Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Computer Networks 41 (2003) 161–176

COMPUTER NETWORKS

www.elsevier.com/locate/comnet

Loss differentiated multicast congestion control

Yung-Sze Gan *, Chen-Khong Tham

Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119260, Singapore

Received 10 March 2002; received in revised form 31 July 2002; accepted 20 September 2002

Responsible Editor: J. Crowcroft

Abstract

A new layered multicast scheme known as loss differentiated multicast congestion control architecture is proposed to provide effective congestion control in heterogeneous multicast networks. It is comprised of two components, the random early detection assisted layered multicast (RALM) and the layer marking discovery protocol (LMDP). The RALM protocol utilises the packet marking and priority dropping mechanisms of the differentiated services architecture to differentiate losses in the layers of a layered multicast session. The LMDP protocol assists in the discovery of the optimal subscription levels to which packets of the layers should be marked. By marking the layers' packets appropriately and dropping them differently during congestion, the RALM protocol guides receivers to their stable optimal subscription levels that satisfy their bandwidth requirements while providing multirate max–min fairness in the network.

Keywords: Layered multicast; Loss differentiation; Multicast congestion control

1. Introduction

The widespread deployment of congestion control protocols like TCP in unicast IP networks has helped them to perform well in the face of ever increasing amount of traffic. Despite the efficiency with which IP multicast utilises network bandwidth, congestion can still occur and hence the success of multicast networks will similarly be dependent on the implementation of equivalent congestion control schemes. Due to the heterogeneous nature of multicast networks, layered multicast protocols had been proposed to fulfill the congestion control needs.

Layered multicast protocols can be classified into receiver-driven and feedback-driven congestion control systems. In receiver-driven systems [5,9], the receivers do not communicate with their sources or intermediate routers during the optimal subscription level discovery process. These systems generally suffer from significant loss rates due to their inducement of congestion to estimate bottleneck bandwidths. Although this bandwidth probing method is critical to the discovery of the optimal subscription level, they could have

* Corresponding author.
E-mail addresses: engp0515@nus.edu.sg (Y.-S. Gan), eletck@nus.edu.sg (C.-K. Tham).

1389-1286/03/$ - see front matter
PII: S1389-1286(02)00372-9

minimised the impact of the induced losses on the data stream decoding by exploiting the relative importance of the cumulative data layers [2] to confine packet losses to the higher layers. They also do not perform well when competing against TCP flows [8] even if they implement TCP-friendly algorithms. And they have difficulty in providing multirate max–min fairness as defined in [6] to all sessions in a network.

In feedback-driven congestion control systems [7,19], the receivers interact with their sources and intermediate routers through feedbacks during the optimal subscription level discovery process. Through close cooperation among receivers, sources and routers, these systems can provide multirate max–min fairness in multicast networks and assure low loss rate in layers constituting the optimal subscription levels through the deployment of priority dropping mechanisms. However, feedback-driven systems suffer from feedback implosion problem and slow reaction due to feedback latency. Thus, complex mechanisms are usually implemented in these systems to overcome their weaknesses.

The loss differentiated multicast congestion control (LDMCC) architecture is a feedback-driven system that employs layer marking and priority dropping mechanisms to discover the optimal subscription levels in a layered multicast session. In many ways, the LDMCC architecture is similar to the receiver-selectable loss priorities (RSLP) and receiver-driven layered multicast with priorities (RLMP) protocols proposed in [7]. Layers that constitute the optimal subscription levels are marked with low drop precedence while other layers are marked with high drop precedence. The receivers provide feedbacks of their bottleneck bandwidth estimates to the network that influence the layers' priority marking. However, the LDMCC architecture does not require all routers to process feedbacks and no modification to the IGMP [13] protocol is needed to support layer priority indication.

The LDMCC architecture has two components, namely the random early detection (RED) assisted layered multicast (RALM) and the layer marking discovery protocol (LMDP). The RALM protocol provides the congestion control algorithm that guide receivers to the optimal subscription levels of a session. The LMDP protocol facilitates the operations of the RALM protocol by discovering the optimal subscription levels in the session. These two protocols are described in details in the next two sections. Finally, simulations are performed using the *ns* simulator [18] to study the characteristics of the LDMCC architecture in the remaining sections.

2. RED assisted layered multicast

In the RALM protocol, an algorithm that provides loss differentiated congestion control functionality is implemented in every receiver. This algorithm requires the cooperation of the source and routers to provide layer differentiation through the use of two-level priority packet marking and dropping mechanisms. The following subsections examine how these mechanisms can be implemented in an IP network, and describe the congestion control algorithm.

2.1. Two-level drop priority support in IP networks

The IETF has defined the differentiated services (DiffServ) architecture [11] and two per-hop forwarding behaviours (PHB) [14,15] to permit the implementation of packet service differentiation in current IP networks. Of the two PHBs defined by the IETF, the assured forwarding (AF) PHB group provides 3 drop precedence levels within 4 AF PHB classes. The AF PHB classes are implemented as four separate physical queues in a DiffServ router [10] in addition to the best-effort droptail queue. The drop precedences in each AF class are in turn implemented as virtual queues in each AF queue through the use of a three-level RED algorithm. By using a single AF PHB class (e.g. AF1x) and restricting the number of drop precedences used to just two (e.g. AF11 and AF12), two-level loss differentiation can be realised in the network routers.

