Selfish Traffic Allocation for Server Farms *

Artur Czumaj
Department of Computer Science
New Jersey Institute of Technology
czumaj@cis.njit.edu

Piotr Krysta
Max-Planck-Institut für Informatik
Saarbrücken, Germany
krysta@mpi-sb.mpg.de

Berthold Vöcking
Max-Planck-Institut für Informatik
Saarbrücken, Germany
voecking@mpi-sb.mpg.de

ABSTRACT

We investigate the price of selfish routing in non-cooperative networks in terms of the coordination and bicriteria ratios in the recently introduced game theoretic network model of Koutsoupias and Papadimitriou. We present the first thorough study of this model for general, monotone families of cost functions and for cost functions from Queueing Theory. Our main results can be summarized as follows.

- We give a precise characterization of cost functions having a bounded/unbounded coordination ratio. For example, cost functions that describe the expected delay in queueing systems have an unbounded coordination ratio.
- We show that an unbounded coordination ratio implies additionally an extremely high performance degradation under bicriteria measures. We demonstrate that the price of selfish routing can be as high as a bandwidth degradation by a factor that is linear in the network size.
- We separate the game theoretic (integral) allocation model from the (fractional) flow model by demonstrating that even a very small, in fact negligible, amount of integrality can lead to a dramatic performance degradation.
- We unify recent results on selfish routing under different objectives by showing that an unbounded coordination ratio under the min-max objective implies an unbounded coordination ratio under the average-cost (or total-latency) objective and vice versa.

Our special focus lies on cost functions describing the behavior of Web servers that can open only a limited number of TCP connections. In particular, we compare the performance of queueing systems that serve all incoming requests with servers that reject requests in case of overload.

From the result presented in this paper we conclude that queuing systems without rejection cannot give any reasonable guarantee on the expected delay of requests under selfish routing even when the injected load is far away from the capacity of the system. In contrast, Web server farms that are allowed to reject requests can guarantee a high quality of service for every individual request stream even under relatively high injection rates.

*Supported in part by NSF grant CCR-0105701, by SBR grant No. 421090, by DFG grant Vo889/1-1, and by the IST program of the EU under contract number IST-1999-14186 (ALCOM-FT).

STOC'02, May 19–21, 2002, Montreal, Quebec, Canada.

1. INTRODUCTION

In large-scale communication networks, like the Internet, it is usually impossible to globally manage network traffic. In the absence of global control it is therefore a reasonable assumption in traffic modeling that network users follow the most rational approach, that is, they behave selfishly to optimize their own individual welfare. This motivates the analysis of network traffic using models from Game Theory in which each player is aware of the situation facing all other players and tries to minimize its cost. Under these assumptions, the routing process should arrive into a so-called *Nash equilibrium* in which no network user has an incentive to change its strategy.

It is well known (and easy to see) that Nash equilibria do not always optimize the overall performance of the system. Therefore, in order to understand the phenomenon of non-cooperative systems Koutsoupias and Papadimitriou [14] initiated investigations of the *coordination ratio*, which is the ratio between the worst possible Nash equilibrium and the social (i.e., overall) optimum. In other words, this analysis seeks the price of uncoordinated selfish decisions ("the price of anarchy"). Koutsoupias and Papadimitriou [14, 19] proposed to investigate the coordination ratio for routing problems in which a set of agents is sending traffic along a set of parallel links with linear cost functions.

In this paper, we generalize the model of Koutsoupias and Papadimitriou towards more realistic cost functions. Our main focus is on a specific example in which parallel links are the natural choice: we investigate the effects of selfish behavior on a *Web server farm*. Suppose some companies maintain a set of servers distributed all over the world and offer content providers to store data for them. Such servers could store, e.g., large pictures and other embedded files, since this kind of data makes up most of the load. The request streams that would normally go to the servers of the content provider must now be redirected to these new Web servers. Clearly, this defines a load balancing problem in which streams must be mapped to the servers such that a high quality of service can be guaranteed for every stream. Important aspects that have to be taken into account is that different streams might have different characteristics, e.g., caused by different file lengths. For practical studies that investigate the reasons and impacts of this variability in traffic see, e.g., [2, 4, 5, 22, 23].

In nowadays server farms the mapping of data streams to servers is typically done by a centralized or distributed algorithm that is under control of the provider of the server farm. We can imagine, however, that such a service can be offered in a completely different way without global control. For example, each stream of requests is managed by a selfish agent (e.g., the content provider) that decides to which server the stream is directed. In this case, every agent would aim to minimize its own cost, e.g., the expected latency experienced by the requests in the stream or the fraction of requests that are rejected.

In this paper, we present the first thorough study of coordination and bicriteria ratios under more realistic cost functions. In previous works on traffic analysis in networks, it has been typically assumed that every data stream is completely described by a single rate. In some of our investigations we will make this assumption, too. In order to incorporate the variability of traffic, however, we will additionally study selfish routing under more complex cost functions that take into account different session length distributions. In fact, we will consider general distributions for session lengths. We further distinguish between *homogeneous traffic* in which all streams have the same session length distribution and *heterogeneous traffic* in which different streams might have different session length distributions.

1.1 Definition of the Routing Problem

The routing problem described above can be formally defined as an assignment problem with n data streams and m servers (or parallel links). The set of streams is denoted by $[n] = \{1, \ldots, n\}$ and the set of servers is denoted by $[m] = \{1, \ldots, m\}$. The data streams shall be mapped to the servers such that a cost function (describing, e.g., waiting or service times) is minimized. We aim at comparing the assignment obtained by selfish agents with a min-max optimal assignment.

A *server farm* is a set of m servers, all using the same policy to serve requests. Different servers, however, may have different *bandwidths*. Let b_j denote the bandwidth of server j.

Data streams are infinite sequences of requests for service (to the server farm). These sequences are assumed to be of a stochastic nature. For simplicity, we make a standard assumption that requests are issued by a large number of independent users and hence they arrive with *Poisson* distribution. Let r_i denote the *injection rate* of data stream i. For the lengths of the sessions, however, we allow general probability distributions. In particular, we assume the *session length* of stream i is determined by an arbitrary probability distribution $\mathcal{D}_i$. We define the *weight* of stream i to be $\lambda_i = r_i \cdot \mathbf{E}[\text{session length wrt. } \mathcal{D}_i]$.

We distinguish between fractional and integral assignments of data streams to the servers. In an *integral assignment*, every stream must be assigned to exactly one server. The mapping is described by an assignment matrix $\mathcal{X} = (x_i^j)_{i\in[n],j\in[m]}$, where x_i^j is an indicator variable with $x_i^j = 1$ if stream i is assigned to server j and 0 otherwise. In a *fractional assignment* the variables x_i^j can take arbitrary real values from $[0, 1]$, subject to the constraint $\sum_{j\in[m]} x_i^j = 1$, for every $i \in [n]$.

The cost occurring at the servers under some fixed assignment is defined by *families of cost functions* $\mathcal{F}_B = \{f_b | b \in B\}$, where B denotes the domain of possible bandwidth values and f_b describes the cost function for servers with bandwidth $b \in B$. Typically, we will assume $B = \mathbb{R}_{>0}$ or $B = \mathbb{N}_{>0}$, but occasionally we will study finite domains of bandwidth. For example, a collection of identical servers with some specified bandwidth b is formally described by a family of cost functions $\mathcal{F}_B$ with $B = \{b\}$.

The *load* of a server j under an assignment $\mathcal{X}$ is defined by $w_j = \sum_{i=1}^{n} \lambda_i \cdot x_i^j$ and the *cost* of server j is defined by

$$C_j = f_{b_j}(x_1^j, \ldots, x_n^j) .$$

Unless otherwise stated, we consider the routing problem with respect to the *min-max objective*. That is, we assume an *optimal assignment* minimizes the maximum cost over all servers:

$$opt = \min_{\mathcal{X}} \max_{j\in[m]} f_{b_j}(x_1^j, \ldots, x_n^j) .$$

We note here, that in the definition of *opt* the minimization is over all matrices $\mathcal{X}$ that are either integral (in the case of integral assignments) or fractional (in the case of fractional assignments). This distinction will be clear from the context.

1.2 Preliminaries in Game Theory

Integral assignments and Nash equilibria. We assume the decision about the assignment of a data stream $i \in [n]$ to a server is performed by an agent i who uses certain strategy to assign its data stream. Game Theory distinguishes between mixed and pure strategies. The set of *pure strategies* for agent $i \in [n]$ is $[m]$, that is, a pure strategy maps every stream to exactly one server and hence can be described by an integral assignment matrix $\mathcal{X}$. A *mixed strategy* is defined to be a probability distribution over pure strategies. In particular, the probability that agent i maps its stream to server j is denoted by p_i^j.

