GI/s/s loss queues, where s denotes the nominal system capacity, i.e., the maximum number of servers that can be active simultaneously. The authors address the coldstart problem by analyzing the steady-state probability of a coldstart occurrence. Due to analytical intractability, they propose an heuristic approximation that validate empirically

by simulations. The model in [31] has been recently extended in [5], [19] to account for the possibility of activating multiple servers upon a coldstart event, instead of just one. The idea is close to what is known in the computer systems literature as *container pre-warming* [38]: anticipating potential coldstarts in the near future while striving an optimal balance between coldstart occurrences and energy consumption. In particular, the authors develop a reinforcement learning algorithm with convergence guarantees that learns over time the optimal number of servers to scale up. A similar direction is followed in [41], where the problem is formulated as a Markov decision process.

Another closely related work to ours is the autoscaling model studied in [28], [29], a two-dimensional Markov chain where Stein's method is applied to characterize its stationary behavior in terms of a non-trivial PDE. Similar to our approach, their analysis considers the system under a diffusion scaling. The main difference lies in the scale-up rule: in their model, new servers are activated by an exogenous Poisson process, independently of the system state – technically, this decoupling simplifies the structure. Such asynchronous autoscaling is employed in some settings; see [4, Section 1.3] for a discussion. However, in practice most autoscaling systems trigger scale-up decisions at job arrival times.

All the works mentioned above analyze properties of the underlying stochastic process in a steady-state regime. In contrast, our study focuses on a performance metric that naturally belongs to the transient regime. Most of the works above assume an underlying M/M/s or M/M/s/s queue, whereas we consider an infinite-server setting. In practice, the nominal capacity in serverless platforms is merely an upper bound specified by the platform user and is rarely, if ever, reached.

B. First-Passage Times

With Theorem 1 and Remark 4, we will reduce our problem to studying the FPT of a one-dimensional Gaussian process to zero; or equivalently, the FPT of a Ornstein–Uhlenbeck (O–U) process to a stochastically decreasing barrier. A large body of literature in applied probability and statistics considered the density and moments of FPTs to a fixed level, say w, for one-dimensional Gaussian processes; some classical references are [21], [2], see also [47] for a recent account. This problem is of broad interest: beyond queueing theory, it arises in several domains such as physics, finance, neuroscience, and biology. Despite the vast literature, explicit solutions are usually expressed in terms of special functions, and no general closed-form solution is known except in special cases [26]. As a result, existing approaches are mainly numerical or rely on asymptotic methods and heuristics.

Existing asymptotic methods are mostly based on long- or short-time asymptotics [32], a very small variance, a constant time horizon or very high barrier levels ($w \to \infty$) [8], [18], [21, Chapter 12]. In our autoscaling setting, the warm-server deactivation rate γ keeps the barrier level uniformly bounded (see (20)), and the meeting time occurs at a time that significantly depends on the initial barrier level. This prevents the application of asymptotic results such as the celebrated Pickands' theorem.

In addition, to apply the general theory, the underlying Gaussian process needs at least one of the following assumptions [9], [35], [36], [34], [12]: i) to possess C^1 sample paths, ii) to be mean-square differentiable, iii) to be centered or stationary, or iv) to have a constant barrier w. None of these assumptions hold for the Gaussian process of our autoscaling setting, defined in (7).

A closely related problem to ours is the mean FPT of an O–U process to an exponentially-decaying (with rate γ) deterministic boundary (in our setting, the barrier is also decreasing, but stochastically). This connection can be shown by replacing our stochastic barrier by its expected value, provided that γ is small. Though a bit simpler because the boundary is deterministic, it is well known that this problem remains analytically intractable [42], even in the case of a constant barrier [26], [32], [3], [47]. In [42], the authors reduce the problem to a partial differential equation that can be solved numerically, and [39] investigates the case where the hitting process is a Wiener process, while ours is an O–U process.

III. Stochastic Framework for Autoscaling

We describe the architecture of the autoscaling system under investigation, define a stochastic model for its dynamics and introduce the problem addressed in this paper.

A. Autoscaling System

We consider a system composed of an unlimited number of *servers*. In serverless computing, a server – also called container, cloud function, instance, or replica – is a flexible software entity that can be rapidly instantiated, even within tens to hundreds of milliseconds. Each server can be in one of three macro states: *cold* (turned off), *warm* (turned on), or *initializing* (transitioning from cold to warm). During the initialization phase, a server performs startup operations such as connecting to databases and is not yet available to process jobs. A warm server can be either *idle-on* (warm but not processing any job) or *busy* (warm and processing a job).

At any point in time, a central controller knows the state of each server.

In agreement with the discussion in the Introduction, we consider an autoscaling system that operates according to the following rules:

- **No autoscaling**: When a job arrives and at least one idle-on server is available, the job is immediately assigned to one of them, making that server busy.
- **Scale-up rule**: If no idle-on server is available upon job arrival, a *coldstart* occurs: a cold server is activated by entering the initializing state, and once its initialization is completed, it becomes busy with the waiting job.
- **Scale-down rule**: Whenever a busy server finishes its job, it becomes idle-on and starts a timer of random duration. If no new job is assigned to it before the timer expires, the server transitions back to the cold state.

In any case, after processing, jobs leave the system and servers transition to idle-on.

Remark 3. *In practice, idle-on server timers are much longer than job service times. For example, in serverless computing platforms such as AWS Lambda, typical service times range between 10 and 30 milliseconds, idle timers lasts 2–5 minutes, and server initialization times are less than a second [43]. In addition, existing platforms are huge in size and able to keep thousands of servers active. This motivates an analysis in a regime where γ (timer expiration rate) is small with respect to μ (job service rate) and n (system size) is large.*

Our analysis will concern the time to observe the next coldstart conditioned on the presence of at least one idle-on server in the current state. Given the autoscaling mechanism above, no scale-up events are triggered within this time interval. Consequently, the initializing state plays no role in our analysis and will be omitted in our model below for simplicity.

B. Queueing Model

We consider a sequence of autoscaling systems indexed by n. In the n-th system, the sequences of job interarrival times, service times and idle-on timer expirations are all i.i.d. and mutually independent. Service times and expiration times are exponentially distributed with rates μ and γ, respectively (these assumptions can be weaked as we will discussed in the Conclusions). Interarrival times follow the distribution of a random variable $A^{(n)}$, which satisfies

$$\sqrt{n}\big(\mathbb{E}[nA^{(n)}] - \lambda^{-1}\big) \to 0, \qquad \mathrm{Var}(nA^{(n)}) \to \nu^2,$$

for some $0 < \lambda < \infty$ and $0 < \nu^2 < \infty$.

Then, we model the n-th autoscaling system by the stochastic process $(X_t^{(n)}, W_t^{(n)})$, where $X_t^{(n)}$ and $W_t^{(n)}$ respectively represent the number of busy and warm servers at time $t \geq 0$. The process $X^{(n)}$ coincides with the number of jobs in a GI/M/∞ queue up to the first coldstart time, defined in the following by $\tau^{(n)}$. This restriction is important: after a coldstart, the job that triggered it would wait for the initialization of a new server before entering service, and an additional initialization state would have to be modeled. Since this paper studies the time to the first coldstart, only the dynamics on $[0, \tau^{(n)}]$, or equivalently the stopped processes, are relevant. The process $(W_t^{(n)})$ satisfies $W_t^{(n)} \geq X_t^{(n)}$ for all $t \geq 0$ and evolves through two types of events. First, when a job arrival occurs at time t and all warm servers are busy (i.e., $W_{t^-}^{(n)} = X_{t^-}^{(n)}$), one additional server is turned on, resulting in an upward jump of $W^{(n)}$. Second, when a server remains idle-on for a random exponential time with rate γ, its idle timer expires and the server switches off, generating a downward jump of $W^{(n)}$. The dynamics of $W^{(n)}$ can be expressed through the following stochastic differential equation:

$$dW_t^{(n)} = \mathbf{1}_{\{W_{t^-}^{(n)} = X_{t^-}^{(n)}\}}\, dN_t^{(n)} - \gamma\big(W_t^{(n)} - X_t^{(n)}\big)\, dt - dM_t^{(n)}, \tag{2}$$

where $N_t^{(n)}$ denotes the job arrival counting process, and $M_t^{(n)}$ is a square-integrable martingale capturing the randomness associated with the timer-expiration events. Equation (2) is to be understood on the pre-coldstart time interval $[0, \tau^{(n)}]$. In particular, on $[0, \tau^{(n)})$ no scale-up event occurs, and the upward-jump term is only relevant to identify the hitting/coldstart event itself.

C. Problem Statement

We focus on the *coldstart hitting time*, the time until the next coldstart event occurs given the current system state. We define $\tau^{(n)}$ as the stopping time at which the number of busy and warm servers coincide:

$$\begin{aligned}\tau^{(n)} &\stackrel{\text{def}}{=} \tau^{(n)}(x, w)\\ &\stackrel{\text{def}}{=} \min\big\{t \geq 0 : X_t^{(n)} = W_t^{(n)} \mid (X_0^{(n)}, W_0^{(n)}) = (x, w)\big\}.\end{aligned} \tag{3}$$

The dependence on the initial condition (x, w) will often omitted if not necessary or clear from the context. If $X_t^{(n)} = x, W_t^{(n)} = w$, it follows that over the interval $[t, t+\tau^{(n)}(x, w)]$, no scale-up actions are initiated, as these are triggered exclusively by arrivals encountering no idle-on servers.

The objective of this paper is to analyze how $\mathbb{E}[\tau^{(n)}]$ is impacted by the idle-on expiration rate γ and the current spare capacity $W_0^{(n)} - X_0^{(n)}$. Both quantities can be controlled by the system manager. Specifically, the parameter γ is typically a configuration constant of the autoscaling platform and evolves on a slow time scale. In contrast, the difference $W_0^{(n)} - X_0^{(n)}$ can be adjusted dynamically by the central controller, which may initialize any number of servers instantaneously. This supports the finetuning or design of autoscaling policies.

IV. Diffusion Approximation

We study $(X^{(n)}, W^{(n)})$ in a limiting regime where n is large. The motivation is twofold. First, cloud platforms typically operate with thousands of servers, so analyzing the system under a large-n assumption is both natural and realistic. Second, an exact analytical characterization of the mean coldstart hitting time $\mathbb{E}[\tau^{(n)}]$ is intractable, whereas examining an appropriate limiting regime can yield a more tractable description.

A common first step in this direction is to investigate the *fluid limit*, which provides a first-order approximation of the system dynamics. Unfortunately, such an approximation is not sufficient to extract meaningful insights on $\mathbb{E}[\tau^{(n)}]$. The first limitation is that if $(X_0^{(n)}/n, W_0^{(n)}/n) \to (x_0, w_0)$ almost surely, with $w_0 > x_0$, then the initial pool of idle-on servers grows linearly with n. This is unrealistic because autoscaling policies are designed to keep this pool small to reduce energy consumption. The second, and more fundamental, limitation is as follows. Assuming Poisson arrivals for simplicity and setting $x_0 = \rho := \lambda/\mu$, one can show that $(X^{(n)}/n, W^{(n)}/n)$ converges to the deterministic process $(x(t), w(t))$ satisfying, for $t \geq 0$,

$$x(t) = \rho, \quad \dot{w}(t) = -\gamma\big(w(t) - \rho\big), \quad w(0) = w_0.$$

Analogously to (3), define the fluid-scale coldstart hitting time as $\tau_{\text{fluid}} \stackrel{\text{def}}{=} \min\{t \geq 0 : x(t) = w(t) \mid (x(0), w(0)) = (\rho, w_0)\}$. The issue is that $\tau_{\text{fluid}} = +\infty$ and so the fluid limit fails to capture the second-order ($\sqrt{n}$) fluctuations that are essential for obtaining a meaningful estimate of $\mathbb{E}[\tau^{(n)}]$. To overcome this difficulty, we now introduce a *diffusion model*.

A. Diffusion Model

The diffusion model for autoscaling is referred to as (X_t, W_t), where X_t and W_t are interpreted as *normalized* versions of the processes of the busy and warm servers, respectively; the precise normalization will be given in (8) and such interpretation will follow from Theorem 1.

For now, we define (X_t) as the Ornstein–Uhlenbeck (O–U) process with initial value $X(0) = x_0$, infinitesimal mean $m(x) = -\mu x$, and infinitesimal variance $\theta := \lambda + \lambda^3\nu^2$. Thus, it satisfies the SDE

$$dX_t = -\mu X_t dt + \sqrt{\theta}\, dB_t, \quad t \geq 0 \tag{4}$$

where B_t denotes a Brownian motion started at zero. This is the clear candidate for the diffusion limit of $X^{(n)}$ because it is well-known that the properly normalized process of the number of jobs in the GI/M/∞ queue converges to a O–U process, e.g., [45]. Then, we define (W_t) as the process satisfying

$$dW_t = -\gamma\,(W_t - X_t)\,dt, \quad t \geq 0, \tag{5}$$

$W(0) = w_0 > x_0$, which is the candidate for the diffusion limit of $W^{(n)}$ in the time window of interest, i.e., before a coldstart happens. Analogously to (3), we define

$$\tau \stackrel{\text{def}}{=} \tau(x,w) \stackrel{\text{def}}{=} \min\{t \geq 0 : X_t = W_t \mid (X_0, W_0) = (x, w)\}, \tag{6}$$

as the coldstart hitting time at the diffusion scale, which is well defined because X_t and W_t are continuous.

