ELSEVIER

Performance Evaluation 38 (1999) 21–44

PERFORMANCE EVALUATION
An International Journal

www.elsevier.com/locate/peva

A generic weight-proportional bandwidth sharing policy for ATM ABR service

Yiwei Thomas Hou [a,*,1], Henry Tzeng [b], Shivendra S. Panwar [c], Vijay P. Kumar [b]

[a] *Fujitsu Laboratories of America, Network Research, 595 Lawrence Expressway, Sunnyvale, CA 94086-3922, USA*

[b] *Bell Laboratories, Lucent Technologies, Holmdel, NJ, USA*

[c] *Polytechnic University, Brooklyn, NY, USA*

Received 25 July 1998; received in revised form 22 December 1998

Abstract

A generic weight-proportional max–min (WPMM) policy has been proposed for the ATM available bit rate (ABR) service. This policy generalizes the classical max–min policy by supporting the minimum cell rate (MCR) requirement, the peak cell rate (PCR) constraint, and a generic weight for each connection. This paper presents a distributed ABR flow control algorithm for the the WPMM policy and gives a formal proof of the distributed algorithm's convergence to the WPMM policy under any network configuration and any set of link distances. Simulation results on various network configurations demonstrate that the distributed algorithm has a very fast convergence property. ©1999 Elsevier Science B.V. All rights reserved.

Keywords: Max–min policy; Minimum rate; Peak rate; Flow control; Convergence; ABR service

1. Introduction

The classical max–minpolicy has been suggested to allocate network bandwidth among ATM ABR connections [1]. Informally, the max–min policy attempts to maximize the smallest rate among all connections; given the best smallest rate, the next smallest rate is maximized, etc. [3].

There are several drawbacks associated with using the classical max–min policy for ABR service. First of all, the max–min policy, as it stands, cannot support the MCR/PCR constraints of each connection. Second, the max–min policy treats each connection with equal priority and thus is not flexible enough for network providers wishing to introduce differential service options to user connections.

Recently, we proposed a generic weight-based network bandwidth sharing policy, also called *weight-proportional max–min* (WPMM), for ATM ABR service [9]. The WPMM policy generalizes the classical

* Corresponding author. Tel.: +1-408-530-4529; fax: +1-408-530-4515

E-mail address: thou@fla.fujitsu.com (Y.T. Hou)

[1] This work was completed while Y.T. Hou was with Polytechnic University, Brooklyn, NY, USA.

PII: S0166-5316(99)00003-6

max–min by supporting the minimum rate requirement and the peak rate constraint for each connection, as well as sharing remaining network bandwidth among all connections based on a flexible weight assignment associated with each connection. [2]

The main contributions of this paper are the design of an ABR flow control protocol to achieve the WPMM policy and the proof that the protocol converges to the WPMM policy through distributed iterations under any network configuration and any set of link distances. More specifically, our ABR algorithm is motivated by the *consistent marking* technique by Charny et al. [5], which was designed to achieve the simple max–min policy. We extend this technique and design a distributed algorithm for our WPMM policy with the support of a minimum rate requirement, a peak rate constraint, and a weight for each connection. We present an extension of the proof given in [5] to show that our distributed algorithm converges to the WPMM policy through distributed and asynchronous iterations.

The rest of this paper is organized as follows. In Section 2, we define our generic weight-proportional max–min (WPMM) policy. In Section 3, we present a distributed protocol to achieve the WPMM policy and give a formal proof of the protocol's convergence. Section 4 shows simulation results of our distributed algorithm under various network configurations. Section 5 concludes this paper and points out future research directions.

2. A generic weight-proportional rate allocation policy

In our model, a network of switches are interconnected by a set of links $\mathcal{L}$. A set of connections $\mathcal{S}$ traverses one or more links in $\mathcal{L}$ and each connection is allocated a specific rate r_s. Denote $\mathcal{S}_l$ the set of connections traversing link $l \in \mathcal{L}$. Then the (aggregate) allocated rate F_l on link $l \in \mathcal{L}$ of the network is $F_l = \sum_{s \in \mathcal{S}_l} r_s$.

Let C_l be the capacity of link $l \in \mathcal{L}$. A link l is *saturated* or *fully utilized* if $F_l = C_l$. Denote MCR_s and PCR_s the minimum rate requirement and the peak rate constraint for each connection $s \in \mathcal{S}$, respectively.

Definition 1. A rate vector $r = \{r_s | s \in \mathcal{S}\}$ is ABR-feasible if the following two constraints are satisfied: (1) $\mathrm{MCR}_s \leq r_s \leq \mathrm{PCR}_s$ for all $s \in \mathcal{S}$; and (2) $F_l \leq C_l$ for all $l \in \mathcal{L}$.

We assume that the sum of all connections' MCR requirements traversing any link does not exceed the link's capacity, i.e. $\sum_{s \in \mathcal{S}_l} \mathrm{MCR}_s \leq C_l$ for every $l \in \mathcal{L}$. This assumption can be enforced by admission control at call setup time to determine whether or not to accept a new connection.

In our generic weight-proportional max–min policy, we associate each connection $s \in \mathcal{S}$ with a weight (or priority) w_s. [3] Informally, the WPMM policy first allocates to each connection its MCR. Then from the remaining network capacity, it allocates additional bandwidth for each connection using a proportional version of the max–min policy based on each individual connection's weight while satisfying its PCR constraint. The final bandwidth for each connection is its MCR plus an additional "weighted" max–min share. Formally, this policy is defined as follows.

Definition 2. A rate vector r is weight-proportional max–min (WPMM) if it is ABR-feasible, and for each $s \in \mathcal{S}$ and every ABR-feasible rate vector $\hat{r}$ in which $\hat{r}_s > r_s$, there exists some connection $t \in \mathcal{S}$ such that $(r_s - \mathrm{MCR}_s)/w_s \geq (r_t - \mathrm{MCR}_t)/w_t$ and $r_t > \hat{r}_t$.

[2] The WPMM policy here is different from the *proportionally fair* policy by Kelly [14].

[3] We assume a positive weight assignment for each connection.

Due to minimum rate requirement, peak rate constraint, and weight associated with each connection, the bottleneck link definition for the classical max–min in [3] cannot be applied to the WPMM policy. In the following, we define a new notion of WPMM-bottleneck link for our WPMM rate allocation policy.

Definition 3. Given an ABR-feasible rate vector r, a link $l \in \mathcal{L}$ is a WPMM-bottleneck link with respect to r for a connection s traversing l if $F_l = C_l$ and $(r_s - \mathrm{MCR}_s)/w_s \geq (r_t - \mathrm{MCR}_t)/w_t$ for all connections t traversing link l.

We would like to point out that in the special case when: (1) each connection's minimum rate requirement is zero; (2) there is no peak rate constraint for each connection; and (3) each connection has equal weight; then the above WPMM rate allocation policy and the WPMM-bottleneck link definitions degenerate into those for the classical max–min, respectively.

The following theorem links the relationship between the above WPMM policy and the WPMM-bottleneck link definitions.

Theorem 1. *An ABR-feasible rate vector r is WPMM if and only if each connection has either a WPMM-bottleneck link with respect to r or a rate assignment equal to its PCR.*

A proof of Theorem 1 was given in [9]. It can be shown that there exists a unique rate vector that satisfies the WPMM rate allocation policy.

The following iterative steps describe how to compute the rate allocation for each connection in any network such that the WPMM policy is satisfied.

Algorithm 1 (A centralized algorithm).

1. Start the rate allocation of each connection with its MCR.
2. Increase the rate of each connection with an increment proportional to its weight until either some link becomes saturated or some connection reaches its PCR, whichever comes first.
3. Remove those connections that either traverse saturated links or have reached their PCRs and the capacity associated with such connections from the network.
4. If there is no connection left, the algorithm terminates; otherwise, go back to step 2 for the remaining connections and remaining network capacity.

A correctness proof that Algorithm 1 achieves the WPMM rate allocation was given in [9]. We use the following simple example to illustrate how Algorithm 1 allocates network bandwidth to achieve the WPMM policy.

Example 1 (Peer-to-peer configuration). In this network (Fig. 1), the output port link of SW1 (Link 12) is the only potential WPMM-bottleneck link for the three connections. Assume that all links are of unit

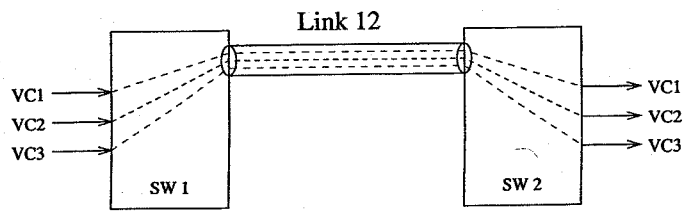


Fig. 1. The peer-to-peer network configuration.

Table 1
MCR requirement, PCR constraint, weight, and WPMM rate allocation for each connection in the peer-to-peer network

Connection	MCR	PCR	Weight	WPMM rate allocation
VC1	0.15	1.00	3	0.525
VC2	0.10	0.30	2	0.300
VC3	0.05	0.50	1	0.175

capacity. The MCR requirement, PCR constraint, and weight of each connection are listed in Table 1, as well as the WPMM rate allocation for each connection.

We would like to point out that our WPMM policy provides an attractive pricing strategy for network service providers. In particular, each connection may be charged a premium rate corresponding to the guaranteed bandwidth (i.e. MCR). Beyond this rate, each connection may be billed an additional tariff for the weight (or priority) to share any additional unguaranteed (or available) network capacity.

The centralized algorithm for the WPMM rate allocation requires global information. The main contributions of this paper are the design and the convergence proof of a distributed ABR flow control algorithm for the WPMM policy, which are presented in the following sections.

3. A distributed ABR flow control algorithm for WPMM

3.1. Previous work

There have been extensive prior efforts on the design of distributed algorithms to achieve the classical max–min policy. In essence, all these schemes maintain some link controls at the switch level and convey some information about these controls to the source by means of feedback. Upon the receipt of the feedback signal, the source adjusts its estimate of the allowed transmission rate according to some rule. These algorithms essentially differ in the particular choices of link controls and the type of feedback provided to the sources by the network.

The work by Hayden [7], Jaffe [10], and Gafni [6] described distributed algorithms of this type. However, these algorithms required synchronization of all nodes during each iteration, which is difficult to achieve in practice.

The work of Mosely [15] was the first on distributed algorithm using asynchronous iterations. Unfortunately, this algorithm's convergence time was rather slow and simulation results showed poor adaptation to any change in the network.