Observe that under these assumptions the load w_j and the cost C_j of server j are random variables wrt. the probabilities p_i^j, $i \in [n]$. Let ℓ_j denote the *expected cost on server j*, that is, $\ell_j = \mathbf{E}[C_j]$. For a stream i, let us define the *expected cost of stream i on server j* by $c_i^j = \mathbf{E}[C_j \,|\, x_i^j = 1]$. Recall that our objective is to minimize the maximum cost over all servers. Therefore, we define the *social cost* of a mixed assignment by

$$C = \mathbf{E}[\max_{j\in[m]} C_j] .$$

If selfish players aim to minimize their individual cost, then the resulting (possible mixed) assignment is in *Nash equilibrium*, that is, $p_i^j > 0$ implies $c_i^j \le c_i^q$, for every $i \in [n]$ and $j, q \in [m]$. In other words, a Nash equilibrium is characterized by the property that there is no incentive for any task to change its strategy.

Coordination and bicriteria ratios. The *coordination ratio for a fixed set of servers and streams* is defined by $\max \frac{C}{opt}$, where the maximum is over all Nash equilibria. Thus, the coordination ratio specifies how many times the cost can increase due to selfish behavior. The *coordination ratio R over a family of cost functions $\mathcal{F}_B$* is defined to be the maximum coordination ratio over all possible sets of streams and servers with cost functions from $\mathcal{F}_B$. Typically R is described by an asymptotic function in m.

In our study, we will identify several instances of cost functions for which R is *unbounded*. In this case, we will investigate *bicriteria* characteristics of the system. Let opt_Γ denote the value of an optimal solution over pure strategies assuming that all the injection rates r_i are increased by a factor of Γ. Then, the *bicriteria ratio* $\overline{R}$ is defined to be the smallest Γ satisfying $C \le opt_\Gamma$ over all Nash equilibria. In other words, the bicriteria ratio describes how many times the injected rates (the amount of traffic in the system) must be decreased so that the worst-case cost in Nash equilibrium cannot exceed the optimal cost for the original rates.

Fractional assignments and selfish flow. The motivation to consider fractional assignments is to assume that every stream consists of infinitely many units each carrying an infinitesimal (and thus negligible) amount of flow (traffic). Each such a unit behaves in a selfish way. Intuitively, we expect each unit to be assigned (selfishly) to a server promising minimum cost, taking into account the behavior of other units of flow. Assuming infinitesimal small units of flow, we come to the following fractional variant of the integral assignment model. (This fractional model has been frequently considered in the literature, see, e.g., [8, 24, 25, 26, 27]).

The fractional model does not distinguish between mixed and pure strategies. There are several equivalent ways to define a Nash equilibrium in this model. We use the characterization of Wardrop [29], see also [24]. A fractional assignment is in *Nash equilibrium* if $x_i^j > 0$ implies $C_j \leq C_q$, for every $i \in [n]$ and $j, q \in [m]$. The *coordination ratio* is defined analogously to the integral assignment model, and the *coordination ratio* over a family of cost functions is denoted by R^*.

1.3 Previous Research

The game theoretical integral assignment model for server farms (or parallel links) as described above, has been introduced by Koutsoupias and Papadimitriou [14]. Their focus is on the integral assignment model and *linear cost functions*, that is, functions of the form $f_b(x_1, \ldots, x_n) = \sum_{i=1}^{n} \lambda_i x_i / b$. Koutsoupias and Papadimitriou give first results for the coordination ratio in this model (e.g., tight bounds on the coordination ratio for two links). These results have been later largely extended in [3, 13, 16], and in particular, tight bounds in this model are established by Czumaj and Vöcking in [3]. We are not aware of any results for non-linear cost functions in this model.

Roughgarden and Tardos [24] study the cost of selfish routing in a general network model, where the streams may be required to be routed through a network from given sources to given destinations. They focus mainly on the fractional flow model making the assumption that each stream consists of infinitely many units, each of which behaves in a selfish way. The traffic network parameter to be minimized is the weighted *average-cost* over all streams instead of the min-max cost over all servers. (We will discuss this objective function in a more detail in Section 2.5.) They showed that when the cost functions of the edges are linear then the average cost in a Nash equilibrium is at most $\frac{4}{3}$ times the average cost in an optimal routing. For arbitrary nondecreasing and continuous cost functions, they show the existence of Nash equilibria whose total cost may be arbitrarily larger than the average cost in an optimal routing. On the other hand, they give a bicriteria result that the average cost in Nash equilibrium is upper bounded by the average cost in an optimal routing for twice the amount of flow. The same objective function has been investigated recently by Friedman [8], who studied how the amount of flow influences the bicriteria ratio under cost functions from Queueing Theory.

Besides the selfish flow, Roughgarden and Tardos [24] consider also integral assignments. They give an example of a network with an unbounded bicriteria ratio. They present also a sufficient condition under which the bicriteria ratio is bounded by a certain function: they prove that the bicriteria ratio is upper bounded by $\frac{\alpha}{2-\alpha}$ if data streams are so small that adding any stream to any server increases the total cost at most by a factor of α, where $\alpha \leq 2$. This condition is restricted to pure assignments only. Furthermore, they remark that a useful application of this condition requires families of cost functions with $f_b(0) > \gamma$, for any $b > 0$ and a fixed $\gamma > 0$.

1.4 Outline

Our new results are presented in the following two sections. In Section 2, we will focus on general, monotone cost functions. In Section 3, we will consider families of cost functions from Queueing Theory. Section 4 contains conclusions. In order to have space for a comprehensive discussion of our results, the technical part containing the proofs is moved to the Appendix. (Missing proofs of Theorems 9 and 11 will appear in the full version.)

2. MONOTONE FAMILIES OF COST FUNCTIONS

A cost function f is called *simple* if it depends only on the injected load, that is, if the cost of a server is a function of the sum of the weights mapped to the server but does not depend on other characteristics like, e.g., the session length distribution. (This is a typical assumption in previous works.) A simple cost function is called *monotone* if it is non-negative, continuous, and nondecreasing. For an ordered set B, a family of simple cost functions $\mathcal{F}_B$ is called *monotone* if (i) f_b is monotone for every $b \in B$ and (ii) the cost functions are non-increasing in b, i.e., $f_b(\lambda) \geq f_{b'}(\lambda)$, for every $\lambda \geq 0$ and $0 < b \leq b'$.

2.1 Fractional assignments

Our first result is that all monotone cost functions behave very well under fractional assignments. Recall that R^* denotes the coordination ratio for fractional assignments.

OBSERVATION 1. *For every server farm whose servers are described by monotone cost functions, $R^* = 1$.*

This result follows almost directly from the definition of Nash equilibria specifying that all servers with positive flow at Nash equilibrium must have same cost, which under monotone cost functions implies that all Nash equilibria have the same social cost $C = opt$ (see, e.g., [24, Lemma 2.5]). The observation separates coordination ratios for the min-max objective investigated by Koutsoupias and Papadimitriou [14] from ratios for the average-cost objective by Roughgarden and Tardos [24] and Friedman [8] (see also Section 2.5). In sharp contrast to Observation 1, the results in [8, 14] show that there exist instances of fractional flow on parallel links in which the average-cost coordination ratio is unbounded [24]. This shows that the average-cost and min-max objective can differ arbitrarily under general, monotone cost functions. In the context that we consider here, that of server farms, the min-max objective seems to be the natural choice as it guarantees fairness and efficiency simultaneously.

2.2 Integral assignments

Now let us consider the integral assignment model. We say a coordination ratio R over a family of cost functions $\mathcal{F}_B$ is *bounded* if for every m and every server farm with m servers with cost functions from $\mathcal{F}_B$ there exists $\Gamma > 0$ such that for every set of streams the value of the worst-case Nash equilibrium is at most $\Gamma \cdot opt$. (Observe that Γ might depend on m. Thus, bounded means bounded by a function in m.) Otherwise, the coordination ratio is *unbounded*. Our first result is an exact characterization of those monotone families of cost functions for which the coordination ratio is bounded.

THEOREM 2. *The coordination ratio R over a monotone family $\mathcal{F}_B$ of cost functions is bounded if and only if*

$$\exists \alpha \geq 1 \quad \forall b \in B \quad \forall \lambda > 0 \quad f_b(2\,\lambda) \leq \alpha \cdot f_b(\lambda) \ .$$

Notice that this characterization of bounded vs. unbounded coordination ratios can be applied also to server farms with identical servers. (Recall that such farms are described by families of cost functions consisting only of a single cost function.)

Clearly, for every family of monotone cost functions one can identify a minimum $\alpha \in \mathbb{R}_{\geq 1} \cup \{\infty\}$ fulfilling the conditions specified in the theorem. A natural question is, how does the coordination ratio depend on this minimum α? — In fact, our analysis (see the proof of Theorem 2 and Lemma 14) shows that the coordination ratio is at most $m^{O(\log \alpha)}$. Observe, if the family of cost functions $\mathcal{F}_B$ is assumed to be fixed, then α is a constant or infinity. Thus, we can conclude that for every fixed family of cost functions the coordination ratio is either unbounded or it is polynomially bounded in the number of servers, m.