Let $D_t \stackrel{\text{def}}{=} W_t - X_t$ be the idle-on server process at diffusion scale. A direct calculation (see Section C-A) gives

$$\begin{aligned} D_t \;&=\; e^{-\gamma t} D_0 + \mu x_0 e^{-\gamma t} \frac{1 - e^{-(\mu-\gamma)t}}{\mu - \gamma} \\ &\quad + \frac{\sqrt{\theta}}{\mu - \gamma} \int_0^t \left(\gamma e^{-\gamma(t-u)} - \mu e^{-\mu(t-u)}\right) dB_u. \end{aligned} \tag{7}$$

Remark 4. *The coldstart time can be written in terms of the idle-server process $D_t := W_t - X_t$, as $\tau = \min\{t \geq 0 : D_t = 0 \mid D_0 = w - x\}$. Thus the two-dimensional problem is equivalently a first-passage-time problem for the one-dimensional Gaussian process D. Although this reformulation places the problem within the broad class of Gaussian first-passage problems, the resulting process does not satisfy the standard assumptions under which classical formulas are usually derived. Indeed, D is non-stationary, non-centered, has no C^1 sample paths, is not Markov, and is not mean-square differentiable. Moreover, the relevant hitting level is not asymptotically high: under the approximation developed below, the level at which the crossing occurs remains uniformly bounded as the initial gap d grows. As discussed in Section II-B, this prevents the direct use of standard high-boundary or smooth-path asymptotics.*

The formulation (7) will be convenient for the construction of our heuristic.

B. Connection between the Diffusion and Queueing Models

Let $\rho \stackrel{\text{def}}{=} \lambda/\mu$,

$$\overline{X}_t^{(n)} \stackrel{\text{def}}{=} (X_t^{(n)} - n\rho)/\sqrt{n} \tag{8a}$$

$$\overline{W}_t^{(n)} \stackrel{\text{def}}{=} (W_t^{(n)} - n\rho)/\sqrt{n} \tag{8b}$$

and let $\Rightarrow$ denote weak convergence (convergence in distribution) both for random variables in $\mathbb{R}$ and stochastic processes in $D[0,\infty)$, the space of càdlàg functions.

For a stochastic process $(Z_t)_{t\geq 0}$ and a stopping time $\mathcal{T}$, let $(Z_{t\wedge\mathcal{T}})_{t\geq 0}$ denote the stopped process at $\mathcal{T}$.

The next result shows the relationship between the queueing model in Section III-B and the diffusion model with stochastic boundary; see Appendix A for a proof.

Theorem 1. *Assume that the initial conditions satisfy*

$$(\overline{X}_0^{(n)}, \overline{W}_0^{(n)}) \;\Rightarrow\; (X_0, W_0), \quad \textit{with } W_0 \geq X_0 \textit{ a.s.},$$

and have uniformly bounded second moments: $\sup_{n\geq 1} \mathbb{E}[|\overline{X}_0^{(n)}|^2 + |\overline{W}_0^{(n)}|^2] < \infty$. *Then, as $n \to \infty$,*

$$\tau^{(n)} \;\Rightarrow\; \tau \tag{9}$$

$$\left(\overline{X}^{(n)}_{\cdot\wedge\tau^{(n)}},\, \overline{W}^{(n)}_{\cdot\wedge\tau^{(n)}}\right) \;\Rightarrow\; \left(X_{\cdot\wedge\tau},\, W_{\cdot\wedge\tau}\right),$$

in $D[0,\infty) \times D[0,\infty)$.

In view of Theorem 1, the diffusion model (X_t, W_t) can be interpreted as a proxy to approximate the dynamics of the original process $(X_t^{(n)}, W_t^{(n)})$ when n is large. This lets us shift the analysis to τ.

V. Analysis of the Coldstart Hitting Time τ

We have established that the coldstart hitting time of our original model, $\tau^{(n)}$, can be approximated by the first-passage time, denoted τ, of a one-dimensional Gaussian process to zero. While this is a classical, long-standing problem, the specific setting of autoscaling prevents the direct application of existing results, as discussed in Section II-B. In this section, we develop a simple approximation for $\mathbb{E}[\tau]$ based on an heuristic argument and then prove an error bound.

A. Heuristic Derivation: A Rice–type Argument

Let (D_t) be as in (7), with $D_0 = d > 0$, and let

$$\tau = \min\{t \geq 0 : D_t = 0 \mid D_0 = d\}. \tag{10}$$

The heuristic argument for our approximation of $\mathbb{E}[\tau]$ is composed of two steps.

Step 1: Use upcrossings. The first-passage time τ corresponds to the point in time where the number of downcrossings of the barrier at level zero reaches one. In view of this observation, the first step is to 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. Though first-passage times and crossings are closely related, these two times are different in general and their relationship in terms of their expectation is unclear [21], [9]. For this reason, this step remains heuristic. In addition, in our setting the situation is even more delicate

because D_t has an infinite number of downcrossings even in short intervals. This leads to the second step.

Step 2: Replace the Rice prefactor with the conditional infinitesimal drift. For a smooth (C^1) Gaussian process Z_t started at z_0 with derivative Z'_t, the expected number of *crossings* (i.e., upcrossings and downcrossings) at level u over the time interval $[0,T]$ is given by the Kac-Rice formula, e.g., [9, Theorem 3.2]:

$$\mathbb{E}[N([0,T],u)] = \int_0^T \mathbb{E}\big[|Z'_t| \mid Z_t = u\big]\, f_{Z(t)}(u)\, dt. \quad (11)$$

where $f_{Z(t)}(u)$ denotes the density of Z_t. For our process (D_t) driven by Brownian motion, D'_t does not exist in the classical sense, and the expected number of downcrossings over any interval of positive length is infinite. Nevertheless, one can heuristically approximate the *probability flux* of a first crossing in a small interval $[t, t+\Delta]$ by considering the process over that interval and relating it to a weaker/different notion of derivative through the barrier. This leads to the key point of our heuristic: replace the drift term $\mathbb{E}\big[|Z'_t| \mid Z_t = u\big]$ in (11) by the *conditional infinitesimal drift*

$$\mathcal{L}_t f(x) \stackrel{\text{def}}{=} \lim_{\delta \downarrow 0} \frac{\mathbb{E}[f(Z_{t+\delta}) \mid Z_t = x] - f(x)}{\delta}$$

of (Z_t) with the test function $f(x) = x$ and evaluated at barrier level $x = 0$.

Remark 5. *Since the process D is not differentiable in time, the classical Rice prefactor $\mathbb{E}[|D'_t| \,|\, D_t = 0]$ is not well defined. The conditional infinitesimal drift $\mathcal{L}_t f(0)$, with $f(x) = x$, provides a weaker notion of velocity through the boundary: it captures the conditional mean displacement of D over an infinitesimal time interval, given that $D_t = 0$. What is lost in this replacement is the fine local oscillatory behavior caused by the Brownian component. In particular, it does not describe the actual pathwise crossing intensity, which is infinite on every non-trivial time interval. Thus the replacement of the Rice prefactor by $|\mathcal{L}_t f(0)|$ should be interpreted as a smoothed mean-flux approximation, rather than as a literal application of the Kac–Rice formula.*

The following proposition provides the conditional infinitesimal drift when $Z_t = D_t$.

Proposition 1. *For $f(x) = x$,*

$$\mathcal{L}_t f(0) = -\mu(\mu-\gamma)\, d\, e^{-\gamma t} \times \frac{\frac{\gamma}{\mu+\gamma}\big(1 - e^{-(\mu+\gamma)t}\big) - \frac{1}{2}\big(1 - e^{-2\mu t}\big)}{\frac{\gamma}{2}\big(1 - e^{-2\gamma t}\big) - \frac{2\gamma\mu}{\gamma+\mu}\big(1 - e^{-(\gamma+\mu)t}\big) + \frac{\mu}{2}\big(1 - e^{-2\mu t}\big)} \quad (12)$$

$$\sim \mu\, d\, e^{-\gamma t}, \quad \text{as } t \to \infty.$$

Here, $f(t) \sim g(t)$ means that $f(t)/g(t) \to 1$ as $t \to \infty$.

Now, letting

$$w(t) \stackrel{\text{def}}{=} d e^{-\gamma t},$$

the density of D_t at level 0 is

$$f_{D_t}(0) = \frac{1}{\sqrt{2\pi\, \sigma_D^2(t)}} \exp\Big(-\frac{w(t)^2}{2\,\sigma_D^2(t)}\Big) \quad (13)$$

with

$$\sigma_D^2(t) = \frac{\theta}{(\mu-\gamma)^2}\Big[\frac{\gamma}{2}\big(1 - e^{-2\gamma t}\big) - \frac{2\gamma\mu}{\gamma+\mu}\big(1 - e^{-(\gamma+\mu)t}\big) + \frac{\mu}{2}\big(1 - e^{-2\mu t}\big)\Big] \sim \frac{\theta}{2} \times \frac{1}{\mu+\gamma}, \qquad \text{as } t \to \infty.$$

Substituting (13) in the RHS of (11) (recall by Remark 3 that γ is much smaller than μ in our autoscaling setting), we finally obtain

$$N_{\text{heu}}(T) := \int_0^T \frac{\mu\, w(t)}{\sqrt{\pi\theta/(\mu+\gamma)}} \exp\Big(-\frac{w(t)^2}{\theta/(\mu+\gamma)}\Big)\, dt, \quad (14)$$

a claimed heuristic approximation for the expected number of "smooth crossings". Then, since we want to impose the expected number of "smooth *down*crossings" to one, we postulate that the number of such upcrossings and downcrossings is equal in a short time interval. This yields the following *Rice–type equation*

$$N_{\text{heu}}(T) = \tfrac{1}{2}, \quad (15)$$

and the T that solves (15), referred to as T^*, is our proxy to $\mathbb{E}[\tau]$. This completes the heuristic justification behind our approximation for the coldstart hitting time.

B. Solving the Equation: The Heuristic Approximation T^*

The nice property of the Rice–type equation (15) is that its solution can be explicited. This is established in the following proposition; see Appendix C-C for a proof.

Proposition 2. *Assume that*

$$0 < \frac{\gamma}{\mu} + \operatorname{erfc}\left(d\sqrt{\frac{\mu+\gamma}{\theta}}\right) < 1. \quad (16)$$

Then, there exists a unique $T^* < \infty$ solution to the equation (15)*, and it is given by*

$$T^*(\gamma, d) = \frac{1}{\gamma}\left(\log d + \tfrac{1}{2}\log\frac{\mu+\gamma}{\theta} - \log z\right) \quad (17)$$

with

$$z := \operatorname{erfc}^{-1}\left(\frac{\gamma}{\mu} + \operatorname{erfc}\left(d\sqrt{\frac{\mu+\gamma}{\theta}}\right)\right). \quad (18)$$

Let us comment on the roles of the control parameters γ and d on T^*.

Since $d > 0$ and $\theta > 0$, the argument of the complementary error function is strictly positive, and therefore

$$\operatorname{erfc}\left(d\sqrt{\frac{\mu+\gamma}{\theta}}\right) \in (0,1).$$

Consequently, for fixed d, μ, θ, the condition (16) is satisfied whenever γ is small enough. This is consistent with the autoscaling regime, where idle-server timers are much longer than service times, i.e., $\gamma \ll \mu$.

Although $\operatorname{erfc}(d\sqrt{(\mu+\gamma)/\theta})$ decays extremely fast as d increases and may appear negligible for moderate to large d, it plays a crucial structural role in the expression of T^*. This

term, through its contribution to z, ensures that T^* remains uniformly bounded in the regime of small γ.

We also remark that, when the left-hand side of the admissibility condition (16) approaches one, the term $z(\gamma, d)$ approaches zero and $T^\star$ diverges to $+\infty$. This divergence is a limitation of the heuristic in that regime. However, this case is of limited interest for autoscaling, where $\mu \gg \gamma$ and d takes moderate to large values, so that the admissibility condition is safely satisfied in the regimes of practical interest.

C. Relationship between $\mathbb{E}[\tau]$ and T^*

At this point, we have developed a simple and interpretable formula for the coldstart hitting time, whose dependence on the model parameters is explicit and transparent. However, we do not have error bounds with respect to $\mathbb{E}[\tau]$.

