Automatic cloud instance provisioning with quality and efficiency

Diego Goldsztajn[a,1], Andrés Ferragut[b,*], Fernando Paganini[b]

[a] *Eindhoven University of Technology, Eindhoven, The Netherlands*
[b] *Universidad ORT Uruguay, Montevideo, Uruguay*

Abstract

A distinctive feature of cloud computing is that it enables customers to dynamically summon server instances. Service providers facing uncertain demand patterns may exploit this feature by setting automatic provisioning rules for right-sizing the capacity contracted from the cloud. This situation can be modeled by a queueing system where the numbers of both jobs and servers evolve in time, the latter subject to delays in creation and deletion. We study in this context different feedback rules with the objective of efficiently matching capacity and load, while simultaneously providing a high quality of service.

These rules are analyzed by means of fluid and diffusion limits for Markov chains. In particular we develop suitable extensions of the classical literature on this topic, required to accommodate non-homogeneous intensity scalings and non-differentiable drift fields. With these tools, our final proposal is shown to exhibit properties akin to the Halfin-Whitt regime, achieved automatically without knowledge of the system load. We further investigate by simulation its behavior under time-varying load, demonstrating the capabilities of our design to provide quality and efficiency in highly dynamic scenarios.

Keywords: cloud computing, right-sizing, fluid and diffusion limits,

1. Introduction

The cloud computing paradigm has significantly changed the way computing services are conceived and deployed. A service provider facing an uncertain demand need no longer invest on a proprietary infrastructure, with the ensuing risk in choosing its appropriate size. Instead, cloud services provide the alternative of an essentially unlimited infrastructure, which can be contracted dynamically as demand arises.

Under idealized circumstances, such a system could be modeled as an infinite-server queue, in which active instances track instantaneously the current job

*Corresponding Author. *Email address:* ferragut@ort.edu.uy
[1] Research conducted when the author was at Universidad ORT Uruguay.

population. There are, however, delays in server creation and deletion, which are non-negligible with respect to the time-scale of jobs. Thus, a model of the resulting queueing system must distinguish the quantities of jobs and servers, and explicitly consider the rules for server creation and deletion, to achieve an efficient use of the infrastructure while guaranteeing quality of service. The analysis and design of such rules is the subject of this paper.

If properly controlled, such a system should operate in the "heavy-traffic" regime [1]; i.e., with load approximately matching capacity. In contrast with most of the literature on this topic, we do not assume heavy-traffic occurs through an exogenous tuning of the load parameter to approach a fixed capacity (a tuning which is typically not given a practical motivation); rather, it is the automatic control of the *capacity* side that provides the "right-sizing" [2] to match the load. This explanatory mechanism for heavy-traffic cannot be studied with tools of fixed server queues: instead, we use a two-dimensional Markov model where the number of servers becomes a state variable, in addition to the number of jobs; the former is subject to exponential startup/deletion times. A mismatch between both populations indicates loss of efficiency (overprovisioned servers) or quality of service (jobs being queued).

We use fluid and diffusion limits to analyze the large-scale behavior of these Markov processes, under different control rules; the analysis requires extensions to the standard theory of [3]. We first consider in Section 2 a basic model which modifies the $M/M/\infty$ queue to incorporate the server dynamics; the resulting fluid model is shown to globally converge to the natural equilibrium, and its diffusion scale approximation exhibits fluctuations in both the overprovisiong and queueing directions. In Section 3 we introduce a bias in the server provisioning control, favoring quality of service (zero queueing). A bias proportional to the current population can be tuned to achieve quality of service with small impact on overprovisiong; however this tuning would depend on the load ρ.

In Section 4 the bias is modified to be proportional to the square-root of the job population; in this way a universal tuning is possible across a wide variety of loads. The steady-state behavior of this system is consistent with the *quality and efficiency driven* (QED) regime of Halfin and Whitt [4]; in particular with $O(\sqrt{\rho})$ idle servers. The distinctive contribution is that we provide a mechanism to reach this operating regime automatically through feedback.

In Section 5 the performance of these methods is demonstrated through extensive simulations. In particular, we validate empirically the behavior of our control rules when subject to non-stationary traffic loads.

Our main contributions are:

- From an applications perspective, we provide control rules for summoning and deleting server instances from a cloud service provider, in order to track an uncertain or time-varying demand. Of the different design choices, we favor rules deliberately biased to guarantee no queueing delay is incurred with high probability, while keeping the overprovisioning level at a minimum. We show this can be achieved in a manner that self-scales to the exogenous load.

- From a theoretical perspective, we extend results of Kurtz and co-authors in regard to scaling limits of continuous time Markov chains [3, 5]. Specifically, we allow transition rates that depend inhomogeneously on the scaling parameter, with potentially differing effects between the fluid and diffusion limits, and we also accommodate non-differentiable drifts. Proofs of these results are outlined in the Appendices.

Conclusions are given in Section 6. Preliminary versions of these results were presented in [6] and an extensive treatment of the mathematics used throughout the paper is provided in the thesis [7].

1.1. Related work

The possibility of controlling in real-time the capacity of service systems has been considered in many contexts. One motivation originates in contact centers where customers consult agents whose time is expensive. In this setting, rather than have the agents work fixed shifts, it can be cheaper to implement strategies for contacting the agents on-demand in a way that minimizes the waiting time of both customers and agents [8, 9]. Another motivation, originating in [10], is the "speed scaling" of hardware (e.g. the microprocessor clock), to trade off processing efficiency with energy consumption. The performance of scaling rules for service capacity as a function of the number of jobs, in combination with job scheduling disciplines, has been analyzed with different tools: [11, 12, 13] study the worst-case over a finite batch of jobs, [14, 15, 16] employ stochastic queueing tools and a control-theoretic viewpoint is given in [17].

A similar kind of tradeoff appears in the situation of a large data center, where servers may enter a power saving mode in times of low demand; the control of active capacity in this context has been termed the "right-sizing" problem [2]. With the emergence of cloud computing these infrastructures have become increasingly distributed, and a great deal of attention has been given to the question of *load balancing* in a large pool of servers; in particular, the issue of finding efficient policies with no centralized queuing and small messaging requirements [18, 19, 20, 21, 22, 23, 24]. The interplay between the right-sizing and load balancing problems has been studied in [25, 26, 27] and is a relevant problem from the perspective of cloud vendors, who need to distribute server instances among multiple customers with time-varying requirements.

In this paper we consider, instead, the viewpoint of cloud clients: i.e., service providers contracting computing capacity from a cloud vendor. Cloud clients typically outsource the load balancing mechanism, and for many applications resort to central queues to store job requests from users [28]. For instance, in e-commerce purchase requests are typically received by web servers which place them in a central queue, from which they are picked by back-end servers for processing; it is only after the purchase has been processed that the user receives an email confirmation with the receipt. A second example are social media sites, which resort to a similar procedure to handle status updates and photo uploads. Another typical example is the web/work split found in several Platform as a Service providers such as Google App Engine or Heroku, where worker processes

can be spawned on demand to serve a time-varying load of time-intensive tasks collected by the central web process. These processes typically communicate through a queue service such as Amazon SQS or RabbitMQ which are central queues where worker processes collect jobs.

In all these examples, the request placed on the queue is a short message describing a potentially time consuming task to be executed by a back-end server; this task is typically resource intensive or depends on a remote service that may not always be available. The central queue architectures described above decouple the collection and execution of user requests, speeding up the first of these processes by allowing web servers to quickly respond to users while the job is processed in the background; for implementation details we refer to [29, 30]. In this context it is natural to control the back-end capacity contracted from the cloud vendor in feedback with the number of job requests in the central queue, and cloud vendors provide tools for doing this [28]. This paper concerns the design of such controls.

The basic mathematical tool for exploring this problem is the theory of fluid and diffusion limits for density dependent families of Jump Markov processes. This theory was established by Kurtz in [5, 31, 32], and is summarized in [3] for a class of density dependent sequences of continuous time Markov chains. These can be used to model a broad class of queueing systems under several large-scale regimes, for which fluid and diffusion approximations follow directly from the standard theory. However, some of the more interesting heavy-traffic regimes, such as the one described in [4], do not fall into this standard framework and have required the use of ad-hoc techniques [33].

In the load balancing references cited above, policies are mostly distinguishable in heavy-traffic; however it is difficult to find a practical explanation for the relevance of this regime. In our model with controlled scaling of the service capacity, heavy-traffic emerges naturally and our analysis extends the methods of Kurtz to these situations, covering in particular the case of [4].

Another limitation in the applicability of Kurtz's results stems from the regularity hypotheses on the drift, which is assumed continuous for fluid limits and continuously differentiable for diffusion approximation. In previous works [34, 35, 36] this has motivated generalizations of the fluid limit in [3] to contemplate discontinuous drifts. In our work, extensions of the diffusion limit are provided for a class of non-smooth drifts.

2. A basic control and its large-scale behavior

Consider a service system where job requests arrive as a Poisson process of intensity λ and have exponential service times of mean $1/\mu$. The number of jobs in the system is denoted by $N(t)$ and the number of active servers, ready to work, is denoted by $M(t)$; we highlight that $M(t)$ may change over time.

Incoming jobs are put in a central queue whenever $N(t) > M(t)$, afterwards these jobs are served in arrival order as servers become available. Instead, when $M(t) > N(t)$, the system is overprovisioned with $M(t) - N(t)$ servers idling. In either case, the total number of active servers is always $\min\{M(t), N(t)\}$.

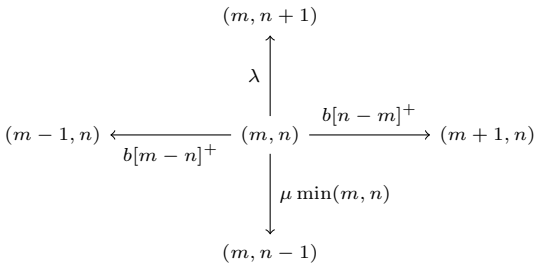


Figure 1: Transition rates of the continuous time Markov chain $X = (M, N)$ corresponding to a system operating under the two-sided backpressure policy.

If the system could create and delete servers instantaneously, it could adapt to an arbitrary load, operating as an infinite-server queue with $M(t) = N(t)$ at all times. However, such fast control of server instances is not possible in practice; in the context of cloud computing, the cloud vendor cannot respond immediately to provisioning requests [24, 25].

In this paper we consider a simple model for the lags in the creation and deletion of servers, assuming they are exponentially distributed with mean $1/b$. By choosing appropriate rules for server creation and deletion, the stochastic process $X = (M, N)$ becomes a continuous time Markov chain on $\mathbb{N}^2$. Below we describe one of such rules, which provisions servers in feedback with the current system occupation.