Ramakrishnan et al. [16] proposed to use a single bit to indicate congestion with the aim of achieving max–min rate allocation. Due to the binary nature of the algorithm, the source's rate exhibited oscillations.

Recent research activities on ABR at the ATM Forum have brought many renewed efforts on the design of distributed algorithms to achieve the classical max–min policy. These algorithms either used heuristics [12,13,17–19] or a theoretical approach [5] and had different performance behaviors and implementation complexities. In particular, the *consistent marking* technique by Charny et al. [5] was one of the few algorithms that were proven to converge to the max–min.

The main contributions of this paper are: (1) We extend the Consistent Marking technique to design a distributed algorithm for our WPMM policy; and (2) We present a formal proof of our distributed algorithm's convergence to the WPMM policy, which is a generalization of the proof given in [5].

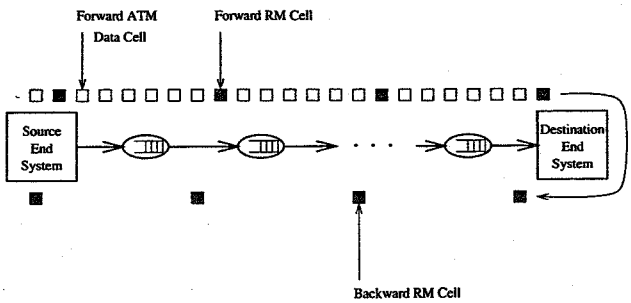


Fig. 2. ABR flow control mechanism.

Our distributed protocol uses the ABR flow control mechanism (Fig. 2), where special control packets called resource management (RM) cells are employed for end-to-end flow control and to convey congestion information from the network to the source. Our work is to design an algorithm at each switch which performs rate calculation for each connection such that our WPMM rate allocation can be achieved globally through distributed iterations. We present our distributed ABR algorithm in Section 3.2 and provide a convergence proof in Section 3.3.

3.2. A distributed protocol

We first specify the source and destination behaviors for each connection [2]. The following are the parameters at a source or in the RM cell.

ACR: Allowed cell rate of a source.
ICR: Initial cell rate of a source.
CCR: Current cell rate field in an RM cell.
ER: Explicit rate field in an RM cell.

Algorithm 2 (End system behaviors).

- *Source behavior*
  - The source starts with ACR := ICR, with ICR $\geq$ MCR;
  - For every N_{rm} transmitted data cells, the source sends a forward RM(CCR, MCR, ER, W) cell with CCR := ACR; MCR := MCR; ER := PCR; W := W;
  - Upon the receipt of a backward RM(CCR, MCR, ER, W) cell from the destination, the ACR at the source is adjusted to: ACR := ER.
- *Destination behavior*
  - The destination end system of a connection simply returns every RM cell back towards the source upon receiving it.

For the design of our switch algorithm, we employ per flow accounting at each output port of a switch. That is, we maintain a table at each output port of a switch to keep track of the state information of each traversing connection. Based on the state information of each connection, we calculate the explicit rate for each connection.

The following are the link parameters and variables used by our switch algorithm:

n_l: Number of connections in $\mathcal{S}_l$, $l \in \mathcal{L}$, i.e. $n_l = |\mathcal{S}_l|$.
r_l^i: CCR value of connection $i \in \mathcal{S}_l$ at link l.

b_l^i: Bit used to mark connection $i \in \mathcal{S}_l$ at link l,

$$b_l^i = \begin{cases} 1 & \text{if connection } i \in \mathcal{S}_l \text{ is } \textit{marked} \text{ at link } l, \\ 0 & \text{if connection } i \in \mathcal{S}_l \text{ is } \textit{unmarked} \text{ at link } l. \end{cases}$$

$\mathcal{M}_l$: Set of connections marked at link l, i.e. $\mathcal{M}_l = \{i | i \in \mathcal{S}_l \text{ and } b_l^i = 1\}$.
$\mathcal{U}_l$: Set of connections unmarked at link l, i.e. $\mathcal{U}_l = \{i | i \in \mathcal{S}_l \text{ and } b_l^i = 0\}$, and $\mathcal{M}_l \cup \mathcal{U}_l = \mathcal{S}_l$.
φ_l: A variable at link l used to estimate the weight-normalized WPMM bottleneck-link rate. It is calculated by the following algorithm, which is an extension of the one given in [5].

Algorithm 3 (Calculation of φ_l).

$$\varphi_l := \begin{cases} \infty & \text{if } n_l = 0;^4 \\ (C_l - \sum_{i \in \mathcal{S}_l} r_l^i)/\sum_{i \in \mathcal{S}_l} w_i + \max_{i \in \mathcal{S}_l}((r_l^i - \text{MCR}^i)/w_i) & \text{if } n_l = |\mathcal{M}_l|; \\ ((C_l - \sum_{i \in \mathcal{S}_l} \text{MCR}^i) - \sum_{i \in \mathcal{M}_l}(r_l^i - \text{MCR}^i))/\sum_{i \in \mathcal{U}_l} w_i & \text{otherwise.} \end{cases}$$

We give some intuitions on Algorithm 3 for the special case when both $\text{MCR}^i = 0$ and $w_i = 1$ for all $i \in \mathcal{S}$, i.e. the max–min case. In this special case, the last expression becomes $\varphi_l := (C_l - \sum_{i \in \mathcal{M}_l} r_l^i)/|\mathcal{U}_l|$ when not all connections are marked, i.e. during transient iteration process. This is precisely the expression commonly used to calculate max–min rate allocation. The second expression for φ_l shows what happens when all connections are marked, which would be the case when the distributed algorithm converges. In this case, $C_l = \sum_{i \in \mathcal{S}_l} r_l^i$ at a saturated link where all connections are marked and the second expression simply becomes $\varphi_l := \max_{i \in \mathcal{S}_l} r_l^i$, i.e. the max–min bottleneck link. This simple max–min case for φ_l was done in [5]. Our construction of φ_l calculation in Algorithm 3 extends that in [5] by taking into account of the weight and MCR of each connection.

We point out that the φ_l calculation for WPMM rate allocation may not be unique. But the specific φ_l calculation which we constructed in Algorithm 3 is provable to converge to the WPMM rate allocation (Section 3.3) when used in conjunction with our switch algorithm (Algorithm 4) below.

The following algorithm specifies the behavior at each output port of a switch. Initially: for each $l \in \mathcal{L}$, $\mathcal{S}_l := \emptyset$; $n_l := 0$; $\varphi_l := \infty$.

Algorithm 4 (Switch behavior).

```
Upon the receipt of a forward RM(CCR, MCR, ER, W) cell from the source of connection i {
  if RM cell signals connection termination {
    S_l := S_l − {i};   n_l := n_l − 1;
    table_update();
    }
  if RM cell signals a new connection initiation {
    S_l := S_l ∪ {i};   n_l := n_l + 1;
    r_l^i := CCR;   MCR^i := MCR;   w_i := W;   b_l^i := 0;
    table_update();
    }
  else /* i.e. RM cell belongs to an ongoing active connection. */ {
    r_l^i := CCR;
    if ((r_l^i − MCR^i)/w_i ≤ φ_l) then b_l^i := 1;
```

[4] In fact, φ_l can be set to any value when $n_l = 0$.

```
    table_update();
    }
  Forward RM(CCR, MCR, ER, W) towards its destination;
}
Upon the receipt of a backward RM(CCR, MCR, ER, W) cell from the destination of connection i {
  ER:= max{min{ER, (φ_l · w_i + MCR^i)}, MCR^i};
  Forward RM(CCR, MCR, ER, W) towards its source;
}
table_update()
{
  rate_calculation_1: use Algorithm 3 to calculate φ_l^1;
  Unmark (i.e. set b_l^i = 0) any connection i ∈ M_l at link l with (r_l^i − MCR^i)/w_i > φ_l^1;
  /* Update φ_l after the above unmarking operation. */
  rate_calculation_2: use Algorithm 3 to calculate φ_l;
  if (φ_l < φ_l^1), then {
    Unmark any connection i ∈ M_l at link l with (r_l^i − MCR^i)/w_i > φ_l;
    rate_calculation_3: use Algorithm 3 to calculate φ_l again;
    }[5]
}
```

We observe that by the operations of Algorithms 2 and 4, we have the following fact for the ACR parameter at the source and the CCR field in the RM cell.

Fact 1. For every connection $s \in \mathcal{S}$, the ACR at the source and the CCR field in the RM cell are ABR-feasible, i.e. $\text{MCR}^s \leq \text{ACR}^s \leq \text{PCR}^s$ and $\text{MCR}^s \leq \text{CCR}^s \leq \text{PCR}^s$.

The space and time complexity of our distributed algorithm is as follows. Since we employ per flow accounting, the memory storage requirement at each output port link is $O(|\mathcal{S}|)$, where $|\mathcal{S}|$ is the number of connections in the network. It is easy to see that the computational complexity of the distributed algorithm is also $O(|\mathcal{S}|)$.

In the following, we give a proof that rate calculated by the above distributed algorithm converges to the WPMM policy through distributed and asynchronous iterations. The objective of our proof is to give a theoretical guarantee that our distributed algorithm converges to the WPMM policy under *any* network configuration and *any* set of link distances. Our proof extends the work by Charny et al. [5], which was done for a simpler algorithm for the max–min case.

3.3. Convergence proof

The convergence proof of our algorithm is based on a sequence of lemmas. The key concept in the proof is the notion of *marking-consistent*, which we define as follows.

[5] Both φ_l^1 and φ_l follow the same φ_l calculation in Algorithm 3. For the classical max–min policy, φ_l calculated by rate_calculation_2 is always greater than or equal to φ_l^1 and rate_calculation_3 is not needed [5]. But for our WPMM policy, φ_l calculated by rate_calculation_2 may be less than φ_l^1 and therefore, another around of unmarking and rate_calculation_3 is necessary (see the proof of Lemma 1 for such a unique case). This is an extension of the consistent marking technique in [5].

Definition 4. Let $\mathcal{M}_l$ be the set of marked connections at link $l \in \mathcal{L}$. We say that the marking of connections at link $l \in \mathcal{L}$ is in the state of marking-consistent if

$$\frac{r_l^i - \mathrm{MCR}^i}{w_i} \leq \varphi_l$$

for every connection $i \in \mathcal{M}_l$.

The following key lemma shows the table marking property at a link after the switch algorithm is performed for a traversing RM cell.