In this context, our main result is the following and establishes that the error of the proposed heuristic is bounded in the initial gap d; see Appendix B for our proof.

Theorem 2. *Let τ be defined as in* (10) *and T^* as in* (17)*. Then*

$$\mathbb{E}[\tau] = T^* + O(1), \tag{19}$$

where the implicit constant does not depend on d.

The reason why classical high-boundary asymptotics cannot be applied is that the level at which the crossing occurs under the approximation is uniformly bounded in d. Indeed,

$$w(T^\star) = z\sqrt{\frac{\theta}{\mu+\gamma}} \le \operatorname{erfc}^{-1}\left(\frac{\gamma}{\mu}\right)\sqrt{\frac{\theta}{\mu+\gamma}}, \tag{20}$$

which does not depend on d. Hence, although the initial gap $D_0 = d$ may be large, the effective crossing level remains $O(1)$. This prevents the direct application of classical theorems and asymptotic results such as Pickands' theorem, as they apply to large barrier levels [21, Chapter 12] and constant time horizon.

Theorem 2 should be interpreted as a consistency result rather than as a sharp error estimate. The bound implies a vanishing relative error of order $O(1/\log d)$ for fixed γ. This result does not, by itself, explain the very small relative errors observed in the simulations. Those errors indicate that, in the autoscaling regimes considered numerically, the bounded remainder is small; quantifying this constant explicitly would require a substantially sharper analysis of the non-smooth Gaussian first-passage problem.

The control parameters of the system are γ and d, and Theorem 2 concerns the regime in which d varies while γ is fixed. Extending (19) to an explicit bound that also depends on γ appears out of reach. From the perspective of autoscaling applications, d (which reflects the fluctuations of the idle-server process) evolves on a much faster time scale than γ, a configuration parameter of the autoscaling platform. In this setting, treating γ as fixed is natural.

VI. Empirical Validation

In Proposition 2, we have developed a simple interpretable formula for the coldstart hitting time, T^*. The objective of this section is to validate its accuracy numerically via simulations of the coldstart hitting time $\mathbb{E}[\tau^{(n)}]$ associated to the original finite system $(X^{(n)}, W^{(n)})$.

Let $\tau_{\text{sim}}^{(n)}$ be the empirical mean coldstart hitting time obtained from Monte Carlo simulations of the n-th system $(X^{(n)}, W^{(n)})$, averaged over M samples; $n = \infty$ refers to the diffusion model. Given approximation 'approx' we evaluate the gap with respect to the true simulated value $\tau_{\text{sim}}^{(n)}$ using the relative percentage error

$$\mathcal{E}^{(n)}(T^*) := \frac{|T^* - \tau_{\text{sim}}^{(n)}|}{\tau_{\text{sim}}^{(n)}} \times 100\%.$$

This is a common error measure in queueing.

A. Autoscaling Parameters and Simulation Setup

We consider a unit service rate $\mu = 1$ and Poisson arrivals with (normalized) rate $\lambda \in \{0.25, 0.75\}$. These scenarios model low (usual) and moderate-to-heavy loads. Since the arrival process is Poisson, $\theta := \lambda + \lambda^3\nu^2 = 2\lambda$. Given $\mu = 1$, a realistic range for the deactivation rate γ is $[10^{-4}, 10^{-3}]$ (see Section III-A); e.g, if $\gamma = 10^{-4}$ and a service time unit corresponds to 30ms, then the mean idle timer duration is 5 minutes. Since $n\lambda/\mu$ corresponds to the equilibrium mean number of busy servers (a basic fact of the GI/GI/∞ queue), a reasonable range for n is $[10^3, 10^4]$. For d, the quantity $d\sqrt{n}$ represents the initial number of idle-on servers, we consider a wide parameter set $d \in [5, 30]$.

Finally, we set the parameter M, the number of simulations, to $M = 5,000$ when $n \le 10^4$. This was enough to get reliable mean error estimates, as the coefficient of variation, i.e., the ratio between the standard deviation and empirical mean, of $\tau_{\text{sim}}^{(n)}$ has been always less than 0.015. When $n = \infty$, the cost of simulations increased and we have used $M = 500$.

B. Validation against the Queueing Model

We now evaluate $\mathcal{E}^{(n)}(T^*)$. The numerical results obtained by varying n, λ, γ and d in the ranges above are shown in Figure 1. First, we observe that the errors do not vary much with the system size parameter n, which validates Theorem 1. We should notice that the error associated with the diffuson model ($n = \infty$) is noisy because of the numerical difficulty of simulating the diffusion model with an adequate discretization time step. We also observe that the proposed approximation becomes increasingly accurate as γ decreases and d increases, with very small errors ($\mathcal{E}^{(n)} \le 1\%$) in the most relevant setting where $\gamma = 10^{-4}$. This can be explained as follows:

- The improved accuracy for small γ is related to the critical heuristic argument in Step 2, where the Rice prefactor is replaced by the conditional infinitesimal drift. Since the latter "smooths out" the fast fluctuations of the O–U process X_t, it is plausible that this replacement performs better when the stochastic barrier is replaced by its expected counterpart, as it is smoother. This should be the case if γ is small, as in this setting the O–U process X_t sees W_t as a slow process.
- The improved accuracy for large d is explained by the heuristic argument in Step 1: when the barrier starts high, the time window in which the expected number

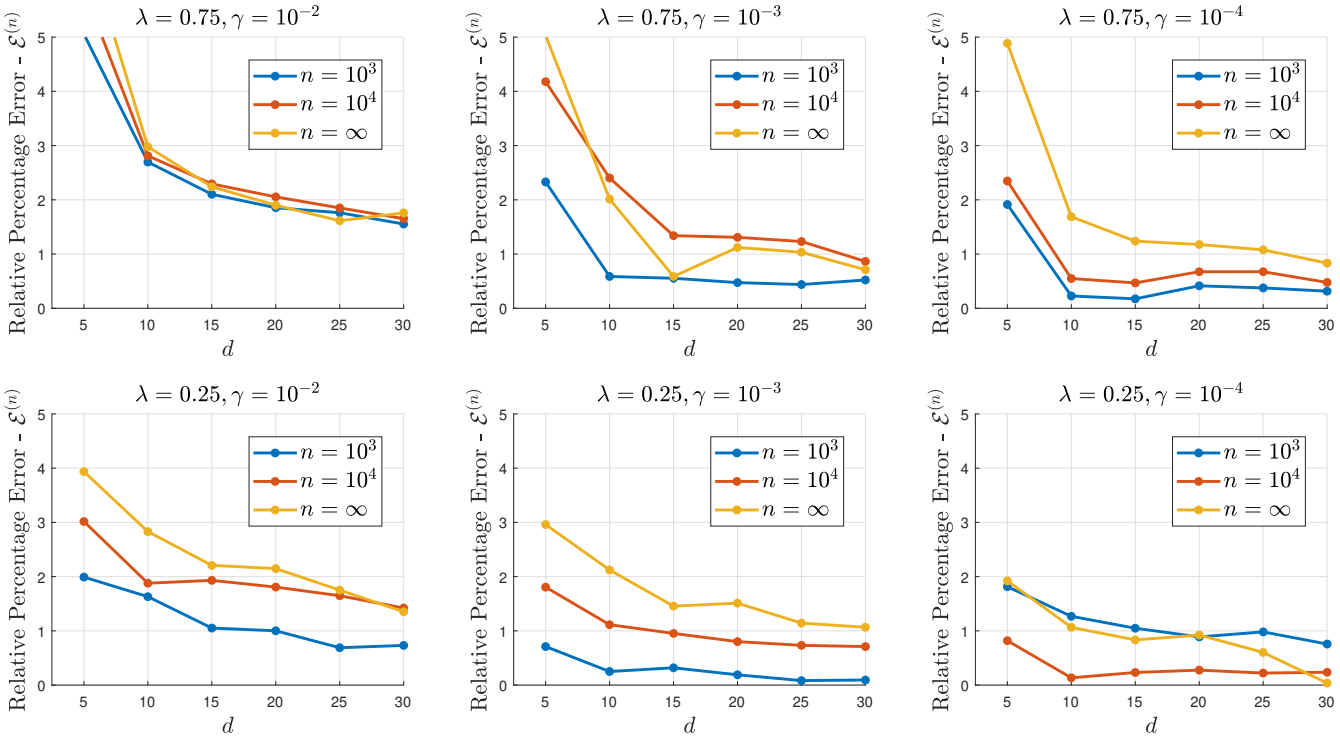


Figure 1: Relative percentage error $\mathcal{E}^{(n)}$ under different parameter settings. The results confirm the validity of the diffusion approximation, which becomes increasingly accurate as d grows as expected.

of "smooth upcrossings" occurs should shrink. In this situation, replacing the FPT by the time at which the expected number of such crossings reaches one becomes more accurate.

Finally, an unexpected non-monotonicity with respect to n can be observed in some parameter regimes: the relative error is sometimes slightly smaller for smaller values of n. The quantity plotted in Figure 1 combines two distinct errors: the finite-size error between $\tau^{(n)}$ and its diffusion limit τ, and the error between $T^\star$ and $\mathbb{E}[\tau]$. These two contributions need not have the same sign. In particular, for some finite values of n, the finite-size correction may partially compensate the Rice-type heuristic error, leading to a smaller observed relative error. This explains the non-monotonicity.

C. Robustness to Non-Poisson Arrivals and Non-Exponential Service Times

We next test the sensitivity of the approximation to two modeling assumptions.

In the first experiment, we keep exponential service times and exponential idle timers, but replace Poisson arrivals by renewal arrivals with squared coefficient of variation $c_a^2 \in \{0.25, 1, 4\}$. Specifically, we use Erlang-4, exponential, and hyperexponential interarrival times, respectively, all normalized to have mean $1/(n\lambda)$. In each case, the approximation $T^\star$ is computed with $\theta = \lambda + \lambda^3\nu^2 = \lambda(1 + c_a^2)$, as predicted by the diffusion scaling.

In the second experiment, we keep Poisson arrivals and exponential idle timers, but replace exponential service times by mean-one service times with different variability: Erlang-4, exponential, and hyperexponential services. The corresponding squared coefficients of variation are $c_s^2 = 0.25$, $c_s^2 = 1$, and $c_s^2 = 4$, respectively. Here $T^\star$ is used as a plug-in approximation with $\mu = 1/\mathbb{E}[S]$.

The results are reported in Figure 2; here, we set $M = 500$, $\lambda = 0.75$, $n = 10^3$ and $\gamma = 10^{-3}$. The approximation remains accurate for renewal arrivals even when the variance correction $\theta = \lambda(1 + c_a^2)$ is used. The errors are also stable under moderate changes in the service-time distribution, although the hyperexponential case, as expected, produces larger deviations. Note that for non-exponential service times the diffusion limit of the busy-server process is no longer a one-dimensional Ornstein–Uhlenbeck process [45]. However, the proposed formula still produces accurate results.

D. Comparison with Alternative Approaches

A first naive approach to approximate $\mathbb{E}[\tau^{(n)}]$ is to consider only its dominant term (in d) – given by $\frac{1}{\gamma}\log d$ by Theorem 2 – and neglecting all other lower-order terms. Thus, we let $T_{1\text{st}} \stackrel{\text{def}}{=} \frac{1}{\gamma}\log d$.

Another natural alternative to approximate $\mathbb{E}[\tau^{(n)}]$ (or its limit $\mathbb{E}[\tau]$) is to replace the moving barrier by a fixed constant level u. For a fixed barrier u, classical asymptotic results such as Pickands' theorem [21], [18] or heuristic approximations [40] can be applied to derive closed-form estimates of $\mathbb{E}[\tau]$.

