An infinite server system with general packing constraints

Alexander L. Stolyar
Bell Labs, Alcatel-Lucent
600 Mountain Ave., 2C-322
Murray Hill, NJ 07974
stolyar@research.bell-labs.com

June 1, 2012

Abstract

We consider a service system model primarily motivated by the problem of efficient assignment of virtual machines to physical host machines in a network cloud, so that the number of occupied hosts is minimized.

There are multiple input flows of different type customers, with a customer mean service time depending on its type. There is infinite number of servers. A server packing *configuration* is the vector $k = \{k_i\}$, where k_i is the number of type i customers the server "contains". Packing constraints must be observed, namely there is a fixed finite set of configurations k that are allowed. Service times of different customers are independent; after a service completion, each customer leaves its server and the system. Each new arriving customer is placed for service immediately; it can be placed into a server already serving other customers (as long as packing constraints are not violated), or into an idle server.

We consider a simple parsimonious real-time algorithm, called *Greedy*, which attempts to minimize the increment of the objective function $\sum_k X_k^{1+\alpha}$, $\alpha > 0$, caused by each new assignment; here X_k is the number of servers in configuration k. (When α is small, $\sum_k X_k^{1+\alpha}$ approximates the total number $\sum_k X_k$ of occupied servers.) Our main results show that certain versions of the Greedy algorithm are *asymptotically optimal*, in the sense of minimizing $\sum_k X_k^{1+\alpha}$ in stationary regime, as the input flow rates grow to infinity. We also show that in the special case when the set of allowed configurations is determined by *vector-packing* constraints, Greedy algorithm can work with *aggregate configurations* as opposed to exact configurations k, thus reducing computational complexity while preserving the asymptotic optimality.

1 Introduction

The primary motivation for this work is the following problem arising in cloud computing: how to assign various types of virtual machines to physical host machines (in a data center) in real time, so that the total number of host machines in use is minimized. It is very desirable that an assignment algorithm is simple, does need to know the system parameters, and makes decisions based on the current system state only. (An excellent overview of this and other resource allocation issues arising in cloud computing can be found in [4].)

A data center (DC) in the "cloud" consists of a number of host machines. Assume that all hosts are same: each of them possesses the amount $B_n > 0$ of resource n, where $n \in \{1, 2, \ldots, N\}$ is a resource index. (For example, resource 1 is CPU, resource 2 is memory, etc.) The DC receives requests for virtual machine (VM) placements; VMs can be of different types $i \in \{1, \ldots, I\}$; a type i VM requires the amounts $b_{i,n} > 0$ of each resource n. Several VMs can share the same host, as long as the host's capacity constraints are not violated; namely, a host can simultaneously contain a set of VMs given by a vector $k = (k_1, \ldots, k_I)$, where k_i is the number of type i VMs, as long as for each resource n

$$\sum_i k_i b_{i,n} \le B_n. \tag{1}$$

Thus, VMs can be assigned to hosts already containing other VMs, subject to the above "packing" constraints. After a certain random sojourn (service) time each VM vacates its host (leaves the system), which increases the "room" for new arriving VMs to be potentially assigned to the host. A natural problem is to find a real-time algorithm for assigning VM requests to the hosts, which minimizes (in appropriate sense) the total number of hosts in use. Clearly, such a scheme will maximize the DC capacity; or, if it leaves a large number of hosts unoccupied, those hosts can be (at least temporarily) turned off to save energy.

More specifically, the model assumptions that we make are as follows:
(a) The exact nature of "packing" constraints will not be important – we just assume that the feasible configuration vectors k (describing feasible sets of VMs that can simultaneously occupy one host) form a finite set $\mathcal{K}$; and assume monotonicity – if $k \in \mathcal{K}$ then so is any $k' \le k$.
(b) There is no limit on the number of hosts that can be used and each new VM is assigned to a host immediately – so it is an *infinite server* model, with no blocking or waiting.
(c) Service times of different VMs are independent of each other, even for VMs served simultaneously on the same host.
(d) We further assume in this paper that the arrival processes of VMs of each type are Poisson and service time distributions are exponential. These assumptions are not essential and can be much relaxed, as discussed in Section 8.2.

The basic problem we address in this paper is:

$$\text{minimize} \quad \sum_k X_k^{1+\alpha}, \tag{2}$$

where $\alpha > 0$ is a fixed parameter, and X_k is the (random) number of hosts having configuration k in the stationary regime. (Clearly, when α is small, $\sum_k X_k^{1+\alpha}$ approximates the total number $\sum_k X_k$ of occupied hosts.) We consider the *Greedy* real-time (on-line) VM assignment algorithm, which, roughly speaking, tries to minimize the increment of the objective function $\sum_k X_k^{1+\alpha}$ caused by each new assignment. Our main results show that certain versions of the Greedy algorithm are *asymptotically optimal*, as the input flow rates become large or, equivalently, the average number of VMs in the system becomes large. We also show (in Section 7) that in the special case when feasible configurations are determined by constraints (1), Greedy algorithm can work with "aggregate configurations" as opposed to exact configurations k, thus reducing computational complexity while preserving the asymptotic optimality.

1.1 Previous work

Our model is related to the vast literature on the classical *stochastic bin packing* problems. (For a good recent review of one-dimensional bin packing see e.g. [2].) In particular, in *online* stochastic bin packing problems, random-size items arrive in the system and need to be placed according to an online algorithm into finite size bins; the items never leave or move between bins; the typical objective is to minimize the number of occupied bins. A bin packing problem is *multi-dimensional*, when bins and item sizes are vectors; the problems with the packing constraints (1) are called *multi-dimensional vector packing* (see e.g. [1] for a recent review). Bin packing *service* systems arise when there is a random in time input flow of random-sized items (customers), which need to be served by a bin (server) and leave after a random service time; the server can simultaneously process multiple customers as long as they can simultaneously fit into it; the customers waiting for service are queued; a typical problem is to determine the maximum throughput under a given ("packing") algorithm for assigning customers for service. (See e.g. [3] for a review of this line of work.) Our model is similar to the latter systems, except there are multiple bins (servers), in fact – infinite number in our case. Models of this type are more recent (see e.g. [5, 6]). Paper [5] addresses a real-time VM allocation problem, which in particular includes packing constraints; the approach of [5] is close in spirit to Markov Chain algorithms used in combinatorial optimization. Paper [6] is concerned mostly with maximizing throughput of a queueing system (where VMs can actually wait for service) with a finite number of bins.

The asymptotic regime in this paper is such that the input flow rates scale up to infinity. In this respect, our work is related to the (also vast) literature on queueing systems in the *many servers* regime. (See e.g. [9] for an overview. The name "many servers" reflects the fact that the number of servers scales up to infinity as well, linearly with the input rates; this condition is irrelevant in our case of infinite number of servers.) In particular, we study *fluid limits* of our system, obtained by scaling the system state down by the (large) total number of customers. We note, however, that packing constraints are not present in the previous work on the many servers regime, to the best of our knowledge.

2 Model and main results

We consider a service system with I input customer flows of different types, indexed by $i \in \{1, 2, \ldots, I\} \equiv \mathcal{I}$. Each flow i is Poisson with rate $\Lambda_i > 0$. Service time of a type i customer is an exponentially distributed random variable with mean $1/\mu_i$. All input flows and customer service times are mutually independent. There is an infinite number of servers. Each server can potentially serve more than one customer simultaneously, subject to the following general "packing" constraints. We say that a vector $k = \{k_i,\ i \in \mathcal{I}\}$ with non-negative integer components is a server *configuration*, if a server can simultaneously serve a combination of different type customers given by vector k. The set of all configurations is finite, and is denoted by $\bar{\mathcal{K}}$. We assume that $k \in \bar{\mathcal{K}}$ implies that all "smaller" configurations $k' \leq k$ belong to $\bar{\mathcal{K}}$ as well. Without loss of generality assume that $e_i \in \bar{\mathcal{K}}$ for all types i, where e_i is the i-th coordinate unit vector (otherwise, i-customers cannot be served at all). By convention, the (component-wise) zero vector $k = 0$ belongs to $\bar{\mathcal{K}}$ – this is the configuration of an "empty" server; we denote by $\mathcal{K} = \bar{\mathcal{K}} \setminus \{0\}$ the set of configurations *not* including the zero configuration.

An important feature of the model is that simultaneous service does *not* affect the service rates of individual customers; in other words, the service time of a customer is unaffected by whether or not there are other customers served simultaneously by the same server. Each arriving customer is immediately placed for service in one of the servers; it can be "added" to an empty or non-empty server as long as configuration feasibility constraint is not violated, i.e. it can be added to any server whose configuration $k \in \bar{\mathcal{K}}$ (before the addition) is such that $k + e_i \in \mathcal{K}$. When the service of an i-customer by the server in configuration k is completed, the customer leaves the system and the server's configuration changes to $k - e_i$. Denote by X_k the number of servers with configuration $k \in \mathcal{K}$. The system state is then the vector $X = \{X_k,\ k \in \mathcal{K}\}$.

A *service discipline* ("packing rule") determines where an arriving customer is placed, as a function of the current system state X. Under any well-defined service discipline, the system state at time t, $X(t)$, is a continuous time, countable Markov chain. It is easily seen to be irreducible and positive recurrent; the positive recurrence follows from the fact that the total number $Y_i(t)$ of type i customers in the system is the process independent of the service discipline, and its stationary distribution is Poisson with mean Λ_i/μ_i. Therefore, the process $X(t),\ t \geq 0$, has a unique stationary distribution.

We are interested in finding a service disciplines minimizing (in a certain sense) the total number of non-empty servers in the stationary regime. For example, an objective can be

$$\min \mathbb{E} \sum_{k \in \mathcal{K}} X_k^{1+\alpha}(\infty), \tag{3}$$

where $\alpha \geq 0$ is a parameter, and $X(\infty)$ denotes the random system state in stationary regime. Another possible objective is

$$\min \mathbb{P}\{\sum_{k \in \mathcal{K}} X_k^{1+\alpha}(\infty) > C\}, \tag{4}$$

where C is a fixed threshold. In both problems, setting $\alpha = 0$ obviously corresponds to the exact objective of minimizing the number of servers in use; however, if a good discipline for $\alpha > 0$ exists, using such discipline with positive α close to 0 would (hopefully) also produce a good solution for the $\alpha = 0$ case.

In this paper we consider objectives (3) and (4) with $\alpha > 0$, and prove the following *Greedy* service discipline (or, rather its slight modification, to be precise) is asymptotically optimal as the flows' input rates become large.

Definition 1 (Greedy discipline)**.**

1. Integral form (Greedy-I). *Define* $F(x) = \sum_{k\in\mathcal{K}}(1+\alpha)^{-1}x_k^{1+\alpha}$, $x \in \mathbb{R}_+^{|\mathcal{K}|}$. *A type i customer arriving at time t is added to an available configuration k (with either $k = 0$ or $X_k(t-) > 0$) such that $k + e_i \in \mathcal{K}$ and the increment $F(X(t)) - F(X(t-))$ is the smallest. (Here $X(t-)$ and $X(t)$ are the states just before and just after the addition; so that $X_{k+e_i}(t) = X_{k+e_i}(t-) + 1$ and, unless $k = 0$, $X_k(t) = X_k(t-) - 1$.) The ties are broken according to an arbitrary deterministic rule.*

2. Differential form (Greedy-D). *For each $k \in \mathcal{K}$, denote $W_k(x) = (\partial/\partial x_x)F(x) = x_k^\alpha$, $x \in \mathbb{R}_+^{|\mathcal{K}|}$. A type i customer arriving at time t is added to an available configuration k (with either $k = 0$ or $X_k(t-) > 0$) such that $k + e_i \in \mathcal{K}$ and the difference $W_{k+e_i}(X(t-)) - I\{k \neq 0\}W_k(X(t-))$ is the smallest. (Here $I\{\cdot\}$ is the indicator function, equal to 1 when the condition holds and 0 otherwise.) The ties are broken according to an arbitrary deterministic rule.*

In the asymptotic analysis of this paper, the two forms of Greedy algorithm are essentially identical, because they induce the same dynamics of the system in the "fluid limit". We will analyze the differential form, as it is slightly more convenient to work with, and is probably more easily implementable in practice; it should be clear that all results (along with essentially same proofs) hold for the integral form as well.

