Fair and Distributed Bandwidth Allocation under Adversarial Timing

Christian Schindelhauer
Dept. of Math. and Comp. Science
and Heinz-Nixdorf-Institut, Paderborn University
33095 Paderborn, Germany
schindel@upb.de

Berthold Vöcking
Max-Planck-Institut für Informatik
66123 Saarbrücken, Germany
voecking@mpi-sb.mpg.de

Abstract

Congestion control algorithms like, e.g., TCP have to meet the demands of high utilization and fairness simultaneously. We study the tradeoff of these two objectives in a plain model consisting of players, shared resources with bounded bandwidth capacities and rate update events, i.e., points of time at which players can adjust their shares of occupied bandwidth. The times at which players can perform their rate update operations is determined by an adversary. As feedback we allow players to receive the size of the unused bandwidth.

We investigate infinite games, where players can enter and leave at any time but focus our analysis on those periods in which the system is closed, i.e., the set of players that perform update operations is fixed. The major novelty of our model is the adversarial timing of the rate update events. Within this model, we present a simple allocation protocol achieving fairness and almost full utilization in the limit. For example, if players interact only on a single channel then their bandwidths converge polynomially to a state that yields fairness and almost full utilization simultaneously. In general networks the players' bandwidths converge to a state that is only a $(1+\epsilon)$-factor away from max-min fairness. Here the speed of convergence is determined by local parameters like the dilation and the congestion of routing paths.

Furthermore, as a lower bound we prove that there is no protocol achieving full utilization and fairness in the limit, if an adversary determines the order of rate update events and one cannot distinguish between slow and stalled players.

1 Introduction

In this paper, we investigate distributed and cooperative bandwidth allocation protocols. A well-known example of such a protocol is TCP in the Internet. This current form of the TCP protocol was introduced when the Internet experienced a severe service degradation or "Internet Meltdown" during the early growth phase of the mid 1980s [Nagle84]. The dynamics of packet forwarding were underestimated which resulted in a "congestion collapse". The fix for the Internet meltdown is the "back off" behavior of TCP [Jaco88]. In simplified form, when TCP suffers a packet loss, it decreases its sending rate (by decreasing its window size by a factor of two), and when a packet is successfully delivered, it increases its sending rate (by increasing its window size by one). The behavior of TCP depends heavily on the speed at which individual players increase their rate. It is known that TCP is inherently unfair to connections with long-round trip times [FlJa92] and the unfairness can sometimes be as bad as the inverse square of round-trip times [LaMa97].

In this paper we concentrate on fairness and full bandwidth utilization. For this, we consider an asynchronous distributed network in a very simplified setting. In contrast to TCP (but along some concurrent concepts), we allow the protocols to see the residual bandwidth, while other information like the allocated bandwidths (or even the number) of competing protocols is not used. Following the ideas of [Boro98, KKPS00] we challenge our protocols by an adversary to ensure robustness and reliability. In [KKPS00] this is modeled in form of the choice of the bandwidth by an adversarial strategy. In our new approach the link bandwidth is fixed and fluctuations in the available bandwidth for individual players is modeled using an adversary that determines when player enter and leave the system and, in especially, controls the timing of rate update operations of individual players. Let us describe this in more detail.

Consider a set of k players that share a single bus of bandwidth B. Each participating player i holds a rate variable r_i describing how much bandwidth the player currently occupies. From time to time new players arrive and claim a fair share of bandwidth while other players leave the system and release allocated bandwidth. Clearly, such a dynamic environment requires a resource management that adapts the bandwidth allocation continuously to the varying circumstances. For example, if several players share a single bus and a new player arrives then the established players have to release parts of their bandwidth so that the newly arrived player can receive a fair amount of bandwidth. Similarly, if some players leave the system then the remaining players can divide up the released bandwidth.

Let us transfer TCP into this model: A player increases its rate by one unit when he observes that its current rate value can actually be realized since $\sum_i r_i \leq B$. Eventually, the rates will be increased by such an amount that the sum of the individual rates exceeds the available bandwidth and the system collapses. This collapse is observed by the individual players and as a response all players halve their rate values. Then players continue with the linear increase and so on.

Recently, some TCP implementations that use more aggressive congestion strategies and increase their rates at higher speed have been suggested. In fact, already today the speed at which players increase their rates depends on many different aspects, especially on the so-called round-trip times, which again depend on the bandwidth utilization and, hence, on the rates chosen by the players.

In order to study the influence of different speeds in our toy model, consider two players X and Y that interact on a bus. Suppose player X increases its bandwidth s times as fast as player Y. Then, on the long run, the average rate of player X will be s times higher than the average rate of player Y. (This is because the ratio between the sum of rate increments and the sum of rate decrements converges against one with time so that the average loss of player X in case of a collapse must be s times higher than the one of player Y, which in turn implies that also the average rate of X must be s times the average rate of Y.)

We summarize that different speeds for updating the bandwidths can result in an unfair bandwidth allocation in practice as in our toy model (see also [ChJa89, MSM 97]). In the following, we will have a closer look at this kind of problems in an adversarial model of time. We start with upper and lower bounds for a very simple model in which players interact on a single bus. Afterwards we generalize our model to general networks.

1.1 Model 1: Fair bandwidth allocation on a single bus

Consider a single bus of bandwidth B. We assume an *open system* in which players can enter and leave the bus continuously. Let K denote the possibly infinite set of players. When players from K *enter* the bus they request a share of its bandwidth, and when they *leave* the bus they release the allocated bandwidth. *Active players* (i.e., players that entered but did not leave the bus) need to agree on the share of bandwidth that they receive. This is done by so-called "rate update operations" that active players can perform in order to adjust their individual share of bandwidth. We formalize this as follows.

