Stochastic Stability Under Network Utility Maximization: General File Size Distribution

Mung Chiang Devavrat Shah Ao Tang

Abstract

We prove the stochastic stability of resource allocation under Network Utility Maximization (NUM) under general arrival process and file size distribution with bounded support, for α-fair utilities with α sufficiently small and possibly different for different sources' utility functions. In addition, our results imply that the system operating under α-fair utility is $1/(1+\alpha)$-approximate stable for any $\alpha \in (0,\infty)$ for any file size distribution with bounded support. Our results are in contrast to the recent stability result of Bramson (2005) for max-min fair (i.e. $\alpha = \infty$) under general arrival process and file size distribution, and that of Massoulie (2006) for proportional fair (i.e. $\alpha = 1$) under Poisson arrival process and phase-type distributions. To obtain our results, we develop an appropriate Lyapunov function for the fluid model established by Gromoll and Williams (2006)[1].

I. Introduction

In 1998, Kelly, Maullo, and Tan [15] identified the current Internet congestion control protocol with an algorithm that allocates rates to flows according to certain 'fairness criteria' reflected through concave utility functions, which are maximized under linear capacity constraints. An extensive amount of research since then has shown many applications of this approach, from reverse-engineering of all major types of TCP congestion control protocols in use today to development of substantially improved new protocols. In [21], [27], an interested reader can find detailed surveys on the philosophy of viewing a resource allocation or congestion control algorithm as implicitly solving a global Network Utility Maximization (NUM) problem. In particular, such optimization problems have been studied as 'monotropic programming' [25] for a long time, and admit a simple, iterative, and distributed solution based on dual decomposition. Over the last several years, this line of work has further evolved to the following view of 'Layering as Optimization Decomposition': the entire protocol stack of network architecture can be thought of as optimizing a generalized network utility function over a constraint set of various types of variables, with different decomposition schemes corresponding to different layering architectural alternatives. Under a particular decomposition, the decomposed subproblems correspond to the functional modules (i.e., layers), and the interfaces among the layers are represented by some specific function of the primal or dual variables. As surveyed in [6], many researchers have contributed to this research area.

However, many results in the area adopt a deterministic NUM formulation. In reality, flows arrive to the network with finite workloads and depart after finishing the work. The service rates are determined by the solution to the NUM problem, which in turn takes in the number of flows as an argument. The key property of stochastic stability has been extensively studied since 1999. In [26], Robert and Massoulie introduced a stochastic dynamic model for Internet congestion control where flows with different service requirement (or file size when flow requests are 'file transfers') arrive, the rate allocation is done according to appropriate NUM, and flows depart on completion on their service. Subsequent to this work, de Veciana, Konstantopoulos and Lee [9], as well as Bonald and Massoulie [2], studied the stability property of the above introduced model under the assumption that arrival process has Poisson distribution while service requirement of flows have exponential distribution. In [9], stability of max-min fair and proportional fairness was established, while in [2], the stability of all weighted α-fair policies, $\alpha \in (0,\infty)$ was established. These results assumed that the rate allocation according the appropriate optimization is done

Author names appear in the alphabetical order of their last names. Affiliations: M. Chiang is with Department of Electrical Engineering and affiliated with the Program in Applied and Computational Mathematics at and Princeton University, D. Shah is with Department of Electrical Engineering and Computer Science at MIT, and A. Tang is with Division of Engineering and Applied Science at Caltech. Email: chiangm@princeton.edu,devavrat@mit.edu, aotang@caltech.edu .

[1] To be precise, our fluid model scaling is different so as to accommodate the case of heterogeneous utility functions of different sources. This scaling allows for possibility of larger class of utility functions as well. However, the justification is the same as that of Gromoll and Williams[13]. Hence, we call it the fluid model of [13]. For completeness, we will include key steps of justification borrowed from [13].

instantaneously. This is called the 'time-scale separation', i.e., the time scale at which rate allocation algorithm operates is extremely fast compared to the time scale of the system dynamics. Lin and Shroff [19], as well as Srikant [28], established stability without time-scale separation assumption for α-fair policies for $\alpha \geq 1$ under the Poisson and exponential distributional assumptions. Natural generalizations of these results to other convex constraint sets were also obtained [31], [20].

While assuming a Markov traffic model (Poisson arrival with exponential file size distribution) leads to analytic tractability, it is widely recognized that file sizes in the Internet or wireless networks do *not* follow the exponential distribution. In this paper, we are interested in answering the question of whether the network is stable, under α-fair rate allocation, for general distributional assumption on arrival process and service requirement of the flows. Here, stability means that the departure rate is the same as the arrival rate, i.e., rate stability, or fluid stability. This question has been of great recent interest as a positive answer will provide justification for using NUM and its generalizations for network resource allocation and architecture design. A stable system essentially means that the capacities that can be utilized in a deterministic NUM can also be utilized in the stochastic setting.

The following is a brief summary of what is currently known about this question (to the best of authors' knowledge based on the available preprints and personal communication). Bramson [5] has established stability for max-min fair (corresponding to $\alpha = \infty$) rate allocation under general arrival and file size distribution, and Massoulie [22] has established stability for proportional fair (corresponding to $\alpha = 1$) rate allocation for Poisson arrival and phase-type service distribution. The result of [22] is established by (a) justifying fluid model for system with exponential and Poisson assumption with routing, (b) establishing stability of this fluid model, and (c) using the known observation that network with phase type distribution for service requirement can be mapped to network with exponential-Poisson assumption and routing. We also make note of the following two results. Lakshmikantha, Beck and Srikant [18] established stability of Proportional fairness for a two resource linear network and 2×2 grid network for Poisson arrival and phase-type distribution of service requirement; Kelly and Williams [16] had formulated a proper fluid model for exponential service requirement to study the 'invariant states' as an intermediate step for obtaining diffusion approximation for all $\alpha \in (0, \infty)$.

Recently, Gromoll and Williams [13] have established fluid model for α-fair rate allocation, $\alpha \in (0, \infty)$, under general distributional condition on arrival process and service distribution. This is a very important step in the process of establishing stability via the means of fluid models. Using this fluid model, they have obtained a characterization of 'invariant states'. This led to stability of network under α-fair allocation, $\alpha \in (0, \infty)$, when the network topology is a tree.

We will establish the approximate stability of any α-fair rate allocation for any network topology under general distribution for $\alpha \in (0, \infty)$. We prove that any network with α-fair rate allocation is $1/(1 + \alpha)$-approximate stable[2] under general distribution conditions. In a stronger characterization, we prove that the system is stable for a continuum of sufficiently small and strictly positive α_i, possibly different α_i for each source i. We will crucially use the fluid model established in [13] to obtain our results.

The paper is organized as follows. In Section II we present notations, technical preliminaries, system description, and stochastic model. In Section III we present the fluid model scaling and formal statement establishing relation between fluid model solutions and the stochastic system. The fluid model scaling presented in the paper is different from that used in [13] or [22] as it allows for heterogeneous utility functions for different sources, with possibly utilities coming from a larger class of utility functions compared to that in [13]. In Section IV, we present the main result of this paper (Corollary 7 and Theorem 6) establishing $1/(1+\alpha)$-approximate stability of network operating under α-fair rate allocation and general distributional conditions. This also implies the stability of network for a range of sufficiently small α. The stability is established by use of a new Lyapunov function, which is inspired by known Lyapunov functions in this research literature. However, as reader will notice, in contrast to the Markov arrival model, it is substantially more challenging to work with the fluid model for general distribution due to limited amount of information about fluid dynamics. We present some simple extensions and limitations of our results along with a discussion on future directions in Section V.

Even though our fluid model scaling is different, its justification is identical to that in [13]. For completeness, we present a sketch of the proof (of Theorem 5) in Appendix. An interested reader is encouraged to read [13] for

[2] Definition of $1/(1 + \alpha)$-approximate stability will be made clear in Corollary 7. Roughly speaking, it means $100/(1 + \alpha)$ % utilization of network's resource.

any of the missing details.

II. Setup

This section describes notation, necessary technical preliminaries, network model with NUM, and the stochastic model that will be studied in the paper. Our notation and representation of variables are almost identical to those in [13], so as to make it easier for an interested reader to re-construct the missing details in the proof of fluid model justification based on [13].

A. *Notation and Technical Preliminaries*

Let the natural number set be $\mathbb{N} = \{1, 2, \dots\}$, and the real number set be $\mathbb{R} = (-\infty, \infty)$ and $\mathbb{R}_+ = \{x \in \mathbb{R} : x \geq 0\} = [0, \infty)$. Let $\mathbb{R}^d$ be d-dimensional Euclidian space; similarly $\mathbb{N}^d$ and $\mathbb{R}^d_+$. Let $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. Let identity function be denoted by χ, i.e. $\chi(x) = x$ for all $x \in \mathbb{R}_+$. Let unit function be denoted by $\mathbf{1}$, i.e. $\mathbf{1}(x) = 1$ for all $x \in \mathbb{R}_+$. For vectors $\mathbf{u} = (u_1, \dots, u_{\mathbf{I}})$ and $\mathbf{v} = (v_1, \dots, v_{\mathbf{I}})$, let $\mathbf{u} \circ \mathbf{v}$ denote component-wise multiplication $(u_1 v_1, \dots, u_{\mathbf{I}} v_{\mathbf{I}})$.

For a real-valued function defined on $\mathbb{R}_+$, say $f : \mathbb{R}_+ \to \mathbb{R}$, its sup-norm is defined as $\|f\|_\infty = \sup_{x \in \mathbb{R}_+} |f(x)|$. Similarly, for $f : [0, T] \to \mathbb{R}$ define $\|f\|_T = \sup_{x \in [0,T]} |f(x)|$. Let f' denote the derivative of f, if exists. For any function f, let $f(\cdot - s), s > 0$, be its shifted copy by s, with the understanding that $f(x - s) = 0$ for all $x < s$.

Let $\mathbf{C}_b(\mathbb{R}_+)$ denote the set of bounded continuous functions defined on $\mathbb{R}_+$, $\mathbf{C}^1(\mathbb{R}_+)$ denote the set of once continuously differentiable functions and $\mathbf{C}^1_b(\mathbb{R}_+)$ denote the set of $f \in \mathbf{C}^1(\mathbb{R}_+)$ that have both f, f' bounded on $\mathbb{R}_+$. Define, $\mathcal{C} = \{f \in \mathbf{C}^1_b(\mathbb{R}_+) : f(0) = 0, f'(0) = 0\}$ and $\mathcal{C}_c = \{f \in \mathcal{C} :\ f \text{ has compact support}\}$.

Let $\mathbf{M}$ be set of finite non-negative measures (not necessarily probability measures) on $\mathbb{R}_+$. Let it be endowed with the topology induced by weak convergence: $\zeta^k \xrightarrow{\mathbf{w}} \zeta$ in $\mathbf{M}$ if and only if $\langle f, \zeta^k \rangle \to \langle f, \zeta \rangle$ for all $f \in \mathbf{C}_b(\mathbb{R}_+)$, where we have used notation[3] that, for $\zeta \in \mathbf{M}$,

$$\langle f, \zeta \rangle = \int_{\mathbb{R}_+} f d\zeta.$$

This topology is induced by the Prohorov's metric defined as follows: for $\zeta, \xi \in \mathbf{M}$, define

$$\mathbf{d}[\zeta, \xi] = \inf\{\varepsilon > 0 : \zeta(B) \leq \xi(B^\varepsilon) + \varepsilon, \text{ and } \xi(B) \leq \zeta(B^\varepsilon) + \varepsilon, \text{for all closed } B \subset \mathbb{R}_+\}, \tag{1}$$

where $B^\varepsilon = \{x \in \mathbb{R}_+ : \inf_{y \in B} |x - y| < \varepsilon\}$. For product space $\mathbf{M}^{\mathbf{I}}$ for any $\mathbf{I} \in \mathbb{N}$, define metric $\mathbf{d}_{\mathbf{I}}$ as follows: for $\zeta = (\zeta_1, \dots, \zeta_{\mathbf{I}}), \xi = (\xi_1, \dots, \xi_{\mathbf{I}}) \in \mathbf{M}^{\mathbf{I}}$,

$$\mathbf{d}_{\mathbf{I}}(\zeta, \xi) = \max_{1 \leq i \leq \mathbf{I}} \mathbf{d}(\zeta_i, \xi_i).$$

It is well known that the metric space $\mathbf{M}^{\mathbf{I}}$ thus defined is a complete and separable, i.e. Polish space.

Let $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$ denote the set of functions from $[0, T]$ to $\mathbf{M}^{\mathbf{I}}$ that are right continuous with left limits, also known as *cadlag* functions. In this paper, the domain $[0, T]$ will be time and hence use of 'time' should not confuse the reader. We will endow $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$ with Skorohod's J_1-topology. Our interest will be in convergence of probability distributions on $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$, for finite (time-interval) T. For this, we will be interested in an appropriate metric on $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$ defined next.

Let Φ be set of nondecreasing function $\varphi : [0, T] \to [0, T]$ with $\varphi(0) = 0, \varphi(T) = T$. Define $\|\varphi\|^o = \sup_{0 \leq s < t \leq T} \left| \log \frac{\varphi(t) - \varphi(s)}{t - s} \right|$. Let $\Phi_b = \{\varphi \in \Phi : \|\varphi\|^o < \infty\}$. Now, for any $\zeta, \xi \in \mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$, the distance between them is defined as

$$d^o(\zeta, \xi) = \inf_{\varphi \in \Phi_b} \left\{ \|\varphi\|^o \vee \left(\sup_{0 \leq t \leq T} \mathbf{d}_{\mathbf{I}}(\zeta(t), \xi(\varphi(t))) \right) \right\}.$$

The space $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$ endowed with the above metric is complete and separable, i.e. Polish. Before we characterize the relatively compact sets in $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$, we define the modulus of continuity for $\zeta \in \mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$. Consider any $\delta \in (0, 1)$ and any sequence $\{t_i\}$ of some $v \leq 2T/\delta$ points, such that $0 = t_0 < t_1 < \cdots < t_v = T$

[3] The notation $\langle f, \zeta \rangle$ for $\zeta = (\zeta_1, \dots, \zeta_d) \in \mathbf{M}^d$ will naturally mean $(\langle f, \zeta_1 \rangle, \dots, \langle f, \zeta_d \rangle)$.

and $\min_i t_i - t_{i-1} > \delta$. Call the set of all such sequences as $\mathbf{T}_\delta$. Then, the modulus of continuity of ζ with δ precision is

$$\mathbf{w}'_T(\zeta,\delta) = \inf_{\{t_i\}\in\mathbf{T}_\delta} \max_i \sup_{s,t\in[t_{i-1},t_i]} \mathbf{d_I}[\zeta(s),\zeta(t)].$$

In $\mathbf{D}([0,T],\mathbf{M^I})$, a set A is relatively compact if the following holds: (1) there exists a compact set $\mathbf{K}\subset\mathbf{M^I}$ such that for any $\zeta\in A$, $\zeta(t)\in\mathbf{K}$ for all $t\in[0,T]$, and (2) $\lim_{\delta\to 0}\sup_{\zeta\in A}\mathbf{w}'_T(\zeta,\delta)=0$. This characterization of relatively compact set suggests the following criteria for proving tightness of a sequence of probability measures $\mathbb{P}_n, n\in\mathbb{N}$, on $D([0,T],\mathbf{M^I})$ as follows: the sequence of probability measures $\mathbb{P}_n, n\in\mathbb{N}$ is tight if (1) for any $\varepsilon>0$ there exists a compact set $\mathbf{K}_\varepsilon\subset\mathbf{M^I}$ such that $\liminf_n \mathbb{P}_n(\zeta(t)\in\mathbf{K}_\varepsilon,\ \forall\ t\in[0,T])\geq 1-\varepsilon$, and (2) for any $\varepsilon>0$, $\lim_{\delta\to 0}\limsup_n \mathbb{P}_n(\{\zeta:\mathbf{w}'_T(\zeta,\delta)\geq\varepsilon\})=0$. This characterization of tightness of probability measures is used crucially in fluid model justification.