Let us illustrate the power of the above theorem by investigating some examples. First, we consider families over *polynomial cost functions*, i.e., functions of the form $\sum_{r=0}^{k} a_r \cdot \lambda^r$, for a fixed $k \geq 0$. For these families we can pick $\alpha = a_k \cdot 2^k$ to conclude that here the coordination ratio is bounded. In contrast, there is no such α for *exponential cost functions*, i.e., cost functions for which an additive increase in the load leads to a multiplicative increase in the cost.

COROLLARY 3. *The coordination ratio R for server farms with polynomial cost functions is bounded, whereas the coordination ratio for server farms of (possibly identical) servers with exponential cost functions is unbounded.*

In the next sections we shall discuss several other, practically motivated examples of families of cost functions with unbounded coordination ratios based on well-known formulas from Queueing Theory. We want to point out that unbounded coordination ratios are not only a special phenomenon of cost functions having a pole or an unbounded first derivative. Later in the paper, we will see a practical example of a family of cost functions (based on the Erlang loss formula) that has an unbounded coordination ratio although the functions in this family as well as their first derivatives are bounded above by a small constant, namely one.

2.3 Integral assignments with negligible weights

If we compare the results in Observation 1 and Theorem 2 then we come to the conclusion that integrality can lead to a dramatic performance degradation. As mentioned before, the fractional flow model is assumed to be a simplification that aims to model the situation in which each stream carries only a negligible fraction of the total load [24]. Therefore, let us investigate the relationship between fractional flow and integral assignments of streams with tiny weights more closely. For this purpose we define the notion of "ϵ-small streams." For a server farm with identical servers we have the following simple definition. Stream i is called *ϵ-small* if

$$\lambda_i \leq \frac{\epsilon}{m} \sum_{j \in [m]} \lambda_j ,$$

that is, the stream has at most an $\frac{\epsilon}{m}$-fraction of the overall weight. In the case of servers with different bandwidths, we use the following, slightly more technical definition. Let us fix a server farm and a set of streams with positive weights. Let *opt** denote the minimum maximum cost over all fractional assignments. For a stream $i \in [n]$, define the *scaled stream* i to be a stream with rate $r'_i = r_i/\epsilon$ and session length distribution $\mathcal{D}'_i = \mathcal{D}_i$. (Observe that this implies that the weight of the scaled stream is $\lambda'_i = \lambda_i/\epsilon$.) Then stream i is called *ϵ-small* if, for every $j \in [m]$, the cost of server j is at most *opt** when this server gets assigned the scaled stream i and no other stream. Now, we define R_ϵ to be the worst-case coordination ratio under the restriction that all streams are ϵ-small.

THEOREM 4. *Given any monotone family of cost functions $\mathcal{F}_B$. For every $\epsilon > 0$, the coordination ratio R_ϵ over ϵ-small streams is bounded if and only if the coordination ratio R is bounded.*

A motivation for considering fractional flow instead of integral assignments is that these two models are sometimes assumed to be "essentially equivalent," see, e.g., Remark 2.3 in [24]. Theorem 4 disproves this equivalence for general cost functions. It implies, that there are cost functions with $\lim_{\epsilon \to 0} R_\epsilon = R = \infty$ and $R^* = 1$. Moreover, the instances proving the characterization of unbounded coordination ratios use only pure strategies. Hence, even pure assignments with negligible weights are different from fractional flow.

2.4 Integral assignments under bicriteria measures

It is not surprising that selfish routing can lead to a dramatic cost increase when the cost function has an ∞-pole. In principle, bicriteria measures can be much more informative as they in some sense filter out the extreme behavior of such cost functions at the pole. The following theorem, however, shows that an unbounded coordination ratio R implies a very poor worst-case behavior under bicriteria measures as well. Recall that $\overline{R}$ denotes the bicriteria ratio over integral assignments, that is, $\overline{R}$ specifies by how much the injected rates must be increased to ensure that the worst-case cost in Nash equilibrium will not exceed the optimal cost for the increased rates.

THEOREM 5. *Consider a server farm with m servers with cost functions drawn from a monotone family $\mathcal{F}_B$. If the coordination ratio R is unbounded, then the bicriteria ratio $\overline{R}$ has value at least m, even if all streams are restricted to be ϵ-small.*

The example proving this bad ratio is a server farm of identical servers. In fact, for the case of identical servers one can easily show that a bicriteria ratio of m is the worst possible. This is because a Nash equilibrium cannot be worse than mapping all streams to the same server, and the cost of this extremely unbalanced solution is bounded above by an optimal assignment for an instance with all weights blown up by a factor of m.

2.5 Min-max versus average-cost objective functions

Besides the min-max objective function investigated above, we study also the *average-cost (or total latency) objective function* that has been investigated by Roughgarden and Tardos [24] (see also, e.g., [6, 28] for related results). This objective function aims at minimizing the expected weighted average cost over all streams. Formally, the cost under this objective function is defined by

$$C_{ave} = \frac{1}{\lambda} \sum_{j \in [m]} \mathbf{E}[w_j \cdot C_j] ,$$

and the social optimum is defined by

$$opt_{ave} = \min_{\mathcal{X}} \left(\frac{1}{\lambda} \sum_{j \in [m]} C_j \cdot f_{b_j}(x_1^j, \ldots, x_n^j) \right) ,$$

where $\lambda = \sum_{i \in [n]} \lambda_i$ is the total injected weight and the minimum is taken over all integral assignment matrices. These definitions are equivalent to the respective definitions in the Roughgarden-Tardos model [24] when normalizing λ to one.

We can consider various coordination ratios for the average-cost objective function similarly as for the min-max objective function. These average-cost coordination ratios are defined in the same way as for the min-max model considered before, the only difference is that now one compares the average cost in Nash equilibrium with the average-cost optimum.

THEOREM 6. *In the case of integral assignments, the average-cost objective leads to exactly the same characterizations for coordination and bicriteria ratios as those given in the Theorems 2, 4 and 5 for the min-max objective.*

We want to point out that this equivalence is non-trivial. In general, the behavior under the two objective functions can be quite different. For example, as described in the discussion below Observation 1, in case of fractional assignments, the coordination ratios can be completely opposite under average-cost and min-max objective functions.

3. COST FUNCTIONS FROM QUEUEING THEORY

A typical example of a monotone family of cost functions that is derived from the formula for the expected system time (delay) on an M/M/1 server with injection rate λ and service rate b, namely $\frac{1}{b-\min\{b,\lambda\}}$. Already Koutsoupias and Papadimitriou in their seminal work [14] ask for the price of selfish routing under cost functions of this form. Our characterization of bounded and unbounded coordination ratios given in Theorem 2 immediately implies that the integral coordination ratios for this family of functions are unbounded, which answers the open question from [14]. Of course, this is only one particular example for cost functions from Queueing Theory. Selfish Routing under similar functions have been widely studied, e.g., in [11, 12, 15, 18, 19, 20]. We want to have a closer look at such functions in various server farm models.

We distinguish two general kinds of servers in server farms. A *parallel server* has multiple service channels. Each *service channel* can serve a session independently from other channels. The number of channels on server j corresponds to its bandwidth b_j and all channels serve requests with the same service rate one. Thus, the time a channel needs to serve a session is equal to the length of the session. (Recall that the session lengths in stream i are determined by a probability distribution $\mathcal{D}_i$.) For example, the number of channels may correspond to the number of TCP connections that are allowed to be opened simultaneously on a Web server. In contrast, a *sequential server* has only one service channel. This channel works at service rate b_j. Thus, the time a channel needs to serve a session is equal to the length of the session divided by b_j. We will consider both farms of parallel servers and farms of sequential servers.

Another important aspect that leads to different cost functions is what happens in case of overload. Again, we distinguish two extreme variants, namely, *queueing* and *rejection*. In the *queuing model*, every server maintains an FCFS queue in which it inserts requests when all available channels are in use by other requests. When a service channel finishes a session and the queue is non-empty the server immediately starts serving the next request in the queue. (All channels on the same server share the same queue.) In this model, the objective is to *minimize the maximum expected delay*. In the *rejection model*, blocked requests are rejected and disappear from the system. A natural objective in this model is to *minimize the number of rejected requests* and hence the cost function should describe the fraction of rejected requests.

Using the standard notation from Queueing Theory, a server with k channels that queues requests in case of overload corresponds to a so-called M/$\mathcal{D}$/k/∞ or short M/$\mathcal{D}$/k process, where $\mathcal{D}$ corresponds to the service time (session length) distribution of the injected request stream. When requests are rejected then the server is described by a so-called M/$\mathcal{D}$/k/k process.

3.1 Queueing systems without rejection

In order to avoid discussions about what is exactly the right queueing model (e.g., M/M/1, M/D/1, M/G/1, M/G/c, ...) and what is exactly the right cost to consider (e.g., expected waiting time or expected system time), let us introduce a generic concept of "monotone queueing functions." A monotone cost function is called a *monotone queueing function* if it satisfies $\lim_{\lambda\to b^-} f_b(\lambda) = \infty$. (Observe that the monotonicity implies $f_b(\lambda) = \infty$, for every $\lambda \geq b$). This assumption is motivated by the fact that the expected waiting time as well as the expected system time in every queueing process without rejection goes to infinity when the injection rate approaches the service rate (or bandwidth) of the server. Clearly, an immediate consequence of the ∞-pole is that the parameter α introduced in Theorem 2 is ∞. Thus, by Observation 1 and Theorems 2 and 5, we obtain the following corollary.