Remark. All results will hold for a more general objective function $F(x) = \sum_{k\in\mathcal{K}} c_k x_k^{1+\alpha}$, with arbitrarily positive weights c_k. The generalization is completely straightforward – we choose to work with $c_k = (1+\alpha)^{-1}$ simply to avoid "carrying" factors $c_k(1 + \alpha)$, which clog notation.

We now define the asymptotic regime. Let $r \to \infty$ be a positive scaling parameter. (To be specific, assume that $r \geq 1$, and r increases to infinity along a discrete sequence.) Input rates scale linearly with r; namely, for each r, $\Lambda_i = \lambda_i r$, where λ_i are positive parameters. Let $X^r(\cdot)$ be the process associated with system with parameter r, and $X^r(\infty)$ be the (random) system state in the stationary regime. For each i denote by $Y_i^r(t) \equiv \sum_{k\in\mathcal{K}} k_i X_k^r(t)$ the total number of customers of type i. Since arriving customers are taken for service immediately and their service times are independent (of the rest of the system), the distribution of $Y_i^r(\infty)$ is Poisson with mean $r\rho_i$, where $\rho_i \equiv \lambda_i/\mu_i$. Moreover, $Y_i^r(\infty)$ are independent across i. Since the total number of occupied servers is no greater than the total number of customers, $\sum_k X_k^r(t) \leq Z^r(t) \equiv \sum_i Y_i^r(t)$, we have a simple upper bound on the total number of occupied servers in steady state, $\sum_k X_k^r(\infty) \leq Z^r(\infty)$, where $Z^r(\infty)$ is a Poisson random variable with mean $r\sum_i \rho_i$. Without loss of generality, from now on in the paper we assume $\sum_i \rho_i = 1$. (This is equivalent to rechoosing parameter r to be $r\sum_i \rho_i$.)

The *fluid scaled* process is $x^r(t) = X^r(t)/r$; for any r, $x^r(t)$ takes values in (the positive orthant of) Euclidean space $\mathbb{R}^{|\mathcal{K}|}$, where $|\mathcal{K}|$ is the cardinality of $\mathcal{K}$. Similarly, $y_i^r(t) = Y_i^r(t)/r$ and $z^r(t) = Z^r(t)/r$. Since $\sum_k x_k^r(\infty) \leq z^r(\infty) = Z^r(\infty)/r$, we see that the random variables $(\sum_k x_k^r(\infty))^{1+\alpha}$ are uniformly integrable in r (for any fixed $\alpha \geq 0$). This in particular implies that the sequence of distributions of $x^r(\infty)$ is tight, and therefore there always exists a limit in distribution $x^r(\infty) \implies x(\infty)$, along a subsequence of r. (The limit depends on the service discipline, of course.) The limit (random) vector $x(\infty)$ satisfies the following conservation laws:

$$\sum_{k\in\mathcal{K}} k_i x_k(\infty) \equiv y_i(\infty) = \rho_i, \quad \forall i, \tag{5}$$

implying, in particular,

$$z_i(\infty) \equiv \sum_i y_i(\infty) \equiv \sum_i \rho_i = 1. \tag{6}$$

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

$$\mathcal{X} \equiv \{x \in \mathbb{R}^{|\mathcal{K}|} \mid x_k \geq 0,\ \forall k \in \mathcal{K};\ \sum_k k_i x_k = \rho_i,\ \forall i \in \mathcal{I}\}.$$

(We will slightly abuse notation by using symbol x for a generic element of $\mathcal{X}$; while $x(\infty)$, and later $x(t)$, refer to random variables taking values in $\mathcal{X}$.)

The asymptotic regime and the associated basic properties (5) and (6) hold *for any service discipline*, not necessarily Greedy-D.

Note that function $F(x)$ with $\alpha > 0$ is strictly convex on $\mathcal{X}$, and therefore there is the unique optimal point x^* minimizing F:

$$x^* = \arg\min_{u \in \mathcal{X}} F(u). \tag{7}$$

(Note that, if $\alpha = 0$, then x^* is an optimal solution of a linear program and therefore might not be unique. We do not consider the $\alpha = 0$ case in this paper.)

Our **main results** are as follows.
Let $\alpha > 0$. We prove that, as $r \to \infty$, the convergence is distribution $x^r(\infty) \implies x^*$ holds in two cases:
(a) For the closed system, with the constant number $Y_i^r = \rho_i r$ of customers of each type, operating under the Greedy-D discipline. (The exact result is Theorem 2.)
(b) For the original system (as defined above), operating under a slightly modified Greedy-D discipline, called Greedy-DM. (The exact result is Theorem 12.)
In addition, in the special case when feasible configurations are determined by vector-packing constraints (1), we prove that essentially same results hold if Greedy algorithm uses quantities X_k^r aggregated over classes of equivalent configurations, thus reducing the total number of system variables the algorithm needs to maintain. (The exact results are Theorems 19 and 21.)

2.1 Basic notation and conventions

Standard Euclidean norm of a vector $x \in \mathbb{R}^n$ is denoted $\|x\|$; the distance from vector x to a set U in a Euclidean space is denoted $d(x, U) = \inf_{u \in U} \|x - u\|$; $\mathbb{R}_+$ is the set of real non-negative numbers; the cardinality of a finite set $\mathcal{N}$ is $|\mathcal{N}|$. Symbol $\to$ means ordinary convergence in $\mathbb{R}^n$; $\implies$ denotes convergence in distribution of random variables taking values in $\mathbb{R}^n$, equipped with the Borel σ-algebra; abbreviation *w.p.1* means convergence *with probability 1*. We often write $x(\cdot)$ to mean the function (or random process) $x(t)$, $t \geq 0$. Abbreviation *u.o.c.* means *uniform on compact sets* convergence of functions, with the argument (usually in $[0, \infty)$) determined by the context. We often write $\{x_k\}$ to mean the vector $\{x_k,\ k \in \mathcal{K}\}$, with the set of indices $\mathcal{K}$ determined by the context.

For a finite set of scalar functions $f_n(t),\ t \geq 0,\ n \in \mathcal{N}$, a point t is called *regular* if for any subset $\mathcal{N}' \subseteq \mathcal{N}$ the proper derivatives

$$\frac{d}{dt} \max_{n \in \mathcal{N}'} f_n(t) \text{ and } \frac{d}{dt} \min_{n \in \mathcal{N}'} f_n(t)$$

exist.

3 Closed system. Greedy-D optimality

In this section we consider a "closed" version of our system. Namely, assume that there is a fixed number $\rho_i r$ customers of type i (in a system with parameter r); there are no exogenous arrivals into the system – when a service of a type i customer is completed, the customer immediately has to be placed into a server for a new service. Service discipline determines where the customer is placed, based on the current system state.

It is easy to see that, for any r, a stationary distribution of the process exists under any given service discipline, because the process in this case is a finite-state continuous time Markov chain.

The main result of this section is the following

Theorem 2. *Consider a sequence of closed systems, indexed by r, and let $x^r(\infty)$ denote the random state of the (fluid-scaled) process is a stationary regime, under the Greedy-D discipline with $\alpha > 0$. Then, as $r \to \infty$,*

$$x^r(\infty) \implies x^*,$$

where x^* is defined in (7).

To prove the theorem we will need to study the transient behavior of the fluid-scaled process and its limits.

Let $\mathcal{M}$ denote the set of pairs (k,i) such that $k \in \mathcal{K}$ and $k - e_i \in \bar{\mathcal{K}}$. Each pair (k,i) is associated with the "edge" $(k - e_i, k)$ connecting configurations $k - e_i$ and k; often we refer to this edge as (k,i). By "arrival along the edge (k,i)" we will mean placement of a type i customer into a server configuration $k - e_i$ to form configuration k; similarly, "departure along the edge (k,i)" is a departure of a type i customer from a server in configuration k, which changes its configuration to $k - e_i$.

For each $(k,i) \in \mathcal{M}$, consider an independent unit-rate Poisson process $\Pi_{ki}(t),\ t \geq 0$. We have the functional strong law of large numbers:

$$\frac{1}{r}\Pi_{ki}(rt) \to t, \quad u.o.c., \quad w.p.1. \tag{8}$$

Without loss of generality, assume that the Markov process $X^r(\cdot)$ for each r is driven by the common set of Poisson processes $\Pi_{ki}(\cdot)$, as follows. For each $(k,i) \in \mathcal{M}$, let us denote by $D^r_{ki}(t)$ the total number of departures along the edge (k,i) in $[0,t]$; then we can assume that

$$D^r_{ki}(t) = \Pi_{ki}(\int_0^t X^r_k(\xi) k_i \mu_i d\xi). \tag{9}$$

Each type-i departure in the closed system is simultaneously a type-i "arrival", which is allocated according to Greedy-D. Thus, the realization of the process is (w.p.1) uniquely determined by the initial state $X^r(0)$ and the realizations of $\Pi_{ki}(\cdot)$. Denote by $A^r_{ki}(t)$ the total number of arrivals allocated along edge (k,i). Obviously, we have the conservation law for each type i: $\sum_k A^r_{ki}(t) \equiv \sum_k D^r_{ki}(t),\ \forall t \geq 0$. In addition to

$$x^r_k(t) = \frac{1}{r} X^r_k(t),$$

we introduce other fluid-scaled quantities:

$$d^r_{ki}(t) = \frac{1}{r} D^r_{ki}(t), \quad a^r_{ki}(t) = \frac{1}{r} A^r_{ki}(t).$$

A set of Lipschitz continuous functions $[\{x_k(\cdot),\ k \in \mathcal{K}\}, \{d_{ki}(\cdot),\ (k,i) \in \mathcal{M}\}, \{a_{ki}(\cdot),\ (k,i) \in \mathcal{M}\}]$ on the time interval $[0,\infty)$ we call a *fluid sample path* (FSP), if there exist realizations of $\Pi_{ki}(\cdot)$ satisfying (8) and a fixed subsequence of r, along which

$$\begin{aligned}&[\{x^r_k(\cdot),\ k \in \mathcal{K}\}, \{d^r_{ki}(\cdot),\ (k,i) \in \mathcal{M}\}, \{a^r_{ki}(\cdot),\ (k,i) \in \mathcal{M}\}] \to \\ &[\{x_k(\cdot),\ k \in \mathcal{K}\}, \{d_{ki}(\cdot),\ (k,i) \in \mathcal{M}\}, \{a_{ki}(\cdot),\ (k,i) \in \mathcal{M}\}], \quad u.o.c.\end{aligned} \tag{10}$$

It is easy to see that the family of all FSPs is *uniformly* Lipschitz.

Lemma 3. *Suppose the initial states $x^r(0)$ are fixed and are such that $x^r(0) \to x(0)$. Then, w.p.1 for any subsequence of r there exists a further subsequence of r, along which the convergence (10) holds, where the limit is an FSP.*

Proof is very standard. Essentially, it suffices to observe that, with probability 1, each sequence (in r) of functions $d^r_{ki}(\cdot)$ is asymptotically Lipschitz, namely for some $C > 0$, and all $0 \le t_1 \le t_2 < \infty$,

$$\limsup_r d^r_{ki}(t_2) - d^r_{ki}(t_1) \le C(t_2 - t_1),$$

which in turn follows from (8). And similarly for functions $a^r_{ki}(\cdot)$, because $a^r_{ki}(t) \le \sum_k d^r_{ki}(t)$. Using this and (10), we easily verify the u.o.c. convergence of all $x^r_k(\cdot), d^r_{ki}(\cdot), a^r_{ki}(\cdot)$, along possibly a further subsequence, and the fact that the limits are Lipschitz. We omit details. □

For an FSP, at a regular time point t, we denote $v_{ki}(t) = (d/dt)a_{ki}(t)$ and $w_{ki}(t) = (d/dt)d_{ki}(t)$. In other words, $v_{ki}(t)$ and $w_{ki}(t)$ are the rates of type i "fluid" arrival and departure along edge (k, i), respectively.

