Queueing Syst (2016) 83:361–397
DOI 10.1007/s11134-016-9491-0

Impact of fairness and heterogeneity on delays in large-scale centralized content delivery systems

Virag Shah[1] · **Gustavo de Veciana**[1]

Received: 10 September 2015 / Revised: 18 April 2016 / Published online: 7 July 2016

Abstract We consider multiclass 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 a growing number of service classes n (for example, files), $m = \lceil bn \rceil$ heterogenous servers with total service rate ξm, and polymatroid capacity resulting from a random bipartite graph $\mathcal{G}^{(n)}$ modeling service availability (for example, placement of files across servers). This models, for example, content delivery systems supporting pooling of server resources, i.e., parallel servicing of a download request from multiple servers. For an appropriate asymptotic regime, we show that the system's capacity region is uniformly close to a symmetric polymatroid—heterogeneity in servers' capacity and file placement disappears. Combining our comparison results and the asymptotic 'symmetry' in large systems, we show that large randomly configured systems with a logarithmic number of file copies are robust to substantial load and server heterogeneities for a class of fairness criteria. If each class can be served by $c_n = \omega(\log n)$ servers, the load per class does not exceed $\theta_n = o\left(\min(\frac{n}{\log n}, c_n)\right)$, mean service requirement of a job is ν, and average server utilization is bounded by $\gamma < 1$, then for each constant $\delta > 1$, the conditional expectation of delay of a typical job with respect to the σ-algebra generated by $\mathcal{G}^{(n)}$ satisfies the following:

✉ Virag Shah
virag@utexas.edu

Gustavo de Veciana
gustavo@ece.utexas.edu

[1] Department of ECE, The University of Texas at Austin, Austin, TX 78712, USA

$$\lim_{n\to\infty} P\left(E[D^{(n)}|\mathcal{G}^{(n)}] \le \delta \frac{\nu}{\xi c_n} \frac{1}{\gamma} \log\left(\frac{1}{1-\gamma}\right)\right) = 1.$$

Keywords Delays · Fairness · Heterogeneity · Robustness · Queueing · Asymptotic symmetry · Content delivery systems

Mathematics Subject Classification Primary 68M20 · Secondary 60K30

1 Introduction

In many shared network systems, service rate is allocated to ongoing jobs based on a fairness criterion, for example, α-fair (αF) (including max-min and proportional fair) as well as balanced fair (BF), and other greedy criteria [26]. When the network loads are stochastic a key open question is how the choice of fairness and network design will impact user perceived performance, for example, 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 centralized content delivery systems where a collection of servers deliver a proportionally large number of files. There has been substantial recent interest in understanding basic design questions for these systems, including, for example, [10,14,20,24] 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?

We consider a centralized system with several collocated servers. The replication of files across servers is kept static. We allow resource pooling, i.e., 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 capacities, as well as to the fairness criterion driving resource allocation.

Some elements of content delivery infrastructure may see less pronounced heterogeneity in demands, for example, a centralized back end used to deliver files that are not available at distributed sites/caches. For such a system, with sufficient redundancy, enabling resource pooling for individual download requests could achieve scalable and robust performance.

1.1 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 Sects. 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 allocation, 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 Sect. 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 the system's service capacity region is polymatroid. We show that for appropriately scaled large systems where the edge set is chosen at random (random file placement) the capacity region is uniformly close to a *symmetric* polymatroid.

(c.) *Performance robustness of large systems* Combining these two results, in Sect. 6 we provide a simple performance bound for large-scale content delivery systems. More specifically, the performance under α-fair rate allocation for a large system is upper-bounded by that under a system with smaller, symmetric, and approximate capacity region. The bound exhibits performance robustness in large systems with respect to variations in total system load, heterogeneity in load across the classes, and heterogeneity in server capacities, for α-fair based resource allocation.

We have deferred some technical results to the appendix. Section 7 concludes the paper.

1.2 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 an $M/G/1$ 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. Recently, [23] 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 [23,26]. For example, the work in [26] 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 Sect. 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 a particular focus on addressing user-performance in a large-scale content delivery system which leverages server diversity, i.e., availability of multiple copies of a file to serve a download request.

From a content delivery perspective, the two works closest to this paper are [24] and [23]. Both adopt a natural model for a content delivery 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 [24] consider a service model where each request can be served by a single server—recall we consider systems allowing parallel download of a file from multiple servers. Resource pooling in our service model leads to a significantly improved mean delay bound. For example, upon availability of c_n servers for each class, our delays scale as $O\left(\frac{1}{c_n}\right)$. Also in our work we are able to address the role of fairness criteria and robustness to heterogeneity in server capacities.

Our service model via resource pooling is same as in [23]. However, our work here is different in several respects. Firstly, in [23] the focus is on mean delays under balanced fair resource allocation, whereas here we directly study the impact of fairness criteria on users delays. Secondly, the system considered was by design symmetric, whereas here 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 $\rho_i = \lambda_i \nu$ is the load associated with class i. For example, if the service requirement of a job is measured in bits then the load for each class is measured in bits per second.

Jobs arrive to the system at total rate $\sum_{i \in F} \lambda_i$. Let u_k denote the job corresponding to the kth arrival after time $t = 0$. 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 $\mathbf{q}(t) = (q_i(t) : i \in F)$. For each $A \subset F$, let $q_A(t) = \cup_{i \in A} q_i(t)$, i.e., the set of all active jobs whose class is in A. 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 $\mathbf{b}(t) = (b_v(t) : v \in q_F(t))$ 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 for example, [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; see Fig. 1.

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

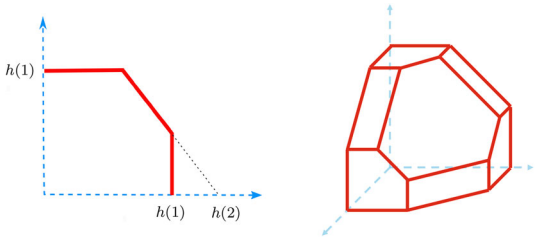

Fig. 1 Symmetric polymatroids in two and three dimensions

symmetric multiaccess channels [26]. Further, we will see in Sect. 5 that certain types of large content delivery systems 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

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

and will see that $\hat{\mathcal{C}}$ 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 $z_{[1]} \geq z_{[2]} \geq \ldots \geq z_{[|I|]}$ 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 ith 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 by $\mathbf{z} \prec \mathbf{z}'$.

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 by $\mathbf{z} \prec_w \mathbf{z}'$.

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 by $\mathbf{z} \prec^w \mathbf{z}'$.

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 by $\tilde{\mathbf{Z}} \leq^{st} \tilde{\mathbf{Z}}'$.

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

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 kth 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 $\lim_{n\to\infty} \frac{g_n}{f_n} = 0$.

We say an event A happens with high probability (denoted as w.h.p.) if $P(A)$ is $1 - o(1)$.

Several parts of the notation above are borrowed from [16,26] and [22].

3 Rate allocation policies: background

There are several possible rate allocation policies, each resulting in potentially different user-perceived delays. Below, we introduce three different policies studied in the 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

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

Equivalently, the sum rate assigned to the k largest queues, namely $\sum_{l=1}^{k} r^G_{(l)}(\mathbf{x})$, is equal to $\mu(\{[1], [2], \ldots, [k]\})$. 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 [26] where the following result was shown.

Proposition 1 ([26]) *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 the greedy and an arbitrary feasible rate allocation, respectively. If* $\mathbf{X}^G(0) \prec_w^{st} \tilde{\mathbf{X}}(0)$ *then* $\mathbf{X}^G(t) \prec_w^{st} \tilde{\mathbf{X}}(t)$ *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, for example, 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 ([23]) *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 is amenable to distributed implementation [12,15] and satisfies natural axioms for fairness [13]. Unfortunately, little is known regarding their performance under stochastic arrivals. 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 mean delay analysis. Formally, the 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 $\sum_{i\in A} r_i(\mathbf{x}) \le \mu(A)$ 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 $(\mathbf{X}^B(t) : t \ge 0)$ 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 the stationary distribution makes balanced fairness amenable to time-averaged performance analysis, a property which we will use extensively in the sequel. While, in general, BF may result in wasteful resource allocation, for example, BF is not Pareto efficient for certain triangle networks studied in [3], for polymatroid capacity regions, BF has been shown to be Pareto efficient:

Proposition 3 ([23]) *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 $\mathbf{x}$.*

Using Pareto optimality, a recursive expression for mean delay was provided in [23] for arbitrary polymatroid capacity region and load. The expression can be significantly simplified under symmetry, as also shown below. First, let

$$\pi_k = \sum_{\mathbf{x}:|A_{\mathbf{x}}|=k} \pi(\mathbf{x}),$$

i.e., π_k is the stationary probability that there are k active classes in the system. Then, under symmetry, the following expression was shown to hold for π_k in [23]. We provide a (slightly different) proof below for the sake of completion.