We model the open system by an adversary that specifies a sequence of events $\sigma = \sigma_1\sigma_2\sigma_3\dots$, where each event σ_t is a tuple (i, x) with $i \in K$ and $x \in \{\text{enter}, \text{leave}, \text{update}\}$. With each player $i \in K$, we associate a positive rate variable r_i whose value is zero if the player is inactive, that is, the initial value of r_i is zero and r_i is reset to zero whenever the adversary calls (i, leave). The adversary calls update operations only for active player. In particular, if the adversary calls (i, update) then player i can set r_i to any positive value. In other words, the adversary determines how often and when players can redefine their rate. At any given time, we define the share of bandwidth b_i that player i receives by

$$b_i = \begin{cases} r_i & \text{if } \sum_{i\in K} r_i \le B, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the share of bandwidth of all players is zero when the system is overloaded. (For analogous models see, e.g., [KKPS00].) A fair and efficient allocation protocol aims to set the rates in such a way that all players in the system get almost the same share of bandwidth and the unused bandwidth is as small as possible.

Clearly, when the adversary frequently changes the set of active players or does not allow to perform a reasonable number of update operations for all active players then it is impossible to achieve a fair and efficient allocation of bandwidths among the active players. Therefore, we focus on periods of times in which the system is closed. A *closed system period* $(\mathcal{I}, \mathcal{K})$ is defined by a possibly infinite interval of time $\mathcal{I}$ and a finite set of players $\mathcal{K} \subseteq K$. During $\mathcal{I}$ there are no players that enter and leave the system and the adversary only allows the players in $\mathcal{K}$ to perform update operations. Our goal is to rapidly approach a fair and efficient allocation of bandwidth in closed system periods. For this purpose, we investigate the following simple protocol which is also known as the Phantom Protocol [AMO96a]. Let $\alpha \ge 1$ denote a global parameter.

The Virtual Player Protocol
Suppose player j performs an update operation. Let $\bar{r} = \max\{B - \sum_{i\in K} r_i, 0\}$ denote the unused bandwidth immediately before the update operation. Then player j sets

$$r_j := \frac{\alpha}{\alpha+1}(r_j + \bar{r}) \enspace .$$

In order to describe the behavior of the virtual player protocol (VPP) in a closed system period $(\mathcal{I}, \mathcal{K})$, let us partition $\mathcal{I}$ into contiguous phases in such a way that each phase contains at least one update operation for each player.

Theorem 1.1 *Let c_0 denote a suitable absolute constant. Consider any closed system period $(\mathcal{I}, \mathcal{K})$. Define $k = |\mathcal{K}|$ and let $R = \frac{B}{1+\alpha k}$. Within the first $c_0(\alpha k^2)\log(\alpha k/\epsilon)$ phases of the interval $\mathcal{I}$ the VPP* *reduces the unused bandwidth to at most* $(1+\epsilon)R$ *and yields*

$$b_i \in \left[(1-\epsilon)\frac{B-R}{k}, (1+\epsilon)\frac{B-R}{k}\right]$$

for all all $i \in \mathcal{K}$*, regardless of the initial rates.*

In other words, the VPP utilizes the available bandwidth almost completely and distributes it in a fair way among the players in $\mathcal{K}$. In fact, one can interpret the unused bandwidth $\bar{r}$ (which is the only feedback used by the VPP) as the rate of an additional virtual player. Suppose $\alpha = 1$. Then an update operation of player j simply brings r_j into line with $\bar{r}$. In this way, the bandwidth will finally be divided up in a fair way among all players in $\mathcal{K}$ and the virtual player. By increasing α, the share of the virtual player can be made arbitrarily small. A formal proof of the theorem is given in Section 2.

Let us measure the *length* of closed system periods in the number of phases that they define. Then the theorem implies that the VPP converges against a completely fair bandwidth allocation in closed system periods of infinite length. In other words,

$$\lim_{\tau\to\infty} \frac{b_i(\tau)}{b_j(\tau)} = 1 \ ,$$

for all players i, j from $\mathcal{K}$. Observe, however, that the VPP does not utilize the full bandwidth. In fact, the wasted bandwidth is $R = \frac{B}{1+\alpha k}$ in the limit. This gives rise to the question whether it is possible to obtain fairness and full utilization simultaneously. The following theorem answers this question negatively and, hence, gives a strong motivation for leaving a small fraction of the bandwidth unused.

Theorem 1.2 *For any bandwidth allocation protocol* $\mathcal{P}$ *that converges against full utilization in closed system periods of infinite length, there is an adversarial sequence* σ *that defines a closed system period* $(\mathcal{I}, \mathcal{K})$ *of infinite length with* $|\mathcal{K}| \le 2$ *that enforces a bandwidth assignment of at most* ϵB $(\epsilon > 0)$ *for one of the players in* $\mathcal{K}$*.*

This surprising impossibility result follows from a simple, elegant diagonalization argument. The corresponding proof is given in Section 3.

Note that the irreconcilability of fairness and full utilization also holds if all players know the complete current status, e.g. for explicit rate based algorithms like in [CRL96, Robe96].

1.2 Model 2: Bandwidth allocation in general networks.

We generalize the above adversarial model to general networks. The network is modeled by a (hyper)graph $G = (V, E)$. Edges represent busses, routers, or other shared resources of limited bandwidth. The bandwidth capacity of edge e is denoted by $B(e)$. Each player comes with a set of edges constituting a *simple path* (i.e., a path in which every edge appears at most once). For player $i \in K$, let $path(i)$ denote the player's path, and for an edge $e \in E$ let $K(e) \subseteq K$ denote the set of those players whose paths contain e.

As before, an adversary determines when players enter and leave the system and when they can update their rates. In this paper, we assume that update operations are performed *atomically*, i.e., an update operation is not performed by the adversary until the previous one has become effective on all edges of the respective path. We generalize the VPP as follows. Let $\alpha \ge 1$ denote a global parameter.

The Virtual Player Protocol for General Networks
Suppose player j performs an update operation. For every edge e, let $\bar{r}(e) = B(e) - \sum_{i\in\mathcal{K}(e)} r_i(e)$ denote the free bandwidth on edge e. Then player j sets

$$r_j := \frac{\alpha}{\alpha+1}(r_j + \min_{e\in path(j)}(\bar{r}(e))) \ ,$$

where $\alpha \geq 1$ denotes a global parameter.

