Impact of Fairness and Heterogeneity on Delays in Large-scale Content Delivery Networks

Virag Shah
The University of Texas at Austin
Texas, USA
virag@utexas.edu

Gustavo de Veciana
The University of Texas at Austin
Texas, USA
gustavo@ece.utexas.edu

ABSTRACT

We consider multi-class queueing systems where the per class service rates depend on the network state, fairness criterion, and is constrained to be in a symmetric polymatroid capacity region. We develop new comparison results leading to explicit bounds on the mean service time under various fairness criteria and possibly heterogeneous loads. We then study large-scale systems with growing numbers of service classes n (e.g., files), heterogenous servers m and polymatroid capacity resulting from a random bipartite graph modeling service availability (e.g., placement of files across servers). This models, for example, a large scale content delivery network (CDN) supporting parallel servicing of a download request. For an appropriate asymptotic regime, we show that the system's capacity region is uniformly close to a symmetric polymatroid – i.e., heterogeneity in servers' capacity and file placement disappears.

Combining our comparison results and the asymptotic 'symmetry' in large systems, we study performance robustness to heterogeneity in per class loads and fairness criteria. Roughly, if each class can be served by $c_n = \omega(\log n)$ servers, the load per class does not exceed $\theta_n = o\left(\min(\frac{n}{\log n}, c_n)\right)$, and average server utilization is bounded by $\gamma < 1$, then mean delay satisfies the following bound:

$$E[D^{(n)}] \leq K \frac{\theta_n}{c_n} \frac{1}{\gamma} \log\left(\frac{1}{1-\gamma}\right),$$

where K is a constant. Thus, large, randomly configured CDNs with a logarithmic number of file copies are robust to substantial load and server heterogeneities for a class of fairness criteria.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of Systems—*Modeling techniques*; G.2 [**Probability and Statistics**]: Queueing theory; H.2.4 [**Information Systems**]: Systems—*Concurrency*

SIGMETRICS'15, June 15–19, 2015, Portland, OR, USA.
Copyright © 2015 ACM 978-1-4503-3486-0/15/06 ...$15.00.
http://dx.doi.org/10.1145/2745844.2745857.

General Terms

Performance, Theory

Keywords

Delays; queuing system; content delivery networks; content placement; parallel downloads; bipartite graphs; robustness; fairness; asymptotics

1. INTRODUCTION

In many shared network systems service rate is allocated to ongoing jobs based on a fairness criterion, e.g., α-fair (αF) (including max-min and proportional fair) as well as Balanced fair (BF), and other Greedy criteria [24]. When the network loads are stochastic a key open question is how the choice of fairness and network design will impact user perceived performance, e.g., job delays, as well as the sensitivity of performance to heterogeneity in network resources and traffic loads. Motivated by this challenge in this paper we take a step towards understanding these issues by investigating performance bounds for an interesting class of stochastic networks with symmetric polymatroid capacity under various fairness criteria.

The second question driving this paper is whether large scale systems can be designed to be inherently robust to heterogeneity and at what cost? Specifically we consider a centralized content delivery infrastructure where a collection of servers store and deliver large files, e.g., scientific datasets/visualization, 3D videos, software updates, other immersive technologies, and are aimed at doing so with small delays. Such centralized infrastructure could, for example, be part of a larger distributed content delivery network (CDN), where requests not currently available at distributed sites are forwarded to the centralized infrastructure which in turn delivers the files to the remote sites and/or users. There has been substantial recent interest in understanding basic design questions for these systems including, see e.g. [10, 14, 20, 23] and references therein: How should the number of file copies scale with the demand? What kinds of hierarchical caching policies are most suitable? How to best optimize storage/backhaul costs for unpredictable time-varying demands? Our focus is on CDNs that permit parallel file downloads from multiple servers – akin to peer-to-peer systems. In principle, with an appropriate degree of storage redundancy, one can achieve much better peak service rates, exploit diversity in service paths, produce robustness to failures, and provide better sharing of pooled server resources.

Intuitively when such systems have sufficient redundancy they will exhibit performance which is robust to limited heterogeneity in demands and server capacity, as well as to the fairness criterion driving resource allocation. Such systems might also circumvent the need for, and overheads (such as backhaul, state update, etc) associated with, dynamic caching. If this is the case, CDNs enabling parallel servicing of individual download requests could be more scalable and robust to serving popular content.

Our Contributions and Organization: The contributions of this paper are threefold, each of independent interest, and collectively, providing a significant step forward over what is known in the current literature.

a.) *Performance bounds:* In Sections 3-4 we consider a class of systems with symmetric polymatroid capacity for which we develop several rate allocation monotonicity properties which translate to performance comparisons amongst fairness policies, and eventually give explicit bounds on mean delays. Specifically we show that under homogeneous loads the mean delay achieved by Greedy and αF rate allocations are bounded by that of BF allocation which is computable. We then extend this upper bound to the case when the load is heterogeneous but 'majorized by a symmetric load'.

b.) *Uniform symmetry in large systems:* In Section 5 we consider a bipartite graph where nodes represent n job classes (files) and m servers with potentially heterogenous service capacity. The graph edges capture the ability of servers to serve the jobs in the given classes. If jobs can be concurrently served by multiple servers then the system's service capacity region is a polymatroid. We show that for appropriately scaled large system where the edge set is chosen at random (random file placement) the capacity region is uniformly close to a *symmetric* polymatroid.

c.) *Performance robustness in large systems:* By combining these two results, in Section 6 we provide a simple performance bound for large-scale content delivery systems. The bound exhibits performance robustness in such systems with respect to variations in total system load, heterogeneity in load across the classes, heterogeneity in server capacities, for α-fair based resource allocation. Specifically it establishes a clear link between the degree of content replication and permissible demand heterogeneity while ensuring performance scalability.

We have endeavored to provide as complete results as possible, and have deferred details to the appendix. Section 7 concludes the paper.

Related work: There is a substantial amount of related work. Yet the link between fairness in resource allocation and job delays in stochastic networks is poorly understood. The only fairness criterion for which explicit expressions or bounds are known is the Balanced Fair rate allocation [3] which generalizes the notion of 'insensitivity' of the processor sharing discipline in $M/G/1$ queuing system. Under balanced fairness, an explicit expression for mean delay was obtained in [5, 6] for a class of wireline networks, namely, those with line and tree topologies. Also, a performance bound for arbitrary polytope capacity region and arbitrary load was provided in [1]. Similarly [11] developed bounds for stochastic networks where flows can be split over multiple paths. These bounds and expressions are either too specific or too loose. In [22] we developed an expression for the mean delay for systems with polymatroid capacity and arbitrary loads under Balanced Fair rate allocations. Unfortunately the result has exponential computational complexity in general. However the symmetric case has low complexity, a fact we use in the sequel.

Balanced fair rate allocation is defined recursively and is difficult to implement. α-fair rate allocations [13, 19] which are based on maximizing a concave sum utility function over the system's capacity region – this includes proportional and max-min fair allocations, are more amenable to implementation [12, 15]. However, the only known explicit performance results for stochastic networks under such fairness criteria are for systems where proportional fair is equivalent to balanced fair [3, 17]. In [2], performance relationship under balanced and proportional fairness for several systems where they are not equivalent was studied through numerical computations, and were found to be relatively close in several scenarios.

In this paper we focus on a class of stochastic networks that can be characterized by a polymatroid capacity region. Such systems have also been considered in [22, 24]. For example, the work in [24] shows that when such systems are symmetric with respect to load and capacity, a greedy rate allocation is delay optimal. However, the result is brittle to asymmetries. We provide more details on greedy and other rate allocations in Section 3.

In summary when it comes to fairness criteria and stochastic network performance there is a gap between what is implementable and what is analyzable. One of the goals of this paper is to provide comparison results which address this gap, with particular focus on addressing user-performance in large-scale CDN systems prevalent today. In this setting the two works closest to this paper are [23] and [22]. Both adopt a natural model for a CDN-like system based on a bipartite graph which captures the availability of files at servers to support the file-download requests. They show that if the graph is chosen at random and scaled appropriately then user performance is robust to load heterogeneity. The authors in [23] consider a service model where each request can be served by a single server without preemption – recall we consider systems allowing parallel downloads. The flexibility of our service model leads to a significantly improved mean delay bound and the resulting robustness. For example, upon availability of c_n servers for each class, the maximum per-class load allowed in [23] is $o\left(\sqrt{\frac{c_n}{\log n}}\right)$ which is significantly lower than our limit of $o\left(\min(\frac{n}{\log n}, c_n)\right)$. Also in our work we are able to address the role of fairness criteria and robustness to heterogeneity in server capacities.

While our service model is similar to that in [22], our work in this paper is different in several respects. Firstly, in [22] we focused on mean delays only for Balanced fair resource allocation whereas in this paper we directly study the impact of fairness criteria on users' delays. Secondly, we consider here a different scaling regime where the number of job classes (files) n grows proportionally to the number of servers m. In our earlier work the system was by design symmetric whereas in this paper we establish the asymptotic symmetry. Thirdly, in this paper we establish new results on robustness to limited heterogeneity in file demands, server capacity and α-fairness criteria by providing a uniform bound on delays.

2. SYSTEM MODEL

Our system consists of a set F of n classes. Jobs for class $i \in F$ arrive as an independent Poisson process of rate λ_i. Let $\boldsymbol{\lambda} = (\lambda_i : i \in F)$. Service requirements of jobs are i.i.d exponential with mean ν. Let $\boldsymbol{\rho} = (\rho_i : i \in F)$, where $\rho_i = \lambda_i \nu$ denotes the load associated with class i.

Let $q_i(t)$ denote the *set* of ongoing jobs of class i at time t, i.e., jobs which have arrived but have not completed service, and $\mathbf{q}(t) = (q_i(t) : i \in F)$. Let $\mathbf{x}(t) = (x_i(t) : i \in F)$, where $x_i(t) \triangleq |q_i(t)|$, i.e., $\mathbf{x}(t)$ captures the number of ongoing jobs in each class.

We refer to $\mathbf{x}(t)$ as the state of the system at time t. Let $\mathbf{X}(t)$ correspond to the random vector describing the state of the system at time t. We refer to the random process $(\mathbf{X}(t) : t \geq 0)$ as the state process. For any $\mathbf{x}(t)$, let $A_{\mathbf{x}(t)}$ denote the set of active classes, i.e., the classes with at least one ongoing job.

Service Model: For any $v \in q_i(t)$, let $b_v(t)$ be the rate at which job v is served at time t. The vector $\mathbf{b}(t) = (b_v(t) : v \in \cup_i q_i(t))$ represents the rates assigned to ongoing jobs at time t. Within each class we assume that each job is allocated equal rate, i.e., $b_v(t) = b_u(t)$ for each $u, v \in q_i(t)$. If job v arrives at time t_v^a and has service requirement η_v, then it departs at time t_v^d such that $\eta_v = \int_{t_v^a}^{t_v^d} b_v(t)dt$. Thus, $t_v^d - t_v^a$ is the delay for job v.

Further, let $r_i(\mathbf{x}')$ be the total rate at which class i jobs are served at time t when $\mathbf{x}(t) = \mathbf{x}'$, i.e., at any time t, $r_i(\mathbf{x}(t)) = \sum_{v \in q_i(t)} b_v(t)$. Let $\mathbf{r}(\mathbf{x}') = (r_i(\mathbf{x}') : i \in F)$. We call the vector function $\mathbf{r}(.)$ the *rate allocation*. Note that the rate allocation at any time t depends only on the $\mathbf{x}(t)$ and thus can not depend on the residual file sizes of ongoing jobs.

Polymatroid Capacity Region: We shall consider systems where rate allocation $\mathbf{r}(\mathbf{x})$ for each $\mathbf{x}$ are constrained to be within a polymatroid capacity region $\mathcal{C}$.