Proposition 4 ([23]) *For a system with symmetric polymatroid capacity region, with load $\rho_i = \rho$ for each class $i \in F$, and with balanced fair rate allocation, we have*

$$\pi_0 = \frac{1}{1 + \sum_{k=1}^{n} \prod_{l=1}^{k} \frac{(n-l+1)\rho}{h(l)-l\rho}}, \tag{5}$$

and for $k = 1, \ldots, n$ we have

$$\pi_k = \frac{(n-k+1)\rho}{h(k)-k\rho}\pi_{k-1}. \tag{6}$$

Equivalently, for $k = 1, \ldots, n$, we have

$$\pi_k = \pi_0 \prod_{l=1}^{k} \frac{(n-l+1)\rho}{h(l)-l\rho}. \tag{7}$$

Proof It is enough to show that for each $k \geq 1$ we have

$$\pi_k h(k) = (n-k+1)\rho\pi_{k-1} + k\rho\pi_k. \tag{8}$$

There are two ways to argue that the above expression holds: (1) using PASTA and time reversibility, and (2) using the stationary distribution expression via the balance function. We summarize both approaches below.

Note that $\pi_k h(k) = \sum_{|\mathbf{x}|:|A_\mathbf{x}|=k} \pi(\mathbf{x})h(k)$ is the total rate of departures from states with k active classes. In reverse time these departures correspond either to (1) the arrivals to the system which see $k-1$ active classes and cause an increase in the number of active classes, or to (2) arrivals which see k active classes and do not cause an increase the number of active classes. Since arrivals in the reverse time form a Poisson process, PASTA holds, and the rates of above transitions is equal to $(n-k+1)\rho\pi_{k-1}$ and $k\rho\pi_k$, respectively. Thus, we get (8).

Alternatively, from the definition and Proposition 3, we have

$$\begin{aligned}\pi_k = \pi_0 \sum_{\mathbf{x}:|A_\mathbf{x}|=k} \Phi(\mathbf{x})\rho^{|\mathbf{x}|} &= \pi_0 \sum_{\mathbf{x}:|A_\mathbf{x}|=k} \frac{\sum_{i\in A_\mathbf{x}} \Phi(\mathbf{x}-\mathbf{e}_i)}{\mu(A_\mathbf{x})}\rho^{|\mathbf{x}|} \\ &= \frac{\pi_0\rho}{h(k)} \sum_{\mathbf{x}:|A_\mathbf{x}|=k} \sum_{i\in A_\mathbf{x}} \Phi(\mathbf{x}-\mathbf{e}_i)\rho^{|\mathbf{x}-\mathbf{e}_i|}.\end{aligned}$$

This can be shown to simplify to the following:

$$\pi_k = \frac{\pi_0\rho}{h(k)}(n-k+1) \sum_{\mathbf{x}:|A_\mathbf{x}|=k-1} \Phi(\mathbf{x})\rho^{|\mathbf{x}|} + \frac{\pi_0\rho}{h(k)}k \sum_{\mathbf{x}:|A_\mathbf{x}|=k} \Phi(\mathbf{x})\rho^{|\mathbf{x}|}.$$

Upon simplification, we get (8). □

Now, let $\beta_k = E\left[|\mathbf{X}|\big||A_\mathbf{X}| = k\right]$, i.e., $\beta_k = \frac{\sum_{\mathbf{x}:|A_\mathbf{x}|=k} |\mathbf{x}|\pi(\mathbf{x})}{\pi_k}$. There exists a surprisingly simple expression for β_k using which an explicit expression for the mean delay can be obtained, as given by the following theorem.

Theorem 1 *Consider a system with symmetric polymatroid capacity region, and with load $\rho_i = \rho$ for each class $i \in F$. Under balanced fair rate allocation, let $\beta_k = E\left[|\mathbf{X}|\big||A_\mathbf{X}| = k\right]$. Then, for $k = 1, \ldots, n$ we have*

$$\beta_k = \sum_{l=1}^{k} \frac{h(l)}{h(l) - l\rho}. \tag{9}$$

Further, if the arrival rate for each class is equal to λ then the mean delay for jobs under balanced fairness is given by

$$E[D^B] = \frac{1}{\lambda n} \sum_{k=1}^{n} \beta_k \pi_k, \tag{10}$$

where π_k can be computed using (5) and (7).

Proof We provide a proof for the expression for β_k. The expression for the mean delay then follows from Little's law. From the definition and Proposition 3 we have

$$(\beta_k - 1)\pi_k = \sum_{\mathbf{x}:|A_\mathbf{x}|=k} (|\mathbf{x}| - 1)\Phi(\mathbf{x})\rho^{|\mathbf{x}|} = \sum_{\mathbf{x}:|A_\mathbf{x}|=k} (|\mathbf{x}| - 1)\frac{\sum_{i\in A_\mathbf{x}} \Phi(\mathbf{x} - \mathbf{e}_i)}{\mu(A_\mathbf{x})}\rho^{|\mathbf{x}|}$$
$$= \frac{\rho}{h(k)} \sum_{\mathbf{x}:|A_\mathbf{x}|=k} (|\mathbf{x}|-1) \sum_{i\in A_\mathbf{x}} \Phi(\mathbf{x}-\mathbf{e}_i)\rho^{|\mathbf{x}-\mathbf{e}_i|}.$$

This can be shown to simplify to the following:

$$(\beta_k - 1)\pi_k = \frac{\rho}{h(k)}(n - k + 1) \sum_{\mathbf{x}:|A_\mathbf{x}|=k-1} |\mathbf{x}|\Phi(\mathbf{x})\rho^{|\mathbf{x}|} + \frac{\rho}{h(k)}k \sum_{\mathbf{x}:|A_\mathbf{x}|=k} |\mathbf{x}|\Phi(\mathbf{x})\rho^{|\mathbf{x}|},$$

which in turn gives

$$\beta_k - 1 = \frac{(n - k + 1)\rho\pi_{k-1}}{\pi_k h(k)}\beta_{k-1} + \frac{k\rho\pi_k}{\pi_k h(k)}\beta_k. \tag{11}$$

Upon further simplification, one obtains

$$\beta_k = \frac{h(k)}{h(k) - k\rho} + \frac{(n - k + 1)\rho}{h(k) - k\rho}\frac{\pi_{k-1}}{\pi_k}\beta_{k-1} = \frac{h(k)}{h(k) - k\rho} + \beta_{k-1},$$

where the last equality follows from (6). From this (9) follows.

Paralleling the discussion for expression (8), (11) can also be argued directly using PASTA and time reversibility. In this case, $\beta_k - 1$ can be interpreted as the mean number of jobs a departure leaves behind it when the system has k active classes. Recalling the argument for (8), in reverse time, $\frac{(n-k+1)\rho\pi_{k-1}}{\pi_k h(k)}$ is the fraction of arrivals which result in k active classes by increasing the number of active classes by 1. Note that the rate of such arrivals does not depend on the precise state of the system. Thus, using a 'ratio of rates' argument, see [25], the mean number of customers seen by these arrivals is β_{k-1}. Similarly, one can argue that the remaining fraction $\frac{k\rho\pi_k}{\pi_k h(k)}$ of arrivals which see k active classes see a mean number of jobs as β_k. Thus, the expression (11) follows. □

In the sequel we use several other properties of balanced fairness and also of other rate allocation policies, some of which are provided in the Appendix (Relative greediness and other rate allocation properties).

4 Performance bounds

Recall that for each rate allocation policy considered in Sect. 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 resource pooling.

Below we develop upper bounds for mean delay 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:* Consider loads $\boldsymbol{\rho}, \boldsymbol{\rho}' \in \hat{\mathcal{C}}$ such that $\boldsymbol{\rho} \leq \boldsymbol{\rho}'$ and $\boldsymbol{\rho}'$ 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 $\boldsymbol{\rho}'$, even if $\boldsymbol{\rho}$ is heterogeneous.

(iii) *Majorization bound:* 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 $\boldsymbol{\rho}'$.

Throughout this section we will assume that the mean service requirements for jobs ν is same for each system. The bound for homogeneous loads and the majorization bound are provided below for both αF and greedy, whereas the dominance bound is provided for αF. Next we will also develop a lower bound for mean delay for each rate allocation policy under arbitrary loads.

Note that using the 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 upper 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 Theorem 1 is an upper bound for that of greedy. Below we show that this performance upper bound via BF also holds for αF rate allocation.

