Feedback control of server instances for right sizing in the cloud

Diego Goldsztajn, Andres Ferragut and Fernando Paganini
Universidad ORT Uruguay

***Abstract*— We consider a computing system based on summoning server instances on the fly, possibly from a remote cloud service. A feedback rule must be designed to track the exogenous load with the right service capacity, taking into account the inherent lags in server creation and deletion. We use fluid and diffusion queueing models to analyze control schemes that manage the tradeoff between job queueing and idle capacity, in the large scale limit. In particular we propose a method in which the system can achieve negligible queueing while minimizing idle capacity. Theoretical results are supported by simulations.**

I. Introduction

With the emergence of the cloud computing model, computing tasks are nowadays often performed by a mutualized infrastructure, where processing capacity, memory and storage can be summoned dynamically by the customers from the cloud service provider. Vendors like Google or Amazon offer a large catalog of computing instances that can be spawned to match a given load of requests on the fly.

In this context, consider a facility offering computing services to an exogenous customer load, by either managing a local server system, or alternatively contracting servers from an external cloud infrastructure. In either case, cost considerations dictate that the active (or contracted) capacity must be "right-sized" to the external demand [9]: a shortage of capacity causes queueing delay in the tasks to be performed, while over-provisioning naturally incurs extra costs. The desirable operating point is with load matching capacity.

In the queueing literature, this situation has been termed *heavy traffic* [4]; classical results (see [7] and references therein) describe the asymptotic behavior of large scale systems with multiple servers, in the limit where load reaches the boundary of their service capacity. From a practical perspective, however, it is highly unlikely that the exogenous demand (measured e.g. in Erlangs of offered traffic) would coincidentally match *exactly* the available service capacity. What makes this situation relevant is the more likely scenario of load matching capacity as a result of the latter being *actively controlled* to follow the (uncertain) load; i.e. when servers are turned on (or remotely summoned) as needed for the current demand. Thus, service capacity becomes *variable* and a feedback rule must be implemented to control it.

A separate aspect of cloud computing is load balancing: how arriving tasks should be distributed between the deployed servers. The very active and recent literature on this problem (e.g. [3]), again typically poses the question in terms of an inelastic number of instances that grows with the scale of the system. Under appropriate scaling conditions, fluid or diffusion approximations to the corresponding stochastic queues can be derived, leading to interesting conclusions on the performance of different job scheduling policies, such as join-the-idle-queue and its variants which have been thoroughly analyzed in this fashion [2], [6], [10], [11].

In this paper, we focus on the problem of controlling the *number* of active computing instances in feedback with the current system occupation. The simplest, classical model of variable service capacity is the infinite-server queue: here for each arriving job, a new server is summoned to take care of it, and the server disappears upon job completion. A computing system that could create/destroy server instances quickly and with no penalty would adapt to an arbitrary uncertain load, with no idle capacity or queueing delay. However, practical considerations stand in the way of such fast control of server instances. In the local server scenario, it takes some time to activate a server which is currently turned off; in the remote case there is also a delay in the response of a cloud provider. Finding adequate control rules under this delayed scenario is the subject of this paper. These control rules can then be matched with suitable load balancing mechanisms. Relevant references in this regard are [5], [12]–[14].

We begin in Section II with a two-state, switched differential equation model of a fluid queue under controlled capacity, summoned with a first-order lag; we show the desired equilibrium is globally stable. In a stochastic setting fluctuations will arise; to analyze these we look at the associated Markov chain model, and its diffusion limit; we find undesirable job queueing. This motivates us to consider in Section III a variation of the control to ensure nearly zero queueing; we analyze the corresponding fluid and diffusion models for this alternative. In Section IV we refine the control rule to achieve this reduction in queue length with minimal over-provisioning, making the system work automatically in a similar way to the well known Halfin-Whitt regime [7]. In Section V we discuss implementation issues. Conclusions are provided in Section VI and some proofs are deferred to the Appendix.

II. On demand servers with a startup lag

We consider a queueing system with arrivals at rate λ jobs/sec, served by a set of computing instances, each with individual capacity μ jobs/sec. The offered traffic load will be denoted $\rho = \lambda/\mu$ Erlangs, i.e. the number of server instances required to cover the load on average.

This work was partially supported by ANII–Uruguay under grants FCE_1_2017_1_136748 and SNB POS_NAC_2016_1_130333.

E-mail: goldsztajn@ort.edu.uy.

At a given time t, there will be $n(t)$ jobs present in the system, and $m(t)$ active servers. Note that $n > m$ means there are $n-m$ jobs waiting with no service, whereas $m > n$ means there are $m-n$ idle servers. In either case $\min\{m, n\}$ is the number of working servers in the system.

A. Differential equation model

Taking these variables as fluid (real-valued) quantities, our basic model for the queue dynamics is

$$\dot{n} = \lambda - \mu \min\{m, n\}. \tag{1}$$

The right-hand side of this ODE is continuous and Lipschitz, but nonsmooth; switching occurs at the line $m = n$.

Assuming an unknown, and perhaps variable, arrival rate of jobs, the active server pool $m(t)$ must be *controlled* in real time to cover the demand. Since idle servers consume resources, and queueing jobs penalizes customers, our ideal objective would be to keep $m(t) = n(t)$ at all times. As mentioned in the introduction, this corresponds to the infinite-server queue that simply summons or kills server instances triggered by job arrivals and departures.

The main hurdle that we encounter is, however, the natural delay in the control of server instances; the simplest possible model is, in transfer function notation, the first order lag

$$\hat{m}(s) = \frac{1}{1+\tau s}\hat{n}(s)$$

with delay parameter τ. Equivalently, letting $\beta = 1/\tau$, we have the time-domain model

$$\dot{m} = \beta(n - m). \tag{2}$$

Our first model for the full dynamics is thus (1)-(2). We are implicitly assuming the dynamics are constrained to the quadrant $\{m \geq 0, n \geq 0\}$; note that the field naturally points inwards at the boundary. For a given constant λ, the dynamics (1)-(2) have a unique equilibrium point x^* with coordinates $m^* = \rho$ and $n^* = \rho$, as desired. We begin by showing that this equilibrium is a global attractor.

Proposition 2.1: The above equilibrium point $x^* = (\rho, \rho)$ of equation (1)-(2) is globally asymptotically stable.