Finally, for completeness we make the following note. We will be interested in probability measures defined on the product of finite number of spaces. Let there be complete separable metric spaces $\mathbf{S}_1,\dots,\mathbf{S}_d$ with metric $\mathbf{d}_{\mathbf{S}_1},\dots\mathbf{d}_{\mathbf{S}_d}$, respectively. Their product space $\mathbf{S}=\mathbf{S}_1\times\cdots\times\mathbf{S}_d$ will be endowed with topology induced by metric $\mathbf{d_S}$ defined as

$$\mathbf{d_S}(\mathbf{a},\mathbf{b}) = \max_{k\leq d}\mathbf{d}_{\mathbf{S}_k}(\mathbf{a}_k,\mathbf{b}_k), \quad \text{for} \ \ \mathbf{a},\mathbf{b}\in\mathbf{S}.$$

Recall that under this metric $\mathbf{S}$ will be complete and separable as well [4].

B. Networks with Rate Allocation By NUM

We consider a connected network $G=(\mathcal{V},\mathcal{J},C,\mathcal{I})$, where $\mathcal{V}$ is the set of all vertices, $\mathcal{J}$ is the set of $\mathbf{J}$ links, $C=(C_j)_{1\leq j\leq\mathbf{J}}$ denote the capacity vector of the links, and $\mathcal{I}$ is the set of $\mathbf{I}$ routes. Let A be $\mathbf{J}\times\mathbf{I}$ routing incidence matrix, with $A_{ji}=1$ if route i passes through link j and 0 otherwise.

In a network, multiple flows can be active on the same route. Further, flows of different routes (or types) can be sharing a link. The links have limited capacity. Hence, network need to assign the rates to the flows passing through it. In this paper, we are interested in bandwidth sharing policies in which each flow of the same type gets the same bandwidth allocated. Let λ_i be net bandwidth allocated to flows on route i. Since the links have limited capacity, we immediately have the following requirement:

$$A\lambda\leq C.$$

The set of all $\lambda=(\lambda_1,\dots,\lambda_{\mathbf{I}})\in\mathbb{R}_+^{\mathbf{I}}$ satisfying the above inequality are called feasible bandwidth allocation.

In this paper, we are interested in the bandwidth allocation policies that maximizes certain network utility. Equivalently, bandwidth allocation corresponds to a solution of an appropriate Network Utility Maximization (NUM) problem. Let $\mathcal{U}_i(x)$ be utility of a flow of type i when it is allocated rate x. If there are z_i flows of type i and each one is allocated rate x_i, then the net bandwidth allocated to flows of type i is $\lambda_i=x_iz_i$. In this paper, we are primarily interested in the α-fair utility function, introduced by Mo and Walrand [23], which is commonly used in studying NUM type network resource allocation. For any $\alpha\in(0,\infty)$, define [5]

$$\varphi^\alpha(x)=\begin{cases}\frac{x^{1-\alpha}}{1-\alpha} & \text{for } \alpha\in(0,\infty)\backslash\{1\}\\ \log x & \text{for } \alpha=1.\end{cases}$$

Then, under unweighted α-fair allocation, the utility of each flow i is $\mathcal{U}_i=\varphi^{\alpha_i},\alpha_i\in(0,\infty)$. In the weighted α-fair allocation, $\mathcal{U}_i=\kappa_i\varphi^{\alpha_i}$, with κ_i some positive weights (constants). In this paper, for simplicity we will assume that $\kappa_i=1$ for all i. However, as it will be clear to the reader that the results of this paper hold true for any choice of $\kappa_i>0$.

Now the bandwidth or rate allocation happens according to an optimization problem which uses number of flows as argument. Let $z=(z_1,\dots,z_{\mathbf{I}})$ be vector of number of flows. Then, each flow of type i is allocated rate $\mathbf{x}_i(z)$,

[4] We refer an interested reader to the book by Billingsley [1] for exposition on the topic of convergence of probability distribution on metric spaces (and some facts stated here).

[5] The general definition of α-fair utility allows for $\alpha=0$, but such linear utility function leads to potential starvation and is not considered here.

$1 \leq i \leq \mathbf{I}$, where $\mathbf{x}(z) = (\mathbf{x}_1(z), \ldots, \mathbf{x}_{\mathbf{I}}(z))$ is a solution to the following optimization problem over $x \geq 0$:

$$\begin{aligned} \text{maximize} \quad & \sum_{i=1}^{\mathbf{I}} \mathcal{U}_i(x_i) z_i \\ \text{subject to} \quad & Ax \circ z \leq C, \\ & x_i = 0 \ \text{ if } \ z_i = 0, \quad \text{for all } i \leq \mathbf{I}, \end{aligned} \tag{2}$$

where $x \circ z = (x_1 z_1, \ldots, x_{\mathbf{I}} z_{\mathbf{I}})$ is the vector of net bandwidth allocated to flows. In this paper the utilities $\mathcal{U}_i(\cdot)$ are strictly concave on $(0, \infty)$. This will imply the uniqueness of the solution of the above optimization problem from standard arguments. Thus, $\mathbf{x}(z)$ can be viewed as a function from $\mathbb{R}_+^{\mathbf{I}}$ to $\mathbb{R}_+^{\mathbf{I}}$. We will assume that choice of utilities is such that $x(\cdot)$ satisfies the following assumptions.

Assumption 1: For each $i \leq \mathbf{I}$, $\mathbf{x}_i(z)$ is a continuous function on $\{z \in \mathbb{R}_+^{\mathbf{I}} : z_i > 0\}$. Further, if $z_i > 0$ then $\mathbf{x}_i(z) > 0$.

Kelly and Williams [16] showed that Assumption 1 holds when for all $i \leq \mathbf{I}$, the utility functions are the same and $\mathcal{U}_i = \varphi^{\alpha}$ for some $\alpha \in (0, \infty)$ for all $i \leq \mathbf{I}$. This assumption was verified by Ye, Qu and Yuan [32] as well. Next, we establish that Assumption 1 is satisfied even when the utilities of flow types are α fair with different α for different flow types.

Lemma 1: The Assumption 1 is satisfied when $\mathcal{U}_i = \varphi^{\alpha_i}$ with $\alpha_i \in (0, 1)$ for all $i \leq \mathbf{I}$.

Proof: In [16] Kelly and Williams established this lemma when $\alpha_i = \alpha \in (0, 1)$ for all $i \leq \mathbf{I}$. However, their proof used only the following key facts: (a) On $(0, \infty)$ the utility function $\mathcal{U}_i$ is continuous and strictly concave for $i \leq \mathbf{I}$; (b) $\mathcal{U}_i'(x) \to \infty$ as $x \to 0$ for $i \leq \mathbf{I}$ and (c) $\mathcal{U}_i(x) > 0$ if $x > 0$, $\mathcal{U}_i(x) \to 0$ as $x \to 0$. Using these facts (especially (b)), they established that $\mathbf{x}_i(z) > 0$ if $z_i > 0$. Similarly, they used them to provide a detailed argument of the continuity of $\mathbf{x}_i(z)$ on $\{z : z_i > 0\}$.

The proof of [16] for the case when $\alpha_i = \alpha \in (0, 1)$ for all i, does not require the fact that all α_i are equal. Hence, their proof establishes this Lemma[6]. We refer reader to [16] for details. ■

Finally, define the vector of bandwidth allocated to flows, when vector of flows is z, as

$$\Lambda(z) = \mathbf{x}(z) \circ z = (\mathbf{x}_1(z) z_1, \ldots, \mathbf{x}_{\mathbf{I}}(z) z_{\mathbf{I}}).$$

We note the following obvious but crucial property of rate-allocation function $\mathbf{x}(\cdot) : \mathbb{R}_+^{\mathbf{I}} \to \mathbb{R}_+^{\mathbf{I}}$.

Lemma 2: For any $z \in \mathbb{R}_+^{\mathbf{I}}$ such that $z_i \geq \varepsilon$, $\mathbf{x}_i(z) \leq \|C\|/\varepsilon$.

Proof: By definition of optimization problem, we have

$$\|\Lambda(z)\| = \max_{i \leq \mathbf{I}} \Lambda_i(z) \leq \max_{j \leq \mathbf{J}} C_j = \|C\|. \tag{3}$$

The above equation states that the simple fact that the net rate allocated to any flow type is at most $\|C\|$. Hence, for $z_i \geq \varepsilon$,

$$\mathbf{x}_i(z) \ \leq \ \|\Lambda(z)\|/z_i \ \leq \|C\|/\varepsilon.$$

■

C. Network Dynamics and Stochastic Model

Let $t \in \mathbb{R}_+$ denote the time index. Let $Z(t) = (Z_1(t), \ldots, Z_{\mathbf{I}}(t))$ denote the vector of the numbers of flows at time t. Let $E(t) = (E_1(t), \ldots, E_{\mathbf{I}}(t))$ be vector of cumulative number of arrivals of flows to the network in $[0, t]$ with $E(0) = \mathbf{0}$. Let U_{ik}, $k \geq 0$, denote the arrival time of k^{th} flow of type i with $U_{i0} = 0$. Each flow arrives with service requirement (or file-size). Let V_{ik} denote the service requirement of k^{th} flow of type i. Denote $V_i = (V_{ik}, k \geq 1)$ and $V = (V_1, \ldots, V_{\mathbf{I}})$. The system is assumed to start empty[7] at time $t = 0$.

Given the bandwidth allocation rule, the dynamics of the whole system can be obtained from the starting condition, arrival process, and service requirement process. We assume that the arrival process and service requirement process

[6] In general, when all α_i are different and possibly $\alpha_i \in (0, \infty)$, one needs to use the argument of [16] for three different region $(0, 1)$, $\{1\}$ and $(1, \infty)$ together and patch them properly. We believe that proof of [16] extends easily but requires detailed argument. We skip it here but it is a good excercise establishing Lemma 1 for different $\alpha_i \in (0, \infty)$ for an interested reader.

[7] Instead of empty, starting condition can be anything that is *not too bad*. Usually, such starting conditions are handled in a standard manner and we refer an interested reader to see [13].

are defined on a common probability space, say $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{E}$ denoting the expectation. To this end, we assume that arrival process is such that inter-arrival times for flow $i \leq \mathbf{I}$, i.e. $U_{ik} - U_{i(k-1)}, k \geq 1$ are independent and identically distributed (i.i.d.) with $\nu_i^{-1} = \mathbb{E}[U_{i1} - U_{i0}] = \mathbb{E}[U_{i1}] \in (0, \infty)$. The service requirements for flow $i \leq \mathbf{I}$, $\{V_{ik}\}$ also form an i.i.d. sequence with density of distribution ϑ_i such that $\langle \mathbf{1}_{\{0\}}, \vartheta_i \rangle = 0$. Let the average service requirement be $\langle \chi, \vartheta_i \rangle = \mu_i^{-1} \in (0, \infty)$. The traffic intensity is defined as $\rho_i = \nu_i / \mu_i$. We assume that system is underloaded, that is,

$$A\rho \;<\; C. \tag{4}$$

The above condition is necessary for stability: the system can become unstable otherwise. We note that we have assumed the arrival and service processes to be i.i.d. just for simplicity. The only requirement is the existence of functional law of large numbers (equivalently, the validity of Lemma 11). As long as it is true, the fluid model (based on the proof in [13]) is justified and result of this paper holds true.

Now, we describe system dynamics that will lead to the definition of a succinct system descriptor. Given the vector of number of flows in the system at time t, $Z(t)$, the rate allocation happens according to mapping $\mathbf{x}(Z(t))$. Define $S_i(t)$ to be the total amount of service allocated to a flows of type i in $[0, t]$. That is,

$$S_i(t) \;=\; \int_0^t \mathbf{x}_i(Z(\tau)) d\tau. \tag{5}$$

Also define $S_i(t, t+\tau) = S_i(t+\tau) - S_i(t)$ for $\tau \in \mathbb{R}_+$. Finally, let $V_{ik}(t)$ be the remaining amount of service of k^{th} flow of type i at time t. Then,

$$V_{ik}(t) = (V_{ik} - S_i(U_{ik}, t)).$$

Let $W_i(t) = \sum_{k=1}^{E_i(t)} V_{ik}^+(t)$ be the total amount of unfinished work in the system at time t, where $x^+ = x\mathbf{1}_{\{x>0\}}$.