Definition 1. *We say that $\mathcal{C}$ is a* **polymatroid** *if it takes the following form:*

$$\mathcal{C} = \left\{ \mathbf{r} \geq \mathbf{0} : \sum_{i \in A} r_i \leq \mu(A),\ \forall A \subset F \right\},$$

where $\mu(.)$ is a set function which satisfies the following properties:
1) Normalized: $\mu(\emptyset) = 0$.
2) Monotonic: if $A \subset B$, $\mu(A) \leq \mu(B)$.
3) Submodular: for all $A, B \subset F$,

$$\mu(A) + \mu(B) \geq \mu(A \cup B) + \mu(A \cap B).$$

The function $\mu(.)$ is called a **rank function***.*

Polymatroids and submodular functions are well studied in literature, see e.g., [9, 21].

Definition 2. *A polymatroid $\mathcal{C}$ is a* **symmetric polymatroid** *if its rank function $\mu(.)$ satisfies the following property: for each $A \subset F$, we have $\mu(A) = h(|A|)$, where $h : \mathbb{Z}_+ \to \mathbb{R}_+$ is a non-decreasing concave function.*

For a given $\mathbf{x}$, we say $\mathbf{r}(\mathbf{x})$ is feasible if $\mathbf{r}(\mathbf{x}) \in \mathcal{C}$; when this is true for all $\mathbf{x}$, we say that the rate allocation $\mathbf{r}(.)$ is feasible. We call $\mathcal{C}$ the capacity region of the system. Symmetric polymatroid capacity regions appear in several systems, for example, Gaussian symmetric multi-access channels [24]. Further, we will see in Section 5 that certain types of large content delivery networks have approximately symmetric polymatroid capacity regions.

Polymatroid capacity regions $\mathcal{C}$ have a special property that for any $\mathbf{r} \in \mathcal{C}$, there exists $\mathbf{r}' \geq \mathbf{r}$ such that $\mathbf{r}' \in \mathcal{D} \triangleq \{\mathbf{r} \in \mathcal{C} : \sum_{i \in F} r_i = \mu(F)\}$ [9, 21]. Also, as evident from the definition, for any $A \subset F$ the set $\{\mathbf{r} \in \mathcal{C} : r_i = 0, \forall i \notin A\}$ is also a polymatroid, with a rank function which is the restriction of $\mu(.)$ to subsets of A.

Further, we let

$$\hat{\mathcal{C}} \triangleq \left\{ \boldsymbol{\rho}' \geq \mathbf{0} : \sum_{i \in A} \rho_i' < \mu(A),\ \forall A \subset F \right\}, \qquad (1)$$

and will see, $\hat{\mathcal{C}}$ is the set of loads which are stabilizable for appropriate rate allocation policies.

Notation for ordering and majorization: In the sequel, we will rely on notation for ordering and majorization which we introduce below.

Let I be a finite arbitrary index set. Consider an arbitrary vector $\mathbf{z} = (z_i : i \in I)$. We let $z_{[1]} \geq z_{[2]} \geq \ldots, z_{[|I|]}$ denote the components of $\mathbf{z}$ in decreasing order. We let $|\mathbf{z}|$ denote $\sum_{i \in I} |z_i|$. We let $\mathbf{e}_i$ denote a vector with 1 at the i^{th} coordinate and 0 elsewhere.

For vectors $\mathbf{z}$ and $\mathbf{z}'$ such that $z_i \leq z_i'$ for each $i \in I$, we write $\mathbf{z} \leq \mathbf{z}'$ and say that $\mathbf{z}$ is *dominated* by $\mathbf{z}'$.

Below we define *majorization* ($\prec$) which describes how 'balanced' a vector is as compared to another vector. In words, by $\mathbf{z} \prec \mathbf{z}'$ we mean that $\mathbf{z}$ is 'more balanced' than $\mathbf{z}'$ but they have the same sum. By $\mathbf{z} \prec_w \mathbf{z}'$ we mean that $\mathbf{z}$ is 'more balanced' and has lower sum than $\mathbf{z}'$. Similarly, by $\mathbf{z} \prec^w \mathbf{z}'$ we mean that $\mathbf{z}$ is 'more balanced' and has larger sum than $\mathbf{z}'$.

Definition 3. *For vectors $\mathbf{z}$ and $\mathbf{z}'$ such that $|\mathbf{z}| = |\mathbf{z}'|$ and $\sum_{l=1}^k z_{[l]} \leq \sum_{l=1}^k z'_{[l]}$ for each $k \in \{1, 2, \ldots, |I|\}$, we say $\mathbf{z}$ is* majorized *by $\mathbf{z}'$, and denote this as $\mathbf{z} \prec \mathbf{z}'$.*

If we have $\sum_{l=1}^k z_{[l]} \leq \sum_{l=1}^k z'_{[l]}$ for each $k \in \{1, 2, \ldots, |I|\}$, we say $\mathbf{z}$ is weak-majorized from below *by $\mathbf{z}'$, and denote this as $\mathbf{z} \prec_w \mathbf{z}'$.*

Similarly, if we have $\sum_{l=0}^k z_{[|I|-l]} \geq \sum_{l=1}^k z'_{[|I|-l]}$ for each $k \in \{0, 1, \ldots, |I|-1\}$, we say $\mathbf{z}$ is weak-majorized from above *by $\mathbf{z}'$, and denote this as $\mathbf{z} \prec^w \mathbf{z}'$.*

The dominance and majorization have an associated stochastic version, defined below.

Definition 4. *Consider random vectors $\mathbf{Z}$ and $\mathbf{Z}'$. If there exist random vectors $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{Z}}'$ such that $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are identically distributed, $\mathbf{Z}'$ and $\tilde{\mathbf{Z}}'$ are identically distributed, and $\tilde{\mathbf{Z}}' \leq \tilde{\mathbf{Z}}'$ almost surely, then we say that $\mathbf{Z}$ is* stochastically dominated *by $\mathbf{Z}'$, and denote this as $\tilde{\mathbf{Z}} \leq^{st} \tilde{\mathbf{Z}}'$.*

Instead, if $\tilde{\mathbf{Z}}' \prec_w \tilde{\mathbf{Z}}'$, then we say that $\mathbf{Z}$ stochastically weak-majorized from below *by $\mathbf{Z}'$, and denote this as $\tilde{\mathbf{Z}} \prec_w^{st} \tilde{\mathbf{Z}}'$.*

In the sequel, it will be useful to introduce following notation. Recall, $\mathbf{r}(\mathbf{x}) = (r_i(\mathbf{x}) : i \in F)$ is the vector of rates allocated to various classes. We define $r_{(k)}(.)$ for each $k \in \{1, \ldots, n\}$ as follows: For a given state $\mathbf{x}$, let i_k be the class corresponding to $x_{[k]}$. Then, $r_{(k)}(\mathbf{x}) = r_{i_k}(\mathbf{x})$. In words, $r_{(k)}(\mathbf{x})$ is the rate allocated to the class with the k^{th} largest number of ongoing jobs.

Notation for scaling: Consider sequences of numbers $(f_n : n \in \mathbb{N})$ and $(g_n : n \in \mathbb{N})$. We say that $f_n = O(g_n)$ if there exists a constant $k > 0$ and an integer n_0 such that for each $n \geq n_0$, we have $f_n \leq k g_n$. We say that $f_n = \Omega(g_n)$ if there exists a constant $k > 0$ and an integer n_0 such that for each $n \geq n_0$, we have $f_n \geq k g_n$.

We say that $f_n = o(g_n)$ if $\lim_{n\to\infty} \frac{f_n}{g_n} = 0$. Similarly, we say that $f_n = \omega(g_n)$ if $\lim_{n\to\infty} \frac{g_n}{f_n} = 0$.

Several notations above are borrowed from [16], [24] and [22].

3. RATE ALLOCATION POLICIES: A BACKGROUND

There are several possible rate allocation policies, each resulting in potentially different user-perceived delays. Below, we introduce three different policies studied in literature, each with its own merits.

1) Greedy rate allocation: Roughly, the Greedy rate allocation policy on a polymatroid capacity region $\mathcal{C}$ assigns the maximum possible rate to the largest queues subject to the capacity constraints. We denote the Greedy rate allocation by $\mathbf{r}^G(.)$ and define it as follows: for each state $\mathbf{x}$, we let

$$\begin{aligned} r^G_{(k)}(\mathbf{x}) = \mu\left(\{[1],[2],\ldots,[k]\}\right) - \mu\left(\{[1],[2],\ldots,[k-1]\right) \\ \text{if } k \in \{1,2,\ldots,|A_{\mathbf{x}}|\}, \\ = 0 \text{ otherwise.} \end{aligned}$$

Equivalently, the sum rate assigned to the k largest queues, namely $\sum_{l=1}^{k} r^G_{(l)}(\mathbf{x})$, is equal to $\mu\left(\{[1],[2],\ldots,[k]\}\right)$. Using a quadratic Lyapunov function, one can show that Greedy rate allocation results in a stationary state process if $\boldsymbol{\rho} \in \hat{\mathcal{C}}$, where $\hat{\mathcal{C}}$ is defined in (1). The Greedy rate allocation for symmetric polymatroid capacity regions was first studied in [24] where the following result was shown.

Proposition 1. ([24]) *Suppose the capacity region $\mathcal{C}$ is a symmetric polymatroid and the load $\boldsymbol{\rho} \in \hat{\mathcal{C}}$ is homogeneous, i.e., $\rho_i = \rho$ for each $i \in F$. Then the following statements hold:*

1. *Let $(\mathbf{X}^G(t) : t \geq 0)$ and $(\tilde{\mathbf{X}}(t) : t \geq 0)$ be state processes under Greedy and an arbitrary feasible rate allocation, respectively. If $\mathbf{X}^G(0) \prec^{st}_{w} \tilde{\mathbf{X}}(0)$ then $\mathbf{X}^G(t) \prec^{st}_{w} \tilde{\mathbf{X}}(t)$ for each $t \geq 0$.*
2. *The mean job delay under Greedy rate allocation is less than or equal to that under any feasible rate allocation.*

Unfortunately, this optimality result for symmetric systems does not provide any explicit performance characterization or bound. Further, the result is brittle to heterogeneity in load or capacity.

2) α-fair rate allocation: As introduced in [19], this policy allocates rates based on maximizing a concave sum utility function subject to the system's capacity region. Formally, for a given $\alpha > 0$, the α-fair (αF) rate allocation $\mathbf{r}^\alpha(.)$, can be defined as follows: for each state $\mathbf{x}$, let

$$\mathbf{r}^\alpha(\mathbf{x}) = \begin{cases} \arg\max_{\hat{\mathbf{r}} \in \mathcal{C}} \sum_{i\in F} \frac{x_i^\alpha\, \hat{r}_i^{1-\alpha}}{1-\alpha} & \text{for } \alpha \in (0,\infty)\backslash\{1\}, \\ \arg\max_{\hat{\mathbf{r}} \in \mathcal{C}} \sum_{i\in F} x_i \log(\hat{r}_i) & \text{for } \alpha = 1. \end{cases} \tag{2}$$

This generalizes various notions of fairness across jobs, e.g., proportional fair and max-min fair allocations are equivalent to the α-fair policy for $\alpha = 1$ and $\alpha \to \infty$, respectively [19]. However, for polymatroid capacity regions the following result has been established.

Proposition 2. ([22]) *All α-fair rate allocations are equivalent for polymatroid capacity regions.*

Further, the stability result in [7] implies that the αF rate allocation results in a stationary state process when $\boldsymbol{\rho} \in \hat{\mathcal{C}}$. The α-fair rate allocation is attractive in that it it is amenable to distributed implementation [12, 15] and satisfies natural axioms for fairness [13]. Unfortunately, little is known regarding their performance under stochastic loads. What has been shown is that for α-fair allocations, the performance is *sensitive* to the distribution of service requirements [3]. Thus, it will be hard to make general claims. This leads us to the Balanced fair rate allocation below.