Lemma 4. *An FSP satisfies the following properties:*

$$y_i(t) = \sum_k k_i x_k(t) \equiv \rho_i;$$

at any regular point t,

$$w_{ki}(t) = k_i \mu_i x_k, \quad \forall (k,i) \in \mathcal{M},$$

$$\sum_{k:(k,i)\in\mathcal{M}} w_{ki}(t) = \sum_{k:(k,i)\in\mathcal{M}} v_{ki}(t) = \lambda_i, \quad \forall i \in \mathcal{I},$$

$$(d/dt)x_k(t) = [\sum_{i:k-e_i\in\bar{\mathcal{K}}} v_{ki} - \sum_{i:k+e_i\in\mathcal{K}} v_{k+e_i,i}] - [\sum_{i:k-e_i\in\bar{\mathcal{K}}} w_{ki} - \sum_{i:k+e_i\in\mathcal{K}} w_{k+e_i,i}], \quad \forall k \in \mathcal{K}.$$

Proof is both standard and obvious. □

4 Characterization of the optimal point x^* and related properties

This section describes properties of the optimal point x^*, and related general properties of the "allocation" vectors, which, roughly speaking, have the meaning of the vector $v(t) = \{v_{ki}(t),\ (k,i) \in \mathcal{M}\}$ of arrival rates. The results of this section are *not* limited the closed system or the Greedy-D algorithm.

Recall that x^* is defined as the unique optimal solution of the convex optimization problem

$$\min_{x\in\mathbb{R}_+^{|\mathcal{K}|}} F(x) \tag{11}$$

subject to

$$\sum_{k\in\mathcal{K}} k_i x_k = \rho_i, \quad \forall i, \tag{12}$$

where $F(x) = \sum_k (1+\alpha)^{-1} x_k^{1+\alpha}$ with $\alpha > 0$.

Using Lagrange multipliers η_i for the constraints (12), the Lagrangian is

$$\sum_k \frac{1}{1+\alpha} x_k^{1+\alpha} + \sum_i \eta_i [\rho_i - \sum_k k_i x_k] = \sum_k [\frac{1}{1+\alpha} x_k^{1+\alpha} - x_k \sum_i \eta_i k_i] + \sum_i \eta_i \rho_i.$$

Therefore, we have the following characterization: vector $x = x^*$ if and only if there exist constants η_i such that

$$x_k^\alpha = \max\{\sum_i k_i \eta_i, 0\}. \tag{13}$$

From here we also observe that at least one of the Lagrange multipliers is strictly positive, $\eta_i > 0$, and therefore, for such i, $x^*_{e_i} > 0$.

For $x \in \mathbb{R}_+^{|\mathcal{K}|}$ and $(k,i) \in \mathcal{M}$ denote

$$\Delta_{ki} = \Delta_{ki}(x) = x_k^\alpha - x_{k-e_i}^\alpha,$$

where, by convention, $x_k = 0$ when $k = 0$.

Definition 5 (Allocations). *A vector $\gamma = \{\gamma_{ki},\ (k,i) \in \mathcal{M}\}$, such that all $\gamma_{ki} \geq 0$,*

$$\sum_{k\ :\ (k,i)\in\mathcal{M}} \gamma_{ki} = \lambda_i, \quad \forall i,$$

we will call an allocation *(of the arrival rates). For a given $x \in \mathcal{X}$, the allocation $\gamma' = \gamma'(x)$, with components $\gamma'_{ki} = k_i \mu_i x_k$, is called* neutral. *An allocation γ is called a* simple improving allocation, *or* SI-allocation, *for a given $x \in \mathcal{X}$, if there exist edges (k,i) and (k',i), with the same i but $k' \neq k$, such that the following conditions hold:*

$$k_i > 0,\ \ x_k > 0,$$

$$\text{either}\ \ k' = e_i \ \ \text{or}\ \ [k'_i > 0 \ \text{and}\ x_{k'-e_i} > 0], \tag{14}$$

$$\Delta_{k'i} < \Delta_{ki}, \tag{15}$$

$$\gamma_{ki} = 0, \quad \gamma_{k'i} = k'_i \mu_i x_{k'} + k_i \mu_i x_k,$$

$$\gamma_{\ell j} = \gamma'_{\ell j} = \ell_j \mu_j x_\ell, \quad \text{for } (\ell,j) \neq (k,i),(k',i).$$

We denote by $\Gamma(x)$ the set consisting of the neutral and all SI-allocations for x.

The meaning of this definition is simple and becomes clear with the use of following notation. For $x \in \mathcal{X}$, denote

$$D(\gamma) = D(\gamma, x) = \sum_{(k,i)\in\mathcal{M}} \Delta_{ki}[\gamma_{ki} - k_i \mu_i x_k].$$

(The meaning of this is: $D(v(t), x(t)) = (d/dt)F(x(t))$ for an FSP at a regular point t.) Clearly, $D(\gamma') = 0$ for the neutral allocation, and $D(\gamma) < 0$ for any SI-allocation. This is because any SI-allocation γ, associated with edges (k,i) and (k',i), is obtained from the neutral by "reallocating" the positive amount $k_i \mu_i x_k$ of type-i input rate from edge (k,i) to edge (k',i) with strictly smaller $\Delta_{k'i}$; condition (14) guarantees that there are servers in the configuration $k' - e_i$, to which these reallocated type-i arrivals can be added.

Lemma 6. *If $x \in \mathcal{X}$ and $x \neq x^*$, then there exists at least one SI-allocation $\gamma \in \Gamma(x)$.*

Proof is by contradiction. Suppose SI-allocations do not exist. Then there must exist at least one $i \in \mathcal{I}$ such that $x_{e_i} > 0$. If not, we pick a minimal k with $x_k > 0$, for which necessarily $\sum_i k_i \geq 2$. Then we could pick i such that $k_i \geq 1$ and construct the SI-allocation γ associated with edges (k,i) and (e_i,i). (I.e., γ is obtained from the neutral allocation by "reallocating" amount $k_i \mu_i x_k$ of type i input from edge (k,i) to edge (e_i,i).) Therefore, indeed, $x_{e_i} > 0$ for at least one i.

Denote by $\mathcal{I}^+$ the set of those i with $x_{e_i} > 0$, and by $\mathcal{I}^- = \mathcal{I} \setminus \mathcal{I}^+$ the set of those i with $x_{e_i} = 0$. Set $\mathcal{I}^+$ is non-empty, while $\mathcal{I}^-$ may or may not be empty. Let us fix the following values of η_i:

$$\eta_i = \begin{cases} x_{e_i}^\alpha, & \text{if } i \in \mathcal{I}^+ \\ \min_{k\ :\ (k,i)\in\mathcal{M}} \Delta_{ki}, & \text{if } i \in \mathcal{I}^- \end{cases}$$

It follows from this definition that $\eta_i > 0$ if and only if $i \in \mathcal{I}^+$.

Let us show that (13) must hold for these η_i. Denote $u_k = \max\{\sum_i k_i \eta_i, 0\}$; so, we will show that $x_k^\alpha = u_k$ for all k. Suppose not. Let us choose a minimal k for which this relation fails. Note that necessarily $\sum_i k_i \geq 2$.

Consider first the case when $k_i = 0$ for all $i \in \mathcal{I}^-$. Pick any $i \in \mathcal{I}^+$ for which $k_i \geq 1$. If $x_k^\alpha > u_k$, then $\Delta_{ki} > \eta_i$; we can then construct the SI-allocation associated with the reallocation from (k,i) to (e_i,i); this is a contradiction. If $x_k^\alpha < u_k$, then $\Delta_{ki} < \eta_i$, and we can construct the SI-allocation, which does the "opposite" reallocation from (e_i, i) to (k, i); again, a contradiction.

The second case is when $k_i \geq 1$ for at least one $i \in \mathcal{I}^-$. We pick such an i. Condition $x_k^\alpha < u_k$ would imply $\Delta_{ki} < \eta_i$ – a contradiction with the definition of η_i. Therefore, $x_k^\alpha > u_k$, and then $x_k > 0$ and $\Delta_{ki} > \eta_i$. If $\eta_i = 0$, then we can construct the SI-allocation associated with the reallocation from (k, i) to (e_i, i). Otherwise, if $\eta_i < 0$, there exists an edge (k', i) such that $\Delta_{k'i} = \eta_i < 0$ and, consequently, $x_{k'-e_i} > 0$; we can construct the SI-allocation associated with the reallocation from (k, i) to (k', i).

We have proved that, indeed, (13) holds for the chosen values of η_i. But this means that $x = x^*$ – this contradiction completes the proof. □

For any x, denote

$$D_{min}(x) = \min_{\gamma \in \Gamma(x)} D(\gamma, x).$$

The minimum is attained and, obviously, $D_{min}(x) \leq 0$ for any x.

Lemma 7. *For any $\epsilon > 0$ there exists $\epsilon_1 > 0$ such that*

$$\|x - x^*\| \geq \epsilon \quad \textit{implies} \quad D_{min}(x) < -\epsilon_1. \tag{16}$$

Proof. We use compactness of $\mathcal{X}$. If the lemma statement does not hold, then, for some fixed $\epsilon > 0$ we can find a converging sequence $x^{(n)} \to x' \in \mathcal{X}$, $n \to \infty$, such that $D_{min}(x^{(n)}) \to 0$ and $\|x - x'\| \geq \epsilon$. By Lemma 6, there exists an SI-allocation $\gamma \in \Gamma(x')$. Since $x'_k > 0$ implies $x_k^{(n)} > 0$ for all large n, we can construct (for all large n) an SI-allocation $\gamma^{(n)} \in \Gamma(x^{(n)})$ associated with the same reallocation as the one producing γ. Then, we have $\gamma^{(n)} \to \gamma$ and $D(\gamma^{(n)}, x^{(n)}) \to D(\gamma, x') < 0$. This contradicts the assumption $D_{min}(x^{(n)}) \to 0$. □

5 Proof of Theorem 2

Lemma 8. *Any FSP (for the closed system under Greedy-D) is such that at any regular point t,*

$$\frac{d}{dt} F(x(t)) = D(v(t), x(t)) \leq D_{min}(x(t)). \tag{17}$$

Proof. Within this proof, $x = x(t)$ and $v = v(t)$. Consider a specific allocation $\gamma \in \Gamma(x)$, for which the minimum in the definition of $D_{min}(x)$ is attained; unless $x = x^*$, γ is an SI-allocation, associated with some fixed reallocation from (k, i) to (k', i). Define the following "distribution function":

$$H(\xi; v) = \sum_{(\ell, j) \in \mathcal{M} \,:\, \Delta_{\ell j} \leq \xi} v_{\ell j}, \quad \xi \in \mathbb{R}.$$

Function $H(\xi; \gamma)$ is defined the same way, by replacing v with γ. Then, we must have

$$H(\xi; v) \geq H(\xi; \gamma), \quad \forall \xi. \tag{18}$$

Indeed, consider any two edges $(\ell, j) \neq (\ell', j)$ with a common j, and suppose $\Delta_{\ell' j} < \Delta_{\ell j}$. Then, in a fixed sufficiently small time interval $[t, t + \epsilon]$, for all pre-limit trajectories with sufficiently large parameter r, any j-customer whose service completes at a server with configuration ℓ' cannot possibly be placed along the edge (ℓ, j). (Here we use the fact that the system is closed: when a customer service is completed, there is always an option to place it back into the same server for the new service.) Furthermore, if either $\ell' = e_j$ or $x_{\ell' - e_j} > 0$, any j-customer completing service in configuration ℓ, will be placed for new service either along the edge (ℓ', j) or possibly along another edge (ℓ'', j) with $\Delta_{\ell'' j} \leq \Delta_{\ell' j}$. This proves (18), from which (17) easily follows. □

As a corollary of Lemma 8 we obtain

Lemma 9. *Any FSP (for the closed system under Greedy-D) is such that*

$$x(t) \to x^*. \tag{19}$$

The convergence is uniform across all initial states $x(0) \in \mathcal{X}$.