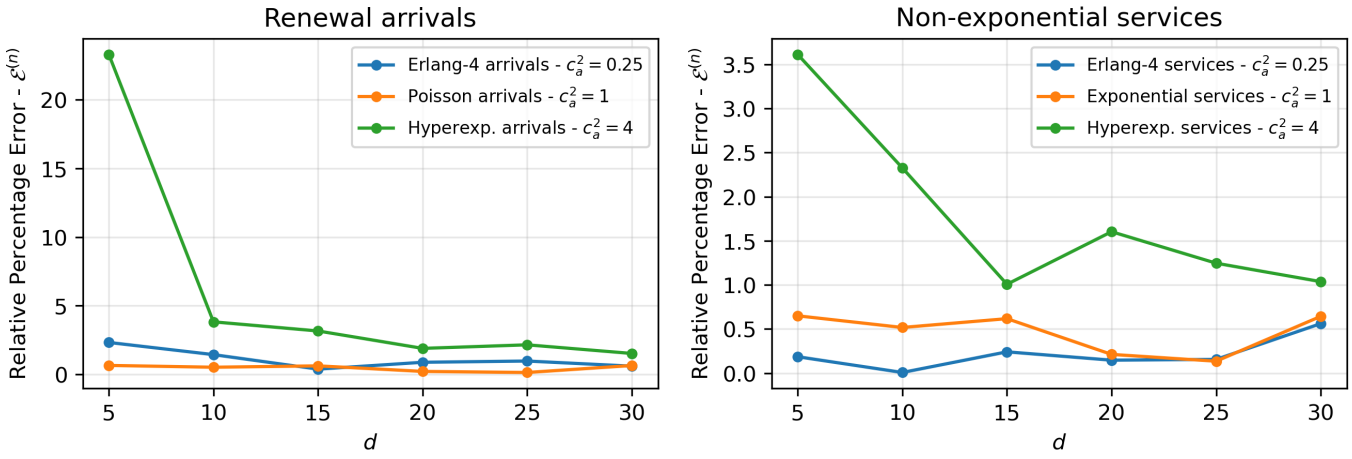


Figure 2: Robustness of the approximation (17). Left: renewal arrivals with different arrival variability c_a^2, using exponential service times. Right: non-exponential service times with mean one, using Poisson arrivals. In all cases, $n = 10^3$, $\lambda = 0.75$, $\gamma = 10^{-4}$ and $M = 500$. The arrival experiment uses the theoretically predicted variance parameter $\theta = \lambda(1 + c_a^2)$; the service-time experiment uses the plug-in value $\mu = 1/\mathbb{E}[S]$.

Here, one possibility is simply to set $u = d$ for all times, which may appear reasonable when γ is small. This yields highly inaccurate, unacceptable results (as the plot below will show). Another option is to choose $u = w(T^*)$, the level at which, under our approximation, the busy-server process is expected to meet the warm-server process.

We now compare our heuristic T^* with T_{1st}, as well as with the first-passage time $T_{w(T^*)}$ of the finite system when the warm-server process is artificially frozen at the level $w(T^*)$. Figure 3 illustrates the overall idea and shows a representative outcome of this comparison. From the figure, we first observe that T^* is substantially closer to τ than both $T_{w(T^*)}$ and T_{1st}. We emphasize that this behavior is typical: in the example shown, the relative percentage errors of T^* (reported in the plot titles) are fully consistent with the results in Figure 1. In addition, we observe that:

- The fixed-barrier approach suffers from a fundamental limitation that becomes more pronounced as d increases: the probability that X_t hits the constant level $w(T^*)$ well before time T^* becomes large;
- The first-order approximation T_{1st} substantially overestimates τ. Here, the uniformly bounded (in d) term $\frac{1}{2}\log\frac{\mu+\gamma}{\theta} - \log z$, equivalent to $T^* - T_{\mathrm{1st}}$, provides a significant and quite accurate correction.

VII. Conclusions

Using a queueing system to model a common autoscaling mechanism, we analyzed the time it takes to observe the next coldstart occurrence conditioned on the current state. At a diffusion scaling, we provided a simple approximation that is accurate in the parameter regime of autoscaling applications. This provided insights (see Remarks 1 and 2) about the influence of input parameters such as d and the idle-off rate γ, which are both under control of the system manager.

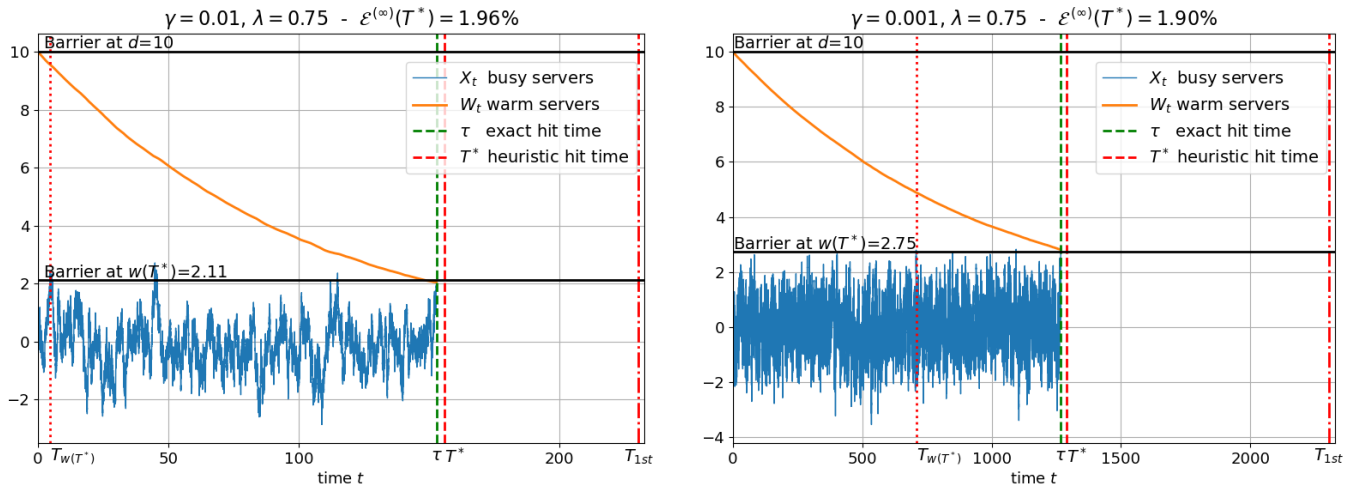
The diffusion approach followed in this paper provides a first analysis for understanding coldstart dynamics in highly reactive autoscaling systems. We outline the following research directions.

Idle-server timer monitoring and/or deterministic timers

Warm servers may dynamically inform the central controller about the state of their idle timers. Such information may be exploited to improve the estimation of the coldstart hitting time: for instance, if most servers are close to finishing their idle timers, the likelihood of observing a coldstart in the near future increases. Within our model, this idea can be incorporated by replacing the exponential idle timers with Erlang-distributed timers having k phases, each with rate γk. Each phase can be interpreted as a separate timer, and when a phase completes, the corresponding server notifies the controller. This leads naturally to a model of the form $(X^{(n)}, W^{(n)}, I_1^{(n)}, \ldots, I_k^{(n)})$, where $I_i^{(n)}$ represents the number of idle-on servers in phase i. Adapting the arguments used to prove Theorem 1, one can show convergence to a $(k+1)$-dimensional diffusion process in which the components of W are split across the k phases, and convergence of the mean coldstart hitting time. Then, the analysis and the heuristic of the diffusion–scaled coldstart hitting time need to be extended to a multidimensional setting.

Idle-server buffer

Another practically relevant extension is to introduce an idle-server buffer: instead of waiting until no idle-on server is available, the controller could trigger a scale-up action when the idle-on pool falls below a prescribed threshold. In the diffusion scaling, a buffer of size $b\sqrt{n}$ leads to the hitting time $\tau_b := \inf\{t \geq 0 : D_t \leq b\}$, $b > 0$, instead of the zero-level hitting time studied in this paper. The diffusion approximation itself extends directly, but the Rice-type approximation would need to be modified because the conditional infinitesimal drift and the Gaussian density must be evaluated at the non-zero boundary b.



(a) G001L75 (b) G01L75

Figure 3: Comparison of the exact coldstart hitting time τ with three approximations on a representative sample path of the busy-server and warm-server processes: the proposed heuristic T^*, the fixed-barrier approximation $T_{w(T^*)}$, and the first-order approximation $T_{1\mathrm{st}}$.

General service times

Our analysis builds on the convergence result developed by Whitt in [45], which applies more generally to phase-type service time distributions (say with k phases). In this case, under diffusion scaling, the diffusion limit becomes a k-dimensional O–U process. Mutatis mutandis, Theorem 1 remains valid under this generalization, provided that X is redefined as this multidimensional process. More precisely, in the equation (5) for W_t, the scalar X_t should be replaced by the L_1 norm of the components of the new X corresponding to jobs completing service after completing their current phase. This extension is not included to avoid overcomplicating the exposition, as the covariance matrix of the resulting O–U process is quite cumbersome.

Heterogeneous server parameters

The considered homogeneous model is a natural first-order abstraction at the level of a single serverless function or application, whose instances typically execute the same code and follow the same keep-alive policy. A heterogeneous model with class-dependent service rates μ_i and expiration rates γ_i would require tracking the busy and idle-on populations in each class. The scalar gap $D_t = W_t - X_t$ would then be replaced by a multidimensional diffusion, and the coldstart event would become a multidimensional boundary hitting problem.

Finite-capacity systems

The infinite-server assumption should be interpreted as a large-capacity approximation for the pre-coldstart dynamics. Initialization delays do not affect this time interval, since they start only after an arrival finds no idle-on server and triggers a coldstart. By contrast, a finite server pool or a concurrency limit would introduce an additional upper boundary for the busy-server process. For instance, if the maximum number of active servers is C_n and

$$\frac{C_n - n\rho}{\sqrt{n}} \to c \in (0, \infty],$$

then, at the diffusion scale, the process must be constrained by the upper level c. When $c = \infty$, or more generally when the upper boundary is large compared with the typical fluctuations reached before the coldstart, the present approximation is expected to remain accurate. When $c < \infty$ is close to the operating point, however, the coldstart time becomes a two-boundary first-passage problem: the idle gap D_t may hit zero, while the busy-server diffusion may also interact with the capacity boundary. This finite-capacity regime requires a different analysis and is left for future work.

REFERENCES

[1] Knative: Configuring scale bounds. https://knative.dev/docs/serving/autoscaling/scale-bounds/, 2025. Online; accessed: 2025-10-10.
[2] R. J. Adler and J. E. Taylor. Gaussian inequalities. In *Random Fields and Geometry*, Springer Monographs in Mathematics. Springer New York, New York, NY, 2007.
[3] L. Alili, P. Patie, and J. L. Pedersen. Representations of the first hitting time density of an ornstein-uhlenbeck process 1. *Stochastic Models*, 21(4):967–980, 2005.
[4] J. Anselmi. Asynchronous load balancing and auto-scaling: Mean-field limit and optimal design. *IEEE/ACM Transactions on Networking*, 32(4):2960–2971, 2024.
[5] J. Anselmi, B. Gaujal, and L.-S. Rebuffi. Non-stationary gradient descent for optimal auto-scaling in serverless platforms. *IEEE Transactions on Networking*, 33(4):1574–1587, 2025.
[6] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. Analysis of a multiserver queue with setup times. *Queueing Syst. Theory Appl.*, 51(1–2):53–76, Oct. 2005.
[7] R. Arumuganathan and S. Jeyakumar. Steady state analysis of a bulk queue with multiple vacations, setup times with n-policy and closedown times. *Applied Mathematical Modelling*, 29(10):972–986, 2005.
[8] J.-M. Azaïs and C. Mercadier. Asymptotic poisson character of extremes in non-stationary gaussian models. *Extremes*, 6(4):301–318, 2003.

[9] J.-M. Azaïs and M. Wschebor. *Level Sets and Extrema of Random Processes and Fields*. John Wiley & Sons, Hoboken, NJ, 2009.

[10] D. Bermbach, A.-S. Karakaya, and S. Buchholz. Using application knowledge to reduce cold starts in faas services. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 134–143, New York, NY, USA, 2020. Association for Computing Machinery.

[11] P. Billingsley. *Convergence of probability measures*. Wiley series in probability and statistics, 1999.

[12] A. Dassios and L. O. Li. Explicit asymptotics on first passage times of diffusion processes. *Advances in Applied Probability*, 52(3):681–704, 2020.

[13] C. Denninnart, T. Chanikaphon, and M. Amini Salehi. Efficiency in the serverless cloud paradigm: A survey on the reusing and approximation aspects. *Software: Practice and Experience*, 53(10):1853–1886, 2023.

[14] J. Dogani and F. Khunjush. Proactive auto-scaling technique for web applications in container-based edge computing using federated learning model. *Journal of Parallel and Distributed Computing*, 187:104837, 2024.

[15] S. W. Fuhrmann and R. B. Cooper. Stochastic decompositions in the m/g/1 queue with generalized vacations. *Operations Research*, 33(5):1117–1129, 1985.

[16] A. Gandhi, S. Doroudi, M. Harchol-Balter, and A. Scheller-Wolf. Exact analysis of the m/m/k/setup class of markov chains via recursive renewal reward. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '13, page 153–166, New York, NY, USA, 2013. Association for Computing Machinery.

[17] A. Gandhi and M. Harchol-Balter. M/G/k with staggered setup. *Operations Research Letters*, 41(4):317–320, July 2013.

[18] V. Giorno, A. G. Nobile, and L. M. Ricciardi. On the asymptotic behaviour of first-passage-time densities for one-dimensional diffusion processes and varying boundaries. *Advances in Applied Probability*, 22(4):883–914, 1990.

[19] A. W. Kambale, J. Anselmi, D. Ardagna, and B. Gaujal. Autoscaling in Serverless Platforms via Online Learning with Convergence Guarantees. working paper or preprint, 2025.