To this 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 a 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 Sect. 4.5.

Theorem 2 *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:*

$$E[D_{\boldsymbol{\rho}}^{\alpha}] \leq E[D_{\boldsymbol{\rho}}^{B}].$$

4.2 Dominance bound

Consider the following rate allocation property. 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 5 (*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.

From the definition of αF, one can check that αF rate allocation satisfies per-job rate monotonicity. This property was used in [4] to provide a comparison result for systems where the rate allocation in one system dominates that in another system for each state $\mathbf{x}$. In contrast, we provide below a comparison result for systems with the same rate allocation policy and capacity region, but with different loads. For such systems, we show that the larger loads result in worse delays if the rate allocation satisfies per-job rate monotonicity. For a proof of the theorem below, see Sect. 4.5.

Theorem 3 *Consider a system with symmetric polymatroid capacity region $\mathcal{C}$. Suppose that the rate allocation $\mathbf{r}(.)$ satisfies per-job rate monotonicity. Let $\boldsymbol{\rho}, \boldsymbol{\rho}' \in \hat{\mathcal{C}}$ (recall, $\hat{\mathcal{C}}$ is the stability region) be such that $\boldsymbol{\rho} \leq \boldsymbol{\rho}'$. 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. *For systems with loads $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$, the mean delays for jobs for each class $i \in F$ satisfy the following:*

$$E\left[D_i^{(\boldsymbol{\rho})}\right] \leq E\left[D_i^{(\boldsymbol{\rho}')}\right].$$

The above result holds for αF since it satisfies per-job rate monotonicity. However, one can check that the greedy rate allocation does not satisfy per-job rate monotonicity in general. Thus, it is not clear if such a bound holds for greedy rate allocation.

Now, if $\boldsymbol{\rho}'$ is homogeneous, then under αF rate allocation we have an explicit bound for mean delays via Theorem 2. Thus, consider the following region:

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

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 3 implies that the mean delay under αF rate allocation for each load $\boldsymbol{\rho} \in \mathcal{B}_D$ can be bounded by that for a corresponding symmetric load $\boldsymbol{\rho}' \in \mathcal{B}_H$, which in turn has an easily computable bound. 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 delay for a system with αF rate allocation for load $\boldsymbol{\rho}$ satisfies the following:*

$$E[D^{\alpha}_{\boldsymbol{\rho}}] \leq E[D^{B}_{\boldsymbol{\rho}'}].$$

4.3 Majorization bound

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 2, 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 $\mathbf{x}'$, then 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 proof of the theorem below, see Sect. 4.5.

Theorem 4 *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 $\mathbf{X}(t) \prec_w^{st} \mathbf{X}'(t)$ 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 4 above is stronger than Theorem 3 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.

For both $\mathbf{r}^G(.)$ and $\mathbf{r}^\alpha(.)$, Theorem 4, along with Theorem 2 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 4 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 2. 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 Theorem 1.

4.4 Lower bound

The following proposition provides a lower bound on the mean delay for any system with symmetric polymatroid capacity region, a feasible rate allocation policy, and with arbitrary loads. See Sect. 4.5 for a proof.

Proposition 5 *Consider a system with a symmetric polymatroid capacity region $\mathcal{C}$ with rank function $\mu(A) = h(|A|)$ for each $A \subset F$, an arbitrary feasible rate allocation policy, and with load $\boldsymbol{\rho} \in \hat{\mathcal{C}}$, i.e., the system is stabilizable. Let the total arrival rate for jobs, i.e. $\sum_{i\in F} \lambda_i$, be equal to λn. Then, the following lower bound on the mean delay holds:*

$$E[D] \geq \frac{1}{\lambda n}\left(\frac{\sum_{k=1}^n k \frac{|\boldsymbol{\rho}|^k}{\prod_{l=1}^k h(l)}}{1 + \sum_{k=1}^n \frac{|\boldsymbol{\rho}|^k}{\prod_{l=1}^k h(l)}}\right).$$

4.5 Proofs of coupling results

Proof of Theorem 2 Consider the following lemma regarding relative greediness of αF and BF.

Lemma 1 *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* $\sum_{l=1}^{k} r_{(l)}^{\alpha}(\mathbf{x}) \geq \sum_{l=1}^{k} r_{(l)}^{B}(\mathbf{y})$.

Roughly, this asserts that if state $\mathbf{x}$ is the 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 $\mathbf{y}$. The proof of this lemma is given in the Appendix (Relative greediness and other rate allocation properties). 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 the 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 $[t' - \epsilon_{t'}, t')$.

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 the interval $(0, \mu(F)]$. For k such that

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

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

Similarly, for k such that

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

let a departure occur in $(\mathbf{X}^{B}(t) : t \geq 0)$ at time t' in the kth 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 1. The following two cases arise.

Case 1 $\sum_{l=1}^{k} X^{\alpha}_{[l]}(t'-\epsilon_{t'}) < \sum_{l=1}^{k} X^{B}_{[l]}(t'-\epsilon_{t'})$. Since a maximum of one departure occurs at time t' in either processes, we clearly 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 1 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 first part of the theorem follows. The second part follows by application of Little's law to $(|\mathbf{X}^{\alpha}(t)| : t \geq 0)$ and $(|\mathbf{X}^{B}(t)| : t \geq 0)$. □

Proof of Theorem 3 Suppose $\mathbf{X}(0) \leq \mathbf{X}'(0)$. Below we couple the arrivals and departures of jobs in $(\mathbf{X}(t) : t \geq 0)$ and $(\mathbf{X}'(t) : t \geq 0)$ such that their marginal distributions remain intact and $\mathbf{X}(t) \leq \mathbf{X}'(t)$ almost surely for each $t \geq 0$.

Since the mean service requirement of jobs ν is same for both the systems, the corresponding arrival rates satisfy $\boldsymbol{\lambda} \leq \boldsymbol{\lambda}'$. For each i let Π_i and Π'_i be the Poisson arrival processes for class i in the respective systems. Let Π_i be obtained by sampling Π'_i. For each class i, the arrivals in $(\mathbf{X}'(t) : t \geq 0)$ at the sampled points, i.e., points in Π_i, see the average delay which is equal to the overall average delay of jobs in Π'_i for this system. Thus, the theorem follows if we couple the departures of jobs in both the systems such that for each point in Π_i, the corresponding job departure in $(\mathbf{X}(t) : t \geq 0)$ is no later than that in $(\mathbf{X}'(t) : t \geq 0)$. By using the per-flow rate monotonicity property, one can couple the service rate of these jobs at each time t so that if such a job departs from $(\mathbf{X}'(t) : t \geq 0)$ than the corresponding job departs from $(\mathbf{X}(t) : t \geq 0)$ as well, if it has not already. □

Proof of Theorem 4 The theorem can be proved in a fashion similar to that of Theorem 2, except for the following changes. For notational convenience, for each time t let $\lambda_{(k)}(t)$ and $\lambda'_{(k)}(t)$ be the arrival rates of the kth largest queues in $\mathbf{X}(t)$ and $\mathbf{X}'(t)$, respectively, with ties broken arbitrarily.

1. *Coupling of arrivals* For each point t' in Π_a, we choose a random variable $Z_{t'}$ independently and uniformly from the interval $(0, |\boldsymbol{\lambda}|]$. For each k such that

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

let an arrival occur in $(\mathbf{X}(t) : t \geq 0)$ at time t' in the kth largest queue of $\mathbf{X}(t' - \epsilon_{t'})$. Similarly, for each k such that

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

let an arrival occur in $(\mathbf{X}'(t) : t \geq 0)$ at time t' in the kth largest queue of $\mathbf{X}'(t' - \epsilon_{t'})$.

2. *Coupling of departures* Similar to that of Theorem 2, except that instead of Lemma 1 for a proof of weak-majorization upon a potential departure, we use the following lemma which asserts that αF and greedy provide a larger rate to longer queues in more balanced states.

Lemma 2 *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* $\sum_{l=1}^{k} r^{\alpha}_{(l)}(\mathbf{x}) \geq \sum_{l=1}^{k} r^{\alpha}_{(l)}(\mathbf{y})$, *and also* $\sum_{l=1}^{k} r^{G}_{(l)}(\mathbf{x}) \geq \sum_{l=1}^{k} r^{G}_{(l)}(\mathbf{y})$.

For $\mathbf{r}^{G}(.)$, it is easy to verify that the lemma holds. For $\mathbf{r}^{\alpha}(.)$, it follows from Lemma 9 in the Appendix (Relative greediness and other rate allocation properties).