Remainder of the proof of Theorem 2. We fix $\epsilon > 0$ and choose $T > 0$ large enough so that, for any FSP we have $\|x(T) - x^*\| \le \epsilon$. We claim that, as $r \to \infty$,

$$\mathbb{P}\{\|x^r(T) - x^*\| > 2\epsilon\} \to 0, \tag{20}$$

where the convergence is uniform across all initial states $x^r(0)$. This claim is true, because for an arbitrary sequence of fixed initial states $x^r(0)$, we must have

$$\limsup_{r\to\infty} \|x^r(T) - x^*\| \le \epsilon, \quad \text{w.p.1},$$

because w.p.1 we can always choose a subsequence of r along which the u.o.c convergence to an FSP holds. Claim (20) implies the result, because ϵ can be arbitrarily small. □

6 Original system. Optimality of a modified version of Greedy-D

We now return to our original "open" system. Unfortunately, the proof of Greedy-D optimality in the closed system does not carry over to the open system. The key reason can be informally described as follows. If $v(t)$ is the exogenous arrival rate allocation vector in the fluid limit, then the result analogous to Lemma 8 no longer holds; namely, property (18) in its proof is not valid. Suppose we have an edge (ℓ, j) on which the unique minimum $\min_{\ell'} \Delta_{\ell' j}$ is attained, $x_{\ell - e_j} = 0$, $x_\ell > 0$ and $\ell_j > 0$. There is the non-zero rate $\mu_j \ell_j x_\ell$ of departures along (ℓ, j). In this case it is possible that $v_{\ell j} < \mu_j \ell_j x_\ell$, because some type j exogenous arrivals will find no servers in configuration $\ell - e_j$. (In the closed system we must have $v_{\ell j} \ge \mu_j \ell_j x_\ell$, because all departures along (ℓ, j) have the option of "coming back" along the same edge.) Therefore, the argument leading to (18), and in fact the property itself, does not hold.

In this section we will prove the asymptotic optimality of a slightly modified version of the Greedy-D algorithm, called Greedy-DM, which, in a sense, "emulates" the behavior of Greedy-D in the corresponding closed system. Informally the key idea of Greedy-DM is to make decisions about placements of new exogenous type i arrivals a little "in advance", at the instants of type i departures. This is achieved by using "placeholders," called tokens: when a type i departure occurs, we "pretend" that immediately a new type i arrival occurs, decide which server this arrival would be placed into according to the Greed-D rule, and place a type i token in that server. When new actual type i customers arrive, they first seek and replace type i tokens if there are any; if no type i token is available, the customer is placed according to the Greedy-D. (In addition to being replaced by actual customers, we make the tokens "impatient" – they leave the system by themselves after a random exponentially distributed time.) The analysis in this section shows that, in the fluid limit, "all" tokens are replaced by actual arrivals and "all" actual arrivals replace tokens. This means that the allocation of actual customer arrival rates is "equal" to that of tokens; the latter, in turn, satisfies same properties as the rate allocation in the closed system.

Definition 10 (Greedy-DM discipline)**.** *Suppose the weights $W_k(x) = x_k^\alpha$ are given, as in the definition of standard Greedy-D. At any given time there are two kinds of type i customers – actual customers being served as usual and* tokens. *For the purposes of defining server configurations k and the system state X, type i tokens are treated the same way as actual type i customers.*
When a departure of actual type i customer from configuration k at time t occurs, the following actions are taken: a new token of type i is immediately created; this new token is treated the same way as a new type i arrival, and is placed for "service" immediately; the token is added to an available configuration k (with either $k = 0$ or $X_k(t-) > 0$) such that $k + e_i \in \mathcal{K}$ and the difference $W_{k+e_i}(X(t-)) - I\{k \ne 0\} W_k(X(t-))$

is the smallest (where $t-$ refers to the time after the service completion, but before the token placement). When a new exogenous arrival of an (actual) type i customer occurs, it replaces an arbitrarily chosen type i token (anywhere in the system), if such is available; otherwise, this arrival is added (as in the usual Greedy-D) to an available configuration k (with either $k = 0$ or $X_k(t-) > 0$) such that $k + e_i \in \mathcal{K}$ and the difference $W_{k+e_i}(X(t-)) - I\{k \neq 0\}W_k(X(t-))$ is the smallest (where $t-$ refers to the time just before the arrival). Any token of any type anywhere in the system "completes service" at the rate $\mu_0 > 0$, independently of anything else (i.e. the probability of service completion in a dt-long interval is $\mu_0 dt + o(dt)$); "service completion" of any token is treated the same way as service completion of an actual customer, except no new token is created.

The random process, describing the evolution of this system is more complex, but it is still an irreducible Markov chain. A *complete server configuration* is by definition a pair $(k, \hat{k})$, where vector $k = (k_1, \ldots, k_I) \in \mathcal{K}$ gives the numbers all customers (both actual and tokens) in a server, while vector $\hat{k} \leq k$, $k \in \bar{\mathcal{K}}$, gives the number of actual customers only. Therefore, the Markov process state at time t is the vector $\{X_{(k,\hat{k})}(t)\}$, where the index $(k, \hat{k})$ takes values that are all possible complete server configurations, as described above. Obviously, $\hat{Y}_i^r(t) \leq Y_i^r(t)$ for all i and t, where $Y_i^r(t)$ and $\hat{Y}_i^r(t)$ is the total number of all and actual type i customers, respectively, and superscript r, as usual, indicates the system with parameter r. Moreover, the behaviors of the processes $(Y_i^r(t), \hat{Y}_i^r(t))$, $t \geq 0$, are independent across all i, with $\hat{Y}_i^r(\infty)$ having Poisson distribution with mean $\rho_i r$. Finally, by the Greedy-DM definition, at any time any existing type i token "completes service" at the rate μ_0. Using these facts, we easily establish the following

Lemma 11. *Markov chain $\{X^r_{(k,\hat{k})}(t)\}$, $t \geq 0$, is positive recurrent for each r. Moreover, the distributions of $\{(\hat{y}_i^r(\infty), y_i^r(\infty)),\ i \in \mathcal{I}\} = (1/r)\{(\hat{Y}_i^r(\infty), Y_i^r(\infty)),\ i \in \mathcal{I}\}$ are tight, and any limit in distribution $\{(\hat{y}_i(\infty), y_i(\infty)),\ i \in \mathcal{I}\}$ is such that $\hat{y}_i(\infty) = \rho_i$ and $y_i(\infty) \leq \rho_i + \lambda_i/\mu_0$ for all i. Consequently, the distributions of $\{x^r_{(k,\hat{k})}(\infty)\} = (1/r)\{X^r_{(k,\hat{k})}(\infty)\}$ are tight.*

Proof. Consider the process $(Y_i^r(\cdot), \hat{Y}_i^r(\cdot))$ for a fixed r and i. Consider a different process $(\bar{Y}_i^r(\cdot), \hat{Y}_i^r(\cdot))$, defined as follows: actual type i customers arrive the same way, and depart after their service completion, so that $\hat{Y}_i^r(t)$ is same as before; upon every service completion of an actual i-customer, one type i token is created, which then stays in the system for a random time, exponentially distributed with mean $1/\mu_0$, independently of anything else, and then leaves the system; $\bar{Y}_i^r(t)$ is the total number of all type i customers (actual and tokens) in the system. Obviously, processes $(Y_i^r(\cdot), \hat{Y}_i^r(\cdot))$ and $(\bar{Y}_i^r(\cdot), \hat{Y}_i^r(\cdot))$ can be constructed on a common probability space so that $Y_i^r(t) \leq \bar{Y}_i^r(t)$. But, $\bar{Y}_i^r(t)$ is simply the number of customers in the infinite-server system with input rate $\lambda_i r$ and mean service time $1/\mu_i + 1/\mu_0$. Therefore, $(\bar{Y}_i^r(\cdot), \hat{Y}_i^r(\cdot))$ is positive recurrent; moreover, in stationary regime, $\bar{Y}_i^r(\infty)$ and $\hat{Y}_i^r(\infty)$ have Poisson distributions with means $\lambda_i r(1/\mu_i + 1/\mu_0)$ and $\rho_i r$, respectively. The lemma results easily follow. □

Let $\{x_{(k,\hat{k})}\}$ denote a vector with non-negative components, with indices $(k, \hat{k})$ being all possible complete configurations. Denote by $x = \{x_k\}$, $y = \{y_i\}$ and $\hat{y} = \{\hat{y}_i\}$ its projections, with components being

$$x_k = \sum_{\hat{k}\,:\,\hat{k} \leq k} x_{(k,\hat{k})}, \quad y_i = \sum_{(k,\hat{k})} k_i x_{(k,\hat{k})} = \sum_k k_i x_k, \quad \hat{y}_i = \sum_{(k,\hat{k})} \hat{k}_i x_{(k,\hat{k})}.$$

Denote by $\bar{\mathcal{X}}$ the set of those values of $\{x_{(k,\hat{k})}\}$ satisfying condition

$$y_i = \hat{y}_i = \rho_i, \quad \forall i.$$

Obviously, for any $\{x_{(k,\hat{k})}\} \in \bar{\mathcal{X}}$, its x-projection is an element of $\mathcal{X}$. Also, note that the condition $y_i = \hat{y}_i$, $\forall i$, is equivalent to

$$x_{(k,\hat{k})} = 0 \ \text{ unless } \ \hat{k} = k, \tag{21}$$

and therefore (21) holds for any $\{x_{(k,\hat{k})}\} \in \bar{\mathcal{X}}$.

The main result of this section is

Theorem 12. *Consider a sequence of original systems, indexed by r, under the Greedy-DM algorithm with $\alpha > 0$. Let $\{x^r_{(k,\hat{k})}(\infty)\}$ denote the random (complete) state of the fluid-scaled process in a stationary regime. Then, as $r \to \infty$,*

$$d(\{x^r_{(k,\hat{k})}(\infty)\}, \bar{\mathcal{X}}) \implies 0, \tag{22}$$

$$x^r(\infty) \implies x^*, \tag{23}$$

where x^* is defined in (7).

For each $(k,i) \in \mathcal{M}$ and $\hat{k} \le k$, consider independent unit-rate Poisson processes $\hat{\Pi}_{(k,\hat{k}),i}(t),\ t \ge 0$, and $\tilde{\Pi}_{(k,\hat{k}),i}(t),\ t \ge 0$; and for each $i \in \mathcal{I}$ – an independent unit-rate Poisson process $\hat{\Pi}_i(t),\ t \ge 0$. Each of these processes satisfies the functional strong law of large numbers (FSLLN) analogous to (8).

