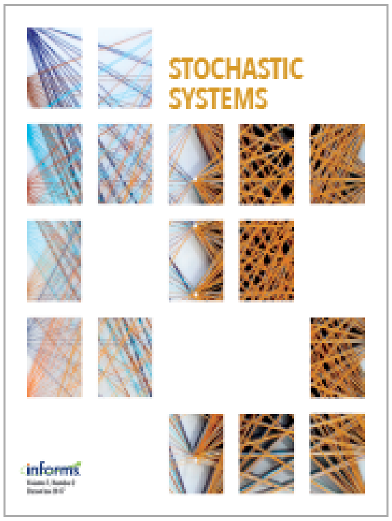

Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Service System with Packing Constraints: Greedy Randomized Algorithm Achieving Sublinear in Scale Optimality Gap

Alexander L. Stolyar, Yuan Zhong

To cite this article:
Alexander L. Stolyar, Yuan Zhong (2021) A Service System with Packing Constraints: Greedy Randomized Algorithm Achieving Sublinear in Scale Optimality Gap. Stochastic Systems 11(2):83-111. https://doi.org/10.1287/stsy.2019.0067

http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
Vol. 11, No. 2, June 2021, pp. 83–111
ISSN 1946-5238 (online)

A Service System with Packing Constraints: Greedy Randomized Algorithm Achieving Sublinear in Scale Optimality Gap

Alexander L. Stolyar,[a] Yuan Zhong[b]

[a] Department of Industrial and Enterprise Systems Engineering and Coordinated Sciences Laboratory, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801; [b] Booth School of Business, University of Chicago, Chicago, Illinois 60637

Contact: stolyar@illinois.edu, https://orcid.org/0000-0002-1496-9803 (ALS); yuan.zhong@chicagobooth.edu (YZ)

Received: November 10, 2015
Revised: March 24, 2018; December 11, 2018
Accepted: July 30, 2019
Published Online in Articles in Advance: February 2, 2021

https://doi.org/10.1287/stsy.2019.0067

Abstract. A service system with multiple types of arriving customers is considered. There is an infinite number of homogeneous servers. Multiple customers can be placed for simultaneous service into one server, subject to general *packing constraints*. The service times of different customers are independent even if they are served simultaneously by the same server; the service time distribution depends on the customer type. Each new arriving customer is placed for service immediately into either an occupied server, that is, one already serving other customers, as long as packing constraints are not violated or into an empty server. After service completion, each customer leaves its server and the system. The basic objective is to minimize the number of occupied servers in steady state. We study a *greedy random* (GRAND) placement (packing) algorithm, introduced in our previous work. This is a simple online algorithm that places each arriving customer uniformly at random into either one of the already occupied servers that can still fit the customer or one of the so-called *zero servers*, which are empty servers designated to be available to new arrivals. In our previous work, a version of the algorithm, labeled GRAND(aZ), is considered, in which the number of zero servers is aZ with Z being the current total number of customers in the system and positive a being an algorithm parameter. GRAND(aZ) is shown in our previous work to be asymptotically optimal in the following sense: (a) the steady-state optimality gap grows linearly in the system scale r (the mean total number of customers in service), that is, as $c(a)r$ for some positive $c(a)$, and (b) $c(a)$ vanishes as a goes to zero. In this paper, we consider the GRAND(Z^p) algorithm, in which the number of zero servers is Z^p, where $p < 1$ is a fixed parameter, sufficiently close to 1. We prove the asymptotic optimality of GRAND(Z^p) in the sense that the steady-state optimality gap is sublinear in the system scale r. This is a stronger form of asymptotic optimality than that of GRAND(aZ).

Keywords: queueing networks • stochastic bin packing • packing constraints • greedy random (GRAND) algorithm • cloud computing • sublinear optimality gap

1. Introduction

Efficient resource allocation in modern cloud computing systems poses many interesting and challenging new problems; see, for example, Gulati et al. (2012) for an overview. One such problem is that of efficient real-time assignment (or packing) of virtual machines to physical machines in a cloud data center, in which a primary objective is to minimize the number of physical machines being used. This leads to stochastic dynamic bin-packing models, in which, in contrast to many classical bin-packing models, "items" (virtual machines) being placed into "bins" (physical machines) do not stay in the system forever but leave after a random "service time." For this reason, such models are naturally viewed as service (or queueing) systems with items and bins being viewed as customers and servers, respectively.

In this paper, we consider a general service system model that has also been studied in Stolyar (2013) and Stolyar and Zhong (2013, 2015). There is a finite number of customer types, and the number of available servers is infinite. Customers arrive to the system over time, and multiple customers can be placed for simultaneous service (or *fit*) into the same server subject to *packing constraints*. We consider *monotone* packing, a general class of packing constraints, in which we only impose the following natural and nonrestrictive *monotonicity* condition: if a certain set of customers can fit into a server, then a subset can fit as well. The servers are *homogeneous* in that they all have the same packing constraints. The service times of different customers are independent even if they are served simultaneously by the same server, and the service time distribution depends only on the customer type. Each new arriving customer is placed for service immediately into either an occupied server, that is, one that is already serving other customers as long as packing constraints are not violated, or into an empty server. After service completion, each customer leaves its server and the system—as mentioned, this is what distinguishes our model from many classical bin-packing models (see, e.g., Csirik et al. 2006, Bansal et al. 2009). As in Stolyar and Zhong (2015), we make Markov assumptions, in which customers of each type arrive as an independent Poisson process, and service time distributions are all exponential.

We are interested in designing customer placement (packing) algorithms that minimize the total number of occupied servers in the steady state. In addition, given the scale at which modern cloud data centers operate, it is highly desirable that a placement algorithm is *online*, that is, it makes decisions based on the current system state only, and *parsimonious*, that is, it requires only minimal knowledge of system structure and state information.

We study a *greedy random* (GRAND) placement algorithm, introduced in Stolyar and Zhong (2015). This is a very parsimonious online algorithm, which places each arriving customer uniformly at random into either one of the already occupied servers (subject to packing constraints) or one of the so-called *zero servers*, which are empty servers designated to be available to new arrivals. In Stolyar and Zhong (2015), a version of the algorithm, which we call GRAND(aZ), is considered, in which the number of zero servers depends on the current system state as aZ with Z being the current total number of customers in the system and $a > 0$ being an algorithm parameter. GRAND(aZ) is shown in Stolyar and Zhong (2015) to be asymptotically optimal in the sense of the following two properties: (a) for each $a > 0$, the steady-state optimality gap is of the form $c^r(a)r$, where r is the system *scale*, defined to be the expectation of Z in the steady state with $c^r(a) \to c(a) > 0$ as $r \to \infty$, and (b) $c(a) \to 0$ as $a \to 0$. In other words, under GRAND(aZ) the optimality gap grows linearly as the system scale r with the linear factor going to zero as the algorithm parameter $a \to 0$.

The focus of this paper is the GRAND(Z^p) algorithm, in which the number of zero servers is Z^p, where $p \in (1 - 1/(8\kappa), 1)$ is an algorithm parameter and $(\kappa - 1)$ is the maximum possible number of customers that a server can fit. Our *main result* is the asymptotic optimality of GRAND(Z^p) in the sense that the steady-state optimality gap is $o(r)$; that is, it is sublinear in the system scale. This is a stronger form of asymptotic optimality than that of GRAND(aZ) because GRAND(Z^p) achieves the sublinear gap simply as $r \to \infty$ without having to take an additional limit on any algorithm parameter. This is in contrast to the case of GRAND(aZ), which achieves asymptotic optimality only by first taking the limit $r \to \infty$ and then the limit $a \to 0$ on the algorithm parameter a.

Let us provide some remarks on the reasons why this stronger form of the asymptotic optimality of GRAND(Z^p) is, on the one hand, natural to expect and, on the other hand, substantially more difficult to rigorously prove. Informally speaking, GRAND(Z^p) can be viewed as GRAND(aZ), where the parameter a, instead of being fixed, is replaced by a variable that decreases to zero with increasing system scale; namely $a = Z^p/Z = Z^{p-1}$. However, the asymptotic optimality of GRAND(aZ) does *not*, of course, imply the (stronger form of) asymptotic optimality of GRAND(Z^p) because the system scale $r \approx Z$ and, therefore, the "parameter" $a = Z^{p-1} \approx r^{p-1}$ depends on and changes with the scale r. The key technical difficulty is that the analysis of GRAND(Z^p) *cannot* be reduced to the analysis of system fluid limits and/or local fluid limits. See Section 5.1 for a more detailed discussion.

The implementation of GRAND(Z^p) is exactly the same as GRAND(aZ) with the only difference being that GRAND(Z^p) designates $\lceil Z^p \rceil$ zero servers instead of $\lceil aZ \rceil$ as in the case of GRAND(aZ). Thus, as with GRAND(aZ), it is extremely simple and easy to implement. It does not need to keep track of the exact states (i.e., "packing configurations") of the servers, and the only information required at any time is which customer types a given server can still accept for service. (For example, for each customer type, a list of servers that can accept an additional customer of this type can be maintained.) As a result, the algorithm only needs to maintain a very small amount of information. Furthermore, the algorithm does not use knowledge of the customer arrival rates or expected service times. We refer the readers to Stolyar and Zhong (2015) for a more detailed discussion of these attractive implementational features.

1.1. Brief Literature Review

Our work is related to several lines of previous research. First, our work is related to the extensive literature on classical bin-packing problems, in which items of various sizes arrive to the system and need to be placed in finite-size bins, according to an online algorithm. Once placed, items never leave or move between bins. The *worst-case analysis* of such problems considers all possible instances of item sizes and sequencings and aims to develop simple algorithms with a performance guarantee over all problem instances. See, for example, Coffman et al. (2013) for a recent, extensive survey. The *stochastic analysis* assumes that item sizes are given according to a probability distribution, and the typical objective is to minimize the expected number of occupied bins. For an overview of results, see, for example, Csirik et al. (2006) and references therein. A recent paper (Gupta and Radovanovic 2012) establishes improved results for the classical stochastic setting and contains some heuristics and simulations for the case with item departures, which is a special case of our model.

Much research on classical bin-packing concerns the one-dimensional case, in which both item and bin sizes are scalars. It is possible to generalize one-dimensional bin packing to higher dimensions in multiple ways. For example, in vector packing problems (Chekuri and Khanna 2004, Stolyar 2013), item and bin sizes are vectors, and in box-packing problems (Bansal et al. 2009), items and bins are rectangles or hyper-rectangles. Let us also remark that the packing constraints in our model include vector and box packing as special cases. See, for example, Bansal et al. (2009) for an overview of multidimensional packing.

Motivated partly by applications to computer storage allocation, a *one-dimensional, dynamic bin-packing* problem is introduced in Coffman et al. (1983), in which, similar to the model of this paper, items leave the system after a finite service time. Coffman et al. (1983) contains a worst-case analysis of the problem, so the techniques used are quite distinct from those in our paper. A recent review of results on worst-case dynamic bin packing can be found in Coffman et al. (2013).

Another related line of works considers bin packing *service* systems, which have one (see, e.g., Coffman and Stolyar 2001, Gamarnik 2004) or several servers (see, e.g., Jiang et al. 2012, Maguluri et al. 2012, Guo et al. 2013, Maguluri and Srikant 2013). In these systems, random-size items (or customers) arrive over time and get placed into servers for processing, subject to packing constraints. Customers waiting for service are queued, and a typical problem is to determine the maximum throughput and/or minimum queueing delay under a packing algorithm. Our model is similar to these systems because they model customer departures, but it is also different, mainly because our system has an infinite number of servers so that there are no queues or problem of stability. As in our work, recent papers on bin-packing service systems with multiple servers (Jiang et al. 2012, Maguluri et al. 2012, Guo et al. 2013, Maguluri and Srikant 2013) are also motivated by real-time virtual machine placement problems.

As mentioned in the introduction, the model in this paper is the same as that studied in Stolyar (2013) and Stolyar and Zhong (2013, 2015). Stolyar (2013) and Stolyar and Zhong (2013) introduce and study different classes of greedy algorithms and prove their asymptotic optimality as the system scale grows to infinity. A greedy algorithm does not use the knowledge of the customer arrival rates or mean service times and makes placement decisions based on the current system state only. However, unlike GRAND, it does need to keep track of the numbers of servers in different *packing configurations*. The number of possible configurations can be prohibitively large in many practical scenarios, a feature that may limit the implementability of greedy.

The relation of our main results to those for GRAND(aZ) in Stolyar and Zhong (2015) has already been discussed in much detail. Also related to the GRAND algorithms is a recent paper (Stolyar 2017) that generalizes GRAND(aZ) and its asymptotic optimality to *heterogeneous* systems with multiple server types, in which packing constraints depend on the server type; it also contains results for heterogeneous systems with *finite pools* of servers of each type. Finally, a randomized version of the best-fit algorithm is studied in Ghaderi et al. 2014, and it considers the model of this paper with specialized packing constraints and is proven to be asymptotically optimal, using techniques similar to those in Stolyar and Zhong (2015).

1.2. Organization

The rest of the paper is organized as follows. In Section 1.3, we introduce basic notation and conventions that are used throughout the paper. The model, the GRAND(Z^p) algorithm, and the asymptotic regime are formally defined in Section 2. We state Theorem 1, the main result on the asymptotic optimality of GRAND(Z^p) in Section 3. In Section 4, we collect some results and observations obtained in Stolyar and Zhong (2015), which are needed for our proofs. The proof of Theorem 1 is in Section 5 with the proofs of some auxilliary results given in the appendix. The paper is concluded in Section 6 with some discussion and suggestions for future work.

1.3. Basic Notation and Conventions

Sets of real and nonnegative real numbers are denoted by $\mathbb{R}$ and $\mathbb{R}_+$, respectively. Similarly, sets of integers and nonnegative integers are denoted by $\mathbb{Z}$ and $\mathbb{Z}_+$, respectively. For $\xi \in \mathbb{R}$, $\lceil \xi \rceil$ denotes the smallest integer greater than or equal to ξ, and $\lfloor \xi \rfloor$ denotes the largest integer smaller than or equal to ξ. For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we use $\boldsymbol{x} \cdot \boldsymbol{y}$ to denote their scalar (dot) product; that is, $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_i x_i y_i$. The standard Euclidean norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}$, and the distance from vector $\boldsymbol{x}$ to a set U in a Euclidean space is denoted by $d(\boldsymbol{x}, U) = \inf_{\boldsymbol{u} \in U} \|\boldsymbol{x} - \boldsymbol{u}\|$. Notation $\boldsymbol{x} \to \boldsymbol{u} \in \mathbb{R}^n$ means ordinary convergence in $\mathbb{R}^n$, and $\boldsymbol{x} \to U \subseteq \mathbb{R}^n$ means $d(\boldsymbol{x}, U) \to 0$. For $\boldsymbol{x} \in \mathbb{R}^n$, we also use $\|\boldsymbol{x}\|_1$ to denote the one-norm of $\boldsymbol{x}$, defined to be $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$. Vector $\boldsymbol{e}_i$ is the ith coordinate unit vector in $\mathbb{R}^n$. Symbol $\Rightarrow$ denotes convergence in distribution of random variables taking values in space $\mathbb{R}^n$ equipped with the Borel σ-algebra. Symbol $\overset{d}{=}$ means *equal in distribution*. We often write $x(\cdot)$ to mean the function (or random process) $\{x(t),\ t \ge 0\}$, and we write $\{x_{\boldsymbol{k}}\}$ to mean the vector $\{x_{\boldsymbol{k}},\ \boldsymbol{k} \in \mathcal{K}\}$, where the set of indices $\mathcal{K}$ is determined by the context. For a function (or random process) $x(\cdot)$, we use $x(t+)$ to denote its right limit at time t, that is, $x(t+) = \lim_{s \downarrow t} x(s)$, and use $x(t-)$ to denote its left limit at time t, that is, $x(t-) = \lim_{s \uparrow t} x(s)$, whenever these limits exist. For a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, its gradient at $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $\nabla f(\boldsymbol{x}) = \{\frac{\partial}{\partial x_i} f(\boldsymbol{x}),\ i = 1, \ldots, n\}$. Indicator function $I\{A\}$ for condition A is equal to one if A holds and zero otherwise. The cardinality of a finite set $\mathcal{N}$ is $|\mathcal{N}|$. Notation $\doteq$ means *is defined to be*. In this paper, we use bold letters to represent vectors and plain letters to represent scalars.

2. System Model

2.1. Infinite Server System with Packing Constraints

The model that we study in this paper is the same as in Stolyar (2013) and Stolyar and Zhong (2013, 2015). We consider a service system with an infinite number of servers. There are I types of arriving customers, indexed by $i \in \{1, 2, \ldots, I\} \doteq \mathcal{I}$. For each $i \in \mathcal{I}$, customers of type i arrive as an independent Poisson process of rate $\Lambda_i > 0$ and have service times that are exponentially distributed with mean $1/\mu_i$. The service times of all customers are mutually independent and independent of the arrival processes. Each customer is placed into one of the servers for processing immediately upon arrival and departs the system after the service completes. Multiple customers can occupy the same server simultaneously, subject to the so-called monotone packing constraint, which we now describe.

Definition 1 (Monotone Packing). A packing constraint is characterized by a finite set $\bar{\mathcal{K}}$, the set of feasible server configurations. A vector $\boldsymbol{k} = \{k_i,\ i \in \mathcal{I}\} \in \bar{\mathcal{K}}$ is a (feasible) server configuration if (a) for each $i \in \mathcal{I}$, $k_i \in \mathbb{Z}_+$, and (b) a server can simultaneously process k_1 customers of type 1, k_2 customers of type 2, $\cdots$, and k_I customers of type I. The packing constraint $\bar{\mathcal{K}}$ is called monotone if the following condition holds: whenever $\boldsymbol{k} \in \bar{\mathcal{K}}$ and $\boldsymbol{k}' \le \boldsymbol{k}$ componentwise, then $\boldsymbol{k}' \in \bar{\mathcal{K}}$ as well.

Without loss of generality, we assume that $\boldsymbol{e}_i \in \bar{\mathcal{K}}$ for all $i \in \mathcal{I}$, where $\boldsymbol{e}_i$ is the ith coordinate unit vector so that customers of all types can be processed. By definition, the componentwise zero vector $\boldsymbol{0}$ belongs to $\bar{\mathcal{K}}$; this is the configuration of an empty server. We denote by $\mathcal{K} = \bar{\mathcal{K}} \setminus \{\boldsymbol{0}\}$ the set of configurations that *do not* include the zero configuration. As discussed in Stolyar (2013) and Stolyar and Zhong (2013), monotone packing includes important special packing constraints, such as vector packing.

An important assumption of the model is that simultaneous service does *not* affect the service time distributions of individual customers. In other words, the service time of a customer is unaffected by whether there are other customers served simultaneously by the same server.

We now define the system state. For each $\boldsymbol{k} \in \mathcal{K}$, let $X_{\boldsymbol{k}}(t)$ denote the number of servers with configuration $\boldsymbol{k}$. Then, the vector $\boldsymbol{X}(t) = \{X_{\boldsymbol{k}}(t), \boldsymbol{k} \in \mathcal{K}\}$ is the *system state* at time t, and we often write $\boldsymbol{X} = \{X_{\boldsymbol{k}}, \boldsymbol{k} \in \mathcal{K}\}$ for a generic state. Note that the system state does not include the number of empty servers, which would always be infinite.