Hence, the result follows. □

Proof of Proposition 5 Consider a queue where the jobs arrive as a Poisson point process with rate λn. The buffer size is finite and equal to n. Thus, an arrival is blocked if there are already n ongoing jobs in the queue. Service requirements of jobs are i.i.d. exponential with rate ν. The total service rate of jobs at each time is state dependent, as follows: if there are $\tilde{x}(t)$ ongoing jobs in the queue at time t then the total service rate at time t is equal to $h(\tilde{x}(t))$. One can check that the mean number of jobs in a stationary regime for this system is given by

$$E[\tilde{X}] = \frac{\sum_{k=1}^{n} k \frac{|\boldsymbol{\rho}|^k}{\prod_{l=1}^{k} h(l)}}{1 + \sum_{k=1}^{n} \frac{|\boldsymbol{\rho}|^k}{\prod_{l=1}^{k} h(l)}}.$$

It is easy to check that for a given total number of ongoing jobs, the overall service rate in the above queue is greater than or equal to that in the original system with symmetric polymatroid capacity region. Thus, one can couple the arrivals and departures of the two systems such that the above queue has a lower than or equal number of active jobs at each time as compared to the original system. The result then follows by applying Little's law to the original system. □

5 Large system has approximately symmetric capacity

In this section we consider a large content delivery system employing resource pooling and show that such a system not only has polymatroid capacity, but under appropriate assumptions becomes approximately symmetric.

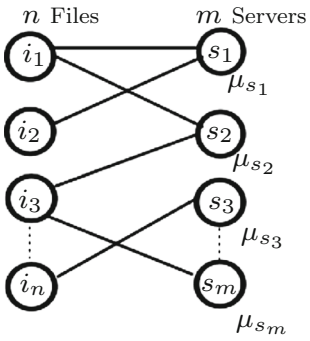

Fig. 2 Graph $G^{(n)} = (F^{(n)} \cup S^{(n)}; E^{(n)})$ modeling the placement of copies of n files across $m = \lceil bn \rceil$ servers with finite service capacities in a content delivery system

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 (see Fig. 2). For each node $s \in S^{(n)}$, let $N_s^{(n)}$ 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 ceiling and floor notation to avoid clutter.

We associate each file in $F^{(n)}$ to a class of jobs where the job corresponds to a download request for that file. The arrival processes and service requirements for the jobs are as described in Sect. 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 $\mathbf{x}$: $\forall A \subset F^{(n)}$,

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

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

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

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, a uniform law of large numbers holds where $\mathcal{C}^{(n)}$ 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 which we denote by $\mathcal{G}^{(n)} = (F^{(n)} \cup S^{(n)}; \mathcal{E}^{(n)})$. 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. Since the probability that a server does not have a copy of a file is equal to $1 - \frac{c_n}{m}$, we have

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

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 that $\bar{\mu}^{(n)}(A)$ 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 5 *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*

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

and

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

This result gives us the 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

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

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 $\mathcal{G}^{(n)}$ 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 5

Here we will only show

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

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

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{-\frac{\epsilon^2}{2}mp_k^{(n)}}, \tag{12}$$

and,

$$P\left(M^{(n)}(A) \leq (1-\epsilon)\bar{\mu}^{(n)}(A)\right) \leq e^{-mH\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right)}, \tag{13}$$

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 (Technical lemmas for proof of Theorem 5).

Lemma 3 *Let a sequence $(g_n : n \in \mathbb{N})$ be such that $g_n = o(c_n)$. Let $\delta_1 < 1$ be a positive constant independent of k and n. 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 4 *There exists a positive constant δ, independent of k and n, 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:

$$g_n = \omega(\log n) \text{ and } g_n = o(c_n). \tag{14}$$

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 the following two cases.

Case 1 $0 \leq k \leq \frac{n}{g_n}$: From Lemma 3, 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 (12), 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

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

for a constant δ_2 less than $\frac{\epsilon^2}{2}\delta_1 b$.

Case 2 $\frac{n}{g_n} < k \leq n$: In this case, we use the concentration result (13). From Lemma 4, there exists a constant δ_6 such that

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

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, $\delta_6 m - \epsilon kc_n \leq -(\epsilon/2)kc_n$ for each k such that $\frac{n}{g_n} < k \leq n$, and consequently there exists a constant δ_7 such that

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

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

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

for a constant δ_8 less than δ_7. 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

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

for a constant δ_{10} less than δ_9. 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 Sects. 4 and 5 to exhibit performance robustness in large-scale content delivery systems. In Sect. 5 we showed that large systems support symmetric polymatroid capacity regions. This allows us to apply the performance bounds developed in Sect. 4 for symmetric polymatroid capacity regions.

However, there is one more hurdle to overcome before we can apply our bounds from Sect. 4. Recall, from Corollary 3, under Assumptions 1 and 2 the random capacity region for a content delivery system *contains* and is *contained by* approximate symmetric polymatroids with high probability. A realization of the random capacity region 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 certain monotonicity conditions for rate allocations. Consider the following monotonicity condition.

Definition 6 (*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.

Further, recall per-job rate monotonicity defined in Sect. 4.2, where the rate allocated to each job (viz., $\frac{r_i(\mathbf{x})}{x_i}$ for jobs in class i) only decreases when an additional job is added into the system. The following lemma can be shown to hold through a simple coupling argument across jobs for arbitrary polymatroid capacity regions.

Lemma 5 *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 5 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. However, the condition $\theta_n = o\left(\min(\frac{n}{\log n}, c_n)\right)$ implies that peak per-class load is limited, i.e., it constrains the heterogeneity in load 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. For a proof, see Sect. 6.2.

Theorem 6 *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. Suppose $c_n = \omega(\log n)$ copies for each file are stored across servers at random. Let $\mathcal{G}^{(n)} = (F^{(n)} \cup S^{(n)}; \mathcal{E}^{(n)})$ represent the associated random bipartite graph representing file placement across servers.*

Given a realization of $\mathcal{G}^{(n)}$, let jobs for each file-class $i \in F^{(n)}$ arrive at rate λ_i. Let $\boldsymbol{\lambda}^{(n)} = (\lambda_i : i \in F^{(n)})$. Let the mean service requirement of jobs be ν, where ν is independent of n. Let $\boldsymbol{\rho}^{(n)} = \nu \boldsymbol{\lambda}^{(n)}$. Suppose that the jobs are served as per α-fair rate allocation.

Let $(\theta_n : n \in \mathbb{N})$ be a sequence such that $\theta_n = o\left(\min(\frac{n}{\log n}, c_n)\right)$. Fix a constant $\gamma < 1$. Let $\mathcal{B}^{(n)} = \{\boldsymbol{\rho} : \max_i \rho_i \leq \theta_n \text{ and } |\boldsymbol{\rho}| \leq \gamma \xi m\}$. Suppose that for each n we have $\boldsymbol{\rho}^{(n)} \in \mathcal{B}^{(n)}$. Fix a constant $\delta > 1$. Let $E[D^{(n)}|\mathcal{G}^{(n)}]$ be the conditional expectation of delay of a typical job with respect to the σ-algebra generated by $\mathcal{G}^{(n)}$. Then, we have

$$\lim_{n \to \infty} P\left(E[D^{(n)}|\mathcal{G}^{(n)}] \leq \delta \frac{\nu}{\xi c_n} \frac{1}{\gamma} \log\left(\frac{1}{1-\gamma}\right)\right) = 1.$$

6.1 Numerical validation and robustness of Theorem 6

The mean delay bound in Theorem 6 holds with high probability when the system size n is large, and when the load heterogeneity θ_n is small as compared to c_n. Below, we numerically explore the impact of the system size and these parameters on performance and our bounds. The motivation for our work is, in part, that simulation of large systems is difficult and it is desirable to reach a rough understanding of how performance scales. To this end, we consider systems using randomized file placement, and assume that

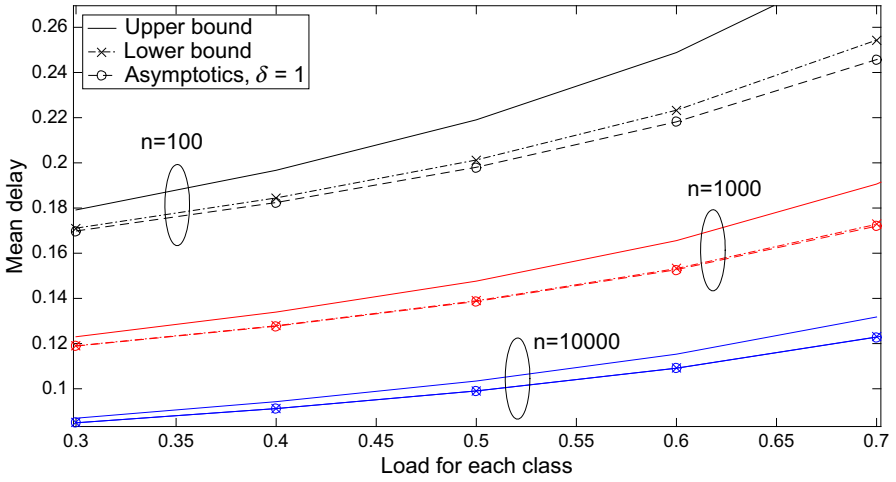

Fig. 3 Convergence of mean delay at different loads for symmetric systems as n increases. $m = n$, $c_n = \lceil \log_2 n \rceil$, $\xi = 1$, $\nu = 1$, and $\delta = 1$. Load is symmetric across classes

the capacity region is essentially symmetric—in our scaling regime, this is known to happen with high probability, see Theorem 5. Symmetry of the capacity region allows us to numerically compute the mean delay, and compare exact results to our asymptotic bounds, for large systems.