The most widely accepted criterion for a fair and efficient bandwidth allocation in networks is the concept of "max-min fairness" [Jaff81, KRT99]. The network is considered to be in a state of *max-min fairness* if it is impossible to infinitesimally increase the rate of any player without exceeding the edge capacities or decreasing the rate of players whose rate is equal or smaller. Our impossibility result for a single edge implies that one cannot converge against max-min fairness in closed system periods. Therefore, we relax the concept of max-min fairness as follows.

For every $\delta > 0$, the network is in a state of *δ-max-min fairness* if it is impossible to increase the rate r of any player by more than a factor of $(1+\delta)$ without exceeding the edge capacities in $path(i)$ or decreasing the rate of players whose rate is at most $(1+\delta)r$. We define that a protocol *converges against δ-max-min fairness* if, given any closed system period $(\mathcal{I}, \mathcal{K})$ of infinite length, the rates converge against a state in which the above criterion is fulfilled among the players in $\mathcal{K}$.

Theorem 1.3 *The VPP converges against $\frac{1}{\alpha}$-max-min fairness.*

The proof of the theorem can be found in Section 4. If $\alpha = 1$ then we can describe the state against which the protocol converges as follows. For every edge e, we define a virtual player whose path contains only the edge e. The rate of this player is defined to be the unused bandwidth on edge e. Then the system converges against a state of max-min fairness among all participating players including the virtual players. Increasing α simply decreases the share of the virtual player and, hence, the wasted bandwidth.

Unfortunately, the analysis showing the convergence does not also prove a fast convergence. For this purpose, we investigate a *discrete variant* of the VPP adopting some ideas of [AwSh98], that is, the rate values of active players are of the form $(1+\epsilon)^q$, for fixed $\epsilon > 0$ and $q \in \mathcal{Z}$. Fix any closed system period $(\mathcal{I}, \mathcal{K})$. Let the *congestion*, $C = C(\mathcal{I}, \mathcal{K})$, denote the maximum number of paths (of participating players) that contain the same edge, and let the *dilation*, $D = D(\mathcal{I}, \mathcal{K})$, denote the maximum length of a path. Furthermore, let $B = B(\mathcal{I}, \mathcal{K})$ denote the ratio between the bandwidth that is available for the participating players on the widest and the narrowest edge.

Theorem 1.4 *For every $\delta > 0$, there is a discrete variant of the VPP that approaches a δ-max-min fair state in any closed system phase. This state is reached after $O(DC^2 + DC(\log B)/\delta^2)$ phases.*

The proof of this theorem is given Section 5. Observe that the performance of the protocol depends only on local parameters like the congestion or the dilation but not on global parameters like the total number of players or the size of the network. Furthermore, the protocol does not need to be parameterized with any other parameter than δ, and the only feedback that a player needs in order to perform an update operation is the unused bandwidth on the narrowest edge on its path.

2 Proof of Theorem 1.1

W.l.o.g., assume $\mathcal{K} = \{1, \ldots, k\}$. We add a virtual player 0 whose rate is defined by $r_0 := \alpha\bar{r}$. In this way, the set of all participating players is $[k+1] := \{0, \ldots, k\}$. Furthermore, we assume that the closed system period starts with σ_1. We want to show that the maximum distance between any pair of rates (including the rate of the virtual player) is at most $(B-R)\epsilon/k^2$ after $T = O(\alpha k^2 \log(\alpha k))$ phases, which implies the theorem.

For every $i \in [k+1]$ and $t \geq 1$, let $r_i(t)$ denote the rate of player i after step t and let $r_i(0)$ denote the initial rate. For $t \geq 0$, let

$$\Delta_t = \max_{i,j \in [k+1]} (r_i(t) - r_j(t))$$

denote the maximum distance after t. We define the following potential function

$$P_t \;=\; \alpha \sum_{\{i,j\}\in\mathcal{K}} |r_i(t) - r_j(t)| + \sum_{i\in\mathcal{K}} |r_i(t) - r_0(t)| \;\; .$$

Observe that $k\Delta_t \le P_t \le \alpha k^2 \Delta_t$, for every $t \ge 0$. Hence, we only have to show that the value of the potential function drops below $(B-R)\epsilon/k$ after $T = O(\alpha k^2 \log(\alpha k))$ phases.

For ≥ 1, define $\delta_t = |r_{i_t}(t) - r_0(t)|$, i.e., the distance between the virtual player and the activated player i_t. We observe $P_t \le P_t - \delta_t$ because

$$|r_{i_t}(t) - r_0(t)| \;=\; |r_{i_t}(t-1) - r_0(t-1)| - \delta_t$$

and, for every $j \in \mathcal{K} \setminus \{i_t\}$,

$$\alpha|r_j(t) - r_{i_t}(t)| + |r_j(t) - r_0(t)| \;\le\; \alpha|r_j(t-1) - r_{i_t}(t-1)| + |r_j(t-1) - r_0(t-1)| \;\; .$$

Thus, the rate of the virtual player changes by $\frac{\alpha}{\alpha+1}\delta_t$ during step t. In other words, the potential decreases by the distance that the virtual player moves times $\frac{\alpha+1}{\alpha}$.