- A server creation request is generated whenever a job arrives and must be queued because there are no available servers.
- A deletion request is generated when a server becomes available.

The number of creation (deletion) requests that are pending at a given time is thus $[N(t) - M(t)]^+$ (respectively, $[M(t) - N(t)]^+$)[2], which multiplied by the exponential rate b will give the horizontal transition rates of Fig. 1. The vertical transitions model the job queue. In this way, the queue length acts as backpressure against the increase in queue length itself, and the same happens with the number of idle servers.

Denoting by $\beta_l(m, n)$ the transition rate in the direction $l \in \mathbb{Z}^2$, we have:

$$\beta_l(m,n) = \begin{cases} b[n-m]^+ & \text{if} \quad l = (1,0), \\ b[m-n]^+ & \text{if} \quad l = -(1,0), \\ \lambda & \text{if} \quad l = (0,1), \\ \mu \min(m,n) & \text{if} \quad l = -(0,1). \end{cases} \tag{1}$$

[2]Strictly speaking, keeping the creation requests aligned with $[N(t) - M(t)]^+$ requires canceling requests if jobs depart, and similarly with deletion requests if jobs arrive.

The invariant distribution of this Markov chain does not admit a closed form expression. Nevertheless, in the context of large-scale cloud systems, it is natural to analyze it through appropriate scaling limits.

2.1. Fluid limit

To understand how the system behaves in a large-scale regime, we introduce a scale parameter k to model the level of demand for service. Specifically, we let the arrival rate be $k\lambda$ and we consider growing values of k approaching infinity, thus modeling systems which receive requests from a large number of users. This scaling is standard for infinite-server queues, see e.g. [37]; note that due to the infinite-server nature of the system, the average number of active servers in steady-state also scales with k. Indeed, while the service capacity of each individual server is kept constant, the overall active capacity grows with k.

We denote the stochastic process corresponding to the scaled system by $\hat{X}_k = (\hat{M}_k, \hat{N}_k)$; the transition rates are the same as in Fig. 1 except for the new arrival rate $k\lambda$. The normalized process $X_k = \hat{X}_k/k$ may be studied with the tools of [3], which are now briefly reviewed.

The Markov chain X_k is called a *density dependent* population process: it has state space in $k^{-1}\mathbb{N}^2$, a lattice in the positive quadrant $\mathbb{R}^2_+$, and the transition rate between two states, x and y, is $q^k_{xy} = k\beta_{k(y-x)}(x)$, invoking (1). Intuitively, transitions occur k-times faster but cover a k-times smaller distance than in the original chain. The process X_k converges with k to a *fluid limit*, a deterministic function on $\mathbb{R}^2_+$, specified as follows.

Introduce the drift vector field

$$F(x) = \sum_{l \in D} l\beta_l(x) \quad \text{for all} \quad x \in \mathbb{R}^2;$$

here D refers to the set of valid transitions (in this case four) of (1). This gives

$$F(m,n) = \begin{bmatrix} b(n-m) \\ \lambda - \mu \min(m,n) \end{bmatrix},$$

which is piecewise linear, thus Lipschitz. We invoke the next result from [3].

Theorem 2.1 (Kurtz)**.** Suppose the maps β_l are bounded on compact sets and the drift F is locally Lipschitz. Assume that the deterministic initial conditions $X_k(0)$ converge to $x_0 \in \mathbb{R}^2_+$ and let $x(t)$ denote the unique solution to $\dot{x} = F(x)$ with initial condition x_0. If this solution is defined on $[0,T]$ then

$$\sup_{t \in [0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad k \to \infty.$$

The function $x(t)$ is called fluid limit.

In our specific case, the fluid limit is the deterministic function denoted $x(t) = (m(t), n(t))$, solution of the differential equation:

$$\dot{m} = b(n-m), \tag{2a}$$

$$\dot{n} = \lambda - \mu \min(m,n). \tag{2b}$$

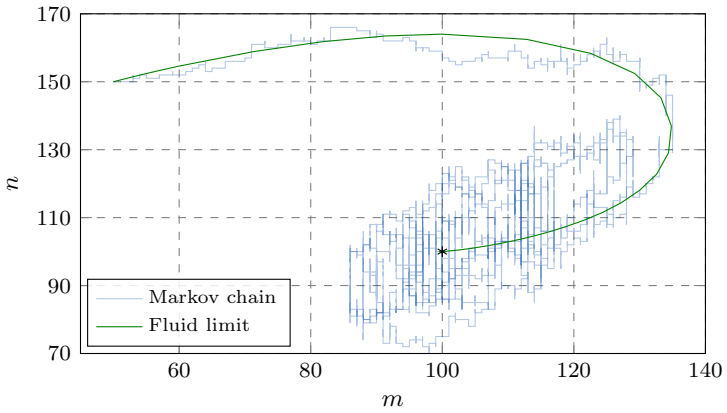


Figure 2: Sample path of the Markov chain $X(t)$ and its fluid limit $x(t)$; the parameters of the simulation are $\lambda = 100$, $\mu = 1$, $b = 1$ and $x_0 = (50, 150)$.

Here we are slightly overloading the notation (m, n), different from its use in Figure 1 where (m, n) represents an element of the state-space of the Markov chain X; the interpretation in each case should be clear from context.

To test how close the deterministic approximation (2) is to the stochastic dynamics shown in Fig. 1, we analyze the next example: Fig. 2 shows a sample path of a system where the offered traffic or load is $\rho := \lambda/\mu = 100$, and the unique solution to (2) with the same initial condition. This system corresponds to a service provider contracting, on average, about a hundred servers from the cloud; this is still an order of magnitude below the maximum number of instances that can be managed through auto-scaling features of, for instance, Microsoft Azure. We see that the fluid model correctly describes the average trend of the stochastic process, converging to an an equilibrium point.

We focus for a while on the fluid dynamics (2), later on returning to the fluctuations. A first simple observation is that they have a unique equilibrium point: $x^* = (\rho, \rho)$, i.e. the number of jobs and servers in the large-scale limit matches the load. Moreover, we have the following stability result.

Proposition 2.2. $x^* = (\rho, \rho)$ is a global attractor of the dynamics (2).

Proof. The dynamics (2) are piecewise linear, switching at the line $m = n$, hence we have different Jacobian matrices in $\{m < n\}$ and $\{m > n\}$, respectively:

$$A_1 = \begin{bmatrix} -b & b \\ -\mu & 0 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -b & b \\ 0 & -\mu \end{bmatrix}.$$

We claim that there exists a common quadratic Lyapunov function. Namely, a positive definite symmetric matrix

$$P = \begin{bmatrix} 1 & q \\ q & r \end{bmatrix}$$

such that $A_i^T P + P A_i$ is negative definite for $i = 1, 2$. Let t_i and d_i denote, respectively, the trace and determinant of $A_i^T P + P A_i$. To prove the claim we must find $q, r \in \mathbb{R}$ such that P is positive definite and:

$$\begin{aligned} t_1(q,r) &= 2(b-\mu)q - 2b < 0, \\ d_1(q,r) &= -4b(b+\mu q)q - (b - \mu r - bq)^2 > 0, \\ t_2(q,r) &= 2b(q-1) - 2\mu r < 0, \\ d_2(q,r) &= -4b(bq - \mu r) - [b - (b+\mu)q]^2 > 0. \end{aligned}$$

The set $\{d_1(q,r) > 0\}$ is the interior of an ellipse, located inside $\{q \leq 0\}$ and tangent to the line $q = 0$ at the point $(0, b/\mu)$, whereas $\{d_2(q,r) > 0\}$ is the open set above the graph of a parabola that contains the point $(0, b/4\mu)$ and has positive concavity. The two sets intersect, moreover, there exists $\delta > 0$ such that $(-\varepsilon, b/\mu)$ lies in the intersection for all $\varepsilon \in (0, \delta)$. Since $t_1(\varepsilon, b/\mu) \to -2b$ and $t_2(\varepsilon, b) \to -4b$ as $\varepsilon \to 0$, there exists some $\varepsilon > 0$ such that

$$P = \begin{bmatrix} 1 & -\varepsilon \\ -\varepsilon & b/\mu \end{bmatrix}$$

satisfies all the conditions listed on the previous paragraph; this matrix is positive definite for all sufficiently small ε. □

The above result implies the system will approach the desired equilibrium, regardless of the initial condition, and remain close afterwards. In addition, the stability result can be extended to the pre-limit process X_k, exploiting the common quadratic Lyapunov function to show that X_k is ergodic, by means of a Foster-Lyapunov argument; this is done in [7, Section 4.2].

The fluid equilibrium has the same number of jobs and servers, but the oscillations of X around x^* will result in queueing at some times and overprovisioning on other occasions. In steady-state, the opposing transition rates in Fig. 1 cancel on average. Particularly, if we look at the transitions in the m-direction, we see that, in the steady-state, the mean queue length equals the average number of idle servers:

$$\mathbb{E}[N - M]^+ = \mathbb{E}[M - N]^+.$$

To better understand these fluctuations we may use a diffusion model.

2.2. Diffusion approximation

Let us introduce the standard central limit scaling of the process X_k around the equilibrium: $Z_k := \sqrt{k}(X_k - x^*)$. We have the next approximation theorem.

Theorem 2.3. Assume the deterministic initial condition of Z_k converges to some $Z_0 \in \mathbb{R}^2$ as $k \to \infty$. Then Z_k has a limit in distribution in the Skorokhod

space $D_{\mathbb{R}^2}[0,\infty)$. This limit is the unique solution $Z = (Z^m, Z^n)$ of the following stochastic differential equation (SDE), with initial condition Z_0.

$$dZ_t^m = b(Z_t^n - Z_t^m)dt, \tag{3a}$$

$$dZ_t^n = -\mu \min(Z_t^m, Z_t^n)dt + \sqrt{2\lambda}dW_t, \tag{3b}$$

with W a standard one-dimensional Wiener process.