All of the above system information can be compactly represented via measure on $\mathbb{R}_+$ as follow: define $\mathcal{Z}(t) = (\mathcal{Z}_1(t), \ldots, \mathcal{Z}_{\mathbf{I}}(t)) \in \mathbf{M}^{\mathbf{I}}$ as

$$\mathcal{Z}_i(t) = \sum_{k=1}^{E_i(t)} \boldsymbol{\delta}^+_{V_{ik}(t)},$$

where $\boldsymbol{\delta}^+_x \in \mathbf{M}$ is a point mass measure at x if $x > 0$ and is $\mathbf{0}$ if $x \leq 0$. The $\mathcal{Z}_i(t)$ puts a unit amount of mass for each flow of type i in the system at time t at the positive value corresponding to the unfinished amount of work of the flow. For example, if the system has two flows of type 1 with remaining amount of work 2 and 4 at time t, then $\mathcal{Z}_1(t) = \boldsymbol{\delta}^+_2 + \boldsymbol{\delta}^+_4$. The $\mathcal{Z}(t)$ is sufficient to recover most of the relevant system information. For example, for $i \leq \mathbf{I}$

$$Z_i(t) \;=\; \langle \mathbf{1}, \mathcal{Z}_i(t) \rangle, \tag{6}$$

$$W_i(t) \;=\; \langle \chi, \mathcal{Z}_i(t) \rangle \;=\; \mathcal{L}_i(t) - T_i(t), \tag{7}$$

where $\mathcal{L}_i(t) = \sum_{k=1}^{E_i(t)} \boldsymbol{\delta}^+_{V_{ik}}$ and the process T_i is defined as follows: let $T(t) = (T_1(t), \ldots, T_{\mathbf{I}}(t))$ track the cumulative amount of work given to flows. That is,

$$T_i(t) \;=\; \int_0^t \mathbf{x}_i(Z(s)) Z_i(s) ds \;=\; \int_0^t \Lambda_i(Z(s)) ds. \tag{8}$$

Similarly, let process U track the cumulative amount of unused bandwidth in the network. That is,

$$U(t) \;=\; Ct - AT(t). \tag{9}$$

In summary, the system is determined by parameters $(A, C, \nu, \vartheta, \mathcal{U})$, where $\mathcal{U} = (\mathcal{U}_1, \ldots, \mathcal{U}_{\mathbf{I}})$. The processes describing system dynamics are (Z, W, T, U), which are induced by $(E, \mathcal{Z})$ and the NUM given that system starts empty, i.e., $\mathcal{Z}(0) = \mathbf{0}$.

III. Fluid Model Scaling

In this section, we describe fluid model scaling by considering a sequence of systems, indexed by scaling parameter $r \in \mathbb{N}$[8]. Specifically, the r^{th} system has corresponding parameters $(A, C^r, \nu^r, \vartheta, \mathcal{U})$ obeying the following relation: $C^r = rC = (rC_1, \ldots, rC_{\mathbf{J}})$ and $\nu^r = r\nu = (r\nu_1, \ldots, r\nu_{\mathbf{I}})$. That is, the capacity of each link and the arrival rate are scaled r times. However, the network routing matrix A, service requirement ϑ, and utility of the network remains the same. We make a quick remark that under this scaling the loading is $\rho^r = r\rho$, and, from (4),

$$rA\rho = A\rho^r < C^r = rC.$$

Now, we describe the arrival process and service requirement process of the r^{th} system. In this notation, the original system corresponds to the r^{th} system with $r = 1$. Recall that the original system's cumulative arrival process is E and its service requirement is given by V. The arrival process of the r^{th} system, denoted by E^r is $E^r(t) = E(rt)$. That is, requests arriving to the original system in time $[0, rt]$ arrive to the r^{th} system in time $[0, t]$. The requests retain their service requirement, that is, $V^r = V$. The stochastics of a system is completely in the arrival and service processes. Given the above described scaling, we have all the r systems, $r \in \mathbb{N}$, living on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Now, we define the scaled system variables. Let $(Z^r, W^r, T^r, U^r, \mathcal{Z}^r)$ be the variables corresponding to the r^{th} system. Now, define the scaled variables as follows. Given (6)-(9), it is sufficient to describe the scaled measure valued descriptor. Let it be defined as

$$\bar{\mathcal{Z}}^r(t) \;=\; \frac{1}{r}\mathcal{Z}^r(t). \tag{10}$$

Given the scaling of (10), we obtain that for other scaled variables

$$\begin{aligned} \bar{Z}^r(t) = \langle \mathbf{1}, \bar{\mathcal{Z}}^r(t)\rangle \;&=\; \frac{1}{r}\langle \mathbf{1}, \mathcal{Z}^r(t)\rangle \;=\; \frac{1}{r}Z^r(t), \\ \bar{W}^r(t) = \langle \chi, \bar{\mathcal{Z}}^r(t)\rangle \;&=\; \frac{1}{r}\langle \chi, \mathcal{Z}^r(t)\rangle \;=\; \frac{1}{r}W^r(t). \end{aligned} \tag{11}$$

Denote by $\mathbf{x}^r : \mathbb{R}^{\mathbf{I}} \to \mathbb{R}^{\mathbf{I}}$ the mapping from vector of number of flows to rates allocated to flows under NUM for r^{th} system with capacities C^r in place of C in (2).

Lemma 3: For any $r > 0$,

$$\mathbf{x}^r(rz) \;=\; \mathbf{x}(z). \tag{12}$$

Proof: Given $r > 0$, consider the following.

$$\begin{aligned} \mathbf{x}^r(rz) &= \arg\max\left\{\sum_{i\le \mathbf{I}} \mathcal{U}_i(x_i) r z_i : Ax \circ rz \le C^r;\; x_i = 0 \text{ if } rz_i = 0 \text{ and } x \ge \mathbf{0}\right\} \\ &= \arg\max\left\{r\sum_{i\le \mathbf{I}} \mathcal{U}_i(x_i) z_i : Ax \circ z \le C^r/r;\; x_i = 0 \text{ if } z_i = 0 \text{ and } x \ge \mathbf{0}\right\} \\ &= \arg\max\left\{\sum_{i\le \mathbf{I}} \mathcal{U}_i(x_i) z_i : Ax \circ z \le C;\; x_i = 0 \text{ if } z_i = 0 \text{ and } x \ge \mathbf{0}\right\} \\ &= \mathbf{x}(z). \end{aligned} \tag{13}$$

This completes the proof of Lemma 3. ■

Lemma 3 implies that for bandwidth allocation

$$\Lambda^r(rz) = \mathbf{x}^r(rz) \circ rz = \mathbf{x}(z) \circ rz = r\mathbf{x}(z) \circ z = r\Lambda(z).$$

[8] We call parameter r instead of n, as scaling parameter is traditionally called r. This notation hopefully will not create much confusion.

Also, from Lemma 3, we have that, for $i \leq \mathbf{I}$ and $t, \tau > 0$,

$$\begin{aligned} \bar{S}_i^r(t, t+\tau) = \int_t^{t+\tau} \mathbf{x}_i(\bar{Z}^r(s))ds \; &= \; \int_t^{t+\tau} \mathbf{x}_i\left(\frac{Z^r(s)}{r}\right)ds \\ = \int_t^{t+\tau} \mathbf{x}_i^r(Z^r(s))ds \; &= \; S_i^r(t, t+\tau), \end{aligned} \tag{14}$$

$$\bar{T}_i^r(t) = \int_0^t \Lambda_i(\bar{Z}^r(s))ds \; = \; \int_0^t \frac{1}{r}\Lambda_i^r(r\bar{Z}^r(s))ds \; = \; \frac{1}{r}\int_0^t \Lambda_i^r(Z^r(s))ds \; = \; \frac{1}{r}T_i^r(t), \tag{15}$$

$$\bar{U}^r(t) \; = \; Ct - A\bar{T}^r(t) \; = \; \frac{1}{r}\left(C^r t - AT^r(t)\right) \; = \; \frac{1}{r}U^r(t), \tag{16}$$

$$\bar{W}^r(t) \; = \; \frac{1}{r}\mathcal{L}^r(t) - \frac{1}{r}T^r(t) \; = \; \bar{\mathcal{L}}^r(t) - \bar{T}^r(t). \tag{17}$$

Here our interest is in studying the behavior of $(\bar{Z}^r, \bar{W}^r, \bar{T}^r, \bar{U}^r, \bar{\mathcal{Z}}^r)$ as $r \to \infty$. Under the stochastic assumptions on the arrival process and service requirement process, we will find that they will satisfy deterministic fluid model equations as defined below almost surely. Before proceeding further, we make the following remark about the scaling considered in this paper.

Remark. The scaling described above is different from the 'standard' fluid model scaling considered in [13], where the r^{th} system is obtained by scaling the variables of original system in time and space. For example, the $\bar{\mathcal{Z}}^r(t) = \mathcal{Z}(rt)/r$. For fluid model to be meaningfully defined, it is required that, for all $i \leq \mathbf{I}$, $\mathcal{U}_i(rz) = g(r)\mathcal{U}_i(z)$ for some function $g(r)$ (same for all $i \leq \mathbf{I}$) such that $g(r) > 0$ when $r > 0$. Instead, here we are scaling capacity, speeding up the arrival process, and scaling down the variables. A main advantage of such scaling is that it does not require the strictly concave utilities to have the above stated property. This allows for considering heterogeneous utility functions unlike in [13]. We again note that, despite of difference in scaling, the proof techniques of [13] are still sufficient here, primarily because the dynamics of our scaled system is the same as those of the scaled system in [13].

Definition 1 (Auxiliary variables): Given function $\zeta : \mathbb{R}_+ \to \mathbf{M}^{\mathbf{I}}$, define (z, w, τ, u) as follows:

$$\begin{aligned} z(t) \; &= \; \langle \mathbf{1}, \zeta(t) \rangle, \\ w(t) \; &= \; \langle \chi, \zeta(t) \rangle, \\ \tau(t) \; &= \; (\tau_i(t))_{i \leq \mathbf{I}}, \text{ where } \tau_i(t) = \int_0^t \left(\mathbf{x}_i(z(s))z_i(s)\mathbf{1}_{\{z_i(s)>0\}} + \rho_i \mathbf{1}_{\{z_i(s)=0\}}\right)ds, \\ u(t) \; &= \; Ct - A\tau(t). \end{aligned}$$

Definition 2 (Fluid model solution): Given system with parameters $(A, C, \nu, \vartheta, \mathcal{U})$, we call $\zeta : \mathbb{R}_+ \to \mathbf{M}^{\mathbf{I}}$ a solution to fluid model equation if ζ and corresponding auxiliary variables (z, w, τ, u) satisfy the following conditions:

(a) ζ is continuous.
(b) $\|\langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle\| = 0$ for all $t \geq 0$.
(c) For any $f \in \mathcal{C}$ and $i \leq \mathbf{I}$,

$$\langle f, \zeta_i(t) \rangle = \nu_i \langle f, \vartheta_i \rangle \left(\int_0^t \mathbf{1}_{\{z_i(s)>0\}}ds\right) - \int_0^t \langle f', \zeta_i(s) \rangle \mathbf{x}_i(z(s))ds. \tag{18}$$

The following are useful properties of the auxiliary variables (z, w) associated with a fluid model solution. These properties are stated in [13] (specifically, Lemma 3.3 for property of w).

Lemma 4: Suppose ζ is a fluid model solution with $\zeta(0) = \mathbf{0}$. Then, for each $i \leq \mathbf{I}, t \geq 0$,

$$z_i(t) \; \leq \; \nu_i t, \tag{19}$$

$$w_i(t) \; = \; \int_0^t \left(\rho_i - z_i(s)\mathbf{x}_i(z(s))\right)\mathbf{1}_{\{z_i(s)>0\}}ds \; = \; \rho_i t - \tau_i(t), \tag{20}$$

$$\tau_i(t) \; \geq \; 0 \text{ and } \tau_i(t) \; \leq \; \left(\|C\| + \|\rho\|\right)t. \tag{21}$$

Proof: We present proof from [13] for completeness. It is easy to show that there exists functions $f_n \in \mathcal{C}$ such that $f_n \uparrow \mathbf{1}_{(0,\infty)}$ and f_n' are non-negative. For such functions, (18) implies that

$$\langle f_n, \zeta \rangle \leq \nu_i \langle f_n, \vartheta_i \rangle \left(\int_0^t \mathbf{1}_{\{z_i(s)>0\}} ds \right) \ \leq \ \nu_i \langle f_n, \vartheta_i \rangle t.$$

Now, use of monotone convergence theorem, property of ζ that $\|\langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle\| = 0$, $\langle \mathbf{1}_{\{0\}}, \vartheta_i \rangle = 0$, and ϑ_i being probability density together give us the desired result (i.e., (19)):

$$z_i(t) = \langle \mathbf{1}_{(0,\infty)}, \zeta_i(t) \rangle \leq \nu_i t.$$

Now, we argue for (20). Again, it can be shown that χ can be approximated by sequence of function $f_n \in \mathcal{C}$ so that $f_n \leq \chi$ and $f_n \uparrow \chi, f_n' \uparrow \mathbf{1}_{(0,\infty)}$. Let ζ be a fluid model solution. From (18)

$$\langle f_n, \zeta \rangle = \nu_i \langle f_n, \vartheta_i \rangle \left(\int_0^t \mathbf{1}_{\{z_i(s)>0\}} ds \right) - \int_0^t \langle f_n', \zeta_i(s) \rangle \mathbf{x}_i(z(s)) ds.$$

Taking the limit as $n \to \infty$, monotone convergence theorem and the property of fluid model solution that $\|\langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle\| = 0$ together imply that

$$w_i(t) = \int_0^t \left(\rho_i - z_i(s) \mathbf{x}_i(z(s)) \right) \mathbf{1}_{\{z_i(s)>0\}} ds$$

Finally, we justify (21). Definition of $\tau_i(t)$ implies $\tau_i(t) \geq 0$. Further,

$$\begin{aligned} \tau_i(t) \ &= \ \int_0^t \left(\mathbf{x}_i(z(s)) z_i(s) \mathbf{1}_{z_i(s)>0} + \rho_i \mathbf{1}_{z_i(s)=0} \right) ds \\ &\leq \ \int_0^t \left(\mathbf{x}_i(z(s)) z_i(s) + \rho_i \right) ds. \end{aligned} \tag{22}$$

Now, by property of optimization problem corresponding to rate allocation function $\mathbf{x}_i(\cdot)$, we have $\mathbf{x}_i(z(s)) z_i(s) \leq \max_{j \leq \mathbf{J}} C_j = \|C\|$. Hence, the above inequality gives us desired conclusion

$$\tau_i(t) \leq \left(\|C\| + \|\rho\| \right) t.$$

■

The following is a direct adaptation of Theorem 4.1 in [13] for the scaling described above. For this, let $\mathbb{P}_r^T$ denote the joint distribution of $(\bar{\mathcal{Z}}^r, E^r, \bar{Z}^r, \bar{W}^r, \bar{T}^r, \bar{U}^r)$ restricted to (compact) time interval $[0, T]$. Note that $\mathbb{P}_r^T$ has its support on the product space $\mathbf{D}([0,T], \mathbf{M}^{\mathbf{I}}) \times \mathbf{D}^5([0,T], \mathbb{R}_+^{\mathbf{I}})$ with the appropriately defined topology as described in Section II.