COROLLARY 7. *For every family $\mathcal{F}_{\mathbb{R}_{>0}}$ of monotone queueing functions, $R^* = 1$, $R = \infty$, and $\overline{R} \geq m$, even under the restriction that all streams are ϵ-small.*

The proof for this negative result uses a Nash equilibrium in which all streams are identical and the total injected load $\sum_{i\in[n]} \lambda_i$ is less than the bandwidth of a single server. Thus, selfish routing can lead to a catastrophic performance degradation even under bicriteria measures in extremely lightly loaded cases.

Recall that the instances proving the unbounded coordination ratio R are constructed using only pure strategies. However, the bad instances for the bicriteria ratio $\overline{R}$ that we have seen until now use mixed strategies. This motivates us to investigate whether the randomness introduced due to the choice of mixed strategies is the only source of troubles under bicriteria measures. The following theorem demonstrates that bicriteria ratios can also be poor in case of pure strategies only.

THEOREM 8. *Let $\mathcal{F}_{\mathbb{R}_{>0}}$ be any family of monotone queueing functions. Suppose m is the number of servers and there exists $b > \sqrt{m}$ such that $f_b\left(\left(1-\frac{b}{m}\right)\cdot b\right) < f_1\left(\frac{b}{m}\right)$, where $f_1, f_b \in \mathcal{F}_{\mathbb{R}_{>0}}$. Then, the bicriteria ratio over pure strategies is at least $\frac{m}{2\,b}$.*

This theorem needs some explanation. For example, the cost function for M/M/1 waiting time is $\frac{\lambda}{b(b-\min\{b,\lambda\})}$ and the cost function for M/M/1 system time is $\frac{1}{b-\min\{b,\lambda\}}$. If we assume the cost function for waiting time then the theorem implies a bicriteria ratio over pure strategies of $\Omega(m^{1/3})$. Similarly, for system time the bicriteria ratio is $\Omega(m^{1/2})$. In both cases, the total injected load in the example that gives these bad results is very small. We investigate this closer, and show that the $\Omega(m^{1/3})$ bound for M/M/1 waiting time is essentially tight, proving the following theorem.

THEOREM 9. *Let us consider the integral allocation model and cost function on a server being the M/M/1 waiting time. The bicriteria ratio with pure strategies, $\overline{R}$, in this model has value at most $O(m^{1/3}\log(m))$, where m is the number of servers.*

Summarizing, even for pure strategies and under a small total injection rate, the slowdown due to the lack of coordination can be dramatic.

An alternative bicriteria measure. For the family of monotone queueing functions there is another interesting bicriteria measure. It is a very natural question to ask by how much one has to decrease the bandwidths of the servers such that an optimal assignment under the decreased bandwidths is at least as expensive as a

Nash equilibrium for the original system. Let $\overline{R}_{\rm bw}$ denote the corresponding worst-case bicriteria ratio.

It turns out that for most functions from Queueing Theory, the effect of changing the bandwidths is larger than the effect of changing the injection rate. In fact, most of these functions show *super-linear scaling*, i.e., $f_b(\lambda) \leq \frac{1}{\alpha} \cdot f_{b/\alpha}(\lambda/\alpha)$, for every $\lambda \in [0, b)$ and $\alpha \geq 1$. Applying this property, we can determine the bicriteria bandwidth ratio exactly.

THEOREM 10. *For every family $\mathcal{F}_B$ of monotone queueing functions with super-linear scaling, $\overline{R}_{\rm bw} = m$, where m is the number of servers.*

We want to emphasize that this theorem gives tight results, e.g., for expected waiting time or system time in the queueing systems M/M/1, M/D/1, M/G/1, M/M/c or M/G/1.

3.2 Queueing under heterogeneous traffic

Until now we implicitly assumed homogeneous traffic, i.e., all streams have the same (general) session length distribution. However, several practical studies show that Internet traffic is far away from being homogeneous, see, e.g., [2, 5, 22]. Following these studies, one has to take into account different session lengths distributions.

The *Pollaczek-Khinchin* (P-K) formula (see, e.g., [10]) describes expected waiting time in M/G/1 queues, that is, the expected delay of requests on sequential servers under *heterogeneous* traffic with arbitrary service time distributions. We can use this formula and transform it into a family of cost functions depending only on two parameters, namely, the weight λ and the variance V, of the combined streams injected into server. Let us describe this in a more detail. Suppose every stream i is characterized by two weights λ_i and V_i corresponding to the *expected load* (i.e., the number of bytes requested per unit of time) and the *variance of the load*. Then, the P-K cost function family $\mathcal{F}_{\mathbb{R}_{>0}}$ can be defined as follows:

$$f_b(x_1, \ldots, x_n) \;=\; \frac{\sum_{i=1}^n V_i\, x_i}{b\left(b - \sum_{i=1}^n \lambda_i\, x_i\right)} \quad .$$

The remarkable fact here is that both parameters, the expected load and the variance, can be aggregated independently in a simple linear fashion. That is, the expected load injected into the server is $\lambda = \sum_{i=1}^n \lambda_i\, x_i$ and the variance of this load is $V = \sum_{i=1}^n V_i\, x_i$.

Observe that if we assume $\lambda_i = V_i$ then we are back in the homogeneous model with identical session length distribution and we obtain a monotone queueing function with only one parameter, λ. Consequently $R = \infty$ and $\overline{R} \geq m$ for the P-K cost function family. In the fractional flow model, however, we will come to different results. Recall that $R^* = 1$ under homogeneous traffic.

THEOREM 11. *The coordination ratio R^* for the P-K cost function family under heterogeneous traffic is unbounded. If the ratio between the bandwidth of the fastest and slowest server is restricted to be at most S then $R^* = S$.*

We conclude that the optimality of fractional flow in Nash equilibrium is a special property of homogeneous traffic on parallel links, and hence one must take into account the heterogeneous nature of Web traffic when studying the price of selfish routing in the Internet.

3.3 Servers with parallel channels and rejection

Until now, we assumed that all requests are served, regardless of how long they have to wait for service. In practice, however, Web servers reject requests when they are overloaded. For simplicity, let us assume that a server rejects requests whenever all service channels are occupied and then these requests disappear from the system. In this case, the fraction of rejected requests is completely independent of the service time distribution. In other words, there is no difference between homogeneous and heterogeneous traffic under this service model. In fact, the fraction of rejected requests can be derived from the *Erlang loss formula*, see, e.g., [8, 9]. We obtain the following cost function family $\mathcal{F}_{\mathbb{N}_{>0}}$ for servers that can open up to b channels simultaneously:

$$f_b(x_1, \ldots, x_n) \;=\; \frac{\lambda^b/b!}{\sum_{k=0}^b \lambda^k/k!} \qquad \text{with } \lambda = \sum_{i=1}^n \lambda_i\, x_i \quad .$$

On the first glance, the family of Erlang loss functions makes an innocent impression. Indeed, these functions are continuous, convex, monotonically increasing in λ and $f_b'(\lambda) \leq 1/b$, for every $\lambda \geq 0$. (Hence $R^* = 1$.) Nevertheless, the following corollary shows that the coordination ratio R for integral assignments is unbounded.

COROLLARY 12. *For the family $\mathcal{F}_{\mathbb{N}_{>0}}$ of Erlang loss cost functions, $R = \infty$ and $\overline{R} \geq m$.*

The corollary follows from Theorems 2 and 5, as $\alpha = \infty$. This can be seen as follows. Let us consider the family of functions $\mathcal{F}_{\mathbb{N}_{>0}}$ with $F_b(x) = f_b(b\,x)$, i.e., the Erlang loss functions in terms of relative load. We observe that for every $x \geq 0$,

$$\lim_{b\to\infty} F_b(x) = \max\left\{0, \frac{x-1}{x}\right\} \quad .$$

Thus, the limit of these functions behaves in a very extreme way at $x = 1$, where the cost suddenly increases by an unbounded factor. This implies that $\alpha = \infty$.

In contrast to the monotone queueing functions from Section 3.1, however, the source of the troubles is not an ∞-pole, but the rapid increase from cost $f_b((1-\epsilon)b) \approx \exp(-b\epsilon^2)$ to cost $f_b((1+\epsilon)b) \approx \epsilon$, or in other words, the rapid increase from tiny to small cost. Hence, one might hope that the absolute cost of selfish routing under the Erlang loss cost function family is small. In fact, this is confirmed by the following theorem.