The proof relies on observing that the dynamics are piecewise-linear, switching at the line $m = n$, and finding a common quadratic Lyapunov function for the dynamics in each subset; we defer it to the Appendix.

An alternative, and somewhat more complex model of the server dynamics, would be

$$\dot{m} = \beta[n - m]^+ - \gamma[m - n]^+ \quad \text{with } \beta, \gamma > 0. \tag{3}$$

The rationale is that the lags $1/\beta$ for server creation and $1/\gamma$ for server destruction could possibly differ. The joint dynamics (1)-(3) are still switching only at the line $m = n$, and have the same equilibrium (ρ, ρ). In this case we do not always have a common quadratic Lyapunov function for the two switched fields, but still it is possible to establish that the equilibrium is a global attractor, the proof is similar to that of Proposition 3.1 below.

B. Modeling stochastic fluctuations

The preceding study shows that, at a macroscopic level, a simple first-order model for the control of server instances appears to achieve the purpose of matching load, leaving no idle servers or queues. In practice, however, fluctuations will occur around the equilibrium values, which warrants a closer look at the dynamics from a microscopic perspective.

A natural stochastic model of our queuing system is to write the continuous time Markov chain $X = (M, N)$ that corresponds to our fluid dynamics; this process with state-space $\mathbb{N}^2$ has the transition rates depicted in Fig. 1.

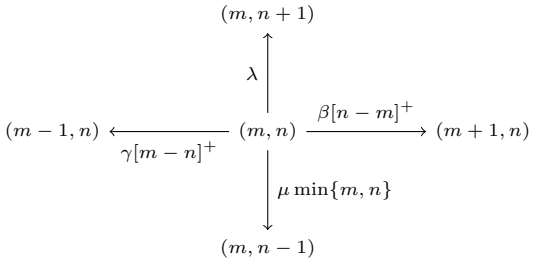

Fig. 1. Transition rates of the Markov chain model.

Here the integer number of jobs n is driven by a Poisson process of arrivals with intensity λ, and service times are exponential with mean $1/\mu$. The server dynamics can be interpreted as follows: a new server is summoned whenever a job has to be queued; the setup[1] time of this server is exponential with mean $\tau = 1/\beta$. Also, each idle server shuts down after an $\exp(\gamma)$ deletion time. The stochasticity in these transition times is especially justified in a cloud environment, where server creation is invoked by a load scaling feature, which is then executed in the infrastructure, where availability is uncertain.

The connection between the Markov and differential equation models can be formalized through a standard fluid limit procedure: define $X_l = (M_l, N_l)$ to be a continuous time Markov chain such that M_l and N_l correspond to the number of active servers and jobs, respectively, for an arrival rate of $l\lambda$. Letting the scale parameter $l \to \infty$ we can model the behavior of large scale systems. Consider now the normalized processes $\bar{X}_l = X_l/l$. Since the field of the dynamics (1)-(3) is Lipschitz, we can establish the following result as a straightforward consequence of the strong law-of-large-numbers for density dependent families of continuous time Markov chains (see [1, Chapter 11]).

Theorem 2.2: Assume that the deterministic initial conditions of the processes $\bar{X}_l$ converge to some $x_0 \in [0, +\infty)^2$ as $l \to \infty$, and let $x(t)$ be the unique solution to (1)-(3) when the initial condition is x_0.

$$\sup_{t \in [0,T]} \left|\left| \bar{X}_l(t) - x(t) \right|\right| \xrightarrow{a.s.} 0 \quad \text{as } l \to \infty \quad \forall\, T \geq 0.$$

[1]If one of the active servers becomes available, this initialization is cancelled before the setup is finished. This keeps the number of summoned servers aligned with $n - m$.

The above law-of-large-numbers type limit helps justify our macroscopic ODE model; however it has removed all stochasticity from the dynamics. To retain a view of local fluctuations around the equilibrium of the macroscopic model one can carry out a *diffusion* limit of the Markov chain by considering the processes $Z_l = \sqrt{l}(\bar{X}_l - x^*)$, which models fluctuations around the equilibrium on the scale of $\sqrt{l}$. We have the following result.

Theorem 2.3: Assume that the deterministic initial conditions of the processes Z_l converge to some $Z_0 \in \mathbb{R}^2$ as $l \to \infty$. Then, the processes Z_l converge weakly, in the Skorokhod space $D_{\mathbb{R}^2}[0,\infty)$, to the process $Z = (\tilde{M}, \tilde{N})$, whose initial condition is Z_0, and solves the following stochastic differential equation (SDE).

$$d\tilde{M} = \left[\beta(\tilde{N} - \tilde{M})^+ - \gamma(\tilde{M} - \tilde{N})^+\right] dt, \tag{4a}$$

$$d\tilde{N} = -\mu \min\{\tilde{M}, \tilde{N}\}dt + \sqrt{2\lambda}dW. \tag{4b}$$

Here $W(t)$ is a standard one-dimensional Wiener process.

The diffusion limit for a Markov chain, around a given fluid limit, is covered in the classical work of Kurtz [8]: the construction rule is to replace each transition of rate $q_i(\tilde{M}, \tilde{N})$ of Fig. 1 with a drift term $q_i dt$ plus a noise term $\sqrt{q_i}dW_i$, where the $W_i(t)$ are independent Wiener processes. These noise coefficients are evaluated at the nominal solution $x(t)$ of the fluid dynamics; in our case we are picking the equilibrium solution $x(t) \equiv (\rho, \rho)$, which implies the noise terms in (4a) disappear. Also the two noise terms in (4b) have been combined into one with twice the variance.

The Proof of Theorem 2.3 requires an extension of the results in [8], due to the nonsmooth vector field of the dynamics (1)-(3); this is discussed in the Appendix.

Finding the stationary distribution of Z in (4) would help us understand the steady-state behavior of X, and compute estimates for the main performance metrics of interest: the mean queue length $\mathbb{E}[N - M]^+$, and the mean number of idle servers in the system $\mathbb{E}[M - N]^+$. Unfortunately, the nonlinear switching in (4) precludes us from finding closed form expressions for this steady-state distribution, or the above expectations. Still, from the symmetric nature of the noise we can expect both of these metrics to be nonzero.