The following notation and terminology is used in the sequel. Define $\mathcal{M} = \{(\boldsymbol{k}, i) : \boldsymbol{k} \in \mathcal{K}, \boldsymbol{k} - \boldsymbol{e}_i \in \bar{\mathcal{K}}\}$. Note that $(\boldsymbol{k}, i) \in \mathcal{M}$ are in one-to-one correspondence with the pairs $(\boldsymbol{k}, \boldsymbol{k} - \boldsymbol{e}_i)$ with $\boldsymbol{k} \in \mathcal{K}$ and $\boldsymbol{k} - \boldsymbol{e}_i \in \bar{\mathcal{K}}$, so $(\boldsymbol{k}, i)$ can be viewed as a shorthand for $(\boldsymbol{k}, \boldsymbol{k} - \boldsymbol{e}_i)$. For this reason, we call a pair $(\boldsymbol{k}, i) \in \mathcal{M}$ an *edge*. The placement of a type-i arrival into a server with configuration $\boldsymbol{k} - \boldsymbol{e}_i$ to form configuration $\boldsymbol{k}$ is called *an arrival along the edge* $(\boldsymbol{k}, i)$, and the departure of a type-i customer from a server with configuration $\boldsymbol{k}$, which changes the server configuration to $\boldsymbol{k} - \boldsymbol{e}_i$, is called *a departure along the edge* $(\boldsymbol{k}, i)$.

2.2. GRAND and GRAND(Z^p) Algorithms

In general, a *placement* (or *packing*) *algorithm* determines the servers into which arriving customers are placed dynamically over time. In this paper, we are interested in *online* placement algorithms, which make placement decisions based only on the current system state $\boldsymbol{X}$. Thus, from now on, we use "placement algorithms" to mean online ones. Our primary objective is the design of placement algorithms that minimize the total number of occupied servers $\sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}}$ in the stationary regime.

Under any well-defined placement algorithm, the process $\{\boldsymbol{X}(t), t \geq 0\}$ is a continuous-time Markov chain with a countable state space, which is irreducible and positive recurrent. The positive recurrence of $\{\boldsymbol{X}(t), t \geq 0\}$ can be argued as follows. First, for each $i \in \mathcal{I}$ and each $t \geq 0$, we let

$$Y_i(t) = \sum_{\boldsymbol{k}\in\mathcal{K}} k_i X_{\boldsymbol{k}}(t) \tag{1}$$

be the number of type-i customers in the system at time t and let

$$Z(t) = \sum_i Y_i(t) \tag{2}$$

be the total number of customers in the system at time t. For each $i \in \mathcal{I}$, $\{Y_i(t) : t \geq 0\}$ describes exactly the dynamics of an independent $M/M/\infty$ queueing system with arrival rate Λ_i and service rate μ_i regardless of the placement algorithm and has a unique stationary distribution. Let us denote by $Y_i(\infty)$ the random value of $Y_i(t)$ in the stationary regime. Then, $Y_i(\infty)$ is a Poisson random variable with mean Λ_i/μ_i for each $i \in \mathcal{I}$. Similarly, the process $\{Z(t) : t \geq 0\}$ also has a unique stationary distribution, and if we let $Z(\infty)$ be the random value of $Z(t)$ in the stationary regime, then $Z(\infty)$ is a Poisson random variable with mean $\sum_{i\in\mathcal{I}} \Lambda_i/\mu_i$. Thus, the positive recurrence of the Markov chain $\{\boldsymbol{X}(t), t \geq 0\}$ follows from the facts that $\sum_{\boldsymbol{k}\in\mathcal{K}} X_{\boldsymbol{k}}(t) \leq \sum_i Y_i(t) = Z(t)$ for all $t \geq 0$ and that the process $\{Z(t) : t \geq 0\}$ is positive recurrent. Consequently, the process $\{\boldsymbol{X}(t),\ t \geq 0\}$ has a unique stationary distribution, and we let $\boldsymbol{X}(\infty) = \{X_{\boldsymbol{k}}(\infty), \boldsymbol{k} \in \mathcal{K}\}$ be the random system state $\boldsymbol{X}(t)$ in the stationary regime.

In Stolyar and Zhong (2015), we introduce a broad class of placement algorithms, called the GRAND algorithms. For completeness, we include the definition here.

Definition 2 (GRAND Algorithm). At any given time t, there is a designated finite set of $X_{\boldsymbol{0}}(t)$ empty servers, called zero servers, for which $X_{\boldsymbol{0}}(t) = f(\boldsymbol{X}(t))$ is a given fixed function of the system state $\boldsymbol{X}(t)$. Suppose that a type i customer arrives at time t. Then, the customer is placed into a server chosen uniformly at random among the zero servers and the occupied servers in which the customer can fit. In other words, the total number of servers available to a type i arrival at time t is

$$X_{(i)}(t) \doteq X_{\boldsymbol{0}}(t) + \sum_{\boldsymbol{k}\in\mathcal{K}:\ \boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}} X_{\boldsymbol{k}}(t).$$

If $X_{(i)}(t) = 0$, then the customer is placed into an empty server. Furthermore, it is important, that immediately after any arrival or departure at time t, the value of $X_{\boldsymbol{0}}(t+)$ is reset (if necessary) to $f(\boldsymbol{X}(t+))$.

Note that the number $X_{\boldsymbol{0}} = f(\boldsymbol{X})$ of zero servers is finite at all times even though there is always an infinite number of empty servers available. Also, recall that a system state $\boldsymbol{X} = \{X_{\boldsymbol{k}},\ \boldsymbol{k} \in \mathcal{K}\}$ only includes the quantities of servers in nonzero configurations, so it does not include $X_{\boldsymbol{0}}$. Clearly, for a given function $f(\cdot)$, GRAND is a well-defined placement algorithm.

The focus of this paper is the following specialization of the general GRAND algorithm.

Definition 3 (GRAND(Z^p) Algorithm). A special case of the GRAND algorithm with the number of zero servers depending only on the total number of customers as

$$X_{\boldsymbol{0}}(t) = \lceil Z(t)^p \rceil \quad \forall t \geq 0,$$

where $p \in (0, 1)$ is a parameter, is called a GRAND(Z^p) algorithm.

In Stolyar and Zhong (2015), we introduce a different specialization of the general GRAND algorithm, namely the GRAND(aZ) algorithm, in which the number of zero servers $X_{\boldsymbol{0}}$ depends on Z, the total number of customers, as $X_{\boldsymbol{0}} = \lceil aZ \rceil$, $a > 0$. GRAND(Z^p) is similar to GRAND(aZ) in that, under both algorithms, $X_{\boldsymbol{0}}$ only depends on the system state $\boldsymbol{X}$ through Z, the total number of customers.

As we can see, GRAND are fairly "blind" algorithms in that, when they place a customer, they do not prefer one configuration over another as long as they can fit this additional customer. Similar to GRAND(aZ), GRAND(Z^p) can be efficiently implemented. At all times, the algorithm only needs to keep track of $I+1$ variables to make placement decisions with the variables being (a) Z, the total number of customers, and (b) for each $i \in \mathcal{I}$, $X_{(i)}$, the total number of servers that can fit an additional type i customer, including zero servers and occupied ones. In contrast to GRAND(Z^p) and GRAND(aZ), the greedy algorithm (Stolyar 2013) and the greedy-with-sublinear-safety-stocks algorithm (Stolyar and Zhong (2013) both need to keep track of $|\mathcal{K}|$ number of variables, which is often prohibitively large in real applications.

2.3. Asymptotic Regime

We are interested in the asymptotic properties of the GRAND(Z^p) algorithms as the arrival rates become large. Specifically, assume that a positive *scaling parameter* r increases to infinity along a discrete sequence. Customer arrival rates scale linearly with r; that is, for each r, $\Lambda_i = \lambda_i r$, where λ_i are fixed positive parameters. Without loss of generality, we assume that $\sum_i \lambda_i/\mu_i = 1$ by suitably redefining r if required. For each r, let $\boldsymbol{X}^r(\cdot)$ be the random process associated with the system parameterized by r, let $\boldsymbol{X}^r(\infty)$ be the (random) system state in the stationary regime, and let $X^r_{\mathbf{0}}(t)$ be the number of zero servers in the system at time t. For each $i \in \mathcal{I}$, let $Y^r_i(t) = \sum_{\boldsymbol{k}\in\mathcal{K}} k_i X^r_{\boldsymbol{k}}(t)$ be the total number of type i customers at time t, and let $Y^r_i(\infty) = \sum_{\boldsymbol{k}\in\mathcal{K}} k_i X^r_{\boldsymbol{k}}(\infty)$. Similarly, let $Z^r(t) = \sum_{i\in\mathcal{I}} Y^r_i(t)$ be the total number of customers at time t and let $Z^r(\infty) = \sum_i Y^r_i(\infty)$. As explained in Section 2.2, for each $i \in \mathcal{I}$, $Y^r_i(\infty)$ is an independent Poisson random variable with mean $r\rho_i$, where $\rho_i \equiv \lambda_i/\mu_i$, and $Z^r(\infty)$ is a Poisson random variable with mean $r\sum_i \rho_i = r$.

The *fluid-scaled* processes are defined to be $\{\boldsymbol{x}^r(t) : t \geq 0\}$ for each r, where $\boldsymbol{x}^r(t) = \boldsymbol{X}^r(t)/r$. We also define $\boldsymbol{x}^r(\infty) = \boldsymbol{X}^r(\infty)/r$. For any r, $\boldsymbol{x}^r(t)$ takes values in the nonnegative orthant $\mathbb{R}^{|\mathcal{K}|}_+$. Similarly, $y^r_i(t) \doteq Y^r_i(t)/r$, $z^r(t) \doteq Z^r(t)/r$, $x^r_{\mathbf{0}}(t) \doteq X^r_{\mathbf{0}}(t)/r$, and $x^r_{(i)}(t) \doteq X^r_{(i)}(t)/r$ for $t \in [0, \infty]$. Because $\sum_{\boldsymbol{k}\in\mathcal{K}} x^r_{\boldsymbol{k}}(\infty) \leq \sum_i y^r_i(\infty) \leq z^r(\infty) = Z^r(\infty)/r$, the family of random variables $\sum_{\boldsymbol{k}\in\mathcal{K}} x^r_{\boldsymbol{k}}(\infty)$ is uniformly integrable in r. This, in particular, implies that the sequence of distributions of $\boldsymbol{x}^r(\infty)$ is tight, so there always exists a limit $\boldsymbol{x}(\infty)$ in distribution, and $\boldsymbol{x}^r(\infty) \Rightarrow \boldsymbol{x}(\infty)$ along a subsequence of r.

Because $Y^r_i(\infty)$ is a Poisson random variable of mean $\rho_i r$, any weak limit point $y_i(\infty)$ of $\{y^r_i(\infty)\}_r$ must concentrate at the constant ρ_i. Thus, the limit (random) vector $\boldsymbol{x}(\infty)$ satisfies the following conservation laws:

$$\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}(\infty) \equiv y_i(\infty) = \rho_i, \quad \forall i, \tag{3}$$

which, in particular, implies that

$$z(\infty) \equiv \sum_i y_i(\infty) \equiv \sum_i \rho_i = 1. \tag{4}$$

Therefore, the values of $\boldsymbol{x}(\infty)$ are confined to the convex compact $(|\mathcal{K}| - I)$-dimensional polyhedron

$$\mathcal{X} \doteq \left\{ \boldsymbol{x} \in \mathbb{R}^{|\mathcal{K}|}_+ \;\middle|\; \sum_{\boldsymbol{k}} k_i x_{\boldsymbol{k}} = \rho_i, \ \forall i \in \mathcal{I} \right\}. \tag{5}$$

We slightly abuse notation by using symbol $\boldsymbol{x}$ for a generic element of $\mathbb{R}^{|\mathcal{K}|}_+$, and $\boldsymbol{x}^r(\infty), \boldsymbol{x}(\infty), \boldsymbol{x}^r(t)$, and later $\boldsymbol{x}(t)$ refer to random vectors taking values in $\mathbb{R}^{|\mathcal{K}|}_+$.

Note that the asymptotic regime and the associated basic properties (3) and (4) hold *for any placement algorithm*. Indeed, (3) and (4) only depend on the fact that $Y^r_i(\infty)$ are mutually independent Poisson random variables with means $\rho_i r$, discussed earlier.

Consider the following linear program (LP) of minimizing $\sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}$, the number of occupied servers on the fluid scale.

$$\text{(LP)} \qquad \text{Minimize} \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}, \tag{6}$$

$$\text{subject to} \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i \quad \forall i \in \mathcal{I}, \tag{7}$$

$$x_{\boldsymbol{k}} \geq 0, \quad \forall \boldsymbol{k} \in \mathcal{K}. \tag{8}$$

Denote by $\mathcal{X}^* \subseteq \mathcal{X}$ the set of its optimal solutions, and by L^* its optimal value. Because constraints (7) hold for $\boldsymbol{x}(\infty)$ under any placement algorithm, the optimal value L^* provides a lower bound on $\sum_{\boldsymbol{k}} x_{\boldsymbol{k}}(\infty)$, the steady-state total number of occupied servers on the fluid scale under any placement algorithm.

3. Main Result

The main result of this paper is the following theorem. It shows that GRAND(Z^p) is asymptotically optimal when parameter p is sufficiently close to one, which depends on the structure of the packing constraint specified by $\mathcal{K}$.

Theorem 1. *Denote $\kappa \doteq 1 + \max_{\boldsymbol{k}} \sum_i k_i$, and assume that $p < 1$ and*

$$1 - \kappa(1 - p) > 7/8, \quad \textit{or equivalently}, \quad p > 1 - 1/(8\kappa). \tag{9}$$

Consider a sequence of systems with parameter $r \to \infty$ operating under the GRAND(Z^p) algorithm. For each r, let $\boldsymbol{x}^r(\infty)$ denote the random state of the fluid-scaled process in the stationary regime. Then as $r \to \infty$,

$$d(\boldsymbol{x}^r(\infty), \mathcal{X}^*) \Rightarrow 0.$$

We would like to compare our main result, Theorem 1, with the main results (theorems 3 and 4) of Stolyar and Zhong (2015) on the asymptotic performance of the GRAND(aZ) algorithm. For any given p that satisfies (9), GRAND(Z^p) is asymptotically optimal in the sense that the *fluid-scaled optimality gap*, as measured by $d(\boldsymbol{x}^r(\infty), \mathcal{X}^*)$, goes to zero as the system scale r grows to infinity. In contrast, GRAND(aZ) is asymptotically optimal in the following (essentially, weaker) sense: for any fixed parameter $a > 0$ and under GRAND(aZ), as $r \to \infty$, the fluid-scaled optimality gap $d(\boldsymbol{x}^r(\infty), \mathcal{X}^*)$ converges to a constant $c(a)$, which is not necessarily (and typically is not) zero; however, $c(a) \to 0$ as $a \to 0$. So, speaking informally, GRAND(Z^p) optimality requires only that $r \to \infty$ although the optimality of GRAND(aZ) requires that $r \to \infty$ and then $a \to 0$.

Let us also remark on the role of condition (9) on the parameter $p \in (0, 1)$ in Theorem 1. This condition is technical and needed for our technical approach to work. It requires p to be *sufficiently close* to one. Roughly speaking, when the parameter p is sufficiently close to one, this guarantees that $x_{\boldsymbol{0}}^r$ is sufficiently "large," approximately r^{p-1}. This induces the steady-state values of $x_{\boldsymbol{k}}^r$ to be also large, at least $r^{7/8-1}$ for *all* configurations $\boldsymbol{k}$ (see Lemma 6). This, in turn, allows us to obtain desired bounds on the drift and increments of an appropriate Lyapunov function under GRAND(Z^p) (see Sections 5.3 and 5.4 for details).

4. Preliminaries

In this section, we present some definitions and facts from Stolyar and Zhong (2015) to provide enough background needed for the proof of Theorem 1, our main result. The main purpose of this section is to introduce the Lyapunov function $L^{(a)}$ defined in (15), parameterized by $a > 0$, together with its "drift" term Ξ, defined in (28) when the system is operating under the corresponding GRAND algorithm. In Stolyar and Zhong (2015), the term Ξ is shown to be the time derivative of $L^{(a)}$ in the fluid limit under the GRAND(aZ) algorithm. For the GRAND(Z^p) algorithm that we study in this paper, we use the Lyapunov function $L^{(a)}$ with a depending on r, namely $a = r^{p-1}$, to prove Theorem 1; in this case, the term Ξ approximates the drift of $L^{(a)}$, and for any large fixed r, the random increment of $L^{(a)}$ can be shown to be well approximated by a time integral of Ξ with high probability. Along the way, we also introduce some useful properties of the LP (6)–(8), $L^{(a)}$ and Ξ.

4.1. Dual Characterization of (LP)

Using the monotonicity of $\bar{\mathcal{K}}$ (cf. Definition 1), it is easy to check that if, in the program (LP) defined by (6)–(8), we replace the equality constraints (7) with the inequality constraints

$$\sum_{\boldsymbol{k} \in \mathcal{K}} k_i x_{\boldsymbol{k}} \geq \rho_i, \quad \forall i, \tag{10}$$

we form a new linear program (LP') with the same optimal value as (LP). More explicitly, (LP') is given by

$$\begin{aligned} \text{(LP')} \quad \text{Minimize} \ & \sum_{\boldsymbol{k} \in \mathcal{K}} x_{\boldsymbol{k}} \\ \text{subject to} \ & \sum_{\boldsymbol{k} \in \mathcal{K}} k_i x_{\boldsymbol{k}} \geq \rho_i \quad \forall i \in \mathcal{I}, \\ & x_{\boldsymbol{k}} \geq 0 \quad \forall \boldsymbol{k} \in \mathcal{K}. \end{aligned}$$

Let $\mathcal{X}^{**}$ denote the set of optimal solutions of (LP′). Then, $\mathcal{X}^{**}$ contains $\mathcal{X}^{*}$, the set of optimal solutions of (LP), or more precisely, $\mathcal{X}^{*} = \mathcal{X}^{**} \cap \mathcal{X}$. The dual program (DUAL′) of (LP′) is given by

$$\text{(DUAL}') \quad \text{Maximize} \quad \sum_{i\in\mathcal{I}} \rho_i \eta_i, \tag{11}$$

$$\text{subject to} \quad \sum_{i\in\mathcal{I}} k_i \eta_i \le 1, \quad \forall \boldsymbol{k} \in \mathcal{K}, \tag{12}$$

$$\eta_i \ge 0, \quad \forall i \in \mathcal{I}. \tag{13}$$

The following lemma is a simple consequence of the Kuhn–Tucker theorem.

Lemma 1. *Vector $\boldsymbol{x}$ is an optimal solution of (LP), that is, $\boldsymbol{x} \in \mathcal{X}^*$, if and only if $\boldsymbol{x} \in \mathcal{X}$, and there exists a vector $\boldsymbol{\eta} = \{\eta_i,\ i \in \mathcal{I}\}$ that satisfies* (12), (13), *and the complementary slackness condition*

$$\textit{for any } \boldsymbol{k} \in \mathcal{K}, \text{ if } \sum_i k_i \eta_i < 1, \textit{ then } x_{\boldsymbol{k}} = 0. \tag{14}$$

(Clearly, any such vector $\boldsymbol{\eta}$ is an optimal solution of (DUAL′).)

Let $\mathcal{H}^*$ be the set of vectors $\boldsymbol{\eta}$ that satisfy (12)–(14) for some $\boldsymbol{x} \in \mathcal{X}$.

4.2. A Useful Lyapunov Function

For any parameter $a \in (0,1)$, define the function $L^{(a)} : \mathbb{R}_+^{|\mathcal{K}|} \to \mathbb{R}$ by

$$L^{(a)}(\boldsymbol{x}) \doteq -\frac{1}{\log a} \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} \log\left(\frac{x_{\boldsymbol{k}} c_{\boldsymbol{k}}}{ea}\right), \tag{15}$$

where $c_{\boldsymbol{k}} \doteq \prod_i k_i!$, $0! = 1$, and we use the convention that $0 \log 0 = 0$.

Function $L^{(a)}(\boldsymbol{x})$ can be viewed as as an approximation to the linear function $\sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}$. Indeed, it has been shown Stolyar and Zhong (2015) that, as $a \to 0$, $L^{(a)}(\boldsymbol{x}) \to \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}$ uniformly over any compact set.