3) Balanced fair rate allocation: As introduced in [3], the Balanced fair (BF) rate allocation is 'insensitive', i.e., performance depends on the job service distribution only through its mean. Further, as we will see, it is more amenable to performance analysis under stochastic loads. Formally, Balanced fair rate allocation $\mathbf{r}^B(.)$ for a polymatroid capacity region $\mathcal{C}$ can be defined as follows, see [3]: for each state $\mathbf{x}$, we have

$$r_i^B(\mathbf{x}) = \frac{\Phi(\mathbf{x} - \mathbf{e}_i)}{\Phi(\mathbf{x})},\ \forall i \in F \tag{3}$$

where the function Φ is called a balance function and is defined recursively as follows: $\Phi(\mathbf{0}) = 1$, and $\Phi(\mathbf{x}) = 0$ $\forall \mathbf{x}$ s.t. $x_i < 0$ for some i, otherwise,

$$\Phi(\mathbf{x}) = \max_{A \subset F} \left\{ \frac{\sum_{i\in A} \Phi(\mathbf{x} - \mathbf{e}_i)}{\mu(A)} \right\}. \tag{4}$$

As shown in [3], (3) ensures the property of insensitivity, while (4) ensures that $\mathbf{r}(\mathbf{x})$ for each $\mathbf{x}$ lies in the capacity region, i.e., the constraints $\sum_{i\in A} r_i(\mathbf{x}) \leq \mu(A)$ are satisfied for each A. It also ensures that there exists a set $B \subset A_{\mathbf{x}}$ for which $\sum_{i\in B} r_i(\mathbf{x}) = \mu(B)$. In fact the BF allocation is the unique policy satisfying the above properties.

It was shown in [2, 3] that if $\boldsymbol{\rho} \in \hat{\mathcal{C}}$, the state process $(\mathbf{X}^B(t) : t \geq 0)$ is asymptotically stationary. Further, under this condition, its stationary distribution is given by

$$\pi(\mathbf{x}) = \frac{\Phi(\mathbf{x})}{G(\boldsymbol{\rho})} \prod_{i\in A_{\mathbf{x}}} \rho_i^{x_i} \quad \text{where} \quad G(\boldsymbol{\rho}) = \sum_{\mathbf{x}'} \Phi(\mathbf{x}') \prod_{i \in A_{x'}} \rho_i^{x'_i}.$$

The existence of such an expression for stationary distribution makes balanced fairness amenable for time-averaged performance analysis, a property we will use extensively in the sequel. While, in general, BF may result in wasteful resource allocation, e.g., BF is not Pareto efficient for certain triangle networks studied in [3], for polymatroid capacity regions BF has been shown to be Pareto efficient which leads us to following two results.

Proposition 3. ([22]) *For polymatroid capacity regions $\mathcal{C}$, BF rate allocation is Pareto efficient, i.e., $\sum_{i\in A_{\mathbf{x}}} r_i^B(\mathbf{x}) = \mu(A_{\mathbf{x}})$ for each $\mathbf{x}$.*

In [22], the following easily computable *exact* expression for mean delay was provided for homogeneous loads.

PROPOSITION 4. ([22]) *Consider a system with symmetric polymatroid capacity region $\mathcal{C}$ and with homogenous load $\boldsymbol{\rho} \in \hat{\mathcal{C}}$, i.e., for all $j \in F$ we have $\rho_j = \rho$. Then, the mean delay under balanced fair resource allocation to serve the jobs is given by,*

$$E[D^B_{\boldsymbol{\rho}}] = \frac{\nu \hat{F}(\rho)}{F(\rho)}, \tag{5}$$

where, $F(\rho)$ and $\hat{F}(\rho)$ can be recursively obtained as follows:

$$F(\rho) = \sum_{k=0}^{n} F_k(\rho), \tag{6}$$

where, $F_0(\rho) = 1$, and for $k \geq 1$,

$$F_k(\rho) = \frac{(n-k+1)\rho F_{k-1}(\rho)}{h(k) - k\rho}. \tag{7}$$

Also,

$$\hat{F}(\rho) = \sum_{k=0}^{n} \frac{k}{n} \hat{F}_k(\rho), \tag{8}$$

where, $\hat{F}_0(\rho) = 0$, and for $k \geq 1$,

$$\hat{F}_k(\rho) = \frac{F_k(\rho) + \frac{n-k+1}{k} F_{k-1}(\rho) + \frac{(n-k+1)(k-1)}{k} \rho \hat{F}_{k-1}(\rho)}{h(k) - k\rho}. \tag{9}$$

An expression for mean delay for a system with an arbitrary polymatroid capacity region and arbitrary load is also provided in [22], but has exponential computation complexity.

4. PERFORMANCE BOUNDS

Recall that for each rate allocation policy considered in Section 3, namely Greedy, αF, and BF, the underlying state process is asymptotically stationary if the load $\boldsymbol{\rho} \in \hat{\mathcal{C}}$. Thus the corresponding mean delays of the system's jobs are finite. In this section, we assume that the **capacity region $\mathcal{C}$ is symmetric**, and develop explicit and easily computable bounds on the mean delay of jobs in systems with Greedy or αF rate allocation under potentially heterogeneous load $\boldsymbol{\rho}$ within a subset of the stability region $\hat{\mathcal{C}}$.

Our goal here is to enable performance analysis for a general enough class of systems so as to allow us to develop quantitative and qualitative insights for large-scale systems prevalent today. For example, the bounds developed below will enable us to later characterize user-performance in downloading files from heterogeneous (in loads and service capacities) large-scale content delivery systems supporting parallel servicing of downloads.

Below we develop performance bounds for the following three cases:

(i) Homogeneous loads: We provide an upper bound for mean delay for loads $\boldsymbol{\rho} \in \hat{\mathcal{C}}$ which are *homogeneous across classes with non-zero entries*, i.e., if A is the set of classes such that $\rho_i > 0$ for each $i \in A$, then $\rho_i = \rho_j$ for each $i, j \in A$.

(ii) Dominance bound: For loads $\boldsymbol{\rho}, \boldsymbol{\rho}' \in \hat{\mathcal{C}}$ such that $\boldsymbol{\rho} \leq \boldsymbol{\rho}'$ we show that the system with load $\boldsymbol{\rho}$ has lower mean delay than that with load $\boldsymbol{\rho}'$. Thus, if $\boldsymbol{\rho}'$ is homogeneous across non-zero entries as described above then we have an upper bound for mean delay for $\boldsymbol{\rho}$ as well, even if $\boldsymbol{\rho}$ is heterogeneous.

(iii) Majorization bound: We consider loads $\boldsymbol{\rho}, \boldsymbol{\rho}' \in \hat{\mathcal{C}}$ such that $\boldsymbol{\rho} \prec \boldsymbol{\rho}'$. Further, suppose that $\boldsymbol{\rho}'$ is homogeneous across non-zero entries as described above. Then, we show that the system with load $\boldsymbol{\rho}$ has lower mean delay than that with load $\boldsymbol{\rho}'$.

Using the above majorization bound, we can bound mean delay for a larger subset of heterogeneous loads as compared to the dominance bound. For example, consider $\boldsymbol{\rho} = (\rho, \frac{1}{2}\rho, \frac{1}{2}\rho)$. Recall, for symmetric rank functions we have $\mu(A) = h(|A|)$ for each $A \subset F$, where $h(.)$ is concave. Now, if $\frac{1}{3}h(3) < \rho < \frac{1}{2}h(2)$, then $\boldsymbol{\rho}' = (\rho, \rho, 0)$ is in $\hat{\mathcal{C}}$ but $\boldsymbol{\rho}'' = (\rho, \rho, \rho)$ is not. Then the majorization bound holds for $\boldsymbol{\rho}$ but the dominance bound does not. Further, even if $\boldsymbol{\rho}''$ is in $\hat{\mathcal{C}}$, the bound obtained through $\boldsymbol{\rho}'$ may be tighter.

The bounds for each case will be obtained through coupling arguments on the corresponding state processes, followed by an application of Little's law.

4.1 Homogeneous Loads

Consider the following set of loads:

$$\mathcal{B}_H \triangleq \{\boldsymbol{\rho} \in \hat{\mathcal{C}} : \exists A \subset F \text{ s.t. } \rho_i = \rho_j \ \forall i, j \in A \text{ and } \rho_i = 0 \ \forall i \in F \backslash A\}.$$

Since by Proposition 1 the Greedy rate allocation is delay optimal for homogeneous loads, for each $\boldsymbol{\rho} \in \mathcal{B}_H$ one can immediately conclude that the performance of BF as obtained in Proposition 4 is an upper bound for Greedy. Below we show that this performance upper bound via BF also holds for αF rate allocation.

To that end we show a coupling result for systems under αF and BF rate allocations. In the process, we prove and use the property that αF is more greedy than BF in the following sense: if the state process corresponding to αF is the same as or more balanced than that of BF, then αF assigns larger rate to bigger queues than BF. This in turn keeps the state process for αF more balanced in the future. For a proof of the theorem below see the Appendix.

Theorem 1. *Consider a system with symmetric polymatroid capacity region and load $\boldsymbol{\rho} \in \mathcal{B}_H$, i.e., $\boldsymbol{\rho}$ is homogeneous across classes with non-zero entries. Then the following statements hold:*

1. *Let $(\mathbf{X}^\alpha(t) : t \geq 0)$ and $(\mathbf{X}^B(t) : t \geq 0)$ be state processes under αF and BF rate allocation. If $\mathbf{X}^\alpha(0) \prec_w \mathbf{X}^B(0)$ then we have $\mathbf{X}^\alpha(t) \prec_w^{st} \mathbf{X}^B(t)$ for each $t \geq 0$.*

2. *The mean delays for systems with αF and BF rate allocation for load $\boldsymbol{\rho} \in \mathcal{B}_H$ satisfy the following:*

$$E[D^\alpha_{\boldsymbol{\rho}}] \leq E[D^B_{\boldsymbol{\rho}}].$$

4.2 Dominance Bounds

The result below maintains that for αF and Greedy rate allocations, if $\boldsymbol{\rho} \leq \boldsymbol{\rho}'$ then the state process of a system under load $\boldsymbol{\rho}$ is stochastically dominated by that under $\boldsymbol{\rho}'$. The result can be shown using a per class coupling argument, along with the following per-class rate monotonicity which is satisfied by both αF and Greedy: for each $i \neq j$ we have $r_i(\mathbf{x}) \geq r_i(\mathbf{x} + \mathbf{e}_j)$. We omit the proof for brevity.

Theorem 2. *Consider a system with symmetric polymatroid capacity region $\mathcal{C}$. The rate allocation $\mathbf{r}(.)$ is either αF or Greedy. Let $\boldsymbol{\rho} \leq \boldsymbol{\rho}'$. Then the following statements hold:*

1. *Let $(\mathbf{X}(t) : t \geq 0)$ and $(\mathbf{X}'(t) : t \geq 0)$ be state processes under loads $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$. If $\mathbf{X}(0) \leq \mathbf{X}'(0)$, then we have $\mathbf{X}(t) \leq^{st} \mathbf{X}'(t)$ for each $t \geq 0$.*

2. *The mean delays for systems with loads $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ satisfy the following:*
$$E[D_{\boldsymbol{\rho}}] \leq E[D_{\boldsymbol{\rho}'}]$$

Theorem 2 along with Theorem 1 allows us to bound the mean delay for any load in the following region:

$$\mathcal{B}_D \triangleq \{\boldsymbol{\rho} \in \hat{\mathcal{C}} : \exists \boldsymbol{\rho}' \in \mathcal{B}_H \text{ s.t. } \boldsymbol{\rho} \leq \boldsymbol{\rho}'\},$$

or equivalently,

$$\mathcal{B}_D \triangleq \left\{\boldsymbol{\rho} \in \hat{\mathcal{C}} : \max_i \rho_i < \frac{h(k)}{k} \text{ where } k = |\{i : \rho_i > 0\}|\right\}.$$

Theorem 2 implies that for αF and Greedy rate allocations, the mean delay for each load $\boldsymbol{\rho} \in \mathcal{B}_D$ can be bounded by that for a corresponding load $\boldsymbol{\rho}' \in \mathcal{B}_H$, which in turn has an easily computable bound through Theorem 1. Thus, we get the following corollary.