This is confirmed by the simulations shown in Fig. 2. Our system incurs both performance penalties: on one hand, the systematic appearance of queueing means that a non-trivial number of jobs must be held before servers become available for them; on the other, idle servers imply an excess provisioning cost. The relative size of these two factors is dictated by the parameter ratio β/γ. Indeed, since the drift of X equals zero in steady-state, then looking at the drift in the m-direction we see that

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

Now, since the above ratio depends on time lags inherent to the system and not under our direct control, we cannot use it as a way of managing the queue length/over-provisioning tradeoff. In the next section, we will discuss an alternative

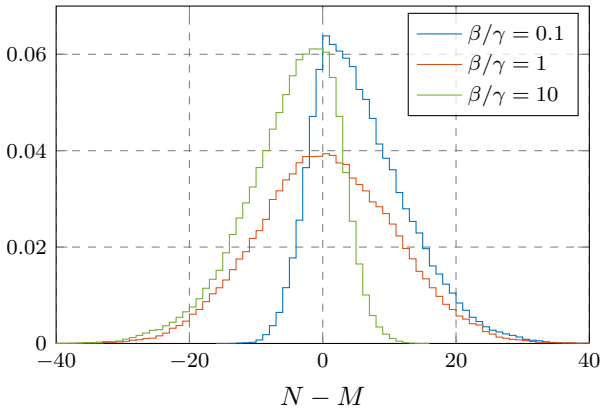

Fig. 2. Empirical distribution of $N - M$ in steady state for the Markov chain of Fig 1. Different values of β/γ, $\lambda = 1000, \mu = 1$.

control strategy for this purpose. For simplicity, then, we will assume $\beta = \gamma$ in the remainder of the paper.

The question arises as to which of the two penalties is more troublesome from a practical perspective. We will aim at the (almost) complete elimination of queueing. The rationale is that in these cloud-based systems the entity in control of the server dynamics is a *dispatcher* which may not have enough local storage, and/or would rather avoid the overhead of holding jobs; this is in line with recent literature on the subject [14].

III. Controlling for zero queue-length

In the search for a variant of our instance controlling rule, we return to the fluid setting of Section II-A. We first note that (1) is our physical model of the queue and cannot be modified; this implies that our equilibrium point will necessarily verify $\min\{m^*, n^*\} = \rho$. Thus, the potential equilibria lie in the L-shaped curve depicted in Fig. 3.

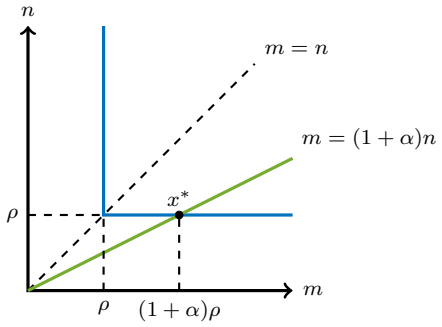

Fig. 3. Feasible equilibria and the result of control (5).

The degree of freedom available to us is the provisioning rule for summoning servers; still, this rule must preserve the time lags that are part of our physical constraints. With this in mind, we propose the following variant of (2). For some constant $\alpha > 0$ define the server provisioning dynamics:

$$\dot{m} = \beta[(1+\alpha)n - m]. \tag{5}$$

The main idea in equation (5) is to anticipate the spawning of servers to account for the delay. We defer for now the discussion on implementation issues. The overall dynamics then become

$$\dot{m} = \beta[(1+\alpha)n - m], \tag{6a}$$
$$\dot{n} = \lambda - \mu \min\{m, n\}. \tag{6b}$$

The unique equilibrium x^* has coordinates $m^* = (1+\alpha)\rho$ and $n^* = \rho$, as shown in Fig. 3. The number of jobs still operates at ρ, which is a hard lower bound; however we are accepting an over-provisioning of $\alpha\rho$ servers in mean, with the aim of avoiding operation in the queuing region $n > m$.

Proposition 3.1: The equilibrium x^* of the dynamics (6) is globally asymptotically stable.

Proof: The dynamics (6) are piecewise-linear switching at the line $m = n$. The Jacobian matrix of the field has negative (real) eigenvalues in $\{m > n\}$. Hence, it is easy to see that solutions starting in this set stay in the same set forever, and approach x^* as $t \to \infty$.

Now consider the restriction of (6) to the set $\{m < n\}$. The linear extension of these dynamics to the entire quadrant would have $(\rho, \rho/(1+\alpha))$ as a global attractor. Since this point is outside the region, solutions starting in $\{m < n\}$ eventually leave this set into the already analyzed contiguous set, where they remain and approach x^* as $t \to \infty$. ■

As in Section II-B, to model stochastic fluctuations we resort to the continuous time Markov chain $X = (M, N)$ corresponding to the dynamics (6). The state-space is $\mathbb{N}^2$, as before, and the transitions for this case are shown in Fig. 4.

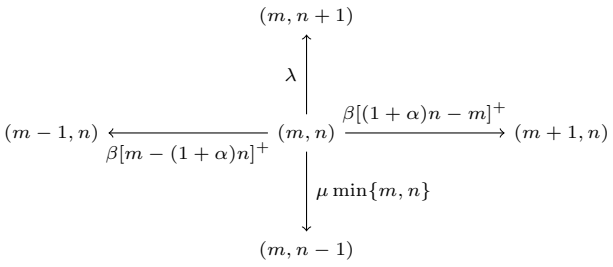

Fig. 4. Transition rates of the modified Markov chain.

A fluid limit, analog of Theorem 2.2, is also possible in this setting. Here again the processes $X_l = (M_l, N_l)$ are defined by scaling the arrival rate to $l\lambda$ and $\bar{X}_l = X_l/l$.

Theorem 3.2: Assume that the deterministic initial conditions of the processes $\bar{X}_l$ converge to some $x_0 \in [0, +\infty)^2$ as $l \to \infty$, and let $x(t)$ be the unique solution to (6) with initial condition x_0.

$$\sup_{t\in[0,T]} \left|\left|\bar{X}_l(t) - x(t)\right|\right| \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty \quad \forall\, T \geq 0.$$

The proof again follows from [1, Chapter11] since the fluid dynamics have a Lipschitz field.

Furthermore, the analog of Theorem 2.3 is also true; a sketch of the proof can be found in the Appendix. In order to state this result consider the processes $Z_l = \sqrt{l}(X_l - x^*)$.