Function $L^{(a)}(\boldsymbol{x})$ is introduced in Stolyar and Zhong (2015) in the analysis of the GRAND(aZ) algorithm. Its unique optimal point $\boldsymbol{x}^{*,a}$ has a product form (see Lemma 2), and $L^{(a)}(\boldsymbol{x})$ serve as a Lyapunov function to show that, in the fluid limit, $\boldsymbol{x}^{*,a}$ is the unique equilibrium point under GRAND(aZ). The rest of this section is devoted to stating properties of—or related to—function $L^{(a)}(\boldsymbol{x})$. For more details on the intuition for the Lyapunov function $L^{(a)}(\cdot)$, see remark 1 of Stolyar and Zhong (2015). Note that, in the analysis of GRAND(Z^p) in this paper, $L^{(a)}(\boldsymbol{x})$ also serve as a Lyapunov function but with the parameter a now depending on r as $a = r^{p-1}$; this is the source of one of the key challenges we face in this paper.

Throughout the paper, we use notation $b = -\log a$. Then, for each $\boldsymbol{k} \in \mathcal{K}$, we have

$$\frac{\partial}{\partial x_{\boldsymbol{k}}} L^{(a)}(\boldsymbol{x}) = \frac{1}{b} \log\left(\frac{c_{\boldsymbol{k}} x_{\boldsymbol{k}}}{a}\right). \tag{16}$$

Note that, if we adopt the convention that

$$\left.\frac{\partial}{\partial x_{\boldsymbol{0}}} L^{(a)}(\boldsymbol{x})\right|_{x_{\boldsymbol{0}}=a} = 0, \tag{17}$$

then (16) is valid for $\boldsymbol{k} = \boldsymbol{0}$ and $x_{\boldsymbol{0}} = a$, which is useful later.

Function $L^{(a)}$ is strictly convex in $\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|}$. Consider the problem $\min_{\boldsymbol{x}\in\mathcal{X}} L^{(a)}(\boldsymbol{x})$. It is the following convex optimization problem (CVX(a)):

$$\text{(CVX}(a)) \qquad \text{Minimize} \quad L^{(a)}(\boldsymbol{x}), \tag{18}$$

$$\text{subject to} \quad \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i, \quad \forall i \in \mathcal{I}, \tag{19}$$

$$x_{\boldsymbol{k}} \ge 0, \quad \forall \boldsymbol{k} \in \mathcal{K}. \tag{20}$$

Denote by $\boldsymbol{x}^{*,a} \in \mathcal{X}$ its unique optimal solution. The following lemma provides a crisp characterization of the point $\boldsymbol{x}^{*,a}$.

Lemma 2. *A point $\boldsymbol{x} \in \mathcal{X}$ is the optimal solution to (CVX(a)) defined by (18)–(20), that is, $\boldsymbol{x} = \boldsymbol{x}^{*,a}$, if and only if it has a product form representation*

$$x_{\boldsymbol{k}}^{*,a} = \frac{a}{c_{\boldsymbol{k}}} \exp\left[b \sum_i k_i \nu_i^{*,a}\right] = \frac{1}{c_{\boldsymbol{k}}} a^{1-\sum_i k_i \nu_i^{*,a}}, \quad \forall \boldsymbol{k} \in \mathcal{K}, \tag{21}$$

for some vector $\boldsymbol{\nu}^{*,a} = \{\nu_i^{*,a},\ i \in \mathcal{I}\}$.

Proof. For each $i \in \mathcal{I}$, let ν_i be the dual variable that corresponds to the equality constraint $\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \rho_i$, and consider the Lagrangian

$$L^{(a)}(\boldsymbol{x}) + \sum_i \nu_i \left(\rho_i - \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}\right).$$

By setting partial derivatives of the Lagrangian to zero, we get

$$\frac{1}{b} \log\left[\frac{x_{\boldsymbol{k}} c_{\boldsymbol{k}}}{a}\right] - \sum_i \nu_i k_i = 0, \quad \forall \boldsymbol{k} \in \mathcal{K}.$$

$\boldsymbol{x}^{*,a} \in \mathcal{X}$ is the optimal solution to (CVX(a)) if and only if there exist Lagrange multipliers $\boldsymbol{\nu}^{*,a} = \{\nu_i^{*,a},\ i \in \mathcal{I}\}$ such that

$$\frac{1}{b} \log\left[\frac{x_{\boldsymbol{k}}^{*,a} c_{\boldsymbol{k}}}{a}\right] - \sum_i \nu_i^{*,a} k_i = 0, \quad \forall \boldsymbol{k} \in \mathcal{K},$$

which leads to the product form representation (21).

A simple consequence of Lemma 2 is that the Lagrange multipliers $\nu_i^{*,a}$ are unique and are equal to $1 - \log(x_{\boldsymbol{e}_i}^{*,a}) / \log a$ by considering (21) for $\boldsymbol{e}_i$, $i \in \mathcal{I}$. The following result is from Stolyar and Zhong (2015).

Theorem 2. *Let $\boldsymbol{x}^{*,a}$ and $\boldsymbol{\nu}^{*,a}$ be as in Lemma 2. Then, as $a \downarrow 0$, $\boldsymbol{x}^{*,a} \to \mathcal{X}^*$ and $\boldsymbol{\nu}^{*,a} \to \mathcal{H}^*$.*

Note that the convergence $\boldsymbol{x}^{*,a} \to \mathcal{X}^*$ is an easy consequence of the fact that $L^{(a)}(\boldsymbol{x}) - \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}} \to 0$ uniform on compact sets as $a \to 0$.

In Stolyar and Zhong (2015), we considered *fluid limit* trajectories $\boldsymbol{x}(\cdot)$, which, speaking informally, arise as limits of the fluid-scaled process $\boldsymbol{x}^r(\cdot)$ as $r \to \infty$. We show there that, for any fixed $a > 0$, under GRAND(aZ), the derivative $(d/dt)L^{(a)}(\boldsymbol{x}(t))$ along a fluid limit trajectory $\boldsymbol{x}(\cdot)$ is always negative whenever $\boldsymbol{x}(t) \neq \boldsymbol{x}^{*,a}$; consequently, $L^{(a)}$ serves as a Lyapunov function, which allows us to prove that $\boldsymbol{x}(t) \to \boldsymbol{x}^{*,a}$ as $t \to \infty$ along any fluid limit trajectory. This, in turn, is the key property that led to establishing the fact that the random steady-state system state $\boldsymbol{x}^r(\infty)$ concentrates on $\boldsymbol{x}^{*,a}$ as $r \to \infty$; together with Theorem 2, this means the asymptotic optimality of GRAND(aZ) if the limit $r \to \infty$ of steady states $\boldsymbol{x}^r(\infty)$ is taken first and the limit $a \to 0$ is taken after that. In this paper, we also make use of the family of Lyapunov functions $L^{(a)}(\boldsymbol{x})$ defined in (15) but use them in a different way because the analysis of GRAND(Z^p) cannot (as is explained later) rely on an analysis of fluid-limit trajectories.

We continue to introduce some expressions and their properties, derived in Stolyar and Zhong (2015) and that are closely related to the Lyapunov function $L^{(a)}$. In Stolyar and Zhong (2015), the expression $\Xi(\boldsymbol{x})$ defined in (28) is the derivative of the Lyapunov function $L^{(a)}(\boldsymbol{x}(t))$ with respect to time along a fluid limit trajectory $\boldsymbol{x}(\cdot)$. In the context of this paper, Ξ is interpreted and used differently; roughly speaking, it is the value of the process generator when applied to $L^{(a)}$.

Consider $\boldsymbol{x} \in \mathbb{R}_+^{|\mathcal{K}|}$ such that $x_{\boldsymbol{k}} > 0$, $\forall \boldsymbol{k} \in \mathcal{K}$. For two edges $(\boldsymbol{k}, i)$ and $(\boldsymbol{k}', i)$, consider the ordered pair $((\boldsymbol{k}, i), (\boldsymbol{k}', i))$, for which we define

$$\xi_{\boldsymbol{k},\boldsymbol{k}',i} = \frac{1}{b}\left[\log\left(k_i' x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}\right) - \log\left(k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}\right)\right] \frac{k_i \mu_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{x_{(i)}}, \tag{22}$$

where, by convention,

$$x_{\boldsymbol{0}} = a \text{ and, correspondingly, } x_{(i)} \doteq x_{\boldsymbol{0}} + \sum_{\boldsymbol{k}\in\mathcal{K}:\ \boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}} x_{\boldsymbol{k}}, \tag{23}$$

and $\xi_{\boldsymbol{k},\boldsymbol{k}',i}$ is defined for cases in which either $\boldsymbol{k}-\boldsymbol{e}_i=\boldsymbol{0}$ or $\boldsymbol{k}'-\boldsymbol{e}_i=\boldsymbol{0}$ as well. It is convenient, for each ordered pair of edges $((\boldsymbol{k},i),(\boldsymbol{k}',i))$, to define

$$\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})=\log\big(k_i'x_{\boldsymbol{k}-\boldsymbol{e}_i}x_{\boldsymbol{k}'}\big)-\log\big(k_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}\big)=\log\left[\frac{x_{\boldsymbol{k}-\boldsymbol{e}_i}}{k_ix_{\boldsymbol{k}}}\Big/\frac{x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{k_i'x_{\boldsymbol{k}'}}\right]. \tag{24}$$

Clearly, $|\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})|$ is the log-scale distance between the ratios $x_{\boldsymbol{k}-\boldsymbol{e}_i}/(k_ix_{\boldsymbol{k}})$ and $x_{\boldsymbol{k}'-\boldsymbol{e}_i}/(k_i'x_{\boldsymbol{k}'})$; in particular, $\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})=0$ if and only if $x_{\boldsymbol{k}-\boldsymbol{e}_i}/(k_ix_{\boldsymbol{k}})=x_{\boldsymbol{k}'-\boldsymbol{e}_i}/(k_i'x_{\boldsymbol{k}'})$. Note also that

$$\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})=-\chi_{\boldsymbol{k}',\boldsymbol{k},i}(\boldsymbol{x}). \tag{25}$$

Then, (22) can be written as

$$\xi_{\boldsymbol{k},\boldsymbol{k}',i}=\frac{k_i\mu_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{bx_{(i)}}\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x}). \tag{26}$$

And we have the following:

$$\begin{aligned}\xi_{\boldsymbol{k},\boldsymbol{k}',i}+\xi_{\boldsymbol{k}',\boldsymbol{k},i}&=\frac{k_i\mu_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{bx_{(i)}}\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})+\frac{k_i\mu_ix_{\boldsymbol{k}'}x_{\boldsymbol{k}-\boldsymbol{e}_i}}{bx_{(i)}}\chi_{\boldsymbol{k}',\boldsymbol{k},i}(\boldsymbol{x})\\&=\frac{k_i\mu_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}}{bx_{(i)}}\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})-\frac{k_i\mu_ix_{\boldsymbol{k}'}x_{\boldsymbol{k}-\boldsymbol{e}_i}}{bx_{(i)}}\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x}),\\&=\frac{\mu_i}{bx_{(i)}}\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})\big[k_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}-k_i'x_{\boldsymbol{k}-\boldsymbol{e}_i}x_{\boldsymbol{k}'}\big]\\&=\frac{\mu_i}{bx_{(i)}}\big[\log\big(k_i'x_{\boldsymbol{k}-\boldsymbol{e}_i}x_{\boldsymbol{k}'}\big)-\log\big(k_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}\big)\big]\big[k_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}-k_i'x_{\boldsymbol{k}-\boldsymbol{e}_i}x_{\boldsymbol{k}'}\big]\le 0.\end{aligned} \tag{27}$$

The inequality in (27) holds because the two terms in the square brackets are nonzero and have opposite signs unless they both are zero; therefore, the inequality in (27) is strict unless $k_i'x_{\boldsymbol{k}-\boldsymbol{e}_i}x_{\boldsymbol{k}'}=k_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}$.

Finally, we define

$$\Xi(\boldsymbol{x})=\sum_i\sum_{\boldsymbol{k},\boldsymbol{k}'}\big[\xi_{\boldsymbol{k},\boldsymbol{k}',i}+\xi_{\boldsymbol{k}',\boldsymbol{k},i}\big], \tag{28}$$

where the summation $\sum_{\boldsymbol{k},\boldsymbol{k}'}$ is over all *unordered* pairs $\{\boldsymbol{k},\boldsymbol{k}'\}$. Observe that $\Xi(\boldsymbol{x})<0$ unless $\boldsymbol{x}$ has a product form representation (21) for some vector $\boldsymbol{v}^{*,a}$. This is because $\Xi(\boldsymbol{x})=0$ if and only if the equality in (27) holds for all unordered pairs of edges $\{(\boldsymbol{k},i),(\boldsymbol{k}',i)\}$, which is true if and only if $k_i'x_{\boldsymbol{k}-\boldsymbol{e}_i}x_{\boldsymbol{k}'}=k_ix_{\boldsymbol{k}}x_{\boldsymbol{k}'-\boldsymbol{e}_i}$ (i.e., $\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x})=0$) for all ordered pairs $((\boldsymbol{k},i),(\boldsymbol{k}',i))$—a condition that is equivalent to a product form presentation (21) of $\boldsymbol{x}$ for some $\boldsymbol{v}^{*,a}$. Let us also note that, here, the vector $\boldsymbol{v}^{*,a}$ need not correspond to the (unique) Lagrange multipliers of the program (CVX(a)) because $\boldsymbol{x}$ does not necessarily satisfy the equalities (19); that is, we do not necessarily have $\boldsymbol{x}\in\mathcal{X}$ (recall the definition of $\mathcal{X}$ in (5)). If $\boldsymbol{x}\in\mathcal{X}$, we have $\boldsymbol{x}=\boldsymbol{x}^{*,a}$, exactly as in (21).

We now provide a more concrete interpretation of (28). Consider a state $\boldsymbol{x}$ and suppose that the time derivatives (or rates of changes) of $\boldsymbol{x}$ are defined by the following dynamics. Along each edge $(\boldsymbol{k},i)$, "mass" is moving from $x_{\boldsymbol{k}}$ to $x_{\boldsymbol{k}-\boldsymbol{e}_i}$ at the rate

$$\tilde{w}_{\boldsymbol{k},i}=k_i\mu_ix_{\boldsymbol{k}}. \tag{29}$$

(Value $\tilde{w}_{\boldsymbol{k},i}$ can be interpreted as the fluid-scaled rate of type i departures along the edge $(\boldsymbol{k},i)$.) In addition, along each edge $(\boldsymbol{k},i)$, mass is moving from $x_{\boldsymbol{k}-\boldsymbol{e}_i}$ to $x_{\boldsymbol{k}}$ at the rate

$$\tilde{v}_{\boldsymbol{k},i}=\frac{\tilde{\lambda}_ix_{\boldsymbol{k}-\boldsymbol{e}_i}}{x_{(i)}}, \tag{30}$$

where we denote

$$\tilde{\lambda}_i=\sum_{\boldsymbol{k}\in\mathcal{K}}k_i\mu_ix_{\boldsymbol{k}}, \tag{31}$$

and $x_{(i)}$ is defined by (23). Values $\tilde{v}_{\boldsymbol{k},i}$ and $\tilde{\lambda}_i$ defined in (30) and (31) can be interpreted as follows. If $\boldsymbol{x}\in\mathcal{X}$, then $\tilde{\lambda}_i=\lambda_i$ for all $i\in\mathcal{I}$ by the equality constraints (19), and $\tilde{v}_{\boldsymbol{k},i}$ of (30) is the fluid-scaled rate of type i arrivals along the edge $(\boldsymbol{k}, i)$ under the GRAND(Z^p) algorithm. In general, this is not necessarily the case because we consider a generic system state $\boldsymbol{x}$, which need not belong to $\mathcal{X}$.

Equations (29) and (30) define the derivatives of $x_{\boldsymbol{k}}$ for all $\boldsymbol{k} \in \mathcal{K}$. By convention, $x_{\mathbf{0}}$ remains constant at a, so it has zero derivative. By (16), let us note that, in the expression of $\xi_{\boldsymbol{k},\boldsymbol{k}',i}$ (recall (22)),

$$\frac{1}{b}\left[\log\left(k_i' x_{\boldsymbol{k}-\boldsymbol{e}_i} x_{\boldsymbol{k}'}\right) - \log\left(k_i x_{\boldsymbol{k}} x_{\boldsymbol{k}'-\boldsymbol{e}_i}\right)\right] = \left[\frac{\partial}{\partial x_{\boldsymbol{k}'}} L^{(a)}(\boldsymbol{x}) - \frac{\partial}{\partial x_{\boldsymbol{k}'-\boldsymbol{e}_i}} L^{(a)}(\boldsymbol{x})\right] - \left[\frac{\partial}{\partial x_{\boldsymbol{k}}} L^{(a)}(\boldsymbol{x}) - \frac{\partial}{\partial x_{\boldsymbol{k}-\boldsymbol{e}_i}} L^{(a)}(\boldsymbol{x})\right]. \tag{32}$$

Then, by Equations (29) and (30) (which define the derivatives of $x_{\boldsymbol{k}}$), the convention (23), and Equation (32), it is not difficult to check that $\Xi(\boldsymbol{x})$ is the time derivative of $L^{(a)}(\boldsymbol{x})$:

$$\Xi(\boldsymbol{x}) = \sum_{(\boldsymbol{k},i)\in\mathcal{M}} \left[\frac{\partial}{\partial x_{\boldsymbol{k}}} L^{(a)}(\boldsymbol{x}) - \frac{\partial}{\partial x_{\boldsymbol{k}-\boldsymbol{e}_i}} L^{(a)}(\boldsymbol{x})\right]\left[\tilde{v}_{\boldsymbol{k},i} - \tilde{w}_{\boldsymbol{k},i}\right], \tag{33}$$

where convention (17) is used.

5. Proof of Theorem 1

This section is organized as follows. In Section 5.1, we present the Lyapunov function that we use for each system indexed by r and derive some basic properties of this Lyapunov function. In particular, we see that the Lyapunov function can be thought of as an approximation to the linear objective $\sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}$. We also point out the main difficulties in establishing the asymptotic optimality of GRAND(Z^p) here. In Section 5.2, we show that some conditions hold for the process $\boldsymbol{x}(\cdot)$ in the stationary regime with high probability for large r. Section 5.3 is the key section in which we introduce an "artificial process," such that the conditions considered in Section 5.2 are "enforced," that is, they hold with probability one (w.p.1) (as opposed to "with high probability" as for the actual process). For this artificial process, we derive high-probability estimates on the change of the Lyapunov function. An informal "road map" is provided at the beginning of Section 5.3 for this derivation. We conclude the proof of Theorem 1 in Section 5.4, in which we use the high-probability bounds in Section 5.2 to show that, with high probability, the artificial and the actual processes coincide. This, in turn, allows us to obtain high-probability estimates on the change of the Lyapunov function under the actual process.

Throughout this section, we drop superscript r in notation—such as that appearing in $\boldsymbol{x}^r(t)$—pertaining to processes with parameter r.

5.1. A Lyapunov Function and Some Basic Properties

We use the Lyapunov function $L^{(a)}$ for the proof of Theorem 1, in which the parameter a depends on the system scale r as $a = r^{p-1}$. Correspondingly, b also depends on r with $b = -\log a = (1-p)\log r$. When it is clear from the context that we are working with $L^{(a)}$ for a given fixed r (and corresponding $a = r^{p-1}$), we often drop the superscript (a).

Let us now provide an important remark regarding the Lyapunov function $L^{(a)}$ with $a = r^{p-1}$ and the GRAND(Z^p) algorithm. On the one hand, for any given system scale parameter r, $L^{(a)}$ has a fixed parameter $a = r^{p-1}$. Furthermore, the expressions (33) and (28) for $\Xi(\boldsymbol{x})$, the time derivative of $L^{(a)}$, make use of the *convention* (23), in which $a = r^{p-1}$. On the other hand, GRAND(Z^p) itself does *not* use or need to know the parameter r because the actual number of zero servers that it uses at time t is $X_{\mathbf{0}}(t) = Z^p(t)$, which only uses the knowledge of $Z(t)$, the total number of customers at time t.