We first consider a large system with both symmetric capacity and symmetric load across classes. Theorem 1, along with Theorem 2, provides an upper bound for mean delay under α-fair rate allocation. Further, Proposition 5 provides a lower bound for the same. Figure 3 exhibits these bounds as a function of load per class for several systems with large n, and $c_n = \lceil \log_2 n \rceil$, and compares it with the asymptotic expression for expected delay given in Theorem 6 (i.e., $\delta \frac{\nu}{\xi c_n} \frac{1}{\gamma} \log\left(\frac{1}{1-\gamma}\right)$) for $\delta = 1$. As can be seen, as n increases, both bounds converge to the asymptotic expression, for example, the relative error of upper bound for $n = 1000$ and $\gamma = 0.6$ is less than 10%. Recall that the expression in Theorem 6 is an asymptotic upper bound for $\delta > 1$ (thus the asymptotic expression shown in the figure for $\delta = 1$ is the most aggressive bound one could hope for). Thus, n needs to be as large as 1000 or more for the asymptotic upper bound to be meaningful at medium loads.

Next we study the impact of load heterogeneity. Recall that in our model for constrained heterogeneity we allow the peak per-class load to be at most θ_n while maintaining the total system load to be less than or equal to $\gamma \xi m$. Thus, the 'worst case' load heterogeneity is when the total system load is equal to $\gamma \xi m$ and there is a subset of classes which have load equal to θ_n, with the remaining classes having a load equal to 0. An upper bound for mean delay for a system with such a worst case load and with α-fair rate allocation can again be obtained via the expression in Theorem 1, with load per class equal to θ_n but with a smaller total number of classes.

Figure 4 exhibits our mean delay upper bound obtained as above as a function of θ_n, and compares it with the asymptotic bound obtained via Theorem 6 for $\delta = 2$. For

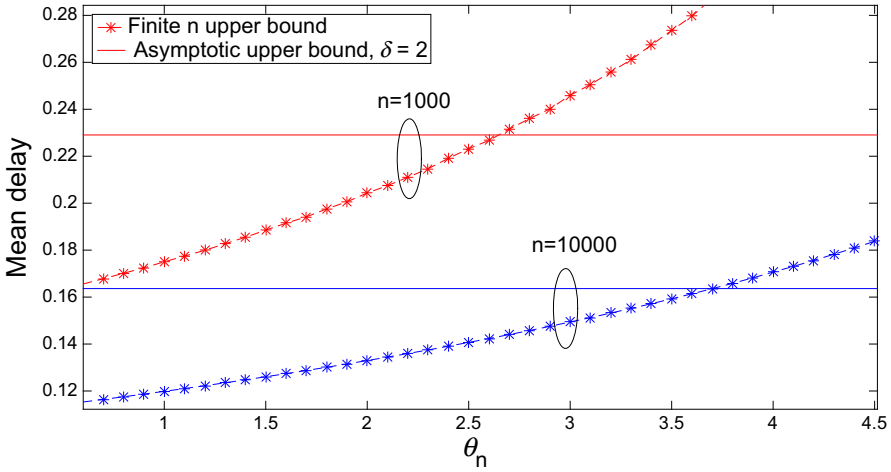

Fig. 4 Impact of heterogeneity θ_n on mean delays. $m = n$, $c_n = \lceil \log_2 n \rceil$, $\xi = 1$, $\nu = 1$, and $\gamma = 0.6$

$n = 10000$, the asymptotic bound holds as θ_n varies from 0.6 to up to 3.7. Note that $\theta_n = 0.6$ corresponds to a system with homogeneous load across classes. Thus, for a large system the asymptotic bound is good as long as the peak per-class load θ_n is no more than six times the per-class load of the homogeneous system.

6.2 Proof of Theorem 6

In light of Corollary 3, we consider a symmetric capacity region which, with high probability, contains the capacity region resulting from randomized file placement. Further, to obtain an upper bound on the mean delay for heterogeneous loads, we consider a system with extremely unbalanced arrivals in that the arrival rate is maximum for a subset of classes and negligible for others. The bound is obtained via the mean delay expression under balanced fairness for the extremely unbalanced system.

Without loss of generality, assume $\delta < \frac{1}{\gamma}$. From Corollary 3 and the definitions of $\bar{\mathcal{C}}^{(n)}$ and $\bar{\mu}^{(n)}(.)$, with high probability the 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

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

Thus, from Lemma 5 and Corollary 3, the theorem follows if we show that for a system with (deterministic) capacity region $\tilde{\mathcal{C}}^{(n)}$ and with α-fair rate allocation the

mean delay is upper bounded by $\delta \frac{\nu}{\xi c_n} \frac{1}{\gamma} \log\left(\frac{1}{1-\gamma}\right)$ for large enough n. Thus, for the rest of the proof we will assume that the system has capacity region $\tilde{\mathcal{C}}^{(n)}$ and α-fair rate allocation, and eventually establish the mean delay bound.

Note that 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

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

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

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

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.

Now we consider a case where certain classes have maximum load (i.e., θ_n) and the rest have load 0, while ensuring that the overall system load is still approximately γm.

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

Henceforth, we assume BF rate allocation and let the load $\boldsymbol{\rho}^{(n)} = \hat{\boldsymbol{\rho}}^{(n)}$. For each n, we invoke Proposition 4 and Theorem 1 with ρ replaced by θ_n and n replaced by t_n, to obtain an expression for $\pi_k^{(n)}$ and $\beta_k^{(n)}$, and eventually mean delay. We first show below concentration for $\pi_k^{(n)}$ using Proposition 4.

Below we refrain from using ceiling and floor to avoid cluttering.

Theorem 7 *Consider a system with capacity region $\tilde{\mathcal{C}}^{(n)}$ and with the load vector $\hat{\boldsymbol{\rho}}^{(n)}$. Under balanced fair rate allocation, $\pi_k^{(n)}$, which represents the stationary probability that k classes are active in the system, satisfies the following concentration result. For any positive constants $\epsilon > 1$ and $\epsilon' < 1$ independent of n, there exists a constant $\tilde{\delta} < 1$ such that for large enough n we have*

$$\sum_{k=\epsilon' b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}}^{\epsilon b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}} \pi_k^{(n)} \geq 1 - \tilde{\delta}^{\frac{m}{c_n}}. \tag{15}$$

Proof From Proposition 4 for $k = 1, \ldots, t_n$ we have

$$\pi_k^{(n)} = \frac{(t_n - k + 1)\theta_n}{h^{(n)}(k) - k\theta_n}\pi_{k-1}^{(n)}. \tag{16}$$

Fix a constant δ_{11} independent of n such that $0 < \delta_{11} < 1$. Let

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

Then, one can check that $h^{(n)}(k_{\downarrow}^{(n)}) \le \gamma'\delta_{11}\xi' m$.

In fact, we have $h^{(n)}(k) \le \gamma'\delta_{11}\xi' m, \quad \forall k \le k_{\downarrow}^{(n)}$. Using (16), for each $k \le k_{\downarrow}^{(n)}$ we have

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

for a positive constant δ_{12} such that $\delta_{11} < \delta_{12} < 1$, and large enough n.

Equivalently, $\pi_k^{(n)} \le \delta_{12}\pi_{k+1}^{(n)} \quad \forall k < k_{\downarrow}^{(n)}$.