In contrast to the fluid limit result, the above theorem is not covered by the classical diffusion limit of [3]; the reason is the vector field $F(x)$ is not differentiable at x^*. In Appendix B we provide a suitable generalization, using a notion of *pseudo-differentiability*: namely, the existence of a local piecewise linear approximation, which in the above case takes the form

$$\partial F(x) = \partial F(m,n) = \begin{bmatrix} -b & b \\ -\mu & 0 \end{bmatrix} x \quad \text{if} \quad m > n, \quad \begin{bmatrix} -b & b \\ 0 & -\mu \end{bmatrix} x \quad \text{if} \quad m < n.$$

The stochastic dynamics (3) are an approximation to the oscillations of X around x^* in the scale of $\sqrt{\rho}$, the approximation becomes exact when $\rho \to \infty$. We are currently unable to compute the invariant distribution of (3) explicitly because the non-linear term $\mu \min(Z_t^m, Z_t^n)$ makes it difficult to solve this SDE. Still, (3a) implies that any invariant distribution of (3) satisfies $\mathbb{E}[Z^m] = \mathbb{E}[Z^n]$, which suggests that X will fluctuate between the overprovisioning and the queueing zone as it approaches the steady-state regime; this is the behavior that we observed in Fig. 2.

In the next sections we explore alternatives to the two-sided backpressure policy, in order to bias the allocation to avoid queueing delays. The design criterion is that our policies should not require prior knowledge of the system's offered load and should cope with possibly large variations of its value.

3. Achieving zero queueing delay

Our first approach to eliminate queueing is better explained by returning to the fluid dynamics (2). The many-server component, captured by (2b), forces possible equilibria to lie in the L-shaped region $\min\{m,n\} = \rho$, as depicted in Fig. 3; the remaining degree of freedom is the rule for provisioning servers. Consider the following alternative dynamics at the fluid level:

$$\dot{m} = b[(1+\delta)n - m], \tag{4a}$$

$$\dot{n} = \lambda - \mu \min\{m,n\}. \tag{4b}$$

The fluid queue is unmodified from (2b), but the rule for summoning servers has been altered, still preserving the time lags that are part of our physical constraints. The new rule (4a), aims for a fraction δn of spare capacity, hedging against full utilization of servers by summoning them before full occupancy.

The unique equilibrium x^* of (4) has coordinates $m^* = (1+\delta)\rho$ and $n^* = \rho$, as shown in Fig. 3. The number of jobs still operates at ρ, which is a hard lower

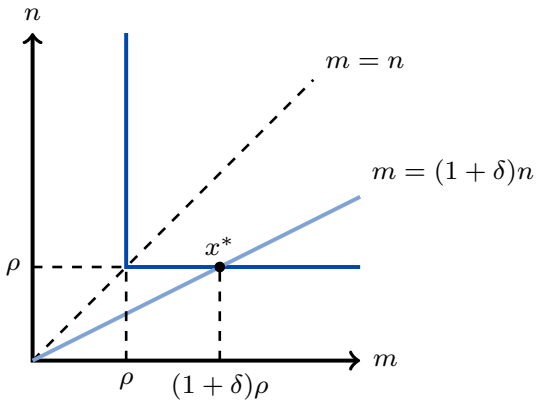


Figure 3: Feasible equilibria and the result of control (4a).

bound; however, we accept an average overprovisioning of about $\delta\rho$ servers, with the aim of avoiding operation in the queuing region $\{n > m\}$. To achieve this control, we modify the policy of Section 2 adding a δn bias term as in (4a); the corresponding chain $X = (M, N)$ has the intensities depicted in Fig. 4.

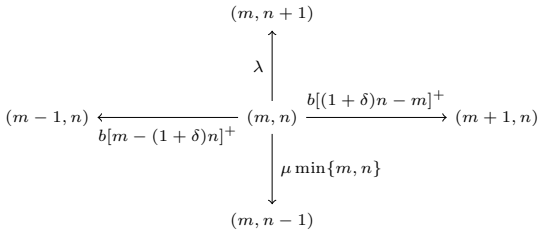


Figure 4: Transition rates of the Markov chain with linear overprovisioning bias.

A similar relationship holds here between Markov and fluid models. Namely, denote again by $\hat{X}_k = (\hat{M}_k, \hat{N}_k)$ the process governed by the transition rates in Fig. 4 but with arrival rate scaled by $k\lambda$. The Markov chain $X_k = \hat{X}_k/k$ is a density dependent population process with drift

$$F(m, n) = \begin{bmatrix} b((1+\delta)n - m) \\ \lambda - \mu \min(m, n) \end{bmatrix}.$$

This vector field is Lipschitz, so Theorem 2.1 implies the following.

Proposition 3.1. Assume $X_k(0)$ converges to some $x_0 \in \mathbb{R}^2_+$ as $k \to \infty$ and let $x = (m, n)$ be the unique solution to (4) with initial condition x_0. Then

$$\sup_{t \in [0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad k \to \infty \quad \text{for all} \quad T \geq 0.$$

The dynamics (4) have switching at the line $m = n$ as in Section 2, but now the equilibrium $x^* = ((1+\delta)\rho, \rho)$ lies on the interior of $\{m > n\}$. Still, there exists a common quadratic Lypaunov function as in Proposition 2.2, which yields the next result; see [7, Proposition 4.3.1] for further details.

Proposition 3.2. x^* is a global attractor of (4) provided that $\delta \in (0, 1)$.

In order to characterize variability around this biased equilibrium, we resort again to a diffusion approximation. In this case, since the equilibrium of the fluid limit lies in the interior of the linear region $\{m > n\}$, the diffusion is given by a linear SDE and its steady-state can be completely characterized.

Consider the process $Z_k = \sqrt{k}(X_k - x^*)$ describing the oscillations of X_k around the fluid equilibrium x^* on the scale of $\sqrt{k}$. We have the following diffusion approximation.

Proposition 3.3. Assume $Z_k(0)$ converges to some $Z_0 \in \mathbb{R}^2$ as $k \to \infty$. Then Z_k converges weakly in $D_{\mathbb{R}^2}[0, \infty)$ to the unique solution $Z = (Z^m, Z^n)$ of the following SDE, with initial condition Z_0.

$$dZ_t^m = b[(1+\delta)Z_t^n - Z_t^m]dt, \tag{5a}$$

$$dZ_t^n = -\mu Z_t^n dt + \sqrt{2\lambda}dW_t, \tag{5b}$$

where W is a standard one-dimensional Wiener process.

The proof follows from the standard results for density dependent population processes in [3] since the drift is linear in a neighborhood of x^*. Equation (5b) is now linear, so the diffusion corresponds to an Ornstein-Uhlenbeck process whose stationary distribution can be computed exactly.

Specifically, let $\eta = \mu/b$ denote the ratio between mean provisioning delays and mean service times; the stationary distribution of Z is a bivariate Gaussian $\mathcal{N}(0, \Sigma)$, with mean zero and covariance matrix

$$\Sigma = \rho \frac{1+\delta}{1+\eta} \begin{bmatrix} 1+\delta & 1 \\ 1 & \frac{1+\eta}{1+\delta} \end{bmatrix}. \tag{6}$$

The latter is obtained by solving the Lyapunov equation $A\Sigma + \Sigma A^T + BB^T = 0$, where $dZ = AZdt + BdW$ is the SDE (5) written in matrix form.

From a practical perspective, the interpretation of Propositions 3.1 and 3.3 is that $\hat{X}_k \approx kx^* + \sqrt{k}Z$ is a reasonable steady-state approximation[3] for large enough k. Recall that $\hat{M}_k$ and $\hat{N}_k$ are the number of servers and jobs, respectively, in a system where the arrival rate is $k\lambda$. The fluid equilibrium of this system is $kx^* = (k\rho, k\rho)$, and the steady-state covariance of $\sqrt{k}Z$ is as in (6) but replacing ρ by $k\rho$. Thus, another way to express this estimate, incorporating the scaling into λ, is to say that $X \approx x^* + Z$ when λ is large enough.

[3]Strictly speaking, invoking this steady-state approximation involves interchanging the diffusion limit with the limit in time. This aspect is currently outside our scope.

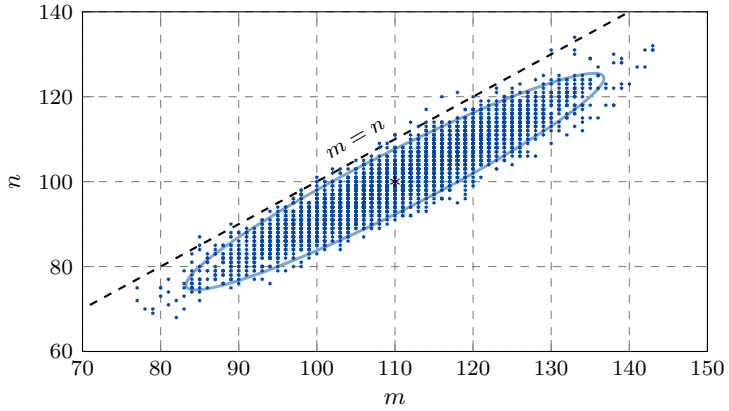


Figure 5: Level set of $(y - x^*)^T \Sigma^{-1} (y - x^*)$ and the states visited by a sample path of X; the parameters of the simulation are $\lambda = 100$, $\mu = 1$, $b = 10$, $\delta = 0.1$ and $x_0 = (100, 110)$.

This is illustrated in Fig. 5, which shows the states visited by a system X operating with about a hundred servers; the states are not connected through lines as in Fig. 2 to make the plot clearer. The plot corresponds to the stationary behavior of the system, which was started from the equilibrium x^*; a level set of the density of the stationary distribution of Z has been drawn to show the similarity with the shape of the cloud of states visited by the system.

3.1. Parameter setting

We now discuss how to set the parameter δ. For this purpose, note that the random variable which measures idle capacity is $[M - N]^+$ and that the above estimate translates into $M - N \approx \mathcal{N}(\delta\rho, \sigma^2)$, where the variance is

$$\sigma^2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \Sigma \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{\delta^2 + \eta}{1 + \eta} \rho. \tag{7}$$

We can use this variance to adjust the parameter δ so that queueing is avoided with high probability. For instance, setting $\delta\rho - 2\sigma \geq 0$ (the equilibrium point two standard deviations away from the diagonal), the probability of negative values for $M - N$ is made very small. From equation (7), this design condition on δ can be rewritten as[4]:

$$\delta \geq \sqrt{\frac{\eta}{\rho(1 + \eta)/4 - 1}}. \tag{8}$$

Given a certain load ρ, the above choice of δ results in the queue remaining empty with very high probability. This is represented in Fig. 6, where we see how $M - N$ fluctuates around $\delta\rho$, hardly ever reaching $\{m < n\}$, the region

[4] We assume the load is large enough, so the denominator is always positive.