At this point, we can highlight the key technical difficulty of the analysis of GRAND(Z^p) compared with that of GRAND(aZ) in Stolyar and Zhong (2015). Under GRAND(aZ), the fluid-scaled quantities $x^r_{\boldsymbol{k}}$ are $O(1)$ for *all* configurations $\boldsymbol{k}$. Consequently, as $r \to \infty$, the fluid-scaled process $\boldsymbol{x}^r(\cdot)$ converges to a well-defined fluid limit $\boldsymbol{x}(\cdot)$ for any well-defined limiting initial state $\boldsymbol{x}(0)$. The Lyapunov function $L^{(a)}$ with fixed $a > 0$ is then used in Stolyar and Zhong (2015) to show the convergence of all fluid trajectories $\boldsymbol{x}(t)$ to the point $\boldsymbol{x}^{*,a}$ as $t \to \infty$, which is the key part of the analysis of GRAND(aZ). Therefore, in Stolyar and Zhong (2015), it suffices to work with fluid limits and derivatives of the Lyapunov function along fluid limit trajectories.

In this paper, to analyze GRAND(Z^p), we use $L^{(r^{p-1})}$ as a Lyapunov function for a system with given r. The key difficulty is that, under GRAND(Z^p), some of the fluid-scaled $x^r_{\boldsymbol{k}}$ are $o(1)$; for example, $x^r_{\mathbf{0}} = O(r^{p-1})$. A further complication is that different $x^r_{\boldsymbol{k}}$ are on *different scales* with respect to r; namely they are $O(r^{s(\boldsymbol{k})})$ with the power $s(\boldsymbol{k})$ depending on $\boldsymbol{k}$, $s(\boldsymbol{k}) \in (0,1]$. Because of the multiscale system dynamics, the conventional fluid limit is not sufficient to establish a negative drift of the Lyapunov function. Moreover, the use of other, *local* fluid limits (obtained under different space/time scalings) also does not appear to be particularly useful, because, again, $x_{\boldsymbol{k}}^r$ evolve on different scales. To overcome this difficulty, in Section 5.3, we use direct (and more involved) probabilistic estimates of the Lyapunov function increments, which are based on martingale theory and which do not involve fluid or local fluid limits.

The following lemmas state some elementary properties related to the Lyapunov function $L^{(a)}$ with $a = r^{p-1}$.

Lemma 3. *Let $\varepsilon \in (0, \frac{1}{2})$. Consider a sequence of points $\boldsymbol{x}^{\circ,a}$ indexed by $a = r^{p-1}$, which satisfy*

$$\left|\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}^{\circ,a} - \rho_i\right| \le \frac{r^{-1/2+\varepsilon}}{I}, \quad \forall i, \text{ for sufficiently large } r, \tag{34}$$

and

$$\chi_{\boldsymbol{k},\boldsymbol{k}',i}(\boldsymbol{x}^{\circ,a}) \to 0 \ \text{ for all pairs of edges } (\boldsymbol{k},i),(\boldsymbol{k}',i),$$

where we recall the definition of $\chi_{\boldsymbol{k},\boldsymbol{k}',i}$ in (24). Then, as $r \to \infty$, $L^{(a)}(\boldsymbol{x}^{\circ,a}) \to L^*$, where L^* is the optimal value of (LP) defined by (6)–(8).

Proof. For each r and the corresponding $a = r^{p-1}$, denote

$$\nu_i^{*,a} = 1 - \frac{\log\left(x_{\boldsymbol{e}_i}^{\circ,a}\right)}{\log a}, \quad \forall i \in \mathcal{I}.$$

Define $\boldsymbol{x}^{*,a} = \{x_{\boldsymbol{k}}^{*,a}, \boldsymbol{k} \in \mathcal{K}\}$ via product form (21) corresponding to these $\nu_i^{*,a}$:

$$x_{\boldsymbol{k}}^{*,a} = \frac{a}{c_{\boldsymbol{k}}} \exp\left[b \sum_i k_i \nu_i^{*,a}\right] = \frac{1}{c_{\boldsymbol{k}}} a^{1-\sum_i k_i \nu_i^{*,a}}, \quad \boldsymbol{k} \in \mathcal{K}. \tag{35}$$

Note that, for any r (and, correspondingly, $a = r^{p-1}$), $\boldsymbol{x}^{*,a}$ and $\boldsymbol{\nu}^{*,a}$ need not be the optimal primal and dual solutions to (CVX(a)), defined by (18)–(20) because $\boldsymbol{x}^{*,a}$ need not belong to $\mathcal{X}$. Instead, point $\boldsymbol{x}^{*,a}$ is the optimal solution for the problem

$$\text{minimize} \quad L^{(a)}(\boldsymbol{x}), \tag{36}$$

$$\text{subject to} \quad \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}} = \sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}^{*,a}, \qquad \forall i \in \mathcal{I}, \tag{37}$$

$$x_{\boldsymbol{k}} \ge 0, \qquad \forall \boldsymbol{k} \in \mathcal{K}. \tag{38}$$

We now claim that, for all $\boldsymbol{k} \in \mathcal{K}$,

$$x_{\boldsymbol{k}}^{\circ,a} / x_{\boldsymbol{k}}^{*,a} \to 1 \text{ as } r \to \infty. \tag{39}$$

By convention, $x_{\boldsymbol{0}}^{\circ,a} = x_{\boldsymbol{0}}^{*,a} = a = r^{p-1}$. It is also easy to check that, for all $i \in \mathcal{I}$, $x_{\boldsymbol{e}_i}^{\circ,a} = x_{\boldsymbol{e}_i}^{*,a}$. Consider a general configuration $\boldsymbol{k} \in \mathcal{K}$. For an edge $(\boldsymbol{k}, i)$, we use $\tilde{\chi}_{(\boldsymbol{k},i)}$ as a shorthand for $\chi_{\boldsymbol{e}_i,\boldsymbol{k},i}(\boldsymbol{x}^{\circ,a})$, and we have

$$\begin{aligned}
\tilde{\chi}_{(\boldsymbol{k},i)} &= \chi_{\boldsymbol{e}_i,\boldsymbol{k},i}(\boldsymbol{x}^{\circ,a}) \\
&= \log\left(k_i x_{\boldsymbol{0}}^{\circ,a} x_{\boldsymbol{k}}^{\circ,a}\right) - \log\left(x_{\boldsymbol{e}_i}^{\circ,a} x_{\boldsymbol{k}-\boldsymbol{e}_i}^{\circ,a}\right) \\
&= \log\left(k_i a x_{\boldsymbol{k}}^{\circ,a}\right) - \log\left(x_{\boldsymbol{e}_i}^{*,a} x_{\boldsymbol{k}-\boldsymbol{e}_i}^{\circ,a}\right).
\end{aligned}$$

By some simple algebraic manipulation, we have the recursion

$$x_{\boldsymbol{k}}^{\circ,a} = \frac{e^{\tilde{\chi}_{(\boldsymbol{k},i)}} x_{\boldsymbol{e}_i}^{*,a} x_{\boldsymbol{k}-\boldsymbol{e}_i}^{\circ,a}}{k_i a} = \frac{1}{k_i} e^{\tilde{\chi}_{(\boldsymbol{k},i)}} a^{-\nu_i^{*,a}} x_{\boldsymbol{k}-\boldsymbol{e}_i}^{\circ,a}. \tag{40}$$

Consider a path that connects configuration $\boldsymbol{k}$ and $\boldsymbol{0}$ using the following sequence of edges: the first k_1 edges are $(\boldsymbol{k},1), (\boldsymbol{k}-\boldsymbol{e}_1,1), \cdots, (\boldsymbol{k}-(k_1-1)\boldsymbol{e}_1,1)$, followed by k_2 edges of the form $(\boldsymbol{k}-(k_1-1)\boldsymbol{e}_1,2), (\boldsymbol{k}-(k_1-1)\boldsymbol{e}_1-\boldsymbol{e}_2,2), \cdots,$

$(\boldsymbol{k}-(k_1-1)\boldsymbol{e}_1-(k_2-1)\boldsymbol{e}_2,2)$, etc. Thus, the path uses $\sum_i k_i$ edges to connect $\boldsymbol{k}$ with $\boldsymbol{0}$. Using $\tilde{e}$ to denote a generic edge that belongs to this path and using the recursion (40), it is easy to see that

$$x_{\boldsymbol{k}}^{\circ,a} = \frac{1}{c_{\boldsymbol{k}}}\left[\prod_{\tilde{e}} \exp(\tilde{\chi}_{\tilde{e}})\right] a^{1-\sum_i k_i \nu_i^{*,a}} = \left[\prod_{\tilde{e}} \exp(\tilde{\chi}_{\tilde{e}})\right] x_{\boldsymbol{k}}^{*,a}.$$

Thus,

$$\frac{x_{\boldsymbol{k}}^{\circ,a}}{x_{\boldsymbol{k}}^{*,a}} = \left[\prod_{\tilde{e}} \exp(\tilde{\chi}_{\tilde{e}})\right] \to 1,$$

because $\tilde{\chi}_{\tilde{e}} \to 0$ for all edges $\tilde{e}$. This completes the proof of the claim.

By (39), we have

$$\left\|\boldsymbol{x}^{\circ,a} - \boldsymbol{x}^{*,a}\right\| \to 0, \quad r \to \infty.$$

In addition, using condition (34), we have that, for each $i \in \mathcal{I}$, the term $\sum_{\boldsymbol{k}\in\mathcal{K}} k_i x_{\boldsymbol{k}}^{*,a}$ on the right-hand side (RHS) of the equality constraint (37) converges to ρ_i as $r \to \infty$. Furthermore, recall that $\left|L^{(a)}(\boldsymbol{x}) - \sum_{\boldsymbol{k}\in\mathcal{K}} x_{\boldsymbol{k}}\right| \to 0$ uniformly on $\boldsymbol{x}$ from a bounded set. Therefore, we must have $L^{(a)}(\boldsymbol{x}^{*,a}) \to L^*$. □

Lemma 3 implies the following property that we actually use.

Lemma 4. *For any $\gamma > 0$, there exists $\delta_1 > 0$ such that, for all sufficiently large r and $a = r^{p-1}$, condition (34) and the condition that*

$$|\chi_{\boldsymbol{k},\boldsymbol{k}',i}| \le \delta_1 \quad \text{for all pairs of edges } (\boldsymbol{k},i),(\boldsymbol{k}',i)$$

imply

$$L^{(a)}(\boldsymbol{x}) - L^* \le \gamma.$$

5.2. High-Probability Steady-State Bounds on $\boldsymbol{x}(\cdot)$

In this section, we show that certain conditions hold under GRAND(Z^p) in steady state with high probability, converging to one as $r \to \infty$. Moreover, these conditions hold over time intervals with length increasing with r as r^α, $\alpha > 0$.

From now on in the paper, we fix p satisfying (9) and a corresponding s satisfying $7/8 < s < 1 - \kappa(1-p)$. We often refer to the following conditions at a given time $t \ge 0$ with some constants $c > 0$ and $\varepsilon > 0$:

$$|Z(t)/r - 1| \le r^{-1/2+\varepsilon}; \tag{41}$$

$$\left|\sum_{\boldsymbol{k}} k_i x_{\boldsymbol{k}}(t) - \rho_i\right| \le \frac{r^{-1/2+\varepsilon}}{I}, \quad \forall i; \tag{42}$$

$$x_{\boldsymbol{k}}(t) \ge c r^{s-1}, \quad \forall \boldsymbol{k}. \tag{43}$$

It is easy to see that (42) implies (41).

Lemma 5. *Let $\alpha > 0$ and $\varepsilon > 0$. Consider our system in the stationary regime for each r. Then,*

$$\mathbb{P}\{\textit{condition (42) holds for all } t \in [0, r^\alpha]\} \to 1, \quad r \to \infty.$$

The proof of Lemma 5 is provided in Appendix A. The statement of Lemma 5 is very intuitive because we know that, in the stationary regime, $Y_i(t)$ has Poisson distribution with mean $\rho_i r$ for any t. Because $Y_i(t) = \sum_{\boldsymbol{k}\in\mathcal{K}} k_i X_{\boldsymbol{k}}(t)$, condition (42) clearly holds for any given t. The proof shows that, in fact, with high probability, (42) holds on any r^α-long time interval.

Lemma 6. *Let $\alpha > 0$. Consider our system in the stationary regime for each r. Then, there exists $c > 0$ such that*

$$\mathbb{P}\{\text{condition (43) holds for } t \in [0, r^\alpha]\} \to 1, \quad r \to \infty.$$

The proof of Lemma 6 is provided in Appendix B, in which we establish a stronger property: with high probability, there exists some $c > 0$ such that, for all $\boldsymbol{k} \in \mathcal{K}$ and for all $t \in [0, r^\alpha]$, $x_{\boldsymbol{k}}(t) \ge c r^{s(\boldsymbol{k})-1}$, where

$s(\boldsymbol{k}) = 1 - (\sum_i k_i + 1)(1-p) \geq 1 - \kappa(1-p) > s$. Here, we provide a sketch of the argument used to establish this stronger property. By Lemma 5, with high probability, the fluid-scaled total number of customers, $Z(t)/r$, is in steady state close to one for all $t \in [0, r^{\alpha}]$. Therefore, $x_{\mathbf{0}}$ is close to $Z^p/r \approx r^{p-1}$, and the property should hold for $\boldsymbol{k} = \mathbf{0}$. Then, we use an induction argument. Fix configuration $\boldsymbol{k} \neq \mathbf{0}$ and suppose the property is true for all configurations $\boldsymbol{k}' \neq \boldsymbol{k}$, $\boldsymbol{k}' \leq \boldsymbol{k}$. We argue that it should then hold for $\boldsymbol{k}$ and some c (possibly rechosen to be smaller) as follows. Pick $\boldsymbol{k}' = \boldsymbol{k} - \boldsymbol{e}_i$ for some i; such $\boldsymbol{k}'$ exists because $\boldsymbol{k} \neq \mathbf{0}$. We account separately for transitions that increase $x_{\boldsymbol{k}}$ and those that decrease $x_{\boldsymbol{k}}$. First, type i arrivals along edge $(\boldsymbol{k}, i)$ increase $x_{\boldsymbol{k}}$ and take place at (unscaled) rate $\lambda_i r x_{\boldsymbol{k}'}/x_{(i)}$. Furthermore, this arrival process is lower bounded by a Poisson process of (unscaled) rate $c_1 r(r^{s(\boldsymbol{k}')}/r) = c_1 r^{s(\boldsymbol{k}')}$ for a constant $c_1 > 0$ because $X_{(i)}/r \leq (X_{\mathbf{0}} + Z)/r$ is essentially upper bounded by a constant and $x_{\boldsymbol{k}'} \geq c(r^{s(\boldsymbol{k}')}/r)$ by the induction hypothesis. Second, transitions that would decrease $x_{\boldsymbol{k}}$ consist of departures from $\boldsymbol{k}$ and arrivals into $\boldsymbol{k}$. When $x_{\boldsymbol{k}}(t) \leq c r^{s(\boldsymbol{k})-1}$, the (unscaled) rates of these transitions are upper bounded by $c_2 c r^{s(\boldsymbol{k})}$ and $c_3 c r(r^{s(\boldsymbol{k})}/r^p) = c_3 c r^{s(\boldsymbol{k})+1-p}$ (because $X_{(i)} \geq X_{\mathbf{0}} \approx r^p$), respectively, for some constants $c_2 > 0$ and $c_3 > 0$. Thus, the transitions that would decrease $x_{\boldsymbol{k}}$ are upper bounded by a Poisson process with (unscaled) rate $c_2 c r^{s(\boldsymbol{k})} + c_3 c r^{s(\boldsymbol{k})+1-p} \leq c_4 c r^{s(\boldsymbol{k})+1-p} = c_4 c r^{s(\boldsymbol{k}')}$ for some $c_4 > 0$. Therefore, if we choose constant c to be sufficiently small, then, when $x_{\boldsymbol{k}}(t) \leq c r^{s(\boldsymbol{k})}/r$, the rate of transitions that increases $x_{\boldsymbol{k}}$ dominates the rate of transitions that decreases it, at least by some factor greater than one. Thus, with high probability, $x_{\boldsymbol{k}}(t)$ *should* stay above $c r^{s(\boldsymbol{k})-1}$.

5.3. Lyapunov Function Drift Estimates for an Artificial Process

We start with an informal overview of this section and how its results are used in the next one.

In this section, we consider an *artificial* process for which conditions (41)–(43) are enforced over a r^{s-1}-long interval; the construction is such that the *original* process pathwise coincides with the artificial one as long as (41)–(43) hold. For the artificial process, we obtain high-probability estimates of the increment of a certain process $F(t)$ over an r^{s-1}-long interval. The construction of $F(t)$ is such that, if sample paths of the artificial and original processes coincide, then $F(t)$ is equal to the increment of the Lyapunov function $L(\boldsymbol{x}(t))$. (Recall that $L = L^{(r^{p-1})}$ depends on r.)

The goal of this section is to obtain high-probability estimates of $F(t)$. To do that, we use the martingale representation of $F(t)$, in which the compensator is $B(t) = \int_0^t AL(\boldsymbol{x}(\xi))d\xi$, where operator A is closely related to the generator of the artificial (and original) process. We then derive bounds on $|AL(\boldsymbol{x}(\xi)) - \Xi(\boldsymbol{x}(\xi))|$, which allows us to obtain a bound on $|B(t) - \int_0^t \Xi(\boldsymbol{x}(\xi))d\xi|$. We also obtain an upper bound on the quadratic characteristic (predictable quadratic variation) of the martingale $M(t) = F(t) - B(t)$, which, by Doob's inequality, gives us a high-probability bound on $|M(t)|$. Combining the latter two bounds, we obtain one on $|F(t) - \int_0^t \Xi(\boldsymbol{x}(\xi))d\xi|$, namely we can claim that it is "small." Finally, we observe that $\Xi(\boldsymbol{x}(\xi)) \leq 0$ for any $\boldsymbol{x}(\xi)$ and use Lemma 4 to show that $\Xi(\boldsymbol{x}(\xi))$ is negative and bounded away from zero as long as $\boldsymbol{x}(\xi)$ is bounded away from the optimal set $\mathcal{X}^*$ of (LP) in (6)–(8). Combining all this, we can claim that, with high probability, $F(t)$ has a negative (bounded away from zero) increment over a r^{s-1}-long interval, in which $\boldsymbol{x}(t)$ "stays away" from the optimal set $\mathcal{X}^*$.

In Section 5.4 we use the fact that (41)–(43) hold for the original process with high probability to couple the artificial and original processes in a way such that they coincide with high probability. Therefore, with high probability, the increment of the Lyapunov function $L(\boldsymbol{x}(t))$ for the original process is equal to $F(t)$, for which the estimates of this section apply.

This ends the informal overview. Let us proceed with formal results.

In Sections 5.3 and 5.4, we use the following convention. By C_g, we denote a generic positive constant, which does not depend on the system parameters or on the scaling parameter r; the value of C_g may be different in different expressions in which it appears.

Consider a single r^{s-1}-long interval and let $T = r^{s-1}$. The initial state is such that, at $t = 0$, conditions (41)–(43) hold for some constant $c > 0$ and a small constant $\varepsilon > 0$, where the magnitude of ε is specified later (at the end of this section).