Lemma 1 (Fundamental lemma). *After the switch algorithm is performed for an RM cell traversing a link, the marking of connections at this link is marking-consistent.*

Proof of Lemma 1. Let $\mathcal{M}_l$ and $\mathcal{U}_l$ be the set of marked and unmarked connections at link l just before rate_calculation_1 is performed, respectively; φ_l^1 be the result by rate_calculation_1 in function table_update(); $\mathcal{Z}_l \subseteq \mathcal{M}_l$ be the set of connections with $(r_l^i - \mathrm{MCR}^i)/w_i > \varphi_l^1, i \in \mathcal{Z}_l$ and therefore, are unmarked by the unmarking operation after rate_calculation_1 in function table_update(); φ_l be the result by rate_calculation_2 in function table_update().

Case 1. If not all connections in $\mathcal{S}_l$ are marked before rate_calculation_1, i.e. $\mathcal{M}_l \neq \mathcal{S}_l$, we have

$$\varphi_l^1 = \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{U}_l} w_i}. \tag{1}$$

After the unmarking operation, φ_l calculated by rate_calculation_2 is

$$\begin{aligned}
\varphi_l &= \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in(\mathcal{M}_l-\mathcal{Z}_l)}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in(\mathcal{U}_l\cup\mathcal{Z}_l)} w_i} \\
&= \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i) + \sum_{i\in\mathcal{Z}_l}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{U}_l} w_i + \sum_{i\in\mathcal{Z}_l} w_i} \\
&\geq \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i) + \varphi_l^1 \cdot \sum_{i\in\mathcal{Z}_l} w_i}{\sum_{i\in\mathcal{U}_l} w_i + \sum_{i\in\mathcal{Z}_l} w_i} = \varphi_l^1.
\end{aligned}$$

The last equality holds because of Eq. (1). Therefore, φ_l calculated by rate_calculation_2 is greater than or equal to φ_l^1 by rate_calculation_1. Since $(r_l^i - \mathrm{MCR}^i)/w_i \leq \varphi_l^1$ for $i \in (\mathcal{M}_l - \mathcal{Z}_l)$, and $\varphi_l^1 \leq \varphi_l$, the marking of these connections continues to satisfy marking-consistent after rate_calculation_2 is performed.

Case 2. If all connections in $\mathcal{S}_l$ are marked before rate_calculation_1, i.e. $\mathcal{M}_l = \mathcal{S}_l$, we have two scenarios. Let the RM cell for which the switch algorithm is performed belong to connection $s \in \mathcal{S}$.

Subcase A. If connection s was not marked before the RM cell's arrival at link l and is marked because of this RM cell's arrival with

$$\frac{r_l^s - \mathrm{MCR}^s}{w_s} \leq \varphi_l = \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{S}_l, i\neq s}(r_l^i - \mathrm{MCR}^i)}{w_s}.$$

After marking $b_l^s = 1$, we have

$$C_l - \sum_{i\in\mathcal{S}_l} r_l^i \geq 0. \tag{2}$$

-During rate_calculation_1:

$$\varphi_l^1 = \frac{C_l - \sum_{i\in\mathcal{S}_l} r_l^i}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{i\in\mathcal{S}_l} \frac{r_l^i - \mathrm{MCR}^i}{w_i}.$$

With (2), we have

$$\varphi_l^1 \geq \max_{i\in\mathcal{S}_l} \frac{r_l^i - \mathrm{MCR}^i}{w_i} \geq \frac{r_l^p - \mathrm{MCR}^p}{w_p} \quad \text{for every connection } p \in \mathcal{S}_l.$$

So all connections in $\mathcal{S}_l$ will remain marked after the unmarking operation. Therefore, φ_l calculated by rate_calculation_2 will be the same as φ_l^1 and the marking of all connections is marking-consistent.

Subcase B. If connection s was already marked before this RM cell arriving at link l, the arrival of this RM cell will not change the advertised rate value if the CCR in this RM cell is the same as r_l^s in the current VC table. On the other hand, if the new CCR is different from the recorded CCR for this connection in the VC table, r_l^s will be updated with this new CCR value. During rate_calculation_1:

$$\varphi_l^1 = \frac{C_l - \sum_{i\in\mathcal{S}_l} r_l^i}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{i\in\mathcal{S}_l} \frac{r_l^i - \mathrm{MCR}^i}{w_i}. \tag{3}$$

Again, let $\mathcal{Z}_l \subseteq \mathcal{M}_l$ denote the set of connections with $(r_l^i - \mathrm{MCR}^i)/w_i > \varphi_l^1$, $i \in \mathcal{Z}_l$ and therefore, are unmarked by the unmarking operation after rate_calculation_1 in function table_update().

If $\mathcal{Z}_l = \emptyset$, i.e. no connection is unmarked, then φ_l calculated by rate_calculation_2 will be the same as φ_l^1 and all connections will remain marking-consistent.

If $\mathcal{Z}_l \neq \emptyset$, then the set of connections in $\mathcal{Z}_l$ will be unmarked since

$$\frac{r_l^i - \mathrm{MCR}^i}{w_i} > \varphi_l^1, \quad i \in \mathcal{Z}_l.$$

During rate_calculation_2, we have

$$\begin{aligned}
\varphi_l &= \frac{(C_l - \sum_{i\in\mathcal{S}_l} \mathrm{MCR}^i) - \sum_{i\in(\mathcal{S}_l - \mathcal{Z}_l)} (r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{Z}_l} w_i} \\
&= \frac{1}{\sum_{i\in\mathcal{Z}_l} w_i} \left[\left(\varphi_l^1 \cdot \sum_{i\in\mathcal{S}_l} w_i - \max_{i\in\mathcal{S}_l} \frac{r_l^i - \mathrm{MCR}^i}{w_i} \cdot \sum_{i\in\mathcal{S}_l} w_i + \sum_{i\in\mathcal{S}_l} r_l^i - \sum_{i\in\mathcal{S}_l} \mathrm{MCR}^i \right) \right. \\
&\quad \left. - \sum_{i\in(\mathcal{S}_l - \mathcal{Z}_l)} (r_l^i - \mathrm{MCR}^i) \right] = \frac{1}{\sum_{i\in\mathcal{Z}_l} w_i} \left\{ \varphi_l^1 \cdot \sum_{i\in\mathcal{S}_l} w_i - \max_{i\in\mathcal{S}_l} \frac{r_l^i - \mathrm{MCR}^i}{w_i} \cdot \sum_{i\in\mathcal{S}_l} w_i \right. \\
&\quad \left. + \sum_{i\in\mathcal{S}_l} (r_l^i - \mathrm{MCR}^i) - \left[\sum_{i\in\mathcal{S}_l} (r_l^i - \mathrm{MCR}^i) - \sum_{i\in\mathcal{Z}_l} (r_l^i - \mathrm{MCR}^i) \right] \right\} \\
&= \frac{1}{\sum_{i\in\mathcal{Z}_l} w_i} \left\{ \varphi_l^1 \cdot \sum_{i\in\mathcal{S}_l} w_i - \max_{i\in\mathcal{S}_l} \frac{(r_l^i - \mathrm{MCR}^i)}{w_i} \cdot \sum_{i\in\mathcal{S}_l} w_i + \sum_{i\in\mathcal{Z}_l} (r_l^i - \mathrm{MCR}^i) \right\}.
\end{aligned}$$

The second equality above follows from Eq. (3). Now let connection $p \in \mathcal{S}_l$ be the connection with $(r_l^p - \mathrm{MCR}^p)/w_p = \max_{i\in\mathcal{S}_l}((r_l^i - \mathrm{MCR}^i)/w_i)$. Then connection p must be in the set of $\mathcal{Z}_l$ since $\mathcal{Z}_l$ contains connections with $(r_l^i - \mathrm{MCR}^i)/w_i > \varphi_l^1, i \in \mathcal{Z}_l$ and connection p has the largest value of $(r_l^p - \mathrm{MCR}^p)/w_p$ among all connections. Therefore,

$$\begin{aligned}
\varphi_l &= \frac{1}{\sum_{i\in\mathcal{Z}_l} w_i}\left\{\varphi_l^1 \cdot \left[\sum_{i\in\mathcal{Z}_l} w_i + \sum_{i\in(\mathcal{S}_l-\mathcal{Z}_l)} w_i\right]\right.\\
&\quad \left. - \frac{r_l^p - \mathrm{MCR}^p}{w_p} \cdot \left[\sum_{i\in\mathcal{Z}_l} w_i + \sum_{i\in(\mathcal{S}_l-\mathcal{Z}_l)} w_i\right] + \sum_{i\in\mathcal{Z}_l}(r_l^i - \mathrm{MCR}^i)\right\}\\
&= \frac{1}{\sum_{i\in\mathcal{Z}_l} w_i}\left\{\varphi_l^1 \cdot \sum_{i\in\mathcal{Z}_l} w_i + \left(\varphi_l^1 - \frac{r_l^p - \mathrm{MCR}^p}{w_p}\right) \cdot \sum_{i\in(\mathcal{S}_l-\mathcal{Z}_l)} w_i\right.\\
&\quad \left. + \sum_{i\in\mathcal{Z}_l}\left[(r_l^i - \mathrm{MCR}^i) - \frac{r_l^p - \mathrm{MCR}^p}{w_p} \cdot w_i\right]\right\}\\
&= \varphi_l^1 + \frac{1}{\sum_{i\in\mathcal{Z}_l} w_i}\left\{\left(\varphi_l^1 - \frac{r_l^p - \mathrm{MCR}^p}{w_p}\right) \cdot \sum_{i\in(\mathcal{S}_l-\mathcal{Z}_l)} w_i\right.\\
&\quad \left. + \sum_{i\in\mathcal{Z}_l}\left[\left(\frac{r_l^i - \mathrm{MCR}^i}{w_i} - \frac{r_l^p - \mathrm{MCR}^p}{w_p}\right) \cdot w_i\right]\right\} < \varphi_l^1.
\end{aligned}$$

The last inequality holds since $\varphi_l^1 < (r_l^p - \mathrm{MCR}^p)/w_p$ and $(r_l^i - \mathrm{MCR}^i)/w_i \le (r_l^p - \mathrm{MCR}^p)/w_p$ for $i \in \mathcal{Z}_l$. This is the only situation where φ_l calculated by rate_calculation_2 is less than φ_l^1. We move on to perform another round of unmarking and φ_l calculation (rate_calculation_3). It is clear that the combined steps of rate_calculation_2, unmarking, and rate_calculation_3 here are equivalent to Case 1. Thus, φ_l calculated by rate_calculation_3 is greater than or equal to φ_l calculated by rate_calculation_2 and the marking of connections is marking-consistent upon the termination of function table_update(). □

Lemma 1 is a fundamental lemma in our convergence proof and will be used by subsequent lemmas. The following lemma gives a lower bound for φ_l at link $l, l \in \mathcal{L}$.