Corollary 1. *Consider a system with symmetric polymatroid capacity region and load $\boldsymbol{\rho} \in \mathcal{B}_D$. Let $\rho' = \max_i \rho_i$. Let $\boldsymbol{\rho}'$ be such that for each $i \in F$ we have $\rho'_i = \rho'$ if $\rho_i > 0$ and $\rho'_i = 0$ if $\rho_i = 0$. Then, mean delays for systems with Greedy and αF rate allocations for load $\boldsymbol{\rho}$ satisfy the following:*

$$E[D^G_{\boldsymbol{\rho}}] \leq E[D^B_{\boldsymbol{\rho}'}], \text{ and } E[D^\alpha_{\boldsymbol{\rho}}] \leq E[D^B_{\boldsymbol{\rho}'}].$$

4.3 Majorization Bounds

The theorem below generalizes the Dominance bound to provide a mean delay bound for a system with load $\boldsymbol{\rho}$ such that there exists $\boldsymbol{\rho}' \in \mathcal{B}_H$ which satisfies $\boldsymbol{\rho} \prec \boldsymbol{\rho}'$.

Its proof is similar to that of Theorem 1, where instead of relative greediness between rate allocations, we use the following balancing property satisfied by both αF and Greedy: if state $\mathbf{x}$ is more balanced than state $\mathbf{x}'$, than the rate allocation $\mathbf{r}(.)$ would provide larger rates to longer queues in state $\mathbf{x}$ as compared to $\mathbf{x}'$, and thus balancing it even further. For a brief discussion of this property, see the Appendix. We omit the proof for brevity.

Theorem 3. *Consider a system with symmetric polymatroid capacity region $\mathcal{C}$. The rate allocation $\mathbf{r}(.)$ is either αF or Greedy. Let $\boldsymbol{\rho}, \boldsymbol{\rho}' \in \hat{\mathcal{C}}$ be such that $\boldsymbol{\rho} \prec \boldsymbol{\rho}'$ and $\boldsymbol{\rho}' \in \mathcal{B}_H$, i.e., $\boldsymbol{\rho}'$ is homogeneous across classes with non-zero entries. Then the following statements hold:*

1. *Let $(\mathbf{X}(t) : t \geq 0)$ and $(\mathbf{X}'(t) : t \geq 0)$ be state processes under loads $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$. If $\mathbf{X}(0) \prec_w \mathbf{X}'(0)$, then we have $\mathbf{X}(t) \prec^{st}_w \mathbf{X}'(t)$ for each $t \geq 0$.*

2. *The mean delays for systems with loads $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ satisfy the following:*
$$E[D_{\boldsymbol{\rho}}] \leq E[D_{\boldsymbol{\rho}'}]$$

Theorem 3 above is stronger than Theorem 2 in the sense that it only requires the condition $\boldsymbol{\rho} \prec_w \boldsymbol{\rho}'$ instead of $\boldsymbol{\rho} \leq \boldsymbol{\rho}'$. However, it is weaker in the sense that it requires $\boldsymbol{\rho}'$ to be in $\mathcal{B}_H$ and that it gives stochastic weak-majorization of the corresponding state processes instead of stochastic dominance.

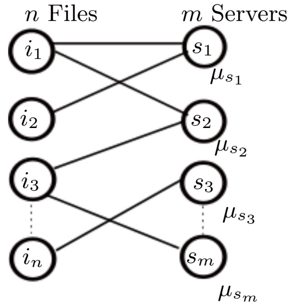

Figure 1: Graph $G^{(n)} = (F^{(n)} \cup S^{(n)}; E^{(n)})$ modeling the placement of copies of n files across $m = \lceil bn \rceil$ servers with finite service capacities in a CDN-like infrastructure.

For both $\mathbf{r}^G(.)$ and $\mathbf{r}^\alpha(.)$, Theorem 3, along with Theorem 1 and Proposition 1, allows us to bound the mean delay for any load in the following region:

$$\mathcal{B}_M \triangleq \{\boldsymbol{\rho} \in \hat{\mathcal{C}} : \exists \boldsymbol{\rho}' \in \mathcal{B}_H \text{ s.t. } \boldsymbol{\rho} \prec \boldsymbol{\rho}'\},$$

or equivalently,

$$\mathcal{B}_M \triangleq \left\{\boldsymbol{\rho} \in \hat{\mathcal{C}} : \exists k \leq n \text{ s.t. } \max_i \rho_i < \frac{h(k)}{k} \text{ and } |\boldsymbol{\rho}| < h(k)\right\}.$$

Theorem 3 implies that for αF and Greedy rate allocation, the mean delay for each load $\boldsymbol{\rho} \in \mathcal{B}_M$ can be bounded by that for a corresponding load $\boldsymbol{\rho}' \in \mathcal{B}_H$, which in turn has an easily computable bound through Theorem 1. Thus, we get the following corollary.

Corollary 2. *Consider a system with symmetric polymatroid capacity region and load $\boldsymbol{\rho} \in \mathcal{B}_M$. Let $\rho' = \max_{i \in F} \rho_i$. Let $k = \min\{l : \rho' \leq \frac{h(l)}{l}$ and $|\boldsymbol{\rho}| \leq h(l)\}$. Let A be an arbitrary subset of F of size k and $\boldsymbol{\rho}'$ be such that $\rho'_i = \rho' \;\; \forall i \in A$ and $\rho'_i = 0$ otherwise. Then, the mean delays for systems with Greedy and αF rate allocations for load $\boldsymbol{\rho}$ satisfy the following:*

$$E[D^G_{\boldsymbol{\rho}}] \leq E[D^B_{\boldsymbol{\rho}'}], \text{ and } E[D^\alpha_{\boldsymbol{\rho}}] \leq E[D^B_{\boldsymbol{\rho}'}].$$

It is easy to check that for each $\boldsymbol{\rho} \in \mathcal{B}_M$ the computation of the mean delay upper bound as given by Corollary 2 has complexity $O(n)$ when computed using Proposition 4.

5. UNIFORM SYMMETRY IN LARGE SYSTEMS

Large content delivery infrastructure, where servers can jointly serve file-download requests, not only have polymatroid capacity but under appropriate assumptions become approximately symmetric.

Consider a sequence of bipartite graphs $G^{(n)} = (F^{(n)} \cup S^{(n)}; E^{(n)})$ where $F^{(n)}$ is a set of n files, $S^{(n)}$ is a set of $m = \lceil bn \rceil$ servers for some constant b, and each edge $e \in E^{(n)}$ connecting a file $i \in F^{(n)}$ and server $s \in S^{(n)}$ implies that a copy of file i is available at server s. For each node $s \in S^{(n)}$, let $N^{(n)}_s$ denote the set of neighbors of server s, i.e., the set of files it stores and can serve. Henceforth, wherever possible, we will avoid the use of ceil and floor notations to avoid clutter.

We associate each file in $F^{(n)}$ with a class of job arrivals each corresponding to a file download request. The arrival processes and service requirements are as described in Section 2, with $\boldsymbol{\lambda}^{(n)}$ and $\boldsymbol{\rho}^{(n)}$ representing the corresponding arrival rates and loads. Further, we let the service capacity of each server $s \in S^{(n)}$ be μ_s bits per second.

We allow each server $s \in S^{(n)}$ to concurrently serve the jobs with classes $N_s^{(n)}$ as long as the total service rate does not exceed μ_s. The service rate for each job is the sum of the rates it receives from different servers. For any $A \subset F^{(n)}$, let $\mu^{(n)}(A)$ be the maximum sum rate at which jobs with file-class in A could be served, i.e.,

$$\mu^{(n)}(A) \triangleq \sum_{s \in S^{(n)}} \mathbf{1}_{\left\{A \cap N_s^{(n)} \neq \emptyset\right\}} \mu_s.$$

Clearly any rate allocation $\mathbf{r}(.)$ for such a system must satisfy the following constraints for each state $\mathbf{x}$: $\forall A \subset F^{(n)}$,

$$\sum_{i \in A} r_i(\mathbf{x}) \le \mu^{(n)}(A).$$

It was shown in [22] that $\mu^{(n)}(.)$ is submodular and that the corresponding polymatroid

$$\mathcal{C}^{(n)} \triangleq \left\{ \mathbf{r} \ge \mathbf{0} : \sum_{i \in A} r_i \le \mu^{(n)}(A), \ \forall A \subset F^{(n)} \right\}$$

is indeed the capacity region for such a system, i.e., each $\mathbf{r} \in \mathcal{C}^{(n)}$ is achievable.

Note that $\mathcal{C}^{(n)}$ will in general be an asymmetric polymatroid depending upon edges $E^{(n)}$ and service capacities μ_s for each $s \in S^{(n)}$. However, we show below that if copies of files are stored across servers at random and scaled appropriately with n then, as n increases, $\mathcal{C}^{(n)}$ gets uniformly close to a symmetric polymatroid, subject to the following assumptions:

Assumption 1 (Heterogeneous server capacities). *$S^{(n)}$ is partitioned into a finite number of groups where each group has $\Omega(n)$ number of servers. Within each group, the server capacities are homogeneous. The server capacities across groups may be heterogeneous such that average of service capacity across servers*

$$\xi \triangleq \frac{1}{m} \sum_{s \in S^{(n)}} \mu_s$$

is independent of n.

Assumption 2 (Randomized file placement). *Let $(c_n : n \in \mathbb{N})$ be a sequence such that*

$$c_n = \omega(\log n).$$

For each file $i \in F^{(n)}$, store a copy in c_n different servers chosen uniformly and independently at random.

A randomized placement of file copies implies a random system configuration, i.e., a random graph. Let $\mathcal{E}^{(n)}$ denote the random set of edges resulting Assumption 2. Similarly, for each $s \in S^{(n)}$, let $\mathcal{N}_s^{(n)}$ denote the random set of neighbors of s, i.e., the random set of files stored in server s. Let $M^{(n)}(.)$ denote the corresponding random rank function, and $\mu^{(n)}(.)$ a possible realization. Then, for each $A \subset F^{(n)}$, we have

$$M^{(n)}(A) = \sum_{s \in S^{(n)}} \mathbf{1}_{\left\{A \cap \mathcal{N}_s^{(n)} \neq \emptyset\right\}} \mu_s,$$

where $\mathbf{1}_{\left\{A \cap \mathcal{N}_s^{(n)} \neq \emptyset\right\}}$ is now a Bernoulli random variable indicating if a copy of at least one of the files in A is placed in s. In fact, for each $A \subset F^{(n)}$ such that $|A| = k$, the set $\left\{ \mathbf{1}_{\left\{A \cap \mathcal{N}_s^{(n)} \neq \emptyset\right\}} : s \in S^{(n)} \right\}$ is a set of m negatively associated Bernoulli$(p_k^{(n)})$ random variables [8] where $p_k^{(n)}$ is the probability that a given server is assigned at least one of the kc_n copies of files in A and is given by

$$p_k^{(n)} \triangleq 1 - \left(1 - \frac{1}{m}\right)^{kc_n} \quad \forall k = 0, 1, \ldots, n.$$

By linearity of expectation, for each $A \subset F^{(n)}$, we have

$$\bar{\mu}^{(n)}(A) \triangleq E[M^{(n)}(A)] = \xi m p_{|A|}^{(n)}.$$

Note, $\bar{\mu}^{(n)}(A)$ depends on A only through $|A|$ and is thus symmetric. The theorem below shows that with high probability we can bound the random rank function $M^{(n)}(.)$ uniformly over all $A \subset F^{(n)}$, from above as well as from below, with a symmetric rank function which is close to $\bar{\mu}^{(n)}(A)$. See Section 5.1 for a proof.