Let $\widehat{\mathcal{X}} = \widehat{\mathcal{X}}^r$ denote the set of states $\boldsymbol{x}$ that satisfy (41)–(43). For $\boldsymbol{x} \in \widehat{\mathcal{X}}$, we have

$$\left|\frac{\partial}{\partial x_{\boldsymbol{k}}} L(\boldsymbol{x})\right| \leq C_g, \tag{44}$$

$$0 \leq \left|\frac{\partial^2}{\partial x_{\boldsymbol{k}}^2} L(\boldsymbol{x})\right| \leq C_g\left[r^{s-1}\log r\right]^{-1} \leq C_g r^{1-s}, \tag{45}$$

for all sufficiently large r. To see inequality (44), first recall from (16) that

$$\frac{\partial}{\partial x_k} L(\boldsymbol{x}) = \frac{1}{b}\log(\frac{c_k x_k}{a}) = \frac{1}{b}\log(c_k x_k) + 1$$

for any $\boldsymbol{k}$ because $b = -\log a$. Also, because $\boldsymbol{x} \in \widehat{\mathcal{X}}$ satisfies condition (41), $x_k \le z \le 2$ for sufficiently large r. Thus, (44) holds for all sufficiently large r. To see inequality (45), note that

$$\frac{\partial^2}{\partial x_k^2} L(\boldsymbol{x}) = \frac{1}{b x_k},$$
$$b = -\log a = (1-p)\log r$$

(because $a = r^{p-1}$), and $x_k \ge c r^{s-1}$ by condition (43).

Our process $\boldsymbol{x}(t)$ is a pure jump process; the jumps are caused by customer arrivals and departures. For the rest of this section, we consider an artificial version of the process, which we now describe. In this artificial process, for which, with some abuse, we keep the same notation $\boldsymbol{x}(\cdot)$, some of the jumps (i.e., transitions) are not allowed to occur. Specifically, if ξ is a point in time when an arrival or departure takes place, we call it a point of *potential transition*. If this potential transition keeps the state within conditions (41)–(43), we allow it to occur, so this becomes an actual transition for the artificial process. If the potential transition were to take us to a state violating (41)–(43), we do *not* allow this transition to actually occur and keep the state unchanged.

We now introduce some notation related to the artificial process. For $\boldsymbol{x} \in \widehat{\mathcal{X}}$, let $\mathcal{T}(\boldsymbol{x}) = \mathcal{T}^r(\boldsymbol{x})$ denote the set of all potential transitions out of it. Any potential transition $\alpha \in \mathcal{T}(\boldsymbol{x})$ uniquely identifies the "from" state, which is $\boldsymbol{x}$, and the "to" state, which we denote by $\boldsymbol{x}_\alpha$. Thus, the sets $\mathcal{T}(\boldsymbol{x})$ for different $\boldsymbol{x}$ are nonintersecting. Let $\mathcal{T} = \mathcal{T}^r = \cup_{\boldsymbol{x}\in\widehat{\mathcal{X}}}\mathcal{T}(\boldsymbol{x})$ be the set of all possible potential transitions from the states $\boldsymbol{x} \in \widehat{\mathcal{X}}$. Note that, for any r, the cardinalities of $\widehat{\mathcal{X}}$ and $\mathcal{T}$ are both finite. Denote by $\widehat{\mathcal{T}}(\boldsymbol{x}) \subseteq \mathcal{T}(\boldsymbol{x})$ the set of all allowed potential transitions out of $\boldsymbol{x}$, that is, those for which $\boldsymbol{x}_\alpha \in \widehat{\mathcal{X}}$. Let $\widehat{\mathcal{T}} = \cup_{\boldsymbol{x}\in\widehat{\mathcal{X}}}\widehat{\mathcal{T}}(\boldsymbol{x}) \subseteq \mathcal{T}$ denote the set of all potential transitions that are allowed. Finally, for a potential transition α, denote by β_α its (infinitesimal) rate.

We can and do construct the artificial process on the following probability space. Let the initial state $\boldsymbol{x}(0) \in \widehat{\mathcal{X}}$ be fixed. For each potential transition type $\alpha \in \mathcal{T}$, there is an independent unit-rate Poisson process $\Pi_\alpha(\cdot)$. The counting process of potential α-transitions is

$$H_\alpha(t) = \Pi_\alpha\left(\int_0^t \beta_\alpha I\{\alpha \in \mathcal{T}(\boldsymbol{x}(\xi))\}d\xi\right).$$

Denote by $\tau_\alpha(t)$ the set of time points at which the potential transitions α and the associated jumps of $H_\alpha(\cdot)$ occur in $[0,t]$. Denote by $\tau(t) = \cup_{\alpha\in\mathcal{T}}\tau_\alpha(t)$ and $\widehat{\tau}(t) = \cup_{\alpha\in\widehat{\mathcal{T}}}\tau_\alpha(t)$ the sets of all potential transition points and allowed (i.e., actual) transition points, respectively, in $[0,t]$. The state $\boldsymbol{x}(t)$ at time t is the state after the last allowed potential transition. More formally, if $\max\widehat{\tau}(t) \in \tau_\alpha(t)$, then $\boldsymbol{x}(t) = \boldsymbol{x}_\alpha$, where $\boldsymbol{x}_\alpha$ is the to state of potential transition α; if $\widehat{\tau}(t)$ is empty, then $\boldsymbol{x}(t) = \boldsymbol{x}(0)$. It is easy to see that, w.p.1, this construction uniquely determines the realization of the process, including all potential transition points.

Consider the following pure-jump process:

$$F(t) = \sum_{\alpha\in\mathcal{T}}\sum_{s\in\tau_\alpha(t)}\Delta_\alpha \equiv \sum_{\alpha\in\mathcal{T}}\Delta_\alpha H_\alpha(t),$$

where

$$\Delta_\alpha = L(\boldsymbol{x}_\alpha) - L(\boldsymbol{x}),$$

and $\boldsymbol{x}$ and $\boldsymbol{x}_\alpha$ are the from and to states of potential transition α. Note that this definition of $F(t)$ has the sum over *all* potential transitions, both allowed and not allowed. We observe for future reference that, if the artificial process realization is such that all potential transitions in $[0,t]$ are allowed, that is, they are actual transitions, then $F(t) = L(\boldsymbol{x}(t)) - L(\boldsymbol{x}(0))$.

Denote

$$AL(\boldsymbol{x}) = \sum_{\alpha\in\mathcal{T}(\boldsymbol{x})}\beta_\alpha\Delta_\alpha \equiv \sum_{\alpha\in\mathcal{T}(\boldsymbol{x})}\beta_\alpha[L(\boldsymbol{x}_\alpha) - L(\boldsymbol{x})].$$

Let us remark that, for a process in which all potential transitions are allowed, AL is the process generator, if L is within its domain. For our artificial process $\boldsymbol{x}(\cdot)$, however, $AL(\boldsymbol{x})$ is just a formally defined expression.

We have

$$F(t) = M(t) + B(t), \tag{46}$$

where, as we see shortly,

$$B(t) = \sum_{\alpha\in\mathcal{T}} \Delta_\alpha \int_0^t \beta_\alpha I\{\alpha \in \mathcal{T}(\boldsymbol{x}(s))\}ds = \int_0^t AL(\boldsymbol{x}(t))dt$$

is the compensator of $F(t)$ and $M(t) = F(t) - B(t)$ is martingale. We use the following additional notation:

$$B(t) = \int_0^t AL(\boldsymbol{x}(t))dt = \sum_{\alpha\in\mathcal{T}} B_\alpha(t), \tag{47}$$

where

$$B_\alpha(t) = \Delta_\alpha \bar{H}_\alpha(t), \quad \bar{H}_\alpha(t) \equiv \int_0^t \beta_\alpha I\{\alpha \in \mathcal{T}(\boldsymbol{x}(s))\}ds; \tag{48}$$

$$M(t) = \sum_{\alpha\in\mathcal{T}} M_\alpha(t), \tag{49}$$

where

$$M_\alpha(t) = \Delta_\alpha\left[\Pi_\alpha\left(\bar{H}_\alpha(t)\right) - \bar{H}_\alpha(t)\right]. \tag{50}$$

Note also that

$$H_\alpha(t) = \Pi_\alpha\left(\bar{H}_\alpha(t)\right).$$

We have (from, e.g., lemmas 3.1 and 3.2 in Pang et al. 2007) that each $M_\alpha(t)$, $t \geq 0$, is a martingale with respect to (w.r.t.) the filtration describing the history of the process up to time t. (This itself is a martingale; there is no need to localize, as in Pang et al. (2007), because, in our case, the total transition rate and all jump sizes are uniformly bounded.) Moreover, its predictable and optional quadratic variations are, respectively,

$$\langle M_\alpha\rangle(t) = \Delta_\alpha^2 \bar{H}_\alpha(t), \quad [M_\alpha](t) = \Delta_\alpha^2 H_\alpha(t).$$

We see that $M(\cdot)$ is also a martingale. Furthermore, because w.p.1 all potential transition time points are distinct, we have

$$[M](t) = \sum_{\alpha\in\mathcal{T}} [M_\alpha](t) = \sum_{\alpha\in\mathcal{T}} \Delta_\alpha^2 H_\alpha(t).$$

Using this fact and the fact that each $M_\alpha(\cdot)$ is a martingale, the same argument as in the proof of lemma 3.1 in Pang et al. (2007) shows that

$$\langle M\rangle(t) = \sum_{\alpha\in\mathcal{T}} \Delta_\alpha^2 \bar{H}_\alpha(t).$$

We have the following deterministic upper bound on the predictable quadratic variation (quadratic characteristic):

$$\langle M\rangle(t) \leq [C_g r]\left(\frac{C_g}{r}\right)^2 t, \tag{51}$$

where $C_g r$ is a uniform upper bound on the total rate of potential transitions from any state, and C_g/r is a uniform upper bound on Δ_α.

By Doob's inequality (see, e.g., theorem 1.9.1 in Liptser and Shiryaev 1989), for any $\delta > 0$,

$$\mathbb{P}\left\{\max_{0\leq\xi\leq t} |M(\xi)| \geq \delta\right\} \leq \frac{\mathbb{E}[\langle M\rangle(t)]}{\delta^2}. \tag{52}$$

In particular, for $\delta = \eta r^{3s-3-\varepsilon}$ with some $\eta \in (0, 1/4)$, and $t = T = r^{s-1}$, we have

$$\mathbb{P}\left\{\max_{0\leq\xi\leq T} |M(\xi)| \geq \eta r^{3s-3-\varepsilon}\right\} \leq \frac{[C_g\lambda r](C_g/r)^2 r^{s-1}}{[\eta r^{3s-3-\varepsilon}]^2} \leq C_g \frac{r^{4-5s+2\varepsilon}}{\eta^2}. \tag{53}$$

Our next goal is to estimate $|AL(\boldsymbol{x}) - \Xi(\boldsymbol{x})|$, where $\Xi(\boldsymbol{x})$ is defined in (28) and (33). Denote by $\bar{A}L(\boldsymbol{x})$ the "linear component" of $AL(\boldsymbol{x})$, that is, the $AL(\boldsymbol{x})$ computed with function L replaced by its linearization at $\boldsymbol{x}$. Specifically,

$$\bar{A}L(\boldsymbol{x}) = \sum_{\alpha \in \mathcal{T}(\boldsymbol{x})} \beta_\alpha (\boldsymbol{x}_\alpha - \boldsymbol{x}) \cdot \nabla L(\boldsymbol{x}),$$

where $\nabla L(\boldsymbol{x})$ is the gradient of L at $\boldsymbol{x}$. Because the difference between $L(\boldsymbol{x}_\alpha) - L(\boldsymbol{x})$ and $(\boldsymbol{x}_\alpha - \boldsymbol{x}) \cdot \nabla L(\boldsymbol{x})$ is second order, using (45), we have

$$|AL(\boldsymbol{x}) - \bar{A}L(\boldsymbol{x})| \le C_g r \cdot \frac{r^{1-s}}{\log r} \cdot \left(\frac{1}{r}\right)^2 < r^{-s}, \tag{54}$$

where the second inequality is true for $r > \exp(C_g)$.

Let us now consider the relation between $\bar{A}L(\boldsymbol{x})$ and $\Xi(\boldsymbol{x})$. The expression for $\bar{A}L(\boldsymbol{x})$ can be rewritten as

$$\bar{A}L(\boldsymbol{x}) = \sum_{(\boldsymbol{k},i) \in \mathcal{M}} \left[\frac{\partial}{\partial x_{\boldsymbol{k}}} L^{(a)}(\boldsymbol{x}) - \frac{\partial}{\partial x_{\boldsymbol{k}-\boldsymbol{e}_i}} L^{(a)}(\boldsymbol{x})\right] \left[v_{\boldsymbol{k},i} - \tilde{w}_{\boldsymbol{k},i}\right], \tag{55}$$

where $v_{\boldsymbol{k},i}$ are obtained from making the following modifications to $\tilde{v}_{\boldsymbol{k},i}$ in (33) and (30). For each i, replace $\tilde{\lambda}_i$ by λ_i (because λ_i are the *actual* fluid-scaled arrival rates), which changes the expression of $\tilde{v}_{\boldsymbol{k},i}$ in (30) accordingly. Next, $x_{(i)}$ is defined as in (23), but $x_{\boldsymbol{0}} = a$ is replaced by the actual fluid-scaled number of zero servers under GRAND(Z^p), that is, $x_{\boldsymbol{0}} = Z^p/r$, which further changes the value of $\tilde{v}_{\boldsymbol{k},i}$ in (30). We keep the convention $(\partial/\partial x_{\boldsymbol{0}})L(\boldsymbol{x}) = 0$ for any $\boldsymbol{x}$. This completes the description of the modifications. We see that $\bar{A}L(\boldsymbol{x})$ is exactly equal to $\Xi(\boldsymbol{x})$ with each $\tilde{v}_{\boldsymbol{k},i}$ replaced by the corresponding $v_{\boldsymbol{k},i}$.

We can now estimate $|\bar{A}L(\boldsymbol{x}) - \Xi(\boldsymbol{x})|$. We know from (42) that

$$\left|\tilde{\lambda}_i - \lambda_i\right| \le r^{-1/2+\varepsilon}.$$

Furthermore, from (41), we have

$$\left|\frac{Z^p}{r^p} - 1\right| \le 2r^{-1/2+\varepsilon}.$$

Therefore, the modified (and *actual*) $x_{(i)}$ with $x_{\boldsymbol{0}} = Z^p/r$ and the value of $x_{(i)}$ in (23), assuming $x_{\boldsymbol{0}} = a = r^{p-1}$, are different by at most $C_g r^{p-1} r^{-1/2+\varepsilon}$. Summarizing these observations, we conclude that vectors $\tilde{\boldsymbol{v}} = \{\tilde{v}_{\boldsymbol{k},i}\}$ and $\boldsymbol{v} = \{v_{\boldsymbol{k},i}\}$ must be close, specifically,

$$\|\tilde{\boldsymbol{v}} - \boldsymbol{v}\| \le C_g r^{-1/2+\varepsilon}. \tag{56}$$

Using the expressions (33) for $\Xi(\boldsymbol{x})$ and (55) for $\bar{A}L(\boldsymbol{x})$, we have

$$\begin{aligned} |\Xi(\boldsymbol{x}) - \bar{A}L(\boldsymbol{x})| &= \left| \sum_{(\boldsymbol{k},i) \in \mathcal{M}} \left[\frac{\partial}{\partial x_{\boldsymbol{k}}} L^{(a)}(\boldsymbol{x}) - \frac{\partial}{\partial x_{\boldsymbol{k}-\boldsymbol{e}_i}} L^{(a)}(\boldsymbol{x})\right] \left[v_{\boldsymbol{k},i} - \tilde{v}_{\boldsymbol{k},i}\right] \right| \\ &\le C_g \|\nabla L^{(a)}(\boldsymbol{x})\| \, \|\tilde{\boldsymbol{v}} - \boldsymbol{v}\| \\ &\le C_g r^{-1/2+\varepsilon}, \end{aligned} \tag{57}$$

where the last inequality follows from (44) and (56).

Summing (54) and (57), recalling that $s > 7/8$, we obtain the estimate

$$|AL(\boldsymbol{x}) - \Xi(\boldsymbol{x})| \le |AL(\boldsymbol{x}) - \bar{A}L(\boldsymbol{x})| + |\bar{A}L(\boldsymbol{x}) - \Xi(\boldsymbol{x})| \le C_g r^{-1/2+\varepsilon}.$$

Then,

$$\int_0^T |AL(\boldsymbol{x}(t)) - \Xi(\boldsymbol{x}(t))| dt \le C_g r^{s-3/2+\varepsilon}. \tag{58}$$

Now, we specify the choice of the constant $\varepsilon > 0$: it has to be small enough to satisfy

$$3s - 3 - \varepsilon > s - 3/2 + \varepsilon \quad (\text{or} \quad \varepsilon < s - 3/4), \text{ and} \tag{59}$$

$$(-3s + 3 + \varepsilon) + (4 - 5s + 2\varepsilon) < 0 \quad (\text{or} \quad \varepsilon < (8s - 7)/3). \tag{60}$$

Because $s > 7/8$, we have both $s - 3/4 > 0$ and $(8s - 7)/3 > 0$. Thus, conditions (59) and (60) are well defined. Now, compare $\eta r^{3s-3-\varepsilon}$, the term on the RHS of the event in bracket in (53), and $C_g r^{s-3/2+\varepsilon}$, the term on the RHS of the error estimate in (58). By condition (59), we have $\eta r^{3s-3-\varepsilon} > C_g r^{s-3/2+\varepsilon}$ for all sufficiently large r. Then, from (46), (53), and (58), we obtain

$$\mathbb{P}\left\{\max_{0\le t\le T}\left|F(t) - \int_0^t \Xi(\boldsymbol{x}(t))dt\right| \ge 2\eta r^{3s-3-\varepsilon}\right\} \le C_g r^{4-5s+2\varepsilon}/\eta^2. \tag{61}$$

Let us remark that to obtain (61), we only need ε to be positive and satisfy condition (59). Condition (60) is needed for the rest of the proof of Theorem 1 in the next section.

We conclude this section with the estimates of $\Xi(\boldsymbol{x})$ and $\int_0^T \Xi(\boldsymbol{x}(\xi))d\xi$. By the remark after (28), we have $\Xi(\boldsymbol{x}) \le 0$. Furthermore, let any $\delta_1 > 0$ be fixed. Then, by (27) and (33), there exists $C = C(\delta_1) > 0$ such that the following holds: if $|\chi_{k,k',i}| \ge \delta_1$ for at least one pair of edges, then

$$\Xi(\boldsymbol{x}) \le -C(1/\log r)\left[r^{s-1}\right]^2 < -r^{2s-2-\varepsilon},$$

for all sufficiently large r. We obtain that, always,

$$\int_0^t \Xi(\boldsymbol{x}(\xi))d\xi \le 0, \quad \forall t \le T, \tag{62}$$

and if $|\chi_{k,k',i}| \ge \delta_1 > 0$ for at least one pair of edges in the entire interval $[0, T]$, then

$$\int_0^T \Xi(\boldsymbol{x}(t))dt \le -C(1/\log r)\left[r^{s-1}\right]^2 r^{s-1} < -r^{3s-3-\varepsilon}, \tag{63}$$

for all sufficiently large r.

5.4. Completing the Proof of Theorem 1

We start with a comment about the probability space constructions. The construction that we use in this section does *not* have to be same as those used in Section 5.3 or Appendices A and B. The probability spaces used in different parts of the paper can be different as long as we use them to obtain statements in the form of probability estimates, which is the case here. Obviously, the probability estimates are valid regardless of the probability space construction used to derive them.