[20] T. Le-Anh and T. Phung-Duc. Energy-performance tradeoffs in server farms with batch services and setup times. *Performance Evaluation*, 168:102468, 2025.

[21] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1983.

[22] H. Levy and U. Yechiali. Utilization of idle time in an m/g/1 queueing system. *Management Science*, 22(2):202–211, 1975.

[23] Y. Li, Y. Lin, Y. Wang, K. Ye, and C.-Z. Xu. Serverless computing: State-of-the-art, challenges and opportunities. *IEEE Transactions on Services Computing*, 2022.

[24] Z. Li, L. L. Guo, J. Cheng, Q. Chen, B. He, and M. Guo. The serverless computing survey: A technical primer for design architecture. *ACM Computing Surveys (CSUR)*, 2021.

[25] P.-M. Lin and A. Glikson. Mitigating cold starts in serverless platforms: A pool-based approach. *arXiv preprint arXiv:1903.12221*, 2019.

[26] A. Lipton and V. Kaushansky. On the First Hitting Time Density of an Ornstein-Uhlenbeck Process. *arXiv preprint arXiv:1810.02390*, 2018.

[27] X. Liu, J. Wen, Z. Chen, D. Li, J. Chen, Y. Liu, H. Wang, and X. Jin. Faaslight: General application-level cold-start latency optimization for function-as-a-service in serverless computing. *ACM Trans. Softw. Eng. Methodol.*, 32(5), July 2023.

[28] Y. Lu. On a stein method based approximation for a two-dimensional markov chain, 2021.

[29] Y. Lu. Performance analysis of a queueing system with server arrival and departure. *SIGMETRICS Perform. Eval. Rev.*, 49(2):21–23, Jan. 2022.

[30] V. Maccio and D. Down. Structural properties and exact analysis of energy-aware multiserver queueing systems with setup times. *Performance Evaluation*, 121-122:48–66, 2018.

[31] N. Mahmoudi and H. Khazaei. Performance modeling of serverless computing platforms. *IEEE Transactions on Cloud Computing*, pages 1–1, 2020.

[32] R. J. Martin, M. J. Kearney, and R. V. Craster. Long- and short-time asymptotics of the first-passage time of the ornstein–uhlenbeck and other mean-reverting processes. *Journal of Physics A: Mathematical and Theoretical*, 52(13):134001, mar 2019.

[33] M. Nazari, S. Goodarzy, E. Keller, E. Rozner, and S. Mishra. Optimizing and extending serverless platforms: A survey. In *Proceedings of the Eighth International Conference on Software Deíned Systems (SDS)*, pages 1–8. IEEE, 2021.

[34] L. Ricciardi and S. Sato. A note on first passage time problems for gaussian processes and varying boundaries (corresp.). *IEEE Trans. Inf. Theor.*, 29(3):454–457, Sept. 2006.

[35] L. M. Ricciardi and L. Sacerdote. The ornstein-uhlenbeck process as a model for neuronal activity. *Biol. Cybern.*, 35(1):1–9, Mar. 1979.

[36] L. M. Ricciardi and S. Sato. First-passage-time density and moments of the ornstein-uhlenbeck process. *Journal of Applied Probability*, 25(1):43–57, 1988.

[37] S. Song, H. Tong, C. Meng, M. Pan, and Y. Yu. Funcscaler: Cold-start-aware holistic autoscaling for serverless resource management. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 1036–1047, 2024.

[38] Y. Sui, H. Yu, Y. Hu, J. Li, and H. Wang. Pre-warming is not enough: Accelerating serverless inference with opportunistic pre-loading. In *Proceedings of the 2024 ACM Symposium on Cloud Computing*, SoCC '24, page 178–195, New York, NY, USA, 2024. Association for Computing Machinery.

[39] M. Tamborrino. Approximation of the first passage time density of a wiener process to an exponentially decaying boundary by two-piecewise linear threshold. application to neuronal spiking activity. *Mathematical Biosciences & Engineering*, 13(3):613–629, 2016.

[40] M. U. Thomas. Some mean first-passage time approximations for the ornstein-uhlenbeck process. *Journal of Applied Probability*, 12(3):600–604, 1975.

[41] T. Tournaire, H. Castel-Taleb, and E. Hyon. Efficient computation of optimal thresholds in cloud auto-scaling systems. *ACM Trans. Model. Perform. Eval. Comput. Syst.*, 8(4), jul 2023.

[42] H. C. Tuckwell and F. Y. M. Wan. First-passage time of markov processes to moving barriers. *Journal of Applied Probability*, 21(4):695–709, 1984.

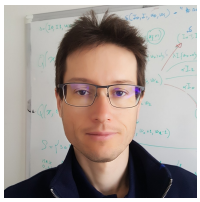
[43] L. Wang, M. Li, Y. Zhang, T. Ristenpart, and M. Swift. Peeking behind the curtains of serverless platforms. In *Proceedings of the 2018 USENIX Conference on Usenix Annual Technical Conference*, USENIX ATC '18, page 133–145, USA, 2018. USENIX Association.

[44] P. D. Welch. On a generalized m/g/1 queuing process in which the first customer of each busy period receives exceptional service. *Operations Research*, 12(5):736–752, Oct. 1964.

[45] W. Whitt. On the heavy-traffic limit theorem for $GI/G/\infty$ queues. *Advances in Applied Probability*, 14(1):171–190, 1982.

[46] J. Williams, M. Harchol-Balter, and W. Wang. M/M/k with deterministic setup times. *Proceedings of the ACM Measurement and Analysis of Computer Systems – SIGMETRICS*, 6(3):56:1–56:33, 2022. Sigmetrics 2023.

[47] C. Zeng. Mean exit time and escape probability for the Ornstein–Uhlenbeck process. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30(9):093127, 09 2020.



Jonatha Anselmi a tenured researcher at Inria, the French National Institute for Research in Digital Science and Technology, since 2014. He previously held research positions at the Basque Center for Applied Mathematics and Inria, after receiving his PhD in computer engineering from Politecnico di Milano in 2009. Working at the intersection of applied mathematics, computer science, and engineering, his research develops scalable methods for decision making under uncertainty, with a focus on reducing congestion and costs in large-scale systems.

APPENDIX A
PROOF OF THEOREM 1: FROM $\tau^{(n)}$ TO τ

Let us write τ instead of $\tau^{(\mathrm{stoch})}$. The proof is structured as follows. First, since the process $(X_t^{(n)})$ evolves autonomously, we can apply [45, Theorem 1] to obtain that $\overline{X}^{(n)} \Rightarrow X$ on $D[0,\infty)$. The upward jumps of $W_t^{(n)}$ are due to scale-up actions. In time interval $[0,\tau^{(n)})$, no scale-up actions are taken by definition of $\tau^{(n)}$. Therefore, the stopped version of the process $(X_t^{(n)}, W_t^{(n)})$ is identical to the stopped version of the same process where $W_t^{(n)}$, instead of being defined by (2), is the unique solution of the SDE

$$dW_t^{(n)} = -\gamma\,(W_t^{(n)} - X_t^{(n)})\,dt - dM_t^{(n)}, \tag{21}$$

with $M_t^{(n)}$ being the same martingale term; and in the remainder of the proof, $W_t^{(n)}$ will satisfy (21). We prove the joint convergence of the stopped process by showing (i) tightness of the pair $(\overline{X}^{(n)}, \overline{W}^{(n)})$, (ii) that any limit point $(\widetilde{X}, \widetilde{W})$ satisfies the desired limit equations, and (iii) that the limit equation for $\widetilde{W}$ has a unique solution given $\widetilde{X}$ and the initial condition; uniqueness gives convergence of the full sequence. Finally, we rely on the previous joint convergence to establish (iv) that $\tau^{(n)}$ converges weakly to τ.

Step 1: Tightness. Using the normalizations from (8) we obtain for every $t \geq 0$,

$$\overline{W}_t^{(n)} = \overline{W}_0^{(n)} - \gamma \int_0^t \big(\overline{W}_s^{(n)} - \overline{X}_s^{(n)}\big)^+ ds - \frac{1}{\sqrt{n}} M_t^{(n)}. \tag{22}$$

We claim that the martingale term vanishes uniformly on compact time intervals:

$$\frac{1}{\sqrt{n}} M^{(n)} \;\Rightarrow\; 0 \quad \text{in } D[0,\infty).$$

Its predictable quadratic variation satisfies for each fixed t

$$\langle M^{(n)}\rangle_t = \gamma \int_0^t (W_s^{(n)} - X_s^{(n)})^+ ds = \gamma\sqrt{n} \int_0^t (\overline{W}_s^{(n)} - \overline{X}_s^{(n)})^+ ds$$

Taking expectations in (22), since the centered initial conditions have uniformly bounded second moments and because W_t here is a pure-death counting process (non-increasing), we obtain $\mathbb{E}[\int_0^t (\overline{W}_s^{(n)} - \overline{X}_s^{(n)})^+ ds] = O(1)$, and thus $\mathbb{E}[\langle M^{(n)}\rangle_t] = O(\sqrt{n})$ for each fixed t. Consequently, by Chebyshev's inequality, Doob's maximal inequality and the martingale isometry,

$$\mathbb{P}\Big(\sup_{s\leq t} \big|\tfrac{1}{\sqrt{n}} M_s^{(n)}\big| > \varepsilon\Big) \leq \frac{1}{\varepsilon^2} \frac{\mathbb{E}[\sup_{s\leq t} |M_s^{(n)}|^2]}{n} \leq \frac{4}{\varepsilon^2} \frac{\mathbb{E}[|M_t^{(n)}|^2]}{n} = \frac{4}{\varepsilon^2} \frac{\mathbb{E}[\langle M^{(n)}\rangle_t]}{n} \longrightarrow 0$$

as $n \to \infty$. So, for any $T > 0$,

$$\sup_{t\leq T} \Big|\frac{1}{\sqrt{n}} M_t^{(n)}\Big| \;\xrightarrow[n\to\infty]{}\; 0 \quad \text{in probability.} \tag{23}$$

Now, we can show tightness of $(\overline{X}^{(n)}, \overline{W}^{(n)})$. Since tightness of $\{\overline{X}^{(n)}\}$ is given by [45], we show tightness of $\{\overline{W}^{(n)}\}$. Using $(a-b)^+ \leq |a| + |b|$ in (22) yields, for all $t \leq T$,

$$|\overline{W}_t^{(n)}| \leq |\overline{W}_0^{(n)}| + \gamma \int_0^t \big(|\overline{W}_s^{(n)}| + |\overline{X}_s^{(n)}|\big)\, ds + \sup_{u\leq T} \Big|\frac{1}{\sqrt{n}} M_u^{(n)}\Big|.$$

Taking the supremum over $t \leq T$ and applying Grönwall's inequality gives

$$\begin{aligned}\sup_{t\leq T} |\overline{W}_t^{(n)}| &\leq e^{\gamma T}\Big(|\overline{W}_0^{(n)}| + \gamma \int_0^T \sup_{u\leq s} |\overline{X}_u^{(n)}|\, ds + \sup_{u\leq T} \Big|\frac{1}{\sqrt{n}} M_u^{(n)}\Big|\Big) \\ &\leq e^{\gamma T}\Big(|\overline{W}_0^{(n)}| + \gamma T \sup_{u\leq T} |\overline{X}_u^{(n)}| + \sup_{u\leq T} \Big|\frac{1}{\sqrt{n}} M_u^{(n)}\Big|\Big).\end{aligned} \tag{24}$$

Since $\{\overline{W}_0^{(n)}\}$ is tight by assumption, $\{\sup_{u\leq T} |\overline{X}_u^{(n)}|\}$ is tight by tightness of $\overline{X}^{(n)}$, and the martingale term vanishes in probability by (23), it follows from (24) that $\{\sup_{t\leq T} |\overline{W}_t^{(n)}|\}$ is tight. To conclude tightness of the process family $\{\overline{W}^{(n)}\}$, in the following we develop a modulus-of-continuity argument. First, let us write (22) as $\overline{W}^{(n)} = \overline{W}_0^{(n)} + A^{(n)} - (1/\sqrt{n}) M^{(n)}$, where

$$A_t^{(n)} := -\gamma \int_0^t (\overline{W}_s^{(n)} - \overline{X}_s^{(n)})^+\, ds.$$

Then $A^{(n)}$ is absolutely continuous and, for $0 \le s < t \le T$,

$$|A_t^{(n)} - A_s^{(n)}| \le \gamma \int_s^t \big(|\overline{W}_u^{(n)}| + |\overline{X}_u^{(n)}|\big)\, du \le \gamma(t-s)\Big(\sup_{u\le T}|\overline{W}_u^{(n)}| + \sup_{u\le T}|\overline{X}_u^{(n)}|\Big). \tag{25}$$