Theorem 5: Fix any $T > 0$. Then, the sequence of probability measures $\mathbb{P}_r^T, r \in \mathbb{N}$ is tight (component-wise and hence with respect to the product topology as well). Hence, any weak limit point is a probability measure on the same space. Under any such weak limit point, with probability 1 the tuple $(\mathcal{Z}, E, z, w, \tau, u)$ is such that $E(t) = \nu t$ and $(\mathcal{Z}(t), z(t), w(t), \tau(t), u(t))$ is a solution to fluid model for all $t \in [0, T]$.

For completeness, we provide some details on proof of Theorem 5 in Appendix, which are based on a direct adaption of arguments for Theorem 4.1 [13]. We refer an interested reader to [13] for a complete treatment.

IV. Main Result

We state and prove the main result of this paper regarding stability of the network operating under α fair utility based rate allocation for strictly positive and sufficiently small α_i, possibly a different α_i for each source i. The general approximate stability result, which is a Corollary of Theorem 6 is stated in the next subsection.

Theorem 6: Consider a sequence of networks $(A, C^r, \nu^r, \vartheta, \mathcal{U}), r \in \mathbb{N}$, as defined in Section III. Further,

(a) there exists $B > 0$ such that $\vartheta_i((B, \infty)) = 0$ for $i \leq \mathbf{I}$;

(b) there exists $\delta > 0$ such that $(1 + \delta) A \rho < C$; and

(c) utility of a flow $i \leq \mathbf{I}$ is $\mathcal{U}_i = \varphi^{\alpha_i}$ so that Assumption 1 is satisfied as well as $\alpha_i < \frac{\delta}{B \mu_i}$.

Then, for any finite $T > 0$ and $\theta > 0$,

$$\liminf_{r\to\infty} \mathbb{P}_r^T \left(\max_{i\leq \mathbf{I}} \sup_{0\leq t\leq T} \bar{Z}_i^r(t) < \theta \right) = 1.$$

Before we dive into the proof of Theorem 6, we explain its consequences. The main claim of Theorem 6 implies that for large enough r, $\bar{Z}^r(\cdot)$ is uniformly close to 0 in the interval $[0, T]$ for any finite T with probability close to 1. Recall that $\bar{Z}^r(\cdot)$, the scaled vector of flows in the system, is equal to the difference between arrivals and departures at any time. Hence, we have that, for the limiting system as $r \to \infty$, the normalized cumulative arrivals is the same as normalized cumulative departures for any time $t \in [0, T]$. That is, the system is rate-stable. The main conditions required to prove the Theorem are uniform boundedness of file-size (or service requirement) by B and the fair utility parameter α_i being small enough (or close to, but strictly greater than, 0).

A. Another Implication of Theorem 6

Theorem 6 can be interpreted as an approximate stability result as well. Before we state a general implication, consider the following example.

Example 1: Suppose ϑ_i be uniform distribution on $[0, B]$ for $i \leq \mathbf{I}$. Then $\mu_i = 2/B$. Then for $\delta = 2$, condition (c) is satisfied for any $\alpha_i \in (0, 1)$. That is, the Theorem 6 proves stability of any ρ such that $3A\rho < C$. Thus, Theorem 6 implies 1/3-approximation of stability for the uniform distribution with bounded file-size.

Next, we state the implication of Theorem 6 that essentially shows that the system is $1/(1+\alpha)$-approximate stable when all $\alpha_i = \alpha \in (0, \infty)$ for *any* system with bounded file size distribution.

Corollary 7: Consider a sequence of networks $(A, C^r, \nu^r, \vartheta, \mathcal{U}), r \in \mathbb{N}$, as defined in Section III. Further,

(d) there exists $0 < b \leq B < \infty$ such that $\vartheta_i([0, b) \cup (B, \infty)) = 0$ for all $i \leq \mathbf{I}$;
(e) utility of flow $i \leq \mathbf{I}$ is $\mathcal{U}_i = \varphi^\alpha$ for $\alpha \in (0, \infty)$; and
(f) $(1 + \alpha)A\rho < C$.

Then, for any finite $T > 0$ and $\theta > 0$,

$$\liminf_{r\to\infty} \mathbb{P}_r^T \left(\max_{i\leq \mathbf{I}} \sup_{0\leq t\leq T} \bar{Z}_i^r(t) < \theta \right) = 1.$$

Proof: The condition (f) implies that there exists an $\varepsilon > 0$ such that

$$(1 + \varepsilon)(1 + \alpha)A\rho \quad < \quad C. \tag{23}$$

That is

$$(1 + \delta)A\rho \quad < \quad C, \tag{24}$$

where $\delta = \alpha(1+\varepsilon)+\varepsilon$. Now define interval $I_k = [b(1+\varepsilon)^k, b(1+\varepsilon)^{k+1})$. Let $K_\varepsilon = \lceil \log(B/b)/\log(1+\varepsilon) \rceil$. Then

$$[b, B) \subset \cup_{k=0}^{K_\varepsilon} I_k.$$

Now consider ϑ_i any $i \leq \mathbf{I}$. Since the support of ϑ_i is contained in $[b, B)$ we can write ϑ_i as follows.

$$\vartheta_i = \sum_{k=0}^{K_\varepsilon} p_{ik} \vartheta_{ik},$$

where

$$p_{ik} = \langle \mathbf{1}_{I_k}, \vartheta_i \rangle \quad \text{and} \quad \vartheta_{ik} = p_{ik}^{-1} \vartheta_i \mathbf{1}_{I_k}.$$

Also, define $\nu_{ik} = \nu_i p_{ik}$, $\mu_{ik}^{-1} = \langle \chi, \vartheta_{ik} \rangle$ and $\hat{\rho}_{ik} = \nu_{ik}/\mu_{ik}$. For completeness, we adopt notation that if $p_{ik} = 0$ then $\vartheta_{ik} = \mathbf{0},\ \hat{\rho}_{ik} = 0$.

The above suggests that flow of type i with parameters ν_i, ϑ_i is equivalent to K_ε different flows, denoted by flow $(i, k),\ 0 \leq k \leq K_\varepsilon$ with parameters $(\nu_{ik}, \vartheta_{ik})_{0\leq k\leq K_\varepsilon}$. That is the original system with $\mathbf{I}$ flows is equivalent to $K_\varepsilon \mathbf{I}$ flows. The $\mathbf{J} \times \mathbf{I}$ routing matrix A naturally extends to $\mathbf{J} \times K_\varepsilon \mathbf{I}$ matrix $\hat{A}$. Then, (24) implies that

$$(1 + \delta)\hat{A}\hat{\rho} \quad < \quad C. \tag{25}$$

We want to remind reader that this newly created system with K_ε times more flows has all the stochastic properties of the original system - primarily the arrival process and service requirement process of each satisfy the functional law of large numbers (as stated in Lemma 11). This can be checked easily given how the construction of new system is done from the original system. Now to complete the proof it is sufficient to show that this new system satisfies the conditions of Theorem 6.

To this end, consider conditions (a)-(c) of Theorem 6. Given (25) it is straightforward to check that conditions (a)-(b) are satisfied and $\mathcal{U}_i = \varphi^\alpha \in (0,\infty)$ for all $i \leq \mathbf{I}$ satisfy Assumption 1 (i.e. all $K_\varepsilon \mathbf{I}$ flows satisfy it as well). Thus we are required to check the second part of condition (c). Now for flow $(i,k), 0 \leq k \leq K_\varepsilon, i \leq \mathbf{I}$, the bound on service requirement is $B_k \overset{\triangle}{=} b(1+\varepsilon)^{k+1}$ while support of ϑ_{ik} is on interval $[b(1+\varepsilon)^k, b(1+\varepsilon)^{k+1})$. Hence, $\mu_{ik}^{-1} \in [b(1+\varepsilon)^k, b(1+\varepsilon)^{k+1})$. That is,

$$\frac{1}{1+\varepsilon} \;\leq\; \frac{1}{\mu_{ik}B_k}. \tag{26}$$

Now the definition of δ in (25) and (26) imply the following.

$$\begin{aligned}
\alpha &< \alpha + \frac{\varepsilon}{1+\varepsilon} \\
&= \frac{\alpha(1+\varepsilon)+\varepsilon}{1+\varepsilon} \\
&= \frac{\delta}{1+\varepsilon} \\
&\leq \frac{\delta}{\mu_{ik}B_k}.
\end{aligned} \tag{27}$$

The (27) completes the verification of the condition (c) of Theorem 6. Given that the sum of the number of flows of $(i,k), 0 \leq k \leq K_\varepsilon$ is the same as the number of flows of type i, conclusion of Theorem 6 implies the desired conclusion of Corollary 7 and thus completes its proof. ■

B. Proof of Theorem 6

We will use Theorem 5 crucially to obtain proof of Theorem 6. We state the following result about the fluid model solutions.

Lemma 8: Consider any system satisfying conditions (a)-(c) of Theorem 6. Let $\mathcal{Z}$ be corresponding fluid model solution with auxiliary variables (z, w, τ, u), and $\mathcal{Z}(0) = \mathbf{0}$ since system starts empty. Then, for any $t \in [0,T]$,

$$\max_{i \leq \mathbf{I}} z_i(t) = 0,$$

where recall that $z_i(t) = \langle \mathbf{1}, \mathcal{Z}_i(t)\rangle$.

The proof of the Lemma 8 will be presented in the next sub-section. First, we use it to complete the proof of Theorem 6. To this end, note that $\mathcal{Z}(\cdot)$ is sufficient to define the fluid model solution with auxiliary variables.

For $\theta > 0$, let $A_\theta = \{\mathcal{Z}(\cdot) : \max_{i\leq \mathbf{I}} \sup_{0\leq t \leq T} \langle \mathbf{1}, \mathcal{Z}_i(t)\rangle < \theta\}$. Then, we claim that A_θ is open in $\mathbf{D}([0,T],\mathbf{M}^\mathbf{I})$. It is justified as follows. Consider $B_\theta = A_\theta^c$. It is sufficient to show that B_θ is closed. Equivalently, it is sufficient to show that if $\zeta^k \to \zeta$ with $\{\zeta^k\} \subset B_\theta$ then $\zeta \in B_\theta$. Since the topology is induced by metric d^o, we have that $d^o(\zeta^k, \zeta) \to 0$.

Proposition 9: For any $\zeta, \xi \in \mathbf{D}([0,T],\mathbf{M}^\mathbf{I})$, let there be B (same as in Theorem 6(a)) such that

$$\zeta_i((B,\infty)) = \xi_i((B,\infty)) = 0.$$

Then,

$$\left|\max_{i\leq\mathbf{I}} \sup_t \langle \mathbf{1}, \zeta_i(t)\rangle - \max_{i\leq\mathbf{I}} \sup_t \langle \mathbf{1}, \xi_i(t)\rangle\right| \leq d^o(\zeta,\xi).$$

Proof: Recall from Section II that

$$d^o(\zeta,\xi) = \inf_{\varphi\in\Phi_b}\left\{\|\varphi\|^o \vee \left(\sup_{t\in[0,T]} \mathbf{d_I}(\zeta(t),\xi(t))\right)\right\}.$$

Note that by definition of Φ_b, all $\varphi \in \Phi_b$ must be continuous in addition to being nondecreasing and $\varphi(0) = 0, \varphi(T) = T$. Hence, every $\varphi \in \Phi_b$ map $[0,T]$ onto $[0,T]$. Given $\delta > 0$ there exists $\varphi \in \Phi_b$ such that

$$\max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \mathbf{d}[\zeta_i(\varphi(t)), \xi_i(t)] \leq d^o(\zeta, \xi) + \delta.$$

Let $\ell_\delta = d^o(\zeta, \xi) + \delta$. Then from above and definition of Prohov's metric $\mathbf{d}(\cdot,\cdot)$ imply that for any Borel set S

$$\zeta_i(\varphi(t))(S) \leq \xi_i(t)(S^{\ell_\delta}) + \ell_\delta; \quad \xi_i(t)(S) \leq \zeta_i(\varphi(t))(S^{\ell_\delta}) + \ell_\delta.$$

Now, since there exists B such that $\zeta_i((B,\infty)) = \xi_i((B,\infty)) = 0$, we have that for $S = [0, B + 2\ell_\delta]$

$$\zeta_i(\varphi(t))(S) = \zeta_i(\varphi(t))(S^{\ell_\delta}); \quad \xi_i(t)(S) = \xi_i(t)(S^{\ell_\delta}).$$

Further, for such choice of S

$$\zeta_i(\varphi(t))(S) = \langle \mathbf{1}, \zeta_i(\varphi(t)) \rangle; \quad \xi_i(t)(S) = \langle \mathbf{1}, \xi_i(t) \rangle.$$

Putting above together, we have that

$$\langle \mathbf{1}, \zeta_i(\varphi(t)) \rangle \leq \langle \mathbf{1}, \xi_i(t) \rangle + \ell_\delta; \quad \langle \mathbf{1}, \xi_i(t) \rangle \leq \langle \mathbf{1}, \zeta_i(\varphi(t)) \rangle + \ell_\delta.$$

Now since φ maps $[0,T]$ onto $[0,T]$, the above implies that

$$\left| \max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \langle \mathbf{1}, \zeta_i(t) \rangle - \max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \langle \mathbf{1}, \xi_i(t) \rangle \right| \leq \ell_\delta \ = \ d^o(\zeta, \xi) + \delta.$$

Since $\delta > 0$ is arbitrary, we conclude that

$$\left| \max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \langle \mathbf{1}, \zeta_i(t) \rangle - \max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \langle \mathbf{1}, \xi_i(t) \rangle \right| \leq d^o(\zeta, \xi).$$

■

From Proposition 9, we obtain that if $\zeta^k \to \zeta$ with $\{\zeta^k\} \subset \mathbf{D}([0,T], \mathbf{M}^{\mathbf{I}})$ then under hypothesis of Theorem 6,

$$\max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \langle \mathbf{1}, \zeta_i^k(t) \rangle \to \max_{i \leq \mathbf{I}} \sup_{t \in [0,T]} \langle \mathbf{1}, \zeta_i(t) \rangle.$$

Hence, if $\{\zeta^k\} \subset B_\theta$ then $\zeta \in B_\theta$. That is, B_θ is closed and hence A_θ is open.