Theorem 3.3: Assume that the deterministic initial conditions of the processes Z_l converge to some $Z_0 \in \mathbb{R}^2$ as $l \to \infty$. Then, the processes Z_l converge weakly, in the Skorokhod space $D_{\mathbb{R}^2}[0, \infty)$, to the process $Z = (\tilde{M}, \tilde{N})$, whose initial condition is Z_0, and solves the following SDE.

$$d\tilde{M} = \beta[(1+\alpha)\tilde{N} - \tilde{M}]dt, \tag{7a}$$
$$d\tilde{N} = -\mu\tilde{N}dt + \sqrt{2\lambda}dW. \tag{7b}$$

Here W is a standard one-dimensional Wiener process.

Note that equation (7b) is now linear, this change with respect to equation (4b) is due to the shift in the operating point x^*, allowing us to compute the stationary distribution of the limiting process Z. Indeed, if we let $\eta = \mu/\beta$ be the ratio between mean setup delays and mean service times, then the stationary distribution of Z is a bivariate Gaussian $\mathcal{N}(0, \Sigma)$, with mean zero and covariance matrix

$$\Sigma = \rho\frac{1+\alpha}{1+\eta}\begin{bmatrix} 1+\alpha & 1 \\ 1 & \frac{1+\eta}{1+\alpha} \end{bmatrix}. \tag{8}$$

The latter is obtained by solving the Lyapunov equation $A\Sigma + \Sigma A^T + BB^T = 0$, where $dZ = AZdt + BdW$ is the SDE (7) written in matrix form.

Theorems 3.2 and 3.3 tell us that $X_l \approx lx^* + \sqrt{l}Z$ is a reasonable steady-state approximation when l is large enough. Recall that M_l and N_l are the number of servers and jobs, respectively, in a system where the arrival rate is $l\lambda$. Also, the fluid equilibrium of this system is $lx^* = (l\rho, l\rho)$, and the steady-state covariance of $\sqrt{l}Z$ is the same as in equation (8) but replacing ρ by $l\rho$. Hence, another way to express this estimate, incorporating the scaling into λ, is to say that $X \approx x^* + Z$, when λ is large enough. Therefore, $M - N \approx \mathcal{N}(\alpha\rho, \sigma^2)$ where the variance is

$$\sigma^2 = \begin{bmatrix} 1 & -1 \end{bmatrix} \Sigma \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{\alpha^2 + \eta}{1+\eta}\rho. \tag{9}$$

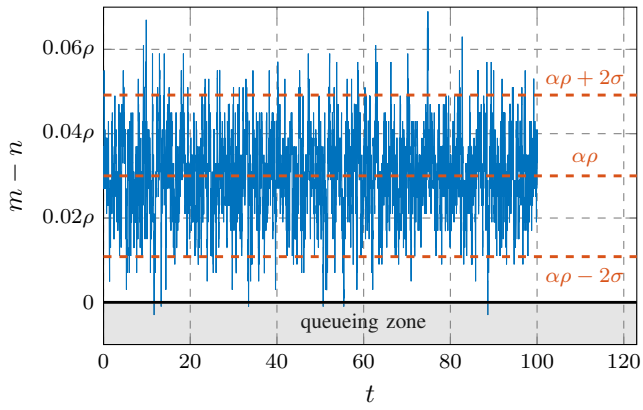

Fig. 5. Simulation for $\lambda = 1000$, $\mu = 1$, $\beta = 10$ and $\alpha = 3\%$.

The simulation of Fig. 5 validates this analysis; in particular it shows that in steady-state $M - N$ spends most of its time within the confidence interval $[\alpha\rho - 2\sigma, \alpha\rho + 2\sigma]$. The parameter α in the simulation has been chosen such that

$\alpha\rho - 2\sigma > 0$; in this way the queue remains empty most of the time as desired. From equation (9), this design condition[2] on α becomes:

$$\frac{1}{\rho(1+\eta)} + \frac{\eta}{\rho\alpha^2(1+\eta)} < \frac{1}{4}. \tag{10}$$

We may also use our Gaussian approximation to estimate the mean queue length. Denoting by φ and ϕ the density and, respectively, cumulative distribution of the standard Gaussian, we have:

$$\mathbb{E}[N-M]^+ \approx \sigma\varphi\left(\frac{\alpha\rho}{\sigma}\right) - \alpha\rho\Phi\left(-\frac{\alpha\rho}{\sigma}\right).$$

This function of α is plotted in Fig. 6, for different values of the load ρ. We see that the mean queue length approaches zero rapidly as α increases. Recall that $\alpha\rho = E[M-N]$, so α reflects the fraction of over-provisioning. We find, for instance, that for the traffic intensity $\rho = 1000$ Erlangs, a 2% over-provisioning yields a mean queue length of order two, and a 5% over-provisioning results in nearly zero queue at the dispatcher.

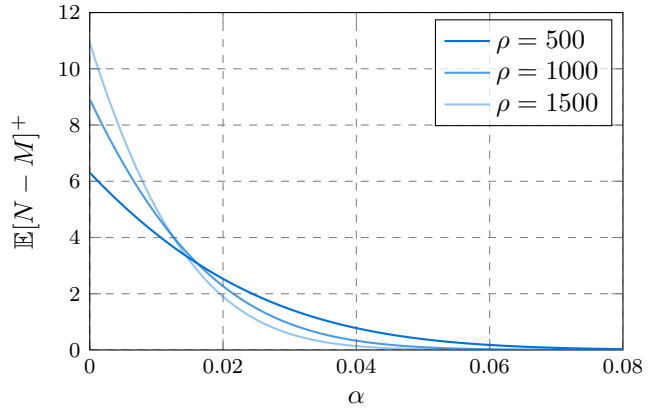

Fig. 6. Mean queue length for $\eta = 1$

For the plots in Fig. 6 we assumed $\eta = 1$, i.e. the mean setup time of servers is equal to the mean job service time. Clearly the performance will not be as good if setup delays are higher. This is captured by our model, since an increase in η causes an increase in the variance σ^2. Still, for reasonable values of η a moderate amount of over-provisioning yields almost zero queue at the dispatcher.

We also note that the queue decays more sharply with α for higher values of ρ. This suggests that perhaps, rather than selecting a fixed fraction of over-provisioning, we could choose one that adapts to the uncertain load ρ. We return to equation (10) with this in mind, and see that in order to satisfy our zero queue condition, one must set $\alpha = O(1/\sqrt{\rho})$. This means that $O(\sqrt{\rho})$ is the minimum over-provisioning needed to avoid queueing almost completely. In the next section we propose a method that self-adjusts the number of idle servers to this level.