Theorem 4. *Fix ϵ independent of n such that $0 < \epsilon < 1$. Consider a sequence of systems with n files and $m = \lceil bn \rceil$ servers, where $b > 0$ is a constant. Under Assumptions 1 and 2, let $M^{(n)}(.)$ be the corresponding random rank function. Then, there exists a sequence $(g_n : n \in \mathbb{N})$ such that $g_n = \omega(\log n)$, and*

$$P\left(\exists A \subset F^{(n)} \text{ s.t. } M^{(n)}(A) \le (1-\epsilon)\bar{\mu}^{(n)}(A) \right) \le e^{-g_n},$$

and

$$P\left(\exists A \subset F^{(n)} \text{ s.t. } M^{(n)}(A) \ge (1+\epsilon)\bar{\mu}^{(n)}(A) \right) \le e^{-g_n}.$$

This result gives us following corollary on the random capacity region associated with $M^{(n)}(.)$ generated by random file placement. Recall, $\bar{\mu}^{(n)}(A) = E[M^{(n)}(A)]$ for all $A \subset F^{(n)}$, and let

$$\bar{\mathcal{C}}^{(n)} \triangleq \left\{ \mathbf{r} \ge \mathbf{0} : \sum_{i \in A} r_i \le \bar{\mu}^{(n)}(A), \ \forall A \subset F^{(n)} \right\}.$$

Thus $\bar{\mathcal{C}}^{(n)}$ is the (symmetric) capacity region associated with the average rank function $\bar{\mu}(.)$. Then, the following holds:

Corollary 3. *Fix ϵ independent of n such that $0 < \epsilon < 1$. Under Assumptions 1 and 2, the random capacity region associated with randomized file placement is a subset of $(1+\epsilon)\bar{\mathcal{C}}^{(n)}$ and a superset of $(1-\epsilon)\bar{\mathcal{C}}^{(n)}$ with high probability.*

Further, under Assumption 1, there exists a deterministic file placement where $c_n = \omega(\log n)$ copies of each file are stored across servers such that the corresponding capacity region $\mathcal{C}^{(n)}$ is a subset of $(1+\epsilon)\bar{\mathcal{C}}^{(n)}$ and a superset of $(1-\epsilon)\bar{\mathcal{C}}^{(n)}$.

5.1 Proof of Theorem 4

Here, we will only show

$$P\left(\exists A \subset F^{(n)} \text{ s.t. } M^{(n)}(A) \le (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \le e^{-g_n},$$

The other bound follows in similar fashion.

For now, suppose $\mu_s = \xi$ for each $s \in S^{(n)}$. We relax this assumption later.

We first provide a bound for $P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right)$ for each $A \subset F^{(n)}$. Then, for each $k = 1, 2, \ldots, n$, we use the union bound to obtain a uniform bound over all sets $A \subset F^{(n)}$ such that $|A| = k$. The bound we provide for $P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right)$ is small enough so that the above union bound is small too. Then, yet another use of the union bound would give us the uniform result over all sets $A \subset F^{(n)}$.

Now, if the random variables $\left\{\mathbf{1}_{\left\{A \cap \mathcal{N}_s^{(n)} \neq \emptyset\right\}} : s \in S^{(n)}\right\}$ were independent Bernoulli($p_k^{(n)}$), then the following two concentration results would hold [18]: Fix $k \in \{1, \ldots, n\}$. For each set $A \subset F^{(n)}$ such that $|A| = k$, we have

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{-\frac{\epsilon^2}{2} m p_k^{(n)}}, \qquad (10)$$

and,

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{-mH\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right)}, \qquad (11)$$

where $H(p||q)$ is the KL divergence between Bernoulli(p) and Bernoulli(q) random variables, given by

$$H(p||q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right).$$

However, in reality, since $\left\{\mathbf{1}_{\left\{A \cap \mathcal{N}_s^{(n)} \neq \emptyset\right\}} : s \in S^{(n)}\right\}$ are negatively associated Bernoulli($p_k^{(n)}$) random variables, the above Chernoff bounds still apply [8].

In the sequel, we will use the following two technical lemmas. Their proofs are provided in the Appendix A.3.

Lemma 1. *Let a sequence $(g_n : n \in \mathbb{N})$ be such that $g_n = o(c_n)$. Let δ_1 be a positive constant independent of n such that $\delta_1 < 1$. Then, for large enough n, we have*

$$p_k^{(n)} \geq \frac{\delta_1 g_n}{n} k \quad \forall k \in \left\{0, 1, \ldots, \left\lfloor \frac{n}{g_n} \right\rfloor\right\}.$$

Lemma 2. *There exists a positive constant δ such that $H\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right) \geq -\delta + \epsilon\frac{kc_n}{m}$.*

Now, let $(g_n : n \in \mathbb{N})$ be a sequence such that $g_n \triangleq (c_n \log n)^{1/2}$ for each n. The following properties of g_n can be easily checked:

$$g_n = \omega(\log n) \text{ and } g_n = o(c_n). \qquad (12)$$

We now provide a uniform bound over all sets $A \subset F^{(n)}$ such that $|A| = k$ for each $k \in \{1, \ldots, n\}$, under following two cases.

Case 1 $0 \leq k \leq \frac{n}{g_n}$: From Lemma 1, for each k we have

$$p_k^{(n)} \geq \delta_1 \frac{kg_n}{n},$$

for a suitably chosen positive constant δ_1 independent of n. In the sequel, δ_i for any $i \geq 1$ will be a suitably chosen positive constant independent of n.

Using the concentration result (10), for $|A| = k$ we get

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{-\frac{\epsilon^2}{2}\delta_1 bkg_n},$$

and using the union bound, we get

$$P\Big(\exists A \subset F^{(n)} \text{ s.t. } |A| = k \text{ and } M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\Big)$$
$$\leq e^{-\frac{\epsilon^2}{2}\delta_1 bkg_n}\binom{n}{k} \leq e^{-\frac{\epsilon^2}{2}\delta_1 bkg_n + k\log n} \leq e^{-\delta_2 kg_n}.$$

Case 2 $\frac{n}{g_n} < k \leq n$: In this case, we use the concentration result (11). From Lemma 2, we get

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{(\delta_6 m - \epsilon k c_n)}.$$

Since $g_n = o(c_n)$, for n large enough we get $\delta_6 m \leq (\epsilon/2)\frac{nc_n}{g_n}$. Also, for each $k > \frac{n}{g_n}$, we have $(\epsilon/2)\frac{nc_n}{g_n} \leq (\epsilon/2)kc_n$. Thus, for large enough n, $\delta_6 m - \epsilon k c_n \leq -(\epsilon/2)kc_n$ for each k such that $\frac{n}{g_n} < k \leq n$, and consequently,

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{-\delta_7 kc_n}$$

By using the union bound, for large enough n, we get

$$P\left(\exists A \subset F^{(n)} \text{ s.t. } |A| = k \text{ and } M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right)$$
$$\leq e^{-\delta_7 kc_n}\binom{n}{k} \leq e^{-\delta_7 kc_n + k\log n} \leq e^{-\delta_8 kc_n}.$$

Combining the above two cases, we can show that for large enough n there exists a positive constant δ_9 such that for each $k \in \{1, \ldots, n\}$ we have

$$P\left(\exists A \subset F^{(n)} \text{ s.t. } |A| = k \text{ and } M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right)$$
$$\leq e^{-\delta_9 g_n}.$$

Using the union bound again, we get

$$P\left(\exists A \subset F^{(n)} \text{ s.t. } M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right)$$
$$\leq ne^{-\delta_9 g_n} \leq e^{-\delta_9 g_n + \log n} \leq e^{-\delta_{10} g_n}.$$

Now, we relax the assumption $\mu_s = \xi$ for each $s \in S^{(n)}$ with Assumption 1. The above proof can then be used to show a similar concentration result for individual groups. The overall result follows by linearity of expectation and yet another use of the union bound. □

6. PERFORMANCE ROBUSTNESS

We now combine results from Section 4 and Section 5 to exhibit performance robustness in large systems. In Section 5, we showed that large systems support symmetric polymatroid capacity regions. This allows us to apply the performance bounds developed in Section 4 for symmetric polymatroid capacity regions.

However, there is one more hurdle to overcome before we can apply our bounds from Section 4. Recall, from Corollary 3, under Assumptions 1 and 2 the random capacity region *contains* and is *contained by* symmetric polymatroids with high probability. The realizations of the random capacity region, themselves, may still not be symmetric. We thus need to show that if the capacity region is bigger then the corresponding mean delay is smaller when subject to the same load.

Intuitively, larger capacity regions may imply larger service rates for each class, and may thus provide better performance. Although intuitively obvious, such results are not always straightforward. We show below that such a comparison result indeed holds under the following monotonicity conditions for rate allocations.

Definition 5 (*Monotonicity w.r.t. capacity region*). *We say that a rate allocation satisfies monotonicity w.r.t. capacity region if for any state* $\mathbf{x}$*, the rate allocation per class for a system with a larger capacity region dominates that with a smaller one.*

Recall, $\frac{r_i(\mathbf{x})}{x_i}$ is the rate allocated to each job in class i when the system is in state $\mathbf{x}$.

Definition 6 (*Per-job rate monotonicity*). *We say that a rate allocation* $\mathbf{r}(.)$ *satisfies per-job rate monotonicity if the following holds for all states* $\mathbf{x}$ *and* $\mathbf{x}'$ *such that* $\mathbf{x} \geq \mathbf{x}'$*: for each class* i*, we have* $\frac{r_i(\mathbf{x})}{x_i} \leq \frac{r_i(\mathbf{x}')}{x_i'}$*. In words, adding jobs into the system only decreases the rate allocated to each job.*

Note that the condition of per-job rate monotonicity is stronger than that of per-class rate monotonicity used in Section 4.2. Per-job rate monotonicity was first used in [4] to provide a comparison result similar to the lemma below. The following lemma can be shown to hold through a simple coupling argument across jobs for arbitrary polymatroid capacity regions.

Lemma 3. *Consider systems with arbitrary polymatroid capacity regions* $\mathcal{C}$ *and* $\tilde{\mathcal{C}}$ *such that* $\mathcal{C} \subset \tilde{\mathcal{C}}$*. Consider a rate allocation which satisfies monotonicity w.r.t. capacity region as well as per-job rate monotonicity. Then, the mean delay for capacity region* $\mathcal{C}$ *under arbitrary load* $\boldsymbol{\rho}$ *upper bounds that for capacity region* $\tilde{\mathcal{C}}$ *under the same load.*

It is easy to check that α-fair rate allocation satisfies per-job rate monotonicity as well as monotonicity w.r.t. capacity region. Thus, Lemma 3 holds for α-fair rate allocation. However, one can show that Greedy rate allocation may not satisfy either property for arbitrary polymatroid capacity regions. This further highlights the brittleness of Greedy rate allocation to asymmetries. Even for Balanced fair rate allocation it is not directly clear if the lemma holds. Thus, henceforth we will only consider α-fair rate allocation.

Now we are indeed ready with all the tools required to exhibit robustness in large scale systems.

Assumption 3 (*Load Heterogeneity*). *We consider a sequence of systems where load* $\boldsymbol{\rho}^{(n)}$ *for each* n *is allowed to be within a set* $\mathcal{B}^{(n)}$ *defined as follows: Consider a sequence* $(\theta_n : n \in \mathbb{N})$ *such that* $\theta_n = \omega(1)$*,* $\theta_n = o(\frac{n}{\log n})$*, and* $\theta_n = o(c_n)$*. Also, fix a constant* $\gamma < 1$ *independent of* n*. For each* n*:*

$$\mathcal{B}^{(n)} \triangleq \left\{ \boldsymbol{\rho} : \max_{i \in F^{(n)}} \rho_i \leq \theta_n \text{ and } |\boldsymbol{\rho}| \leq \gamma \xi m \right\}.$$

The condition $|\boldsymbol{\rho}| \leq \gamma\xi m$ implies that we allow load to increase linearly with system size. Also, since $\theta_n = \omega(1)$, the condition $\max_i \rho_i \leq \theta_n$ implies that we allow load across servers to be increasingly heterogeneous. The condition $\theta_n = o(\frac{n}{\log n})$ limits the heterogeneity allowed in the system. Further, the condition $\theta_n = o(c_n)$ would allow us to claim stability, and to show that the mean delay of the system tends to 0 as n increases.