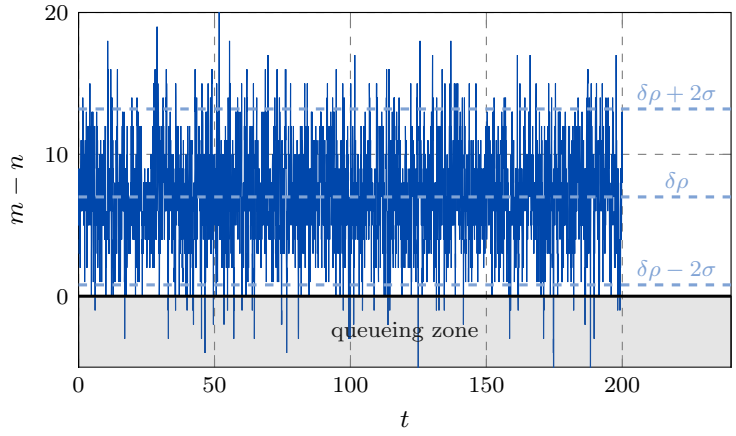


Figure 6: Simulation of the system X showing the overprovisioning level and the avoidance of the queueing zone; the parameters of the simulation are $\lambda = 100$, $\eta = 0.1$ and $\delta = 0.07$.

where incoming jobs are queued upon arrival. The Gaussian approximation adopted to derive the design criterion (8) can also be used to estimate the mean number of job requests waiting in the central queue in steady-state; this estimate is provided in [6], where it is also evaluated for different values of δ and ρ.

This design rule has an important drawback: it requires a priori knowledge of the load ρ. If δ is set for a given estimate of ρ, performance can deteriorate if the offered load changes, either in the direction of excessive overprovisioning, or incurring in undesired queueing. This is why in the next section we propose an automatic policy which can be tuned independently of the load.

4. Automatic efficient scaling

We develop here an automatic rule, independent of the load, with the aim of achieving an optimal provisioning level. Equation (8) suggests that the overprovisioning fraction δ should be of the order of $1/\sqrt{\rho}$, or equivalently, the absolute overprovisioning in equilibrium should be of order $\sqrt{\rho}$. This is exactly what is suggested by the square-root staffing rule of Halfin and Whitt [4] in the context of many-server queues with a *static* number of servers. However, applying such a rule in practice requires the designer to know the load it will be facing. In what follows we propose instead an *automatic* rule for provisioning the number of servers, based on the current job occupation; this rule is shown to achieve, in equilibrium, the same overprovisioning level without previous explicit knowledge of the system load.

The proposal is to take the current occupation n as proxy for the load, making the backpressure bias of order $\sqrt{n}$. This alternate dynamics are represented in the transitions diagram of Fig. 7. Comparing with Fig. 4, the bias term δn has been replaced by $\varepsilon\sqrt{n}$. Consider the scaled processes $\hat{X}_k = (\hat{M}_k, \hat{N}_k)$ where, as before, the arrival rate λ is replaced by $k\lambda$ in Fig. 7. Consider also

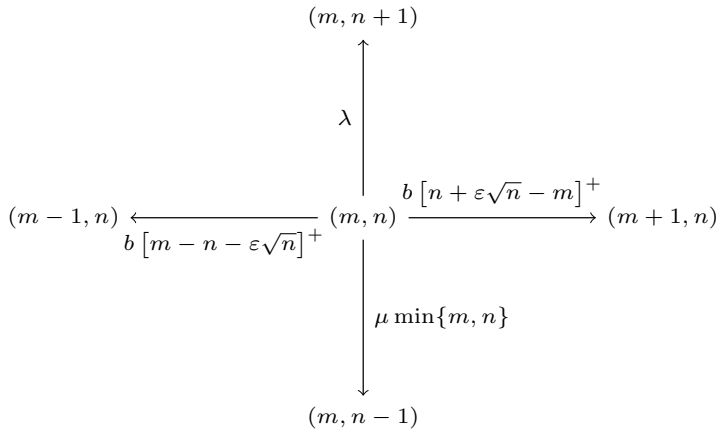


Figure 7: Transition rates for the rule with square-root overprovisioning bias.

the normalized processes $X_k = \hat{X}_k/k$. A first result is that the $\sqrt{n}$ terms in the transition rates of Fig. 7 disappear from the fluid scale as $k \to \infty$.

Theorem 4.1. Assume the deterministic initial conditions $X_k(0)$ converge to some $x_0 \in \mathbb{R}^2_+$ as $k \to \infty$ and let $x = (m, n)$ be the unique solution to:

$$\dot{m} = b(n - m), \tag{9a}$$

$$\dot{n} = \lambda - \mu \min\{m, n\}, \tag{9b}$$

with initial condition x_0. Then the processes X_k satisfy:

$$\sup_{t\in[0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad k \to \infty \quad \text{for all} \quad T \geq 0.$$

Note the coincidence of (9) with our first fluid model (2); in the fluid scale we see the same system as in Section 2, with zero overprovisioning. The intuition behind this result is that the bias $\varepsilon\sqrt{n}$ is of order $\sqrt{\rho}$, and hence negligible in the asymptotic regime $\rho \to \infty$ with respect to the system's operating point (ρ, ρ).

The previous result does not follow directly from Kurtz's Theorem 2.1 as in Section 2. In this case we do not have a density dependent population process, as assumed there, because of the $\sqrt{n}$ terms in the transition rates of Fig. 7. A suitable modification of Theorem 2.1 is provided in Appendix A and the previous theorem is a particular case of this modification.

To see how the system counteracts queueing delay, we look into the oscillations around the fluid equilibrium in the diffusion scale, considering again the processes $Z_k = \sqrt{k}(X_k - x^*)$. In this scale, the system manages to move away from the queueing region $\{n > m\}$, minimizing delay while the overprovisioning is of order $\sqrt{\rho}$. Formally, we have the following result.

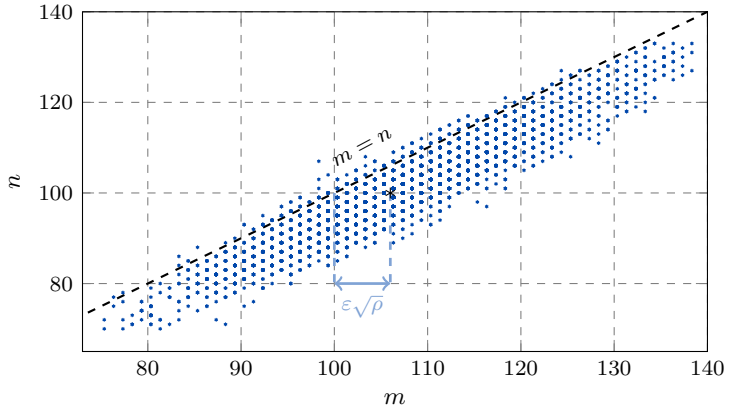


Figure 8: Simulation of the Markov chain of Fig. 7, the plot shows the states the system visited. The parameters of the simulation are $\lambda = 100$, $\mu = 1$, $b = 10$ and $\varepsilon = 0.6$.

Theorem 4.2. Assume $Z_k(0)$ converges to $Z_0 \in \mathbb{R}^2$ as $k \to \infty$. Then Z_k has a limit in distribution in the Skorokhod space $D_{\mathbb{R}^2}[0, \infty)$. This limit is the unique solution $Z = (Z^m, Z^n)$ of the following SDE with initial condition Z_0.

$$dZ_t^m = b[Z_t^n - Z_t^m + \varepsilon\sqrt{\rho}]dt, \tag{10a}$$

$$dZ_t^n = -\mu \min\{Z_t^m, Z_t^n\}dt + \sqrt{2\lambda}dW_t, \tag{10b}$$

where W is a standard one-dimensional Wiener process.

Again, the proof of Theorem 4.2 requires modifications on the standard diffusion limits for density dependent population processes. One reason is the vector field for the drift in (10) is not differentiable, as was already the case for Theorem 2.3. Here we have the additional issue that X_k is not a density dependent population process, as it was explained before. In Appendix B we prove a generalization of the classical diffusion theorem in [3] to inhomogeneous scalings and pseudo-differentiable drifts, and in Appendix C we use this generalization to prove Theorem 4.2.

This extension has, in addition to technical difficulties, non-trivial consequences. In particular, if we compare equations (9) and (10), there is an extra drift term in (10a); removing the noise from the diffusion does *not* give the same result as the incremental fluid model. Equation (10a) implies $\mathbb{E}[Z^m - Z^n] = \varepsilon\sqrt{\rho}$ for any steady-state distribution, so the $O(\sqrt{\rho})$ overprovisioning, invisible in the fluid scale, appears in the diffusion scale. Fig. 8 shows this overprovisioning in a system with a moderately high load; the system is in steady-state, hovering around $(\rho + \varepsilon\sqrt{\rho}, \rho)$, thus operating with an overprovisioning of $\varepsilon\sqrt{\rho}$ servers.

Remark 4.3. The result from Appendix B implies the $\varepsilon\sqrt{\rho}$ term in (10a) disappears when the bias $\varepsilon\sqrt{n}$ is replaced by any $o(\sqrt{n})$ expression. Therefore, $\varepsilon\sqrt{n}$ is the minimal amount of bias that results in diffusion scale overprovisioning.

4.1. Parameter setting

Recall the steady-state approximation $X \approx x^* + Z$ introduced in Section 3, in this case with $x^* = (\rho, \rho)$ the unique equilibrium of (9) and Z the stationary distribution of (10). In contrast to Section 3, now we cannot derive a closed form expression for Z because (10) contains switching; this means we do not have an exact analytic expression to set the parameter ε.

However, the solution of (10) remains within $\{m > n\}$ most of the time for all large enough ε, as we may infer from the plot of Fig. 8. This suggests a further approximation, $X \approx x^* + \tilde{Z}$ with $\tilde{Z}$ the stationary distribution of the Ornstein-Uhlenbeck process that solves:

$$d\tilde{Z}_t^m = b[\tilde{Z}_t^n - \tilde{Z}_t^m + \varepsilon\sqrt{\rho}]dt, \tag{11a}$$

$$d\tilde{Z}_t^n = -\mu\tilde{Z}_t^n dt + \sqrt{2\lambda}dW_t. \tag{11b}$$

Note that (11a) corresponds to (10a) exactly, whereas in (11b) we have replaced the non-linear term $\min\{Z_t^m, Z_t^n\}$ of equation (10b) with $\tilde{Z}_t^n$. This approximation will be evaluated in the following section through extensive simulations.