The Markov process $\{X^r_{(k,\hat{k})}(\cdot)\}$ for each r is driven by the common set of independent Poisson processes $\hat{\Pi}_{(k,\hat{k}),i}(\cdot)$, $\tilde{\Pi}_{(k,\hat{k}),i}(\cdot)$ and $\hat{\Pi}_i(\cdot)$, in the natural way, as follows:

$$\hat{D}^r_{(k,\hat{k}),i}(t) = \hat{\Pi}_{(k,\hat{k}),i}(\int_0^t X^r_{(k,\hat{k})}(\xi)\hat{k}_i\mu_i d\xi),$$

$$\tilde{D}^r_{(k,\hat{k}),i}(t) = \tilde{\Pi}_{(k,\hat{k}),i}(\int_0^t X^r_{(k,\hat{k})}(\xi)(k_i - \hat{k}_i)\mu_0 d\xi),$$

$$\hat{A}^r_i(t) = \hat{\Pi}_i(r\lambda_i t),$$

where $\hat{D}^r_{(k,\hat{k}),i}(t)$ and $\tilde{D}^r_{(k,\hat{k}),i}(t)$ is the number of the type-i departures from the configuration $(k,\hat{k})$, due to the service completions of actual customers and tokens, respectively, and $\hat{A}^r_i(t)$ is the number of the exogenous type-i arrivals of actual customers; all in the interval $[0,t]$. Clearly, the entire process sample path is (w.p.1) uniquely determined by the initial state $\{X^r_{(k,\hat{k})}(0)\}$, the realizations of the driving Poisson processes and the Greedy-DM discipline. In particular, the realizations of the following processes are uniquely determined: the number of type-i departures from configuration k due to actual and token service completions, and their total

$$\hat{D}^r_{k,i}(t) = \sum_{\hat{k}} \hat{D}^r_{(k,\hat{k}),i}(t),\ \tilde{D}^r_{k,i}(t) = \sum_{\hat{k}} \tilde{D}^r_{(k,\hat{k}),i}(t),\ D^r_{k,i}(t) = \hat{D}^r_{k,i}(t) + \tilde{D}^r_{k,i}(t);$$

the number of type-i token "arrivals" allocated (upon type-i actual departures) along edge (k,i), $\tilde{A}^r_{k,i}(t)$; the number of type-i actual exogenous arrivals allocated along edge (k,i), *without replacing an existing i-token*, $\hat{A}^{**,r}_{k,i}(t)$ (such arrivals change the complete configuration from $(k-e_i, \hat{k}-e_i)$ to $(k,\hat{k})$; the number of type-i actual exogenous arrivals, *that replace an existing token in configuration* k, $\hat{A}^{*,r}_{k,i}(t)$ (such arrivals change the complete configuration from $(k, \hat{k}-e_i)$ to $(k,\hat{k})$ – so these do *not* count as arrivals into k); the total number of all type-i arrivals into configuration k, $A^r_{k,i}(t)(t) = \tilde{A}^r_{k,i}(t) + \hat{A}^{**,r}_{k,i}(t)$. The following relations obviously hold:

$$\hat{A}^r_i(t) = \sum_{k:(k,i)\in\mathcal{M}} [\hat{A}^{*,r}_{k,i}(t) + \hat{A}^{**,r}_{k,i}(t)], \quad \forall i,$$

$$\sum_{k:(k,i)\in\mathcal{M}} \tilde{A}^r_{k,i}(t) = \sum_{k:(k,i)\in\mathcal{M}} \hat{D}^r_{k,i}(t), \quad \forall i. \tag{24}$$

We introduce fluid-scaled processes:

$$x^r_k(t) = \frac{1}{r}X^r_k(t), \quad x^r_{(k,\hat{k})}(t) = \frac{1}{r}X^r_{(k,\hat{k})}(t), \quad \hat{a}^r_i(t) = \frac{1}{r}\hat{A}^r_i(t),$$

and similarly defined

$$\hat{d}^r_{ki}(t),\ \tilde{d}^r_{ki}(t),\ d^r_{ki}(t),\ \ \tilde{a}^r_{ki}(t),\ \hat{a}^{*,r}_{ki}(t),\ \hat{a}^{**,r}_{ki}(t),\ a^r_{ki}(t).$$

A set of Lipschitz continuous functions $[\{x_k(\cdot)\}, \{x_{(k,\hat{k})}(\cdot)\}, \{\hat{a}_i(\cdot)\}, \{\hat{d}_{ki}(\cdot)\}, \ldots]$ on the time interval $[0, \infty)$ we call a *fluid sample path* (FSP), if there exist realizations of driving Poisson processes, satisfying the FSLLN analogous to (8), and a fixed subsequence of r, along which

$$[\{x^r_k(\cdot)\}, \{x^r_{(k,\hat{k})}(\cdot)\}, \{\hat{a}^r_i(\cdot)\}, \{\hat{d}^r_{ki}(\cdot)\}, \ldots] \to$$
$$[\{x_k(\cdot)\}, \{x_{(k,\hat{k})}(\cdot)\}, \{\hat{a}_i(\cdot)\}, \{\hat{d}_{ki}(\cdot)\}, \ldots], \quad u.o.c. \tag{25}$$

Lemma 13. *Suppose the initial states $\{x^r_{(k,\hat{k})}(0)\}$ are fixed and are such that $\{x^r_{(k,\hat{k})}(0)\} \to \{x_{(k,\hat{k})}(0)\}$. Then, w.p.1 for any subsequence of r there exists a further subsequence of r, along which the convergence (25) holds, where the limit is an FSP.*

Proof is, again, very standard – it is a more general version of that of Lemma 3. We omit details. □

Lemma 14. *Consider an FSP with the initial state $\{x_{(k,\hat{k})}(0)\}$. Recall notation:*

$$y_i(t) = \sum_{(k,\hat{k})} k_i x_{(k,\hat{k})}(t), \quad \hat{y}_i(t) = \sum_{(k,\hat{k})} \hat{k}_i x_{(k,\hat{k})}(t),$$

and denote $\tilde{y}_i(t) = y_i(t) - \hat{y}_i(t)$. Then, at any regular point t, for any i,

$$(d/dt)\hat{y}_i(t) = \lambda_i - \mu_i \hat{y}_i(t), \tag{26}$$

$$(d/dt)\tilde{y}_i(t) = \begin{cases} -\lambda_i + \mu_i \hat{y}_i(t) - \mu_0 \tilde{y}_i(t), & \text{if } \tilde{y}_i(t) > 0 \\ \max\{0, -\lambda_i + \mu_i \hat{y}_i(t) - \mu_0 \tilde{y}_i(t)\}, & \text{if } \tilde{y}_i(t) = 0 \end{cases} \tag{27}$$

In particular, for any i, the convergence

$$(\hat{y}_i(t), \tilde{y}_i(t)) \to (\rho_i, 0), \quad \forall i, \tag{28}$$

holds and is uniform in initial states $\{x_{(k,\hat{k})}(0)\}$ from a compact set, and

$$(\hat{y}_i(0), \tilde{y}_i(0)) = (\rho_i, 0) \quad \text{implies} \quad (\hat{y}_i(t), \tilde{y}_i(t)) = (\rho_i, 0), \ \forall t. \tag{29}$$

Proof. Equation (26) is very standard, describing an FSP for an $M/M/\infty$ system. Equation (27) for the $\tilde{y}_i(t) > 0$ case is also a very basic; (27) for the $\tilde{y}_i(t) = 0$ case is easily verified by considering the behavior of pre-limit trajectories is a small interval $[t, t+\Delta t]$, and considering the three cases, $-\lambda_i + \mu_i \hat{y}_i(t) - \mu_0 \tilde{y}_i(t) < 0$, $= 0$ and > 0, separately. (See e.g. [8] for this type of argument in more detail.) We omit details. □

Lemma 15. *Consider a sequence of original systems, indexed by r, under the Greedy-DM algorithm. Let $\{x^r_{(k,\hat{k})}(\infty)\}$ denote the random complete state of the fluid-scaled process in a stationary regime. Then, any subsequence of r has a further subsequence, such that*

$$\{x^r_{(k,\hat{k})}(\infty)\} \implies \{x_{(k,\hat{k})}(\infty)\},$$

where $\{x_{(k,\hat{k})}(\infty)\} \in \bar{\mathcal{X}}$ w.p.1.

Proof. Fix arbitrary $\delta > 0$, and a sufficiently large compact B so that for all large r,

$$\mathbb{P}[\{x^r_{(k,\hat{k})}(\infty)\} \in B] \geq 1 - \delta. \tag{30}$$

(We can do that by Lemma 11.) Fix arbitrary $\epsilon > 0$ and choose $T > 0$ large enough so that for any FSP with $\{x_{(k,\hat{k})}(0)\} \in B$, we have $d(\{x_{(k,\hat{k})}(T)\}, \bar{\mathcal{X}}) \leq \epsilon$. (We can do that by Lemma 14.) Fix arbitrary $\delta_1 > 0$. We claim that for all sufficiently large r,

$$\{x^r_{(k,\hat{k})}(0)\} \in B \quad \text{implies} \quad \mathbb{P}\{d(\{x^r_{(k,\hat{k})}(T)\}, \bar{\mathcal{X}}) > 2\epsilon\} \leq \delta_1. \tag{31}$$

This claim is true, because for an arbitrary sequence of fixed initial states $\{x^r_{(k,\hat{k})}(0)\} \in B$, we must have

$$\limsup_{r\to\infty} d(\{x_{(k,\hat{k})}(T)\}, \bar{\mathcal{X}}) \leq \epsilon, \quad \text{w.p.1.}$$

(This follows from Lemma 13.) By (30) and (31), for all large r, a stationary version of the process is such that

$$\mathbb{P}\{d(\{x^r_{(k,\hat{k})}(T)\}, \bar{\mathcal{X}}) \leq 2\epsilon\} \geq (1-\delta)(1-\delta_2).$$

Therefore, any limit-in-distribution $\{x_{(k,\hat{k})}(T)\}$ is such that

$$\mathbb{P}\{d(\{x_{(k,\hat{k})}(T)\}, \bar{\mathcal{X}}) \leq 2\epsilon\} \geq \limsup \mathbb{P}\{d(\{x^r_{(k,\hat{k})}(T)\}, \bar{\mathcal{X}}) \leq 2\epsilon\} \geq (1-\delta)(1-\delta_2).$$

Since $\delta, \delta_2, \epsilon$ are arbitrary positive, $\mathbb{P}[\{x_{(k,\hat{k})}(T)\} \in \bar{\mathcal{X}}] = 1$. □