Let the constants $\varepsilon > 0$, $\eta > 0$, and $c > 0$ be those chosen in Section 5.3. More specifically, $\varepsilon > 0$ satisfies conditions (59) and (60), $\eta \in (0, 1/4)$, and $c > 0$ is such that Lemma 6 holds. Fix $C' > 0$. (The exact choice of C' is specified later.) Consider the stationary version of the original process on $C' r^{-3s+3+\varepsilon}$ consecutive r^{s-1}-long intervals starting at time 0 (i.e., the distribution at initial time 0 is the stationary distribution of the process). These r^{s-1}-long intervals we call *subintervals*. The proof of Theorem 1 is completed if we prove the following assertion. For any fixed $\gamma > 0$ and any fixed subsequence of r, there exists a further subsequence of r along which, w.p.1, for all sufficiently large r, the following properties hold for the original process:

a. Properties (41)–(43) hold at all times within all $C' r^{-3s+3+\varepsilon}$ subintervals.

b. At the end of the last subinterval, $\sum_{k\in\mathcal{K}} x_k - L^* < \gamma$.

To prove this assertion, fix any $\gamma > 0$ and any subsequence of r. In addition to the original process, consider the artificial process coupled to it as follows. The initial state of the artificial process is equal (w.p.1) to that of the original one. If the initial state (of both processes) satisfies (41)–(43), then the artificial process evolves as it is defined, and it is coupled to be equal to the original process until the first time when (41)–(43) is violated (for the original process). By convention, if the initial state (of both processes) violates (41)–(43), then the artificial process is "frozen," that is, remains equal to the initial state at all times.

Let us focus on the artificial process and apply the estimate (61) to each subinterval. (When (61) is applied to a given subinterval, the time is shifted so that $t = 0$ is the beginning of that subinterval.) Then, by (61) and a simple union bound, we see that the probability that the event in the brackets in the left-hand side (LHS) of (61) holds for at least one of the subintervals is upper bounded by

$$C_g C' r^{-3s+3+\varepsilon} r^{4-5s+2\varepsilon}/\eta^2 = C_g C' r^{7-8s+3\varepsilon}/\eta^2.$$

Because ε satisfies (60), $C_g C' r^{7-8s+3\varepsilon}/\eta^2 \to 0$ as $r \to \infty$. Consider a further subsequence of r, increasing fast enough, for example, $r = r(n) \geq e^n$, so that the sum of these probabilities is finite. Then, for the artificial process, by the Borel–Cantelli lemma w.p.1 for all large r, the condition

$$\max_{0\le t\le T}\left|F(t) - \int_0^t \Xi(\boldsymbol{x}(t))dt\right| < 2\eta r^{3s-3-\varepsilon} \tag{64}$$

holds simultaneously for all $C' r^{-3s+3+\varepsilon}$ subintervals; furthermore, we have (62) and (63) for all these subintervals simultaneously.

By Lemmas 5 and 6 and the Borel–Cantelli lemma, we can choose a further subsequence of r along which the original process is such that, w.p.1, for all large r, conditions (41)–(43) hold on all $C' r^{-3s+3+\varepsilon}$ subintervals, and therefore, the artificial and original processes coincide. Therefore, along this subsequence w.p.1 for all large r, $F(t)$ (which we defined for the artificial process) is equal to the increment of L, $F(t) = L(\boldsymbol{x}(t)) - L(\boldsymbol{x}(0))$ for the original process for all subintervals simultaneously, and we also have (62) and (63) for all subintervals simultaneously.

Then, w.p.1 for all large r, the following occurs for the original process. If, at the beginning of a subinterval, $\Delta L \equiv L - L^* \geq \gamma$, then either condition $\Delta L \leq \gamma$ is "hit" within the subinterval, or at the end of the subinterval ΔL is smaller by at least $r^{3s-3-\varepsilon}/2 > 0$. This follows from (63), (64), condition $\eta < 1/4$, and the fact that (by Lemma 4) $\Delta L \geq \gamma$ implies that $|\chi_{k,k',i}| \geq \delta_1 > 0$ for some $\delta_1 = \delta_1(\gamma)$ that (just like γ) does not depend on r. In addition, if, at the beginning of a subinterval or any other point in it, $\Delta L \leq \gamma$, then, at the end of the subinterval, $\Delta L < 2\gamma$.

Note that, w.p.1 for all large r, at the beginning of the first subinterval, $L(\boldsymbol{x}(0)) - L^* \leq C''$ for some constant $C'' > 0$ independent of r. (Recall that $L(\boldsymbol{x}) = L^{(r^{p-1})}(\boldsymbol{x})$, and it converges to $\sum_{k\in\mathcal{K}} x_k$ uniformly on compact sets.) We now specify the choice of C': it is any constant satisfying $C' > 2C''$. Given this choice, we see that (w.p.1 for all large r) condition $\Delta L \leq \gamma$ is, in fact, hit within one of the subintervals; this, in turn, implies that $\Delta L < 2\gamma$ must hold at the end of the last subinterval. Recall again that $L(\boldsymbol{x}) = L^{(r^{p-1})}(\boldsymbol{x})$ converges to $\sum_{k\in\mathcal{K}} x_k$ uniformly on compact sets. We finally obtain that, w.p.1 for all large r, $\sum_{k\in\mathcal{K}} x_k - L^* < 3\gamma$ at the end of the last subinterval, which (by rechoosing γ) completes the proof of the assertion and of Theorem 1. □

6. Discussion

This paper continues the line of work originated in Stolyar and Zhong (2015) that shows, surprisingly, extremely simple dynamic placement (packing) algorithms, such as GRAND, can be asymptotically optimal. We show that the GRAND(Z^p) algorithm, which is as simple as GRAND(aZ) in Stolyar and Zhong (2015), is asymptotically optimal in a stronger sense than GRAND(aZ). The analysis of GRAND(Z^p) is substantially more involved technically than that of GRAND(aZ) because it cannot be reduced to the analysis of fluid limits and/or local fluid limits.

We believe that our main result, Theorem 1, holds for any $p \in (0, 1)$ without the additional condition (9) that requires p to be *sufficiently close* to one. This additional condition is needed for our technical approach to work and is probably just technical. Removing or relaxing the additional condition on p is interesting and important from both practical and methodological points of view and may be a subject of future work.

Appendix A. Proof of Lemma 5

We start with a general outline. First, in Section A.1, we provide some useful concentration inequalities of Poisson random variables together with some notation and terminology that are useful for the proofs of both Lemmas 5 and 6. The actual proof of Lemma 5 is in Section A.2; its outline is as follows. First, in steady state, property (42) holds with high probability (w.h.p.) for each t because $Y_i(t) = \sum_k k_i X_k(t)$ is a Poisson random variable with mean $\lambda_i r$. To show that, w.h.p., property (42) holds uniformly over a polynomially long interval $[0, r^\alpha]$, we divide the interval $[0, r^\alpha]$ into shorter subintervals of length $r^{-1/2}$ and show that, w.h.p., property (42) holds uniformly over each subinterval. In fact, the probability of (42) *not* holding uniformly over a subinterval decays fast enough with r so that we can use a union bound to claim that, w.h.p., (42) holds for all subintervals simultaneously. To obtain a single subinterval bound, we first observe that, at the beginning of a subinterval, property (42) holds w.h.p. by stationarity. Then, using the fact that $Y_i(\cdot)$ follows the evolution of an $M/M/\infty$ queue, we can upper bound the increase of Y_i resulting from arrivals and lower bound the decrease of Y_i resulting from departures over a subinterval. We have chosen the length $r^{-1/2}$ of each subinterval small enough so that increases and decreases in Y_i are small as well, and (by possibly rechoosing ε) property (42) would hold uniformly in a subinterval w.h.p.

A.1. Preliminaries

In this section, we recall some basic concentration inequalities for Poisson random variables. We then describe a convention that we follow in this section and Appendix B.

Proposition A.1 (Theorem 5.4, Mitzenmacher and Upfal 2005). *Let V be a Poisson random variable with mean ν. If $x > \nu$, then*

$$\mathbb{P}(V \geq x) \leq \frac{e^{-\nu}(e\nu)^x}{x^x}; \tag{A.1}$$

if $x < \nu$, then

$$\mathbb{P}(V \leq x) \leq \frac{e^{-\nu}(e\nu)^x}{x^x}. \tag{A.2}$$

A consequence of Proposition A.1 is the following concentration bounds, which are used extensively for the proofs of Lemmas 5 and 6.

Corollary A.1. *Let V be a Poisson random variable with mean ν and let $w \in [0, \nu]$. Then,*

$$\mathbb{P}(V \geq \nu + w) \leq e^{-\frac{w^2}{4\nu}}, \text{ and} \tag{A.3}$$

$$\mathbb{P}(V \leq \nu - w) \leq e^{-\frac{w^2}{4\nu}}. \tag{A.4}$$

Proof sketch. Let $w \in [0, \nu]$. Then, by (A.1),

$$\mathbb{P}(V \geq \nu + w) \leq \frac{e^{-\nu}(e\nu)^{\nu+w}}{(\nu+w)^{\nu+w}} = \exp\left[w + (\nu+w)\log\frac{\nu}{\nu+w}\right].$$

To establish (A.3), it suffices to prove that, for all $w \in [0, \nu]$,

$$w + (\nu+w)\log\frac{\nu}{\nu+w} \leq -\frac{w^2}{4\nu}.$$

The inequality can be established by observing that, when $w = 0$, LHS = RHS = 0 and that, for $w \in [0, \nu]$,

$$\frac{d}{dw}\left[w + (\nu+w)\log\frac{\nu}{\nu+w}\right] = \log\frac{\nu}{\nu+w} \leq -\frac{w}{\nu+w} \leq -\frac{w}{2\nu} = \frac{d}{dw}\left[-\frac{w^2}{4\nu}\right].$$

Equation (A.4) can be established in a similar way. We omit further details.

To prove Lemmas 5 and 6, we often consider probability bounds for a sequence of events, indexed by r. To simplify exposition, we use the following notation. For a sequence of constants $\{C^{(r)}\}$ with $C^{(r)} \in [0, 1]$ for each r, we write

$$C^{(r)} \leq e^{-\text{poly}(r)} \left(\text{respectively}, C^{(r)} \geq 1 - e^{-\text{poly}(r)}\right), \tag{A.5}$$

if there exists a positive constant γ such that, for all sufficiently large r,

$$C^{(r)} \leq e^{-r^{\gamma}} \left(\text{respectively}, C^{(r)} \geq 1 - e^{-r^{\gamma}}\right).$$

It is useful to think of $C^{(r)}$ as probabilities of events indexed by r. Note that, if $\{C_1^{(r)}\}_r$ and $\{C_2^{(r)}\}_r$ are two sequences with $C_1^{(r)} \leq e^{-\text{poly}(r)}$ and $C_2^{(r)} \leq e^{-\text{poly}(r)}$, then we also have

$$C_1^{(r)} + C_2^{(r)} \leq e^{-\text{poly}(r)}.$$

Similarly, if $C_1^{(r)} \geq 1 - e^{-\text{poly}(r)}$ and $C_2^{(r)} \geq 1 - e^{-\text{poly}(r)}$, then

$$C_1^{(r)} + C_2^{(r)} \geq 1 - e^{-\text{poly}(r)}.$$

Finally, we also often use the expression "with probability $1 - e^{-\text{poly}(r)}$" to mean that the probability is at least $1 - e^{-r^{\gamma}}$ for some $\gamma > 0$ and for all sufficiently large r.

A.2. Proof of Lemma 5

Our general strategy for establishing Lemma 5 is to divide the interval $[0, r^{\alpha}]$ into shorter subintervals of length $r^{-1/2}$ and show that (42) holds with high probability over each subinterval. More specifically, let us define events

$$E_{j,i}^r = \left\{ \sup_{t \in \left[(j-1)r^{-\frac{1}{2}}, jr^{-\frac{1}{2}}\right]} |Y_i^r(t) - \rho_i r| \geq \frac{r^{1/2+\varepsilon}}{I} \right\}, \tag{A.6}$$

for $i \in \mathcal{I}$, $j \in \{1, 2, \ldots, r^{\alpha+1/2}\}$,[1] and for each r. Note that, because, for each r, the system is in the stationary regime, we have that, for any $i \in \mathcal{I}$,

$$\mathbb{P}(E_{1,i}^r) = \mathbb{P}(E_{2,i}^r) = \cdots = \mathbb{P}\left(E_{r^{\alpha+1/2},i}^r\right). \tag{A.7}$$

Thus, if we can show that

$$\mathbb{P}\big(E_{1,i}^r\big) \le e^{-\mathrm{poly}(r)}, \forall i \in \mathcal{I}, \tag{A.8}$$

then, by (A.7) and (A.8),

$$\mathbb{P}\Big(\cup_{1\le j\le r^{\alpha+1/2}, i\in\mathcal{I}} E_{j,i}^r\Big) \le e^{-\mathrm{poly}(r)} \to 0 \quad \text{as} \quad r \to \infty,$$

which establishes Lemma 5 immediately.

It is now left to show that $\mathbb{P}(E_{1,i}^r) \le e^{-\mathrm{poly}(r)}$ for each $i \in \mathcal{I}$.

Lemma A.1. *Let $i \in \mathcal{I}$, and consider the system indexed by r in the stationary regime. Then, for sufficiently large r,*

$$\mathbb{P}\big(E_{1,i}^r\big) = \mathbb{P}\left(\sup_{t\in[0,r^{-1/2}]} |Y_i^r(t) - \rho_i r| \ge \frac{r^{1/2+\varepsilon}}{I}\right) \le e^{-poly(r)}. \tag{A.9}$$

Proof. For notational convenience, we drop the subscript i. The proof of Lemma A.1 consists of establishing (a) a probability tail bound of $|Y^r(0) - \rho r|$, (b) a probability tail bound of $\sup_{t\in[0,r^{-1/2}]}(Y^r(t) - \rho r)$, and (c) a probability tail bound of $\inf_{t\in[0,r^{-1/2}]}(Y^r(t) - \rho r)$.

a. We first show that

$$\mathbb{P}\big(|Y^r(0) - \rho r| \ge r^{1/2+\varepsilon/3}\big) \le e^{-\mathrm{poly}(r)}. \tag{A.10}$$

Because $Y^r(0)$ is a Poisson random variable with mean ρr, we apply the concentration bounds (A.3) and (A.4). First by (A.3), we have that, for sufficiently large r, $r^{1/2+\varepsilon/3} \le \rho r$, and

$$\mathbb{P}\big(Y^r(0) - \rho r \ge r^{1/2+\varepsilon/3}\big) \le \exp\left(-\frac{r^{1+2\varepsilon/3}}{4\rho r}\right) = \exp\left(-\frac{r^{2\varepsilon/3}}{4\rho}\right).$$

Similarly, for sufficiently large r,

$$\mathbb{P}\big(Y^r(0) - \rho r \le -r^{1/2+\varepsilon/3}\big) \le \exp\left(-\frac{r^{2\varepsilon/3}}{4\rho}\right).$$

Thus, by a simple union bound, for sufficiently large r,

$$\mathbb{P}\big(|Y^r(0) - \rho r| \ge r^{1/2+\varepsilon/3}\big) \le 2\exp\left(-\frac{r^{2\varepsilon/3}}{4\rho}\right) \le e^{-\mathrm{poly}(r)}. \tag{A.11}$$

This completes part (a).

b. We then show that

$$\mathbb{P}\left(\sup_{t\in[0,r^{-1/2}]} \big(Y^r(t) - \rho r\big) \ge r^{1/2+2\varepsilon/3}\right) \le e^{-\mathrm{poly}(r)}. \tag{A.12}$$

To establish (A.12), we use the following representation of the process $Y^r(\cdot)$. Note that the process $Y^r(\cdot)$ describes the steady-state evolution of a $M/M/\infty$ queueing system, so we can represent $Y^r(\cdot)$ as (see, e.g., Pang et al. 2007)

$$Y^r(t) = Y^r(0) + \Pi(\lambda r t) - \widetilde{\Pi}\left(\int_0^t \mu Y^r(s)ds\right), \tag{A.13}$$

where $Y^r(0)$ is a Poisson random variable with mean ρr and $\Pi(\cdot)$ and $\widetilde{\Pi}(\cdot)$ are independent unit-rate Poisson processes that are also independent from $Y^r(0)$.

By (A.13), we have that, w.p.1, for all $t \in [0, r^{-1/2}]$,

$$Y^r(t) \le Y^r(0) + \Pi(\lambda r t) \le Y^r(0) + \Pi\big(\lambda r \cdot r^{-1/2}\big) = Y^r(0) + \Pi\big(\lambda r^{1/2}\big).$$

Using (A.3) and the fact that $\Pi(\lambda r^{1/2})$ is Poisson with mean $\lambda r^{1/2}$, we have that

$$\mathbb{P}\big(\Pi\big(\lambda r^{1/2}\big) \ge 2\lambda r^{1/2}\big) \le \exp\left(-\frac{\lambda r^{1/2}}{4}\right) \le e^{-\mathrm{poly}(r)}. \tag{A.14}$$

For sufficiently large r, if $Y^r(0)+\Pi(\lambda r^{1/2}) \geq \rho_i r + r^{1/2+2\varepsilon/3}$, then we must have $Y^r(0) \geq \rho_i r + r^{1/2+\varepsilon/3}$ or $\Pi(\lambda r^{1/2}) \geq 2\lambda r^{1/2}$. Thus,

$$\begin{aligned}
&\mathbb{P}\big(Y^r(0)+\Pi\big(\lambda r^{1/2}\big) \geq \rho_i r + r^{1/2+2\varepsilon/3}\big)\\
&\quad\leq \mathbb{P}\big(Y^r(0) \geq \rho_i r + r^{1/2+\varepsilon/3} \text{ or } \Pi\big(\lambda r^{1/2}\big) \geq 2\lambda r^{1/2}\big)\\
&\quad\leq \mathbb{P}\big(Y^r(0) \geq \rho_i r + r^{1/2+\varepsilon/3}\big) + \mathbb{P}\big(\Pi\big(\lambda r^{1/2}\big) \geq 2\lambda r^{1/2}\big)\\
&\quad\leq e^{-\mathrm{poly}(r)},
\end{aligned}$$

where the last inequality follows from (A.10) and (A.14).

Because, with probability 1, for all $t \in [0, r^{-1/2}]$,

$$Y^r(t) \leq Y^r(0) + \Pi\big(\lambda r^{1/2}\big),$$

it immediately follows that (A.12) holds. This completes part (b).

c. We now show that

$$\mathbb{P}\left(\inf_{t\in[0,r^{-1/2}]}\big(Y^r(t)-\rho r\big) \leq -r^{1/2+2\varepsilon/3}\right) \leq e^{-\mathrm{poly}(r)}. \tag{A.15}$$

Using representation (A.13), we have that, with probability 1, for all $t \in [0, r^{-1/2}]$,

$$Y^r(t) \geq Y^r(0) - \widetilde{\Pi}\left(\int_0^t \mu Y^r(s)ds\right) \geq Y^r(0) - \widetilde{\Pi}\left(\int_0^{r^{-1/2}} \mu Y^r(s)ds\right).$$

Consider events F^r and G^r defined to be

$$F^r = \left\{\sup_{t\in[0,r^{-1/2}]} Y^r(t) < 2\rho r\right\}, \quad \text{and} \quad G^r = \left\{\widetilde{\Pi}\big(r^{1/2+\varepsilon/3}\big) < 2r^{1/2+\varepsilon/3}\right\}.$$

For sufficiently large r, we have that, under event F^r,

$$\int_0^{r^{-1/2}} \mu Y^r(s)ds < \mu r^{-1/2}\cdot\big(2\rho r\big) \leq r^{1/2+\varepsilon/3}.$$

Thus, for sufficiently large r, we have that, under the event $F^r \cap G^r$,

$$\widetilde{\Pi}\left(\int_0^{r^{-1/2}} \mu Y^r(s)ds\right) \leq \widetilde{\Pi}\big(r^{1/2+\varepsilon/3}\big) < 2r^{1/2+\varepsilon/3}.$$

By (A.12), we know that $\mathbb{P}(F^r) \geq 1 - e^{-\mathrm{poly}(r)}$, and using (A.3), we can easily show that $\mathbb{P}(G^r) \geq 1 - e^{-\mathrm{poly}(r)}$. Thus, we have $\mathbb{P}(F^r \cap G^r) \geq 1 - e^{-\mathrm{poly}(r)}$, and by considering the complement, we have

$$\mathbb{P}\left(\widetilde{\Pi}\left(\int_0^{r^{-1/2}} \mu Y^r(s)ds\right) \geq 2r^{1/2+\varepsilon/3}\right) \leq e^{-\mathrm{poly}(r)}. \tag{A.16}$$

For sufficiently large r, if $Y^r(0) - \widetilde{\Pi}(\int_0^{r^{-1/2}} \mu Y^r(s)ds) \leq \rho_i r - r^{1/2+2\varepsilon/3}$, then we have $Y^r(0) \leq \rho_i r - r^{1/2+\varepsilon/3}$ or $\widetilde{\Pi}(\int_0^{r^{-1/2}} \mu Y^r(s)ds) \geq 2r^{1/2+\varepsilon/3}$. Thus, similar to the argument in part (b), we have that

$$\begin{aligned}
&\mathbb{P}\left(Y^r(0) - \widetilde{\Pi}\left(\int_0^{r^{-1/2}} \mu Y^r(s)ds\right) \leq \rho_i r - r^{1/2+2\varepsilon/3}\right)\\
&\quad\leq \mathbb{P}\big(Y^r(0) \leq \rho_i r - r^{1/2+\varepsilon/3}\big) + \mathbb{P}\left(\widetilde{\Pi}\left(\int_0^{r^{-1/2}} \mu Y^r(s)ds\right) \geq 2r^{1/2+\varepsilon/3}\right)\\
&\quad\leq e^{-\mathrm{poly}(r)},
\end{aligned}$$

where the last inequality follows from (A.10) and (A.16). This establishes (A.15), and completes part (c).