Now suppose Theorem 6 is false for a given $\theta > 0$. Then, from above discussion it must be that there is a sequence $r_q, q \in \mathbb{N}$, $r_q \to \infty$, such that $\mathbb{P}^T_{r_q}(A_\theta) < 1$ for all q. By Theorem 5, there exists a further subsequence $r_{q_m}, m \in \mathbb{N}$ of $r_q, q \in \mathbb{N}$, so that $\mathbb{P}^T_{r_{q_m}}$ converges to some $\mathbb{P}^T_\star$ under which the system satisfies fluid model solution with probability 1. By Lemma 8, we have that, for any $\theta > 0$,

$$\mathbb{P}^T_\star(A_\theta) = 1.$$

By Portmantau's characterization of weak-convergence and A_θ being open we have that

$$\liminf_{r_{q_m}} \mathbb{P}^T_{r_{q_m}}(A_\theta) \quad \geq \quad \mathbb{P}^T_\star(A_\theta) = 1. \tag{28}$$

This contradicts our assumption that Theorem 6 is false. This completes the proof of Theorem 6.

C. *Proof of Lemma 8*

Consider a system satisfying hypothesis (a)-(c) of the Theorem 6. Let $\mathcal{Z}$ be a fluid model solution with its auxiliary variables (z, w, τ, u), and $\mathcal{Z}(0) = \mathbf{0}$. Let $y_i(t) = (1+\delta)\rho_i / z_i(t)$ for $i \leq \mathbf{I}$. Define the following Lyapunov function

$$L(t) = \sum_{i \leq \mathbf{I}} L_i(t), \quad \text{where} \quad L_i(t) = w_i(t) \mathcal{U}_i'(y_i(t)).$$

Now

$$\mathcal{U}_i'(x) = x^{-\alpha_i} \ \text{ and } \ \mathcal{U}_i''(x) = -\alpha_i x^{-1-\alpha_i}.$$

In what follows, we wish to upper bound $\limsup_{h\to 0^+} \frac{L(t+h)-L(t)}{h}$ for all t. By Fatou's Lemma,

$$\limsup_{h\to 0^+} \frac{L(t+h)-L(t)}{h} \quad \leq \quad \sum_{i\leq \mathbf{I}} \limsup_{h\to 0^+} \frac{L_i(t+h)-L_i(t)}{h}. \tag{29}$$

Next, we bound $\limsup_{h\to 0^+} \frac{L_i(t+h)-L_i(t)}{h}$. To this end, replacing the value of $\mathcal{U}_i'(\cdot)$ and using simple manipulation give us

$$\frac{L_i(t+h)-L_i(t)}{h} \quad = \quad y_i^{-\alpha_i}(t)\frac{w_i(t+h)-w_i(t)}{h} + w_i(t+h)\frac{y_i^{-\alpha_i}(t+h)-y^{-\alpha_i}(t)}{h}. \tag{30}$$

Next, we bound (30) as $h \to 0^+$ in many steps as follows.

Step 1. Bound on $y_i^{-\alpha_i}(t)$*:* We have $y_i^{-\alpha_i}(t) = z_i^{\alpha_i}(t)(1+\delta)^{-\alpha_i}\rho_i^{-\alpha_i}$. For any $t \in [0,T]$, (19) of Lemma 4 imply that $z_i(t) \leq \nu_i t \leq \nu_i T$. Putting this together, we have that, for any $t \in [0,T]$,

$$y_i^{-\alpha_i}(t) \quad \leq \quad (1+\delta)^{-\alpha_i}\rho_i^{-\alpha_i}\nu_i^{\alpha_i}T^{\alpha_i} \quad \stackrel{\triangle}{=} K_1^T. \tag{31}$$

Step 2. Bound on $w_i(t)$*:* From (20) of Lemma 4, for any $t \in [0,T]$, we have

$$w_i(t) \quad \leq \quad \rho_i t \quad \leq \quad \rho_i T. \tag{32}$$

Step 3. Bound on $\limsup_{h\to 0^+} \frac{w_i(t+h)-w_i(t)}{h}$*:* The (20) and (21) of Lemma 4 imply that $w_i(\cdot)$ is a Lipschitz continuous function with constant $(\|C\|+\|\rho\|)$. It is well-known that Lipschitz continuous function are differentiable almost everywhere. Since we have finite $\mathbf{I}$, we have that all w_i, $i \leq \mathbf{I}$, are differentiable almost everywhere. Such t are called regular points. At such t, the term $\limsup_{h\to 0^+} \frac{w_i(t+h)-w_i(t)}{h} = \frac{dw_i(t)}{dt}$. From Lemma 4, for such regular point t, we have

$$\limsup_{h\to 0^+} \frac{w_i(t+h)-w_i(t)}{h} \quad = \quad \frac{dw_i(t)}{dt} \quad = \quad (\rho_i - x_i(z(t))z_i(t))\,\mathbf{1}_{z_i(t)>0} \;\leq\; \rho_i - x_i(z(t))z_i(t). \tag{33}$$

Here the last inequality follows from the fact that, for $z_i(t) = 0$, $x_i(z(t))z_i(t) = 0$. Note that Lipschitz continuity of $w_i(\cdot)$ implies that, for all t, we have $\limsup_{h\to 0^+} \frac{w_i(t+h)-w_i(t)}{h} \leq (\|C\|+\|\rho\|)$ and $\lim_{h\to 0^+} w_i(t+h) = w_i(t)$.

Step 4. Bound on $\limsup_{h\to 0^+} \frac{y_i^{-\alpha_i}(t+h)-y^{-\alpha_i}(t)}{h}$*:* Consider the following.

$$y_i^{-\alpha_i}(t+h) - y^{-\alpha_i}(t) \quad = \quad \frac{z_i^{\alpha_i}(t+h) - z_i^{\alpha_i}(t)}{(1+\delta)^{\alpha_i}\rho_i^{\alpha_i}} \quad \leq \quad \frac{(z_i(t)+h\nu_i)^{\alpha_i} - z_i^{\alpha_i}(t)}{(1+\delta)^{\alpha_i}\rho_i^{\alpha_i}}, \tag{34}$$

where the last inequality follows from Lemma 4. Taking $h \to 0^+$ in (34), we obtain

$$\limsup_{h\to 0^+} \frac{y_i^{-\alpha_i}(t+h)-y^{-\alpha_i}(t)}{h} \quad \leq \quad \frac{\alpha_i\nu_i}{(1+\delta)^{\alpha_i}\rho_i^{\alpha_i}z_i^{1-\alpha_i}(t)} \quad = \quad \frac{\alpha_i\nu_i y_i^{1-\alpha_i}}{(1+\delta)\rho_i}. \tag{35}$$

Using the bounds from Steps 1-4, we obtain the following: for almost every t,

$$\limsup_{h\to 0^+} \frac{L(t+h)-L(t)}{h} \quad \leq \quad \sum_{i\leq\mathbf{I}} \left[(\rho_i - x_i(z(t))z_i(t))y_i^{-\alpha_i}(t) + \frac{w_i(t)\alpha_i\nu_i y_i^{1-\alpha_i}}{(1+\delta)\rho_i}\right]. \tag{36}$$

Further, for any $t \in [0,T]$, there exists a finite constant $K_2^T < \infty$ such that

$$\limsup_{h\to 0^+} \frac{L(t+h)-L(t)}{h} \quad \leq \quad K_2^T. \tag{37}$$

Next, we study the bound on the right hand side (RHS) of (36) with goal of establishing it to be negative if

$\max_{i \leq \mathbf{I}} z_i(t)$ is positive. To this end let $z_i(t) > 0$. Consider term $\sum_{i \leq \mathbf{I}} (\rho_i - x_i(z(t))z_i(t)) y_i^{-\alpha_i}(t)$. For an $i \leq \mathbf{I}$,

$$\begin{aligned} (\rho_i - x_i(z(t))z_i(t))\, y_i^{-\alpha_i}(t) &= -\delta\rho_i y_i^{-\alpha_i}(t) + z_i(t) \left(\frac{(1+\delta)\rho_i}{z_i(t)} - x_i(z(t)) \right) y_i^{-\alpha_i}(t) \\ &= -\delta\rho_i y_i^{-\alpha_i}(t) + z_i(t)\, (y_i(t) - x_i(z(t)))\, y_i^{-\alpha_i}(t). \end{aligned} \tag{38}$$

From the hypothesis of Theorem 6, $(1+\delta)A\rho < C$. Hence, the vector $y(t) = (y_1(t), \ldots, y_{\mathbf{I}}(t))$ is a feasible rate allocation. Now, utility function of i^{th} flow corresponds to α_i fair utility. Hence, it is strictly concave as discussed earlier. That is, given $z(t)$, the rate allocation vector $x(z(t))$ is unique and satisfies the zero gradient condition. From this, using standard argument it follows that

$$\sum_{i \leq \mathbf{I}} z_i(t)(y_i(t) - x_i(z(t))) y_i^{-\alpha_i}(t) \quad \leq \quad 0. \tag{39}$$

Therefore,

$$\sum_{i \leq \mathbf{I}} (\rho_i - x_i(z(t))z_i(t)) y_i^{-\alpha_i}(t) \quad \leq \quad -\delta \sum_{i \leq \mathbf{I}} \rho_i y_i^{-\alpha_i}. \tag{40}$$

Now, the term $\sum_{i \leq \mathbf{I}} \frac{w_i(t)\alpha_i\nu_i y_i^{1-\alpha_i}(t)}{(1+\delta)\rho_i}$. For this, note that we have $\vartheta_i((B,\infty)) = 0$ from the hypothesis of Theorem 6. Subsequently, $\langle \mathbf{1}_{(B,\infty)}, \mathcal{Z}_i(t) \rangle = 0$ for all $i \leq \mathbf{I}$ and all $t > 0$. Hence,

$$w_i(t) \;=\; \langle \chi, \mathcal{Z}_i(t) \rangle \;\leq\; B\langle \mathbf{1}, \mathcal{Z}_i(t) \rangle \;=\; Bz_i(t). \tag{41}$$

Using (41) and recalling the definition of $y_i(t)$, we obtain

$$\sum_{i \leq \mathbf{I}} \frac{w_i(t)\alpha_i\nu_i y_i^{1-\alpha_i}(t)}{(1+\delta)\rho_i} \quad \leq \quad \sum_{i \leq \mathbf{I}} B\alpha_i\nu_i y_i^{-\alpha_i}. \tag{42}$$

Combining (40) and (42) in (36), we obtain that, for almost all $t \in [0,T]$,

$$\limsup_{h \to 0^+} \frac{L(t+h) - L(t)}{h} \quad \leq \quad -\sum_{i \leq \mathbf{I}} (\delta\rho_i - \alpha_i B\nu_i) y_i^{-\alpha_i}(t) \quad = \quad -\sum_{i \leq \mathbf{I}} \frac{(\delta\rho_i - \alpha_i B\nu_i) z_i^{\alpha_i}}{(1+\delta)^{\alpha_i} \rho_i^{\alpha_i}}. \tag{43}$$

By property of fluid model solution, we have $\|\langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle\| = 0$ for all $t \in [0,T]$. Also, $w_i(t) \leq Bz_i(t)$. Hence, we have that

$$L(t) = 0 \Leftrightarrow z(t) = \mathbf{0}.$$

This can be justified as follows. Define $f_m = \mathbf{1}_{[0,\frac{1}{m}]}$. By definition $f_m \to \mathbf{1}_{\{0\}}$ point-wise as $m \to \infty$. Further, $f_m \leq \mathbf{1}_{[0,1]}$ for $m \geq 1$ and $\zeta(t)$ is non-negative measure with $\langle \mathbf{1}_{[0,1]}, \zeta(t) \rangle < \infty$ by definition of fluid model solution. Then, Dominated convergence theorem implies that $\langle f_m, \zeta(t) \rangle \to \langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle$. But $\langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle = 0$ for all $t \in [0,T]$ by property of fluid model solution. That is,

$$\lim_{m \to \infty} \langle f_m, \zeta(t) \rangle = 0.$$

Equivalent, for any $\varepsilon > 0$ there exists $m(\varepsilon)$ such that for all $m \geq m(\varepsilon)$, $\langle f_m, \zeta(t) \rangle \leq \varepsilon$. Now, $w(t) = 0$ implies $\langle \chi, \zeta(t) \rangle = 0$. Consider the following (with $m \geq m(\varepsilon)$)

$$\begin{aligned} z(t) &= \langle \mathbf{1}, \zeta(t) \rangle \;=\; \langle f_m, \zeta(t) \rangle + \langle \mathbf{1}_{(\frac{1}{m},\infty)}, \zeta(t) \rangle \\ &\leq \varepsilon + m\langle \chi, \zeta(t) \rangle \\ &= \varepsilon. \end{aligned} \tag{44}$$

Thus, $z(t) \leq \varepsilon$ for any $\varepsilon > 0$ if $w(t) = 0$. This completes the proof that if $w(t) = 0$ then $z(t) = 0$. That is, if $L(t) = 0$ then $z(t) = 0$ given $\langle \mathbf{1}_{\{0\}}, \zeta(t) \rangle = 0$.

In summary, we have the following: (1) for almost all $t \in [0,T]$, $\limsup_{h \to 0^+} \frac{L(t+h)-L(t)}{h} < 0$ if $L(t) > 0$; (2) $\limsup_{h \to 0^+} \frac{L(t+h)-L(t)}{h} \leq K_2^T$ for all $t \in [0,T]$. Given (1) and (2), a simple analysis in Lemma 10 in the next subsection implies that $L(t) = 0$ for all $t \in [0,T]$ given that $L(0) = 0$. This immediately implies that $z(t) = \mathbf{0}$ for all $t \in [0,T]$. This completes the proof of Lemma 8.

D. Remaining Lemma

We state and prove the following remaining lemma used in proving Lemma 8. This result is standard analytic result and can be found in literature in different guises.

Lemma 10: Let $f : [0,T] \to \mathbb{R}_+$ be any measurable function with properties: (1) for almost all $t \in [0,T]$, $\limsup_{h\to 0^+} \frac{f(t+h)-f(t)}{h} < 0$ if $f(t) > 0$, and (2) $\limsup_{h\to 0^+} \frac{f(t+h)-f(t)}{h} < A$ for all $t \in [0,T]$ with some finite $A > 0$. Let $f(0) = 0$. Then $f(t) = 0$ for all $t \in [0,T]$.