[2]The constant on the right can be adjusted for other confidence levels.

IV. Automatic control of the over-provisioning

We now focus on an automatic rule, independent of the load, with the aim of achieving the desired over-provisioning level of $O(\sqrt{\rho})$ servers. The main idea is to replace the terms αn in the transitions of Fig. 4 with $\delta\sqrt{n}$; the most suitable value of the constant δ will depend on the ratio η, as discussed in the previous section. Note that we are just approximating $\sqrt{\rho}$ by its instantaneous estimate $\sqrt{n}$, the current occupation level. This modification is inspired in the Halfin-Whitt regime of [7], but adapted to the automatic feedback setting of this paper.

A. Fluid approximation

Consider the Markov chain depicted in Fig. 7, where we have replaced the terms αn in Fig. 4 by $\delta\sqrt{n}$. Also, let $X = (M, N)$ denote this process (overloading the notation).

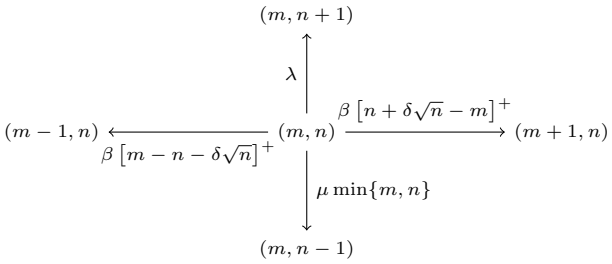

Fig. 7. Transition rates for the $O(\sqrt{n})$ control rule.

As before, consider the scaled processes $X_l = (M_l, N_l)$ where λ is replaced by $l\lambda$ in Fig. 7, and the renormalized processes $\bar{X}_l = X_l/l$. In the macroscopic level we recover the fluid behavior of Section II.

Theorem 4.1: Assume that the deterministic initial conditions of the processes $\bar{X}_l$ converge to some $x_0 \in [0, +\infty)^2$ as $l \to \infty$, and let $x(t)$ be the unique solution to:

$$\dot{m} = \beta(n - m), \tag{11a}$$

$$\dot{n} = \lambda - \mu\min\{m, n\}, \tag{11b}$$

with initial condition x_0. The processes $\bar{X}_l$ satisfy:

$$\sup_{t\in[0,T]} \left|\left|\bar{X}_l(t) - x(t)\right|\right| \xrightarrow{a.s.} 0 \quad \text{as } l \to \infty \quad \forall\, T \geq 0.$$

The proof is sketched in the Appendix, and relies on a suitable modification to the strong law-of-large-numbers in [1, Chapter 11]. The intuition behind this result is that the set point $\delta\sqrt{n}$ 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). As a result, in the fluid scale we see the same system as in Section II-B, with zero over-provisioning. To see how this system counteracts the queueing delay we need to look into the diffusion scale.

B. Diffusion approximation

We now analyze fluctuations around the fluid equilibrium by considering again the processes $Z_l = \sqrt{l}(\bar{X}_l - x^*)$. In this scale, the system manages to move away from the region

$\{n > m\}$, thus minimizing delay and achieving an over-provisioning of the order $\sqrt{\rho}$. More formally, we have the following result; whose proof is sketched in the Appendix.

Theorem 4.2: Assume that the deterministic initial conditions of the processes Z_l converge to some $Z_0 \in \mathbb{R}^2$ as $l \to \infty$. Then, the processes Z_l converge weakly, in the Skorokhod space $D_{\mathbb{R}^2}[0,\infty)$, to the process $Z = (\tilde{M}, \tilde{N})$, whose initial condition is Z_0, and solves the following SDE.

$$d\tilde{M} = \beta[\tilde{N} - \tilde{M} + \delta\sqrt{\rho}]dt, \tag{12a}$$

$$d\tilde{N} = -\mu \min\{\tilde{M}, \tilde{N}\}dt + \sqrt{2\lambda}dW. \tag{12b}$$

Here W is a standard one-dimensional Wiener process.

If we compare equations (11) and (12), there is an extra drift term in (12a) that accounts for the $O(\sqrt{\rho})$ over-provisioning. Indeed, (12a) implies that $\mathbb{E}[\tilde{M} - \tilde{N}] = \delta\sqrt{\rho}$ for the steady-state distribution. Thus, the system's over-provisioning, which disappeared in the fluid scale, becomes apparent in the diffusion scale.

Unfortunately, the switching in equation (12b) precludes us from computing the stationary distribution of the above process. Nevertheless, the analysis at the end of Section III tells us that the system operates with nearly zero queue at the dispatcher. The main difference between this system and the one in Section III is that we have replaced the constant α in Fig. 4 by a function $\alpha(n) = \delta/\sqrt{n}$ that tracks $\delta/\sqrt{\rho}$. Thus, the system's mean queue length should be approximately as in Fig. 6 for $\alpha = \delta/\sqrt{\rho}$.

The advantage of the present provisioning rule is that we do not need to know the traffic intensity that the system will face. Furthermore, when the load changes, the control that we have designed adapts to the new load automatically.

V. Implementation

We begin with an implementation of the provisioning rule that we described in Section III. Note that the boundary case $\alpha = 0$ corresponds to the control of (1)-(2), whose implementation was discussed in Section II-A. Unfortunately, an exact implementation is not possible for a generic $\alpha > 0$. This would require to keep a number of $(1+\alpha)n - m$ servers in the setup stage, waiting to join the working servers, that in general will not be an integer.

Nevertheless, an approximate implementation is possible. Indeed, define $q(m,n) = (1+\alpha)n - m$ and consider the following provisioning rule.

- If a job arrives while $q(m,n) \geq 0$, then one server is summoned with probability $1-\alpha$ and two servers are summoned with probability α.
- If a job departs while $q(m,n) \geq 0$ and there are servers undergoing the setup stage, then one of them is dismissed with probability $1-\alpha$ and two of them are dismissed with probability α, if possible.
- While $q(m,n) < 0$ the dispatcher maintains a list of idle servers with $m - (1+\alpha)n$ (rounded to the closest integer) entries. These servers are dismissed, shutting down at rate β.

With this policy we aim to keep $[(1+\alpha)n - m]^+$ servers in the setup stage, on average, and a set of $[m - (1+\alpha)n]^+$ idle servers that shut down at rate β. Thus, we expect to see the same performance as in Section III.