As in Section 3, the stationary distribution of the linear SDE (11) is Gaussian. In this case with mean $(\varepsilon\sqrt{\rho}, 0)$ and covariance matrix $\tilde{\Sigma}$ as in (6), but with δ replaced by zero. Consequently, $M - N \approx \mathcal{N}(\varepsilon\sqrt{\rho}, \sigma^2)$ with

$$\sigma^2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \tilde{\Sigma} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{\eta}{1+\eta}\rho.$$

As in Section 3.1, we can now choose ε such that $\varepsilon\sqrt{\rho} - 2\sigma \geq 0$, so that the mean of $M - N$ is two standard deviations away from zero, thus avoiding the queueing region with high probability. This condition translates into

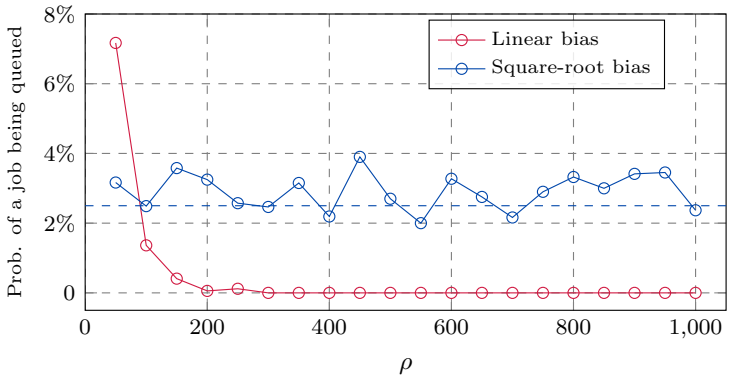
$$\varepsilon \geq 2\sqrt{\frac{\eta}{1+\eta}}. \tag{12}$$

Comparing with (8), the parameter selection no longer depends on the load of the system, ε can be set without knowing ρ.

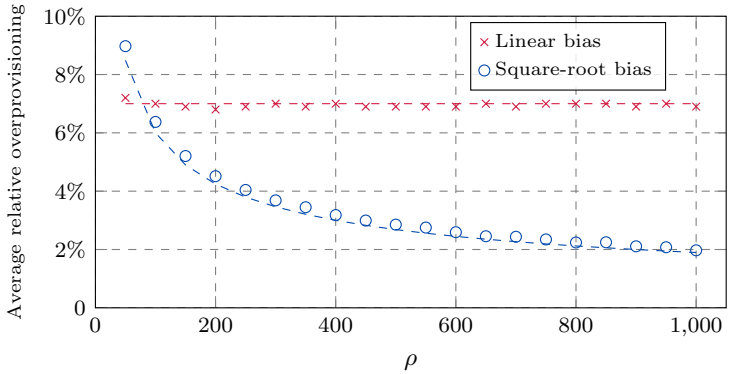
5. Performance evaluation under varying demand

The analysis of sections 3 and 4 was developed assuming a stationary setting: i.e., the demand profile does not change over time. However, the most important application of auto-scaling is tracking a time-varying load while keeping the number of active instances close to the number of jobs at all times.

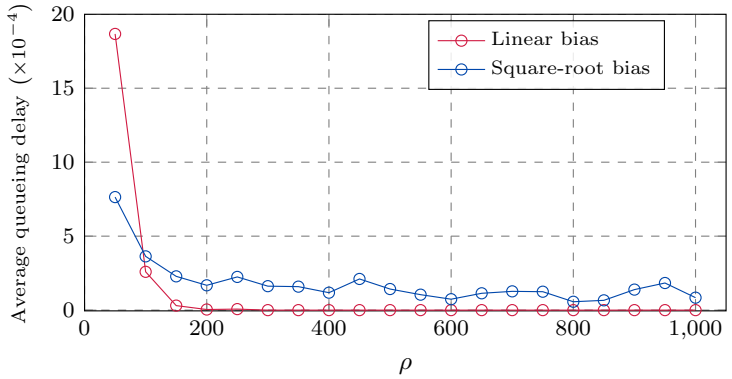
From an implementation perspective, the linear bias rule shown in Fig. 4 only depends on the number of active jobs and servers, as well as the design parameter δ that controls overprovisioning. The controller or *automatic scaler* in cloud settings, should keep track of both magnitudes and request the allocation of $\lceil N(t)(1+\delta) - M(t) \rceil$ new server instances whenever this number is greater than zero, and request deactivation of $\lceil M(t) - N(t)(1+\delta) \rceil$ instances in the



(a) Probability of a job being queued and reference design (dashed).

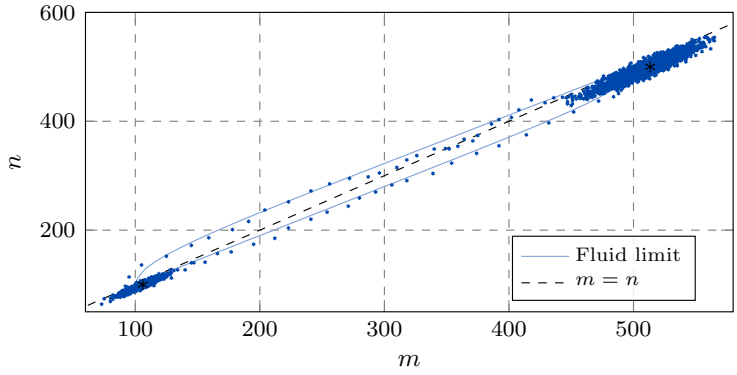


(b) Average relative overprovisioning in the systems and their fluid estimates (dashed).

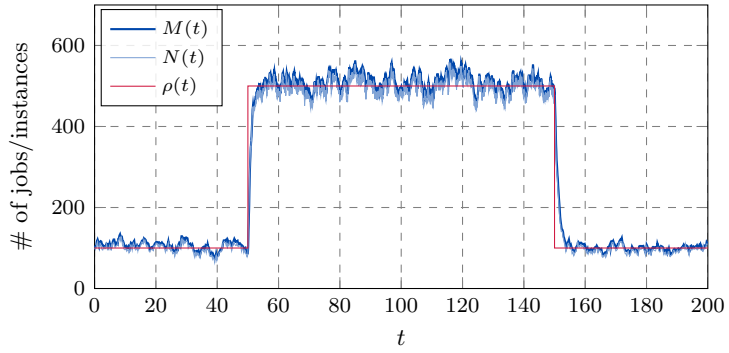


(c) Average queuing delay experienced by customers.

Figure 9: Simulations for two systems with $\mu = 1$, $b = 10$ and different arrival rates. The linear bias scaling uses $\delta = 0.07$ and the square-root bias scaling $\varepsilon = 0.6$, both designed for approximately 2.5% queueing probability at $\rho = 100$. The marks in the plots correspond to time averages from simulations.



(a) Phase diagram.



(b) Time evolution.

Figure 10: Simulation of the square-root bias rule under an abrupt increase in demand: $\lambda(t) = 500$ for all $t \in [50, 150]$ and $\lambda(t) = 100$ otherwise. The remaining parameters of the simulation are $\mu = 1$, $b = 10$ and $\varepsilon = 0.6$.

overprovisioning situation. The exact same discussion applies for the square-root bias rule depicted in Fig. 7. Note that whenever the load grows, the queue will grow faster and the scaler will command more servers. When load decreases the system will detect the overprovisioning and adjust accordingly.

One of the advantages of the square-root rule is it can be designed to automatically provide the same quality of service across different loads. This is explored in Fig. 9, where the linear rule is designed as in (8), assuming a nominal load $\rho = 100$, and the square-root bias rule is designed as in (12), which is independent of ρ. As load varies, we see that the square-root rule keeps a constant queueing probability while the overprovisioning, relative to the load, decreases as $O(1/\sqrt{\rho})$. The linear rule has a similar performance for the design load $\rho = 100$, but the queueing probability becomes too high when the load is below the nominal value, and goes to zero as the load grows, maintaining an unnecessary amount of idle instances. Also, in both cases the average queueing delay experienced by users is small and decreasing as a function of ρ.

Given the simplicity of the square-root rule, and its robustness to load changes, we now explore how the system behaves under time-varying arrivals using this rule. Our first scenario, depicted in Fig. 10, analyzes the case of an abrupt 5-fold increase in the load from a nominal value $\rho = 100$, achieved by changing the job arrival rate. From the phase diagram, we can see that after some transient behavior the system automatically readjusts to the new situation, keeping the right overprovisioning to have approximately the same queueing probability, and then returns to normal as the load decreases again. The second plot emphasizes that the transients are indeed very short and mostly governed by the instance creation and recall delay $1/b$.

Our second scenario, depicted in Fig. 11, shows the behavior of the square-root bias mechanism when the job arrival rate slowly varies in time. We chose $\mu = 1$ and $b = 10$ as before, and $\lambda(t)$ follows a sinusoidal pattern between $\rho = 50$ and $\rho = 150$. As the phase diagram shows, the system mostly operates in the overprovisioning zone, following the $m \approx n+\varepsilon\sqrt{n}$ bias and keeping the queueing probability small, even though ρ is constantly changing. The second plot shows that the system is indeed capable of following the load changes smoothly.

6. Conclusions

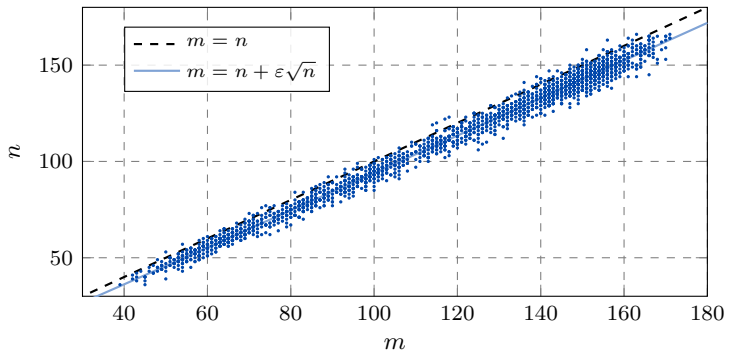
In this paper we analyzed certain feedback policies through which cloud clients may automatically control the deployment of server instances; these polices were designed to cope with variable demands, and our analysis considers the startup lag of servers. We derived simple control rules that explore the tradeoffs between queueing and overprovisioning. Based on both classical and new fluid models and diffusion approximations of the underlying queueing processes, we showed that it is possible to work under a reduced amount of queueing delay, provided the overprovisioning is appropriately scaled with demand. In particular we showed that a simple dynamic version of the square-root staffing rule of [4] achieves nearly zero queueing while keeping the overprovisioning scaling sublinearly with the load. Simulation analysis also showed the devised rule is robust against load variations, keeping track of demand in real time and providing uniform quality of service across all loads.