Lemma 2. *There exists some time t_0 such that for $t \ge t_0$:*

$$\varphi_l \ge \frac{C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i}{\sum_{i\in\mathcal{S}_l} w_i}$$

for every $l \in \mathcal{L}$.

For a proof of Lemma 2, see Appendix A.

In the special case when both $\mathrm{MCR}^i = 0$ and $w_i = 1$ for all $i \in \mathcal{S}$, i.e. the simple max–min case, Lemma 2 says that φ_l is greater than or equal to C_l/n_l at link $l \in \mathcal{L}$.

Let K be the total number of iterations needed to execute the centralized algorithm for the WPMM policy (Algorithm 1). As we have shown in the correctness proof of Algorithm 1, $K \le |\mathcal{S}|$, where $|\mathcal{S}|$ is the

total number of connections in the network. Let $\mathcal{S}_i$, $1 \le i \le K$ be the set of connections being removed at the end of the ith iteration, i.e. connections in $\mathcal{S}_i$ have either reached their WPMM-bottleneck link rate or their PCRs during the ith iteration of Algorithm 1. Let $\mathcal{L}_i$, $1 \le i \le K$ be the set of links traversed by connections in $s \in \mathcal{S}_i$. Note that $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K$ are mutually exclusive and the sum of $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_K$ is $\mathcal{S}$ while $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K$ may be mutually inclusive. That is, there may be links belonging to both $\mathcal{L}_i$ and $\mathcal{L}_{i+1}$.[6]

Let σ_i, $1 \le i \le K$ be defined as follows:

$$\sigma_i = \frac{r^s - \mathrm{MCR}^s}{w_s} \quad \text{for every } s \in \mathcal{S}_i, \quad 1 \le i \le K,$$

where r^s is the final WPMM rate allocation for connection s by Algorithm 1. By the operation of Algorithm 1, for a connection $p \in \mathcal{S}$ which has not yet gone through a saturated link or reached its PCR, its $(r^p - \mathrm{MCR}^p)/w_p$ increases at each iteration. Therefore, we have the following property for σ_i, $1 \le i \le K$:

$$\sigma_1 < \sigma_2 < \cdots < \sigma_K.$$

The following lemma states the inequality between φ_l and σ_1 on every link $l \in \mathcal{L}$ in the network.

Lemma 3. *Let t_0 be defined as in Lemma 2.*

1. *If $\sigma_1 = (C_l - \sum_{i \in \mathcal{S}_l} \mathrm{MCR}^i)/\sum_{i \in \mathcal{S}_l} w_i \le (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s$ for $s \in \mathcal{S}_1$, i.e. connections $s \in \mathcal{S}_1$ reach the WPMM-bottleneck link rate before their PCRs in the centralized algorithm, then for any $t > t_0$, $\varphi_l \ge \sigma_1$ for every $l \in \mathcal{L}_1$ and $\varphi_l > \sigma_1$ for every $l \in (\mathcal{L} - \mathcal{L}_1)$.*
2. *If $\sigma_1 = (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s < (C_l - \sum_{i \in \mathcal{S}_l} \mathrm{MCR}^i)/\sum_{i \in \mathcal{S}_l} w_i$ for $s \in \mathcal{S}_1$, i.e. connections $s \in \mathcal{S}_1$ reach their PCRs before the WPMM-bottleneck link rate in the centralized algorithm, then for any $t > t_0$, $\varphi_l > \sigma_1$ for every $l \in \mathcal{L}$.*

The proof of Lemma 3 is given in Appendix B.

The following lemma shows that the rate allocation for connection $s \in \mathcal{S}_1$ will eventually converge to the WPMM rate with its bit marked permanently on every link along its traversing path.

Lemma 4 (Base case). *There exists a $T_1 \ge 0$ such that:*

1. *If $\sigma_1 = (C_l - \sum_{i \in \mathcal{S}_l} \mathrm{MCR}^i)/\sum_{i \in \mathcal{S}_l} w_i < (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s$ for $s \in \mathcal{S}_1$, i.e. connections $s \in \mathcal{S}_1$ reach the WPMM-bottleneck link rate before their PCRs in the centralized algorithm, then for $t \ge T_1$, the following statements hold for the distributed algorithm:*
   - 1.1. *$\varphi_l = \sigma_1$ for every link $l \in \mathcal{L}_1$.*
   - 1.2. *The ER field of every returning RM cell of connection $i \in \mathcal{S}_1$ satisfies $\mathrm{ER} = \sigma_1 \cdot w_i + \mathrm{MCR}^i$.*
   - 1.3. *The ACR at source of every connection $i \in \mathcal{S}_1$ satisfies $\mathrm{ACR} = \sigma_1 \cdot w_i + \mathrm{MCR}^i$.*
   - 1.4. *$b_l^i = 1, r_l^i = \sigma_1 \cdot w_i + \mathrm{MCR}^i$ for every connection $i \in \mathcal{S}_1$ and every link l traversed by connection $i \in \mathcal{S}_1$.*
   - 1.5. *The ER field of every returning RM cell of connection $j \in (\mathcal{S} - \mathcal{S}_1)$ satisfies $\mathrm{ER} > \sigma_1 \cdot w_j + \mathrm{MCR}^j$.*
   - 1.6. *The ACR at source of every connection $j \in (\mathcal{S} - \mathcal{S}_1)$ satisfies $\mathrm{ACR} > \sigma_1 \cdot w_j + \mathrm{MCR}^j$.*
   - 1.7. *The recorded CCR of connection $j \in (\mathcal{S} - \mathcal{S}_1)$ satisfies $r_l^j > \sigma_1 \cdot w_j + \mathrm{MCR}^j$ at every link l traversed by connection j.*

[6] This happens when connections in $\mathcal{S}_i$ reaching their PCRs before saturating link $l \in \mathcal{L}_i$ and link $l \in \mathcal{L}_i$ becomes part of $\mathcal{L}_j$, $j > i$.

2. *If* $\sigma_1 = (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s \leq (C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i$ *for* $s \in \mathcal{S}_1$, *i.e. connections* $s \in \mathcal{S}_1$ *reach their PCRs before the WPMM-bottleneck link rate in the centralized algorithm, then for* $t \geq T_1$, *the following statements hold for the distributed algorithm:*
   2.1. $\varphi_l > \sigma_1$ *for every link* $l \in \mathcal{L}_1$.
   2.2. *The ER field of every returning RM cell of connection* $i \in \mathcal{S}_1$ *satisfies* $\mathrm{ER} = \mathrm{PCR}^i$.
   2.3. *The ACR at source of every connection* $i \in \mathcal{S}_1$ *satisfies* $\mathrm{ACR} = \mathrm{PCR}^i$.
   2.4. $b_l^i = 1, r_l^i = \mathrm{PCR}^i$ *for every connection* $i \in \mathcal{S}_1$ *and every link* l *traversed by connection* $i \in \mathcal{S}_1$.
   2.5. 2.7. *Same as statements* 1.5–1.7, *respectively.*

The proof of Lemma 4 is given in Appendix C. Note that Lemma 4 states that not only connections $p \in \mathcal{S}_1$ have reached their WPMM rates ($\sigma_1 \cdot w_p + \mathrm{MCR}^p$ in case 1 or PCR^p in case 2), but that their rates will never change and such connection will remain marked at every link along its path.

The result of Lemma 4 will now be used as the base case for induction on the index i of $\mathcal{S}_i$.

Lemma 5 (Induction). *Suppose for some* $1 \leq i \leq K-1$, *there exists a* $T_i \geq 0$ *such that:*

1. *If* $\sigma_j < (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s$ *for* $s \in \mathcal{S}_j, 1 \leq j \leq i$, *i.e. connections* $s \in \mathcal{S}_j$ *reach the WPMM-bottleneck link rate before their PCRs in the centralized algorithm, then and for* $t \geq T_i$, *the following statements hold in the distributed algorithm:*
   1.1. $\varphi_l = \sigma_j$ *for every link* $l \in \mathcal{L}_j$.
   1.2. *The ER field of every returning RM cell of connection* $p \in \mathcal{S}_j$ *satisfies* $\mathrm{ER} = \sigma_j \cdot w_p + \mathrm{MCR}^p$.
   1.3. *The ACR at source of every connection* $p \in \mathcal{S}_j$ *satisfies* $\mathrm{ACR} = \sigma_j \cdot w_p + \mathrm{MCR}^p$.
   1.4. $b_l^p = 1, r_l^p = \sigma_j \cdot w_p + \mathrm{MCR}^p$ *for every connection* $p \in \mathcal{S}_j$ *and every link* l *traversed by connection* $p \in \mathcal{S}_j$.
   1.5. *The ER field of every returning RM cell of connection* $p \in (\mathcal{S} - (\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_i))$ *satisfies* $\mathrm{ER} > \sigma_i \cdot w_p + \mathrm{MCR}^p$.
   1.6. *The ACR at source of every connection* $p \in (\mathcal{S} - (\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_i))$ *satisfies* $\mathrm{ACR} > \sigma_i \cdot w_p + \mathrm{MCR}^p$.
   1.7. *The recorded CCR of connection* $p \in (\mathcal{S} - (\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_i))$ *satisfies* $r_l^p > \sigma_i \cdot w_p + \mathrm{MCR}^p$ *at every link* l *traversed by connection* p.
2. *If* $\sigma_j = (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s$ *for* $s \in \mathcal{S}_j, 1 \leq j \leq i$, *i.e. connections* $s \in \mathcal{S}_j$ *reach their PCRs before the WPMM-bottleneck link rate in the centralized algorithm, then for* $t \geq T_i$, *the following statements hold:*
   2.1. $\varphi_l > \sigma_j$ *for every link* $l \in \mathcal{L}_j$.
   2.2. *The ER field of every returning RM cell of connection* $p \in \mathcal{S}_j$ *satisfies* $\mathrm{ER} = \mathrm{PCR}^p$.
   2.3. *The ACR at source of every connection* $p \in \mathcal{S}_j$ *satisfies* $\mathrm{ACR} = \mathrm{PCR}^p$.
   2.4. $b_l^p = 1, r_l^p = \mathrm{PCR}^p$ *for every connection* $p \in \mathcal{S}_j$ *and every link* l *traversed by connection* $p \in \mathcal{S}_j$.
   2.5. 2.7 *Same as statements* 1.5–1.7, *respectively.*

Then there exists a $T_{i+1} \geq 0$ *such that for* $t \geq T_{i+1}$, *all statements in* 1 *and* 2 *hold for* $i+1$.

For a proof of Lemma 5, see Appendix D.