Now, for $T \ge 1$, let $P_{(T)}$ and $\Delta_{(T)}$ denote the potential and the maximum distance, resp., at the end of phase T, and let $P_{(0)}$ and $\Delta_{(0)}$ denote the corresponding initial values. Observe that the distance traveled by the virtual player in phase T is at least $\frac{\Delta_{(T-1)}\alpha}{\alpha+1}$ because its rate is averaged with the smallest and the largest rate in every phase. As a consequence,

$$P_{(T)} \;\le\; P_{(T-1)} - \frac{\alpha+1}{\alpha} \cdot \frac{\Delta_{(T-1)}\alpha}{\alpha+1} \;=\; P_{(T-1)} - \Delta_{(T-1)} \;\; .$$

Applying $\Delta_{(T-1)} \ge \frac{P_{(T-1)}}{\alpha k^2}$ and $P_{(0)} \le \alpha k^2 B$ gives

$$P_{(T)} \;\le\; \alpha k^2 B \left(1 - \frac{1}{\alpha k^2}\right)^T \;\; .$$

Finally, we observe that

$$P_{(T)} \;\le\; \frac{\epsilon B}{2k} \;\le\; \frac{(B-R)\epsilon}{k}$$

for $T = O(\alpha k^2 \log(\alpha k/\epsilon))$. This completes the proof of Theorem 1.1.

Technical remarks. On the first view it might seem that the speed of convergence should be polylogarithmic rather than polynomial in k. In fact, under a randomized sequence of activations of players the rates would converge within $O(\log k)$ phases. A simple counterexample, however, shows that the adversary can force the process to take a linear number of phases until all players come close. This counterexample is given in Appendix A.

The system of rates can also be interpreted as a simple physical systems in which we are given $k+1$ perfectly isolated rooms that initially have different temperatures. The rooms 1 to k have a door leading to room 0. If such a door is opened than the temperatures in both rooms are averaged. Clearly, if all doors are used frequently then the temperatures in all rooms will come closer and closer. In other words, the entropy of the physical system decreases.

This metaphor suggests to consider the entropy as a potential function, e.g., in form of the sum of the squares of the rates or if form of the relative entropy (Kullback Leibler divergence). In fact, both of these potential functions can also be used in order to show the convergence. However, these functions do not decrease as fast as the potential function P (e.g. on the initial instances of the counter example given in Appendix A) and, hence, lead to a slightly weaker upper bounds on the performance.

3 Proof of Theorem 1.2

Assume that such a protocol $\mathcal{P}$ exists. We start with two players Tom and Tina. At the beginning Tom allocates all the bandwidth and Tina none at all. The adversary activates Tina only if the free bandwidth is smaller than $\frac{\epsilon}{(k+1)^2}$, where k denotes the number of Tinas active rounds. Particularly this implies that Tina is activated again if she allocates more than the free bandwidth (system overload). Since the protocol has to resolve this blockade, we consider only the last allocation of Tina in this sequence.

If the protocol converges to full utilization, Tina is activated infinitely often. If not, Tom would remain alone in a closed system period where the wasted bandwidth never falls below a constant value which contradicts our assumption.

So, Tina can allocate additional bandwidth of at most $\frac{\epsilon}{k^2}$ in her k-th active round. Hence, her overall bandwidth is bounded by $\sum_{i=1}^{\infty} \frac{\epsilon}{k^2} = \frac{\pi^2}{6}\epsilon$.

4 Proof of Theorem 1.3

Fix a closed system phase $(\mathcal{I}, \mathcal{K})$. W.l.o.g., we assume $\mathcal{K} = \{1, \ldots, k\}$ and all other players have rate zero. We show that the VPP converges against a particular state S that we describe in the following paragraph.

For every edge e, we define an additional, *virtual player* whose path contains only edge e. The rate of this player is defined by the unused bandwidth of edge e times α. The set of virtual players is called $\mathcal{K}'$. Now let us imagine for a moment that virtual players have a rate that is independent from the unused bandwidth of the respective edge, that is we want to treat virtual players like original players, except that the bandwidth used by a virtual player is only $\frac{1}{\alpha}$ times its rate. Suppose we increment all rates including the rates of the virtual players in round-robin fashion with infinitesimal increments, starting with all rates being zero, until the bandwidth capacities of the narrowest edges are reached. At this point, we stop incrementing the rates for all paths that use one of these edges and continue with the remaining paths in the same fashion until all rates are settled. Let us denote the final state of this process by S.

We observe that S utilizes the bandwidth of all edges if we take into account also the bandwidth occupied by the virtual players. From now on, we consider the bandwidths occupied by the virtual players again as unused bandwidth. For player i, let $e(i)$ denote one of its *bottleneck edge*, i.e., an edge because of which it stopped increasing the bandwidth. By our incremental construction, the rate of player i in state S is equal to the final rate of the virtual player of $e(i)$. In other words, the rate of every player i in S is α times the unused bandwidth on its bottleneck edge $e(i)$. Furthermore, the values of the unused bandwidth on all other edges on *path*(i) are not smaller than this value. This implies that S is a *fixed point*, i.e., the VPP does not diverge from state S once it reaches this state. Furthermore, we can observe that S satisfies $\frac{1}{\alpha}$-max-min fairness as increasing the rate of a player by more than a $(1 + \frac{1}{\alpha})$ would exceed the capacity on its bottleneck edge. (In fact, S yields min-max fairness if we take into consideration also the rates of the virtual players.) Therefore, it remains only to show that the VPP converges against the fixed point S.

For an edge $e \in E$, let $r_0^*(e)$ denote the value of the rate of the virtual player on e in the steady state S. Define $R^* = \{r_0^*(e) | e \in E\}$. Define $m = |R^*|$. (Observe that possibly $m < |E|$.) Let $r[1], \ldots, r[m]$ denote the elements from R^* in increasing order, and define $E[\ell] = \{e \in E | r_0^*[e] = r(\ell)\}$, for $1 \le \ell \le m$. Furthermore, let $\mathcal{K}[\ell] \subseteq \mathcal{K}$ denote those players whose bottleneck edge is in $E[\ell]$, i.e., the set of players whose steady state rate is equal to $r[\ell]$. We will show by induction on ℓ that the rates of the players in $K[\ell]$ will converge against $r[\ell]$.