A direction of future work, from the theoretical standpoint, would be to establish the interchange between the diffusion and stationary limits, as mentioned in Section 3. We also plan to analyze the performance of the feedback control rules when the scaling feature is combined with distributed load balancing mechanisms such as Join-the-Idle-Queue or power-of-d choices.

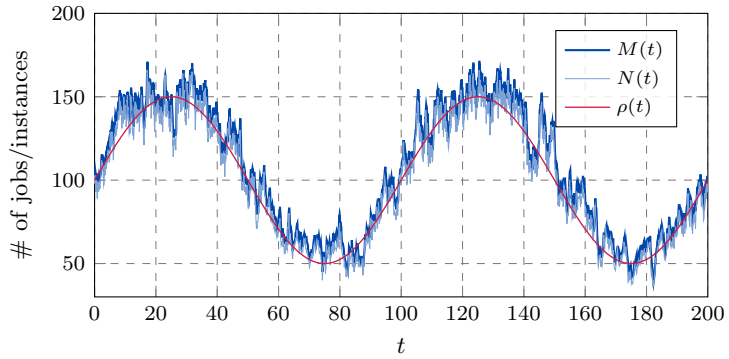
Appendix A. Fluid and diffusion limits for inhomogeneous scalings

For each $k \geq 1$ let us consider a collection of non-negative maps β_l^k, indexed on $D \subset \mathbb{Z}^d$, having a common domain $E \subset \mathbb{R}^d$, and such that

$$x + \frac{l}{k} \in E \cap \frac{\mathbb{Z}^d}{k} \quad \text{for all} \quad l \in D \quad \text{and all} \quad x \in E \cap \frac{\mathbb{Z}^d}{k} \quad \text{such that} \quad \beta_l^k(x) > 0.$$



(a) Phase diagram.



(b) Time evolution.

Figure 11: Simulation of the square-root bias rule under time-varying arrival rate of tasks: $\lambda(t) = 100 + 50\sin(\pi t/50)$. The other parameters are $\mu = 1$, $b = 10$ and $\varepsilon = 0.6$.

Under this hypothesis we may consider the continuous time Markov chains $X_k(t)$ with state-space $S_k = E \cap k^{-1}\mathbb{Z}^d$ and transition rates given by

$$q_{xy}^k = \begin{cases} 0 & \text{if} \quad k(y-x) \notin D, \\ k\beta_{k(y-x)}^k(x) & \text{if} \quad k(y-x) \in D. \end{cases}$$

Remark A.1. The subsequent results hold under mild hypotheses on E. For instance, it is enough to let E be either open or closed and convex; in the previous sections we took $E = \mathbb{R}^2_+$. Also, we assume below that $X_k(t)$ is well-defined for all $t \geq 0$ for simplicity, precluding the possibility of finite explosion times. The results hold in general though, for details we refer to [7]. We further assume that D is a finite set.

The maps $\beta_l^k(x)$ describing the intensities may be decomposed in two terms:

$$\beta_l^k(x) = \gamma_l(x) + \delta_l^k(x),$$

the first term is homogeneous in k and the second depends on k; below we will assume that $\delta_l^k(x)$ converges to zero with k, which implies the decomposition is unique. When the inhomogeneous terms are identically zero, the processes defined above are called density dependent population processes. Fluid and diffusion limits for density dependent population processes were provided by [3] under mild assumptions. In this appendix we extend these results to the broader class of processes introduced above, and we adopt weaker assumptions as explained in the introduction.

Let the initial conditions $X_k(0)$ be deterministic The processes $X_k(t)$ may be constructed as in [3, 7] on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, from a set of independent Poisson processes $\{\mathcal{N}_l\}_{l \in D}$ of intensity one, such that

$$X_k(t) = X_k(0) + \sum_{l \in D} \frac{l}{k} \mathcal{N}_l \left(\int_0^t k\beta_l^k(X_k(\tau))d\tau \right) \quad \text{for all} \quad t \geq 0 \quad \text{and} \quad k \geq 1.$$

It is convenient to let $Y_l(t) = \mathcal{N}_l(t) - t$, so that we may write

$$\sum_{l \in D} \frac{l}{k} \mathcal{N}_l \left(\int_0^t k\beta_l^k(X_k(\tau))d\tau \right) = \sum_{l \in D} \frac{l}{k} Y_l \left(\int_0^t k\beta_l^k(X_k(\tau))d\tau \right) + \int_0^t \sum_{l \in D} l\gamma_l(X_k(\tau))d\tau + \int_0^t \sum_{l \in D} l\delta_l^k(X_k(\tau))d\tau.$$

Definition A.2. The homogeneous and inhomogeneous drifts of the process $X_k(t)$ are, respectively, the vector fields $F, G_k : E \longrightarrow \mathbb{R}^d$ such that

$$F(x) = \sum_{l \in D} l\gamma_l(x) \quad \text{and} \quad G_k(x) = \sum_{l \in D} l\delta_l^k(x).$$

For all $t \geq 0$ and all $k \geq 1$, we may now write

$$\begin{aligned} X_k(t) = X_k(0) + \sum_{l \in D} \frac{l}{k} Y_l \left(\int_0^t k\beta_l^k(X_k(\tau))d\tau \right) \\ + \int_0^t F(X_k(\tau))d\tau + \int_0^t G_k(X_k(\tau))d\tau. \end{aligned} \tag{A.1}$$

The next result generalizes the fluid limit provided in [3, Chapter 11] for density dependent population processes: the case where the maps $\delta_l^k(x)$ are identically zero. The proof is similar, see [7, Theorem 2.2.5] for details.

Theorem A.3. Suppose that the maps $\gamma_l(x)$ are bounded on compact sets, that the drift $F(x)$ is locally Lipschitz and that for each compact $K \subset E$:

$$\begin{aligned} &\sup_{x \in K} |\delta_l^k(x)| < \infty \quad \text{for all} \quad l \in D \quad \text{and} \quad k \geq 1, \\ &\lim_{k \to \infty} \sup_{x \in K} |\delta_l^k(x)| = 0 \quad \text{for all} \quad l \in D. \end{aligned} \tag{A.2}$$

In addition, assume that $X_k(0) \to x_0 \in E$ as $k \to \infty$, and that there exists a unique solution $x : [0, T] \longrightarrow E$, with initial condition x_0, to $\dot{x} = F(x)$. Then

$$\sup_{t \in [0,T]} ||X_k(t) - x(t)|| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad k \to \infty.$$

Condition (A.2) means that the maps $\delta_l^k(x)$ vanish in the fluid scale. As a result, the fluid limit $x(t)$ is determined by the homogeneous drift alone, the inhomogeneous drift has no impact on it. The most interesting behavior corresponds to the case where the maps $\delta_l^k(x)$ disappear in the fluid scale but not in the diffusion scale, as below.

Theorem A.4. Suppose that the hypotheses of Theorem A.3 hold, that the maps $\gamma_l(x)$ are locally Lipschitz and that for each compact $K \subset E$ we have

$$\lim_{k \to \infty} \sup_{x \in K} k^\alpha |\delta_l^k(x)| = 0 \quad \text{for all} \quad \alpha \in [0, 1/2) \quad \text{and} \quad l \in D.$$

Let us define $Z_k(t) = \sqrt{k}[X_k(t) - x(t)]$. Also, assume that $F(x)$ is continuously differentiable and that there exists a continuous vector field $G : E \longrightarrow \mathbb{R}^d$ such that for each compact $K \subset E$ we have

$$\lim_{k \to \infty} \sup_{x \in K} \left|\left| \sqrt{k} G_k(x) - G(x) \right|\right| = 0 \quad \text{for all} \quad l \in D.$$

Suppose that $Z_k(0) \to Z_0 \in \mathbb{R}^d$ as $k \to \infty$, and let Z be the solution to

$$dZ_t = [F'(x(t))Z_t + G(x(t))]dt + B_t dW_t \tag{A.3}$$

with initial condition Z_0, where $F'(x)$ is the Jacobian matrix of the homogeneous drift, W is a d-dimensional standard Wiener process and

$$B_t = \sqrt{\sum_{l \in D} ll^T \gamma_l(x(t))}.$$

Then $Z_k \Rightarrow Z$ in the Skorokhod space $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$.

This diffusion limit has its counterpart in [3, Chapter 11] as a special case where $G(x)$ is identically zero. When the inhomogeneous drifts behave asymptotically[5] as $1/\sqrt{k}$, equation (A.3) has the additional term $G(x(t))dt$, the effect of the inhomogeneous drifts on the processes is captured by the diffusion limit, even though it is not apparent in the fluid limit. This has interesting consequences for applications, as in Section 4 for example.

Proof. Let us introduce the process

$$U_k(t) = \sum_{l \in D} \frac{l}{\sqrt{k}} Y_l \left(\int_0^t k\beta_l^k(X_k(\tau))d\tau \right). \tag{A.4}$$

Using equation (A.1) and the definition of $Z_k(t)$ we see that

$$Z_k(t) = Z_k(0) + U_k(t) + \int_0^t \sqrt{k}[F(X_k(\tau)) - F(x(\tau))]d\tau + \int_0^t \sqrt{k}G_k(X_k(\tau))d\tau.$$

Furthermore, we have

$$\begin{aligned}\sqrt{k}[F(X_k(t)) - F(x(t))] &= \sqrt{k}F'(x(t))[X_k(t) - x(t)] + \sqrt{k}R_t(X_k(t)) \\ &= F'(x(t))Z_k(t) + \sqrt{k}R_t(X_k(t)),\end{aligned}$$

where $R_t(x)$ is the first order Taylor remainder of $F(x)$ at the point $x(t)$.

Introducing the process

$$\nu_k(t) = \int_0^t \sqrt{k}\left[G_k(X_k(\tau)) + R_\tau(X_k(\tau))\right] d\tau,$$

we may write the following equation:

$$Z_k(t) = Z_k(0) + U_k(t) + \nu_k(t) + \int_0^t F'(x(\tau))Z_k(\tau)d\tau. \tag{A.5}$$

The strategy is the same as in [3]. We will write Z_k as a continuous function of $V_k = Z_k(0) + U_k + \nu_k$ and we will prove that V_k has a limit in distribution, the statement will then follow from the continuous mapping theorem.
Claim I: Z_k is a continuous function of V_k.

Let $\Gamma(s,t)$ be the unique solution to

$$\frac{\partial \Gamma}{\partial t}(s,t) = F'(x(t))\Gamma(s,t) \quad \text{and} \quad \Gamma(s,s) = \mathrm{Id} \quad \text{for all} \quad s,t \in [0,T].$$

[5]The hypothesis concerning the maps $\delta_l^k(x)$ is slightly weaker, terms of higher order could cancel when one computes $G_k(x)$.