It should be clear by now that in the distributed algorithm, the convergence/marking of higher level WPMM-bottleneck link rates/connections depend on the convergence/marking of lower level WPMM-bottleneck link rates, which is similar to the case of rate calculation in the centralized algorithm.

The following theorem is the main result of this section.

Theorem 2 (Convergence theorem). *After the number of active connections in the network stabilizes, the rate allocation for each connection by the distributed protocol converges to the WPMM policy.*

Proof of Theorem 2. This theorem follows from Lemmas 4 and 5. □

We reiterate that our distributed algorithm allows the initiation of a new connection and the termination of an existing connection. Also, the ACR value of each connection is always guaranteed between its MCR and PCR. The point is that our distributed algorithm is always in the state of distributed iterations and once the set of connections in the network remain stable for a period of time, the rate allocation computed by our distributed protocol is able to converge to the WPMM policy.

Corollary 1. *Let K be the total number of iterations needed to execute the centralized algorithm for WPMM policy (Algorithm 1) and denote D the maximum round-trip time among all connections. Then an upper bound on the convergence time to the WPMM policy by our distributed algorithm from the time when the number of active connections in the network stabilizes is given by $2.5\,KD$.*

This corollary follows from the proofs of Lemmas 4 and 5. It is worthwhile to point out that this upper bound for the convergence time is a loose one. In practice, the actual convergence time of our distributed algorithm is expected to be much faster since: (1) The *maximum* RTT (D) among *all* connections is used as the worst case upper bound for each individual connection; and (2) since the ER setting in our switch algorithm (Algorithm 4) is performed on backward RM cells (rather than forward RM cells), the effective control loop for a connection is, therefore, between the source and the particular switch, rather than the full source–destination round trip used in Corollary 1.

4. Simulation results

Our work in Section 3.3 gives a proof that our distributed ABR flow control algorithm in Section 3.2 converges to the WPMM policy through distributed and asynchronous iterations. This gives us a theoretical guarantee that our distributed algorithm converges to the WPMM rate allocation under *any* network configuration and *any* set of link distances. In this section, we implement our switch algorithm on our network simulator [8] and perform simulations on various network configurations. The purpose of this section is to demonstrate the fast convergence property of our distributed algorithm.

The network configurations that we use are the peer-to-peer configuration (Fig. 1), the *parking lot* (Fig. 4), and the *generic fairness* (Fig. 6) configurations.

The ATM switches in all the simulations are assumed to have output port buffering with internal switching capacity equal to the aggregate rates of its input ports. Each output port buffer of a switch employs the simple FIFO queuing discipline and is shared by all connections going through that port. We set the link capacity to be 150 Mbps. For stability, we set the target link utilization to be 0.95. That is, we set $C_l = 0.95 \times 150\,\text{Mbps} = 142.5\,\text{Mbps}$ at every link $l \in \mathcal{L}$ for the ER calculation. By setting a target link utilization strictly less than 1, we ensure that the potential buffer build up during transient period will be emptied (or drained) upon convergence. The cell transfer delay within a switch is assumed to be 4 μs (not including queueing delay at an output port).

The distance from an end system (source or destination) to the switch is 1 km and the link distance between the switches is 1000 km (corresponding to a wide area network) and we assume that the propagation delay is 5 μs/km.

At each source, we set ICR to the MCR of the connection (or any small rate when MCR is zero) and N_{rm} to 32.

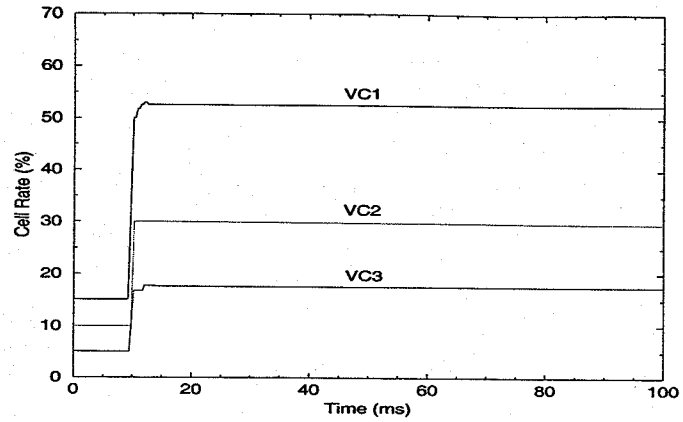


Fig. 3. The cell rates of all connections for the peer-to-peer network.

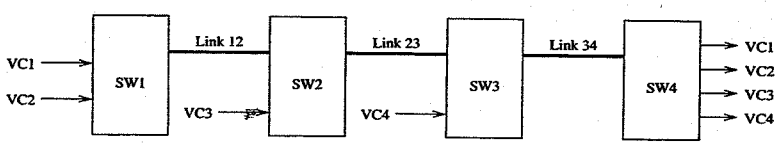


Fig. 4. A parking lot network.

4.1. Peer-to-peer network

For this network (Fig. 1), the output port link of SW1 is the only potential WPMM-bottleneck link for all connections. Under a normalized unit link capacity, the minimum rate requirement, peak rate constraint, weight, and WPMM rate allocation of each connection are listed in Table 1.

Fig. 3 shows the ACR at source for connections VC1, VC2, and VC3, respectively. The cell rates shown in the figure are normalized with respect to the capacity C_l (142.5 Mbps) for easy comparison with those values obtained with our centralized algorithm under unit link capacity in Table 1. Each connection starts with its MCR. The first RM cell for each connection returns to the source after one round trip time (RTT), or 10 ms. After initial iterations, we see that the cell rate of each connection converges to its WPMM rate listed in Table 1.

During the course of distributed iterations, the ACR of each connection in Fig. 3 maintains ABR-feasibility, i.e. MCR $\leq$ ACR $\leq$ PCR.

Also shown in Fig. 3 is that the convergence time of our ABR algorithm is much faster than the upper bound given in Corollary 1. Here the RTT is 10 ms and it takes less than 15 ms for our distributed algorithm to converge.

4.2. Parking lot network configuration

The specific parking lot network that we use is shown in Fig. 4 [11], where connections VC1 and VC2 start from the first switch and go to the last switch, and connections VC3 and VC4 start from SW2 and SW3, respectively, and terminate at the last switch.

Table 2
MCR requirement, PCR constraint, weight, and WPMM rate allocation for each connection in the parking lot network

Connection	MCR	PCR	Weight	WPMM rate allocation
VC1	0.15	0.35	4	0.2543
VC2	0.10	0.20	2	0.1522
VC3	0.10	0.50	8	0.3087
VC4	0.05	0.50	9	0.2848

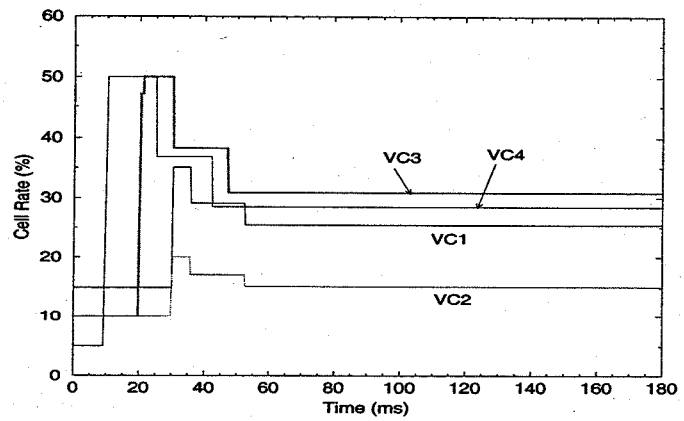


Fig. 5. The cell rates of all connections for the parking lot network.

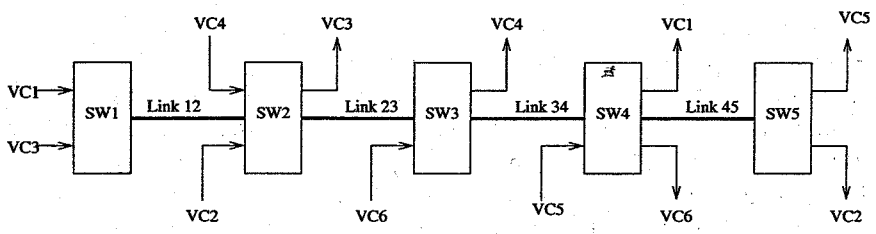


Fig. 6. The generic fairness network configuration.

Table 2 lists the MCR requirement, PCR constraint, weight, and WPMM rate allocation (obtained through the centralized algorithm) for each connection in the parking lot network under unit link capacity.

Fig. 5 shows the normalized ACR of each connection under our distributed algorithm. We find that the ACR of each connection converges to its WPMM rate listed in Table 2. Here the maximum RTT (D) among all connections is 30 ms (VC1 and VC2) and it takes our distributed algorithm less than $2D$ to converge to the final WPMM rates.

4.3. Generic fairness network configuration

The specific generic fairness configuration that we use is shown in Fig. 6 where there are five switches interconnected in a chain with six paths traversing these switches and sharing link capacity [4].

Table 3
MCR requirement, PCR constraint, weight, and WPMM rate allocation for each connection in the generic fairness network

Connection	MCR	PCR	Weight	WPMM rate allocation
VC1	0.10	1.00	4.5	0.3077
VC2	0.20	1.00	4.0	0.3846
VC3	0.20	0.60	2.0	0.6000
VC4	0.05	0.55	2.5	0.3077
VC5	0.05	0.85	4.0	0.6154
VC6	0.10	1.00	4.5	0.3077

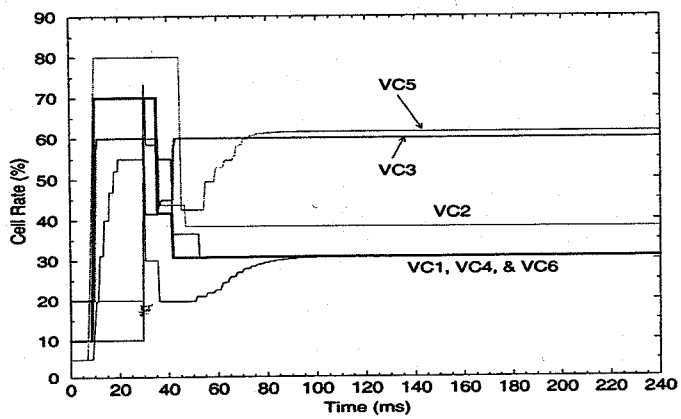


Fig. 7. The cell rates of all connections for the generic fairness network.

Table 3 lists the MCR requirement, PCR constraint, weight, and WPMM rate allocation (obtained from the centralized algorithm) for each connection under unit link capacity.