Lemma 16. *Consider an FSP with the initial state* $(x_{(k,\hat{k})}(0)) \in \bar{\mathcal{X}}$. *(In particular,* $x(0) \in \mathcal{X}$.*) Then* $(x_{(k,\hat{k})}(t)) \in \bar{\mathcal{X}}$ *for all* $t \geq 0$. *In addition, at any regular point* t*, using notation* $w_{ki}(t) = (d/dt)d_{ki}(t)$, $\hat{w}_{ki}(t) = (d/dt)\hat{d}_{ki}(t)$, $v_{ki}(t) = (d/dt)a_{ki}(t)$, $\tilde{v}_{ki}(t) = (d/dt)\tilde{a}_{ki}(t)$, *we have*

$$w_{ki}(t) = \hat{w}_{ki}(t) = k_i\mu_i x_k, \quad \forall (k,i) \in \mathcal{M}, \tag{32}$$

$$\sum_{k:(k,i)\in\mathcal{M}} \hat{w}_{ki}(t) = \sum_{k:(k,i)\in\mathcal{M}} \tilde{v}_{ki}(t) = \lambda_i, \quad \forall i \in \mathcal{I}, \tag{33}$$

$$v_{ki}(t) = \tilde{v}_{ki}(t), \quad \forall (k,i) \in \mathcal{M}, \tag{34}$$

$$(d/dt)x_k(t) = [\sum_{i:k-e_i\in\bar{\mathcal{K}}} v_{ki} - \sum_{i:k+e_i\in\mathcal{K}} v_{k+e_i,i}] - [\sum_{i:k-e_i\in\bar{\mathcal{K}}} w_{ki} - \sum_{i:k+e_i\in\mathcal{K}} w_{k+e_i,i}], \quad \forall k \in \mathcal{K}, \tag{35}$$

$$\frac{d}{dt}F(x(t)) = D(v(t), x(t)) \leq D_{min}(x(t)). \tag{36}$$

Proof. Since we have (29), $\{x_{(k,\hat{k})}(t)\} \in \bar{\mathcal{X}}$ holds by definition. Relation (32) holds because

$$\hat{w}_{ki}(t) = \sum_{\hat{k}\leq k} \hat{k}_i\mu_i x_{(k,\hat{k})}(t), \quad w_{ki}(t) - \hat{w}_{ki}(t) = \sum_{\hat{k}\leq k} (k_i - \hat{k}_i)\mu_0 x_{(k,\hat{k})}(t),$$

and $\{x_{(k,\hat{k})}(t)\} \in \bar{\mathcal{X}}$. We obtain (33) from the limit form of (24), namely

$$\sum_{k:(k,i)\in\mathcal{M}} \tilde{a}_{k,i}(t) = \sum_{k:(k,i)\in\mathcal{M}} \hat{d}_{k,i}(t),$$

and from $\sum_k w_{ki}(t) = \sum_k \hat{w}_{ki}(t) = \mu_i \hat{y}_i(t) = \lambda_i$. Relation (34) follows from the fact that $v_{ki}(t) \geq \tilde{v}_{ki}(t)$, and the strict inequality cannot hold for any (k,i), because otherwise we would have for at least one i

$$(d/dt)y_i(t) = \sum_k v_{ki}(t) - \sum_k w_{ki}(t) > \sum_k \tilde{v}_{ki}(t) - \sum_k \hat{w}_{ki}(t) = 0.$$

Equation (35) is automatic: the RHS is just the difference between arrival and departure rates to/from configuration k. Finally, since $v(t) = \tilde{v}(t)$, and the rates $\tilde{v}_{ki}(t)$ are those of "arriving" tokens (which immediately follow service completions of actual customers), the argument in the proof of Lemma 17 (for the closed system) applies, and we obtain (36). □

As a corollary we obtain the analog of Lemma 9.

Lemma 17. *Consider an FSP (for the original system under Greedy-DM algorithm) with the initial state* $\{x_{(k,\hat{k})}(0)\} \in \bar{\mathcal{X}}$. *(In particular,* $x(0) \in \mathcal{X}$.*) Then*

$$x(t) \to x^*. \tag{37}$$

The convergence is uniform across all initial states in $\bar{\mathcal{X}}$.

Proof of Theorem 12. Convergence (22) has already been proved in Lemma 15. We fix $\epsilon > 0$ and choose $T > 0$ large enough so that for any FSP with $\{x_{(k,\hat{k})}(0)\} \in \bar{\mathcal{X}}$, we have $\|x(T) - x^*\| \leq \epsilon$. We claim that for any $\delta_1 > 0$ there exists a sufficiently small $\delta_2 > 0$ such that for all sufficiently large r,

$$d(\{x^r_{(k,\hat{k})}(0)\}, \bar{\mathcal{X}}) \leq \delta_2 \text{ implies } \mathbb{P}\{\|x^r(T) - x^*\| > 2\epsilon\} \leq \delta_1, \tag{38}$$

This claim is true, because for an arbitrary sequence of fixed initial states $\{x^r_{(k,\hat{k})}(0)\} \to \bar{\mathcal{X}}$, we must have

$$\limsup_{r\to\infty} \|x^r(T) - x^*\| \leq \epsilon, \quad \text{w.p.1.}$$

Constants ϵ and δ_1 can be arbitrarily small; we also know that for any δ_2, $\mathbb{P}\{d(\{x^r_{(k,\hat{k})}(\infty)\}, \bar{\mathcal{X}}) \leq \delta_2\} \to 1$ as $r \to \infty$. Therefore, claim (38) implies (23). □

7 The case of vector-packing constraints (1): Greedy algorithm with aggregate configurations

So far in the paper we did not exploit a possible underlying structure of packing constraints. Instead, we worked with a formally defined set $\bar{\mathcal{K}}$ of possible configurations. Now we will consider a special case: suppose the configuration set $\bar{\mathcal{K}}$ is defined by vector packing constraints (1).

We say that two configurations k and k' are equivalent if they require same total amounts of resources of each type:

$$\sum_i k_i b_{i,n} = \sum_i k'_i b_{i,n}, \quad \forall n. \tag{39}$$

A class of equivalent configurations is denoted q; we will call it *aggregate configuration*, or *a-configuration*. Zero a-configuration, denoted (with some notation abuse) by $q = 0$, is the one containing the sole configuration $k = 0$; by convention $X_0 = 0$, where the subscript 0 can refer to either zero configuration or zero a-configuration. The sets of all a-configurations and non-zero a-configurations are denoted by $\bar{\mathcal{Q}}$ and $\mathcal{Q}$, respectively. We write $q(k)$ for the aggregate configuration containing k. We use notation

$$X_q = \sum_{k\in q} X_k,$$

and similarly for other quantities summed up over an a-configuration q. Clearly, vector $\{X_q,\ q \in \mathcal{Q}\}$ is a projection of $X = \{X_k,\ k \in \mathcal{K}\}$.

In this section we show that (versions of) the Greedy algorithm, using quantities X_q instead of X_k, asymptotically minimizes

$$\sum_{q\in\mathcal{Q}} X_q^{1+\alpha}(\infty),$$

which, again, approximates (when $\alpha > 0$ is small) the total number of occupied servers $\sum_q X_q(\infty) \equiv \sum_k X_k(\infty)$ in a stationary regime. The difference, however, is that $|\mathcal{Q}|$ can be much smaller than $|\mathcal{K}|$, thus making the Greedy algorithm easier to implement in practice.

7.1 Results.

For $x \in \mathbb{R}_+^{|\mathcal{K}|}$ consider function

$$\Phi(x) = \sum_q (1+\alpha)^{-1} x_q^{1+\alpha} \equiv \sum_q (1+\alpha)^{-1} [\sum_{k\in q} x_k]^{1+\alpha}$$

with parameter $\alpha > 0$, and the following the convex optimization problem

$$\min_{x \in \mathbb{R}_+^{|\mathcal{K}|}} \Phi(x) \tag{40}$$

subject to

$$\sum_{k \in \mathcal{K}} k_i x_k = \rho_i, \quad \forall i. \tag{41}$$

We denote by $\mathcal{X}^*$ the set of optimal solutions of this problem; obviously, $\mathcal{X}^* \subseteq \mathcal{X}$.

Definition 18 (Greedy discipline with a-configurations)**.**

1. Integral form (Greedy-I-AC). *A type i customer arriving at time t is added to an available a-configuration q (with either $q = 0$ or $X_q(t-) > 0$) such that the addition does not violate the vector packing constraints and the increment $\Phi(X(t)) - \Phi(X(t-))$ is the smallest. The ties* between a-configurations *are broken according to an arbitrary deterministic rule. The choice of a server* within *the chosen a-configuration is random uniform.*

2. Differential form (Greedy-D-AC). *For each $q \in \mathcal{Q}$, denote $W_q(x) = (\partial/\partial x_q)\Phi(x) = x_q^\alpha$, $x \in \mathbb{R}_+^{|\mathcal{K}|}$. A type i customer arriving at time t is added to an available configuration k (with either $q = 0$ or $X_q(t-) > 0$) such that the addition does not violate the vector packing constraints and the difference $W_{q+e_i}(X(t-)) - I\{q \neq 0\} W_q(X(t-))$ is the smallest. [Here $q+e_i$ denotes the a-configuration containing configurations $k + e_i$, $k \in q$ (and possibly other configurations).] The ties* between a-configurations *are broken according to an arbitrary deterministic rule. The choice of a server* within *the chosen a-configuration is random uniform.*

Theorem 19. *Consider a sequence of closed systems, indexed by r, and let $x^r(\infty)$ denote the random state of the (fluid-scaled) process is a stationary regime, under the Greedy-D-AC discipline with $\alpha > 0$. Then, as $r \to \infty$,*

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

Definition 20 (Greedy-DM-AC discipline)**.** *This discipline, which uses tokens, is the modification of Greedy-D-AC, completely analogous to the modification of Greedy-D that leads to Greedy-DM.*

Theorem 21. *Consider a sequence of original systems, indexed by r, under the Greedy-DM-AC algorithm with $\alpha > 0$. Let $\{x^r_{(k,\hat{k})}(\infty)\}$ denote the random (complete) state of the fluid-scaled process in a stationary regime. Then, as $r \to \infty$,*

$$d(\{x^r_{(k,\hat{k})}(\infty)\}, \bar{\mathcal{X}}) \implies 0,$$
$$d(x^r(\infty), \mathcal{X}^*) \implies 0.$$

In the rest of this section we will consider the closed system under Greedy-D-AC, and will prove Theorem 19. We will omit the proof of Theorem 21 which is "obtained from" that of Theorem 19 in exactly same way as the proof of Theorem 12 was obtained from that of Theorem 2.

7.2 Optimal set characterization and related properties.

Using Lagrange multipliers η_i for the constraints (41), the Lagrangian of the problem (40)-(41) is

$$\sum_q \frac{1}{1+\alpha} [\sum_{k \in q} x_k]^{1+\alpha} + \sum_i \eta_i [\rho_i - \sum_k k_i x_k].$$

We obtain the following characterization: vector $x \in \mathcal{X}$ is an optimal solution of (40)-(41) (i.e. $x \in \mathcal{X}^*$) if and only if there exist constants η_i such that (using notation $u_k = \max\{\sum_i k_i \eta_i, 0\}$)

$$x_q^\alpha = \max_{k \in q} u_k, \quad \forall q \in \mathcal{Q}, \tag{42}$$

$$\left(u_k < \max_{k' \in q(k)} u_{k'} \quad \text{imples} \quad x_k = 0\right), \quad \forall k \in \mathcal{K}. \tag{43}$$

More notation. Consider the following order relation on $\bar{\mathcal{Q}}$: $q' \leq q$ if $k' \leq k$ for some $k' \in q'$ and $k \in q$. $q' < q$ means $q' \leq q$ and $q' \neq q$. If a-configuration q contains at least one k with $k_i > 0$, we use a (slightly abusive) notation $q - e_i$ for the a-configuration containing $k - e_i$, i.e. $q - e_i \doteq \{k - e_i \mid k \in q, k_i > 0\}$; otherwise, $q - e_i = \emptyset$. Denote by $\mathcal{M}^a$ the set of pairs (q, i) such that $q - e_i \neq \emptyset$. For $x \in \mathbb{R}_+^{|\mathcal{K}|}$ and $(q, i) \in \mathcal{M}^a$ denote

$$\Delta_{qi} = \Delta_{qi}(x) = x_q^\alpha - x_{q-e_i}^\alpha.$$

Lemma 22. *Consider the following property of an element $x \in \mathcal{X}$. (We will refer to it as* NSI-property *– "No Simple Improving allocation"). For any two elements $(q, i), (q', i) \in \mathcal{M}^a$ (with $q \neq q'$, but a common i), condition*

$$\Delta_{q'i} < \Delta_{qi} \tag{44}$$

implies either

$$x_k = 0 \quad \text{for all } k \in q \text{ such that } k_i > 0 \tag{45}$$

or

$$q' - e_i \neq 0 \quad \text{and} \quad x_{q'-e_i} = 0. \tag{46}$$

If $x \in \mathcal{X}$ satisfies the NSI-property, then condition (42) holds.