Claim 4.1 *Let $\gamma > 0$ denote any positive real number. For every $\ell \in \{1, \ldots, m\}$, there exists $\tau \ge 1$ such that, after phase τ, the rates of all players in $K[\ell]$ are in the interval $[r[\ell] - \gamma, r[\ell] + \gamma]$.*

In the rest of the remaining analysis we will show this claim using induction. Let $B^*(e)$ denote the unused bandwidth on edge e if we assume that the players $\mathcal{K}[1] \cup \cdots \cup \mathcal{K}[\ell-1]$ have bandwidths as described by S and all other players have rate zero. In fact, we can assume by induction that the rates of all players in $\mathcal{K}[1] \cup \cdots \cup \mathcal{K}[\ell-1]$ deviate at most by $\pm\beta/k$ from their values in S for any $\beta > 0$. Under this assumption, the bandwidth available for the players in $\mathcal{K}^* = K[\ell] \cup \cdots \cup K[m]$ on edge e fluctuates only within the interval $[B^*(e) - \beta, B^*(e) + \beta]$. Observe that we can choose β arbitrarily small. Nevertheless, we need to take into account these fluctuations explicitly because phases can have arbitrarily length so that a small change in the bandwidth at any given time potentially has vast consequences on the system of rates in later time steps that might be even in the same phase.

In the following, we consider only the players in $\mathcal{K}^*$, that is, we ignore the players from $\mathcal{K}[1] \cup \cdots \cup \mathcal{K}[\ell-1]$ but we take into account the small fluctuations that they cause as follows. We define that the maximal available bandwidth on edge e is $B'(e) = B^*(e) + \beta$ but, in each step t, players may observe a slightly disturbed bandwidth $B'_t(e) \in [B'(e) - 2\beta, B'(e)]$. By our construction, none of the players in $\mathcal{K}^*$ uses an edge from $E[1] \cup \cdots \cup E[\ell-1]$. Therefore, we can restrict our attention to the set of edges $E^* = E \setminus (E[1] \cup \cdots \cup E[\ell-1])$. Let $\mathcal{K}''$ denote the set of virtual players of edges in E^*.

Now fix an edge e. Let C denote the number of player on this edge. Let $B = B'(e)$ denote the maximal bandwidth of this edge, and C the number of players whose paths contain e. If an external observer only sees the behavior of the rates on edge e without knowing any details about the rest of the network then he can observe a behavior that is covered by the following protocol.

Adversarial VPP
Suppose player j performs update operation σ_t. Let $\bar{r}(t-1) = B - \sum_{i=1}^{C} r_i(t-1)$. Then player j sets

$$r_j(t) := \frac{\alpha}{\alpha+1}(r_j(t-1) + \bar{r}(t-1) - X_t) \ ,$$

where $X_t \in [0, \bar{r} + 2\beta]$ is selected by an adversary.

The adversarial sequence X models the disturbing influence due to other edges and bandwidth fluctuations simultaneously. Let $r_0 = \alpha\bar{r}$ denote the bandwidth of the virtual player, also called player 0. Furthermore, let Z denote the fix point of the protocol under the assumption that $X_t = 0$, for all t, that is,

$$Z \ = \ \frac{B\alpha}{\alpha k + 1} \ = \ \frac{(B^*(e) + \beta)\alpha}{\alpha k + 1} \ \geq \ r[\ell] \ .$$

Observe that at least one player $i \in [C+1]$ satisfies $r_i \geq Z$ at any given time. Define $r_{\min} = \min_{i \in [C+1]}(r_i(0))$, i.e., the smallest initial rate.

Lemma 4.2 *Assume $r_{\min} \leq Z - \epsilon$, for any $\epsilon > 0$. Then in every time step after performing one phase, $r_0 \geq r_{\min} + \epsilon 2^{-C-1}$.*

Proof. The lemma follows because of the following monotonicity property of the VPP on single edges: Given an adversarial sequence X, increasing X_t, for any t, increases $r_0(t')$ and does not increase $r_i(t')$, for every $t' \geq t$, $1 \leq i \leq C$. (This property can be shown easily by induction. Observe that monotonicity against adversarial bandwidth fluctuations holds only for single edges. In networks with several edges, reducing the bandwidth of a single edge can decrease and increase rates on other edges. In fact, a small local change in bandwidth can have large influences on the rates on remote edges. For an example showing exponential effects in a similar context see [AMO96b]. Here we cover these vast interdependencies among different edges by worst-case assumptions based on the adversarial sequence X.) Because of this monotonicity property, we can assume in the following that $X_t = 0$, for all t, without increasing the rate of the virtual player.

Next we observe that either the initial value of r_0 is at least Z or there is another player with rate at least Z that performs an update during the first executed phase. Consequently, there is a step t^* in the first phase that yields

$$r_0(t^*) \;\geq\; \frac{\alpha(Z + r_{\min})}{\alpha+1} \;\geq\; r_{\min} + \frac{\epsilon}{2} \enspace .$$

Once more, we apply monotonicity and assume, w.l.o.g., that all updates σ_t $(t > t^*)$ that move the virtual player upward are skipped, i.e., all updates with $r_j(t-1) \geq r_0(t-1)$. Under this assumption, each player is called at most once after t^* because $r_j \geq r_0$ in all time steps after its first update. Now a straightforward induction shows that

$$r_{\min} + \frac{\epsilon}{2^{i+1}} \enspace ,$$

after the ith of at most C updates. Clearly, this proves the lemma. □

Now let us take into account all edges again. We consider *double phases*, i.e., pairs of contiguous phases. Let $R_{\min}$ denote the minimal rate over all players in $\mathcal{K}^* \cup \mathcal{K}''$ at the beginning of a double phase. Suppose $R_{\min} \leq r[\ell] - \epsilon$. Then Lemma 4.2 gives a lower bound on the rates of the virtual players after the first phase, namely $r_0^*(e) \geq R_{\min} + \epsilon 2^{-C-1}$, for every $e \in E^*$, where C denotes the maximum number of players on the same edge. Thus, in the second phase, every player is averaged with a virtual player of value at least $R_{\min} + \epsilon 2^{-C-1}$ so that, after the execution of one double phase, the minimum rate over all player increases to

$$\frac{\alpha R_{\min} + (R_{\min} + \epsilon 2^{-C-1} - 2\beta)}{\alpha+1} \geq R_{\min} + \frac{\epsilon 2^{-C-2}}{\alpha+1} \enspace ,$$

provided $\beta \leq \epsilon 2^{-C-4}$. Consequently, all rates will have value at least $r[\ell] - \epsilon$ after a finite number of phases.