Fix a positive constant $\epsilon_1 < 1$. Then, for all $k < \epsilon_1 k_{\downarrow}^{(n)}$ we have

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

Now, fix a constant δ_{13} independent of n such that $\gamma' < \delta_{13} < 1$ and let

$$k_{\uparrow}^{(n)} = \frac{m}{c_n}\log\left(\frac{1}{1-\gamma'/\delta_{13}}\right).$$

Then, one can check that $\frac{h^{(n)}(k_{\uparrow}^{(n)})}{\xi' m} \to \gamma'/\delta_{13}$ as $n \to \infty$. Thus, for some constant δ'_{13} such that $\delta_{13} < \delta'_{13} < 1$, we have $h^{(n)}(k_{\uparrow}^{(n)}) \ge \gamma'\xi' m/\delta'_{13}$. In fact, for all $k \ge k_{\uparrow}^{(n)}$ we have $h^{(n)}(k) \ge \gamma'\xi' m/\delta'_{13}$.

Now, for large enough n, $\gamma'\xi' m/\delta'_{13} \ge \gamma'\xi' m + \theta_n \ge (t_n + 1)\theta_n$.

Thus, for large enough n, we have $h^{(n)}(k) - k\theta_n \ge (t_n - k + 1)\theta_n \quad \forall k \ge k_{\uparrow}^{(n)}$, or equivalently from (16),

$$\pi_k^{(n)} \le \pi_{k-1}^{(n)} \quad \forall k \ge k_{\uparrow}^{(n)}. \tag{17}$$

In fact, for a fixed positive constant $\epsilon_2 > 1$, for all k such that $k_\uparrow^{(n)} \le k \le \epsilon_2 k_\uparrow^{(n)}$ we have

$$\pi_k^{(n)} \le \frac{(t_n - k + 1)\theta_n}{\gamma'\xi' m/\delta'_{13} - k\theta_n}\pi_{k-1}^{(n)} \le \frac{t_n\theta_n}{\gamma'\xi' m/\delta'_{13} - \epsilon_2 k_\uparrow^{(n)}\theta_n}\pi_{k-1}^{(n)}$$
$$\le \frac{\gamma'\xi' m}{\gamma'\xi' m/\delta'_{13} - o(n)}\pi_{k-1}^{(n)} \le \delta_{14}\pi_{k-1}^{(n)},$$

for a positive constant δ_{14} such that $\delta'_{13} < \delta_{14} < 1$, and for large enough n. Thus, $\pi_{\epsilon_2 k_\uparrow^{(n)}}^{(n)} \le \delta_{14}^{(\epsilon_2-1)k_\uparrow^{(n)}}\pi_{k_\uparrow^{(n)}}^{(n)}$ for large enough n. Further, using (17) we get

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

Thus, we get

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

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

We now provide a bound for $\beta_k^{(n)}$. From (9), for $k = 1, \ldots, t_n$ we have

$$\beta_k^{(n)} = \sum_{l=1}^{k}\frac{h^{(n)}(l)}{h^{(n)}(l) - l\theta_n} = \sum_{l=1}^{k}\frac{1}{1 - \frac{l\theta_n}{h^{(n)}(l)}}. \tag{18}$$

Using $g_n = \frac{\theta_n}{\gamma'\xi' b}$ in Lemma 3, we get $h^{(n)}(k) = \xi' bnp_k^{(n)} \ge \frac{\delta_{18}}{\gamma'}k\theta_n$ for large enough n and some constant δ_{18} such that $\gamma' < \delta_{18} < 1$. From (18), for each $k = 1, \ldots, t_n$,

for large enough n we have

$$\beta_k^{(n)} \leq \delta_{19} k$$

for some constant δ_{19} which is greater than 1.

The above bound for $\beta_k^{(n)}$ is somewhat loose, especially for lower values of k. Recall, the concentration result, namely Theorem 7, implies that the number of active classes is smaller than $\epsilon b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}$ with high probability. The bound on $\beta_k^{(n)}$ can be further improved for the smaller values of k as follows.

Suppose $h^{(n)}(.)$ is a continuous function, i.e., $h^{(n)}(t) = \xi' m \left(1 - e^{-\frac{tc_n}{m}}\right)$ for each $t \in \mathbb{R}^+$. Then, by concavity of $h^{(n)}(t)$ and noting that $h^{(n)}(0) = 0$, we get $\frac{h^{(n)}(t)}{t} \geq \frac{d}{dt}h^{(n)}(t)$. Further, by concavity, for each $k \leq \epsilon b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}$ we have $\frac{h^{(n)}(k)}{k} \geq \frac{d}{dt}h^{(n)}(t)\Big|_{t=k} \geq \frac{d}{dt}h^{(n)}(t)\Big|_{t=b\log(\frac{1}{1-\gamma'})\frac{n}{c_n}} = \xi' c_n (1-\gamma')^{-\epsilon}$.

From (18), for $k = 1, \ldots, \epsilon b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}$, we have

$$\beta_k^{(n)} \leq \sum_{l=1}^{k} \frac{1}{1 - \frac{\theta_n}{\xi' c_n (1-\gamma')^{-\epsilon}}} = k \frac{1}{1-o(1)}.$$

We are now ready to bound mean delay. For large enough n, we have

$$\begin{aligned}
\frac{t_n \theta_n}{\nu} E[D^{(n)}] &= \sum_{k=1}^{t_n} \beta_k^{(n)} \pi_k^{(n)} = \sum_{k=1}^{\epsilon b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}} \beta_k^{(n)} \pi_k^{(n)} + \sum_{k=\epsilon' b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}+1}^{t_n} \beta_k^{(n)} \pi_k^{(n)} \\
&\leq \sum_{k=1}^{\epsilon b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}} k \frac{1}{1-o(1)} \pi_k^{(n)} + \sum_{k=\epsilon' b \log(\frac{1}{1-\gamma'})\frac{n}{c_n}+1}^{t_n} \delta_{19} k \pi_k^{(n)} \\
&\leq \epsilon b \log\left(\frac{1}{1-\gamma'}\right) \frac{n}{c_n} \frac{1}{1-o(1)} + \delta_{19} t_n \tilde{\delta}^{\frac{m}{c_n}}.
\end{aligned}$$

The theorem thus follows from the definition of t_n, γ' and ξ', and the fact that ϵ, δ, and $\tilde{\delta}$ were chosen arbitrarily. □

7 Conclusions

Our main conclusions address both practical and theoretical aspects associated with such systems. We show that an infrastructure which allows a user to download in parallel from a pool of servers can achieve scalable performance under limited heterogeneity in file demands. Some elements of content delivery infrastructure such as a centralized back end which handles cache misses at distributed sites may see less pronounced heterogeneity in demands. Our results suggest that pooling of server resources

is a scalable approach towards delivering content for such centralized systems without requiring complex caching strategies internally.

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.

Appendix

Relative greediness and other rate allocation properties

Below, we provide a proof of Lemma 1 which asserts that αF is more greedy than BF. Along the way, we develop several other properties of the rate allocation policies.

The proof of Lemma 1 stems from the properties (1) and (2) below on per-job rate assignment for αF and BF.

(1.) *αF gives the most balanced per-job rate allocation* This property follows from the fact that αF is equivalent to max-min fair rate allocation; see Proposition 2. Formally,

Lemma 6 *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 majorization from above.*

Proof Let the set of flows be $q_{A_\mathbf{x}}$. It is easy to show that $\mathbf{b}^\alpha$ is the unique solution to the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \operatorname{sign}(1-\alpha) \sum_{u \in q_{A_\mathbf{x}}} \hat{b}_u^{1-\alpha}, \\ \text{subject to} \quad & \sum_{u \in q_A} \hat{b}_u \le \mu(A), \ \forall A \subset A_\mathbf{x}, \\ & \hat{b}_u \ge 0, \ \forall u \in q_F. \end{aligned}$$

Also, since $\tilde{\mathbf{b}}$ is feasible, it satisfies the constraints of the above problem. The result then follows by noting that the objective function of the above problem is monotonic and Schur-concave in $(\hat{b}_u : u \in q_{A_\mathbf{x}})$ [13,16]. □

(2.) *In αF and BF, longest queues have smallest 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 given in Appendix (In BF, longest queues have smallest per-job rates). Formally,

Lemma 7 *αF and BF rate allocations satisfy the following property for any state $\mathbf{x}$: if $x_i > x_j$ then $\frac{r_i(\mathbf{x})}{x_i} \le \frac{r_j(\mathbf{x})}{x_j}$.*

Proof Below we prove the lemma for αF rate allocation. For a proof of this lemma for BF rate allocation, see the Appendix (In BF, longest queues have smallest per-job rates).