Proof: In what follows, we use the standard properties of continuous functions, Fatou's Lemma, and the fact that $g(x) \overset{\triangle}{=} \limsup_{n\to\infty} n(f(x+1/n) - f(x))$ is measurable function if f is measurable. Define $F(t) = f^2(t)$ and consider the following:

$$\begin{aligned} F(t) - F(s) &= \lim_{n\to\infty} n \left[\int_t^{t+1/n} F(z)dz - \int_s^{s+1/n} F(z)dz \right] \\ &= \lim_{n\to\infty} \int_s^t n\left(F(z+1/n) - F(z)\right) dz \\ &\leq \int_s^t \limsup_{n\to\infty} n\left(F(z+1/n) - F(z)\right) dz \\ &= \int_s^t 2f(z) \limsup_{n\to\infty} n\left(f(z+1/n) - f(z)\right) dz \\ &= \int_s^t 2f(z)g(z)dz. \end{aligned} \tag{45}$$

Let $B = \{u \in [s,t] :$ Condition (1) holds at $u\ \}$. By the hypothesis of the Lemma, we have the Lebesgue measure of B, $\mu(B) = t - s$, and $\mu(B^c) = 0$ where $B^c = [s,t] - B$. The hypothesis of the Lemma implies that

$$\begin{aligned} & f(z)g(z)\mathbf{1}_{\{z\in B\}} = f(z)g(z)\mathbf{1}_{\{z\in B; f(z)>0\}} \ \leq \ 0, \\ & f(z)g(z)\mathbf{1}_{\{z\notin B\}} < Af(z). \end{aligned} \tag{46}$$

Therefore,

$$\begin{aligned} F(t) - F(s) &\leq 2\int_s^t f(z)g(z)\mathbf{1}_{\{z\in B\}}dz + 2\int_s^t f(z)g(z)\mathbf{1}_{\{z\notin B\}}dz \\ &\leq 0\mu(B) + A\|f\|_T \mu(B^c) \ = \ 0. \end{aligned} \tag{47}$$

Here we used the fact that in $[0,T]$ the function f is bounded above by AT from the hypothesis of the Lemma. Now using $F(0) = 0$ and replacing $s = 0$ in (47), we get $F(t) \leq 0$ for all $t \in [0,T]$. But by definition $F(t) \geq 0$. That is, $F(t) = 0$. Equivalently, $f(t) = 0$ for all $t \in [0,T]$. ■

V. Concluding Remarks and Future Work

Deterministic versions of NUM and its generalizations have been extensively used in many network designs recently. However, most results on stochastic stability of NUM rely on the assumption of exponentially distributed file sizes. In this paper, we have established the stability of network operating under α-fair rate allocation with general file size distributions, when the α corresponding to each flow is close to 0 and the service requirement has bounded size. In addition, our results imply $1/(1+\alpha)$-approximate stability of network with any α-fair utility under general file size distribution. Our method was based on Lyapunov function analysis for fluid model solution of the scaled system. Due to different scaling, we could establish fluid model (and subsequently stability) for heterogeneous α-fair utilities for different flows. Our Lyapunov function is naturally valid for the fluid model scaling of Gromoll and Williams [13] since the fluid model solutions are identical.

It is straightforward to extend Theorems 5 and 6 with same utility functions and convex constraints such that $\mathbf{0}$ is a feasible point under constraints. Extending these results for the case of general set of concave utilities beyond α-fair would also be an interesting task.

The special cases of $\alpha = \infty$ and $\alpha = 1$ have recently been tackled in the preprints of [5] and [22], respectively. In contrast, this paper provides guarantees for a continuum of $\alpha \in (0, \infty)$, including stability for heterogeneous and sufficiently small α_i. However, stability for all $\alpha \in (0, \infty)$ and general file size distribution is still open. Note that we have ignored any dynamical information about the fluid quantity $z(t)$ by upper bounding its rate of change by ν. Further progress can be made by studying the details of the dynamics of $z(t)$, which can be obtained based on the fluid model solution of Theorem 5.

Acknowledgements

It is also our pleasure to thank H. Christian Gromoll and Ruth Williams for providing us with the preprint of their paper [13]. Devavrat Shah is grateful to Ruth Williams for her detailed reading of the initial draft and suggestions that has led to much improvement in the readability and quality of this paper.

References

[1] P. Billingsley. Convergence of Probability Measures. *John Wiley & Sons, Inc.*, Second edition, 1999.

[2] T. Bonald and L. Massoulie. Impact of Fairness on Internet Performance. *Proceedings of ACM Sigmetrics*, 2001.

[3] T. Bonald, L. Masoulie, A. Proutiere, J. Virtamo, A Queueing Analysis of Max-Min Fairness, Proportional Fairness and Balanced Fairness. To appear in *Queueing Systems: Theory and Applications*, 2006.

[4] T. Bonald and A. Proutiere. Insensitive Bandwidth Sharing in Data Networks. *Queueing systems*, 44(1):69-100, 2003.

[5] M. Bramson. Stability of networks for max-min fair routing. Presentation at the 13th INFORMS Applied Probability Conference, Ottawa, 2005.

[6] M. Chiang, S. H. Low, and A. R. Calderbank, and J. C. Doyle. Layering as optimization decomposition. To appear in *Proceedings of IEEE*, January 2007. Shorter version in *Proc. Conf. Inform. Sciences and Syst.*, Princeton, March 2006.

[7] J. G. Dai. On Positive Harris Recurrence of Multiclass Queueing Networks: A Unified Approach Via Fluid Limit Models. *Annals of Applied Probability*, vol. 5, pp. 49-77, 1995.

[8] D.A. Dawson. Measure-valued Markov processes. *In P.L. Hennequin, editor, cole d't de probabilits de Saint Flour XXI-1991*, Volume 1541 of Lecture Notes Math., pages 1-260. Springer-Verlag, Berlin, 1993.

[9] G. de Veciana, T. Lee and T. Konstantopoulos. Stability and Performance Analysis of Networks Supporting Elastic Services. *IEEE/ACM Transactions on Networking*, 9(1):2-14, February 2001.

[10] S. N. Ethier and T. G. Kurtz. Markov Processes: Characterization and Convergence. In Wiley, New York, 1986.

[11] H. C. Gromoll. Diffusion approximation for a processor sharing queue in heavy traffic. *Annals of Applied Probability*, 14 : 555-611, 2004.

[12] H. C. Gromoll, A. Puha and R. Williams. Fluid Limit of a Processor Sharing Queue. *Annals of Applied Probability*, 2002.

[13] H. C. Gromoll and R. Williams. Fluid Limit of a Network With Fair Bandwidth Sharing and General Document Size Distribution. *Preprint and Personal Communication*, 2006.

[14] O. Kallenberg. Random Measures. *Academic Press*, New York, 1986.

[15] F. P. Kelly, A. Maullo, and D. Tan. Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Hournal of the Operational Research Society*, vol. 49, pp. 237-252, 1998.

[16] F. P. Kelly and R. J. Williams. Fluid model for a network operating under a fair bandwidth-sharing policy. *The Annals of Applied Probability*, vol. 14, no. 3, pp. 1055-1083, 2004.

[17] P. Key and L. Massoulie. Fluid Limits and Diffusion Approximzations for Integrated Traffic Models. To appear in *Queueing Systems: Theory and Applications*, 2006.

[18] A. Lakshmikantha, C. L. Beck and R. Srikant. Connection level stability analysis of the Internet using the sum of squares (sos) techniques. *Proceedings of Conference on Information Sciences and Systems*, Princeton, 2004.

[19] X. Lin and N. Shroff. On the Stability Region of Congestion Control. *Proceedings of Allerton Conference*, 2004.

[20] J. Liu, M. Chiang, and H. V. Poor. Stochastic stability of general optimization-based network resource allocation. *Preprint*, March 2006.

[21] S. H. Low. A duality model of TCP and queue management algorithms. *IEEE/ACM Trans. on Networking*, vol. 11, no. 4, pp. 525-536, Aug. 2003.

[22] L. Massoulie. Structural Properties of Proportional Fairness: Stability And Insensitivity. *Submitted*, 2006.

[23] J. Mo and J. Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Transactions on Networking*, 8(5):556–567, October 2000.

[24] A. L. Puha and R. J. Williams. Invariant states and rates of convergence for a critical fluid model of a processor sharing queue. *Annals of Applied Probability*, 14 : 517-554, 2004.

[25] R. T. Rockafellar. Network Flows and Monotropic Optimization. *John Wiley & Sons, Inc.*, 1984.

[26] J. Roberts and L. Massoulie. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15: 185-201, 2000.

[27] R. Srikant. The Mathematics of Internet Congestion Control. Birkhauser, 2004.

[28] R. Srikant. On the Positive Recurrence of a Markov Chain Describing File Arrivals and Departures in a Congestion-controlled Network. *IEEE Computer Communication Workshop*, 2004.

[29] A. Tang, J. Wang, and S. H. Low. Counter-intuitive throughput behaviors in networks under end-to-end control. *IEEE/ACM Transactions on Networking*, 14(2):355-368, April 2006.

[30] H. Varian. Microeconomic analysis. W. W. Norton & Company, 3 edition, 1992.

[31] H. Ye. Stability of Data Networks under an Optimization-Based Bandwidth Allocation. *IEEE Transactions on Automatic Control*, 48(7): 1238-1242, July 2003.

[32] H. Ye, J. Qu, X. Yuan. Stability of Data Networks: Starionary and Bursty Models. *Operations Research*, vol. 53, pp. 107-125, 2005.

Appendix
Fluid Model Justification

For completeness, we present a proof of Theorem 5 and thus justify fluid model solutions in this appendix. As explained earlier, although our scaling is somewhat different, the proof is essentially the same as that of Theorem 4.1 of [13]. Hence, we only provide key steps of the proof. In what follows, whenever details are missing we provide specific pointers in [13] for an interested reader. For a curious reader we make a note that the fluid model techniques used in [13] have also been used in other similar contexts, e.g., [12], [24], [11]. The philosophy of using measure valued descriptor for dynamical systems dates way back. Lecture notes by Dawson [8] is a good reference for this.

Theorem 5 requires establishing two key properties: (1) Tightness of probability measures $\mathbb{P}_r^T, r \in \mathbb{N}$, and (2) Almost sure deterministic dynamics under any weak limit of these measures. We go through various key steps to first establish (1) and then (2) in the following sections.

A. Limit of Primitive Processes

This section looks at the limiting primitive processes, the arrival process, and service process. Given the r^{th} system, recall that

$$\mathcal{L}_i^r(t) = \sum_{k=1}^{E_i^r(t)} \delta_{V_{ik}}, \quad i \leq \mathbf{I}.$$

Let $\nu(t) = \nu t$ and $\rho(t) = \rho t$ with notation $\nu = (\nu_1, \ldots, \nu_{\mathbf{I}})$ and $\rho = (\rho_1, \ldots, \rho_{\mathbf{I}})$. Define the fluid scaled quantity

$$\bar{\mathcal{L}}_i^r(t) = \frac{1}{r}\mathcal{L}_i^r(t).$$

Now $\bar{\mathcal{L}}^r(\cdot) = (\bar{\mathcal{L}}_1^r(\cdot), \ldots, \bar{\mathcal{L}}_{\mathbf{I}}^r(\cdot))$ is a function from $[0, \infty)$ to $\mathbf{M}^{\mathbf{I}}$. Under stochastic assumption that arrival process and service requirement process are i.i.d., the standard functional law of large numbers lead to the following result.

Lemma 11 (Theorem 5.1,[13]): As $r \to \infty$,

$$\left(\bar{\mathcal{L}}^r(\cdot), \langle \chi, \bar{\mathcal{L}}^r(\cdot)\rangle\right) \Rightarrow \left(\nu(\cdot)\vartheta, \rho(\cdot)\right).$$

Here convergence in Lemma 11 is uniformly on compact time intervals on space $\mathbf{D}([0,\infty), \mathbf{M}^{\mathbf{I}})$. We skip the proof and reader can see Appendix in [13] (or any other known standard arguments, e.g. [1])

B. Dynamic Equations

This section provides an important characterization of system dynamics that will be crucial in proving the characterization of fluid model solutions. The result stated next is a clone of Lemma 5.2 [13] for our scaling.

Lemma 12: Fix an $r > 0$. Consider any $i \leq \mathbf{I}$ and $f \in \mathcal{C}_c$. Then almost surely, for all $[s,t] \subset [0,\infty)$ satisfying $\inf_{u \in [s,t]} \bar{Z}_i^r(u) > 0$,

$$\langle f, \bar{\mathcal{Z}}_i^r(t)\rangle = \langle f, \bar{\mathcal{Z}}_i^r(s)\rangle - \int_s^t \langle f', \bar{\mathcal{Z}}_i^r(u)\rangle \mathbf{x}_i(\bar{Z}^r(u))du + \langle f, \bar{\mathcal{L}}_i^r(t)\rangle - \langle f, \bar{\mathcal{L}}_i^r(s)\rangle.$$

Proof: We first connect the dynamics of the r^{th} system to the fluid scaled system. For this, recall that system is described completely by $\mathcal{Z}^r(\cdot)$. Now,

$$\mathcal{Z}_i^r(t) = \sum_{k=1}^{E_i^r(t)} \boldsymbol{\delta}^+_{V_{ik}(t)}.$$

Consider any two points $0 \leq q < t$. From definition,

$$V_{ik}(t) = (V_{ik}(q) - S_i^r(q,t)) \;\text{ for }\; k \leq E_i^r(q); \quad V_{ik}(t) = (V_{ik} - S_i^r(U_{ik},t)) \;\text{ for }\; E_i^r(q) < k \leq E_i^r(t).$$

Now, for any $f \in \mathcal{C}_c$

$$\begin{aligned}
\langle f, \mathcal{Z}_i^r(t)\rangle &= \sum_{k=1}^{E_i^r(t)} \langle f, \boldsymbol{\delta}^+_{V_{ik}(t)}\rangle = \sum_{k=1}^{E_i^r(q)} \langle f, \boldsymbol{\delta}^+_{V_{ik}(q)-S_i^r(q,t)}\rangle + \sum_{k=E_i^r(q)+1}^{E_i^r(t)} \langle f, \boldsymbol{\delta}^+_{V_{ik}-S_i^r(U_{ik},t)}\rangle \\
&= \sum_{k=1}^{E_i^r(q)} \langle f(\cdot - S_i^r(q,t)), \boldsymbol{\delta}^+_{V_{ik}(q)}\rangle + \sum_{k=E_i^r(q)+1}^{E_i^r(t)} f(V_{ik} - S_i^r(U_{ik},t)) \\
&= \langle f(\cdot - S_i^r(q,t)), \mathcal{Z}_i^r(q)\rangle + \sum_{k=E_i^r(q)+1}^{E_i^r(t)} f(V_{ik} - S_i^r(U_{ik},t)). \qquad (48)
\end{aligned}$$

Now applying scaling by dividing both sides of (48) by r and (14), we have

$$\langle f, \bar{\mathcal{Z}}_i^r(t)\rangle = \langle f(\cdot - \bar{S}_i^r(q,t)), \bar{\mathcal{Z}}_i^r(q)\rangle + \frac{1}{r}\sum_{k=E_i^r(q)+1}^{E_i^r(t)} f(V_{ik} - \bar{S}_i^r(U_{ik},t)). \qquad (49)$$

Given (49) that holds for any $q < t$, to establish the result of the Lemma, we will need to use it for q, t such that $|q - t| \to 0$ and apply the properties listed below.