THEOREM 13. *Let δ satisfy $\delta \geq 2/\log_2 m$. Consider a server farm of m servers with bandwidths $b_1 \geq \cdots \geq b_m$ and cost functions from the family $\mathcal{F}_{\mathbb{N}_{>0}}$ of Erlang loss cost functions. Suppose $\sum_{i\in[n]} \lambda_i \leq \frac{1}{6e} \sum_{j\in[m]} b_j$ and $\max_{i\in[n]} \lambda_i \leq \frac{b_m}{3\delta \log_2 m}$. Then, any Nash equilibrium has social cost at most $m^{-\delta+1} + m \cdot 2^{-b_m/4}$.*

Hence, if the total injected load is at most a constant fraction of the total bandwidth and every stream has not too large weight, that is, streams are $O(\frac{1}{\log m})$-small, then the fraction of rejected requests is at most $m^{-\delta+1} + m\,2^{-b_m/4}$, assuming constant δ. Under the same conditions, an optimal assignment would reject a fraction of $2^{-\Theta(b_1)}$ packets. Taking into account that typical Web servers can open several hundred TCP connections simultaneously, so that b_m can be assumed to be quite large, we conclude that the cost of selfish routing is very small in absolute terms, even though the coordination ratio comparing this cost with the optimal cost is unbounded.

4. CONCLUSIONS

In this paper we present the first thorough theoretical study of the price of selfish routing in server farms for general cost function. In our investigations we paid special attention on cost functions from Queueing Theory. Our results have some important algorithmic consequences in these models. They show that the choice of the queueing discipline should take into account the possible performance degradation due to selfish and uncoordinated behavior of network users.

We have shown that the coordination ratio for queueing systems without rejection is unbounded. The same is true for server farms that reject requests in case of overload. However, there is a fundamental difference between these two kinds of queueing policies. Because of the infinity pole, the delay under selfish routing in the queuing systems without rejection is in general unbounded. In fact, we have explicitly shown that the selfish routing in such queueing systems can lead to an arbitrary large delay even when the total injected load can potentially be served by a single server. In contrast, the fraction of rejected requests under selfish routing can be bounded above by a function that is exponentially small in the number of TCP connections that can be opened simultaneously.

We conclude that server farms that serve all requests, regardless of how long requests have to wait, cannot give any reasonable guarantee on the quality of service when selfish agents manage the traffic. However, if requests are allowed to be rejected, then it is possible to guarantee a high quality of service for every individual request stream. Thus, the typical practice of rejecting requests in case of overload is a necessary condition to ensure efficient service under game theoretic measures.

5. REFERENCES

[1] M. Beckmann, C. B. McGuire, and C. B. Winston. *Studies in the Economimices of Transportation*. Yale University Press, 1956.

[2] M. E. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6):835–846, 1997.

[3] A. Czumaj and B. Vöcking. Tight bounds for worst-case equilibria. In *Proc. 13th ACM-SIAM SODA*, 2002.

[4] F. Douglis, A. Feldmann, B. Krishnamurthy, and J. Mogul. Rate of change and other metrics: a live study of the World Wide Web. In *Proc. USENIX Symposium on Internet Technologies and Systems*, pp. 147–158, December 1997.

[5] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger. Dynamics of IP traffic: A study of the role of variability and the impact of control. In *Proc. ACM SIGCOMM'99*, pp. 301–313, 1999.

[6] M. Frank. The Braess paradox. *Mathematical Programming Study*, 20:283–302, 1981.

[7] P. Franken, D. Konig, U. Arndt, and V. Schmidt. *Queues and Point Processes*. Wiley, Chichester, 1982.

[8] E. Friedman. A generic analysis of selfish routing. Manuscript, 2001.

[9] D. Gross and C. M. Harris. *Queueing Theory. Third Edition* John Wiley & Sons, New York, NY, 1998.

[10] L. Kleinrock. *Queueing Systems. Volume I: Theory.* John Wiley & Sons, New York, NY, 1975.

[11] Y. A. Korilis, A. A. Lazar, and A. Orda. Capacity allocation under noncooperative routing. In *IEEE Transactions on Automatic Control*, 42(3):309–325, 1997.

[12] Y. A. Korilis, A. A. Lazar, and A. Orda. Avoiding the Braess paradox in noncooperative networks. In *Journal of Applied Probability*, 36(1):211–212, 1999.

[13] E. Koutsoupias, M. Mavronicolas, and P. Spirakis. Personal communication, 2001.

[14] E. Koutsoupias and C. H. Papadimitriou. Worst-case equilibria. In *Proc. 16th STACS*, pp. 404–413, 1999.

[15] A. A. Lazar, A. Orda, and D. E. Penderakis. Virtual path bandwidth allocation in multiuser networks. In *IEEE/ACM Transactions on Networking*, 5:861–871, 1997.

[16] M. Mavronicolas and P. Spirakis. The price of selfish routing. In *Proc. 33rd ACM STOC*, pp. 510–519, 2001.

[17] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, New York, NY, 1995.

[18] A. Orda, R. Rom, and N. Shimkin. Competitive routing in multi-user communication networks. In *IEEE/ACM Transactions on Networking*, 1:510–521, 1993.

[19] C. H. Papadimitriou. Algorithms, games, and the Internet. In *Proc. 33rd ACM STOC*, pp. 749–753, 2001.

[20] C. H. Papadimitriou. Game theory and mathematical economics: A theoretical computer scientist's introduction (Tutorial). In *Proc. 42th IEEE FOCS*, pp. 4–8, 2001.

[21] C. H. Papadimitriou and M. Yannakakis. On complexity as bounded rationality. In *Proc. 26th ACM STOC*, pp. 726–733, 1994.

[22] K. Park, G. Kim, and M. E. Crovella. On the relationship between file sizes, transport protocols, and self-similar network traffic. In *Proc. IEEE International Conference on Network Protocols*, pp. 171–180, 1996.

[23] V. Paxson and S. Floyd. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.

[24] T. Roughgarden and É. Tardos. How bad is selfish routing? In *Proc. 41st IEEE FOCS*, pp. 93–102, 2000.

[25] T. Roughgarden. Stackelberg scheduling strategies. In *Proc. 33rd ACM STOC*, pp. 104–113, 2001.

[26] T. Roughgarden. Designing networks for selfish users is hard. In *Proc. 42nd IEEE FOCS*, pp. 472–481, 2001.

[27] T. Roughgarden. How unfair is optimal routing? In *Proc. 13th ACM-SIAM SODA*, 2002.

[28] R. Steinberg and W. I. Zangwill. The prevalence of Braess' paradox. *Transportation Science*, 17(3):301–318, 1983.

[29] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proc. Institution of Civil Engineers*, Part II, volume 1, pp. 325–362, 1952.

APPENDIX

A. PROOF OF THEOREMS 2 AND 4

We prove Theorems 2 and 4 in a single proof using two lemmas. The first lemma proves sufficient conditions for a bounded coordination ratio, while the second lemma proves necessary conditions.

LEMMA 14. *Suppose we are given a server farm with m servers having cost functions from a fixed, monotone family $\mathcal{F}_B$ satisfying*

$$\exists \alpha \geq 1 \quad \forall b \in B \quad \forall \lambda > 0 \quad f_b(2\,\lambda) \;\leq\; \alpha f_b(\lambda) \enspace .$$

Then, for every set of streams the worst-case cost over all Nash equilibria is upper bounded by $opt \cdot m^{O(1)}$.

PROOF. Fix an arbitrary allocation in Nash equilibrium. Let C denote the cost of this allocation. We will show that $C \leq m \cdot \alpha^{\lceil \log m \rceil} \cdot opt = m^{O(1)} \cdot opt$. First, we observe that

$$f_b(s\,\lambda) \leq \alpha^{\lceil \log s \rceil} \cdot f_b(\lambda) \enspace ,$$

for every $s \geq 1$. Let $X := \sum_{i \in [n]} \lambda_i$ denote the total injected load. Let x denote the load on server 1 (with the biggest bandwidth) under an *optimal fractional allocation*, i.e., an allocation with minimum maximum cost over all servers assuming that streams can be split arbitrarily. Without loss of generality, we assume that server 1 has the maximum load over all servers, and hence $X \leq m\,x$. (Recall that since server 1 is the server with the biggest bandwidth and each function $f_b \in \mathcal{F}_B$ is nondecreasing in λ and nonincreasing in b, there exists an optimal fractional allocation with maximum load for server 1.) In this way,

$$f_{b_1}(X) \ \leq \ f_{b_1}(m\,x) \ \leq \ \alpha^{\lceil \log m \rceil} \cdot f_{b_1}(x) \ \leq \ \alpha^{\lceil \log m \rceil} \cdot opt \ .$$

Let M denote the set of servers $j \in [m]$ with $\sum_{i \in [n]} p_i^j > 0$. Pick any server j in M. Let $i \in [n]$ denote a stream with $p_i^j > 0$. Then, the Nash equilibrium property guarantees that $c_i^j \leq c_i^1$. Hence, for every $j \in M$,

$$\mathbf{E}\left[C_j\right] \leq \mathbf{E}\left[C_j | x_i^j = 1\right] = c_i^j \leq c_i^1 \leq f_{b_1}(X) \leq \alpha^{\lceil \log m \rceil} \cdot opt \ .$$

Furthermore, for $j \in [m] \setminus M$, we have $\mathbf{E}\left[C_j\right] = f_{b_j}(0) \leq opt \leq \alpha^{\lceil \log m \rceil} opt$. As a consequence,

$$C = \mathbf{E}\left[\max_{j \in [m]} C_j\right] \leq m \cdot \max_{j \in [m]} \mathbf{E}\left[C_j\right] \leq m\,\alpha^{\lceil \log m \rceil}\, opt = m^{O(1)}\, opt,$$

which completes the proof of Lemma 14. □

Now, we prove a sufficient condition for an unbounded coordination ratio. Observe that the negation of the property gives the sufficient condition for a bounded coordination ratio as specified in Theorem 2.