Fix $\varepsilon > 0$. By tightness of $\overline{X}^{(n)}$ in $D[0,T]$, there exists a compact $K_X \subset D[0,T]$ such that $\inf_n \mathbb{P}(\overline{X}^{(n)} \in K_X) \ge 1-\varepsilon/4$. Since K_X is compact, $B_X := \sup_{x\in K_X}\|x\|_\infty < \infty$. By tightness of $\overline{W}_0^{(n)}$, let $B_0 < \infty$ be such that $\inf_n \mathbb{P}(|\overline{W}_0^{(n)}| \le B_0) \ge 1-\varepsilon/4$. Finally, by (23), pick n_ε such that for all $n \ge n_\varepsilon$, $\mathbb{P}(\sup_{t\le T}|M_t^{(n)}|/\sqrt{n} \le \varepsilon/4) \ge 1-\varepsilon/4$. On the event

$$E_n := \Big\{\overline{X}^{(n)} \in K_X,\ |\overline{W}_0^{(n)}| \le B_0,\ \sup_{t\le T}|M_t^{(n)}|/\sqrt{n} \le \varepsilon/4\Big\},$$

the sup–norm bound (24) implies $\sup_{t\le T}|\overline{W}_t^{(n)}| \le B_W$ for some deterministic $B_W < \infty$ (depending only on T, γ, B_0, B_X). Hence, for $n \ge n_\varepsilon$ and any $0 \le s < t \le T$ with $t - s \le \delta$, combining (22) and (25) yields on E_n,

$$|\overline{W}_t^{(n)} - \overline{W}_s^{(n)}| \le |A_t^{(n)} - A_s^{(n)}| + \Big|\frac{1}{\sqrt{n}}M_t^{(n)} - \frac{1}{\sqrt{n}}M_s^{(n)}\Big| \le \gamma\delta\,(B_W + B_X) + 2\sup_{u\le T}\Big|\frac{1}{\sqrt{n}}M_u^{(n)}\Big| \le \gamma\delta\,(B_W + B_X) + \varepsilon/2.$$

Now choose $\delta = \delta(\varepsilon)$ small enough so that $\gamma\delta(B_W + B_X) \le \varepsilon/2$. With these choices and for $n \ge n_\varepsilon$,

$$\mathbb{P}\Big(\omega_{\overline{W}^{(n)}}(\delta) \ge \varepsilon\Big) \le \mathbb{P}(E_n^c) \le \varepsilon,$$

where $\omega_f(\delta) := \sup\{|f(t) - f(s)| : 0 \le s < t \le T,\ t - s \le \delta\}$ is the (uniform) modulus of continuity. Together with the sup–norm tightness above, this implies that $\{\overline{W}^{(n)}\}$ is C-tight in $D[0,T]$ [11]. Here, note that Billingsley's tightness criterion in $D[0,T]$ is stated using the Skorokhod modulus $\omega'_f(\delta)$,. However, since for any càdlàg f and any $\delta > 0$ one has $\omega'_f(\delta) \le \omega_f(\delta)$ and therefore a bound on $\omega_{\overline{W}^{(n)}}(\delta)$ implies the required bound on $\omega'_{\overline{W}^{(n)}}(\delta)$. So, it suffices to bound the uniform modulus $\omega_f(\delta)$.

Step 2: Identification of limit points. Let $(\widetilde{X}, \widetilde{W})$ be any subsequential weak limit of $(\overline{X}^{(n)}, \overline{W}^{(n)})$ in $D[0,\infty) \times D[0,\infty)$. By the established convergence of the martingale term, passing to the limit in (22) yields, for every $t \ge 0$,

$$\widetilde{W}_t = \widetilde{W}_0 - \gamma \int_0^t \big(\widetilde{W}_s - \widetilde{X}_s\big)^+ ds, \qquad \text{a.s.}$$

Here, we have used the Skorokhod representation theorem to work almost surely along the subsequence and dominated convergence / continuity of the integral operator to pass the limit inside the integral; the vanishing martingale term contributes 0 in the limit. Thus, the limit obeys the *deterministic* integral equation

$$\widetilde{W}_t = \widetilde{W}_0 - \gamma \int_0^t (\widetilde{W}_s - \widetilde{X}_s)^+ ds, \qquad t \ge 0. \tag{26}$$

We therefore have shown that any limit point $(\widetilde{X}, \widetilde{W})$ satisfies $\widetilde{X} \overset{d}{=} X$ and $\widetilde{W}$ satisfies (26).

Step 3: Uniqueness. Fix $x \in D[0,\infty)$ and an initial value $w_0 \in \mathbb{R}$. Consider the map Ψ which to (x, w_0) associates the unique function $w \in C[0,\infty)$ solving

$$w(t) = w_0 - \gamma \int_0^t \big(w(s) - x(s)\big)^+ ds, \qquad t \ge 0.$$

A standard argument shows this solution exists, is unique, and depends continuously on (x, w_0) in the uniform-on-compact topology (consider two solutions, bound their diffenrence and use Gronwall's inequality to claim they coincide).

To conclude the proof, we use the continuous mapping theorem. Combining the previous steps, any subsequential limit $(\widetilde{X}, \widetilde{W})$ must satisfy $\widetilde{W} = \Psi(\widetilde{X}, \widetilde{W}_0)$ almost surely. Since Ψ is a deterministic and continuous map and $(\overline{X}^{(n)}, \overline{W}_0^{(n)}) \Rightarrow (X, W_0)$ with X the OU limit for the $\overline{X}^{(n)}$-coordinate, the continuous mapping theorem implies the joint convergence $(\overline{X}^{(n)}, \overline{W}^{(n)}) \Rightarrow (X, W)$, where $W = \Psi(X, W_0)$ is precisely the unique process satisfying (5).

Step 4: Convergence of $\tau^{(n)}$. Let $D_t^{(n)} \overset{\text{def}}{=} \overline{W}_t^{(n)} - \overline{X}_t^{(n)}$ and $D_t \overset{\text{def}}{=} W_t - X_t$. Here the continuous mapping theorem ensures that $D^{(n)} \Rightarrow D$ in $D[0,\infty)$. Now, define the functional $\Phi : D[0,\infty) \to [0,\infty]$, $\Phi(f) \overset{\text{def}}{=} \inf\{t \ge 0 : f(t) = 0\}$ with the convention $\inf \varnothing = +\infty$. We apply the continuous mapping theorem to Φ. The map Φ is not continuous on the whole space $D[0,\infty)$ but it is continuous at every function f that satisfies the following crossing property: there exists a time $\tau(f) = \Phi(f) < \infty$ such that

$$f(t) > 0 \ \text{ for } t \in [0, \tau(f)) \quad \text{and} \quad f(\tau(f)) = 0.$$

At such an f the map Φ is continuous with respect to uniform-on-compact convergence (hence also w.r.t. Skorokhod J_1 convergence because the limit is continuous on compacts). Now, we notice that the set of paths of D at which Φ is discontinuous

has probability zero. Indeed, $\mathbb{P}(\tau < \infty) = 1$, because (X_t, D_t) is a positive Harris recurrent Gaussian diffusion and therefore hits $\{D \le 0\}$ almost surely. Moreover, since the D-coordinate has a non-degenerate Brownian component, the boundary is crossed immediately after the hitting time, so tangential contacts have probability zero. Therefore, the continuous mapping theorem yields $\tau^{(n)} = \Phi(D^{(n)}) \Rightarrow \Phi(D) = \tau$, and hence $\tau^{(n)} \Rightarrow \tau$ in distribution.

The proof is thus concluded: the results above, together with the continuous mapping theorem and the almost sure continuity of (X, W), immediately imply the convergence of the stopped processes.

APPENDIX B
PROOF OF THEOREM 2: LOGARITHMIC GROWTH OF $\mathbb{E}[\tau]$

Let us recall that X denotes the Ornstein–Uhlenbeck process started from $X_0 = 0$

$$dX_t = -\mu X_t\, dt + \sqrt{\theta}\, dB_t, \tag{27}$$

D is the (unique strong) solution of the linear SDE

$$dD_t = \mu X_t\, dt - \gamma D_t\, dt - \sqrt{\theta}\, dB_t, \qquad D_0 = d > 0, \tag{28}$$

and

$$\tau := \inf\{t \ge 0 :\ D_t \le 0\}.$$

For convenience, in the following we let $t_d := \frac{1}{\gamma} \ln d$, so that $T^* = t_d + O(1)$.

Our proof proceeds by establishing, via different arguments, an upper and a lower bound on $\mathbb{E}[\tau]$ and then show that these two bounds coincide up to an additive constant. The upper bound follows from a Foster–Lyapunov contraction (plus a minorization/return argument) that controls the overshoot beyond t_d. The lower bound shows that τ cannot be much smaller than t_d and the key tool is the Borell–TIS inequality.

A. Upper Bound

We first show that as $d \to \infty$,

$$\mathbb{E}\big[\tau\big] \ \le\ \frac{1}{\gamma} \ln d \ +\ O(1), \tag{29}$$

where the $O(1)$ term depends only on (μ, γ, θ). Using the representation (56), (28) gives

$$D_t = e^{-\gamma t} d + Z_t, \qquad Z_t := \mu \int_0^t e^{-\gamma(t-s)} X_s\, ds - \sqrt{\theta} \int_0^t e^{-\gamma(t-s)}\, dB_s. \tag{30}$$

Using Itô isometry and $\mathbb{E}[X_s^2] = \frac{\theta}{2\mu}(1 - e^{-2\mu s}) \le \frac{\theta}{2\mu}$,

$$\begin{aligned}
\mathbb{E}[Z_t^2] &\le 2\mu^2\, \mathbb{E}\Big[\Big(\int_0^t e^{-\gamma(t-s)} X_s\, ds\Big)^2\Big] + 2\theta\, \mathbb{E}\Big[\Big(\int_0^t e^{-\gamma(t-s)}\, dB_s\Big)^2\Big] \\
&\le 2\mu^2 \Big(\int_0^t e^{-\gamma(t-s)}\, ds\Big)\Big(\int_0^t e^{-\gamma(t-s)} \mathbb{E}[X_s^2]\, ds\Big) + 2\theta \int_0^t e^{-2\gamma(t-s)}\, ds \\
&\le 2\mu^2 \cdot \frac{1}{\gamma} \cdot \Big(\frac{\theta}{2\mu}\, \frac{1}{\gamma}\Big) + 2\theta \cdot \frac{1}{2\gamma} \ =\ \theta\Big(\frac{\mu}{\gamma^2} + \frac{1}{\gamma}\Big) =: C_Z < \infty.
\end{aligned} \tag{31}$$

In particular, $\sup_{t \ge 0} \mathbb{E}[Z_t^2] \le C_Z$.

For some $c > 0$, in the following we will use

$$V(x, d) := x^2 + c d^2, \tag{32}$$

which will act as Lyapunov function.

Let us introduce the following two lemmas. The first lemma provides a Lyapunov drift for (X_t, D_t) with the quadratic function V, while the second builds on the first to give a structural bound on the hitting time τ.

Lemma 1. *Recall V is defined in* (32)*. There exists a constant $c > 0$ such that*

$$\mathcal{L}V(x, d) \le -\kappa V(x, d) + b \qquad \textit{for all } (x, d) \in \mathbb{R}^2, \tag{33}$$

for some $\kappa, b > 0$, where $\mathcal{L}$ is the generator of (X_t, D_t).