For the policy in Section IV we exploit the same idea. The algorithm is as above but replacing the constant α with the function $\alpha(n) = \delta/\sqrt{n}$, as in Section IV. The experiment in Fig. 8 compares the ideal system described in Section IV with the proposed implementation, note the similarity.

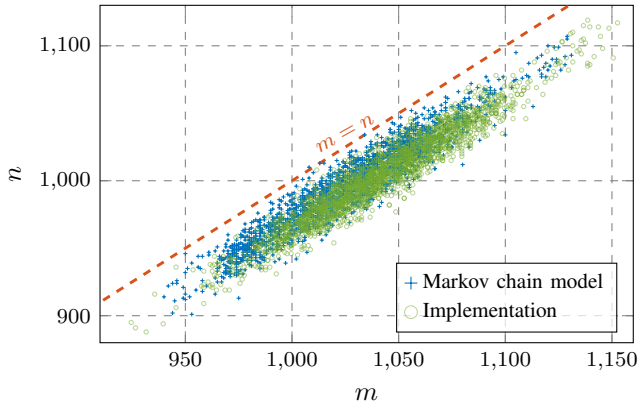

Fig. 8. Approximate implementation of the $O(\sqrt{n})$ control rule; simulation for $\lambda = 1000$, $\mu = 1$, $\beta = 10$ and $\delta = 1$.

At the end of Section IV we pointed out that the (ideal) provisioning rule, that we proposed in the same section, should perform similarly to the rule that we analyzed in Section III for $\alpha = \delta/\sqrt{\rho}$. This claim is supported by the following table, which compares time averages of the implementation of the policy of Section IV with the estimates of Section III. Thus, not only the the system proposed in Section IV behaves as described in Section III, but also its implementation.

	Time averages	**Estimates**
$\mathbb{E}[M]$	1041	1032
$\mathbb{E}[N]$	1004	1000
$\mathbb{E}[N-M]^+$	2.10^{-5}	10^{-3}

Note that the algorithms described above do not require the use of a significant amount of additional resources. Clearly, the dispatcher should keep track of the variables $\alpha(n)$ and $q(m,n)$, updating them whenever m or n change, which is inexpensive in terms of resources. Also, the list of idle servers, that the dispatcher already needs to maintain, should incorporate a tag indicating whether an idle server can be dismissed or is reserved as over-provisioning.

VI. Conclusions

In this paper, we analyzed the dynamic behavior of feedback policies to scale instance deployment in the cloud under startup lag. We focused on deriving simple control rules that explore the tradeoffs between queueing delay and overprovisioning. Based on fluid models and difussion 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 the square-root staffing rule of [7] reduces queueing delay to zero while keeping the overprovisioning scaling sublinearly with the load.

In future work, we plan to analyze the performance of this feedback control rule when combined with distributed load balancing algorithms.

APPENDIX A - STABILITY OF SWITCHED DYNAMICS

Proof: [of Proposition 2.1] The dynamics (2)-(1) are piecewise-linear, switching at the line $m = n$, hence we have different Jacobian matrices in the sets $\{m < n\}$ and $\{m > n\}$, respectively:

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

Note that the state variable is $[m\ n]^T$.

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 the matrix $A_i^T P + P A_i$. We must find $q, r \in \mathbb{R}$ such that P is positive definite and the next equations hold:

$$\begin{aligned} &T_1(q,r) = 2(\beta - \mu)q - 2\beta < 0, \\ &D_1(q,r) = -4\beta(\beta + \mu q)q - (\beta - \mu r - \beta q)^2 > 0, \\ &T_2(q,r) = 2\beta(q-1) - 2\mu r < 0, \\ &D_2(q,r) = -4\beta(\beta q - \mu r) - [\beta - (\beta + \mu)q]^2 > 0. \end{aligned}$$

The set $\{D_1(q,r) > 0\}$ is the interior of an ellipse, located inside the set $\{q \leq 0\}$ and tangent to the line $q = 0$ at the point $(0, \beta/\mu)$. Also, $\{D_2(q,r) > 0\}$ is the open set above the graph of a parabola, that contains the point $(0, \beta/4\mu)$, and has positive concavity. As a result, the two sets intersect, moreover there exists $\delta > 0$ such that $(-\varepsilon, \beta/\mu)$ lies in the intersection for all $\varepsilon \in (0, \delta)$. Since $T_1(\varepsilon, \beta/\mu) \to -2\beta$ and $T_2(\varepsilon, \beta) \to -4\beta$ as $\varepsilon \to 0$, there exists some $\varepsilon > 0$ such the desired matrix is

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

which is clearly positive definite for small values of ε. ■

APPENDIX B - STOCHASTIC LIMITS

We describe here how the methodology in [1, Chapter 11] may be extended to prove the results of Sections II-B and IV.

To begin, we consider the Markov chain $X = (M, N)$ of Section IV, whose transition rates appear in Fig. 7. Let $X_l = (M_l, N_l)$ be the scaled process, obtained by replacing λ by $l\lambda$ in Fig. 7, and denote by $q_v^{(l)}(x)$ the transition rate of this process in the direction $v \in \mathbb{Z}^2$ away from state x.

A. Fluid limit

Following [1, Chapter 11] the Markov chain dynamics can be equivalently written as:

$$X_l(t) = X_l(0) + \sum_v v \mathcal{N}_v \left(\int_0^t q_v^{(l)}(X_l(\tau)) d\tau \right),$$

where $\mathcal{N}_v$ are independent Poisson processes with intensity one, defined over some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. It is convenient to introduce the centered process $Y_v(t) = \mathcal{N}_v(t) - t$. Defining the rescaled process $\bar{X}_l = X_l/l$ as before, we have:

$$\begin{aligned} \bar{X}_l(t) = \bar{X}_l(0) &+ \sum_v \frac{v}{l} Y_v \left(\int_0^t q_v^{(l)}(l\bar{X}_l(\tau)) d\tau \right) \\ &+ \sum_v \frac{v}{l} \int_0^t q_v^{(l)}(l\bar{X}_l(\tau)) d\tau. \end{aligned} \tag{13}$$