Finally, we observe that this lower bound on the minimal rates also upper-bounds the maximal rate for edges from $E[\ell]$. For $\alpha = 1$, the maximal rate is $r[\ell] + k(\epsilon+\beta)$ as $r[\ell]$ denotes the average rate over all players. For general $\alpha > 1$, a small calculation shows that the maximal rate is $r[\ell] + \gamma$ with $\gamma = O(\alpha k \epsilon)$. This proves Claim 4.1 and, hence, completes the proof of Theorem 1.3.

5 Proof of Theorem 1.4

Now we introduce a discrete version of the VPP that guarantees to reach a fair and efficient allocation within a small number of phases. Here "discrete" means that rates of active players are of the form $(1+\epsilon)^s$, for integral s and positive, real ϵ. We use $\lceil \cdot \rceil$ to indicate upward rounding w.r.t. this representation. We will show that the following protocol reaches a δ-max-min fair state after $O(DC^2 + DC(\log B)/\delta^2)$ phases, for $0 < \delta \leq 1$. Besides δ, let $\epsilon > 0$ and $\alpha \geq 1$ denote global parameters whose values, however, depend on δ and will be determined during the analysis.

Discrete Virtual Player Protocol
Suppose player j performs update operation σ_t.
For every edge e, let $\bar{r}(e) = B(e) - \sum_{i \in \mathcal{K}(e)} r_i(e)$ denote the free bandwidth on edge e.
Set $M = \min_{e \in path(j)}(\alpha \bar{r}(t-1))$.
If $r_j(t-1) \notin [M, (1+\delta)M]$ then set $R = \left\lceil \dfrac{\alpha r_j(t-1) + M}{\alpha+1} \right\rceil$.
Finally, set $r_j(t) = \lceil R \rceil$.

For analyzing the discrete variant of the VPP, we use a similar approach as for the original, fractional VPP. We define a virtual player for each edge e. The rate of this player is denoted by r_e and we define $r_e(t) = \alpha\bar{r}_e(t)$. The minimal bandwidth over all players excluding virtual players is denoted by $r_{\min}(t)$, for $t \geq 0$.

Lemma 5.1 *After a single, initial phase, the sequence of rates $r_{\min}$ is non-decreasing and, for every $e \in E$, $r_e(t) > (1-(\alpha+1)\epsilon)r_{\min}(t)$.*

Proof. If j updates its rate r_j in step t then r_j and M (i.e., the minimal rate r_e on j's paths) are averaged. The weighted average of these rates is denoted by R. (If we would skip the final rounding step and set $r_j(t) = R$ then the rate of j and the minimal virtual rate on its path would be exactly equal.) Then we set $r_j(t) = \lceil R \rceil$. Let $x = \lceil R \rceil - R$ denote the additive increment due to the rounding. We observe

$$x \;<\; \epsilon R \;\leq\; \frac{\epsilon r_j(t)}{1+\epsilon}$$

as $R \leq \frac{1}{1+\epsilon} r_j(t)$. Consequently,

$$r_e(t) \;\geq\; r_j(t) - (\alpha+1)x \;>\; r_j(t) - (\alpha+1)\frac{\epsilon r_j(t)}{1+\epsilon} \;\geq\; (1-(\alpha+1)\epsilon)r_j(t) \;,$$

which implies, $r_e(t) > (1-(\alpha+1)\epsilon)r_{\min}(t)$.

It remains to show that $r_{\min}$ is non-decreasing after a first, initial phase. We observe, that $r_e(t) > r_j(t)-(\alpha+1)x$ after player j has been activated on edge e in step t. Hence, if an edge e has been activated at least once until step t then there exists j and $x < (1-\frac{1}{1-\epsilon})r_j(t)$ such that $r_e(t) \geq r_j(t) - (\alpha+1)x$. Now assume player i is activated in step $t+1$. Let j denote the player satisfying the above properties. Then

$$r_i(t+1) \;\geq\; \left\lceil \frac{\alpha r_j(t) + r_e(t)}{\alpha+1} \right\rceil \;\geq\; \lceil r_j(t) - x \rceil \;\geq\; r_j(t) \;,$$

which implies $r_{\min}(t+1) \geq r_{\min}(t)$, for every step t after the first phase. □

In the following, we ignore the initial phase and assume that $r_{\min}$ is non-decreasing. Let us partition time into super-phases. Each of these super-phases consists of $2CD$ phases or CD double phases. We will use the discrete rates in order to show that $r_{\min}$ increases by a factor of $1+\epsilon$ in each super-phase until the system of rates runs into a bottleneck. More formally, for $T \geq 0$, let $r_{\min}^{(T)}$ denote the value of $r_{\min}$ at the end of super-phase T. We will show by induction that $r_{\min}^{(T)} \geq r_{\min}^{(0)}(1+\epsilon)^T$, for $0 \leq T \leq T^*$ with T^* denoting the first super-phase in which at least one edge "settles down". In super-phase T, an edge is called *settled* if the rates of all non-virtual players on the edge are within the interval $[r_{\min}^{(T-1)}, r_{\min}^{(T-1)}(1+\delta)]$ and the rate of the virtual player is at most $r_{\min}^{(T-1)}$.

Observation 5.2 *Once an edge settles during any super-phase, the rates of the players that cross this edge are fixed forever.*

Now let us fix an arbitrary super-phase. We assume that there is no settled edge at the beginning of the super-phase. W.l.o.g., the super-phase starts with update σ_1 and the smallest initial rate in the super-phase is $r_{\min}^{(T-1)} = 1$. We need the following three lemmas in order to show $r_{\min} \geq (1+\epsilon)$ at the end of the super-phase.