1. On a time a interval $[s,t]$, $\bar{Z}_i^r(\cdot)$ is right continuous with left limits. Since its infimum is strictly positive, there exists $\varepsilon, M > 0$ such that

$$0 < \varepsilon \leq \inf_{u\in[s,t]} \bar{Z}_i^r(u) \leq \sup_{u\in[s,t]} \bar{Z}_i^r(u) \leq M. \qquad (50)$$

   Further, we can assume that M is such that $\sup_{u\in[s,t]} \|\bar{Z}^r(u)\| \leq M$.
2. From Lemma 2, $\mathbf{x}_i(z) \leq \|C\|/\varepsilon$ on compact set $\{z : \|z\| \leq M,\ z_i \geq \varepsilon\}$. Further, by Lemma 1 $\mathbf{x}_i(\cdot)$ is continuous on this set.
3. For $f \in \mathcal{C}_c$, there exists non-decreasing continuous function $\psi_f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\psi_f(0) = 0$ and

$$\sup_{x\in\mathbb{R}} |f'(x+h) - f'(x)| \leq \psi_f(|h|).$$

Now, rest of the proof is similar to establishing the validity of 'Riemann Integration' using appropriate filtration for bounded continuous functions. Here is a quick sketch. Consider partition of interval $[s,t]$ into n sub-intervals of equal-size $(t-s)/n$ and apply (49). Then using properties 1-3, standard convergence theorems for integration, Taylor's expansion, and definitions, the desired conclusion of Lemma follows. ■

C. Tightness

Now we prove that sequence of measures $\mathbb{P}_r^T, r \geq 1$, is tight. That is, for any $\varepsilon > 0$ there exists a compact set $\mathbb{K}_\varepsilon \subset \mathbf{D}([0,T], \mathbf{M}^{\mathbf{I}})$ such that $\inf_{r\geq 1} \mathbb{P}_r^T(\mathbb{K}_\varepsilon) \geq 1 - \varepsilon$. A known characterization of compact sets in metric space $\mathbf{D}([0,T], \mathbf{M}^{\mathbf{I}})$ suggests the following are sufficient conditions to establish tightness: for any $\varepsilon > 0$,

(T1). Compact containment: there exists compact set $\mathbf{K}_\varepsilon \subset \mathbf{M}^{\mathbf{I}}$ such that, for all $r \geq 1$,

$$\liminf_r \mathbb{P}_r^T(\bar{\mathcal{Z}}^r(t) \in \mathbf{K}_\varepsilon, \quad \forall\ t \in [0,T]) \geq 1 - \varepsilon,$$

(T2). Bounded oscillation: for any $\delta > 0, \eta > 0$,

$$\limsup_r \mathbb{P}_r^T(\{\bar{\mathcal{Z}}^r : \mathbf{w}'_T(\bar{\mathcal{Z}}^r, \delta) \geq \varepsilon\}) < \eta.$$

In what follows, we present the above two sufficient properties in order to establish tightness of the sequences. Then, we will present proof of tightness that will complete the loop.

We want to note the following. The sufficient conditions (T1)-(T2) stated above suggest that it is okay to establish the existence of $\mathbb{K}_\varepsilon$ such that $\liminf_r \mathbb{P}_r^T(\mathbb{K}_\varepsilon) \geq 1 - \varepsilon$. It is well known that each probability measure is tight by itself (i.e., most of its mass is inside a compact set) since the underlying metric space is complete and separable. This fact, together with the (T1)-(T2), implies the sufficiency for tightness.

D. Compact Containment

For any finite $G > 0$, define

$$\mathbf{K}(G) = \{\zeta \in \mathbf{M}^{\mathbf{I}} : \|\langle \mathbf{1}, \zeta \rangle\| \vee \|\langle \chi, \zeta \rangle\| \leq G\}.$$

Then closure of $\mathbf{K}(G)$ is compact (e.g., Theorem 15.7.5 [14] as stated in [13]). The following Lemma (similar to Lemma 5.3 [13]) establishes the required compact containment property.

Lemma 13: Let $T > 0$ be given. Consider any $\varepsilon > 0$. Then, there exists a compact set $\mathbf{K} \subset \mathbf{M}^{\mathbf{I}}$ such that

$$\liminf_{r} \mathbb{P}_r^T(\bar{\mathcal{Z}}^r(t) \in \mathbf{K}, \ \forall \ t \in [0, T]) \geq 1 - \varepsilon,$$

where $\mathbf{K}$ is closure of $\mathbf{K}(2T(\|\nu\| + \|\rho\|) + 1)$.
The proof follows from the following facts: (a) $\bar{\mathcal{Z}}^r(0) = \mathbf{0}$ and Lemma 11, (b) $\langle \mathbf{1}, \bar{\mathcal{Z}}_i^r(t) \rangle \leq \langle \mathbf{1}, \bar{\mathcal{L}}^r(T) \rangle, \ \forall \ t \in [0, T]$ and (c) $\langle \chi, \bar{\mathcal{Z}}_i^r(t) \rangle \leq \langle \chi, \bar{\mathcal{L}}^r(T) \rangle, \ \forall \ t \in [0, T]$.

E. Regularity near 0

The following property, which is the most crucial in establishing the fluid model justification, is called asymptotic regularity near 0 of $\bar{\mathcal{Z}}^r(\cdot)$ [13].

Lemma 14: Let $T > 0$ be given. Consider any $\varepsilon, \eta > 0$. Then there exists $a > 0$ such that

$$\liminf_{r} \mathbb{P}_r^T \left(\sup_{t \in [0,T]} \max_{i \leq \mathbf{I}} \langle \mathbf{1}_{[0,a]}, \bar{\mathcal{Z}}_i^r(t) \rangle \leq \varepsilon \right) \geq 1 - \eta.$$

We will skip the details of the proof and again refer reader to consult [13]. However, we present some key observations of [13] used in proving the Lemma. In what follows, fix an $i \leq \mathbf{I}$. We want to show that for appropriate selection of $a > 0$, for any $t \in [0, T]$, $\langle \mathbf{1}_{[0,a]}, \bar{\mathcal{Z}}_i^r(t) \rangle \leq \varepsilon$ with probability at least $1 - \eta$. For this, consider large enough r so that the estimates of Lemmas 11 and 13 hold simultaneously with probability at least $1 - \eta/2$. We restrict our attention to this high probability event.

Call $\bar{\mathcal{Z}}_i^r(\cdot)$ *ε-regular* at t, if $\bar{Z}_i^r(t) \leq \varepsilon/8$. First note that, since $\bar{\mathcal{Z}}^r(0) = \mathbf{0}$, trivially $\bar{\mathcal{Z}}_i^r(\cdot)$ is $\varepsilon/8$-regular at $t = 0$. For $t > 0$, either $\bar{\mathcal{Z}}^r(\cdot)$ is $\varepsilon/8$-regular or not. If yes, then trivially we have $\langle \mathbf{1}_{[0,a]}, \bar{\mathcal{Z}}_i^r(t) \rangle \leq \langle \mathbf{1}, \bar{\mathcal{Z}}_i^r(t) \rangle < \varepsilon$. If not, then consider θ, the supermum of $0 \leq s < t$ such that $\bar{\mathcal{Z}}^r(\cdot)$ was $\varepsilon/8$-regular at s. Now in interval $[\theta, t]$, $\inf_{u \in [\theta,t]} \bar{Z}_i^r(u) \geq \varepsilon/8$. From Lemma 2, $\mathbf{x}_i(u) \leq 8\|C\|/\varepsilon$ for $u \in [\theta, t]$. Since the estimates of Lemma 11 hold, $\sup_{u \in [\theta,t]} \mathbf{x}_i(u) \leq 8\|C\|/\varepsilon$ and $|t - \theta| \leq t \leq T$, the increment $\langle \mathbf{1}_{[0,a]}, \bar{\mathcal{Z}}_i^r(t) - \bar{\mathcal{Z}}_i^r(\theta) \rangle$ is bounded above by γa, where γ is a finite number dependent on $T, \nu, \vartheta, \mathbf{I}, \varepsilon$. So appropriate choice of a can make the increment smaller than $\varepsilon/2$ and establish the desired result.

F. Bounded Oscillations

For $T > 0$, $\delta \in [0, T]$ and $\zeta \in \mathbf{D}([0, \infty), \mathbf{M}^{\mathbf{I}})$, define

$$\mathbf{w}_T(\zeta(\cdot), \delta) = \sup_{s,t \in [0,T] : |s-t| < \delta} \mathbf{d}_{\mathbf{I}}[\zeta(s), \zeta(t)].$$

Lemma 15: Let $T > 0$ be given. Then for any $\varepsilon, \eta \in (0, 1)$, there exists $\delta > 0$ such that

$$\liminf_{r} \mathbb{P}_r^T \left(\mathbf{w}_T(\bar{\mathcal{Z}}^r(\cdot), \delta) \leq \varepsilon \right) \geq 1 - \eta.$$

The proof of Lemma 15 is identical to the proof of Lemma 5.6 [13]. We skip it but present some basic ingredients. To prove the Lemma, it is sufficient to show that there exists $\delta > 0$ such that for r large enough with probability at least $1 - \eta$ the following holds: for any $0 \leq s \leq t \leq T, |s - t| < \delta$, and any closed set $B \subset \mathbb{R}_+$,

$$\begin{aligned} \langle \mathbf{1}_B, \bar{\mathcal{Z}}_i^r(s) \rangle &\leq \langle \mathbf{1}_{B^\varepsilon}, \bar{\mathcal{Z}}_i^r(t) \rangle + \varepsilon, && (51) \\ \langle \mathbf{1}_B, \bar{\mathcal{Z}}_i^r(t) \rangle &\leq \langle \mathbf{1}_{B^\varepsilon}, \bar{\mathcal{Z}}_i^r(s) \rangle + \varepsilon. && (52) \end{aligned}$$

The inequality (52) holds with high enough probability for large $r, \delta \leq \varepsilon$, due to Lemma 11 and the fact that

$$\langle \mathbf{1}_B, \bar{\mathcal{Z}}_i^r(t) \rangle \leq \langle \mathbf{1}_{B^\varepsilon}, \bar{\mathcal{Z}}_i^r(s) \rangle + \langle \mathbf{1}, \bar{\mathcal{L}}_i^r(t) - \bar{\mathcal{L}}_i^r(s) \rangle.$$

Proof of (51) is a bit more tricky. In a nutshell, it follows using argument similar to that used to establish Lemma 14.

G. Tightness: Closing the Loop

Now we are ready to establish the tightness of the sequence of probability measures of interest, which will settle the first part of the statement of Theorem 5.

Lemma 16 (Theorem 5.7 [13]): Let $T > 0$ be given. Then under probability measure $\mathbb{P}^T_r, r \in \mathbb{N}$, the sequence $\{(\bar{\mathcal{Z}}^r, \bar{Z}^r, \bar{W}^r, \bar{T}^r, \bar{U}^r)\}$ is tight[9].

Proof: First, tightness of $\bar{\mathcal{Z}}^r(\cdot)$ on $[0, T]$. From definitions, it can be shown (e.g. see [1]) that for any $0 \leq \delta \leq T$,

$$\mathbf{w}'_T(\zeta, \delta) \quad \leq \quad \mathbf{w}_{T+\delta}(\zeta, \delta). \tag{53}$$

Now (53), along with Lemmas 13 and 15, satisfies properties (T1) and (T2) which are sufficient to establish the tightness of $\bar{\mathcal{Z}}^r(\cdot)$ on $[0, T]$. By continuity of mapping $\zeta \to \langle \mathbf{1}, \zeta \rangle$, we obtain that $\bar{Z}^r(\cdot)$ is tight as well on $[0, T]$.

Now, by (3) we have $\|\Lambda_i(z)\| \leq \|C\|$ and $\Lambda_i(z) \geq 0$ for all $i \leq \mathbf{I}$ and $z \in \mathbb{R}^{\mathbf{I}}_+$. Hence, from (15) we have that $\bar{T}^r_i(\cdot)$ is non-decreasing Lipschitz continuous with Lipschitz constant $\|C\|$ for all r. By an application of Arzela-Ascolli's Theorem it follows that the sequence $\bar{T}^r_i(\cdot)$ is tight for any compact time interval $[0, T]$. The tightness of $\bar{T}^r(\cdot)$ implies the tightness of $\bar{U}^r(\cdot)$ on interval $[0, T]$ as well. Finally, tightness of $\bar{W}^r(\cdot)$ will follow from: $\bar{\mathcal{Z}}^r(0) = \mathbf{0}$, i.e., $\bar{W}^r_i(0) = 0$ for all $i \leq \mathbf{I}$, Lemma 11, and the following relation:

$$\bar{W}^r_i(t) = \langle \chi, \bar{\mathcal{L}}^r_i(t) \rangle - \bar{T}^r_i(t).$$

■

H. Characterization of Limit Points

The tightness established in Lemma 16 implies that, for every subsequence $\mathbb{P}^T_{r_q}$, there is a further subsequence $\mathbb{P}^T_{r_{q_m}}$ whose limit point, say $\mathbb{P}^T_\star$, is a probability distribution on $\mathbf{D}([0, T], \mathbf{M}^{\mathbf{I}})$. Such distributional limits are also called weak limit points of sequence $\mathbb{P}^T_r$. Now we wish to establish the second part of Theorem 5 about the support of such limit point being subsumed by the set of fluid model solutions.

In what follows, we are interested in any of the weak limit points. For that matter, fix a converging subsequence $\{r_q\} \subset \mathbb{N}$ so that $\mathbb{P}^T_{r_q} \overset{\mathbf{w}}{\to} \mathbb{P}^T_\star$. For ease of notation, let $(\bar{\mathcal{Z}}^q, \bar{Z}^q, \bar{W}^q, \bar{T}^q, \bar{U}^q, \bar{\mathcal{L}}^q)$ correspond to r_q-system with probability distributions induced by $\mathbb{P}^T_{r_q}$ and $(\mathcal{Z}, z, w, \tau, u, \mathcal{L})$ correspond to the limiting system with probability distribution induced by $\mathbb{P}^T_\star$. Next, we state Lemmas that will together lead to the completion of the proof of Theorem 5. The first one (and its proof) is identical to Lemma 5.8 [13].