The following is the main result of this section.

Theorem 5. *Consider a sequence of systems with* n *files* $F^{(n)}$ *and* $m = \lceil bn \rceil$ *servers* $S^{(n)}$*, where* b *is a constant. For each* n*, let the total service capacity of servers be* ξm*, where* ξ *is independent of* n*.* $S^{(n)}$ *is partitioned into a finite number of heterogeneous groups, each with* $\Omega(n)$ *servers and equal per-server capacity. Let* $(c_n : n \in \mathbb{N})$ *be a sequence such that* $c_n = \omega(\log n)$*. We allow* c_n *copies of each file to be placed across the servers.*

Let the mean service requirement of file-download jobs be ν*, where* ν *is independent of* n*. Let* $(\theta_n : n \in \mathbb{N})$ *be such that* $\theta_n = \omega(1)$*, but* $o\left(\min(\frac{n}{\log n}, c_n)\right)$*. Fix a constant* $\gamma < 1$*. Let* $\mathcal{B}^{(n)} = \{\boldsymbol{\rho} : \max_i \rho_i \leq \theta_n$ *and* $|\boldsymbol{\rho}| \leq \gamma\xi m\}$*. For each* n*, let load across file classes be* $\boldsymbol{\rho}^{(n)} \in \mathcal{B}^{(n)}$*.*

Fix a constant $\delta > 1$*. Then, there exists an integer* n_δ *such that for each* $n \geq n_\delta$ *the following holds: there exists at least one file placement policy such that the mean delay for file-download jobs with* α*-fair rate allocation satisfies the following bound:*

$$E[D^{(n)}] \leq \delta \frac{\nu\theta_n}{\xi c_n} \frac{1}{\gamma} \log\left(\frac{1}{1-\gamma}\right).$$

Further, for each $n \geq n_\delta$*, if the* c_n *copies of each file are stored uniformly at random across servers, then the above bound holds with high probability.*

6.1 Proof of Theorem 5

We first show the existence of a file placement policy such that the mean delay bound is satisfied. Without loss of generality, assume $\delta < \frac{1}{\gamma}$.

From Corollary 3, and definitions of $\bar{\mathcal{C}}^{(n)}$ and $\bar{\mu}^{(n)}(.)$, for large enough n there exists a file placement such that the corresponding capacity region contains the following symmetric polymatroid:

$$\tilde{\mathcal{C}}^{(n)} \triangleq \left\{ \mathbf{r} \geq \mathbf{0} : \sum_{i \in A} r_i \leq h^{(n)}(|A|),\ \forall A \subset F^{(n)} \right\},$$

where

$$h^{(n)}(k) \triangleq (1/\delta)\xi m \left(1 - e^{-\frac{kc_n}{m}}\right) \ \ \forall k = 0, 1, \ldots, n.$$

Thus, from Lemma 3, for α-fair rate allocations it is sufficient to consider $\tilde{\mathcal{C}}^{(n)}$. Further, since $\tilde{\mathcal{C}}^{(n)}$ is monotonic in c_n, it is sufficient to assume that $c_n = o(\frac{n}{\log n})$ since, if it is not, we can set c_n to be equal to $\sqrt{\frac{n}{\log n}\theta_n}$ and all the assumptions still hold. Thus, henceforth we assume that

$$c_n = o(\frac{n}{\log n}).$$

Let $\xi' \triangleq \xi/\delta$. Thus, we get

$$h^{(n)}(k) = \xi' m \left(1 - e^{-\frac{kc_n}{m}}\right) \ \ \forall k = 0, 1, \ldots, n.$$

Since $\gamma\xi m < \xi' m$ and $\theta_n = o(c_n)$, one can check that $B^{(n)}$ is a subset of $\tilde{\mathcal{C}}^{(n)}$ for large enough n, and we get stability.

Let $t_n \triangleq \left\lceil \frac{\gamma\xi' m}{\theta_n} \right\rceil$. Let $A^{(n)}$ be an arbitrary subset of $F^{(n)}$ such that $|A^{(n)}| = t_n$. Let $\hat{\boldsymbol{\rho}}^{(n)} = (\hat{\rho}_i^{(n)} : i \in F^{(n)})$ where

$\hat{\rho}_i^{(n)} = \theta_n$ if $i \in A^{(n)}$ and 0 otherwise. Then, it is easy to show that for each n, we have

$$B^{(n)} \subset \left\{ \boldsymbol{\rho} : \boldsymbol{\rho} \prec_w \hat{\boldsymbol{\rho}}^{(n)} \right\}.$$

Thus, from Theorem 3, it is sufficient to show that the bound on mean delay holds for balanced fair rate allocation under load $\boldsymbol{\rho}^{(n)} = \hat{\boldsymbol{\rho}}^{(n)}$.

Henceforth, we assume BF rate allocation and let load $\boldsymbol{\rho}^{(n)} = \hat{\boldsymbol{\rho}}^{(n)}$. For each n, we invoke Proposition 4 with ρ replaced by θ_n and n replaced by t_n, where for each $k = 0, 1, \ldots, t_n$ we let[1]

$$\pi_k^{(n)} \triangleq \frac{F_k(\theta_n)}{F(\theta_n)}, \text{ and } \tau_k^{(n)} \triangleq \frac{\hat{F}_k(\theta_n)}{F(\theta_n)}.$$

Then, we have

$$E[D^{(n)}] = \nu \sum_{k=1}^{t_n} \frac{k}{t_n} \tau_k^{(n)}. \tag{13}$$

Also, we have $\tau_0^{(n)} = 0$, $\pi_0^{(n)} = 1/F(\theta_n)$, and for each $k = 1, \ldots, t_n$ we have

$$\pi_k^{(n)} = \frac{(t_n - k + 1)\theta_n}{h^{(n)}(k) - k\theta_n} \pi_{k-1}^{(n)}, \tag{14}$$

and

$$\tau_k^{(n)} = \frac{\pi_k^{(n)} + \frac{t_n - k + 1}{k} \pi_{k-1}^{(n)} + \frac{(t_n - k + 1)(k-1)}{k} \theta_n \tau_{k-1}^{(n)}}{h^{(n)}(k) - k\theta_n}. \tag{15}$$

First, we show the following result.

Theorem 6. *For any positive constants $\epsilon > 1$ and $\epsilon' < 1$ independent of n, there exists a constant $\delta' < 1$ such that for large enough n we have*

$$\sum_{k=\epsilon' b \log(\frac{1}{1-\gamma}) \frac{n}{c_n}}^{\epsilon b \log(\frac{1}{1-\gamma}) \frac{n}{c_n}} \pi_k^{(n)} \geq 1 - \delta'^{\frac{m}{c_n}}. \tag{16}$$

Proof. Fix a constant $0 < \delta_{11} < 1$. Let

$$k_\downarrow^{(n)} = \frac{m}{c_n} \log \left(\frac{1}{1 - \gamma\delta_{11}} \right).$$

Then, we have $h^{(n)}(k_\downarrow) = \gamma\delta_{11}\xi' m$. In fact, we have $h^{(n)}(k) \leq \gamma\delta_{11}\xi' m, \ \forall k \leq k_\downarrow^{(n)}$. Using (14), for each $k \leq k_\downarrow^{(n)}$, we have

$$\begin{aligned} \pi_k^{(n)} \geq \frac{(t_n - k + 1)\theta_n}{\gamma\delta_{11}\xi' m - k\theta_n} \pi_{k-1}^{(n)} &\geq \frac{t_n\theta_n - (k_\downarrow^{(n)} - 1)\theta_n}{\gamma\delta_{11}\xi' m} \pi_{k-1}^{(n)} \\ &= \frac{\gamma\xi' m - o(n)}{\gamma\delta_{11}\xi' m} \pi_{k-1}^{(n)} \geq \frac{1}{\delta_{12}} \pi_{k-1}^{(n)}, \end{aligned}$$

for a positive constant δ_{12} such that $\delta_{11} < \delta_{12} < 1$, and large enough n. Equivalently, $\pi_k^{(n)} \leq \delta_{12}\pi_{k+1}^{(n)} \quad \forall k < k_\downarrow^{(n)}$. Fix a positive constant $\epsilon_1 < 1$. Then, for all $k < \epsilon_1 k_\downarrow^{(n)}$, we have

$$\pi_k^{(n)} \leq \delta_{12}^{(1-\epsilon_1)k_\downarrow^{(n)}} \pi_{k_\downarrow^{(n)}}^{(n)}$$

Similarly, for a constant δ_{13} such that $\gamma < \delta_{13} < 1$, a constant $\epsilon_2 > 1$, and $k_\uparrow^{(n)} = \frac{m}{c_n} \log \left(\frac{1}{1 - \gamma/\delta_{13}} \right)$ one can show that there exists a constant $\delta_{14} < 1$ such that

$$\pi_k^{(n)} \leq \delta_{14}^{(\epsilon_2 - 1)k_\uparrow^{(n)}} \pi_{k_\uparrow^{(n)}}^{(n)} \quad \forall k > \epsilon_2 k_\uparrow^{(n)}$$

Thus, we get

$$\begin{aligned} 1 = \sum_{k=0}^{t_n} \pi_k^{(n)} &= \sum_{k=0}^{\epsilon_1 k_\downarrow^{(n)} - 1} \pi_k + \sum_{k=\epsilon_1 k_\downarrow^{(n)}}^{\epsilon_2 k_\uparrow^{(n)}} \pi_k^{(n)} + \sum_{\epsilon_2 k_\uparrow^{(n)} + 1}^{t_n} \pi_k^{(n)} \\ &\leq (\epsilon_1 k_\downarrow^{(n)}) \delta_{12}^{(1-\epsilon_1)k_\downarrow^{(n)}} + \sum_{k=\epsilon_1 k_\downarrow^{(n)}}^{\epsilon_2 k_\uparrow^{(n)}} \pi_k^{(n)} + \left(t_n - \epsilon_2 k_\uparrow^{(n)} \right) \delta_{14}^{(\epsilon_2 - 1)k_\uparrow^{(n)}} \\ &\leq n\delta_{12}^{(1-\epsilon_1)k_\downarrow^{(n)}} + n\delta_{14}^{(\epsilon_2 - 1)k_\uparrow^{(n)}} + \sum_{k=\epsilon_1 k_\downarrow^{(n)}}^{\epsilon_2 k_\uparrow^{(n)}} \pi_k^{(n)} \\ &= \delta_{12}^{\delta_{15} \frac{m}{c_n} - \log_{\delta_{12}} n} + \delta_{14}^{\delta_{17} \frac{m}{c_n} - \log_{\delta_{14}} n} + \sum_{k=\epsilon_1 k_\downarrow^{(n)}}^{\epsilon_2 k_\uparrow^{(n)}} \pi_k^{(n)}, \end{aligned}$$

for suitably chosen positive constants δ_{15}, and δ_{17}. Thus, the theorem follows by noting that $\epsilon_1, \epsilon_2, \delta_{11}$, and δ_{13} can be chosen arbitrarily close to 1. □

We now use (15) to provide a slightly simpler bound on $\tau_k^{(n)}$.

Lemma 4. *For large enough n, we get,*

$$\tau_k^{(n)} \leq \frac{(t_n - k + 1)\theta_n}{h^{(n)}(k) - k\theta_n} \left(\frac{1}{k} \pi_{k-1}^{(n)} + \frac{k-1}{k} \tau_{k-1}^{(n)} \right),$$

for each $k = 1, \ldots, t_n$.