Proof. Using Itô's formula and a standard calculation, the generator of (X_t, D_t) acting on C^2 test functions $f = f(x, d)$ is

$$\mathcal{L}f(x, d) = (-\mu x)\partial_x f + (\mu x - \gamma d)\partial_d f + \frac{\theta}{2}\Big(\partial_{xx} f + \partial_{dd} f - 2\partial_{xd} f\Big). \tag{34}$$

Substituting the derivatives

$$\partial_x V = 2x, \quad \partial_{xx} V = 2, \quad \partial_d V = 2cd, \quad \partial_{dd} V = 2c, \quad \partial_{xd} V = 0$$

into (34) gives

$$\mathcal{L}V(x,d) = -2\mu x^2 + 2c\mu xd - 2c\gamma d^2 + \theta(1+c). \tag{35}$$

We dominate the cross-term $2c\mu xd$ by Young's inequality: using that $2|xd| \le \varepsilon x^2 + \varepsilon^{-1} d^2$ for any $\varepsilon > 0$, in (35) yields, for all (x,d),

$$\mathcal{L}V(x,d) \le -(2\mu - c\mu\varepsilon)x^2 - \big(2c\gamma - c\mu\varepsilon^{-1}\big)d^2 + \theta(1+c). \tag{36}$$

Now, letting $\varepsilon := 1/c$, (36) becomes

$$\mathcal{L}V(x,d) \le -\mu x^2 - \big(2c\gamma - c^2\mu\big)d^2 + \theta(1+c) \tag{37}$$

Then, we choose $c > 0$ small enough so that $2c\gamma - c^2\mu \ge c\gamma$, i.e.

$$2c\gamma - c^2\mu \;\ge\; c\gamma \quad\Longleftrightarrow\quad c(\gamma - c\mu) \ge 0 \quad\Longleftrightarrow\quad c \le \gamma/\mu.$$

For instance, we can take

$$c := \frac{\gamma}{2\mu}. \tag{38}$$

Then, (37) implies

$$\mathcal{L}V(x,d) \le -\mu x^2 - c\gamma d^2 + \theta(1+c). \tag{39}$$

Finally, since $V(x,d) = x^2 + cd^2$, we have

$$\mu x^2 + c\gamma d^2 \;\ge\; \min\{\mu,\gamma\}\,(x^2 + cd^2) = \kappa\, V(x,d), \qquad \kappa := \min\{\mu,\gamma\}.$$

Combining with (39) yields the desired Foster–Lyapunov drift in (33):

$$\mathcal{L}V(x,d) \le -\kappa V(x,d) + b, \qquad \kappa := \min\{\mu,\gamma\}, \quad b := \theta(1+c), \tag{40}$$

valid for all $(x,d) \in \mathbb{R}^2$, with c given by (38). □

Lemma 2. *There exists constant $C < \infty$ that depends only on μ, γ and θ such that*

$$\mathbb{E}_{x,d}[\tau] \;\le\; C\big(1 + V(x,d)\big), \qquad (x,d) \in \mathbb{R} \times [0,\infty). \tag{41}$$

Proof. Let $Y_t := (X_t, D_t)$. Let $R > b/\kappa$, $K_R := \{(x,d) \in \mathbb{R} \times [0,\infty) : V(x,d) \le R\}$ and $\tau_R := \inf\{t \ge 0 :\ Y_t \in K_R\}$. On $\{t < \tau_R\}$ we have $V(Y_t) > R$, hence, using the drift condition (33),

$$\mathcal{L}V(Y_t) \le -\kappa R + b = -\delta$$

where $\delta := \kappa R - b > 0$. Dynkin's formula applied to $V(Y_{t\wedge\tau_R})$ yields

$$0 \le \mathbb{E}_{x,d}[V(Y_{t\wedge\tau_R})] \le V(x,d) - \delta\, \mathbb{E}_{x,d}[t \wedge \tau_R],$$

so letting $t \to \infty$ gives

$$\mathbb{E}_{x,d}[\tau_R] \le \delta^{-1} V(x,d). \tag{42}$$

Next, we show a one-step minorization from K_R into the set $\{d \le 0\}$. Fix any $T > 0$. By linearity/Gaussianity, for each initial (x,d), D_T is Gaussian with mean $\mathbb{E}_{x,d}[D_T]$ (which is affine in (x,d)) and variance $\sigma_T^2 := \mathrm{Var}(D_T) > 0$ depending only on (μ,γ,θ,T). Since K_R is compact, $M_R := \sup_{(x,d)\in K_R} \mathbb{E}_{x,d}[D_T] < \infty$, hence

$$p := \inf_{(x,d)\in K_R} \mathbb{P}_{x,d}(D_T \le 0) \;\ge\; \Phi\!\left(\frac{-M_R}{\sigma_T}\right) \;>\; 0, \tag{43}$$

where Φ is the standard normal cdf. Since $\{D_T \le 0\} \subseteq \{\tau \le T\}$, (43) implies

$$\inf_{y\in K_R} \mathbb{P}_y(\tau \le T) \ge p. \tag{44}$$

We now show that

$$M := \sup_{y\in K_R} \mathbb{E}_y[\tau] \in [0,\infty] < \infty.$$

Towards this purpose, for $n \in \mathbb{N}$ let

$$M_n := \sup_{y\in K_R} \mathbb{E}_y[\tau \wedge n] \in [0,n],$$

and for $y \in K_R$ let

$$\rho := \inf\{t \geq 0 : \ Y_{T+t} \in K_R\}$$

be the (post-T) return time to K_R. Using the strong Markov property at time T and then at time $T+\rho$,

$$\begin{aligned}\mathbb{E}_y[\tau \wedge n] &\leq \mathbb{E}_y\big[(\tau \wedge n) \wedge T\big] + \mathbb{E}_y\big[((\tau \wedge n) - T)^+\big] = \mathbb{E}_y\big[(\tau \wedge n) \wedge T\big] + \mathbb{E}_y\Big[\mathbf{1}_{\{\tau > T\}} \mathbb{E}_{Y_T}[\tau \wedge n]\Big] \\ &\leq T + \mathbb{E}_y\Big[\mathbf{1}_{\{\tau > T\}} \big(\rho + \mathbb{E}_{Y_{T+\rho}}[\tau \wedge n]\big)\Big] \\ &\leq T + \mathbb{E}_y\big[\rho\big] + (1-p)\, M_n, \end{aligned} \tag{45}$$

since on $\{\tau > T\}$ we have $Y_{T+\rho} \in K_R$ and $\mathbb{E}_{Y_{T+\rho}}[\tau \wedge n] \leq M_n$, and also $\mathbb{P}_y(\tau > T) \leq 1-p$ for $y \in K_R$ (by (44)). It remains to bound $\sup_{y \in K_R} \mathbb{E}_y[\rho]$. Conditioning on Y_T and applying (42) gives

$$\mathbb{E}_y[\rho] = \mathbb{E}_y\big[\mathbb{E}_{Y_T}[\tau_R]\big] \leq \delta^{-1}\, \mathbb{E}_y\big[V(Y_T)\big], \qquad y \in K_R. \tag{46}$$

Applying Dynkin's formula to the exponentially-weighted[1] process $e^{\kappa t} V(Y_t)$ and using (33), we obtain for any deterministic $T > 0$,

$$\mathbb{E}_y\big[e^{\kappa T} V(Y_T)\big] \leq V(y) + b \int_0^T e^{\kappa s}\, ds = V(y) + \frac{b}{\kappa}\big(e^{\kappa T} - 1\big),$$

which implies, for $y \in K_R$,

$$\mathbb{E}_y[V(Y_T)] \leq e^{-\kappa T} V(y) + \frac{b}{\kappa}\big(1 - e^{-\kappa T}\big) \leq e^{-\kappa T} V(y) + \frac{b}{\kappa} \leq e^{-\kappa T} R + \frac{b}{\kappa} =: \bar{V}_T < \infty.$$

Therefore, $\sup_{y \in K_R} \mathbb{E}_y[\rho] \leq \delta^{-1} \bar{V}_T < \infty$ (see (46)), and taking the supremum over $y \in K_R$ in (45) yields, for every n,

$$M_n \leq T + \delta^{-1} \bar{V}_T + (1-p) M_n \quad \Longrightarrow \quad M_n \leq \frac{T + \delta^{-1} \bar{V}_T}{p}.$$

By monotone convergence, $\mathbb{E}_y[\tau \wedge n] \uparrow \mathbb{E}_y[\tau]$ for each y, hence

$$M = \sup_{y \in K_R} \mathbb{E}_y[\tau] = \sup_{y \in K_R} \lim_{n \to \infty} \mathbb{E}_y[\tau \wedge n] \leq \lim_{n \to \infty} M_n \leq \frac{T + \delta^{-1} \bar{V}_T}{p} < \infty.$$

Finally, for arbitrary (x, d), decomposing at τ_R we obtain

$$\mathbb{E}_{x,d}[\tau] = \mathbb{E}_{x,d}[\tau_R] + \mathbb{E}_{x,d}\Big[\mathbb{E}_{Y_{\tau_R}}[\tau]\Big] \leq \mathbb{E}_{x,d}[\tau_R] + \sup_{y \in K_R} \mathbb{E}_y[\tau] = \mathbb{E}_{x,d}[\tau_R] + M,$$

and together with (42), the finiteness of M gives

$$\mathbb{E}_{x,d}[\tau] \leq \delta^{-1} V(x,d) + M \leq C\big(1 + V(x,d)\big),$$

for a finite constant C depending only on (μ, γ, θ). □

Now, to prove the upper bound (29), it is enough to control the right tail of τ around a logarithmic time scale. By the Markov property at time t_d and because $(\tau - t_d)^+ = \mathbf{1}_{\{\tau > t_d\}} \inf\{s \geq 0 : D_{t_d+s} \leq 0\}$, it holds that

$$\mathbb{E}\big[(\tau - t_d)^+\big] \leq \mathbb{E}\Big[\mathbb{E}_{X_{t_d},\, D_{t_d}}[\tau]\Big]$$

and using (41) gives

$$\mathbb{E}\big[(\tau - t_d)^+\big] \leq C\Big(1 + \mathbb{E}\big[V(X_{t_d}, D_{t_d})\big]\Big). \tag{47}$$

By (30), $D_{t_d} = 1 + Z_{t_d}$, hence

$$\sup_{d \geq 1} \mathbb{E}\big[(D_{t_d})^2\big] \leq 2 + 2 \sup_{t \geq 0} \mathbb{E}[Z_t^2] \leq 2 + 2C_Z, \qquad \sup_{d \geq 1} \mathbb{E}[X_{t_d}^2] \leq \sup_{t \geq 0} \mathbb{E}[X_t^2] \leq \frac{\theta}{2\mu}.$$

where C_Z was defined in (31). Therefore $\sup_{d \geq 1} \mathbb{E}[V(X_{t_d}, D_{t_d})] < \infty$, and (47) yields

$$\sup_{d \geq 1} \mathbb{E}\big[(\tau - t_d)^+\big] < \infty. \tag{48}$$

Finally, since $\tau \leq t_d + (\tau - t_d)^+$, we obtain (29).

[1]We could apply it to $V(Y_t)$ as well but then we would need to use a further argument involving Grönwall's inequality.

B. Lower Bound

We now prove

$$\mathbb{E}[\tau] \;\geq\; \frac{1}{\gamma}\ln d \;+\; O(1). \tag{49}$$

We do this by showing that $\sup_{d\geq 1}\mathbb{E}[(t_d-\tau)^+]<\infty$, and our strategy is as follows:

- Show that if τ occurs earlier than t_d-u, then Z (see (30)) must make a downward excursion of size at least $e^{\gamma u}$.
- Partition the time interval backward from t_d into unit blocks.
- Apply Borell–TIS uniformly on each block and sum the resulting super-exponential tail bounds.
- Integrate the tail to bound $\mathbb{E}[(t_d-\tau)^+]$.

Let $u\geq 0$. On the event $\{\tau\leq t_d-u\}$, we have $\inf_{0\leq t\leq t_d-u} D_t\leq 0$, hence there exists $t\leq t_d-u$ such that $D_t = e^{-\gamma t}d+Z_t\leq 0$. Since $t\mapsto e^{-\gamma t}d$ is decreasing,

$$e^{-\gamma t}d \;\geq\; e^{-\gamma(t_d-u)}d \;=\; e^{\gamma u},$$

and therefore $Z_t\leq -e^{\gamma u}$. Consequently,

$$\{\tau\leq t_d-u\}\subseteq\Big\{\inf_{0\leq t\leq t_d-u} Z_t\leq -e^{\gamma u}\Big\}. \tag{50}$$

Now, the idea is to apply Borell–TIS inequality. To do this, we partition $[0,T]$ into unit subintervals applying Borell–TIS on each block $[k,k+1]$ backwards from t_d, and then we combine the block estimates via a union bound.