Lemma 17: Under $\mathbb{P}^T_\star$, the following properties are statisfied with probability 1: for $t \geq 0$,

(i) $\|\langle \mathbf{1}_{\{0\}}, \mathcal{Z}(t) \rangle\| = 0$,
(ii) $z(t) = \langle \mathbf{1}, \mathcal{Z}(t) \rangle$,
(iii) $u(t) = Ct - A\tau(t)$,
(iv) $w(t) = \rho t - \tau(t)$,
(v) $w(t) = \langle \chi, \mathcal{Z}(t) \rangle$,
(vi) w is uniformly Lipschitz continuous with Lipschitz constant $\|\rho\| + \|C\|$,
(vii) for $i \leq \mathbf{I}$,

$$\tau_i(t) = \int_0^t \left(\Lambda_i(z(s)) \mathbf{1}_{\{z_i(s) > 0\}} + \rho_i \mathbf{1}_{\{z_i(s) = 0\}} \right),$$

(viii) u_j is non-decreasing for all $j \leq \mathbf{J}$.

Proof: We are given $T > 0$ and sequence $\{r_q\} \subset \mathbb{N}$ so that $\mathbb{P}^T_{r_q}$ converges to $\mathbb{P}^T_\star$ of our interest. We sketch main arguments proving properties (i)-(viii) as follows.

Proof of (i). It is sufficient to prove (i) for $t \in [0, T)$ given the $\langle \mathbf{1}_{\{0\}}, \vartheta \rangle = 0$ and stochastic assumption on the primitive processes. Using Lemma 14, it follows that there exists sequence $\{a_n : n \in \mathbb{N}\}$ so that $a_n > 0$ and

$$\liminf_{n \to \infty} \mathbb{P}^T_{r_q} \left(\sup_{t \in [0,T]} \|\langle \mathbf{1}_{[0, a_n)}, \bar{\mathcal{Z}}^q \rangle\| \leq \frac{1}{n} \right) \geq 1 - \frac{1}{n^2}.$$

[9] Also called $\mathbf{C}$-tight since its about tightness on compact time-intervals.

Define sets

$$\mathbf{A}_n = \left\{ \zeta \in \mathbf{M}^{\mathbf{I}} : \|\langle \mathbf{1}_{[0,a_n)}, \zeta \rangle\| \le \frac{1}{n} \right\}, \quad \mathbf{B}_n = \left\{ \zeta(\cdot) \in \mathbf{D}([0,\infty), \mathbf{M}^{\mathbf{I}}) : \zeta(t) \in \mathbf{A}_n \text{ for all } t \in [0,T] \right\}.$$

It can be checked that both $\mathbf{A}_n$ and $\mathbf{B}_n$ are closed in their respectively topologies. Since $\mathbb{P}^T_{r_q} \Rightarrow \mathbb{P}^T_\star$, we have that (Portmantau's characterization)

$$\mathbb{P}^T_\star(\mathcal{Z} \in \mathbf{B}_n) \ge \limsup_{q\to\infty} \mathbb{P}^T_{r_q}(\bar{\mathcal{Z}}^q \in \mathbf{B}_n) \ge \liminf_{q\to\infty} \mathbb{P}^T_{r_q}(\bar{\mathcal{Z}}^q \in \mathbf{B}_n) \ge 1 - \frac{1}{n^2}.$$

By standard application of Borel-Cantelli's Lemma and the fact that $\{0\} \subset [0,a_n)$ for all n, we have the desired conclusion that

$$\mathbb{P}^T_\star\left(\|\langle \mathbf{1}_{\{0\}}, \mathcal{Z}(t)\rangle\| = 0\right) = 1.$$

Proof of (ii). Since $\mathbf{1}$ is a bounded continuous function and under $\mathbb{P}^T_{r_q} \Rightarrow \mathbb{P}^T_\star$, we have $\bar{\mathcal{Z}}^q \Rightarrow \mathcal{Z}$. By definition of weak convergence on $\mathbf{D}([0,T], \mathbf{M}^{\mathbf{I}})$, we have $\langle \mathbf{1}, \bar{\mathcal{Z}}^q(\cdot)\rangle \to \langle \mathbf{1}, \mathcal{Z}(\cdot)\rangle$. But $\bar{Z}^q(\cdot) = \langle \mathbf{1}, \bar{\mathcal{Z}}^q(\cdot)\rangle$ and $\bar{Z}^q(\cdot) \Rightarrow z(\cdot)$. This proves (ii).

Proof of (iii)-(iv). The (iii)-(iv) follow from: (a) under $\mathbb{P}^T_{r_q} \Rightarrow \mathbb{P}^T_\star$, we have $(\bar{W}^q, \bar{T}^q, \bar{U}^q) \Rightarrow (w, \tau, u)$, (b) dynamic relations (16) and (17), and (c) Lemma 11.

Proof of (v). By definition, $\bar{W}^q(t) = \langle \chi, \bar{\mathcal{Z}}^q(t)\rangle$ and $w(t) = \langle \chi, \mathcal{Z}(t)\rangle$. Under Skorohod's topology, $\mathbb{P}^T_{r_q} \Rightarrow \mathbb{P}^T_\star$ implies $\bar{\mathcal{Z}}^q(t) \Rightarrow \mathcal{Z}(t)$ for all $t \in [0,T]$, it suffices to show that $\{\bar{\mathcal{Z}}^q(t)\}$ are uniformly integrable for all t so as to establish convergence $\langle \chi, \bar{\mathcal{Z}}^q(t)\rangle \to \langle \chi, \mathcal{Z}(t)\rangle$. But we know that for $x > 0$,

$$\langle \chi \mathbf{1}_{[x,\infty)}, \bar{\mathcal{Z}}^q(t)\rangle \le \langle \chi \mathbf{1}_{[x,\infty)}, \bar{\mathcal{L}}^q(t)\rangle.$$

Now use of Lemma 11 will imply the desired uniform integrability and hence the property (v).

Proof of (vi). Follows from (iv) and Lipschitz continuity of $\bar{T}^q(\cdot)$ (and hence of $\tau(\cdot)$).

Proof of (vii). This requires use of a standard, simple, but very insightful trick, which is described in this paragraph. To this end, recall that w, τ are Lipschitz continuous and hence differentiable almost everywhere in $[0,T]$. Let t be a point where both w, τ are differentiable (i.e. t is regular point for both w, τ). Then by (iv), we have $dw_i(t)/dt = \rho_i - d\tau_i(t)/dt$. If $z_i(t) = 0$ then $w_i(t) = 0$. We claim that due to non-negativity of $w_i(\cdot)$, if $dw_i(t)/dt$ exists then it is equal to 0 at $w_i(t) = 0$. This is justified next. Suppose $dw_i(t)/dt \neq 0$. Let $dw_i(t)/dt > 0$. Then

$$\lim_{h\to 0^+} \frac{w_i(t) - w_i(t-h)}{h} > 0.$$

But, $w_i(t) = 0$ and hence $w_i(t-h) < 0$ for some $h > 0$. This is not possible since $w_i(\cdot)$ is non-negative. That is, $dw_i(t)/dt > 0$ is not possible at $w_i(t) = 0$. Similarly, $dw_i(t)/dt < 0$ is not possible as well. Hence, we have showed that $dw_i(t)/dt = 0$ when $w_i(t) = 0$. This implies that $d\tau_i(t)/dt = \rho_i$ when $z_i(t) = 0$. If $z_i(t) > 0$, then $d\tau_i(t)/dt = \Lambda_i(z(t))$ follows from: (a) continuity of $z(\cdot)$ given the continuity of $\mathcal{Z}(\cdot)$, (b) continuity of $\Lambda_i(z)$ on $\{z : z_i > 0\}$ as stated in Assumption 1, (c) boundedness of $\Lambda_i(z) \le \|C\|$, and (d) dynamic equation (15) along with bounded convergence theorem.

Proof of (viii). u_j is non-decreasing because of the dynamic equation (16) and the fact that $\bar{U}^q \Rightarrow u$ uniformly on compact intervals. ■

Lemma 18 (Lemma 5.10, [13]): Fix $f \in \mathcal{C}_c$ and $i \le \mathbf{I}$. for all intervals $[s,t] \subset \mathbb{R}_+$ such that $\inf_{u\in[s,t]} z_i(u) > 0$,

$$\langle f, \mathcal{Z}_i(t)\rangle = \langle f, \mathcal{Z}_i(t)\rangle - \int_s^t \langle f', \mathcal{Z}_i(u)\rangle \mathbf{x}_i(z(u)) + \nu_i(t-s)\langle f, \vartheta_i\rangle.$$

The proof of this is identical to that of Lemma 5.10 [13] using Lemma 12.

Lemma 19 (Theorem 5.9, [13]): Almost surely, for all $i \le \mathbf{I}$, $f \in \mathcal{C}$ and $t \ge 0$,

$$\langle f, \mathcal{Z}_i(t)\rangle = -\int_0^t \langle f', \mathcal{Z}_i(u)\rangle \mathbf{x}_i(z(u)) + \nu_i \langle f, \vartheta_i\rangle \int_0^t \mathbf{1}_{\{z_i(s)>0\}} ds.$$

Proof: As in [13], the proof for $f \in \mathcal{C}$ can be obtained by first proving it for $f \in \mathcal{C}_c$ and then using standard truncation-style argument along with convergence theorems to obtain the result for all $f \in \mathcal{C}$.

Now, let $f \in \mathcal{C}_c$. A key property of $f \in \mathcal{C}_c$ that is very useful is that there exists constant κ_f so that $f(x) \le \kappa_f x$ for all x (and f is Lipschitz continuous with constant κ_f).

Rest of the proof first establishes that, for such $f \in \mathcal{C}_c$, $\langle f, \mathcal{Z}_i(\cdot)\rangle$ is Lipschitz continuous. Then, establish that $|\langle f, \mathcal{Z}_i(t) - \mathcal{Z}_i(s)\rangle| \leq \boldsymbol{\kappa}_f |w_i(t) - w_i(s)|$ using property of $f \in \mathcal{C}_c$ for some constant $\boldsymbol{\kappa}_f$. Finally, using argument similar to that used for establishing property (vii) in Lemma 17, one will be able to complete the proof.

First, we establish Lipschitz continuity. For this, the proof requires handling following two cases separately: (a) $\{s : z_i(s) > 0\}$ and (b) $\{s : z_i(s) = 0\}$. Consider an interval say $[s, t]$. Let $\gamma_0 = \inf\{u \in [s, t] : z_i(u) = 0\}$, with the definition that $\gamma_0 = t$ if $\inf_{u \in [s,t]} z_i(u) > 0$. Given this, for any $[s, \gamma_1] \subset [s, \gamma_0)$, we have $\inf_{u \in [s, \gamma_1]} z_i(u) > 0$. For any such γ_1, Lemma 18 implies that

$$|\langle f, \mathcal{Z}_i(\gamma_1)\rangle - \langle f, \mathcal{Z}_i(s)\rangle| \;\leq\; (\gamma_1 - s)\left(\|f'\|\|C\| + \|f\|\nu_i\right).$$

By continuity of the terms above and letting $\gamma_1 \uparrow \gamma_0$, we obtain

$$|\langle f, \mathcal{Z}_i(\gamma_0)\rangle - \langle f, \mathcal{Z}_i(s)\rangle| \;\;\leq\;\; (\gamma_0 - s)\left(\|f'\|\|C\| + \|f\|\nu_i\right). \tag{54}$$

Now, if $\gamma_0 < t$, then by Lipschitz continuity property of $f \in \mathcal{C}_c$, Lemma 17(vi), and the fact that $w(\gamma_0) = 0$, we have

$$\begin{aligned} |\langle f, \mathcal{Z}_i(t)\rangle - \langle f, \mathcal{Z}_i(\gamma_0)| \;&\leq\; \kappa_f |\langle \chi, \mathcal{Z}_i(t)\rangle| \;=\; \kappa_f w_i(t) \\ &=\; \kappa_f |w_i(t) - w_i(\gamma_0)| \;\leq\; \kappa_f(\|\rho\| + \|C\|)(t - \gamma_0). \end{aligned} \tag{55}$$

Note that (54) for $\gamma_0 = s$ and (55) for $\gamma_0 = t$ work as well. Given (54)-(55), we have

$$|\langle f, \mathcal{Z}_i(t)\rangle - \langle f, \mathcal{Z}_i(s)| \;\;\leq\;\; (t - s)\boldsymbol{\kappa}_f, \tag{56}$$

where $\boldsymbol{\kappa}_f = \kappa_f(\|\rho\| + \|C\|) + \|f'\|\|C\| + \|f\|\|\nu\|$.

The Lipschitz continuity of $\langle f, \mathcal{Z}_i(\cdot)\rangle$ implies that it is differentiable for almost everywhere in $[0, T]$. Let t be a regular point for $\langle f, \mathcal{Z}_i(\cdot)\rangle$ as well as $w(\cdot)$. To evaluate the differential of $\langle f, \mathcal{Z}_i(t)\rangle$ consider two cases separately: (a) If $z_i(t) > 0$, then use Lemma 18 to evaluate the differential of $\langle f, \mathcal{Z}_i(\cdot)\rangle$. (b) If $z_i(t) = 0$, use the fact that $w_i(t) = 0$ implies (using argument similar to that in Lemma 17(vi)) $dw_i(t)/dt = 0$. But the uniform bound on $\langle f, \mathcal{Z}_i(t)\rangle$ in terms of $w_i(t)$ implies that $d\langle f, \mathcal{Z}_i(t)\rangle/dt = 0$ as well. Putting the above together, the proof of the claimed Lemma follows. ■

I. Proof of Theorem 5

Here we wrap up the proof of Theorem 5 by consolidating the above justifications. Note that the Theorem 5 has two main claims: (a) Tightness of measures $\mathbb{P}_r^T$ and (b) Characterization of weak limit points as satisfying fluid model solution with probability 1. Lemma 16 establishes (a). Lemmas 17 and 19 together establish the (b).

Finally, we remind reader once again that the above justification of fluid model (proof of Theorem 5) is identical to that of proof of Theorem 4.1[13], with the only difference in the scaling (over capacity and arrival rates in our scaling, and over time and space in [13]) to facilitate heterogeneous utility functions. More generally, proof of fluid model of [13] (along with our scaling) relies primarily on and Assumption 1 and uniqueness of optimal solution in the NUM based network resource allocation. Hence, the techniques of [13] are likely to extend for many other problems.