Fig. 7 shows the normalized ACR of each connection under our distributed algorithm. Again, the rate of each connection converges to its WPMM rate listed in Table 3. Here the maximum RTT (D) among all connections is 30 ms (VC1 and VC2) and it takes less than $4D$ for our distributed algorithm to converge.

In summary, based on the simulation results in this section, we have demonstrated that our distributed algorithm converges to the WPMM policy with fast convergence time.

5. Concluding remarks

The main contributions of this paper are listed as follows. We designed a distributed algorithm using the ABR flow control protocol to achieve the WPMM policy. Our switch algorithm is an extension of previous work in [5] (for the simple classical max–min policy) by integrating a connection's weight into rate calculation and proper handling of each connection's MCR and PCR constraints. We provided a formal proof that our distributed algorithm converges to the WPMM under any network configuration and any set of link distances through distributed and asynchronous iterations. Our proof generalizes the proof in [5] for the simple classical max–min policy. Simulation results demonstrated the fast convergence property of our distributed algorithm.

We stress that even though our distributed algorithm employs ABR mechanism (and thus can be applied to ATM ABR service), the underlying framework of this paper (network bandwidth sharing policy

and its distributed implementation) is fundamental, and, therefore, can be applied to any flow-oriented packet-switched networks.

Acknowledgements

This work was supported by a National Science Foundation (NSF) Graduate Research Traineeship and in part by the New York State Center for Advanced Technology in Telecommunications (CATT), Polytechnic University, Brooklyn, NY, USA.

Appendix A. Proof of Lemma 2

Let time t_0 be the time immediately after the switch algorithm is performed for an RM cell at link l. Let $\mathcal{M}_l$ and $\mathcal{U}_l$ denote the set of marked and unmarked connections at link l. By Lemma 1, the marking of connections at link l is marking-consistent. That is, any marked connection i at link l satisfies $(r_l^i - \mathrm{MCR}^i)/w_i \le \varphi_l, i \in \mathcal{M}_l$.

Case A.1. If some connections in $\mathcal{S}_l$ are not marked, i.e. $\mathcal{M}_l \neq \mathcal{S}_l$:

$$\varphi_l = \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{U}_l} w_i} \ge \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{M}_l}(\varphi_l \cdot w_i)}{\sum_{i\in\mathcal{U}_l} w_i}.$$

Then, we have,

$$\varphi_l \cdot \left(\sum_{i\in\mathcal{U}_l} w_i + \sum_{i\in\mathcal{M}_l} w_i\right) \ge C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i$$

or

$$\varphi_l \cdot \sum_{i\in\mathcal{S}_l} w_i \ge C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i.$$

Hence,

$$\varphi_l \ge \frac{C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i}{\sum_{i\in\mathcal{S}_l} w_i}.$$

Case A.2. If all connections in $\mathcal{S}_l$ are marked, i.e. $\mathcal{M}_l = \mathcal{S}_l$, we have

$$\varphi_l = \frac{C_l - \sum_{i\in\mathcal{S}_l} r_l^i}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{p\in\mathcal{S}_l}\frac{r_l^p - \mathrm{MCR}^p}{w_p}.$$

To show that

$$\varphi_l = \frac{C_l - \sum_{i\in\mathcal{S}_l} r_l^i}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{p\in\mathcal{S}_l}\frac{r_l^p - \mathrm{MCR}^p}{w_p} \ge \frac{C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i}{\sum_{i\in\mathcal{S}_l} w_i}$$

is equivalent to showing that

$$C_l - \sum_{i\in\mathcal{S}_l} r_l^i + \max_{p\in\mathcal{S}_l}\frac{r_l^p - \mathrm{MCR}^p}{w_p}\cdot\sum_{i\in\mathcal{S}_l} w_i \geq C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i,$$

which is equivalent to showing that

$$\sum_{i\in\mathcal{S}_l}\left[\left(\max_{p\in\mathcal{S}_l}\frac{r_l^p - \mathrm{MCR}^p}{w_p} - \frac{r_l^i - \mathrm{MCR}^i}{w_i}\right)\cdot w_i\right] \geq 0,$$

which trivially holds since $\max_{p\in\mathcal{S}_l}(r_l^p - \mathrm{MCR}^p)/w_p \geq (r_l^i - \mathrm{MCR}^i)/w_i$ for every $i \in \mathcal{S}_l$.

Appendix B. Proof of Lemma 3

1. In this case, first consider link $l \in \mathcal{L}_1$. Since $(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i = \sigma_1$ for $l \in \mathcal{L}_1$, and by Lemma 2, $\varphi_l \geq (C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i$, we have $\varphi_l \geq \sigma_1$ for every $l \in \mathcal{L}_1$.

   Now consider link $l \in (\mathcal{L} - \mathcal{L}_1)$, since $(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i > \sigma_1$ for $l \in (\mathcal{L} - \mathcal{L}_1)$, we have $\varphi_l > \sigma_1$ for every $l \in (\mathcal{L} - \mathcal{L}_1)$.
2. In this case, since $(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i > \sigma_1$ and by Lemma 2, $\varphi_l \geq (C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i$ for every $l \in \mathcal{L}$, we have $\varphi_l > \sigma_1$ for every $l \in \mathcal{L}$.

Appendix C. Proof of Lemma 4

1. In this case, by Lemma 3(1), there exists a $t_1 \geq 0$ such that for $t \geq t_1$,

$$\varphi_l \geq \sigma_1 \quad \text{for every } l \in \mathcal{L}_1, \tag{C.1}$$

$$\varphi_l > \sigma_1 \quad \text{for every } l \in (\mathcal{L} - \mathcal{L}_1). \tag{C.2}$$

We will show that there exists a time t_2, such that for $t \geq t_2$, the following statements hold:

$$\text{The ER field of every returning RM cell of connection } i \in \mathcal{S}_1 \text{ satisfies ER } \geq \sigma_1 \cdot w_i + \mathrm{MCR}^i. \tag{C.3}$$

$$\text{The recorded CCR satisfies } r_l^i \geq \sigma_1 \cdot w_i + \mathrm{MCR}^i \text{ at every link } l \text{ traversed by connection } i \in \mathcal{S}_1. \tag{C.4}$$

To see that (C.3) and (C.4) hold, consider that the first RM cell of connection $i \in \mathcal{S}_1$ leaves the source after time t_1. When this RM cell returns to the source at some time $t_1^{RTT} \geq t_1$, the ER field is set to

$$\mathrm{ER} := \max\{\min\{\mathrm{PCR}^i, \min_{l \text{ traversed by } i}(\varphi_l \cdot w_i + \mathrm{MCR}^i)\}, \mathrm{MCR}^i\} \tag{C.5}$$

Since (1) $(\mathrm{PCR}^i - \mathrm{MCR}^i)/w_i \geq \sigma_1$ for $i \in \mathcal{S}_1$, i.e. $\mathrm{PCR}^i \geq \sigma_1 \cdot w_i + \mathrm{MCR}^i$; and (2) $\varphi_l \geq \sigma_1$ for every $l \in \mathcal{L}_1$, we have that for $t \geq t_1^{RTT}$

$$\mathrm{ER} \geq \sigma_1 \cdot w_i + \mathrm{MCR}^i \quad \text{for } i \in \mathcal{S}_1.$$

Note that any feedback RM cell arriving at the source after time t_1^{RTT} corresponds to a forward RM cell which left the source after time t_1. Apply the above arguments to *any* such returning RM cell of connection $i \in \mathcal{S}_1$ and note that (C.1) holds for $t \geq t_1$, we have that (C.3) is true for $t \geq t_1^{RTT}$.

At time t_1^{RTT}, the ACR at the source is set to ER and $\mathrm{ACR}(t_1^{RTT}) \geq \sigma_1 \cdot w_i + \mathrm{MCR}^i$. Since (C.3) holds for $t \geq t_1^{RTT}$, we have that the ACR at source of connection $i \in \mathcal{S}_1$ satisfies $\mathrm{ACR}(t) \geq \sigma_1 \cdot w_i + \mathrm{MCR}^i$ for $t \geq t_1^{RTT}$.

Let $t_1^{1.5RTT}$ denote the time when an RM cell arrives at its destination after it leaves the source after time t_1^{RTT}. The recorded rate of connection $i \in \mathcal{S}_1$ at every link on its way is set after $t_1^{1.5RTT}$. It has already been shown that every RM cell of connection $i \in \mathcal{S}_1$ leaving the source has its CCR rate set to ACR, which is greater than or equal to $\sigma_1 \cdot w_i + \mathrm{MCR}^i$ for any time $t \geq t_1^{RTT}$. Hence the recorded rate r_l^i satisfies (C.4) for $t \geq t_1^{1.5RTT}$. Let $t_2 = t_1^{1.5RTT}$ and we have proved (C.3) and (C.4).

To prove statement 1.1 of Lemma 4, consider any link $l \in \mathcal{L}_1$. Note that in this case only connections from $\mathcal{S}_1$ traverse links of $\mathcal{L}_1$. Let $\mathcal{M}_l$ and $\mathcal{U}_l$ be the set of marked and unmarked connections, respectively. Then:

Case C.1. If not all connections are marked, then

$$\varphi_l = \frac{(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i) - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{U}_l} w_i}.$$

Since $(C_l - \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i)/\sum_{i\in\mathcal{S}_l} w_i = \sigma_1$ for $l \in \mathcal{L}_1$, we have

$$\varphi_l = \frac{\sigma_1 \cdot \sum_{i\in\mathcal{S}_l} w_i - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{U}_l} w_i} = \frac{\sigma_1 \cdot \left(\sum_{i\in\mathcal{U}_l} w_i + \sum_{i\in\mathcal{M}_l} w_i\right) - \sum_{i\in\mathcal{M}_l}(r_l^i - \mathrm{MCR}^i)}{\sum_{i\in\mathcal{U}_l} w_i}$$
$$= \sigma_1 + \frac{\sum_{i\in\mathcal{M}_l}[\sigma_1 \cdot w_i - (r_l^i - \mathrm{MCR}^i)]}{\sum_{i\in\mathcal{U}_l} w_i} \leq \sigma_1.$$

The last inequality follows from (C.4) for $t \geq t_1^{1.5RTT}$. By (C.1), $\varphi_l \geq \sigma_1$ for every link $l \in \mathcal{L}_1$ for $t \geq t_1$, we must have $\varphi_l = \sigma_1$ for $t \geq t_1^{1.5RTT}$.