Let $\mathbf{b}^\alpha = (b^\alpha_u : u \in q_{A_\mathbf{x}})$ represent the rates assigned to ongoing flows under αF rate allocation in state $\mathbf{x}$. Suppose $x_i > x_j$, but $\frac{r^\alpha_i(\mathbf{x})}{x_i} > \frac{r^\alpha_j(\mathbf{x})}{x_j}$. Then, for each $u' \in q_i$ and $v' \in q_j$, we have $b^\alpha_{u'} > b^\alpha_{v'}$. Let $\tilde{\mathbf{b}} = (\tilde{b}_u : u \in q_{A_\mathbf{x}})$, where $\tilde{b}_u = b^\alpha_u$ for each $u \in q_{A_\mathbf{x}\setminus\{i,j\}}$ and $\tilde{b}_u = \frac{r^\alpha_i(\mathbf{x})+r^\alpha_j(\mathbf{x})}{x_i+x_j}$ for each $u \in q_{\{i,j\}}$. It can be checked that $\tilde{b}_u$ is feasible and that $\tilde{\mathbf{b}} \prec^w \mathbf{b}^\alpha$. This contradicts Lemma 6. Hence the result. □

Now, let us study what the above properties imply for per-class rate allocation. Consider a state $\mathbf{x}$. Lemma 7 above implies that the most disadvantaged jobs are the ones which belong to longest queues for both BF and αF. This, along with Lemma 7, implies that αF provides larger rate to longest queues. Thus, we get the following property:

(3.) *αF provides a larger rate to longest queues compared to BF* Formally, this property can be stated as follows:

Lemma 8 *For any state $\mathbf{x}$, we have $\sum_{l=1}^k r^\alpha_{(l)}(\mathbf{x}) \geq \sum_{l=1}^k r^B_{(l)}(\mathbf{x})$ for each $k \in \{1, 2, \ldots, n\}$.*

Proof Let $u_1, u_2, \ldots, u_{x_{[1]}}$ be the flows in the class corresponding to $x_{[1]}$. Similarly, for each $k \in \{2, \ldots, n\}$, let $u_{\sum_{l=1}^{k-1} x_{[l]}+1}, \ldots, u_{\sum_{l=1}^{k} x_{[l]}}$ be the flows in the class corresponding to $x_{[k]}$. From Lemma 7, under both BF and αF rate allocation we have $b_{u_1} \leq b_{u_2} \leq \ldots \leq b_{u_{|x|}}$. Thus, it is enough to show that $\mathbf{b}^\alpha \prec^w \mathbf{b}^B$. However, this follows from Lemma 6. □

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 provides a larger rate to longer queues in more balanced states* Formally, this property can be stated as follows:

Lemma 9 *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 $\sum_{l=1}^k r^\alpha_{(l)}(\mathbf{x}) \geq \sum_{l=1}^k r^\alpha_{(l)}(\mathbf{y})$.*

Proof Due to monotonicity of $\mathbf{r}^\alpha(\mathbf{y})$ with respect to components of $\mathbf{y}$, it is enough to show the result for the case where $\mathbf{x} \prec \mathbf{y}$. Assume $\mathbf{x} \prec \mathbf{y}$. Let $u_1, u_2, \ldots, u_{x_{[1]}}$ be the flows in the class corresponding to $x_{[1]}$. Similarly, let $u_{\sum_{l=1}^{k-1} x_{[l]}+1}, \ldots, u_{\sum_{l=1}^{k} x_{[l]}}$ be the flows in the class corresponding to $x_{[k]}$ for each $k \in \{2, \ldots, n\}$. Let the corresponding rates assigned to flows under αF rate allocation be given by $\mathbf{b}^{(\mathbf{x})}$. Using Lemma 7, we have $b_{u_1} \leq b_{u_2} \leq \ldots \leq b_{u_{|\mathbf{x}|}}$. Similarly, let $v_1, v_2, \ldots, v_{|\mathbf{y}|}$ be the flows corresponding to state $\mathbf{y}$ and construct the corresponding $\mathbf{b}^{(\mathbf{y})}$.

One can check that $\tilde{\mathbf{b}}^{(\mathbf{x})} = (\tilde{b}^{(\mathbf{x})}_{u_k} : k \in \{1, 2, \ldots, |\mathbf{x}|\})$, where $\tilde{b}^{(\mathbf{x})}_{u_k} = b^{(\mathbf{y})}_{v_k}$ for each $k \leq |\mathbf{x}|$, is feasible under state $\mathbf{x}$ as well. Thus, from Lemma 6, we have $\mathbf{b}^{(\mathbf{x})} \prec^w \tilde{\mathbf{b}}^{(\mathbf{x})}$. From this, the result follows. □

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

(5.) *αF is more greedy than BF* We now prove Lemma 1. Consider states $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x} \prec_w \mathbf{y}$. From Lemma 9 we have $\sum_{l=1}^{k} r^{\alpha}_{(l)}(\mathbf{x}) \geq \sum_{l=1}^{k} r^{\alpha}_{(l)}(\mathbf{y})$, and from Lemma 8 we have $\sum_{l=1}^{k} r^{\alpha}_{(l)}(\mathbf{y}) \geq \sum_{l=1}^{k} r^{B}_{(l)}(\mathbf{y})$. Hence, Lemma 1 holds.

In BF, longest queues have smallest per-job rates

Lemma 10 *For any state $\mathbf{x}$, if $x_i > x_j$ then $\frac{r^B_i(\mathbf{x})}{x_i} \leq \frac{r^B_j(\mathbf{x})}{x_j}$.*

Proof Using the definition of balanced fairness, we have $\frac{r^B_i(\mathbf{x})}{r^B_j(\mathbf{x})} = \frac{\Phi(\mathbf{x}-\mathbf{e}_i)}{\Phi(\mathbf{x}-\mathbf{e}_j)}$. Thus, we need to show that $\frac{\Phi(\mathbf{x}-\mathbf{e}_i)}{\Phi(\mathbf{x}-\mathbf{e}_j)} \leq \frac{x_i}{x_j}$. It is thus sufficient to prove that $\frac{\Phi(\mathbf{x}+\mathbf{e}_i)}{\Phi(\mathbf{x}+\mathbf{e}_j)} \geq \frac{x_j+1}{x_i+1}$ holds for each $\mathbf{x}$ since the result follows when $\mathbf{x}$ is replaced with $\mathbf{x} - \mathbf{e}_i - \mathbf{e}_j$.

We show below that $\frac{\Phi(\mathbf{x}+\mathbf{e}_i)}{\Phi(\mathbf{x}+\mathbf{e}_j)} \geq \frac{x_j+1}{x_i+1}$ holds for each $\mathbf{x}$.

Fix $i, j \in F$. By symmetry of balanced fairness and the capacity region, the result holds for each $\mathbf{x}$ such that $x_i = x_j$. We show that the result holds for each $\mathbf{x}$ such that $x_i \geq x_j$ using induction on $|\mathbf{x}|$. We will use the following recursive expression for $\Phi(.)$ which we get from the definition of balanced fair and Proposition 3: For each state $\mathbf{x}$, we have

$$\Phi(\mathbf{x}) = \frac{\sum_{i' \in A_{\mathbf{x}}} \Phi(\mathbf{x} - \mathbf{e}_{i'})}{\mu(A_{\mathbf{x}})}. \tag{19}$$

The result clearly holds for the base case of $|\mathbf{x}| = 0$. Assume that the result holds for all states $\mathbf{x}'$ such that $|\mathbf{x}'| < |\mathbf{x}|$. We prove that the result holds for the state $\mathbf{x}$ under each of the following two possible cases for $\mathbf{x}$:

Case 1 $A_{\mathbf{x}+\mathbf{e}_i} \subsetneq A_{\mathbf{x}+\mathbf{e}_j}$: This case is possible only if $x_i > 0$ and $x_j = 0$. Thus, $\mu(A_{\mathbf{x}+\mathbf{e}_i}) \leq \mu(A_{\mathbf{x}+\mathbf{e}_j})$. Using (19), we get

$$\frac{\Phi(\mathbf{x}+\mathbf{e}_i)}{\Phi(\mathbf{x}+\mathbf{e}_j)} \geq \frac{\Phi(\mathbf{x}) + \sum_{i' \in A_{\mathbf{x}} \setminus \{i\}} \Phi(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_{i'})}{\Phi(\mathbf{x}) + \Phi(\mathbf{x} + \mathbf{e}_j - \mathbf{e}_i) + \sum_{i' \in A_{\mathbf{x}} \setminus \{i\}} \Phi(\mathbf{x} + \mathbf{e}_j - \mathbf{e}_{i'})}.$$

Using the induction hypothesis, we have $\frac{\Phi(\mathbf{x}+\mathbf{e}_i-\mathbf{e}_{i'})}{\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_{i'})} \geq \frac{x_j+1}{x_i+1}$ for each $i' \in A_{\mathbf{x}} \setminus \{i\}$. Thus, using the fact that $\frac{a_1+a_2}{b_1+b_2} \geq \frac{x}{y}$ if $\frac{a_k}{b_k} \geq \frac{x}{y}$ for each $k \in \{1, 2\}$, the result follows if we show that $\frac{\Phi(\mathbf{x})}{\Phi(\mathbf{x})+\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_i)} \geq \frac{x_j+1}{x_i+1}$. This in turn follows since $x_j = 0$ and $\frac{\Phi(\mathbf{x})}{\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_i)} \geq \frac{1}{x_i}$ holds by the induction hypothesis.