Proof. Using (14) in (15), we get

$$\begin{aligned} \tau_k^{(n)} &= \frac{\left(\begin{aligned} \frac{(t_n - k + 1)\theta_n}{h^{(n)}(k) - k\theta_n} \pi_{k-1}^{(n)} + \frac{t_n - k + 1}{k} \pi_{k-1}^{(n)} \\ + \frac{(t_n - k + 1)(k-1)}{k} \theta_n \tau_{k-1}^{(n)} \end{aligned} \right)}{h^{(n)}(k) - k\theta_n} \\ &= \frac{(t_n - k + 1)\theta_n}{h^{(n)}(k) - k\theta_n} \Bigg(\left(\frac{1}{h^{(n)}(k) - k\theta_n} + \frac{1}{k\theta_n} \right) \pi_{k-1}^{(n)} \\ &\qquad + \frac{k-1}{k} \tau_{k-1}^{(n)} \Bigg). \end{aligned}$$

Now, we have the lemma if we show that for large enough n, we have $\left(\frac{1}{h^{(n)}(k) - k\theta_n} + \frac{1}{k\theta_n} \right) \leq \frac{1}{k}$ for each $k = 1, \ldots, t_n$. This can be shown as follows.

One can show that Lemma 1 holds even when $p_k^{(n)} = 1 - e^{-\frac{kc_n}{m}}$. Using $g_n = \frac{\theta_n}{\gamma\xi' b}$, we get $h^{(n)}(k) = \xi' bnp_k^{(n)} \geq \frac{\delta_{20}}{\gamma} k\theta_n$ for large enough n and some constant δ_{20} such that $\gamma < \delta_{20} < 1$. Thus, $(h^{(n)}(k) - k\theta_n) \geq (\frac{\delta_{20}}{\gamma} - 1)k\theta_n$. For large enough n, $(\frac{\delta_{20}}{\gamma} - 1)\theta_n \geq 2$, and thus, $(h^{(n)}(k) - k\theta_n) \geq 2k$. Similarly, for large enough n, $k\theta_n \geq 2k$. Hence the lemma. □

[1] If $\pi^{(n)}(\mathbf{x})$ stationary distribution of the queue length process for the n^{th} system, then $\pi_k^{(n)}$ has the following interpretation: $\pi_k^{(n)} = \sum_{\mathbf{x}:|A_\mathbf{x}|=k} \pi^{(n)}(\mathbf{x})$ for $k = 1, \ldots, t_n$.

Following lemma provides an even simpler bound on $\tau_k^{(n)}$.

Lemma 5. *For large enough n, we get,*

$$\tau_k^{(n)} \leq \pi_k^{(n)}$$

for each $k \in \{1, \ldots, t_n\}$.

Proof. Fix n large enough such that the bound in Lemma 4 holds. We prove the result using induction on k. Consider the base case of $k = 1$. From Lemma 4 and (14) we have

$$\tau_1^{(n)} \leq \frac{(t_n - 1 + 1)\theta_n}{h^{(n)}(1) - \theta_n}\pi_0^{(n)} = \pi_1^{(n)}.$$

Now, let us assume that the lemma holds for $k = k' - 1$, i.e., $\tau_{k'-1}^{(n)} \leq \pi_{k'-1}^{(n)}$. Using this, we show below that the lemma holds for $k = k'$ as well.

From Lemma 4 and induction hypothesis we have

$$\tau_{k'}^{(n)} \leq \frac{(t_n - k' + 1)\theta_n}{h^{(n)}(k') - k'\theta_n}\left(\frac{1}{k'}\pi_{k'-1}^{(n)} + \frac{k'-1}{k'}\pi_{k'-1}^{(n)}\right) = \pi_{k'}^{(n)},$$

where the last equality follows from (14). Hence, the lemma. □

Using above lemma and (13), we get

$$E[D^{(n)}] \leq \nu \sum_{k=1}^{t_n} \frac{k}{t_n}\pi_k^{(n)}.$$

Or equivalently,

$$\frac{1}{\nu}E[D^{(n)}] = \sum_{k=1}^{\epsilon b \log(\frac{1}{1-\gamma})\frac{n}{c_n}} \frac{k}{t_n}\pi_k^{(n)} + \sum_{k=\epsilon' b \log(\frac{1}{1-\gamma})\frac{n}{c_n}+1}^{t_n} \frac{k}{t_n}\pi_k^{(n)}.$$

We now use Theorem 6 to prove the main result. From Theorem 6, we have

$$\begin{aligned}\frac{1}{\nu}E[D^{(n)}] &\leq \sum_{k=1}^{\epsilon b \log(\frac{1}{1-\gamma})\frac{n}{c_n}} \frac{k}{t_n}\pi_k^{(n)} + \delta'^{\frac{m}{c_n}} \\ &\leq \epsilon b \log\left(\frac{1}{1-\gamma}\right)\frac{n}{c_n t_n} \sum_{k=1}^{\epsilon b \log(\frac{1}{1-\gamma})\frac{n}{c_n}} \pi_k^{(n)} + \delta'^{\frac{m}{c_n}} \\ &\leq \epsilon \log\left(\frac{1}{1-\gamma}\right)\frac{\theta_n}{\gamma\xi' c_n} + \delta'^{\frac{m}{c_n}},\end{aligned}$$

where in last inequality we used definition of t_n. The first part of the theorem thus follows from definition of ξ', and the fact that ϵ and δ' where chosen arbitrarily.

Further, from Corollary 3, upon randomly placing c_n copies of each file, the associated random capacity region contains $\tilde{\mathcal{C}}^{(n)}$ with high probability. Hence, the second part follows as well. □

7. CONCLUSIONS

Our main conclusions address both practical and theoretical aspects associated with such systems. CDN systems address potential high demands by maintaining multiple copies of content – server diversity. At the same time there has been increasing interest in adopting multipath transport protocols to improve reliability and throughput. Our results show that infrastructure combining multipath transport with server diversity scales well even when subject to substantial variations in per file demands. Specifically, if CDN capacity (servers) scale with total load, and memory per server is proportional to worst case per file demand variability then average download delays will scale, i.e., are asymptotically negligible. This suggests a scalable approach towards addressing the delivery of popular content without requiring complex caching strategies.

On the theoretical side we have established: (1) basic new results linking fairness in resource allocation to delays and (2) the asymptotic symmetry of randomly configured large-scale systems with heterogenous components. Together these results suggest large systems might eventually be robust to heterogeneity and fairness criterion.

8. REFERENCES

[1] T. Bonald. Throughput performance in networks with linear capacity contraints. In *Proceedings of CISS*, pages 644 –649, 2006.

[2] T. Bonald, L. Massoulié, A. Proutière, and J. Virtamo. A queueing analysis of max-min fairness, proportional fairness and balanced fairness. *Queueing Systems*, 53:65–84, 2006.

[3] T. Bonald and A. Proutière. Insensitive bandwidth sharing in data networks. *Queueing Systems*, 44:69–100, 2003.

[4] T. Bonald and A. Proutière. On stochastic bounds for monotonic processor sharing networks. *Queueing Systems*, 47:81–106, 2004.

[5] T. Bonald, A. Proutière, J. Roberts, and J. Virtamo. Computational aspects of balanced fairness. In *Proceedings of ITC*, 2003.

[6] T. Bonald and J. Virtamo. Calculating the flow level performance of balanced fairness in tree networks. *Perform. Eval.*, 58(1):1–14, Oct. 2004.

[7] G. de Veciana, T.-J. Lee, and T. Konstantopoulos. Stability and performance analysis of networks supporting elastic services. *IEEE/ACM Transactions on Networking*, 9(1):2–14, Feb. 2001.

[8] D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2):99–124, 1998.

[9] J. Edmonds. Submodular functions, matroids, and certain polyhedra. In *Proceedings of Calgary International Conference on Combinatorial Structures and Applications*, pages 69–87, 1969.

[10] B. Frank, I. Poese, G. Smaragdakis, A. Feldmann, B. M. Maggs, S. Uhlig, V. Aggarwal, and F. Schneider. Collaboration opportunities for content delivery and network infrastructures. In H. Haddadi and O. Bonaventure, editors, *Recent Advances in Networking*, pages 305–377. ACM SIGCOMM ebook, 2013.

[11] V. Joseph and G. de Veciana. Stochastic networks with multipath flow control: Impact of resource pools on flow-level performance and network congestion. In *Proceedings of the ACM Sigmetrics*, pages 61–72, 2011.

[12] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *The Journal of the Operational Research Society*, 49(3):237–252, 1998.

[13] T. Lan, D. Kao, M. Chiang, and A. Sabharwal. An axiomatic theory of fairness in network resource

allocation. In *Proceedings of IEEE Infocom*, pages 1–9, March 2010.

[14] M. Leconte, M. Lelarge, and L. Massoulié. Adaptive replication in distributed content delivery networks. *arXiv preprint arXiv:1401.1770*, 2014.

[15] X. Lin and N. Shroff. Utility maximization for communication networks with multipath routing. *IEEE Transactions on Automatic Control*, 51(5):766 – 781, May 2006.

[16] A. W. Marshall, I. Olkin, and B. C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

[17] L. Massoulié and J. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15(1-2):185–201, 2000.

[18] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

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

[20] S. Moharir, J. Ghaderi, S. Sanghavi, and S. Shakkottai. Serving content with unknown demand: The high-dimensional regime. In *Proceedings of ACM Sigmetrics*, pages 435–447, 2014.

[21] G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*, volume 18. Wiley, 1988.

[22] V. Shah and G. de Veciana. Performance evaluation and asymptotics for content delivery networks. In *IEEE Infocom*, pages 2607–2615, 2014.

[23] J. N. Tsitsiklis and K. Xu. Queueing system topologies with limited flexibility. In *Proceedings of ACM Sigmetrics*, pages 167–178, 2013.

[24] E. Yeh. *Multiaccess and fading in communication networks*. PhD thesis, Massachusetts Institute of Technology, 2001.

APPENDIX

A.1 Proof of Theorem 1

Consider the following lemma regarding relative greediness of αF and BF.

Lemma 6. *Consider states* $\mathbf{x}$ *and* $\mathbf{y}$ *such that* $\mathbf{x} \prec_w \mathbf{y}$*. For each* k *such that* $\sum_{l=1}^{k} x_{[l]} = \sum_{l=1}^{k} y_{[l]}$*, we have* $\sum_{l=1}^{k} r^{\alpha}_{(l)}(\mathbf{x}) \geq \sum_{l=1}^{k} r^{B}_{(l)}(\mathbf{y})$.

Roughly, it asserts that if state $\mathbf{x}$ is same or more balanced than state $\mathbf{y}$, then the sum rate assigned to larger queues by αF to state $\mathbf{x}$ is greater than that by BF to state $\mathbf{y}$. Proof of this lemma is summarized in A.2. Below, we provide a detailed coupling argument showing stochastic weak-majorization using this lemma.

Coupling Argument: Without loss of generality, assume $\nu = 1$. Suppose $\mathbf{X}^{\alpha}(0) \prec_w \mathbf{X}^{B}(0)$. Below, we couple the arrivals and departures of processes $(\mathbf{X}^{\alpha}(t) : t \geq 0)$ and $(\mathbf{X}^{B}(t) : t \geq 0)$ such that their marginal distributions remain intact and $\mathbf{X}^{\alpha}(t) \prec_w \mathbf{X}^{B}(t)$ almost surely for each $t \geq 0$.

Let Π_a be a Poisson point process with rate $\sum_{i \in F} \lambda_i$, and let Π_d be Poisson point process with rate $\mu(F)$. The points in these processes are the times of 'potential events' in $(\mathbf{X}^{B}(t) : t \geq 0)$ and $(\mathbf{X}^{\alpha}(t) : t \geq 0)$. We use Π_a to couple arrivals and Π_d to couple departures. For each time t' when a potential event occurs, let $\epsilon_{t'}$ be a small enough number such that no potential event occurred in the time interval of $[t' - \epsilon_{t'}, t')$.