Case C.2. If all connections are marked, then

$$\varphi_l = \frac{C_l - \sum_{i\in\mathcal{S}_l} r_l^i}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{i\in\mathcal{S}_l}\frac{r_l^i - \mathrm{MCR}^i}{w_i}$$
$$= \frac{\sigma_1 \cdot \sum_{i\in\mathcal{S}_l} w_i + \sum_{i\in\mathcal{S}_l}\mathrm{MCR}^i - \sum_{i\in\mathcal{S}_l} r_l^i}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{i\in\mathcal{S}_l}\frac{r_l^i - \mathrm{MCR}^i}{w_i}$$
$$= \frac{\sum_{i\in\mathcal{S}_l}[\sigma_1 \cdot w_i - (r_l^i - \mathrm{MCR}^i)]}{\sum_{i\in\mathcal{S}_l} w_i} + \max_{i\in\mathcal{S}_l}\frac{r_l^i - \mathrm{MCR}^i}{w_i} \leq \max_{i\in\mathcal{S}_l}\frac{r_l^i - \mathrm{MCR}^i}{w_i}.$$

The last inequality follows from (C.4) for $t \geq t_1^{1.5RTT}$. Since all connections are marked, $\varphi_l \geq \max_{i\in\mathcal{S}_l}((r_l^i - \mathrm{MCR}^i)/w_i)$. Thus, we must have $\varphi_l = \max_{i\in\mathcal{S}_l}((r_l^i - \mathrm{MCR}^i)/w_i)$ and $\sum_{i\in\mathcal{S}_l}[\sigma_1 \cdot w_i - (r_l^i - \mathrm{MCR}^i)] = 0$. But by (C.1), $r_l^i - \mathrm{MCR}^i \geq \sigma_1 \cdot w_i$ for connections $i \in \mathcal{S}_1$ at $t \geq t_1^{1.5RTT}$. Therefore, we must have $r_l^i - \mathrm{MCR}^i = \sigma_1 \cdot w_i$ or $(r_l^i - \mathrm{MCR}^i)/w_i = \sigma_1$ for $i \in \mathcal{S}_1$, and

$$\varphi_l = \max_{i \in \mathcal{S}_l} \frac{r_l^i - \mathrm{MCR}^i}{w_i} = \max_{i \in \mathcal{S}_l} \sigma_1 = \sigma_1$$

for connections $i \in \mathcal{S}_1$ at $t \geq t_1^{1.5RTT}$.

Combining Cases C.1 and C.2 above, statement 1.1 of Lemma 4 holds for $t \geq t_1^{1.5RTT}$.

Note that $i \in \mathcal{S}_1$ traverses at least one link $l \in \mathcal{L}_1$. By (C.5) and statement 1.1 of this lemma, any RM cell that left the destination after $t_1^{1.5RTT}$ returns to the source with the ER field set to $\sigma_1 \cdot w_i + \mathrm{MCR}^i$, $i \in \mathcal{S}_1$. Denote the time of the return of this feedback RM cell to the source by t_1^{2RTT}. This shows that statement 1.2 of the lemma is true for $t \geq t_1^{2RTT}$. It then follows that for $t \geq t_1^{2RTT}$, the ACR at the source is set to $\sigma_1 \cdot w_i + \mathrm{MCR}^i$, $i \in \mathcal{S}_1$, which is statement 1.3 of the lemma.

Let $t_1^{2.5RTT}$ be the time of an RM cell arriving at its destination after leaving the source after t_1^{2RTT}. Then by the operation of the algorithm, every connection $i \in \mathcal{S}_1$ will be marked with $b_l^i = 1$ at every link it traverses and will remain marked ever after as long as the set of connections remain unchanged for $t \geq t_1^{2.5RTT}$. Thus statement 1.4 of Lemma 4 also holds.

So far we have proved that statements (i)(a)–(d) of Lemma 4 hold for $t \geq t_1^{2.5RTT}$.

To see that statement 1.5 of Lemma 4 is true, consider that the first RM cell of connection $j \in (\mathcal{S} - \mathcal{S}_1)$ leaves the source after time t_1. When this RM cell returns to the source at some time $t_1^{RTT} \geq t_1$, the ER field is set to

$$\mathrm{ER} := \max\{\min\{\mathrm{PCR}^j, \min_{l \text{ traversed by } j} (\varphi_l \cdot w_j + \mathrm{MCR}^j)\}, \mathrm{MCR}^j\}.$$

Since (1) $(\mathrm{PCR}^j - \mathrm{MCR}^j)/w_j > \sigma_1$ for $j \in (\mathcal{S} - \mathcal{S}_1)$, i.e. $\mathrm{PCR}^j > \sigma_1 \cdot w_j + \mathrm{MCR}^j$; and (2) $\varphi_l > \sigma_1$ for every $l \in (\mathcal{L} - \mathcal{L}_1)$, we have that for $t \geq t_1^{RTT}$,

$$\mathrm{ER} > \sigma_1 \cdot w_j + \mathrm{MCR}^j \quad \text{for } j \in (\mathcal{S} - \mathcal{S}_1).$$

Now using similar arguments as above for the proofs of (C.3) and (C.4), and taking (C.2) into account, it can be shown that statements 1.5–1.7 hold for $t \geq t_1^{1.5RTT}$.

Let $T_1 = t_1^{2.5RTT}$ and all statements of Lemma 4(1.1–1.7) are proved.

2. In this case, by Lemma 3(2), there exists a $t_1 \geq 0$ such that for $t \geq t_1$,

$$\varphi_l > \sigma_1 \quad \text{for every } l \in \mathcal{L}. \tag{C.6}$$

Therefore, statement 2.1 of Lemma 4 is true for $t \geq t_1$. To see that statements 2.2–2.4 of Lemma 4 hold, consider the first RM cell of connection $i \in \mathcal{S}_1$ leaves the source after time t_1. When this RM cell returns to the source at some time $t_1^{RTT} \geq t_1$, the ER field is set to

$$\mathrm{ER} := \max\{\min\{\mathrm{PCR}^i, \min_{l \text{ traversed by } i} (\varphi_l \cdot w_i + \mathrm{MCR}^i)\}, \mathrm{MCR}^i\}.$$

By (C.6), $\varphi_l > \sigma_1 = (\mathrm{PCR}^i - \mathrm{MCR}^i)/w_i$ for $i \in \mathcal{S}_1$, we have for $t \geq t_1^{RTT}$,

$$\mathrm{ER} = \mathrm{PCR}^i \quad \text{for } i \in \mathcal{S}_1. \tag{C.7}$$

Note that any feedback RM cell arriving at the source after time t_1^{RTT} corresponds to a forward RM cell which left the source after time t_1. Apply the above arguments to *any* such returning RM cell of connection $i \in \mathcal{S}_1$ and note that (C.6) holds for $t \geq t_1$, we have that (C.7) is true for $t \geq t_1^{RTT}$.

At time t_1^{RTT}, the ACR at the source is set to ER and $\text{ACR}(t_1^{RTT}) = \text{PCR}^i$. Since (C.7) holds for $t \geq t_1^{RTT}$, we have that ACR at source satisfies $\text{ACR}(t) = \text{PCR}^i$ for $t \geq t_1^{RTT}$.

Let $t_1^{1.5RTT}$ denote the time when an RM cell arrives at its destination after it leaves the source after time t_1^{RTT}. The recorded rate of connection $i \in \mathcal{S}_1$ at every link on its way is set after $t_1^{1.5RTT}$. It has already been shown that every RM cell of connection $i \in \mathcal{S}_1$ leaving the source has its CCR rate equal to PCR for $t \geq t_1^{RTT}$. Hence the recorded rate r_l^i satisfies $r_l^i = \text{PCR}^i$ for $t \geq t_1^{1.5RTT}$. In addition, since $(\text{PCR}^i - \text{MCR}^i)/w_i = \sigma_1 < \varphi_l$ for $i \in \mathcal{S}_1$ in this case, the b_l^i bit at link l for connection $i \in \mathcal{S}_1$ is marked to 1 for $t \geq t_1^{1.5RTT}$. Thus, we have shown that statements 2.1–2.4 of Lemma 4 hold for $t \geq t_1^{1.5RTT}$.

Similarly, to see that statements 2.5–2.7 of Lemma 4 hold, consider the first RM cell of connection $j \in (\mathcal{S} - \mathcal{S}_1)$ leaves the source after time t_1. When this RM cell returns to the source at some time $t_1^{RTT} \geq t_1$, the ER field is set to

$$\text{ER} := \max\{\min\{\text{PCR}^j, \min_{l \text{ traversed by } j}(\varphi_l \cdot w_j + \text{MCR}^j)\}, \text{MCR}^j\}.$$

Since (1) $(\text{PCR}^j - \text{MCR}^j)/w_j > \sigma_1$ for $j \in (\mathcal{S} - \mathcal{S}_1)$, i.e. $\text{PCR}^j > \sigma_1 \cdot w_j + \text{MCR}^j$; and (2) $\varphi_l > \sigma_1$ for every $l \in \mathcal{L}$, we have that for $t \geq t_1^{RTT}$,

$$\text{ER} > \sigma_1 \cdot w_j + \text{MCR}^j \quad \text{for } j \in (\mathcal{S} - \mathcal{S}_1). \tag{C.8}$$

Using similar arguments for the proof of statements 2.3 and 2.4 of Lemma 4, together with (C.8) above, we see that statement 2.6 holds for $t \geq t_1^{RTT}$ and statement 2.7 holds for $t \geq t_1^{1.5RTT}$.

Let $T_1 = t_1^{1.5RTT}$ and Lemma 4 (2.1–2.7) is proved.

Remark C.1. We have just shown that it takes at most two and a half maximum round-trip time (RTT) for every connection in $\mathcal{S}_1$ to reach its WPMM rate and is marked with this rate at all links along its path.

Appendix D. Proof of Lemma 5

By the induction hypothesis, for $t \geq T_i$, (1) every connection $p \in \mathcal{S}_j, 1 \leq j \leq i$ has reached its WPMM rate $\sigma_j \cdot w_p + \text{MCR}^p$ (in Lemma 4(1)) or PCR^p (in Lemma 4(2)) and these rates do not change as long as the set of connections in the network remain unchanged; and (2) every connection $p \in \mathcal{S}_j, 1 \leq j \leq i$ is marked with its WPMM rate $\sigma_j \cdot w_p + \text{MCR}^p$ or PCR^p along its traversing links.