Lemma 5.3 *For every non-virtual player i, if $r_i(t) \geq 1+\epsilon$ then $r_i(t+1) \geq 1+\epsilon$, for every $t \geq 0$.*

Proof. Because of Lemma 5.1 all virtual players have rate at least $1-(\alpha+1)\epsilon$. The rate $r_i(t+1)$ is minimal if $r_i(t)$ is averaged with a virtual player having exactly this value. In this case, a necessary condition for updating the rate r_i is $r_i(t) \geq (1+\delta)(1-(\alpha+1)\epsilon)$, so that

$$r_i(t+1) \geq \frac{\alpha r_i(t)+r_e(t)}{\alpha+1} \geq \frac{\alpha(1+\delta)(1-(\alpha+1)\epsilon)+1-(\alpha+1)\epsilon}{\alpha+1} \geq 1-(\alpha+1+\delta)\epsilon+\delta \ .$$

This term is lower-bounded by $1+\epsilon$ if we assume $\delta \geq (\alpha+2+\delta)\epsilon$. We come back to this constraint later. □

Lemma 5.4 *For every non-virtual player i, if i is called in step $t+1$ with $r_i(t)=1$ and all virtual player on its paths have rate at least $1+\epsilon$ then $r_i(t+1) \geq 1+\epsilon$, for $t \geq 0$.*

Proof. Under the above assumptions,

$$r_i(t+1) \geq \left\lceil \frac{\alpha+M}{\alpha+1} \right\rceil \geq \left\lceil \frac{\alpha+1+\epsilon}{\alpha+1} \right\rceil \geq 1+\epsilon \ .$$

□

Lemma 5.5 *For every non-settled edge e, in every phase there is at least one update σ_t after or before which the virtual player of e has rate larger than one.*

Proof. At any step in the considered phase, either the virtual player has rate larger than one or at least one of the players must have rate larger than $1+\delta$, otherwise the edge would be settled. Let us assume that the virtual player has rate at most one. Let j denote a player with rate $(1+\epsilon)^s \geq 1+\delta$, During the phase, j updates its rate at least once. Let t denote the corresponding time step. Then

$$r_i(t+1) \leq \left\lceil \frac{\alpha(1+\epsilon)^s+1}{\alpha+1} \right\rceil \ .$$

This term is lower-bounded by $(1+\epsilon)^{s-1}$, if we assume that $(1+\epsilon)^s \geq (\frac{\alpha+1}{1+\epsilon}-\alpha)^{-1}$. Under this assumption, the rate of player j is decreased by more than ϵ during step $t+1$, which in turn implies that the rate of the virtual player is increased by more that $\alpha\epsilon$, so that

$$r_e(t+1) > r_e(t)+\alpha\epsilon \geq (1-\alpha\epsilon)+\alpha\epsilon = 1 \ .$$

It remains to choose the parameters δ, α, and ϵ such that $(1+\epsilon)^s \geq (\frac{\alpha+1}{1+\epsilon}-\alpha)^{-1}$. Recall $(1+\epsilon)^s \geq 1+\delta$. Hence, it is sufficient to choose the parameters such that

$$\delta \geq \left(\frac{\alpha+1}{1+\epsilon}-\alpha\right)^{-1} - 1 = \frac{\epsilon+\alpha\epsilon}{1-\alpha\epsilon}$$

We come back to this constraint later. □

Lemma 5.3 implies that we only need to show that each player with rate one at the beginning of the super-phase is *lifted up* (i.e., its rate is set to at least $1+\epsilon$) once during the super-phase in order to show that all players have rate at least $1+\epsilon$ at the end of the super-phase. Now consider a double phase. Lemma 5.5 shows that every virtual player gets *loaded* (i.e., the virtual rate becomes larger than one) at least once during the first phase of the double phase. Furthermore, Lemma 5.4 shows that a player with rate one is lifted up if all virtual players on its paths are loaded. We conclude that a player that is called with rate one during the second phase is lifted up unless it observes a virtual player on its path that has lifted up another player before. Let us call players that share an edge to be *neighbors*. Then a player with initial rate one is lifted up during a double phase or one of its neighbors is lifted up. As a consequence, all players are lifted up during a super-phase consisting of DC double phases as each player has at most $D(C-1)$ neighbors.

We can summarize that the minimum rate $r_{\min}$ increases by a factor of $1+\epsilon$ in every super-phase until at least one edge settles down.

Lemma 5.6 *The set of settled edges and players satisfy δ-max-min fairness.*

Proof. By definition, the players on settled edges have a rate in $[1,(1+\epsilon)^\gamma]$ and the unused bandwidth is at most $\frac{1}{\alpha} \leq \delta$, assuming $\delta \leq \frac{1}{\alpha}$. Hence, one cannot increase the rate r of one of the players by a factor of $1+\frac{1}{\alpha} = 1+\delta$. without exceeding the unused bandwidth or decreasing the bandwidth of a player with rate $(1+\epsilon)^\gamma r \leq (1+\epsilon)^{3(\alpha+1)} \leq (1+6(\alpha+1)\epsilon)r \leq (1+\delta)r$, which corresponds to the definition of δ-max-min fairness. □

Now suppose one or more of the edges settle down. Then we can exclude these edges and the rates of those players that use one of them from our considerations as the corresponding rates are fixed forever. Hence, we can treat the system of remaining edges and players analogously to the original system. In this way, we continue following the allocation process until we find that the rates of all players are fixed in a δ-max-min fair state, provided the constraints that we derived during the analysis are fulfilled, namely $\delta \geq (\alpha+2+\delta)\epsilon$, $\delta \geq \frac{\epsilon+\alpha\epsilon}{1-\alpha\epsilon}$ and $\delta \leq \frac{1}{\alpha}$. It is easy to check, that these constraints are fulfilled if we set $\alpha = \frac{1}{\delta}$, and $\epsilon = \frac{\delta^2}{4}$.