Coupling of arrivals: For each point t' in Π_a, do the following: Choose a random variable $Z_{t'}$ independently and uniformly from $\{1, \ldots, n\}$. Let an arrival occur in $(\mathbf{X}^{\alpha}(t) : t \geq 0)$ at time t' in the $Z_{t'}^{\text{th}}$ largest queue of $\mathbf{X}^{\alpha}(t' - \epsilon_{t'})$. Ties are broken uniformly at random. Similarly, let an arrival occur in $(\mathbf{X}^{\alpha}(t) : t \geq 0)$ at time t' in the $Z_{t'}^{\text{th}}$ largest queue of $\mathbf{X}^{\alpha}(t' - \epsilon_{t'})$. Again, ties are broken uniformly at random.

Coupling of departures: For each point t' of increment in Π_d, do the following: Choose a random variable $Z_{t'}$ independently and uniformly from interval $(0, \mu(F)]$. For k such that

$$Z_{t'} \in \left(\sum_{l=1}^{k-1} r^{\alpha}_{(l)}(X^{\alpha}(t' - \epsilon_{t'})), \sum_{l=1}^{k} r^{\alpha}_{(l)}(X^{\alpha}(t' - \epsilon_{t'})) \right],$$

let a departure occur in $(\mathbf{X}^{\alpha}(t) : t \geq 0)$ at time t' in the k^{th} largest queue of $\mathbf{X}^{\alpha}(t' - \epsilon_{t'})$, with ties broken uniformly and independently at random.

Similarly, for k such that

$$Z_{t'} \in \left(\sum_{l=1}^{k-1} r^{B}_{(l)}(\mathbf{X}^{B}(t' - \epsilon_{t'})), \sum_{l=1}^{k} r^{B}_{(l)}(\mathbf{X}^{B}(t' - \epsilon_{t'})) \right],$$

let a departure occur in $(\mathbf{X}^{B}(t) : t \geq 0)$ at time t' in the k^{th} largest queue of $\mathbf{X}^{B}(t' - \epsilon_{t'})$, with ties broken uniformly and independently at random. Note that in both cases it is possible that no such k exists since some classes may not be active and the total service rate may be less than $\mu(F)$. In that case, no departure occurs.

It can be checked that the marginal distributions of $(\mathbf{X}^{\alpha}(t) : t \geq 0)$ and $(\mathbf{X}^{B}(t) : t \geq 0)$ remain intact. We now show that $\mathbf{X}^{\alpha}(t) \prec_w \mathbf{X}^{B}(t)$ almost surely for each t.

It is easy to check that if an arrival occurred at time t' and if $\mathbf{X}^{\alpha}(t) \prec_w \mathbf{X}^{B}(t)$ for each $t < t'$, then $\mathbf{X}^{\alpha}(t') \prec_w \mathbf{X}^{B}(t')$ as well. We now show that the same holds for points of Π_d as well.

Suppose a potential departure occurred at t', and $\mathbf{X}^{\alpha}(t) \prec_w \mathbf{X}^{B}(t)$ for each $t < t'$. We show below that $\sum_{l=1}^{k} X^{\alpha}_{[l]}(t') \leq \sum_{l=1}^{k} X^{B}_{[l]}(t')$ for each k. Here, we use Lemma 6. Following two cases arise.

Case 1: $\sum_{l=1}^{k} X^{\alpha}_{[l]}(t' - \epsilon_{t'}) < \sum_{l=1}^{k} X^{B}_{[l]}(t' - \epsilon_{t'})$. A maximum of one departure occurs at time t' in either processes. Thus we have $\sum_{l=1}^{k} X^{\alpha}_{[l]}(t') \leq \sum_{l=1}^{k} X^{B}_{[l]}(t')$.

Case 2: $\sum_{l=1}^{k} X^{\alpha}_{[l]}(t' - \epsilon_{t'}) = \sum_{l=1}^{k} X^{B}_{[l]}(t' - \epsilon_{t'})$. By using $\mathbf{X}^{\alpha}(t - \epsilon_{t'}) \prec_w \mathbf{X}^{B}(t - \epsilon_{t'})$ in Lemma 6 and from the definition of the coupling at time t', it can be shown that if a departure occurs from any of the k largest queues in $\mathbf{X}^{B}(t' - \epsilon_{t'})$, then it also occurs in one of the k largest queues in $\mathbf{X}^{\alpha}(t' - \epsilon_{t'})$. Thus, $\sum_{l=1}^{k} X^{\alpha}_{[l]}(t') \leq \sum_{l=1}^{k} X^{B}_{[l]}(t')$.

Hence the result. □

A.2 Relative greediness and other rate allocation properties

Below, we outline a proof of Lemma 6 which asserts that αF is more greedy than BF. Along the way, we develop several other properties of the rate allocation policies.

Proof of Lemma 6 stems on the following two fundamental properties of per-job rate assignment for αF and BF.

1.) *αF gives the most balanced per-job rate allocation:* The property follows from the fact that αF is equivalent to max-min fair rate allocation, see Proposition 2. Formally, this property implies the following:

Lemma 7. *Let $\mathbf{b}^\alpha$ represent a vector of rates assigned to a set of flows under αF rate allocation. Let $\tilde{\mathbf{b}}$ be the rates assigned to the same set of flows under any other feasible rate allocation. Then, $\mathbf{b}^\alpha \prec^w \tilde{\mathbf{b}}$, i.e., weak majorized from above.*

2.) *In αF and BF, the longest queues have smaller per-job rates:* For αF, this property again follows from the fact that it is equivalent to max-min fair, and that the capacity region is convex and symmetric. For BF, the proof for this property is technical and we omit its discussion here for brevity. Formally, this property implies the following:

Lemma 8. *αF and BF rate allocations satisfy the following property for any state $\mathbf{x}$: if $x_i > x_j$ then $\frac{r_i(\mathbf{x})}{x_i} \le \frac{r_j(\mathbf{x})}{x_j}$.*

Now, let us study what the above properties imply for per-class rate allocation. Consider a state $\mathbf{x}$. Lemma 8 above implies that the most disadvantaged jobs are the ones which belong to longest queues for both, BF and αF. This, along with Lemma 8, implies that αF provides larger rate to longest queues. Thus we get the following property.

3.) *αF gives larger rate to the longest queues as compared to BF:* Formally, this property implies the following:

Lemma 9. *For any state $\mathbf{x}$, $\sum_{l=1}^k r^\alpha_{(l)}(\mathbf{x}) \ge \sum_{l=1}^k r^B_{(l)}(\mathbf{x})$ for each $k \in \{1, 2, \ldots, n\}$.*

Now, we focus on αF and study how it allocates rates across classes for states $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \prec \mathbf{y}$. Intuitively, jobs in longer queues in state $\mathbf{y}$ are more constrained than those in $\mathbf{x}$. Again using the fact that αF is equivalent to max-min fair, the most constrained jobs in state $\mathbf{y}$ have smaller rate than those in state $\mathbf{x}$. By monotonicity of αF, this holds even when $\mathbf{x} \prec_w \mathbf{y}$. When translated to per-class rate allocation in states $\mathbf{x}$ and $\mathbf{y}$, this argument leads us to the following property:

4.) *αF gives larger rate to longer queues in more balanced states:* Formally, this property implies the following:[2]

Lemma 10. *Consider states $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \prec_w \mathbf{y}$. For each k such that $\sum_{l=1}^k x_{[l]} = \sum_{l=1}^k y_{[l]}$, we have $\sum_{l=1}^k r^\alpha_{(l)}(\mathbf{x}) \ge \sum_{l=1}^k r^\alpha_{(l)}(\mathbf{y})$.*

Finally, we are ready to study relative greediness of αF and BF.

5.) *αF is more greedy than BF:* We now prove Lemma 6. Consider states $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \prec_w \mathbf{y}$. From Lemma 10 we have $\sum_{l=1}^k r^\alpha_{(l)}(\mathbf{x}) \ge \sum_{l=1}^k r^\alpha_{(l)}(\mathbf{y})$, and from Lemma 9 we have $\sum_{l=1}^k r^\alpha_{(l)}(\mathbf{y}) \ge \sum_{l=1}^k r^B_{(l)}(\mathbf{y})$. Hence, Lemma 6 holds.

A.3 Technical Lemmas for proof of Theorem 4

Lemma 1. Let a sequence $(g_n : n \in \mathbb{N})$ be such that $g_n = o(c_n)$. Let δ_1 be a positive constant independent of n such that $\delta_1 < 1$. Then, for large enough n, we have

$$p_k^{(n)} \ge \frac{\delta_1 g_n}{n} k \quad \forall k \in \left\{0, 1, \ldots, \left\lfloor \frac{n}{g_n} \right\rfloor \right\}.$$

Proof. Consider a sequence of functions $\left(f^{(n)}(.)\right)_{n\ge1}$ where for each n, $f^{(n)}(t) = 1 - (1 - 1/(bn))^{tc_n}$ for each $t \in \mathbb{R}_+$. Then,

$$f^{(n)}\left(n/g_n\right) = 1 - (1 - 1/(bn))^{\frac{nc_n}{g_n}} \overset{n\to\infty}{\longrightarrow} 1.$$

Thus, there exists an integer n' such that $f^{(n)}\left(n/g_n\right) \ge \delta_1$ for all $n \ge n'$. Also, $f^{(n)}(0) = 0$ for each n. Using concavity of $f^{(n)}(.)$, for each $n \ge n'$ we have

$$f^{(n)}\left(t\right) \ge \frac{f^{(n)}\left(n/g_n\right)}{\left(n/g_n\right)} t, \quad \forall t \text{ s.t. } 0 \le t \le n/g_n.$$

Hence, the lemma. □

Lemma 2. There exists a positive constant δ such that $H\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right) \ge -\delta + \epsilon\frac{kc_n}{m}$.

Proof. From definition,

$$\begin{aligned} H\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right) &= p_k^{(n)}(1-\epsilon)\log(1-\epsilon) \\ &\quad + (1 - p_k^{(n)}(1-\epsilon))\log\left(\frac{1 - p_k^{(n)}(1-\epsilon)}{1 - p_k^{(n)}}\right) \end{aligned}$$

Here, the term $p_k^{(n)}(1-\epsilon)\log(1-\epsilon)$, while negative, is greater than $(1-\epsilon)\log(1-\epsilon)$, a constant. Similarly, the term $(1 - p_k^{(n)}(1-\epsilon))\log\left(1 - p_k^{(n)}(1-\epsilon)\right)$ is negative, but can be upper-bounded by a constant as follows:

$$\begin{aligned} (1-p_k^{(n)}(1-\epsilon))\log\left(1 - p_k^{(n)}(1-\epsilon)\right) &\ge \log\left(1 - p_k^{(n)}(1-\epsilon)\right) \\ &\ge \log(1-(1-\epsilon)) = \log\epsilon \end{aligned}$$

Thus, we have

$$\begin{aligned} H\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right) &\ge -\delta + (1 - p_k^{(n)}(1-\epsilon))\log\left(\frac{1}{1-p_k^{(n)}}\right) \\ &\ge -\delta + (1-(1-\epsilon))\log\left(\frac{1}{1-p_k^{(n)}}\right) \\ &= -\delta + \epsilon\log\left(\frac{1}{1-p_k^{(n)}}\right) \\ &\ge -\delta + \epsilon\frac{kc_n}{m}, \end{aligned}$$

where in the last inequality we used the fact that $1 - p_k^{(n)} \le e^{-\frac{kc_n}{m}}$. □

[2] This property is also used to prove Majorization bound described in Section 4.3. Also, one can check that this property is satisfied by Greedy rate allocation as well.