LEMMA 15. *Let $\epsilon > 0$ be chosen arbitrarily. Suppose we are given a server farm with only two identical servers, each with the same monotone cost function f satisfying*

$$\forall \alpha \geq 1 \quad \exists \lambda > 0 \quad f(2\,\lambda) > \alpha \cdot f(\lambda) \ .$$

Then, for every $\Gamma \geq 1$, there exists a pure Nash equilibrium over ϵ-small streams with cost $C > \Gamma \cdot opt$.

PROOF. First, let us show that the above property of the function f implies that

$$\forall \alpha \geq 1 \quad \beta > 0 \quad \exists \lambda > 0 \quad f((1+\beta)\cdot\lambda) > \alpha \cdot f(\lambda) \ . \qquad (1)$$

Indeed, if we consider the negations of the two statements then we observe that $\exists \alpha \geq 1, \beta > 0 \quad \forall \lambda > 0 \quad f((1+\beta)\,\lambda) \leq \alpha\, f(\lambda)$ implies that $\exists \alpha' \geq 1 \quad \forall \lambda > 0 \quad f(2\,\lambda) \leq \alpha' f(\lambda)$, e.g., we can set $\alpha' = \alpha^{\lceil 1/\log(1+\beta) \rceil}$.

Now, consider a server farm with two identical servers, each with the same monotone cost function f satisfying condition (1). For the purpose of contradiction, assume there exists Γ with $C \leq \Gamma \cdot opt$ for every Nash equilibrium over streams of maximum weight ϵ. Therefore, by (1), there exists $\lambda > 0$ with $f((1+\epsilon/10)\,\lambda) > \Gamma \cdot f(\lambda)$. Using this assumption, we will define a Nash equilibrium over streams of maximum weight $\epsilon > 0$ with cost $C > \Gamma \cdot opt$.

First, let us consider streams of identical weight $w \in [\lambda\epsilon/2, \lambda\epsilon]$ and assign them to the servers so that the cost on each server is exactly $f(\lambda)$. Let τ be the number of streams per server in this allocation. Now, let us slightly change the instance by taking two streams, one from each server, and "break" each of them into two smaller streams, one of weight $\frac{3}{5}\,w$, the other of weight $\frac{2}{5}\,w$. It is easy to see that the optimal allocation for this instance has cost $opt = f(\lambda)$.

Let us consider a different allocations of the streams to the servers. We assign $\tau - 1$ streams of weight w and two streams of weight $\frac{3}{5}\,w$ to the first server and the remaining streams to the second server. In this way, the first server has cost $f(\lambda + \frac{1}{5}\,w)$ whereas the second server has cost $f(\lambda - \frac{1}{5}\,w)$. This allocation defines a Nash equilibrium, because the streams have minimum weight $\frac{2}{5}\,w$ and therefore there is no incentive for any of them to change its strategy. The cost of this Nash equilibrium, however, is

$$C \ = \ f \cdot \left(\lambda + \frac{w}{5}\right) \ \geq \ f \cdot \left(\lambda + \frac{\lambda \cdot \epsilon}{10}\right) \ > \ \Gamma \cdot f(\lambda) \ = \ \Gamma \cdot opt \ .$$

Clearly this contradicts our initial assumption that $C \leq \Gamma \cdot opt$ for any Nash equilibrium over streams of maximum weight ϵ. This completes the proof of Lemma 15. □

Theorems 2 and 4 follow immediately from Lemmas 14 and 15. □

B. PROOF OF THEOREM 5

Let n be the number of streams and let $\alpha = m^{n-1}$. Assume, that n is such that $\frac{m}{n} \leq \epsilon$. Since ratio R is unbounded, Theorem 2 implies that there exist $f_b \in \mathcal{F}_B$ and $\lambda > 0$ with $f_b(2\lambda) > \alpha f_b(\lambda)$.

Assume that we have m identical servers, each with bandwidth $b_j = b$ ($j \in [m]$), and a set of n identical data streams, each having weight of $\lambda_i = \frac{\lambda m}{\Gamma n}$ ($i \in [n]$), where $\Gamma > 0$ will be specified later. Define the probabilities p_j^i as $p_j^i = \frac{1}{m}$, for each $i, j \in [m]$. These probabilities define a Nash equilibrium, since all the expected costs c_j^i have the same value.

Let us fix a server j. The probability that all the data streams are assigned to server j is $\Pi_i p_j^i = m^{-n}$. In this case, the cost on server j is $f_b\left(\frac{\lambda m}{\Gamma}\right)$, since we have n streams, each of weight $\frac{\lambda m}{\Gamma n}$, and so their total weight is $\frac{\lambda m}{\Gamma n} n = \frac{\lambda m}{\Gamma}$. Therefore, with probability m^{-n}, the cost on a particular server j is at least $f_b\left(\frac{\lambda m}{\Gamma}\right)$, and thus also $\max_j C_j \geq f_b\left(\frac{\lambda m}{\Gamma}\right)$, over all $j \in [m]$. Additionally, these events corresponding to different servers are pairwise disjoint. Using these observations, we can estimate $C = \mathbf{E}\left[\max_j C_j\right]$ as

$$\mathbf{E}\left[\max_j C_j\right] \geq m \cdot f_b\left(\frac{\lambda m}{\Gamma}\right) \cdot m^{-n} = \frac{f_b\left(\frac{\lambda m}{\Gamma}\right)}{m^{n-1}} \ .$$

We want to show that if $C \leq opt_\Gamma$ then $\Gamma \geq m$. We proceed by contradiction and let us assume there is Γ such that $C \leq opt_\Gamma$ and $\Gamma < m$. Then, we obtain

$$\frac{f_b\left(\frac{\lambda m}{\Gamma}\right)}{m^{n-1}} \leq opt_\Gamma \leq f_b\left(\frac{\lambda m}{\Gamma n} \cdot \frac{n}{m} \cdot \Gamma\right) = f_b\left(\lambda\right) \ ,$$

where the last inequality follows by observing that the value of opt_Γ is at most the value of a solution in which we assign $\frac{n}{m}$ (we can assume that $\frac{n}{m}$ is a positive integer) data streams to each server, after blowing up each data stream by Γ. Then, we obtain

$$f_b\left(\frac{\lambda m}{\Gamma}\right) \leq m^{n-1} \cdot f_b\left(\lambda\right) \ . \qquad (2)$$

By our assumption, $\forall \alpha \geq 1 \exists \lambda > 0 f_b(2\lambda) > \alpha f_b(\lambda)$. This, by the argument from the proof of Lemma 15, yields $\forall \alpha \geq 1, \beta > 0 \ \ \exists \lambda > 0 \ \ f_b((1+\beta)\lambda) > \alpha f_b(\lambda)$. Now, we plug $\alpha = m^{n-1}$, $1 + \beta = \frac{m}{\Gamma}$, to obtain a contradiction with inequality (2).

Since all the servers are identical, a data stream of weight λ_i is ϵ-small if $\lambda_i \leq \epsilon \frac{\sum_j \lambda_j}{m}$. Since in our case $\lambda_i = \frac{\lambda m}{\Gamma n}$ for all $i \in [n]$, the last condition is equivalent to $\frac{\lambda m}{\Gamma n} \leq \epsilon \frac{\lambda}{\Gamma}$, which gives $\frac{m}{n} \leq \epsilon$. The last inequality is true by the choice of n, and so our data streams are ϵ-small. □

C. PROOF OF THEOREM 6

Fix a monotone family $\mathcal{F}_B = \{f_b | b \in B\}$ of cost functions. First let us show that the coordination ratio under average-cost objective, denoted by Σ, for this family is bounded if the coordination ratio R is bounded. Clearly, if R is bounded then we obtain from Theorem 2

$$\exists \alpha \geq 1 \quad \forall b \in B \quad \forall \lambda > 0 \quad f_b(2\lambda) \leq \alpha f_b(\lambda) \ .$$

We will show that Σ is bounded provided this property is given. Fix any Nash equilibrium with probabilities p_i^j, $i \in [n]$, $j \in [m]$. We have to give an upper bound on the ratio between the expected total latency given by the probabilities p_i^j, on one hand, and the optimal total latency, on the other hand. Let $w = w_1, \ldots, w_m$ denote a vector of random variables with w_j describing the injected load of server j. Let w^* denote a corresponding load vector of an optimal allocation that minimizes the total latency. We have to show that there exists $\Gamma \geq 1$ such that

$$\mathbf{E}\left[\sum_{j\in[m]} w_j f_{b_j}(w_j)\right] \leq \Gamma \sum_{j\in[m]} w_j^* f_{b_j}(w_j^*) \ .$$