It remains to analyze how many super-phases it takes until all players are settled. W.l.o.g., let us assume that the capacities of the edges are from the interval $[1, B]$. Then one can show that the minimum rate after the execution of only one double phase is $\Omega(2^{-C}/k)$. (This follows analogously to the lower bound on the increase of rates per double phase that we have done for the fractional VPP.) Furthermore, after the last super-phase the minimum rate among the players that survived until the end of our construction is $O(B/k)$. As the minimum rate among the surviving players increases by a factor of $1+\epsilon$ per super-phase, we conclude that the process settles down after $O(C+(\log B)/\epsilon)$ super-phases, which corresponds to $O(DC^2+DC(\log B)/\epsilon) = O(DC^2+DC(\log B)/\delta^2)$ phases. Thus, Theorem 1.4 is shown.

References

[AMO96a] Y. Afek, Y. Mansour, Z. Ostfeld. Phantom: A Simple and Effective Flow Control Scheme. SIGCOMM 96, 169–182, 1996.

[AMO96b] Y. Afek, Y. Mansour, Z. Ostfeld. Convergence Complexity of Optimistic Rate Based Flow Control Algorithms. STOC, 89–98, 1996.

[AwSh98] B. Awerbuch, Y. Shavitt. Converging to Approximated Max-Min Fairness in Logarithmic Time. IEEE INFOCOM'98, 1350–1357. April 1998.

[Boro98] Allan Borodin, Ran El-Yaniv. Online Computation and Competitive Analysis. Cambridge University Press, 1998.

[ChJa89] D.-M. Chiu, R. Jain. Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks. Computer Networks and ISDN System 17, 1–14, 1989.

[CRL96] A. Charny, K.K. Ramakrishnan, A. G. Lauck, Scalability Issues for Distributed Explicit Rate Allocation in ATM Networks. In INFOCOM96, 1198–1205, March 1996.

[FlJa92] S. Floyd and V. Jacobson, On Traffic Phase Effects in Packet Switched Gateways. Internetworking: Research and Experience. 3(3):115–156, September 1992.

[Jaco88] V. Jacobson. Congestion Avoidance and Control. ACM SIGCOMM'88, 314–329, August 1988.

[Jaff81] J. M. Jaffe. Bottleneck flow control. IEEE Transactions on Communications. COM-29(7):954-962, July 1981.

[KKPS00] R. Karp, C. Papadimitriou, Elias Koutsoupias, Scott Shenker. Optimization Problems in Congestion Control, FOCS'00, 66-74, 2000.

[KRT99] J. Kleinberg, Y. Rabani, É. Tardos. Fairness in Routing and Load Balancing. FOCS'99, 568-578, 1999.

[LaMa97] T.V. Lakshman, U. Madhow. The Performance of TCP/IP for Networks with high Bandwidth-Delay Products and Random Loss. IEEE/ACM Transactions on Networking, vol. 5 no 3, pp. 336-350, July 1997.

[MSM 97] M. Mathis, J. Semke, J. Mahdavi. The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. ACM SIGCOMM, Vol. 27, No. 3, July 1997.

[Nagle84] J. Nagle. Congestion Control in IP/TCP. RFC-896, January 1984.

[Robe96] L. G. Roberts: Performance of Explicit Rate Flow Control in ATM. CompCom'96, 22-25.

A A counterexample

The k players $\{1,\dots,k\}$ start with bandwidths $0,1,2,\dots,k-1$ with no wasted bandwith, i.e. $r_i(0)=0$ and $B=k(k-1)/2$. In round t we activate player $j=(t-1) \bmod k+1$ and update his bandwidth by $r_j=\frac{1}{2}(r_j+r_0)$. This implements the Virtual Player Protocol for $\alpha=\frac{1}{2}$. So, player 1 is the first to begin in a phase, which consists of k rounds.

Lemma A.1 *At the end of phase τ we have for the bandwidths r_i of players $i \geq 2\tau\log k$:*

$$i-\tau-\frac{\tau}{k} \leq r_i \leq i-\tau+\frac{\tau}{k} .$$

Proof. We prove this claim by induction. For the first round observe that for $t\in\{1,\dots,k\}$: $r_i(t) = r_0(t)=t-1+t2^{-t}$.

For the inductive step we know that $r_0(\tau k+2\tau\log k)\in[0,k-1]$ (From now on we use interval arithmetik and $x\pm y$ as convenient notation for $[x-y,x+y]$). We claim that for $i\geq 0$:

$$r_{\tau k+2\tau\log k+i}((\tau+1)k)=r_0(\tau k+2\tau\log k+i)\in i-\tau\pm\left(k+\frac{\tau}{k}\right)2^{-i} .$$

This follows by

$$\begin{aligned}
r_{\tau k+2\tau\log k+i}(\tau k) &= \frac{1}{2}r_{(\tau-1)k+2\tau\log k+i}((\tau+1)k)+\frac{1}{2}r_0(\tau k+2\tau\log k+i)\\
&\in \frac{1}{2}\left(i-(\tau-1)\pm\left(k+\frac{\tau-1}{k}\right)2^{-i-2\log k}\right)+\frac{1}{2}\left(i-1-\tau\pm\left(k+\frac{\tau}{k}\right)2^{-i+1}\right)\\
&\subseteq i-\tau\pm\left(\frac{1}{k}2^{-i}+\frac{\tau}{k^2}2^{-i}+k2^{-i}+\frac{\tau}{2k}2^{-i}\right)2^{-i}\pm k2^{-i} \subseteq i-\tau\pm\left(k+\frac{\tau}{k}\right)2^{-i}
\end{aligned}$$

□

This Lemma implies that VPP cannot half the maximum bandwidth difference within $\frac{k}{4\log k}$ phases and $\frac{k^2}{4\log k}$ rounds.

A similar but more lengthy proof improves this bound to $\frac{k-2\log n}{4}$ phases. Then, we replace the activation schedule by double-phases of sequences $1,2,\dots,k,k\dots,2,1$ for the same start configuration.