Proof. Consider $x \in \mathcal{X}$ satisfying the NSI-property. For each i define

$$\underline{\xi}_i = \min_{k:\ k_i > 0,\ x_k > 0} \Delta_{q(k),i}, \quad \bar{\xi}_i = \max_{k:\ k_i > 0,\ x_k > 0} \Delta_{q(k),i}.$$

It is easy to check that we cannot have $\bar{\xi}_i > 0$ and $\underline{\xi}_i \leq 0$, because this would violate the NSI-property. Then, we can further define

$$\eta_i = \begin{cases} \bar{\xi}_i, & \text{if } \bar{\xi}_i > 0 \\ \underline{\xi}_i, & \text{if } \underline{\xi}_i \leq 0 \end{cases} \tag{47}$$

Denote by $\mathcal{I}^+$ the subset of those i with $\eta_i > 0$, and by $\mathcal{I}^- = \mathcal{I} \setminus \mathcal{I}^+$ the remaining subset. It is easy to check that for any fixed $i \in \mathcal{I}^-$, we must have

$$\Delta_{q(k),i} = \eta_i \quad \text{for all } k \text{ such that } k_i > 0,\ x_k > 0, \tag{48}$$

otherwise a contradiction to NSI-property is obtained.

Using notation $u_k = \max\{\sum_i k_i \eta_i, 0\}$, let us define the values x_q^0 via

$$[x_q^0]^\alpha = \max_{k \in q} u_k.$$

It is easy to check that

$$[x_q^0]^\alpha - [x_{q-e_i}^0]^\alpha \geq \eta_i, \quad \forall (q, i) \in \mathcal{M}^a. \tag{49}$$

Let us prove (42), which is equivalent to $x_q = x_q^0$, $\forall q$. Suppose this is not true. Consider a minimal q for which $x_q \neq x_q^0$.
Case (c1): suppose $x_q > x_q^0$. Then necessarily $x_q > 0$. Consider any $k \in q$ with $x_k > 0$.
Sub-case (c1.1): suppose $k_i > 0$ for some $i \in \mathcal{I}^-$. Fix this i and denote $q' = q - e_i$. If $q' = 0$, we obtain a contradiction to the definition of η_i, and so $q' \neq 0$ must hold. Then $x_{q'} = x_{q'}^0$ by definition of q (as a minimal counterexample). Using (49) we obtain $x_q^\alpha - x_{q'}^\alpha = \Delta_{q,i} > \eta_i$ – a contradiction to (48). Thus sub-case (c1.1) is impossible.
Sub-case (c1.2): suppose $k_i > 0$ implies $i \in \mathcal{I}^+$ and then $\eta_i > 0$. Fix an i with $k_i > 0$, and consider $q' = q - e_i$. If $q' = 0$, we obtain a contradiction to the definition of η_i, and so $q' \neq 0$ must hold. We have $x_{q'} = x_{q'}^0$ by definition of q; therefore, using (49), $x_q^\alpha - x_{q'}^\alpha = \Delta_{q,i} > \eta_i$ – a contradiction with the definition

of η_i. Sub-case (c1.2), and then case (c1), is impossible.
Case (c2): suppose $x_q < x_q^0$. Then necessarily $x_q^0 > 0$. Let us fix a $k \in q$, on which the $\max_{k\in q} u_k > 0$ is attained, and so $u_k > 0$.
Sub-case (c2.1): suppose $k_i > 0$ for some $i \in \mathcal{I}^-$. Fix this i and denote $q' = q - e_i$. We cannot have $q' = 0$, because that would imply $\eta_i > 0$. Therefore, $q' \neq 0$ and $[x_q^0]^\alpha - [x_{q'}^0]^\alpha = \eta_i$, implying in particular $x_{q'}^0 > 0$. But, $x_{q'} = x_{q'}^0$ and therefore $x_q^\alpha - x_{q'}^\alpha = \Delta_{q,i} < \eta_i$. Recalling the definition of η_i, we obtain a contradiction to NSI-property. Thus, sub-case (c2.1) is impossible.
Sub-case (c2.2): suppose $k_i > 0$ implies $i \in \mathcal{I}^+$ and then $\eta_i > 0$. Fix an i with $k_i > 0$, and consider $q' = q - e_i$. If $q' = 0$, we have $\Delta_{q,i} < \eta_i$ – a contradiction to NSI-property. Therefore, $q' \neq 0$ and $x_{q'} = x_{q'}^0 > 0$ (because $\eta_i > 0$ for all i with $k_i > 0$). Then, $\Delta_{q,i} < \eta_i$ and we, again, obtain a contradiction to NSI-property. Sub-case (c2.2), and then case (c2), is impossible.
The proof of (42) is complete. □

7.3 Fluid sample paths.

We now define fluid sample paths for the closed system under Greedy-D-AC algorithm. First, we will specify the construction of the process itself. In addition to the set of unit-rate Poisson processes, driving the service completions, we define primitive processes (common for each r), driving the random uniform assignment of customers "within" each a-configuration q. Namely, for each q we define an i.i.d. sequence $\xi_q(1), \xi_q(2), \ldots$ of random variables, uniformly distributed in $[0, 1]$. The configurations $k \in q$ are indexed by $1, 2, \ldots, |q|$ (in arbitrary fixed order). When an m-th customer of any type is assigned to a-configuration q (with m referring to the order of assignment since initial time 0, and not to the customer type), this customer is assigned to a server in configuration k' indexed by 1 if

$$\xi_q(m) \in [0, X_{k'}^r / X_q^r],$$

it is assigned to a server in configuration k'' indexed by 2 if

$$\xi_q(m) \in (X_{k'}^r / X_q^r, (X_{k'}^r + X_{k''}^r)/X_q^r],$$

and so on. (Note that necessarily $X_q^r > 0$ – otherwise there would be no assignment to a-configuration q.) Denote

$$g_q^r(s, \zeta) \doteq \sum_{m=1}^{\lfloor rs \rfloor} I\{\xi_q(m) \leq \zeta\},$$

where $s \geq 0$, $0 \leq \zeta \leq 1$, and $\lfloor \cdot \rfloor$ denotes the integer part of a number. Obviously, from the strong law of large numbers (SLLN) and the monotonicity of $g_q^r(s, \zeta)$ on both arguments, we have the following functional SLLN

$$g_q^r(s, \zeta) \to s\zeta, \quad \text{u.o.c.} \quad \text{w.p.1} \tag{50}$$

Clearly, the realization of the process is uniquely determined by the initial state and the realizations of driving processes $\Pi_{ki}(\cdot)$ and $(\xi_q(1), \xi_q(2), \ldots)$.

A set of Lipschitz continuous functions $[\{x_k(\cdot),\ k \in \mathcal{K}\}, \{d_{ki}(\cdot),\ (k, i) \in \mathcal{M}\}, \{a_{ki}(\cdot),\ (k, i) \in \mathcal{M}\}]$ on the time interval $[0, \infty)$ we call a *fluid sample path* (FSP), if there exist realizations of $\Pi_{ki}(\cdot)$ satisfying (8), realizations of $(\xi_q(1), \xi_q(2), \ldots)$ satisfying (50), and a fixed subsequence of r, along which convergence (10) holds.

It is easy to see that the family of all FSPs is *uniformly* Lipschitz.

We can easily verify that Lemmas 3 and 4 hold as is for Greedy-D-AC algorithm. Further, the following lemma is analogous to Lemma 8 (and has essentially same proof).

Lemma 23. *Any FSP is such that at any regular point t,*

$$\frac{d}{dt} F(x(t)) \leq 0. \tag{51}$$

Moreover, unless $x(t)$ satisfies NSI-condition, the inequality (51) is strict.

We will need the following FSP property, which follows from the random uniform rule of Greedy-D-AC for assignments within each a-configuration.

Lemma 24. *Consider an FSP. Suppose that at some $t > 0$, $x = x(t)$ is such that for some $k \in \mathcal{K}$ and i we have: $x_k > 0$, $k_i > 0$, $k' = k - e_i \neq 0$, $x_{k'} = 0$, and $x_{q'} > 0$ where $q' = q(k')$. Then t is not a regular point.*

Proof. Suppose t is a regular point. We must have $\sum_{i'} v_{k'+e_{i'},i'}(t) = 0$. Indeed, in a small interval $[t, t + \delta]$, for all sufficiently large r, the pre-limit sample paths defining the FSP are such that

$$\frac{x^r_{k'}}{x^r_{q'}} < \frac{C\delta}{x_{q'} - C\delta},$$

where $C > 0$ is some constant (depending on the Lipschitz constants for FSP components). This, along with (50), implies that the fraction of the customers added in $[t, t + \delta]$ to servers in configuration k', among those added to a-configuration q', is upper bounded by the RHS, which can be made arbitrarily small by choosing sufficiently small δ. This means that $\sum_{i'}(a_{k'+e_{i'},i'}(t + \delta) - a_{k'+e_{i'},i'}(t))/\delta \downarrow 0$ as $\delta \to 0$, which implies $\sum_{i'} v_{k'+e_{i'},i'}(t) = 0$. Since $x_{k'} = 0$, obviously, $\sum_{i'} w_{k',i'}(t) = 0$. However, $w_{k,i}(t) = \mu_i k_i x_k > 0$. Therefore, $(d/dt)x_{k'}(t) > 0$. This is a contradiction, because if t is regular, $x_{k'}(t) = 0$ implies $(d/dt)x_{k'}(t) = 0$. □

We will also need the *continuity* and *shift* properties of the FSPs, which are quite generic. (See Sections 5 and 6 in [7]. Although our model is different, essentially same proofs as in [7] apply.) The *time shift by $\theta \geq 0$*, applied to an FSP $[\{x_k(\cdot)\}, \{d_{ki}(\cdot)\}, \{a_{ki}(\cdot)\}]$, produces the set of functions with the same time argument $t \geq 0$, but with $x_k(t)$ replaced by $x_k(\theta + t)$, $d_{ki}(t)$ replaced by $d_{ki}(\theta + t) - d_{ki}(\theta)$, $a_{ki}(t)$ replaced by $a_{ki}(\theta + t) - a_{ki}(\theta)$.

Lemma 25. *The family of FSPs satisfies the following properties.*
(i) Continuity: If there is a converging sequence of FSPs, indexed by β, namely

$$[\{x^{(\beta)}_k(\cdot)\}, \{d^{(\beta)}_{ki}(\cdot)\}, \{a^{(\beta)}_{ki}(\cdot)\}] \to [\{x_k(\cdot)\}, \{d_{ki}(\cdot)\}, \{a_{ki}(\cdot)\}], \quad \textit{u.o.c.},$$

then the limit is also an FSP.
(ii) Shift (or "Memoryless"): The time shift of any FSP by any $\theta \geq 0$ is also an FSP.

Proof. (i) For each fixed index β, and the FSP associated with it, consider a sequence of (scaled) sample paths of the process, that define this FSP:

$$[\{x^{(\beta,r)}_k(\cdot)\}, \{d^{(\beta,r)}_{ki}(\cdot)\}, \{a^{(\beta,r)}_{ki}(\cdot)\}] \to [\{x^{(\beta)}_k(\cdot)\}, \{d^{(\beta)}_{ki}(\cdot)\}, \{a^{(\beta)}_{ki}(\cdot)\}], \quad \text{u.o.c.}, \quad \text{as } r \to \infty.$$

Then, we can choose a subsequence of r, and the corresponding $\beta = \beta(r)$, so that

$$[\{x^{(\beta(r),r)}_k(\cdot)\}, \{d^{(\beta(r),r)}_{ki}(\cdot)\}, \{a^{(\beta(r),r)}_{ki}(\cdot)\}] \to [\{x_k(\cdot)\}, \{d_{ki}(\cdot)\}, \{a_{ki}(\cdot)\}], \quad \text{u.o.c.},$$

and therefore the limit satisfies the definition of an FSP.
(ii) We pick a sequence of (scaled) sample paths of the process, that define the FSP. It is easy to see that the time shifts of these sample paths define the FSP which is the time shift of the original one. □

We are now in position to prove the following lemma, which is key (along with Lemmas 22 and 24) in our analysis of Greedy-D-AC algorithm.