Define $X = \sum_{j\in[m]} w_j = \sum_{j\in[m]} w_j^*$. Then $f_{b_1}(X)$ corresponds to the cost that is obtained by assigning all the load to the fastest server. Using the same arguments as in the proof of Lemma 14, we obtain $\mathbf{E}\left[f_{b_j}(w_j)\right] \leq f_{b_1}(X)$, for every $j \in [m]$, such that

$$\begin{aligned}\mathbf{E}\left[\sum_{j\in[m]} w_j f_{b_j}(w_j)\right] &\leq \sum_{j\in[m]} X\, f_{b_1}(X) \\ &\leq m^2\, \alpha^{\lceil 1+\log m\rceil} \frac{X}{m}\, f_{b_1}\!\left(\frac{X}{m}\right) \\ &\leq m^2\, \alpha^{\lceil 1+\log m\rceil} \sum_{j\in[m]} w_j^* f_{b_j}(w_j^*) \ .\end{aligned}$$

The second inequality follows from our assumptions on the family of cost functions. The third inequality needs some more explanation. Observe that there exists $j \in [m]$ with $w_j^* \geq X/m$. Applying monotonicity of our cost function family, we obtain $\frac{X}{m} f_{b_1}(\frac{X}{m}) \leq w_j^* f_{b_j}(w_j^*)$, which yields the inequality. As a consequence, $\Sigma \leq m^2\, \alpha^{\lceil 1+\log m\rceil}$, that is, Σ is bounded.

It remains to show that an unbounded coordination ratio R implies an unbounded ratio Σ. If R is unbounded then the family of cost functions satisfies $\forall \alpha \geq 1\ \exists b \in B\ \exists \lambda > 0\ f_b(2\lambda) > \alpha f_b(\lambda)$. In Lemma 15, we described an instance with two identical servers and a set of ϵ-small streams such that both servers have identical load in an optimal allocation but there is a Nash equilibrium in which one of the servers receives more load than the other server. Using the above property we show that if one server has load at least $(1+\epsilon/10)$ times average load then the cost on this server can deviate by an arbitrary large factor from the optimal cost, which then proves that R is unbounded. A straightforward adaption of these arguments show also that if the above condition is fulfilled then Σ is unbounded. The proof of Theorem 5 in the average-cost case is basically the same as in the min-max case, since the values of the two objectives are the same in the lower bound instance. This completes the proof of Theorem 6. □

D. PROOF OF THEOREM 8

We prove the following lemma that directly implies Theorem 8.

LEMMA 16. *Let $\mathcal{F}_{\mathbb{R}_{>0}}$ be a family of monotone queueing functions. Assume we are given $m+2$ servers with bandwidths $1 = b_1 = b_2 = \ldots = b_m \leq b_{m+1} \leq b_{m+2} = b$, such that $f_{b_j} \in \mathcal{F}_{\mathbb{R}_{>0}}$, for each $j = 1, 2, \ldots, m+2$. Assume, moreover, that there exists an $\varepsilon > 0$ such that functions f_1 and f_b fulfill*

$$f_b\left(\left(1-\frac{1}{2\Gamma}\right)b+\varepsilon\right) \leq f_1\left(\frac{1}{2\Gamma}\right) \ , \qquad (3)$$

where $\Gamma = \frac{m}{2b}$ and $\frac{b}{2\Gamma} - \varepsilon/2 > 1$. Then, in this system, $C > opt_\Gamma$, where C is the maximum value over all Nash equilibria assuming only pure strategies. In particular, $\overline{R} \geq \frac{m}{2b}$.

PROOF. We define the following instance of the problem. Let the servers $1, 2, \ldots, m$ be called slow. We also have a fast server with $b_{m+2} = b$ and an additional server with bandwidth $b_{m+1} = \frac{b}{2\Gamma}$. Assume that each slow server holds one small data stream with weight $\frac{1}{2\Gamma}$ and let the fast server have one large data stream with weight $\frac{b}{2\Gamma} - \varepsilon/2$. The additional server holds no data stream.

In the instance we have just defined, each slow server has cost (after blowing the streams up by Γ) $f_1\left(\frac{1}{2}\right)$ and the fast server has cost $f_b\left(\frac{b}{2} - \frac{\varepsilon\Gamma}{2}\right)$. Let

$$cost(\varepsilon) = \max\left\{f_{1/\Gamma}(1/(2\,\Gamma)), f_{b/\Gamma}(b/(2\,\Gamma) - \varepsilon/2)\right\} \ ,$$

and observe that by the definition of monotone queueing functions, $cost(\varepsilon)$ is finite even if $\varepsilon \to 0$, and obviously $opt_\Gamma \leq cost(\varepsilon)$.

Let us define a Nash equilibrium for our instance. We assign to the fast server a total amount of $\left(1 - \frac{1}{2\Gamma}\right)b + \varepsilon$ of small streams. Observe that we have enough small streams to achieve this, since their total size is equal to $\frac{m}{2\Gamma} > \left(1 - \frac{1}{2\Gamma}\right)b$ (notice that if ε is not small enough, then we can use a smaller ε still satisfying (3) and $\frac{b}{2\Gamma} - \varepsilon/2 > 1$ guaranteed by the monotonicity of function f_b). One can also easily show that the large stream must be assigned to the additional server $m+1$.

We claim we have defined a Nash equilibrium. First, none of the small streams would go from the fast server to a slow server, since by (3) we have $f_b\left(\left(1 - \frac{1}{2\Gamma}\right)b + \varepsilon\right) \leq f_1\left(\frac{1}{2\Gamma}\right)$. Also, none of the small streams would go from the fast server to the additional server, since the remaining space on the additional server is $\varepsilon/2$ and it can be made arbitrarily small. Finally, the large stream cannot go from the additional server to any slow server (since $\frac{b}{2\Gamma} - \varepsilon/2 > 1$) nor to the fast server (since it would exceed the capacity on the fast server). (Notice that it is possible that a small remaining data stream, if any, can go from a slow server to the fast server. But then our construction works as well.)

The cost on the additional server in this Nash equilibrium is $f_{b/(2\Gamma)}\left(\frac{b}{2\Gamma} - \varepsilon/2\right)$ and so by the properties of monotone queueing functions, $\lim_{\varepsilon\to 0} f_{b/(2\Gamma)}\left(\frac{b}{2\Gamma} - \varepsilon/2\right) = \infty$. This implies that there exists an $\varepsilon > 0$, such that C is arbitrarily large, and in particular larger than $cost(0)$. By monotonicity, we have $cost(0) \geq cost(\varepsilon)$. Thus we obtain $opt_\Gamma \leq cost(\varepsilon) \leq cost(0) < C$. □

E. PROOF OF THEOREM 10

Let OPT_Γ denote the optimum solution in a system where bandwidths of all servers are slowed down by a factor of Γ.

LEMMA 17. *Fix an arbitrary monotone queueing function $f = f_b$. Consider a server farm of m identical servers with cost function f. Then for every $\epsilon > 0$ and $\Gamma < m$ there exists a Nash equilibrium over ϵ-small streams such that $C > OPT_\Gamma$.*

PROOF. Assume we have m identical servers, each with bandwidth $b_j = b$, and a set of n identical data streams, each of weight $\lambda_i = \frac{b-\delta}{n}$, where n is s.t. $\frac{b}{n} \leq \epsilon$, and $\delta \in [0, b)$ will be specified later. Let $p_j^i = \frac{1}{m}$, for each $i, j \in [m]$. Since all the expected costs c_j^i have the same value, p_j^i's define a Nash equilibrium.

Fix a server $j \in [m]$. The probability that all data streams are assigned to server j is $\Pi_i p_j^i = m^{-n}$. In this case, the cost on server j is $f_b(b-\delta)$, since the total weight of the streams is $\frac{b-\delta}{n}n = b-\delta$.