Consider a reduced network $\hat{\mathcal{N}}$ with links $\hat{\mathcal{L}} = \mathcal{L}_{i+1} \cup \mathcal{L}_{i+2} \cup \cdots \cup \mathcal{L}_K$,[7] connections $\hat{\mathcal{S}} = \mathcal{S} - (\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_i) = \mathcal{S}_{i+1} \cup \cdots \cup \mathcal{S}_K$ and link capacities $\hat{C}_l = C_l - \sum_{p \in (\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_i) \text{ traversing link } l} r_l^p, l \in \hat{\mathcal{L}}$. Note that it is legitimate to consider the reduced network because by the induction hypothesis every connection in $(\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_i)$ has reached its WPMM rate and is marked at all the traversing links along its path with its WPMM rate for $t \geq T_i$. Denote $\hat{n}_l$ the number of connections traversing link l in the reduced network $\hat{\mathcal{N}}$.

[7] Note that $\hat{\mathcal{L}} = \mathcal{L}_{i+1} \cup \mathcal{L}_{i+2} \cup \cdots \cup \mathcal{L}_K$ may not be the same as $\mathcal{L} - (\mathcal{L}_1 \cup \mathcal{L}_2 \cup \cdots \cup \mathcal{L}_i)$ since links in $\mathcal{L}_i$ may be part of $\mathcal{L}_{i+1}$.

For the reduced network $\hat{\mathcal{N}}(\hat{\mathcal{L}}, \hat{\mathcal{S}}, \hat{C})$, reapplying the arguments used in the proof of Lemma 2, we have

$$\varphi_l \geq \frac{\hat{C}_l - \sum_{i \in \hat{\mathcal{S}}_l} \mathrm{MCR}^i}{\sum_{i \in \hat{\mathcal{S}}_l} w_i} \quad \text{for every } l \in \hat{\mathcal{L}}.$$

Using similar arguments as for the proof of Lemma 3, it is straightforward to show that statements similar to Lemma 3 hold for the reduced network. That is,

1. If $\sigma_{i+1} = (\hat{C}_l - \sum_{i \in \hat{\mathcal{S}}_l} \mathrm{MCR}^i)/\sum_{i \in \hat{\mathcal{S}}_l} w_i \leq (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s$ for $s \in \mathcal{S}_{i+1}$, i.e. connections $s \in \mathcal{S}_{i+1}$ reach the WPMM-bottleneck link rate before their PCRs, then,

$$\varphi_l \geq \sigma_{i+1} \quad \text{for every } l \in \mathcal{L}_{i+1}, \qquad \varphi_l > \sigma_{i+1} \quad \text{for every } l \in (\hat{\mathcal{L}} - \mathcal{L}_{i+1}).$$

2. If $\sigma_{i+1} = (\mathrm{PCR}^s - \mathrm{MCR}^s)/w_s < (\hat{C}_l - \sum_{i \in \hat{\mathcal{S}}_l} \mathrm{MCR}^i)/\sum_{i \in \hat{\mathcal{S}}_l} w_i$ for $s \in \mathcal{S}_{i+1}$, i.e. connections $s \in \mathcal{S}_{i+1}$ reach their PCRs before the WPMM-bottleneck link rate, then,

$$\varphi_l > \sigma_{i+1} \quad \text{for every } l \in \hat{\mathcal{L}}.$$

Now repeat the proof of Lemma 4 for the reduced network, and we can show that all the statements of Lemma 5 hold for $i + 1$.

References

[1] A. Arulambalam, X. Chen, N. Ansari, Allocating fair rates for available bit rate service in ATM networks, IEEE Commun. Mag. 34 (1996) 92–100.

[2] The ATM Forum Technical Committee, Traffic Management Specification, Version 4.0, ATM Forum Contribution, AF-TM 96-0056.00, 1996.

[3] D. Bertsekas, R. Gallager, Data Networks, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[4] F. Bonomi, K.W. Fendick, The rate-based flow control framework for the available bit rate ATM service, IEEE Network Mag. 9 (1995) 25–39.

[5] A. Charny, D. Clark, R. Jain, Congestion control with explicit rate indication, Proceedings of the IEEE International Conference on Communication, 1995, pp. 1954–1963.

[6] E.M. Gafni, The integration of routing and flow control for voice and data in a computer communication network, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, 1982.

[7] H.P. Hayden, Voice flow control in integrated packet networks, M.S. Thesis, Department of Electrical Engineering and Computer Science, MIT, 1981.

[8] A. Heybey, The network simulator, version 2.1, Laboratory of Computer Science, MIT, 1990.

[9] Y.T. Hou, H. Tzeng, S.S. Panwar, V.P. Kumar, ATM ABR traffic control with a generic weight-based bandwidth sharing policy: theory and a simple implementation, IEICE Trans. Commun. E81-B (1998) 958–972.

[10] J.M. Jaffe, Bottleneck flow control, IEEE Trans. Commun. COM-29 (1981) 954–962.

[11] R. Jain, Congestion control and traffic management in ATM networks: recent advances and a survey, Comput. Networks ISDN Syst. 29 (1997) 887–895.

[12] R. Jain et al., ERICA Switch Algorithm: A Complete Description, ATM Forum Contribution, AF-TM 96-1172, 1996.

[13] L. Kalampoukas, A. Varma, K.K. Ramakrishnan, Dynamics of an explicit rate allocation algorithm for available bit rate (ABR) service in ATM networks, Proceedings of the IFIP International Conference on High Performance Networking, 1995, pp. 143–154.

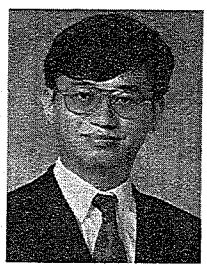
[14] F.P. Kelly, Charging and rate control for elastic traffic, Eur. Trans. Telecommun. 8 (1997) 33–37.

[15] J. Mosely, Asynchronous distributed flow control algorithms, Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, 1984.

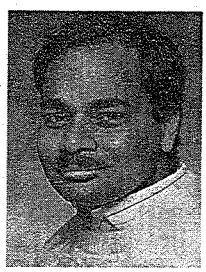
[16] K.K. Ramakrishnan, R. Jain, D.-M. Chiu, Congestion avoidance in computer networks with a connectionless network layer – part IV: a selective binary feedback scheme for general topologies methodology, Tech. Report DEC-TR-510, Digital Equipment Corporation, 1987.

[17] L. Roberts, Enhanced PRCA (Proportional Rate Control Algorithm), ATM Forum Contribution, AF-TM 94-0735R1, 1994.
[18] K.-Y. Siu, H.-Y. Tzeng, Intelligent congestion control for ABR service in ATM networks, ACM SIGCOMM Comput. Commun. Rev. 24 (1994) 81–106.
[19] N. Yin, M.G. Hluchyj, On closed-loop rate control for ATM cell relay networks, Proceedings of the IEEE INFOCOM, 1994, pp. 99–108.

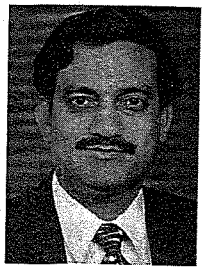
Yiwei Thomas Hou obtained his B.E. degree (Summa Cum Laude) from the City College of New York in 1991, the M.S. degree from Columbia University in 1993, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1997, all in Electrical Engineering. He was awarded a National Science Foundation Graduate Research Traineeship for pursuing Ph.D. degree in high speed networks, and was the recipient of Alexander Hessel award for outstanding Ph.D. dissertation in 1998 from Polytechnic University. While a graduate student, he worked at AT&T Bell Labs, Murray Hill, NJ, during the summers of 1994 and 1995, on implementations of IP and ATM internetworking; he also worked at Bell Labs, Lucent Technologies, Holmdel, NJ, during the summer of 1996, on fundamental problems on network traffic management. Since September 1997, he has been a Research Staff Member at Fujitsu Laboratories of America, Sunnyvale, CA. His current research interests are in the areas of next generation Internet architecture, protocols, and implementations for differentiated and integrated services. He is a member of the IEEE, ACM, and Sigma Xi.



Henry Tzeng received his B.S. degree from the Tatung Institute of Technologies, Taiwan, Republic of China, in 1988. He received his M.S. and Ph.D. degrees in Electrical Engineering from the University of California, Irvine, in 1993 and 1995, respectively. He was a recipient of the 1997 IEEE Browder J. Thompson Memorial Prize Award and also the University of California Regent's Dissertation Fellowship in 1995. Since 1995, he has been a Member of Technical Staff at Bell Labs, Lucent Technologies, Holmdel, NJ. His research interests are in the area of algorithm design for high-speed networks and fault-tolerant distributed systems. In particular, he has been working on the issues related to reliable multicast, congestion control protocols for unicast/multicast ATM ABR services, TCP performance over ATM networks. He is currently working on the high performance routing technologies for the Internet.



Shivendra S. Panwar received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Kanpur, in 1981, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Massachusetts, Amherst, in 1983 and 1986, respectively. He joined the Department of Electrical Engineering at the Polytechnic Institute of New York, Brooklyn (now Polytechnic University), where he is now an Associate Professor. Since 1996, he has served as the Director of the New York State Center for Advanced Technology in Telecommunications (CATT). He spent the summer of 1987 as a Visiting Scientist at the IBM T.J. Watson Research Center, Yorktown Heights, NY, and has been a Special Consultant to AT&T Bell Labs, Holmdel, NJ. He has also worked with NYNEX Science and Technology, Fujitsu Network Communications, Sumitomo Electric Industries, Lucent Technologies, Chase Manhattan Bank, Securities Industries Automation Corporation, Frequency Electronics, LNR, Inc., and the US Department of State, on a variety of projects. His research interests include the performance analysis and design of high speed networks, which has been supported by the National Science Foundation. He is a member of Tau Beta Pi and Sigma Xi and a senior member of the IEEE. He has served as the Secretary of the Technical Affairs Council of the IEEE Communications Society (1992–1993). He was the Co-chairman of the Technical Program Committee and a member of the Organizing Committee of the Second IEEE Network Management and Control Workshop, Tarrytown, NY, September 1993. He has also been a member of the Technical Program Committee of INFOCOM'90, '93 and '97. He is the co-editor of two books, Network Management and Control, vol. II, and Multimedia Communications and Video Coding, both published by Plenum.



Vijay P. Kumar is currently Director of Advanced Internetworking Systems at Lucent Technologies and Head of High Speed Networks Research Department at Bell Labs. He is responsible for research and development in algorithms, architectures, protocols, chips and systems for high speed data networking. He obtained his B.E. degree in Electronics and Communication Engineering from Osmania University, Hyderabad, India, in 1980, and M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Iowa, Iowa City, IA, in 1982 and 1985, respectively. He has been with Bell Labs since 1985, where he has conducted and led research activities in VLSI switch architectures, multicast routing, traffic management, and VLSI yield enhancement. His work has resulted in three generations of ATM switching chip sets and a high speed, QoS-capable IP router.