For each $f \in D_{\mathbb{R}^d}[0,T]$ there exists a unique $\phi_f \in D_{\mathbb{R}^d}[0,T]$ such that

$$\phi_f(t) = f(t) + \int_0^t F'(x(\tau))\phi_f(\tau)d\tau \quad \text{for all} \quad t \in [0,T].$$

This function may be given explicitly in terms of f as follows:

$$\phi_f(t) = f(t) + \int_0^t \Gamma(\tau,t)F'(x(\tau))f(\tau)d\tau.$$

The corresponding mapping $\phi : D_{\mathbb{R}^d}[0,T] \longrightarrow D_{\mathbb{R}^d}[0,T]$ is continuous in the Skorokhod topology; we refer to [7, Lemma 2.3.4]. Since $V_k, Z_k \in D_{\mathbb{R}^d}[0,T]$, then $Z_k = \phi(V_k)$ by definition of ϕ, and Claim I follows.
Claim II: $V_k = Z_k(0) + U_k + \nu_k$ has a limit in distribution in $D_{\mathbb{R}^d}[0,T]$.

Let $\{W_l\}_{l\in D}$ be independent Wiener processes, and let us define

$$V(t) = Z_0 + U(t) + \int_0^t G(x(\tau))d\tau \quad \text{with} \quad U(t) = \sum_{l\in D} lW_l\left(\int_0^t \gamma_l(x(\tau))d\tau\right).$$

From Theorem A.3 and the functional central limit theorem for the Poisson process we see that $U_k \Rightarrow U$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$. Moreover, we also have

$$\sup_{t\in[0,T]} \left|\left|\nu_k(t) - \int_0^t G(x(\tau))d\tau\right|\right| \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad k \to \infty.$$

The proofs of these facts are provided in [7, Theorem 3.2.3, Lemma 3.2.4], respectively, and are technical. From [7, Proposition A.1.8] it follows that the convergence of U_k in distribution and the convergence of ν_k in probability, to a deterministic process, imply $V_k \Rightarrow V$ in $D_{\mathbb{R}^d}[0,T]$, which establishes the claim.

Invoking the continuous mapping theorem we see that $Z_k = \phi(V_k)$ converges in distribution to $Z = \phi(V)$, which by definition of ϕ is the unique continuous process such that

$$Z(t) = Z_0 + U(t) + \int_0^t \left[F'(x(\tau))Z(\tau) + G(x(\tau))\right]d\tau \quad \text{for all} \quad t \in [0,T].$$

This process has the same law as the solution of (A.3). □

We note that equation (A.3) defines a time inhomogeneous Gaussian process. Its mean is the solution to the integral equation

$$\mu(t) = Z_0 + \int_0^t \left[F'(x(\tau))\mu(\tau) + G(x(\tau))\right]d\tau,$$

and its covariance is the same as in the diffusion limit of [3], namely:

$$\Sigma(s,t) = \int_0^s \Gamma(\tau,s)B_\tau B_\tau^T \Gamma(\tau,t)d\tau \quad \text{for all} \quad 0 \le s < t \le T;$$

for details see [7, Section 3.2].

Appendix B. Extension for pseudo-differentiable drifts

Here we provide a version of Theorem A.4 for the case where $F(x)$ is not differentiable. Specifically, we suppose that the nominal solution to the fluid dynamics is an equilibrium point, and we obtain a diffusion limit assuming only that $F(x)$ admits a *pseudo-differential* at this point.

Definition B.1. A Lipschitz field $\partial F : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ is a pseudo-differential of F at x if the subsequent conditions hold.

1. $\partial F(y)$ is positively homogeneous, in the sense that $\partial F(\alpha y) = \alpha \partial F(y)$ for all $y \in \mathbb{R}^d$ and all $\alpha \geq 0$.
2. The remainder $R(y) = F(y) - F(x) - \partial F(y - x)$ is such that

$$\lim_{y \to x} \frac{R(y)}{||y - x||} = 0.$$

The definition is intended to cover fields F which exhibit switching around the point x between a set of differentiable functions; this includes the important special case of piecewise linear fields, which appear in this paper. It is easy to check that ∂F is unique when it exists, and such that

$$\partial F(y) = \lim_{h \to 0^+} \frac{F(x + hy) - F(x)}{h}.$$

An explicit formula for ∂F is available for a field F that has lateral directional derivatives in a set of basis directions; see [7, Proposition 3.4.2].

We tackle now the extension of the diffusion results to this setting. Suppose that the fluid dynamics $\dot{x} = F(x)$ have an equilibrium at $x^* \in E$, and that $X_k(0) \to x^*$ as $k \to \infty$. By Theorem A.3 this implies that

$$\sup_{t \in [0,T]} ||X_k(t) - x^*|| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad k \to \infty \quad \text{for all} \quad T \geq 0.$$

Theorem B.2. Suppose that the maps $\gamma_l(x)$ are locally Lipschitz, that $F(x)$ has a pseudo-differential $\partial F(y)$ at x^* and that for each compact $K \subset E$:

$$\sup_{x \in K} |\delta_l^k(x)| < \infty \quad \text{for all} \quad l \in D \quad \text{and} \quad k \geq 1,$$
$$\lim_{k \to \infty} \sup_{x \in K} k^\alpha |\delta_l^k(x)| = 0 \quad \text{for all} \quad \alpha \in [0, 1/2) \quad \text{and} \quad l \in D.$$

Assume that there exists a continuous field $G : E \longrightarrow \mathbb{R}^d$ such that

$$\lim_{k \to \infty} \sup_{x \in K} \left|\left| \sqrt{k} G_k(x) - G(x) \right|\right| = 0$$

for all compact sets $K \subset E$, and that $Z_k(0) \to Z_0 \in \mathbb{R}^d$ as $k \to \infty$. In addition, let Z be the solution to

$$dZ_t = [\partial F(Z_t) + G(x^*)]dt + B dW_t \tag{B.1}$$

with initial condition Z_0, where W is a d-dimensional Wiener process and

$$B = \sqrt{\sum_{l \in D} l l^T \gamma_l(x^*)}.$$

Then $Z_k \Rightarrow Z$ in the Skorokhod space $D_{\mathbb{R}^d}[0,\infty)$ as $k \to \infty$.

Proof. By [38, Theorem 16.7], the convergence in $D_{\mathbb{R}^d}[0,\infty)$ will be established if we show that $Z_k \Rightarrow Z$ in $D_{\mathbb{R}^d}[0,T]$ for an arbitrary $T \geq 0$.

Recalling (A.1), the definition of Z_k leads to

$$Z_k(t) = Z_k(0) + U_k(t) + \int_0^t \sqrt{k}[F(X_k(\tau)) - F(x^*)]d\tau + \int_0^t \sqrt{k} G_k(X_k(\tau))d\tau,$$

where U_k is defined as in equation (A.4) and we note that $F(x^*) = 0$. Also,

$$\begin{aligned}\sqrt{k}[F(X_k(t)) - F(x^*)] &= \sqrt{k}\partial F(X_k(t) - x^*) + \sqrt{k}R(X_k(t)) \\ &= \partial F(Z_k(t)) + \sqrt{k}R(X_k(t)),\end{aligned}$$

where $R(y)$ is the remainder that appears in Definition B.1 and the last equality follows from the positive homogeneity of $\partial F(y)$.

Consider now the process

$$\nu_k(t) = \int_0^t \sqrt{k}\left[G_k(X_k(\tau)) + R(X_k(\tau))\right] d\tau.$$

The following equation is the analog of (A.5):

$$Z_k(t) = Z_k(0) + U_k(t) + \nu_k(t) + \int_0^t \partial F(Z_k(\tau))d\tau.$$

The strategy now is as in the proof of Theorem A.4. Namely, we will show that Z_k is a continuous function of $V_k = Z_k(0) + U_k + \nu_k$ and that V_k has a limit in distribution, to then apply the continuous mapping theorem. The main difference appears in the proof of Claim I of Theorem A.4. There we were able to use the continuous differentiability of F to provide an explicit expression for the mapping ϕ, simplifying the analysis; this is no longer possible.
Claim I: For each $f \in D_{\mathbb{R}^d}[0,T]$ there exists a unique $\phi_f^T \in D_{\mathbb{R}^d}[0,T]$ such that

$$\phi_f^T(t) = f(t) + \int_0^t \partial F\left(\phi_f^T(\tau)\right) d\tau \quad \text{for all} \quad t \in [0,T].$$

Furthermore, the mapping $\phi^T : D_{\mathbb{R}^d}[0,T] \longrightarrow D_{\mathbb{R}^d}[0,T]$ is continuous.

The proof of the claim is provided in [33] for $d = 1$; for the general case we refer to [7, Section 3.3]. Since $V_k, Z_k \in D_{\mathbb{R}^d}[0,T]$ then $Z_k = \phi^T(V_k)$.

The limit in distribution of V_k is as in Claim II of Theorem A.4. Consider an independent family $\{W_l\}_{l \in D}$ of Wiener processes and let

$$V(t) = Z_0 + U(t) + tG(x^*) \quad \text{with} \quad U(t) = \sum_{l \in D} l W_l\left(t\gamma_l(x^*)\right).$$

By [7, Theorem 3.2.3] we have $U_k \Rightarrow U$ in $D_{\mathbb{R}^d}[0,T]$ and by [7, Lemma 3.2.4]

$$\sup_{t\in[0,T]} ||\nu_k(t) - tG(x^*)|| \xrightarrow{\mathbb{P}} 0 \quad \text{as} \quad k \to \infty.$$

It follows that $V_k \Rightarrow V$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$, by [7, Proposition A.1.8].