The classical density-dependent result from [1] is based on the homogeneity condition $q_v^{(l)}(lx) = l q_v(x)$, for some suitable maps q_v; the main difference is that here homogeneity holds only in the limit. To address this we decompose the transition rates into a homogeneous part and a perturbation:

$$q_v^{(l)}(lx) = l\gamma_v(x) + l\delta_v^{(l)}(x).$$

Then equation (13) can be rewritten as:

$$\begin{aligned} \bar{X}_l(t) = \bar{X}_l(0) &+ \sum_v \frac{v}{l} Y_v \left(\int_0^t q_v^{(l)}(l\bar{X}_l(\tau)) d\tau \right) \\ &+ \sum_v v \int_0^t \gamma_v(\bar{X}_l(\tau)) d\tau + \sum_v v \int_0^t \delta_v^{(l)}(\bar{X}_l(\tau)) d\tau. \end{aligned}$$

Alternatively, we write

$$\begin{aligned} \bar{X}_l(t) = \bar{X}_l(0) &+ \sum_v \frac{v}{l} Y_v \left(\int_0^t q_v^{(l)}(l\bar{X}_l(\tau)) d\tau \right) \\ &+ \int_0^t F(\bar{X}_l(\tau)) d\tau + \int_0^t G_l(\bar{X}_l(\tau)) d\tau, \end{aligned} \tag{14}$$

where we have introduced the vector fields:

$$F(x) = \sum_v v\gamma_v(x), \quad G_l(x) = \sum_v v\delta_v^{(l)}(x).$$

To prove asymptotic results based on this representation, the key is to control the perturbation terms contained in the vector field $G_l(x)$. For this purpose we make the following set of assumptions.

Assumption 6.1:

- The maps γ_v are bounded in compact sets, and locally Lipschitz; particularly F is locally Lipschitz.
- The maps δ_v^l are bounded in compact sets and

$$\lim_{l \to \infty} \sup_{x \in K} ||G_l(x)|| \to 0 \quad K \text{ compact.}$$

Note that these assumptions are valid for the Markov chain considered in Section IV, where in particular the non-homogenous term affects only the horizontal transitions:

$$G_l(m,n) = \beta \frac{\delta\sqrt{n}}{\sqrt{l}} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Fix some initial condition $x_0 \in [0,\infty)^2$ and suppose that the unique solution x to the initial value problem $\dot{x} = F(x)$ is defined on the interval $[0,T]$.

Theorem 6.2: Assume that $\bar{X}_l(0) \to x_0$ as $l \to \infty$, then:

$$\sup_{t\in[0,T]} \left|\left|\bar{X}_l(t) - x(t)\right|\right| \xrightarrow{a.s.} 0 \quad \text{as } l \to \infty.$$

The proof is essentially as in [1, Chapter 11]; the last term in equation (14) vanishes after taking the limit as a result of Assumption 6.1. Also, the strong law-of-large-numbers for the Poisson process is key to control the summation in (14).

B. Diffusion approximation

Before proceeding with the diffusion approximation, we state the following Lemma which serves as a refinement of the weak law-of-large-numbers of the Poisson process. Its proof follows after applying Doob's maximal inequality.

Lemma 6.3: Consider a centered Poisson process Y with unitary intensity.

$$\sup_{t\in[0,T]} \left|\frac{Y(lt)}{l^{1-\alpha}}\right| \xrightarrow{\mathbf{P}} 0 \quad \forall\, T \geq 0, \quad \forall\, \alpha \in [0,1/2).$$

The previous lemma allows to prove the following extension of Theorem 6.2, which will be useful later.

Theorem 6.4: Fix some $\alpha \in [0,1/2)$. Furthermore, assume that $l^\alpha \left|\left|\bar{X}_l(0) - x_0\right|\right| \to 0$ as $l \to \infty$, then:

$$\sup_{t\in[0,T]} l^\alpha \left|\left|\bar{X}_l(t) - x(t)\right|\right| \xrightarrow{\mathbf{P}} 0 \quad \text{as } l \to \infty.$$

Let x^* be an equilibrium point of the dynamics $\dot{x} = F(x)$, and define the processes $Z_l = \sqrt{l}(\bar{X}_l - x^*)$ and

$$U_l(t) = \sum_v \frac{v}{\sqrt{l}} Y_v \left(\int_0^t q_v^{(l)}(l\bar{X}_l(\tau))d\tau \right).$$

The following relation follows from equation (14).

$$\begin{aligned} Z_l(t) = Z_l(0) + U_l(t) + \int_0^t \sqrt{l} F(\bar{X}_l(\tau))d\tau \\ + \int_0^t \sqrt{l} G_l(\bar{X}_l(\tau))d\tau. \end{aligned} \tag{15}$$

Assumption 6.5: Suppose that there exists a Lipschitz field $\partial F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ with the two following properties.

- $\partial F(\lambda x) = \lambda \partial F(x)$ for all $\lambda \geq 0$ and $x \in \mathbb{R}^2$.
- The remainder $R(x) = F(x) - \partial F(x - x^*)$ is such that the following limit holds:

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

Moreover, assume that there exists a field $G : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, locally Lipschitz at x^*, such that:

$$\lim_{l\to\infty} \sup_{x\in K} \left|\left|\sqrt{l} G_l(x) - G(x)\right|\right| \to 0 \quad K \text{ compact}.$$

The first part of the previous assumption holds for the density dependent families in sections II-B and IV. They have the same drift, with $\gamma = \beta$ in Section IV, which is piecewise-linear, switching at the line perpendicular to the vector $v = [-1\ 1]^T$. Thus, we may take:

$$\partial F(u) = \begin{bmatrix} -\beta & \beta \\ -\mu & 0 \end{bmatrix} u\mathbf{1}_{\langle u,v\rangle>0} + \begin{bmatrix} -\gamma & \gamma \\ 0 & -\mu \end{bmatrix} u\mathbf{1}_{\langle u,v\rangle<0},$$

which satisfies Assumption 6.5, as a matter of fact $R(u) \equiv 0$.

The second part of this assumption, which applies only to Section IV, also holds taking:

$$G(m,n) = \beta\delta\sqrt{n} \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Returning to the general setting, note that we may write

$$Z_l(t) = Z_l(0) + U_l(t) + \delta_l(t) + \int_0^t \partial F(Z_l(\tau))d\tau,$$

where δ_l is the following process

$$\delta_l(t) = \int_0^t \sqrt{l} \left[G_l(\bar{X}_l(\tau)) + R(\bar{X}_l(\tau))\right] d\tau.$$

Now consider an independent family $\{W_v\}_v$ of Wiener processes and define the processes:

$$U(t) = \sum_v W_v(\gamma_v(x^*)t), \quad \tilde{U}(t) = U(t) + tG(x^*).$$

Also, let Z be the solution to the following SDE, which we state in integral form:

$$Z(t) = Z_0 + U(t) + \int_0^t \partial F(Z(\tau)) + G(x^*)d\tau.$$

Existence and uniqueness follows from standard results.