For $s\geq 0$ define $Z_t^{(s)}:=Z_{s+t}$ on $t\in[0,1]$, and set

$$m(s):=\mathbb{E}\Big[\sup_{t\in[0,1]}|Z_{s+t}|\Big],\qquad m_*:=\sup_{s\geq 0} m(s).$$

Lemma 3 (Uniform bound on $m(s)$ and uniform Borell–TIS)**.** *There exists $m_*<\infty$ depending only on (μ,γ,θ) such that $m(s)\leq m_*$ for all $s\geq 0$. Moreover, for all $s\geq 0$ and all $r\geq 0$,*

$$\mathbb{P}\left(\sup_{t\in[0,1]}|Z_{s+t}|\geq m_*+r\right)\leq 2\exp\left(-\frac{r^2}{2C_Z}\right). \tag{51}$$

Proof. First, note that Z satisfies the linear SDE

$$dZ_t=\mu X_t\,dt-\gamma Z_t\,dt-\sqrt{\theta}\,dB_t,\qquad Z_0=0,$$

hence for $t\in[0,1]$,

$$Z_{s+t}=e^{-\gamma t}Z_s+\mu\int_0^t e^{-\gamma(t-r)}X_{s+r}\,dr-\sqrt{\theta}\int_0^t e^{-\gamma(t-r)}\,dB_{s+r}.$$

Taking sup over $t\in[0,1]$ and using $e^{-\gamma(t-r)}\leq 1$ gives

$$\sup_{t\in[0,1]}|Z_{s+t}|\leq|Z_s|+\mu\int_0^1|X_{s+r}|\,dr+\sqrt{\theta}\sup_{t\in[0,1]}\left|\int_0^t e^{-\gamma(t-r)}\,dB_{s+r}\right|.$$

Taking expectations, Cauchy–Schwarz, $\sup_{t\geq 0}\mathbb{E}[Z_t^2]\leq C_Z$ from (31), and $\sup_{t\geq 0}\mathbb{E}[X_t^2]\leq\theta/(2\mu)$ yield

$$m(s)\leq\sqrt{C_Z}+\mu\int_0^1\sqrt{\mathbb{E}[X_{s+r}^2]}\,dr+\sqrt{\theta}\,\mathbb{E}\sup_{t\in[0,1]}\left|e^{-\gamma t}\int_0^t e^{\gamma r}\,dB_{s+r}\right|.$$

Let $\widetilde{B}_t:=B_{s+t}-B_s$, so that $(\widetilde{B}_t)_{t\geq 0}$ is a standard Brownian motion and

$$\int_0^t e^{\gamma r}\,dB_{s+r}=\int_0^t e^{\gamma r}\,d\widetilde{B}_r=:M_t,\qquad t\in[0,1].$$

Then $(M_t)_{t\in[0,1]}$ is a continuous martingale with quadratic variation $\langle M\rangle_t=\int_0^t e^{2\gamma r}\,dr$. By the Burkholder-Davis-Gundy (BDG) inequality,

$$\mathbb{E}\sup_{t\in[0,1]}\left|\int_0^t e^{\gamma r}\,dB_{s+r}\right|=\mathbb{E}\sup_{t\in[0,1]}|M_t|\leq C_{\mathrm{BDG}}\,\mathbb{E}\big[\langle M\rangle_1^{1/2}\big]=C_{\mathrm{BDG}}\left(\int_0^1 e^{2\gamma r}\,dr\right)^{1/2},$$

where C_{BDG} is universal. Hence $\sup_{s\geq 0}m(s)\leq m_*<\infty$ with m_* depending only on (μ,γ,θ).

For the tail bound, let us fix $s \geq 0$. Applying Borell–TIS [2, Theorem 2.1.1] to the centered Gaussian field $\{(\varepsilon, t) \mapsto \varepsilon Z_{s+t} : \varepsilon \in \{\pm 1\},\, t \in [0,1]\}$ yields, for all $r \geq 0$,

$$\mathbb{P}\left(\sup_{t\in[0,1]} |Z_{s+t}| \geq m(s) + r \right) \leq \exp\left(-\frac{r^2}{2\sigma_s^2} \right), \qquad \sigma_s^2 := \sup_{t\in[0,1]} \mathrm{Var}(Z_{s+t}).$$

Since Z is centered and $\sup_{t\geq 0} \mathrm{Var}(Z_t) = \sup_{t\geq 0} \mathbb{E}[Z_t^2] \leq C_Z$, we have $\sigma_s^2 \leq C_Z$ for all s. Finally, $m(s) \leq m_*$ implies

$$\Big\{ \sup_{t\in[0,1]} |Z_{s+t}| \geq m_* + r \Big\} \subseteq \Big\{ \sup_{t\in[0,1]} |Z_{s+t}| \geq m(s) + r \Big\},$$

and (51) follows. □

Therefore, for all $s \geq 0$ and all $a \geq m_*$, (51) implies

$$\mathbb{P}\left(\inf_{t\in[s,s+1]} Z_t \leq -a \right) \leq \mathbb{P}\left(\sup_{t\in[s,s+1]} |Z_t| \geq a \right) \leq 2\exp\left(-\frac{(a-m_*)^2}{2C_Z} \right). \tag{52}$$

Now we partition backwards from t_d and combine the block estimates via a union bound. Since

$$\{\tau \leq t_d - u\} \subset \bigcup_{\lfloor u \rfloor \leq j \leq \lfloor t_d \rfloor} \left\{ \inf_{t\in[t_d-(j+1),\, t_d-j]\cap[0,\infty)} Z_t \leq -e^{\gamma j} \right\}. \tag{53}$$

applying (52) to each unit interval $[t_d - (j+1),\, t_d - j]$ with $a = e^{\gamma j}$ gives, for all j large enough,

$$\sup_{d\geq 1} \mathbb{P}\left(\inf_{t\in[t_d-(j+1),\, t_d-j]} Z_t \leq -e^{\gamma j} \right) \leq 2\exp\left(-\frac{(e^{\gamma j} - m_*)^2}{2C_Z} \right) \leq 2\exp\big(-c\, e^{2\gamma j} \big),$$

for some $c > 0$ depending only on (μ, γ, θ). Combining with (53) yields the uniform tail bound

$$\sup_{d\geq 1} \mathbb{P}(\tau \leq t_d - u) \leq \sum_{j \geq \lfloor u \rfloor} 2\exp\big(-c\, e^{2\gamma j} \big), \qquad u \geq 0. \tag{54}$$

Finally,

$$\sup_{d\geq 1} \mathbb{E}[(t_d - \tau)^+] = \sup_{d\geq 1} \int_0^\infty \mathbb{P}(\tau \leq t_d - u)\, du \leq \sum_{k=0}^{\infty} \sup_{d\geq 1} \mathbb{P}(\tau \leq t_d - k),$$

and by (54),

$$\sup_{d\geq 1} \mathbb{P}(\tau \leq t_d - k) \leq \sum_{j \geq k} 2\exp\big(-c\, e^{2\gamma j} \big).$$

Therefore,

$$\sup_{d\geq 1} \mathbb{E}[(t_d - \tau)^+] \leq \sum_{k=0}^{\infty} \sum_{j\geq k} 2\exp\big(-c\, e^{2\gamma j} \big) = \sum_{j=0}^{\infty} 2(j+1)\exp\big(-c\, e^{2\gamma j} \big) < \infty,$$

since $\exp(-ce^{2\gamma j})$ decays super-exponentially in j. This proves $\mathbb{E}[\tau] \geq t_d - O(1)$.

APPENDIX C
ADDITIONAL PROOFS

A. *Proof of Integral Representation* (7)*:* D_t

Let us provide explicit representations for the processes $D_t = W_t - X_t$. First, it is well-known that the general solution of X_t is

$$X_s = x_0 e^{-\mu s} + \sqrt{\theta} \int_0^s e^{-\mu(s-u)} dB_u \tag{55}$$

A direct calculation yields

$$D_t = e^{-\gamma t} D_0 + \mu \int_0^t e^{-\gamma(t-s)} X_s ds - \sqrt{\theta} \int_0^t e^{-\gamma(t-s)} dB_s. \tag{56}$$

$$\mathbb{E}[D_t] = e^{-\gamma t} D_0 + \mu x_0 e^{-\gamma t} \frac{1 - e^{-(\mu-\gamma)t}}{\mu - \gamma},$$

Substituting the explicit O–U formula (55) into the previous integral representation, we get the decomposition into a deterministic term and a single stochastic convolution in (7).

B. Proof of Proposition 1

For $f(x) = x$, Dynkin's formula gives, for any x,

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

Evaluating at $x = 0$ yields

$$\mathcal{L}_t f(0) = \mu\, \mathbb{E}[X_t \mid D_t = 0].$$

Both X_t and D_t are Gaussian under the linear dynamics. Use the integral representations

$$X_t = \sqrt{\theta} \int_0^t e^{-\mu(t-u)}\, dB_u, \qquad D_t = e^{-\gamma t} d + \sqrt{\theta} \int_0^t k(t,u)\, dB_u,$$

where

$$k(t,u) = \frac{\gamma}{\mu - \gamma} e^{-\gamma(t-u)} - \frac{\mu}{\mu - \gamma} e^{-\mu(t-u)}, \qquad 0 \le u \le t.$$

The Gaussian regression formula gives

$$\mathbb{E}[X_t \mid D_t = 0] = \mathbb{E}[X_t] + \frac{\mathrm{Cov}(X_t, D_t)}{\mathrm{Var}(D_t)} \big(0 - \mathbb{E}[D_t]\big) = -e^{-\gamma t} d\, \frac{\mathrm{Cov}(X_t, D_t)}{\mathrm{Var}(D_t)},$$

because $\mathbb{E}[X_t] = 0$ and $\mathbb{E}[D_t] = e^{-\gamma t} d$. Hence

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

Now, using

$$\int_0^t e^{-a(t-u)} e^{-b(t-u)}\, du = \frac{1 - e^{-(a+b)t}}{a+b},$$

one obtains after elementary algebra

$$\mathrm{Cov}(X_t, D_t) = \frac{\theta}{\mu - \gamma} \left(\gamma \frac{1 - e^{-(\mu+\gamma)t}}{\mu + \gamma} - \tfrac{1}{2}\big(1 - e^{-2\mu t}\big) \right),$$

$$\mathrm{Var}(D_t) = \frac{\theta}{(\mu - \gamma)^2} \left(\frac{\gamma}{2}\big(1 - e^{-2\gamma t}\big) - \frac{2\gamma\mu}{\gamma + \mu}\big(1 - e^{-(\gamma+\mu)t}\big) + \frac{\mu}{2}\big(1 - e^{-2\mu t}\big) \right).$$

Substituting these expressions into the formula for $\mathcal{L}_t f(0)$ yields (12).

C. Proof of Proposition 2: Rice Solution

Let $\sigma^2 := \frac{\theta}{2(\mu+\gamma)}$. Let us make the change of variable

$$y := \frac{w(t)^2}{2\sigma^2} = \frac{d^2 e^{-2\gamma t}}{2\sigma^2}, \qquad A := \frac{d^2}{2\sigma^2}.$$

Then $y = A e^{-2\gamma t}$, $dy = -2\gamma y\, dt$, $dt = -\frac{dy}{2\gamma y}$. Since $w(t) = \sqrt{2\sigma^2 y}$ and $e^{-w^2/(2\sigma^2)} = e^{-y}$, the integrand of (14) becomes

$$\mu w(t)\, \frac{1}{\sqrt{2\pi}\sigma}\, e^{-w(t)^2/(2\sigma^2)} = \mu \frac{\sqrt{2\sigma^2 y}}{\sqrt{2\pi}\sigma} e^{-y}\, dt = -\frac{\mu}{2\gamma\sqrt{\pi}}\, y^{-1/2} e^{-y}\, dy.$$

Therefore

$$\int_0^T \mu w(t) \frac{1}{\sqrt{2\pi}\sigma} e^{-w(t)^2/(2\sigma^2)}\, dt = \frac{\mu}{2\gamma\sqrt{\pi}} \int_{A e^{-2\gamma T}}^{A} y^{-1/2} e^{-y}\, dy.$$

Using the relation between the incomplete gamma and the error function $\Gamma(\frac{1}{2}, y) = \sqrt{\pi}\, \mathrm{erfc}(\sqrt{y})$, we obtain

$$\int_u^A y^{-1/2} e^{-y}\, dy = \Gamma(\tfrac{1}{2}, u) - \Gamma(\tfrac{1}{2}, A) = \sqrt{\pi}\, \big(\mathrm{erfc}(\sqrt{u}) - \mathrm{erfc}(\sqrt{A}) \big)$$

Thus our equation boils down to

$$\frac{\mu}{\gamma} \big(\mathrm{erfc}(\sqrt{A} e^{-\gamma T}) - \mathrm{erfc}(\sqrt{A}) \big) = 1.$$

Solving for $\mathrm{erfc}(\sqrt{A} e^{-\gamma T})$ gives

$$\mathrm{erfc}\big(\sqrt{A} e^{-\gamma T}\big) = \frac{\gamma}{\mu} + \mathrm{erfc}(\sqrt{A}).$$

Assumption (16) ensures that we can invert the erfc function. Recalling $\sqrt{A} = d\sqrt{(\mu+\gamma)/\theta}$, we obtain

$$\sqrt{A}\, e^{-\gamma T} = z, \qquad z := \operatorname{erfc}^{-1}\left(\frac{\gamma}{\mu} + \operatorname{erfc}(d\sqrt{(\mu+\gamma)/\theta})\right)$$

Hence

$$e^{-\gamma T} = \frac{z}{d\sqrt{(\mu+\gamma)/\theta}} \quad \Longrightarrow \quad T = \frac{1}{\gamma}\Big(\ln d + \tfrac{1}{2}\ln\frac{\mu+\gamma}{\theta} - \ln z\Big).$$

For uniqueness, the integrand is nonnegative and continuous in t; moreover it is strictly positive for all finite t. Hence the left-hand side $F(T) = \int_0^T (\cdots)\, dt$ is continuous and strictly increasing. Since the argument of erfc^{-1} belongs to $(0, 1)$, we have $z > 0$, and the solution displayed above is the unique finite solution of $F(T^\star) = 1/2$.