Using Claim I we may define Z such that $Z|_{[0,t]} = \phi^t\left(V|_{[0,t]}\right)$ for all $t \geq 0$, and this process satisfies the integral equation

$$Z(t) = Z_0 + U(t) + \int_0^t [\partial F(Z(\tau)) + G(x^*)]\, d\tau \quad \text{for all} \quad t \geq 0.$$

So Z has the same law as the solution to (B.1). Since $Z|_{[0,T]} = \phi^T\left(V|_{[0,T]}\right)$, the continuous mapping theorem yields: $Z_k \Rightarrow Z$ in $D_{\mathbb{R}^d}[0,T]$ as $k \to \infty$. □

Appendix C. Proof of Theorem 4.2

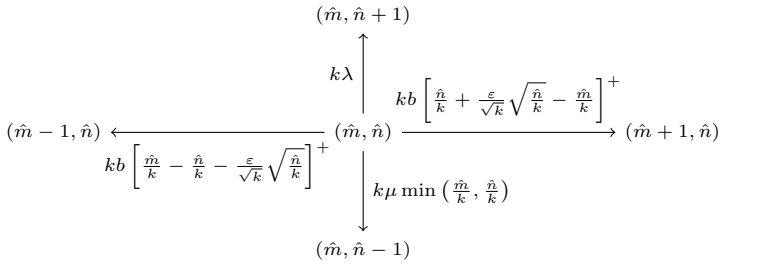


Figure C.12: Transition rates of the processes $\hat{X}_k$.

We begin by writing down the intensities of the processes $\hat{X}_k$ of Section 4, these are shown in Fig. C.12. The intensities of the normalization $X_k = \hat{X}_k/k$ are obtained through the change of variables $m = \hat{m}/k$ and $n = \hat{n}/k$; they can be expressed in terms of maps $\beta_l^k = \gamma_l + \delta_l^k$ on $\mathbb{R}^2_+$, explicitly:

$$\gamma_l(m,n) = \begin{cases} b[n-m]^+ & \text{if} \quad l = (1,0), \\ b[m-n]^+ & \text{if} \quad l = -(1,0), \\ \lambda & \text{if} \quad l = (0,1), \\ \mu \min(m,n) & \text{if} \quad l = -(0,1), \end{cases} \quad \text{and}$$

$$\delta_l^k(m,n) = \begin{cases} b\left[n + \frac{\varepsilon}{\sqrt{k}}\sqrt{n} - m\right]^+ - b[n-m]^+ & \text{if} \quad l = (1,0), \\ b\left[m - n - \frac{\varepsilon}{\sqrt{k}}\sqrt{n}\right]^+ - b[m-n]^+ & \text{if} \quad l = -(1,0), \\ 0 & \text{if} \quad l = (0,1), \\ 0 & \text{if} \quad l = -(0,1). \end{cases}$$

Furthermore, the homogeneous and inhomogeneous drifts are, respectively,

$$F(m,n) = \begin{bmatrix} b(n-m) \\ \lambda - \mu \min(m,n) \end{bmatrix} \quad \text{and} \quad G_k(m,n) = \frac{b\varepsilon}{\sqrt{k}} \begin{bmatrix} \sqrt{n} \\ 0 \end{bmatrix}.$$

In order to prove Theorem 4.2, it is enough to verify that all the hypotheses of Theorem B.2 hold. To begin, we observe that $\gamma_l(m,n)$ is Lipschitz and

$$\left|\delta_l^k(m,n)\right| \leq \frac{\varepsilon}{\sqrt{k}}\sqrt{n} \quad \text{for all} \quad (m,n) \in \mathbb{R}_+^2,$$

so the supremum of $\left|\delta_l^k(m,n)\right|$ on a compact set is $O\left(1/\sqrt{k}\right)$. In addition,

$$\sqrt{k}G_k(m,n) = b\varepsilon \begin{bmatrix} \sqrt{n} \\ 0 \end{bmatrix} = G(m,n) \quad \text{for all} \quad (m,n) \in \mathbb{R}_2^+,$$

and last, $F(m,n)$ admits a pseudo-differential at (ρ,ρ), given by

$$\partial F(m,n) = \begin{bmatrix} -b & b \\ -\mu & 0 \end{bmatrix} \begin{bmatrix} m \\ n \end{bmatrix} \quad \text{if} \quad m > n, \quad \partial F(m,n) = \begin{bmatrix} -b & b \\ 0 & -\mu \end{bmatrix} \begin{bmatrix} m \\ n \end{bmatrix} \quad \text{else.}$$

Acknowledgments

The authors were supported by ANII under grants POS_NAC_2016_1_130333, FCE_1_2017_1_136748 and FCE_1_2019_1_156666.

References

[1] A. Borovkov, On limit laws for service processes in multi-channel systems, Siberian Mathematical Journal 8 (5) (1967) 746–763.

[2] M. Lin, A. Wierman, L. L. Andrew, E. Thereska, Dynamic right-sizing for power-proportional data centers, IEEE/ACM Transactions on Networking (TON) 21 (5) (2013) 1378–1391.

[3] S. N. Ethier, T. G. Kurtz, Markov processes: characterization and convergence, Vol. 282, John Wiley & Sons, 2009.

[4] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Operations research 29 (3) (1981) 567–588.

[5] T. G. Kurtz, et al., Strong approximation theorems for density dependent markov chains, Stochastic Processes and their Applications 6 (3) (1978) 223–240.

[6] D. Goldsztajn, A. Ferragut, F. Paganini, Feedback control of server instances for right sizing in the cloud, in: 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, 2018, pp. 749–756.

[7] D. Goldsztajn, Limit theorems for continuous time markov chains and applications to large scale queueing systems, Master's thesis, Universidad de la República, Montevideo, Uruguay (2018).
URL http://archive.cmat.edu.uy/biblioteca/documentos/tesis/maestria/Gol2018

[8] L. M. Nguyen, A. L. Stolyar, A service system with randomly behaving on-demand agents, ACM SIGMETRICS Performance Evaluation Review 44 (1) (2016) 365–366.

[9] L. M. Nguyen, A. L. Stolyar, A queueing system with on-demand servers: local stability of fluid limits, Queueing Systems 89 (3-4) (2018) 243–268.

[10] F. Yao, A. Demers, S. Shenker, A scheduling model for reduced CPU energy, in: Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on, IEEE, 1995, pp. 374–382.

[11] S. Albers, H. Fujiwara, Energy-efficient algorithms for flow time minimization, ACM Transactions on Algorithms (TALG) 3 (4) (2007) 49.

[12] K. Pruhs, P. Uthaisombut, G. Woeginger, Getting the best response for your erg, ACM Transactions on Algorithms (TALG) 4 (3) (2008) 38.

[13] N. Bansal, H.-L. Chan, K. Pruhs, Speed scaling with an arbitrary power function, in: Proceedings of the twentieth annual ACM-SIAM symposium on discrete algorithms, Society for Industrial and Applied Mathematics, 2009, pp. 693–701.

[14] L. L. Andrew, M. Lin, A. Wierman, Optimality, fairness, and robustness in speed scaling designs, ACM SIGMETRICS Performance Evaluation Review 38 (1) (2010) 37–48.

[15] A. Wierman, L. L. Andrew, A. Tang, Power-aware speed scaling in processor sharing systems: Optimality and robustness, Performance Evaluation 69 (12) (2012) 601–622.

[16] L. Chen, N. Li, On the interaction between load balancing and speed scaling, IEEE Journal on Selected Areas in Communications 33 (12) (2015) 2567–2578.

[17] D. Goldsztajn, A. Ferragut, F. Paganini, A feedback control approach to dynamic speed scaling in computing systems, in: 2017 51st Annual Conference on Information Sciences and Systems (CISS), IEEE, 2017, pp. 1–6.

[18] D. Gamarnik, J. N. Tsitsiklis, M. Zubeldia, Delay, memory, and messaging tradeoffs in distributed service systems, ACM SIGMETRICS Performance Evaluation Review 44 (1) (2016) 1–12.

[19] S. Foss, A. L. Stolyar, Large-scale join-idle-queue system with general service times, Journal of Applied Probability 54 (4) (2017) 995–1007.

[20] V. Gupta, N. Walton, Load balancing in the non-degenerate slowdown regime, arXiv preprint arXiv:1707.01969.

[21] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, A. Greenberg, Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services, Performance Evaluation 68 (11) (2011) 1056–1071.

[22] M. Mitzenmacher, The power of two choices in randomized load balancing, IEEE Transactions on Parallel and Distributed Systems 12 (10) (2001) 1094–1104.

[23] N. D. Vvedenskaya, R. L. Dobrushin, F. I. Karpelevich, Queueing system with selection of the shortest of two queues: An asymptotic approach, Problemy Peredachi Informatsii 32 (1) (1996) 20–34.

[24] M. van der Boor, S. C. Borst, J. S. van Leeuwaarden, D. Mukherjee, Scalable load balancing in networked systems: A survey of recent advances, arXiv preprint arXiv:1806.05444.

[25] D. Mukherjee, S. Dhara, S. C. Borst, J. S. van Leeuwaarden, Optimal service elasticity in large-scale distributed systems, Proceedings of the ACM on Measurement and Analysis of Computing Systems 1 (1) (2017) 25.

[26] D. Goldsztajn, A. Ferragut, F. Paganini, M. Jonckheere, Controlling the number of active instances in a cloud environment, ACM SIGMETRICS Performance Evaluation Review 45 (2) (2018) 15–20.

[27] F. Paganini, D. Goldsztajn, A. Ferragut, An optimization approach to load balancing, scheduling and right sizing of cloud computing systems with data locality, in: 2019 IEEE 58th Conference on Decision and Control (CDC), IEEE, 2019, pp. 1114–1119.

[28] B. Wilder, Cloud architecture patterns: using microsoft azure, " O'Reilly Media, Inc.", 2012.

[29] Amazon Web Services, What is a Message Queue.
URL `https://aws.amazon.com/message-queue/`

[30] Microsoft, Introduction to Azure Queue Storage.
URL `https://docs.microsoft.com/en-us/azure/storage/queues/storage-queues-introduction`

[31] T. G. Kurtz, Solutions of ordinary differential equations as limits of pure jump markov processes, Journal of applied Probability 7 (1) (1970) 49–58.

[32] T. G. Kurtz, Limit theorems for sequences of jump markov processes, J Appl Probab 8 (2) (1971) 344–356.

[33] G. Pang, R. Talreja, W. Whitt, et al., Martingale proofs of many-server heavy-traffic limits for markovian queues, Probability Surveys 4 (2007) 193–267.

[34] N. Gast, B. Gaujal, Markov chains with discontinuous drifts have differential inclusion limits, Performance Evaluation 69 (12) (2012) 623–642.

[35] L. Bortolussi, Hybrid limits of continuous time markov chains, in: 2011 Eighth International Conference on Quantitative Evaluation of SysTems, IEEE, 2011, pp. 3–12.

[36] L. Bortolussi, Hybrid behaviour of markov population models, Information and Computation 247 (2016) 37 – 86.

[37] P. Robert, Stochastic networks and queues, Springer-Verlag, Berlin, 2003.

[38] P. Billingsley, Convergence of probability measures, 2nd Edition, John Wiley & Sons, 1999.