Lemma 26. *Consider an FSP. Suppose t is a regular point and $x(t) \notin \mathcal{X}^*$. Then*

$$(d/dt)\Phi(x(t)) < 0. \tag{52}$$

Proof. Suppose not, namely $(d/dt)\Phi(x(t)) = 0$. Then for $x = x(t)$ the NSI-condition holds, and therefore (42) holds as well. We will obtain a contradiction. Condition (43) does *not* hold (otherwise, (42) and (43) would imply $x \in \mathcal{X}^*$). Then, consider a minimal a-configuration q for which (43) is violated, namely: for some $k \in q$,

$$u_k < \max_{k' \in q} u_{k'}, \quad x_k > 0. \tag{53}$$

We will show that this is impossible. First, obviously, $x_q > 0$.
Case (c3.1): suppose $k_i > 0$ for some $i \in \mathcal{I}^-$. Fix this i and consider $q' = q - e_i$. If $q' = 0$, we have $\Delta_{q,i} > 0$, which means η_i cannot be negative – a contradiction. Therefore, $q' \neq 0$ and we must have $x_{q'} > 0$ (because otherwise $\Delta_{q,i} > 0$ leads, again, to the contradiction with $\eta_i < 0$). Consider the set $p \doteq \{k' + e_i \mid k' \in q'\} \subseteq q$; i.e., these are the configurations in q that are obtained by adding one type i customer to configurations in q' – it may or may not be a strict subset of q. If $\max_{k' \in p} u_{k'} < \max_{k' \in q} u_{k'}$, then, since $\max_{k' \in q'} u_{k'} + \eta_i \leq \max_{k' \in p} u_{k'}$, we obtain $\Delta_{q,i} > \eta_i$, which, along with $x_k > 0$, leads to the contradiction with NSI-property. Therefore, $\max_{k' \in p} u_{k'} = \max_{k' \in q} u_{k'}$. Then, there exists $k'' \in \arg\max_{k' \in q'} u_{k'}$ such that $x_{k''} > 0$ and $k'' + e_i \in \arg\max_{k' \in q} u_{k'}$. (Here we used the fact that for any $k' \in q'$, condition (43) does hold – recall that q is a minimal a-configuration for which (43) is violated.) Note also that $u_{k-e_i} < \max_{k' \in q'} u_{k'} = u_{k''}$ and $x_{k-e_i} = 0$ (otherwise, again, (43) would be violated at $q' < q$). We see that x and the edge (k, i) satisfy conditions of Lemma 24. This means t cannot be regular. Thus, case (c3.1) is impossible.
Case (c3.2): suppose for any i with $k_i > 0$ we have $i \in \mathcal{I}^+$. Fix one such i and consider $q' = q - e_i$. If $q' = 0$, we have $\Delta_{q,i} > \eta_i > 0$ – a contradiction with the definition of η_i. Therefore, $q' \neq 0$ and then we must have $x_{q'} > 0$ (because $\eta_j > 0$ for each j with $k_j > 0$). Consider the set $p \doteq \{k' + e_i \mid k' \in q'\} \subseteq q$. Note that $\max_{k' \in q'} u_{k'} > 0$ and $\max_{k' \in q'} u_{k'} + \eta_i = \max_{k' \in p} u_{k'} > 0$. If $\max_{k' \in p} u_{k'} < \max_{k' \in q} u_{k'}$, then $\Delta_{q,i} > \eta_i > 0$, which (along with $x_k > 0$) contradicts the definition of η_i. Therefore, $\max_{k' \in p} u_{k'} = \max_{k' \in q} u_{k'}$. From this point on, the argument leading to a contradiction repeats that in the case (c3.1) verbatim. Thus, the case (c3.2) is impossible. The proof is complete. □

Lemma 27. *For any $T > 0$ and $\epsilon > 0$, there exists $\delta > 0$ such that the following property holds uniformly on all FSPs and all $t_0 \geq 0$:*

$$d(x(t), \mathcal{X}^*) \geq \epsilon, \ t \in [t_0, t_0 + T] \quad \textit{implies} \quad \Phi(x(t_0 + T)) - \Phi(x(t_0)) \leq -\delta. \tag{54}$$

Proof. If (54) would not hold, we would be able to construct a sequence of FSPs converging u.o.c. to an FSP such that

$$d(x(t), \mathcal{X}^*) \geq \epsilon \ \text{ and } \ \Phi(x(t)) = \Phi(x(0)), \ t \in [0, T].$$

(Here we use the shift, continuity and uniform Lipschitz properties of the family of FSPs, and the fact that $\mathcal{X}$ is compact.) This is not possible, because by Lemma 26 we must have $(d/dt)\Phi(x(t)) < 0$ at every regular point in $[0, T]$. □

As a corollary, we obtain the following analog of Lemma 9.

Lemma 28. *Any FSP is such that*

$$d(x(t), \mathcal{X}^*) \to 0. \tag{55}$$

The convergence is uniform across all initial states $x(0) \in \mathcal{X}$.

7.4 Proof of of Theorem 19.

The rest of the proof of Theorem 19 is same as that of Theorem 2.

8 Some generalizations

A number of generalizations of our results are not difficult to obtain. We will discuss Theorems 2 and 12 to be specific, but analogous generalizations apply to Theorems 19 and 21.

8.1 A different procedure for placing arrivals.

Theorems 2 and 12 require that when a type i customer (in Theorem 2) or a type i token (in Theorem 12) is assigned for service, it is placed along the edge (k, i) minimizing the weight differential $\Delta_{ki} = \Delta_{ki}(X(t-))$. The procedure of choosing the edge to place a customer (or token) can be replaced by the following one, which might be easier to implement in some scenarios. We compare $\Delta_{k'i}$ for the edge (k', i) along which a type i departure just occurred, to the Δ_{ki} for one edge, selected randomly as follows: with probability $\epsilon \in (0,1)$ we select edge (e_i, i); with probability $1-\epsilon$ we pick a non-empty server uniformly at random and, if its configuration ℓ is such that $k = \ell + e_i \in \mathcal{K}$, we select edge (k, i). (ϵ is a fixed parameter.) Now, if $\Delta_{ki} < \Delta_{k'i}$ for the selected edge (k, i) (if any), we place the customer (or token) along (k, i); otherwise, we place it "back" along (k', i). It is not difficult to see that the proofs of Theorems 2 and 12 still hold when Greedy-D and Greedy-DM algorithms, respectively, are adjusted as described above.

The described alternative procedure generalizes the results in the sense that we can, for example, use this procedure with a fixed probability $\delta \in [0, 1]$ and use the the "old" procedure (picking the smallest differential Δ_{ki}) with probability $1 - \delta$, and the results still hold.

8.2 More general input processes and service time distributions.

Theorems 2 and 12 still hold for much more general input processes and service time distributions (as opposed to Poisson and exponential, respectively). For example, a simple (but still far reaching) generalization is for the case when, for each customer type i, the input process is renewal (i.i.d. interarrival times, with mean $1/(\lambda_i r)$ and finite variance) and the service time distribution $G_i(\xi)$ (with mean $\int_0^\infty \xi dG(\xi) = 1/\mu_i$) has the "hazard rate" lower bounded by $\mu_i^{min} \in (0, \mu_i]$: $dG(\xi)/[1-G(\xi)] \geq \mu_i^{min} d\xi$, $\forall \xi \geq 0$. In this case, we observe that the key conservation laws still hold for the fluid limit of the stationary system: (a) the "amount" of type i fluid is ρ_i and remains constant and (b) the total rate of (actual) type i departures is λ_i and remains constant. In addition, say in the proof of Theorem 12 to be specific, the corresponding FSPs are such that (actual) type i departure rate from state $(k, \hat{k})$ is lower bounded by $\mu_i^{min} \hat{k}_i x_{(k,\hat{k})}(t)$. (Of course, the FSPs need to be defined more generally, to account for elapsed service times.) Given these properties, the entire argument goes through essentially as is. And, clearly, these properties hold under the input flow and service time assumptions still far more general than in the simple case described above.

9 Discussion

We have shown that (versions of) the Greedy algorithm are asymptotically optimal in the sense of minimizing the objective function $\sum_k X_k^{1+\alpha}$ with $\alpha > 0$. When α is small (but positive), the algorithms produce an approximation of a solution minimizing the linear objective $\sum_k X_k$, i.e. the total number of occupied servers. If $\sum_k X_k$ is the "real" underlying objective, the "price" we pay by applying Greedy algorithm with small $\alpha > 0$ is that the algorithm will keep non-zero amounts ("safety stocks") of servers in many "unnecessary" (from the point of view of linear objective) configurations k, including many – potentially all – non-maximal configurations in $\mathcal{K}$. What we gain for this "price" is the simplicity and agility of the algorithm. "True" minimization of the linear objective $\sum_k X_k$ requires that a linear program is solved (via explicit offline or implicit dynamic approach), so that the system is prevented from using "unnecessary" configurations k, not employed in optimal LP solutions.

The Greedy algorithm with $\alpha > 0$ is asymptotically optimal as the average number r of customers in the system goes to infinity. The fact that it maintains safety stocks of many configurations, means in particular that the algorithms' performance is close to optimal when the ratio $r/|\mathcal{K}|$ is sufficiently large, so that there is enough customers in the system to keep non-negligible safety stocks of servers in potentially all configurations. If the number $|\mathcal{K}|$ of configurations is large, then r needs to be very large to achieve near-optimality. The use of aggregate configurations in the special case of vector-packing constraints alleviates this scalability issue when the number $|\mathcal{Q}|$ of aggregate configurations is substantially smaller than $|\mathcal{K}|$.

Finally, we note that the closed system, considered in Theorems 2 and 19, is not necessarily artificial. For example, it models the scenario where VMs do not leave the system, but can be moved ("migrated") from one host to another. In this case, a "service completion" is a time point when a VM migration can be attempted.

References

[1] N. Bansal, A. Caprara, M. Sviridenko. A New Approximation Method for Set Covering Problems, with Applications to Multidimensional Bin Packing. *SIAM J. Comput.*, 2009, Vol.39, No.4, pp.1256-1278.

[2] J. Csirik, D. S. Johnson, C. Kenyon, J. B. Orlin, P. W. Shor, and R. R. Weber. On the Sum-of-Squares Algorithm for Bin Packing. *J.ACM*, 2006, Vol.53, pp.1-65.

[3] D.Gamarnik. Stochastic Bandwidth Packing Process: Stability Conditions via Lyapunov Function Technique. *Queueing Systems*, 2004, Vol.48, pp.339-363.

[4] A. Gulati, A. Holler, M. Ji, G. Shanmuganathan, C. Waldspurger, X. Zhu. VMware Distributed Resource Management: Design, Implementation and Lessons Learned. *VMware Technical Journal*, 2012, Vol.1, No.1, pp. 45-64. http://labs.vmware.com/publications/vmware-technical-journal

[5] J.W. Jiang, T. Lan, S. Ha, M. Chen, M. Chiang. Joint VM Placement and Routing for Data Center Traffic Engineering. *INFOCOM-2012.*

[6] S.T. Maguluri, R. Srikant, L.Ying. Stochastic Models of Load Balancing and Scheduling in Cloud Computing Clusters. *INFOCOM-2012.*

[7] A.L. Stolyar. On the Stability of Multiclass Queueing Networks: A Relaxed Sufficient Condition via Limiting Fluid Processes. *Markov Processes and Related Fields*, Vol. 1(4), 1995, pp.491-512.

[8] A. L. Stolyar. MaxWeight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Annals of Applied Probability*, 2004, Vol.14, No.1, pp.1-53.

[9] A. L. Stolyar, T. Tezcan. Shadow routing based control of flexible multi-server pools in overload. *Operations Research*, 2011, Vol.59, No.6, pp.1427-1444.