Case 2 $A_{\mathbf{x}+\mathbf{e}_i} = A_{\mathbf{x}+\mathbf{e}_j}$: Again using (19), we get

$$\frac{\Phi(\mathbf{x}+\mathbf{e}_i)}{\Phi(\mathbf{x}+\mathbf{e}_j)} = \frac{\Phi(\mathbf{x}) + \Phi(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_j) + \sum_{i' \in A_{\mathbf{x}} \setminus \{i,j\}} \Phi(\mathbf{x} + \mathbf{e}_i - \mathbf{e}_{i'})}{\Phi(\mathbf{x}) + \Phi(\mathbf{x} + \mathbf{e}_j - \mathbf{e}_i) + \sum_{i' \in A_{\mathbf{x}} \setminus \{i,j\}} \Phi(\mathbf{x} + \mathbf{e}_j - \mathbf{e}_{i'})}.$$

Again, using the induction hypothesis we have $\frac{\Phi(\mathbf{x}+\mathbf{e}_i-\mathbf{e}_{i'})}{\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_{i'})} \geq \frac{x_j+1}{x_i+1}$ for each $i' \in A_{\mathbf{x}}\backslash\{i, j\}$. Thus, we only need to show that $\frac{\Phi(\mathbf{x})+\Phi(\mathbf{x}+\mathbf{e}_i-\mathbf{e}_j)}{\Phi(\mathbf{x})+\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_i)} \geq \frac{x_j+1}{x_i+1}$. We show this below.

By the induction hypothesis, we have $\frac{\Phi(\mathbf{x}+\mathbf{e}_i-\mathbf{e}_j)}{\Phi(\mathbf{x})} \geq \frac{x_j}{x_i+1}$ and $\frac{\Phi(\mathbf{x})}{\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_i)} \geq \frac{x_j+1}{x_i}$. Thus, we get

$$\frac{\Phi(\mathbf{x})+\Phi(\mathbf{x}+\mathbf{e}_i-\mathbf{e}_j)}{\Phi(\mathbf{x})+\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_i)} = \frac{1+\frac{\Phi(\mathbf{x}+\mathbf{e}_i-\mathbf{e}_j)}{\Phi(\mathbf{x})}}{1+\frac{\Phi(\mathbf{x}+\mathbf{e}_j-\mathbf{e}_i)}{\Phi(\mathbf{x})}} \geq \frac{1+\frac{x_j}{x_i+1}}{1+\frac{x_j+1}{x_i}} = \frac{x_j+1}{x_i+1}.$$

Hence, the result. □

Technical lemmas for proof of Theorem 5

Lemma 3 Let a sequence $(g_n : n \in \mathbb{N})$ be such that $g_n = o(c_n)$. Let $\delta_1 < 1$ be a positive constant independent of k and n. Then, for large enough n, we have

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

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

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

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

$$f^{(n)}(t) \geq \frac{f^{(n)}(n/g_n)}{(n/g_n)} t, \quad \forall t \text{ s.t. } 0 \leq t \leq n/g_n.$$

Hence, the lemma. □

Lemma 4 There exists a positive constant δ, independent of k and n, such that $H\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right) \geq -\delta + \epsilon\frac{kc_n}{m}$.

Proof From the definition,

$$H\left(p_k^{(n)}(1-\epsilon)||p_k^{(n)}\right) = p_k^{(n)}(1-\epsilon)\log(1-\epsilon)$$
$$+(1-p_k^{(n)}(1-\epsilon))\log\left(\frac{1-p_k^{(n)}(1-\epsilon)}{1-p_k^{(n)}}\right).$$

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:

$$(1-p_k^{(n)}(1-\epsilon))\log\left(1-p_k^{(n)}(1-\epsilon)\right) \geq \log\left(1-p_k^{(n)}(1-\epsilon)\right) \geq \log(1-(1-\epsilon)) = \log.$$

Thus, we have

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

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

References

1. Bonald, T.: Throughput performance in networks with linear capacity contraints. In: Proceedings of CISS, pp. 644 –649 (2006)
2. Bonald, T., Massoulié, L., Proutière, A., Virtamo, J.: A queueing analysis of max-min fairness, proportional fairness and balanced fairness. Queueing Syst. **53**, 65–84 (2006)
3. Bonald, T., Proutière, A.: Insensitive bandwidth sharing in data networks. Queueing Syst. **44**, 69–100 (2003)
4. Bonald, T., Proutière, A.: On stochastic bounds for monotonic processor sharing networks. Queueing Syst. **47**, 81–106 (2004)
5. Bonald, T., Proutière, A., Roberts, J., Virtamo, J.: Computational aspects of balanced fairness. In: Proceedings of ITC (2003)
6. Bonald, T., Virtamo, J.: Calculating the flow level performance of balanced fairness in tree networks. Perform. Eval. **58**(1), 1–14 (2004)
7. de Veciana, G., Lee, T.J., Konstantopoulos, T.: Stability and performance analysis of networks supporting elastic services. IEEE/ACM Trans. Netw. **9**(1), 2–14 (2001)
8. Dubhashi, D., Ranjan, D.: Balls and bins: A study in negative dependence. Random Struct. Algorithms **13**(2), 99–124 (1998)
9. Edmonds, J.: Submodular functions, matroids, and certain polyhedra. In: Proceedings of Calgary International Conference on Combinatorial Structures and Applications, pp. 69–87. Gordon and Breach, New York (1969)
10. Frank, B., Poese, I., Smaragdakis, G., Feldmann, A., Maggs, B.M., Uhlig, S., Aggarwal, V., Schneider, F.: Collaboration opportunities for content delivery and network infrastructures. In: H. Haddadi, O. Bonaventure (eds.) Recent Advances in Networking, pp. 305–377 (2013)
11. Joseph, V., de Veciana, G.: Stochastic networks with multipath flow control: Impact of resource pools on flow-level performance and network congestion. In: Proceedings of the ACM Sigmetrics, pp. 61–72 (2011)
12. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control for communication networks: shadow prices, proportional fairness and stability. J. Oper. Res. Soc. **49**(3), 237–252 (1998)
13. Lan, T., Kao, D., Chiang, M., Sabharwal, A.: An axiomatic theory of fairness in network resource allocation. In: Proceedings of IEEE Infocom, pp. 1–9 (2010)
14. Leconte, M., Lelarge, M., Massoulié, L.: Adaptive replication in distributed content delivery networks. arXiv preprint arXiv:1401.1770 (2014)
15. Lin, X., Shroff, N.: Utility maximization for communication networks with multipath routing. IEEE Trans. Autom. Control **51**(5), 766–781 (2006)
16. Marshall, A.W., Olkin, I., Arnold, B.C.: Inequalities: Theory of Majorization and Its Applications, 2nd edn. Springer, New York (2011)
17. Massoulié, L., Roberts, J.: Bandwidth sharing and admission control for elastic traffic. Telecommun. Syst. **15**(1–2), 185–201 (2000)

18. Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press, Cambridge (2005)
19. Mo, J., Walrand, J.: Fair end-to-end window-based congestion control. IEEE/ACM Trans. Netw. **8**(5), 556–567 (2000)
20. Moharir, S., Ghaderi, J., Sanghavi, S., Shakkottai, S.: Serving content with unknown demand: The high-dimensional regime. In: Proceedings of ACM Sigmetrics, pp. 435–447 (2014)
21. Nemhauser, G.L., Wolsey, L.A.: Integer and combinatorial optimization, vol. 18. Wiley, New York (1988)
22. Shah, V., de Veciana, G.: Performance evaluation and asymptotics for content delivery networks. In: IEEE Infocom, pp. 2607–2615 (2014)
23. Shah, V., de Veciana, G.: High performance centralized content delivery infrastructure: models and asymptotics. IEEE/ACM Trans. Netw. **23**, 1674 (2015)
24. Tsitsiklis, J.N., Xu, K.: Flexible queueing architectures. arXiv preprint arXiv:1505.07648 (2015)
25. Walrand, J.: An Introduction to Queueing Networks. Prentice Hall, Englewood Cliffs (1988)
26. Yeh, E.: Multiaccess and fading in communication networks. Ph.D. thesis, Massachusetts Institute of Technology (2001)