Combining (A.10), (A.12), and (A.15), we have

$$\mathbb{P}\left(\sup_{t\in[0,r^{-1/2}]} |Y^r(t) - \rho r| \geq r^{1/2+2\varepsilon/3}\right) \leq e^{-\mathrm{poly}(r)}.$$

Because, for sufficiently large r, $\frac{r^{1/2+\varepsilon}}{I} \geq r^{1/2+2\varepsilon/3}$, we also have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{t\in[0,r^{-1/2}]} |Y^r(t) - \rho r| \geq \frac{r^{1/2+\varepsilon}}{I}\right) \\
&\quad \leq \mathbb{P}\left(\sup_{t\in[0,r^{-1/2}]} |Y^r(t) - \rho r| \geq r^{1/2+2\varepsilon/3}\right) \\
&\quad \leq e^{-\mathrm{poly}(r)}.
\end{aligned}$$

This establishes Lemma A.1. □

Appendix B. Proof of Lemma 6

We already gave an informal sketch of this proof immediately after stating Lemma 6 at the end of Section 5.2.

For each $\boldsymbol{k} \in \bar{\mathcal{K}}$, define constant $s(\boldsymbol{k}) = 1 - (1 + \sum_{i\in\mathcal{I}} k_i)(1-p)$.

To establish Lemma 6, we consider instead the condition

$$X^r_{\boldsymbol{k}}(t) \geq cr^{s(\boldsymbol{k})} \tag{B.1}$$

for each $\boldsymbol{k} \in \bar{\mathcal{K}}$ at any time $t \geq 0$ and prove the following stronger result.

Lemma B.1. *Let $\alpha > 0$. Consider our sequence of systems in the stationary regime under the GRAND(Z^p) algorithm. Then, for each $\boldsymbol{k} \in \bar{\mathcal{K}}$, there exists a constant $c > 0$ such that, as $r \to \infty$,*

$$\mathbb{P}(\text{Condition (B.1) holds for all } t \in [0, r^\alpha]) \geq 1 - e^{-\mathrm{poly}(r)}. \tag{B.2}$$

To prove Lemma B.1, we use the following construction of the underlying probability space. For each $(\boldsymbol{k}, i) \in \mathcal{M}$, let $\Pi_{(\boldsymbol{k},i)}(\cdot)$ and $\widetilde{\Pi}_{(\boldsymbol{k},i)}(\cdot)$ be unit-rate Poisson processes. Furthermore, all $\Pi_{(\boldsymbol{k},i)}(\cdot)$ and $\widetilde{\Pi}_{(\boldsymbol{k},i)}(\cdot)$ are independent. We use $\Pi_{(\boldsymbol{k},i)}(\cdot)$ as the driving processes for customer arrivals and $\widetilde{\Pi}_{(\boldsymbol{k},i)}(\cdot)$ to drive departures. Furthermore, processes $\Pi_{(\boldsymbol{k},i)}(\cdot)$ and $\widetilde{\Pi}_{(\boldsymbol{k},i)}(\cdot)$ are all independent from the initial random system state $\boldsymbol{X}^r(0)$.

More specifically, let $D^r_{(\boldsymbol{k},i)}(t)$ denote the total number of departures along the edge $(\boldsymbol{k}, i)$ in $[0, t]$. Then,

$$D^r_{(\boldsymbol{k},i)}(t) = \widetilde{\Pi}_{(\boldsymbol{k},i)}\left(\int_0^t X^r_{\boldsymbol{k}}(u) k_i \mu_i du\right). \tag{B.3}$$

Similarly, let $A^r_{(\boldsymbol{k},i)}(t)$ denote the total number of arrivals along the edge $(\boldsymbol{k}, i)$ in $[0, t]$. Under the GRAND algorithm, a type i customer that arrives at time s is placed in a server of configuration $\boldsymbol{k} - \boldsymbol{e}_i$ with probability $X^r_{\boldsymbol{k}-\boldsymbol{e}_i}(s)/X^r_{(i)}(s)$. Then, we can write

$$A^r_{(\boldsymbol{k},i)}(t) = \Pi_{(\boldsymbol{k},i)}\left(\int_0^t \lambda_i r \cdot \frac{X^r_{\boldsymbol{k}-\boldsymbol{e}_i}(u)}{X^r_{(i)}(u)} du\right). \tag{B.4}$$

Thus, for each $\boldsymbol{k} \in \mathcal{K}$, we can write

$$X^r_{\boldsymbol{k}}(t) - X^r_{\boldsymbol{k}}(0) \tag{B.5}$$

$$\begin{aligned}
&= \left[\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} A^r_{(\boldsymbol{k},i)}(t) + \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}} D^r_{(\boldsymbol{k}+\boldsymbol{e}_i,i)}(t)\right] \\
&\quad - \left[\sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}} A^r_{(\boldsymbol{k}+\boldsymbol{e}_i,i)}(t) + \sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} D^r_{(\boldsymbol{k},i)}(t)\right] \\
&\geq \left[\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} A^r_{(\boldsymbol{k},i)}(t)\right] - \left[\sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}} A^r_{(\boldsymbol{k}+\boldsymbol{e}_i,i)}(t) + \sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} D^r_{(\boldsymbol{k},i)}(t)\right] \\
&= S_1(t) - S_2(t) - S_3(t),
\end{aligned} \tag{B.6}$$

where

$$S_1(t) = \sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} A^r_{(\boldsymbol{k},i)}(t); \tag{B.7}$$

$$S_2(t) = \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}} A^r_{(\boldsymbol{k}+\boldsymbol{e}_i,i)}(t); \tag{B.8}$$

$$S_3(t) = \sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} D^r_{(\boldsymbol{k},i)}(t). \tag{B.9}$$

Equality (B.5) follows by accounting for all arrivals and departures that contribute to $X_{\boldsymbol{k}}^r(t) - X_{\boldsymbol{k}}^r(0)$, the change in $X_{\boldsymbol{k}}^r$. For example, the term $\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}} A_{(\boldsymbol{k},i)}^r(t)$ accounts for arrivals along the edges $(\boldsymbol{k}, i) \in \mathcal{M}$, which increases the number of servers of configuration $\boldsymbol{k}$.

Proof of Lemma B.1. We prove Lemma B.1 by induction on $\|\boldsymbol{k}\|_1 = \sum_i k_i$.

Base case: $\|\boldsymbol{k}\|_1 = 0$. If $\|\boldsymbol{k}\|_1 = 0$, then $\boldsymbol{k} = \boldsymbol{0}$, and $s(\boldsymbol{0}) = 1 - (1-p) = p$. Then, we show that

$$\mathbb{P}\left(X_{\boldsymbol{0}}^r(t) \geq \frac{1}{2}r^p,\ \forall t \in [0, r^\alpha]\right) \geq 1 - e^{-\text{poly}(r)}.$$

To this end, note that $X_{\boldsymbol{0}}^r(t) = \lceil (Z^r(t))^p \rceil$ for all t. By Lemma 5, we know that, for any $\varepsilon > 0$ as $r \to \infty$,

$$\mathbb{P}\left(|Y_i^r(t) - \rho_i r| \leq \frac{r^{1/2+\varepsilon}}{I},\ \forall t \in [0, r^\alpha], \forall i \in \mathcal{I}\right) \geq 1 - e^{-\text{poly}(r)}.$$

In particular, by setting $\varepsilon = 1/4$, we have

$$\mathbb{P}\left(|Y_i^r(t) - \rho_i r| \leq \frac{r^{3/4}}{I},\ \forall t \in [0, r^\alpha], \forall i \in \mathcal{I}\right) \geq 1 - e^{-\text{poly}(r)}.$$

For sufficiently large r, $r/2^{1/p} \leq r - r^{3/4}$. Because $Z^r(t) = \sum_{i\in\mathcal{I}} Y_i^r(t)$ for all t, we have

$$\begin{aligned}&\mathbb{P}\left(Z^r(t) \geq \frac{r}{2^{1/p}},\ \forall t \in [0, r^\alpha]\right)\\ &\geq \mathbb{P}\left(|Z^r(t) - r| \leq r^{3/4},\ \forall t \in [0, r^\alpha]\right)\\ &\geq \mathbb{P}\left(|Y_i^r(t) - \rho_i r| \leq \frac{r^{3/4}}{I},\ \forall t \in [0, r^\alpha], \forall i \in \mathcal{I}\right) \geq 1 - e^{-\text{poly}(r)}.\end{aligned}$$

This implies that

$$\mathbb{P}\left(X_{\boldsymbol{0}}^r(t) \geq \frac{1}{2}r^p,\ \forall t \in [0, r^\alpha]\right) \geq \mathbb{P}\left(Z^r(t) \geq \frac{r}{2^{1/p}},\ \forall t \in [0, r^\alpha]\right) \geq 1 - e^{-\text{poly}(r)},$$

establishing the base case.

Induction step: Let $\ell > 0$. Suppose that, for all $\boldsymbol{k}'$ with $\|\boldsymbol{k}'\|_1 \leq \ell - 1$, there exists $c' > 0$ such that

$$\mathbb{P}\left(X_{\boldsymbol{k}'}^r(t) \geq c' r^{s(\boldsymbol{k}')},\ \forall t \in [0, r^\alpha]\right) \geq 1 - e^{-\text{poly}(r)}.$$

Let $\boldsymbol{k} \in \mathcal{K}$ be a configuration with $\|\boldsymbol{k}\|_1 = \ell > 0$. Then, there exists $\iota \in \mathcal{I}$ such that $k_\iota \geq 1$. Let $\tilde{\boldsymbol{k}} = \boldsymbol{k} - \boldsymbol{e}_\iota$ so that $\|\tilde{\boldsymbol{k}}\|_1 = \ell - 1$. For notational convenience, write $s_1 = s(\tilde{\boldsymbol{k}})$ and $s_2 = s_1 - (1-p)$ so that $s_2 = s(\boldsymbol{k})$.

By the induction hypothesis, we can assume that

$$\mathbb{P}\left(X_{\tilde{\boldsymbol{k}}}^r(t) \geq \tilde{c} r^{s_1},\ \forall t \in [0, r^\alpha]\right) \geq 1 - e^{-\text{poly}(r)} \tag{B.10}$$

for some $\tilde{c} > 0$. Furthermore, define E^r to be the event

$$E^r = \left\{X_{\tilde{\boldsymbol{k}}}^r(t) \geq \tilde{c} r^{s_1},\ \forall t \in [0, r^\alpha]\right\}. \tag{B.11}$$

Then by (B.10) and (B.11), $\mathbb{P}(E^r) \geq 1 - e^{-\text{poly}(r)}$.

Consider the interval $[0, r^\alpha + 1]$ and divide it into $[r^\alpha + 1]/T$ subintervals of length $T = r^{s_2 - 1 - \varepsilon'}$ (namely subintervals $[0, T], [T, 2T], [2T, 3T], \cdots$), where $0 < \varepsilon' < s_2 - p/2$. We claim the following.

Claim 1: There exist positive constants c and $\bar{c}$ such that, with probability $1 - e^{-\text{poly}(r)}$, for each $j \in \{1, 2, \ldots, [r^\alpha + 1]/T\}$,

i. If $X_{\boldsymbol{k}}^r((j-1)T) \leq 4cr^{s_2}$, then

$$X_{\boldsymbol{k}}^r(jT) - X_{\boldsymbol{k}}^r((j-1)T) \geq \bar{c} r^{2s_2 - p - \varepsilon'};$$

ii. If $X_{\boldsymbol{k}}^r((j-1)T) \geq 2cr^{s_2}$, then

$$\inf_{t\in[(j-1)T, jT]} X_{\boldsymbol{k}}^r(t) \geq \frac{1}{2} X_{\boldsymbol{k}}^r((j-1)T).$$

Assuming the validity of claim 1, we now complete the rest of the induction step and defer the proof of the claim.

Suppose that claim 1 is true. For each r, consider a sample path for which both statements (i) and (ii) of claim 1 hold for all $j \in \{1, 2, \ldots, [r^\alpha + 1]/T\}$ and let j_0 be minimal such that

$$X^r_{\boldsymbol{k}}(j_0 T) \geq 4cr^{s_2}. \tag{B.12}$$

Then, for sufficiently large r, $j_0 \leq 1/T$. If $X^r_{\boldsymbol{k}}(0) \geq 4cr^{s_2}$, then $j_0 = 0 \leq 1/T$. If $X^r_{\boldsymbol{k}}(0) < 4cr^{s_2}$, then we also have $j_0 \leq 1/T$ because, by comparing the threshold $4cr^{s_2}$ and increment $\bar{c}r^{2s_2-p-\varepsilon'}$ from $X^r_{\boldsymbol{k}}(jT)$ to $X^r_{\boldsymbol{k}}((j-1)T)$ in statement (i), we have

$$\frac{4cr^{s_2}}{\bar{c}r^{2s_2-p-\varepsilon'}} = \frac{4c}{\bar{c}} \cdot \frac{1}{r^{s_2-p-\varepsilon'}} = \frac{4c}{\bar{c}} \cdot \frac{r^{p-1}}{T} \leq \frac{1}{T}.$$

Next, claim for any $j \in \{j_0, j_0+1, \ldots, [r^\alpha + 1]/T\}$, we have

$$X^r_{\boldsymbol{k}}(jT) \geq 2cr^{s_2}, \tag{B.13}$$

and we establish this claim by induction (note that this induction is distinct from the induction on $\|\boldsymbol{k}\|$ that we use to prove Lemma B.1). First, (B.13) is true for $j = j_0$ by (B.12). Next, for $j > j_0$, suppose by induction hypothesis that $X^r_{\boldsymbol{k}}((j-1)T) \geq 2cr^{s_2}$. Then, we consider two cases: (a) if $X^r_{\boldsymbol{k}}((j-1)T) \geq 4cr^{s_2}$, then by statement (ii) (B.13) holds for j; (b) if $X^r_{\boldsymbol{k}}((j-1)T) \in [2cr^{s_2}, 4cr^{s_2})$, then by statement (i) (B.13) holds for j as well. This completes the induction step and the proof of (B.13) for all $j \in \{j_0, j_0+1, \ldots, [r^\alpha+1]/T\}$. Now, by statement (ii), for each $j \in \{j_0+1, \ldots, [r^\alpha+1]/T\}$,

$$\inf_{t \in [(j-1)T, jT]} X^r_{\boldsymbol{k}}(t) \geq \frac{1}{2} X^r_{\boldsymbol{k}}((j-1)T) \geq \frac{1}{2} \cdot 2cr^s = cr^s,$$

where the second inequality follows by (B.13). Thus,

$$\begin{aligned} \inf_{t \in [1, r^\alpha+1]} X^r_{\boldsymbol{k}}(t) &\geq \inf_{t \in \left[j_0 T, \frac{r^\alpha+1}{T} \cdot T\right]} X^r_{\boldsymbol{k}}(t) \\ &= \inf_{j \in \left\{j_0+1, \ldots, \frac{r^\alpha+1}{T}\right\}} \left(\inf_{t \in [(j-1)T, jT]} X^r_{\boldsymbol{k}}(t) \right) \geq cr^s; \end{aligned}$$

that is, for all $t \in [1, r^\alpha + 1]$,

$$X^r_{\boldsymbol{k}}(t) \geq cr^{s_2}. \tag{B.14}$$

In summary, we have established that, with probability $1 - e^{-\text{poly}(r)}$, for all $t \in [1, r^\alpha + 1]$,

$$X^r_{\boldsymbol{k}}(t) \geq cr^{s_2} = cr^{s(\boldsymbol{k})}.$$

By the stationarity of the processes $\boldsymbol{X}^r(\cdot)$, we can conclude that

$$\mathbb{P}\left(X^r_{\boldsymbol{k}}(t) \geq cr^{s_2},\ \forall t \in [0, r^\alpha]\right) \geq 1 - e^{-\text{poly}(r)}.$$

This completes the induction step, assuming that claim 1 holds. It is now left to prove claim 1.

Proof of claim 1: Let us first focus on the subinterval $[0, T]$. Again, because the systems indexed by r are in the stationary regime, probability bounds that we derive over the subinterval $[0, T]$ extend naturally to other subintervals $[jT, (j+1)T]$, $j = 1, 2, \cdots$. Claim 1 is an easy consequence of the following claim.

Claim 2: There exist positive constants c and $\bar{c}$ such that, with probability $1 - e^{-\text{poly}(r)}$, the following holds.

1. If $X^r_{\boldsymbol{k}}(0) \leq 4cr^{s_2}$, then

$$X^r_{\boldsymbol{k}}(T) - X^r_{\boldsymbol{k}}(0) \geq \bar{c}r^{2s_2-p-\varepsilon'}. \tag{B.15}$$

2. If $X^r_{\boldsymbol{k}}(0) \geq 2cr^{s_2}$, then

$$\inf_{t \in [0,T]} X^r_{\boldsymbol{k}}(t) \geq \frac{1}{2} X^r_{\boldsymbol{k}}(0). \tag{B.16}$$

To see how claim 2 implies claim 1, recall that our systems are in the stationary regime. Thus, for any other subinterval $[jT, (j+1)T]$, with the same probability $1 - e^{-\text{poly}(r)}$, statements (1) and (2) of claim 2 hold when we replace zero by jT and T by $(j+1)T$. Because there are a polynomial (in r) number of such subintervals, claim 1 then follows immediately. We now prove claim 2.