Theorem 6.6: Assume that $Z_l(0) \to Z_0$ as $l \to \infty$. Then, $Z_l \Rightarrow Z$ in the Skorokhod space $D_{\mathbb{R}^2}[0,\infty)$ as $l \to \infty$.

It is enough to show that $Z_l \Rightarrow Z$ in $D_{\mathbb{R}^2}[0,T]$ for all $T \geq 0$. The strategy is, first, to prove that there exists a continuous function $\phi : D_{\mathbb{R}^2}[0,T] \longrightarrow D_{\mathbb{R}^2}[0,T]$ such that:

$$Z_l = \phi(Z_l(0) + U_l + \delta_l), \quad Z = \phi(Z_0 + \tilde{U}).$$

Afterwards one shows that $U_l + \delta_l \Rightarrow \tilde{U}$ in $D_{\mathbb{R}^2}[0,T]$, and the claim follows from the continuous mapping theorem.

In order to show that such map ϕ exists, we fix some $f \in D_{\mathbb{R}^2}[0,T]$ and use the fixed point theorem, as in the classical proof of Picard's theorem, to prove local existence and uniqueness of solutions to:

$$\varphi(t) = x_0 + f(t) - f(t_0) + \int_{t_0}^t \partial F(\varphi(\tau))d\tau.$$

Since ∂F is uniformly Lipschitz, the size of the neighborhood where local solutions exist, and are unique, is independent of the initial condition (t_0, x_0). This allows to prove that solutions defined in $[0,T]$ exist and are unique. Hence, there exists a unique $\phi(f) \in D_{\mathbb{R}^2}[0,T]$ such that:

$$\phi(f)(t) = f(t) + \int_0^t \partial F(\phi(f)(\tau))d\tau.$$

Moreover, it is possible to prove that ϕ is continuous in the Skorokhod topology.

In order to show that $U_l + \delta_l \Rightarrow \tilde{U}$ in $D_{\mathbb{R}^2}[0,T]$, it is enough to prove that $U_l \Rightarrow U$ in $D_{\mathbb{R}^2}[0,T]$ and

$$\sup_{t\in[0,T]} ||\delta_l(t) - tG(x^*)|| \xrightarrow{\mathbf{P}} 0 \quad \text{as } l \to \infty. \tag{16}$$

By the central limit theorem for the Poisson process:

$$\sum_v \frac{v}{\sqrt{l}} Y_l\left(l\gamma_v(x^*)t\right) \Rightarrow U(t) \quad \text{in } D_{\mathbb{R}^2}[0,T] \quad \text{as } l \to \infty.$$

Also, picking some $\alpha \in (0, 1/2)$ and using Theorem 6.4, and Assumption 6.1, we see that we may write:

$$\left| \int_0^t q_v^{(l)}(l\bar{X}_l(\tau))d\tau - l\gamma_v(x^*)t \right| = O(l^{1-\alpha}) \quad \text{in } \Omega_l^c,$$

for some sets such that $\mathbf{P}(\Omega_l) \to 0$ as $l \to \infty$. Elaborating on this it is possible to show that:

$$\sup_{t\in[0,T]} \left|\left| U_l(t) - \sum_{v\in D} \frac{v}{\sqrt{l}} Y_l\left(l\gamma_v(x^*)t\right) \right|\right| \xrightarrow{\mathbf{P}} 0 \quad \text{as } l \to \infty.$$

This implies that $U_l \Rightarrow U$ in $D_{\mathbb{R}^2}[0,T]$ as $l \to \infty$.

Finally, we may prove (16) using Assumption 6.5.

REFERENCES

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

[2] S. Foss and A. L. Stolyar, "Large-scale join-idle-queue system with general service times," *Journal of Applied Probability*, vol. 54, no. 4, pp. 995–1007, 2017.

[3] D. Gamarnik, J. N. Tsitsiklis, and M. Zubeldia, "Delay, memory, and messaging tradeoffs in distributed service systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1, pp. 1–12, 2016.

[4] A. J. Ganesh, N. O'Connell, and D. J. Wischik, *Big queues*. Springer, 2004.

[5] D. Goldsztajn, A. Ferragut, F. Paganini, and M. Jonckheere, "Controlling the number of active instances in a cloud environment," *ACM SIGMETRICS Performance Evaluation Review*, vol. 45, no. 2, pp. 15–20, 2018.

[6] V. Gupta and N. Walton, "Load balancing in the non-degenerate slowdown regime," *arXiv preprint arXiv:1707.01969*, 2017.

[7] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Operations research*, vol. 29, no. 3, pp. 567–588, 1981.

[8] T. G. Kurtz *et al.*, "Strong approximation theorems for density dependent markov chains," *Stochastic Processes and their Applications*, vol. 6, no. 3, pp. 223–240, 1978.

[9] M. Lin, A. Wierman, L. L. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," *IEEE/ACM Transactions on Networking (TON)*, vol. 21, no. 5, pp. 1378–1391, 2013.

[10] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," *Performance Evaluation*, vol. 68, no. 11, pp. 1056–1071, 2011.

[11] D. Mukherjee, S. Borst, J. Van Leeuwaarden, and P. Whiting, "Universality of power-of-d load balancing schemes," *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 2, pp. 36–38, 2016.

[12] D. Mukherjee, S. Dhara, S. C. Borst, and J. S. van Leeuwaarden, "Optimal service elasticity in large-scale distributed systems," *ACM SIGMETRICS Performance Evaluation Review*, vol. 1, no. 1, p. 25, 2017.

[13] L. M. Nguyen and A. L. Stolyar, "A service system with randomly behaving on-demand agents," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 44, no. 1. ACM, 2016, pp. 365–366.

[14] M. van der Boor, S. C. Borst, J. S. van Leeuwaarden, and D. Mukherjee, "Scalable load balancing in networked systems: A survey of recent advances," *arXiv preprint arXiv:1806.05444*, 2018.