Therefore, with probability m^{-n}, the cost on a particular server j is at least $f_b(b-\delta)$, and thus also $\max_{j\in[m]} C_j$ is at least $f_b(b-\delta)$. Additionally, the random events corresponding to different servers are pairwise disjoint. Using these observations we obtain

$$C = \mathbf{E}\left[\max_j C_j\right] \geq m \cdot f_b(b-\delta)\cdot m^{-n} = \frac{f_b(b-\delta)}{m^{n-1}} \ .$$

We want to show that if $C \leq OPT_\Gamma$, then $\Gamma \geq m$. Assume towards a contradiction that $C \leq OPT_\Gamma$ and $\Gamma < m$. Then we have

$$\frac{f_b(b-\delta)}{m^{n-1}} \leq OPT_\Gamma \leq f_{b/\Gamma}\left(\frac{b-\delta}{n}\cdot\frac{n}{m}\right) = f_{b/\Gamma}\left(\frac{b-\delta}{m}\right) \ ,$$

where the last inequality follows by observing that the value of OPT_Γ is at most the value of a solution in which we assign $\frac{n}{m}$ (we can assume that $\frac{n}{m}$ is a positive integer) data streams to each server, after slowing the server down by Γ. Then, by the last inequality and by continuity of functions f_b and $f_{b/\Gamma}$, we have

$$\lim_{\delta\to 0}\frac{f_b(b-\delta)}{m^{n-1}} \leq \lim_{\delta\to 0} f_{b/\Gamma}\left(\frac{b-\delta}{m}\right) = f_{b/\Gamma}\left(\frac{b}{m}\right) \ .$$

By our assumption, the RHS is finite, but the LHS is infinite by the properties of function f_b. Therefore, there exists a small enough $\delta > 0$, such that $\frac{f_b(b-\delta)}{m^{n-1}} > f_{b/\Gamma}\left(\frac{b}{m}\right) \geq f_{b/\Gamma}\left(\frac{b-\delta}{m}\right)$ (by monotonicity). This yields a contradiction. □

We now prove Theorem 10. Lemma 17 implies that $\overline{R} \geq m$. Therefore, we have to show that $C \leq OPT_m$. Assume that b_j, $j \in [m]$, are the bandwidths of the servers. Let $\Lambda = \sum_i \lambda_i$ and observe that we can assume $\Lambda \leq \frac{1}{m}\sum_j b_j$. Otherwise there is a server, say j, in OPT_m solution with load strictly greater than mb_j, so $OPT_m = \infty$ and we are done. By this assumption we see that $\sum_i \lambda_i \leq \max_j b_j$. Therefore, we can assign all the data streams (deterministically) to the fastest server, say 1, and let $b = b_1 = \max_j b_j$.

Now, using linearity of expectation, we have the following

$$C = \mathbf{E}\left[\max_j C_j\right] \leq \mathbf{E}\left[\sum_j C_j\right] = \sum_j \mathbf{E}\left[C_j\right] \leq m\cdot\max_j \mathbf{E}\left[C_j\right] \ .$$

We now claim that $\max_j \mathbf{E}\left[C_j\right] \leq f_b(\Lambda)$. Indeed, if not, then there is a server, say j_0, such that $\mathbf{E}\left[C_{j_0}\right] > f_b(\Lambda)$. Then, obviously, there is a data stream, say i_0, with $p_{i_0}^{j_0} > 0$. We also have $c_{i_0}^{j_0} = \mathbf{E}\left[C_{j_0}|x_{i_0}^{j_0}=1\right] \geq \mathbf{E}\left[C_{j_0}\right]$ which follows from the fact that the random variable $\left(C_{j_0}|x_{i_0}^{j_0}=1\right)$ can not assume smaller values than the random variable C_{j_0}. Similarly, $f_b(\Lambda) \geq c_{i_0}^1 = \mathbf{E}\left[C_1|x_{i_0}^1 = 1\right]$, since the value $f_b(\Lambda)$ corresponds to the situation where all probabilities $p_i^1 = 1$ for all i, and $p_i^j = 1$ for all i and $j \neq 1$ (recall that 1 is the fastest server). Thus, $p_{i_0}^{j_0} > 0$ and $c_{i_0}^{j_0} > c_{i_0}^1$, which is a contradiction with the Nash equilibrium property. Therefore, we have

$$C \leq m\cdot\max_j \mathbf{E}\left[C_j\right] \leq m\cdot f_b(\Lambda) \ .$$

By the super-linear scaling we obtain

$$C \leq m\cdot f_b(\Lambda) \leq f_{b/m}\left(\frac{\Lambda}{m}\right) \ .$$

Let us now consider $OPT_m^{\mathtt{frac}}$, which is the cost of an optimum solution for the fractional flow model in the system, where the bandwidths of all servers are reduced by a factor of m. We claim that $f_{b/m}\left(\frac{\Lambda}{m}\right) \leq OPT_m^{\mathtt{frac}}$. To show this we first argue that in the optimal fractional solution the fastest server has largest load. Let us fix two servers with bandwidths b_1, b_2 such that $b_1 < b_2$ and let x_1, x_2 be the loads on these servers in the optimal fractional solution. It is easy to observe that $f_{b_1}(x_1) = f_{b_2}(x_2)$. Furthermore, by the super-linear scaling we obtain $f_{b_2}\left(\frac{b_2}{b_1}x_1\right) \leq \frac{b_1}{b_2}f_{b_1}(x_1) < f_{b_2}(x_2)$, which by the monotonicity property implies that $x_1 \leq x_2$.

By the averaging argument there is a server in the optimal fractional solution with load at least $\frac{\Lambda}{m}$. By our discussion above the fastest server (with bandwidth b) must have load at least $\frac{\Lambda}{m}$. Hence $f_{b/m}\left(\frac{\Lambda}{m}\right) \leq OPT_m^{\mathtt{frac}}$. We finish the proof by observing that

$$C \leq f_{b/m}\left(\frac{\Lambda}{m}\right) \leq OPT_m^{\mathtt{frac}} \leq OPT_m \ . \qquad \square$$

F. PROOF OF THEOREM 13

Suppose that $\sum_{i\in[n]}\lambda_i \leq \sum_{j\in[m]} b_j/6\,e$. Suppose the maximum weight over all streams is $W = b_m/(3\,\delta\,\log m)$ for some given $\delta > 0$ with $\delta\log m \geq 2$. We have to show that for every Nash equilibrium $C \leq m^{-\delta+1} + m\cdot 2^{-b_m/4}$.

For $j \in [m]$, let X_j be the random variable describing the injected load on server j. A simple averaging argument shows the existence of a server $q \in [m]$ with

$$\mathbf{E}[X_q] \ = \ \sum_{i\in[n]} p_i^q\,\lambda_i \ \leq \ \frac{b_q}{6\,e} \ .$$

This server has expected cost $\mathbf{E}[C_q] = \mathbf{E}\left[f_{b_q}(X_q)\right]$. Consider an arbitrary server $j \in [m]$. Suppose there exists $i \in [n]$ with $p_i^j > 0$. Then the Nash equilibrium property gives $c_i^j \leq c_i^q$ so that

$$\begin{aligned}\mathbf{E}[C_j] \ &\leq \ \mathbf{E}[C_j|x_i^j = 1] = c_i^j \leq c_i^q = \mathbf{E}[C_q|x_i^q=1] \\ &\leq \ \mathbf{E}\left[f_{b_q}(X_q+W)\right] \ .\end{aligned}$$

If $p_i^j = 0$, for every $i \in [n]$, then $\mathbf{E}[C_j] = 0 \leq \mathbf{E}\left[f_{b_q}(X_q+W)\right]$ as well. Furthermore, observe that $C_j \leq 1$ by the definition of the cost function. Hence, for any server $j \in [m]$ we have

$$\begin{aligned}\mathbf{E}[C_j] \ &\leq \ \mathbf{Pr}\left[X_q \leq \frac{b_q}{3}\right]\cdot f_{b_q}\left(\frac{b_q}{3}+W\right) + \mathbf{Pr}\left[X_q > \frac{b_q}{3}\right]\cdot 1 \\ &\leq \ f_{b_q}\left(\frac{b_q}{3}+W\right) + \mathbf{Pr}\left[X_q > \frac{b_q}{3}\right] \ .\end{aligned}$$

Let us estimate these two terms as follows. Observe that $W \leq b_m/6 \leq b_q/6$. Consequently,

$$\begin{aligned}f_{b_q}\left(\frac{b_q}{3}+W\right) \ &\leq \ f_{b_q}\left(\frac{b_q}{2}\right) \ \leq \ \frac{\left(\frac{b_q}{2}\right)^{b_q}/b_q!}{\sum_{k=0}^{b_q}\left(\frac{b_q}{2}\right)^k/k!} \\ &\leq \ \frac{\left(\frac{b_q}{2}\right)^{\lfloor b_q/2\rfloor}\left\lceil\frac{b_q}{2}\right\rceil!}{b_q!} \ \leq \ 2^{-b_q/4} \ \leq \ 2^{-b_m/4} \ .\end{aligned}$$

Furthermore, using $\mathbf{E}[X_q] \leq b_q/(6e)$ and the upper bound for the maximum weight over all streams of W, we can apply the standard Hoeffding bound (see, e.g., [3]) to obtain

$$\mathbf{Pr}\left[X_q > \frac{b_q}{3}\right] \ \leq \ 2^{-b_q/3W} \ \leq \ m^{-\delta}$$

because $b_q \geq b_m$ and $W = b_m/(3\,\delta\,\log m)$. Hence, we can conclude

$$C \ = \ \mathbf{E}\left[\max_{j\in[m]} C_j\right] \ \leq \ m\cdot\max_{j\in[m]}\left(\mathbf{E}[C_j]\right) \ \leq \ m^{-\delta+1} + m\,2^{-b_m/4} \ .$$

This completes the proof of Theorem 13. □