Proof of claim 2: To bound $X^r_{\boldsymbol{k}}(\cdot)$, we consider terms $S_1(T), S_2(T)$, and $S_3(T)$ defined in (B.7)–(B.9) separately.

a. First, consider the term $S_1(T)$ defined in (B.7). Focus on $A^r_{(\boldsymbol{k},\iota)}(T)$, the cumulative arrivals along the edge $(\boldsymbol{k},\iota)$, where we recall that $k_\iota \geq 1$, and $\tilde{\boldsymbol{k}} = \boldsymbol{k} - \boldsymbol{e}_\iota$. We have

$$A^r_{(\boldsymbol{k},\iota)}(T) = \Pi_{(\boldsymbol{k},\iota)}\left(\int_0^T \lambda_\iota r \frac{X^r_{\tilde{\boldsymbol{k}}}(u)}{X^r_{(\iota)}(u)} du\right) \leq S_1(T).$$

Let F^r_1 be the event defined by

$$F^r_1 = \left\{\frac{1}{2} r \leq Z^r(u) \leq \frac{3}{2} r, \text{ for all } u \in [0,T]\right\}.$$

Then, by inspecting the proof of the base case, it is easy to see that $\mathbb{P}(F^r_1) \geq 1 - e^{-\text{poly}(r)}$. Furthermore, for sufficiently large r, under the event F^r_1, we have that, for all $u \in [0,T]$,

$$X^r_{(\iota)}(u) \leq Z^r(u) + [Z^r(u)]^p \leq 2r.$$

Thus, for sufficiently large r, under the event $E^r \cap F^r_1$, where we recall the definition of E^r in (B.11), we have that, for all $u \in [0,T]$,

$$\frac{X^r_{\tilde{\boldsymbol{k}}}(u)}{X^r_{(\iota)}(u)} \geq \frac{\tilde{c} r^{s_1}}{2r} = \frac{\tilde{c}}{2} r^{s_1 - 1},$$

from which it follows that

$$\int_0^T \lambda_\iota r \frac{X^r_{\tilde{\boldsymbol{k}}}(u)}{X^r_{(\iota)}(u)} du \geq \lambda_\iota r \cdot \frac{\tilde{c}}{2} r^{s_1-1} \cdot T = c_1 r^{s_1+s_2-1-\varepsilon'}.$$

Here, $c_1 = (1/2)\lambda_\iota \tilde{c}$. We also define event G^r_1 to be

$$G^r_1 = \left\{\Pi_{(\boldsymbol{k},\iota)}\left(c_1 r^{s_1+s_2-1-\varepsilon'}\right) \geq \frac{1}{2} c_1 r^{s_1+s_2-1-\varepsilon'}\right\}.$$

Then, $\mathbb{P}(G^r_1) \geq 1 - e^{-\text{poly}(r)}$, and under the event $E^r \cap F^r_1 \cap G^r_1$, we have that

$$A^r_{(\boldsymbol{k},\iota)}(T) \geq \frac{1}{2} c_1 r^{s_1+s_2-1-\varepsilon'}. \tag{B.17}$$

Thus,

$$\mathbb{P}\left(S_1(T) \geq \frac{1}{2} c_1 r^{s_1+s_2-1-\varepsilon'}\right) \geq \mathbb{P}\left(A^r_{(\boldsymbol{k},\iota)}(T) \geq \frac{1}{2} c_1 r^{s_1+s_2-1-\varepsilon'}\right) \geq 1 - e^{-\text{poly}(r)}. \tag{B.18}$$

This completes our probability bound for $S_1(T)$ and part (a).

b. Next, we consider the term $S_2(T)$ defined in (B.8). We have

$$S_2(T) = \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i \in \mathcal{K}} A^r_{(\boldsymbol{k}+\boldsymbol{e}_i,i)}(T) = \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i \in \mathcal{K}} \Pi_{(\boldsymbol{k}+\boldsymbol{e}_i,i)}\left(\int_0^T \lambda_i r \cdot \frac{X^r_{\boldsymbol{k}}(u)}{X^r_{(i)}(u)} du\right).$$

Instead of deriving probability bounds for $S_2(T)$, in part (b), we only derive a probability bound for

$$\sum_{i:\boldsymbol{k}+\boldsymbol{e}_i \in \mathcal{K}} \int_0^T \lambda_i r \cdot \frac{X^r_{\boldsymbol{k}}(u)}{X^r_{(i)}(u)} du. \tag{B.19}$$

The reason is twofold. First, the term (B.19) captures the order of magnitude of $S_2(T)$. Second, the bound that we derive for the term (B.19) depends on $X^r_{\boldsymbol{k}}(0)$. Because we consider two separate cases, depending on the magnitude of $X^r_{\boldsymbol{k}}(0)$, we have separate probability bounds for $S_2(T)$, which we derive after considering the term $S_3(T)$. Consider

$$\int_0^T \lambda_i r \frac{X^r_{\boldsymbol{k}}(u)}{X^r_{(i)}(u)} du,$$

a generic summand of the term (B.19). By the base case and using the fact that $X^r_{(i)}(u) \geq X^r_{\mathbf{0}}(u)$ for all u, we have that

$$\mathbb{P}\left(X^r_{(i)}(u) \geq \frac{1}{2}r^p\right) \geq \mathbb{P}\left(X^r_{\mathbf{0}}(u) \geq \frac{1}{2}r^p\right) \geq 1 - e^{-\mathrm{poly}(r)}. \tag{B.20}$$

Furthermore, we claim that there exists a constant $c_2 > 0$ such that

$$\mathbb{P}\big(X^r_{\mathbf{k}}(u) - X^r_{\mathbf{k}}(0) \leq c_2 rT, \quad \forall u \in [0, T]\big) \geq 1 - e^{-\mathrm{poly}(r)}. \tag{B.21}$$

We establish (B.21) as follows. Let N^r be the total number of arrivals to and departures from the system up to time T. Then, it is clear that, for all $u \in [0, T]$, $X^r_{\mathbf{k}}(u) - X^r_{\mathbf{k}}(0) \leq N^r$. We now obtain a probability bound on N^r. Because we are only interested in probability bounds, we can and will at different points of the proof use different underlying probability space constructions as long as they produce the same—in law—system process. At this point, we use the following different probability space construction. We associate a driving unit-rate Poisson process $\Pi(\cdot)$ for all the arrivals to the system and an independent unit-rate Poisson process $\widetilde{\Pi}(\cdot)$ to drive all the departures. The total arrival rate at all times is r, and the total departure rate at time t is given by $\sum_i \mu_i Y^r_i(t)$. Thus, N^r has the same distribution as $\Pi(rT) + \widetilde{\Pi}(\int_0^T \sum_i \mu_i Y^r_i(u)du)$. With probability $1 - e^{-\mathrm{poly}(r)}$, $\Pi(rT) \leq 2rT$. By Lemma 5, it is easy to see that, with probability $1 - e^{-\mathrm{poly}(r)}$, for all $u \in [0, T]$, $\sum_i \mu_i Y^r_i(u)du \leq \sum_i \mu_i(2\rho_i r) = 2\sum_i \lambda_i r = c_2' r$ for $c_2' = 2\sum_i \lambda_i$. This implies that, with probability $1 - e^{-\mathrm{poly}(r)}$,

$$\widetilde{\Pi}\left(\int_0^T \sum_i \mu_i Y^r_i(u)du\right) \leq \widetilde{\Pi}\left(\int_0^T c_2' r du\right) = \widetilde{\Pi}\big(c_2' rT\big) \leq 2c_2' rT.$$

Thus, with probability $1 - e^{-\mathrm{poly}(r)}$, for all $u \in [0, T]$,

$$\begin{aligned} X^r_{\mathbf{k}}(u) - X^r_{\mathbf{k}}(0) &\leq N^r \overset{d}{=} \Pi(rT) + \widetilde{\Pi}\left(\int_0^T \sum_i \mu_i Y^r_i(u)du\right) \\ &\leq 2rT + 2c_2' rT = c_2 rT. \end{aligned}$$

Here, $c_2 = 2 + 2c_2'$. This establishes (B.21).

By (B.20) and (B.21), we can bound (B.19) as follows. With probability $1 - e^{-\mathrm{poly}(r)}$,

$$\begin{aligned} \int_0^T \lambda_i r \frac{X^r_{\mathbf{k}}(u)}{X^r_{(i)}(u)} du &\leq \int_0^T \lambda_i r \frac{X^r_{\mathbf{k}}(0) + c_2 T}{r^p/2} du \\ &\leq c_3 T r^{1-p} X^r_{\mathbf{k}}(0) + c_4 T^2 r^{2-p} \\ &= c_5 r^{s_2 - p - \varepsilon'}\big(X^r_{\mathbf{k}}(0) + r^{s_2 - \varepsilon'}\big) \end{aligned}$$

for some positive constants c_3, c_4, and c_5. It then follows immediately that, with probability $1 - e^{-\mathrm{poly}(r)}$,

$$\sum_{i:\mathbf{k}+\mathbf{e}_i \in \mathcal{K}} \int_0^T \lambda_i r \frac{X^r_{\mathbf{k}}(u)}{X^r_{(i')}(u)} du \leq c_6 r^{s_2 - p - \varepsilon'}\big(X^r_{\mathbf{k}}(0) + r^{s_2 - \varepsilon'}\big), \tag{B.22}$$

for some positive constant c_6. This completes part (b).

c. We now consider the term $S_3(T)$ defined in (B.9), which is given by

$$S_3(T) = \sum_{i:\mathbf{k}-\mathbf{e}_i \in \bar{\mathcal{K}}} D^r_{(\mathbf{k},i)}(T) = \sum_{i:\mathbf{k}-\mathbf{e}_i \in \bar{\mathcal{K}}} \widetilde{\Pi}_{(\mathbf{k},i)}\left(\int_0^T X^r_{\mathbf{k}}(u) k_i \mu_i du\right).$$

Similar to part (b), we only derive a probability bound for

$$\sum_{i:\mathbf{k}-\mathbf{e}_i \in \bar{\mathcal{K}}} \int_0^T X^r_{\mathbf{k}}(u) k_i \mu_i du. \tag{B.23}$$

By (B.21), we have that, with probability $1 - e^{-\mathrm{poly}(r)}$,

$$\begin{aligned} \int_0^T k_i \mu_i X^r_{\mathbf{k}}(u) du &\leq \int_0^T k_i \mu_i \big(X^r_{\mathbf{k}}(0) + c_2 rT\big) du \\ &= c_7 T X^r_{\mathbf{k}}(0) + c_8 r T^2 \\ &= c_9 r^{s_2 - 1 - \varepsilon'}\big(X^r_{\mathbf{k}}(0) + r^{s_2 - \varepsilon'}\big), \end{aligned}$$

and

$$\sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}}\int_0^T k_i\mu_i X_{\boldsymbol{k}}^r(u)du \le c_{10} r^{s_2-1-\varepsilon'}\left(X_{\boldsymbol{k}}^r(0)+r^{s_2-\varepsilon'}\right), \tag{B.24}$$

for some positive constants c_7, c_8, c_9, and c_{10}. This completes part (c).

Observe that, by (B.22) and (B.24), we have that, with probability $1-e^{-\mathrm{poly}(r)}$,

$$\begin{aligned}\sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}}\int_0^T \lambda_i r\frac{X_{\boldsymbol{k}}^r(u)}{X_{(i)}^r(u)}du + \sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}}\int_0^T k_i\mu_i X_{\boldsymbol{k}}^r(u)du\\ \le c_{11} r^{s_2-p-\varepsilon'}\left(X_{\boldsymbol{k}}^r(0)+r^{s_2-\varepsilon'}\right),\end{aligned}$$

for some positive constant c_{11}. For notational convenience, we use $\widetilde{S}$ to denote the LHS of the preceding inequality, that is,

$$\widetilde{S} = \sum_{i:\boldsymbol{k}+\boldsymbol{e}_i\in\mathcal{K}}\int_0^T \lambda_i r\frac{X_{\boldsymbol{k}}^r(u)}{X_{(i)}^r(u)}du + \sum_{i:\boldsymbol{k}-\boldsymbol{e}_i\in\bar{\mathcal{K}}}\int_0^T k_i\mu_i X_{\boldsymbol{k}}^r(u)du, \tag{B.25}$$

and we have

$$\mathbb{P}\left(\widetilde{S}\le c_{11}r^{s_2-p-\varepsilon'}\left(X_{\boldsymbol{k}}^r(0)+r^{s_2-\varepsilon'}\right)\right)\ge 1-e^{-\mathrm{poly}(r)}. \tag{B.26}$$

Let us also recall $S_2(T)$ and $S_3(T)$ defined in (B.8) and (B.9) and consider the distribution of $S_2(T)+S_3(T)$. Note that, by our construction of the probability space, $S_2(T)$ and $S_3(T)$ are sums of terms that correspond to arrivals and departures driven by independent underlying Poisson processes. Because we are only interested in the distribution of $S_2(T)+S_3(T)$, at this point, consider the following alternative construction of the probability space, in which all arrivals and departures that appear in the summations of $S_2(T)$ and $S_3(T)$ are driven by a *common* unit-rate Poisson process $\Pi'(\cdot)$. Then, $\widetilde{S}$ has the same distribution under the original and alternative constructions, and $S_2(T)+S_3(T)$ under the original construction has the same distribution as $\Pi'(\widetilde{S})$ under the alternative construction.

We now complete the proof of claim 2, mainly making use of (B.18) and (B.26).

Let $c=c_1/[64c_{11}]$ and $\bar{c}=c_1/4$. We first consider case 1 of claim 2 and suppose that $X_{\boldsymbol{k}}^r(0)\le 4cr^{s_2}$. For sufficiently large r, $4cr^{s_2}\ge r^{s_2-\varepsilon'}$. Thus, by (B.26), with probability $1-e^{-\mathrm{poly}(r)}$,

$$\widetilde{S}\le c_{11}r^{s_2-p-\varepsilon'}\left(X_{\boldsymbol{k}}^r(0)+r^{s_2-\varepsilon'}\right)\le c_{11}r^{s_2-p-\varepsilon'}\cdot(8cr^{s_2})\le\frac{c_1}{8}r^{2s_2-p-\varepsilon'}=\frac{\bar{c}}{2}r^{2s_2-p-\varepsilon'}.$$

Because $\varepsilon'<s_2-p/2$, $2s_2-p-\varepsilon'>0$, and with probability $1-e^{-\mathrm{poly}(r)}$, we have

$$\Pi'\left(\widetilde{S}\right)\le\Pi'\left(\frac{\bar{c}}{2}r^{2s_2-p-\varepsilon'}\right)\le\bar{c}r^{2s_2-p-\varepsilon'}.$$

Thus, the event

$$\Pi'\left(\widetilde{S}\right)\le\frac{c_1}{4}r^{2s_2-p-\varepsilon'},$$

and then also the event

$$S_2(T)+S_3(T)\le\frac{c_1}{4}r^{2s_2-p-\varepsilon'},$$

hold with probability $1-e^{-\mathrm{poly}(r)}$. By (B.22), we have that, with probability $1-e^{-\mathrm{poly}(r)}$,

$$\begin{aligned}X_{\boldsymbol{k}}^r(T)-X_{\boldsymbol{k}}^r(0)&\ge S_1(T)-S_2(T)-S_3(T)\\&\ge\frac{c_1}{2}r^{s_1+s_2-1-\varepsilon'}-\bar{c}r^{2s_2-p-\varepsilon'}\\&=\bar{c}r^{2s_2-p-\varepsilon'},\end{aligned}$$

where the last equality follows from the fact that $s_1=s_2+(1-p)$. This completes the proof of case 1.

Next, consider case 2 and suppose that $X_k^r(0) \geq 2cr^{s_2}$. For sufficiently large r, $X_k^r(0) \geq 2cr^{s_2} \geq r^{s_2-\varepsilon'}$. Then, with probability $1 - e^{-\mathrm{poly}(r)}$,

$$\widetilde{S} \leq c_{11} r^{s_2-p-\varepsilon'}\left(X_k^r(0) + r^{s_2-\varepsilon'}\right) \leq 2c_{11} r^{s_2-p-\varepsilon'} X_k^r(0) \leq \frac{1}{4} X_k^r(0),$$

and with probability $1 - e^{-\mathrm{poly}(r)}$,

$$\Pi'\left(\widetilde{S}\right) \leq \Pi'\left(\frac{1}{4} X_k^r(0)\right) \leq \frac{1}{2} X_k^r(0).$$

Thus, with probability $1 - e^{-\mathrm{poly}(r)}$, for every $u \in [0, T]$,

$$\begin{aligned} X_k^r(u) - X_k^r(0) &\geq -S_2(u) - S_3(u) \\ &\geq -S_2(T) - S_3(T) \\ &\geq -\frac{1}{2} X_k^r(0), \end{aligned}$$

where we note that the first two inequalities hold with probability one. This establishes (B.16) and completes the proof of case 2. This concludes the proof of claim 2. □

Endnote

[1] We treat $r^{\alpha+1/2}$ as if it were guaranteed to be an integer. Similarly, in the sequel, whenever Cr^{β} is used as an index for some constants $\beta > 0$ and $C > 0$, we treat it as an integer. Rounding them up or down to a nearest integer would overburden our notation but would not affect our order-of-magnitude estimates.

References

Bansal N, Caprara A, Sviridenko M (2009) A new approximation method for set covering problems, with applications to multidimensional bin packing. *SIAM J. Comput.* 39(4):1256–1278.
Chekuri C, Khanna S (2004) On multidimensional packing problems. *SIAM J. Comput.* 33(4):837–851.
Coffman EG Jr, Stolyar AL (2001) Bandwidth packing. *Algorithmica* 29(1–2):70–88.
Coffman EG Jr, Garey MR, Johnson DS (1983) Dynamic bin packing. *SIAM J. Comput.* 12(2):227–258.
Coffman EG Jr, Csirik J, Galambos G, Martello S, Vigo D (2013) Bin packing approximation algorithms: Survey and classification. Pardalos PM, Du DZ, Graham RL, eds. *Handbook of Combinatorial Optimization* (Springer, New York), 455–531.
Csirik J, Johnson DS, Kenyon C, Orlin JB, Shor PW, Weber RR (2006) On the sum-of-squares algorithm for bin packing. *J. ACM.* 53(1):1–65.
Gamarnik D (2004) Stochastic bandwidth packing process: Stability conditions via Lyapunov function technique. *Queueing Systems* 48:339–363.
Ghaderi J, Zhong Y, Srikant R (2014) Asymptotic optimality of best-fit for stochastic bin packing. *ACM SIGMETRICS Performance Eval. Rev.* 42(2):64–66.
Gulati A, Holler A, Ji M, Shanmuganathan G, Waldspurger C, Zhu X (2012) VMware distributed resource management: Design, implementation and lessons learned. *VMware Tech. J.* 1(1):45–64.
Guo Y, Stolyar AL, Walid A (2015) Shadow-routing based dynamic algorithms for virtual machine placement in a network cloud. *IEEE Trans. Cloud Comput.* 6(1):209–220.
Gupta V, Radovanovic A (2012) Online stochastic bin packing. Preprint, submitted November 12, https://arxiv.org/abs/1211.2687.
Jiang JW, Lan T, Ha S, Chen M, Chiang M (2012) Joint VM placement and routing for data center traffic engineering. *2012 Proc. IEEE INFOCOM*, 2876–2880.
Liptser RS, Shiryaev AN (1989) *Theory of Martingales* (Kluwer Academic Publishers, Dordrecht, Netherlands).
Maguluri ST, Srikant R (2013) Scheduling jobs with unknown duration in clouds. *IEEE/ACM Trans. Networking* 22(6):1938–1951.
Maguluri ST, Srikant R, Ying L (2012) Stochastic models of load balancing and scheduling in cloud computing clusters. *2012 Proc. IEEE* INFOCOM, 702–710.
Mitzenmacher M, Upfal E (2005) *Probability and Computing: Randomized Algorithms and Probabilistic Analysis* (Cambridge University Press, Cambridge, UK).
Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.
Stolyar AL (2013) An infinite server system with general packing constraints. *Oper. Res.* 61(5):1200–1217.
Stolyar AL (2017) Large-scale heterogeneous service systems with general packing constraints. *Adv. Appl. Probab.* 49(1):61–83.
Stolyar AL, Zhong Y (2013) A large-scale service system with packing constraints: Minimizing the number of occupied servers. *ACM SIGMETRICS Performance Eval. Rev.* 41(1):41–52.
Stolyar AL, Zhong Y (2015) Asymptotic optimality of a greedy randomized algorithm in a large-scale service system with general packing constraints. *Queueing Systems* 79(2):117–143.