In the RALM protocol, the layer markings are made by the source of the layered multicast session or its access router, and packet markers along its multicast tree. When the source generates the

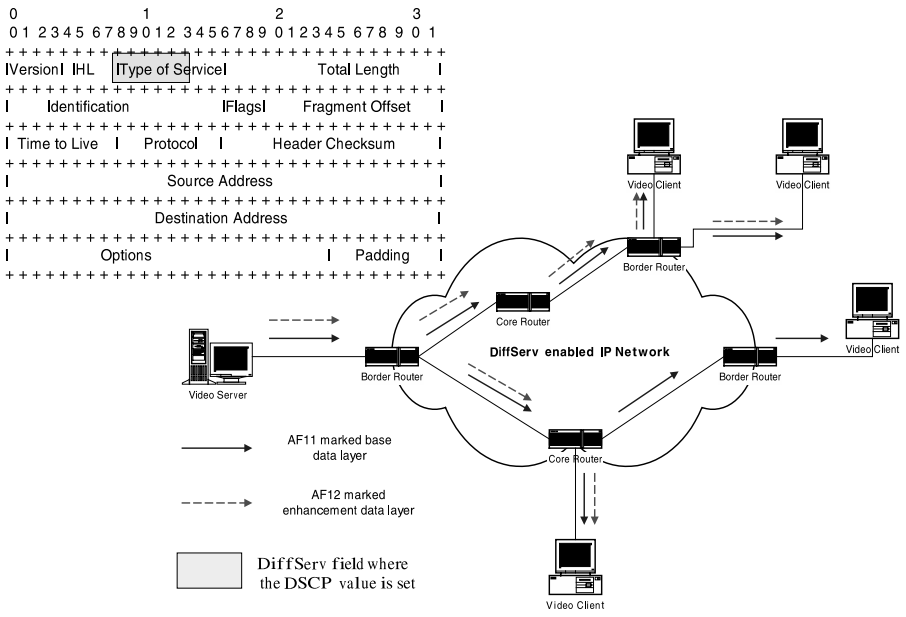

Fig. 1. Forwarding packets marked with different DSCP values through a DiffServ network to receivers of an RALM session.

layers, it marks the IP headers of packets in each layer with either an AF11 or AF12 DiffServ codepoint (DSCP) value before transmitting them into the network. However, the source may lack the packet marking functionality and thus its access router must perform the DSCP marking of the layers on its behalf. Packet markers along the multicast tree can adjust the layer markings to match the optimal subscription levels of their downstream bottlenecks. According to RFC 2597, packets marked with AF12 DSCP value will be dropped in preference to packets marked with AF11 DSCP value when congestion occurs. The DiffServ routers in the multicast tree can just forward the packets in each layer based on their DSCP values as shown in Fig. 1.

To provide loss differentiation within a layered multicast session, the layers within the optimal subscription level should be marked with the AF11 DSCP value while other layers are marked with the AF12 DSCP value. As the bottleneck router forwards the AF11 layers at the expense of the AF12 layers, the layers that constitute the optimal subscription level will be protected from losses induced by oversubscription or traffic fluctuations. These losses are absorbed by the AF12 layers. Therefore, the AF11 layers are called protected layers and the AF12 layers are called sacrificial layers. The AF DSCP values used here may be replaced by alternative DSCP values as long as the drop precedences of the protected and sacrificial layers are preserved.

In the RALM protocol, all multicast packets are marked with an AF DSCP value while the best-effort traffic is not marked at all (i.e. DSCP value is zero). Therefore, the layered multicast traffic in the DiffServ network is completely segregated from the best-effort traffic as they are served from different physical queues in the routers as illustrated in Fig. 2. Traffic segregation

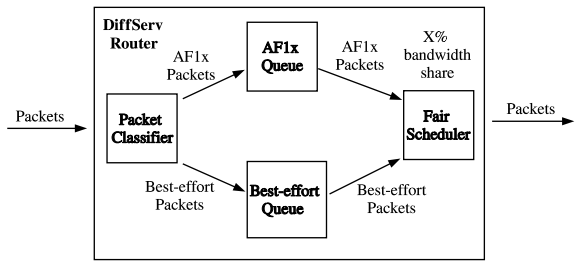

Fig. 2. Segregation of RALM and best-effort traffic in a DiffServ router.

decouples the RED queues from any TCP flows that might traverse the network as they are part of the best-effort traffic. Thus, there is no necessity to consider how the TCP flows will be affected by the RED parameters configured solely on the basis of the RALM protocol for the AF queues. This is beneficial as it has been shown that setting proper RED parameters to allow TCP flows to work with different bandwidth adaptive flows is difficult [4]. The DiffServ router can partition the link bandwidth between the multicast and unicast traffic fairly through the use of a fair queue scheduler. In this paper, a simple round robin scheduler is used to serve the two queues, dividing the link bandwidth equally between them.

The RED parameters of the AF virtual queues are not configured based on any strict criteria. The only two requirements that must be kept in mind when choosing their values are that they must be close to the values recommended by the network community [3,17] and that the AF12 queue must be penalised with significantly higher losses than the AF11 queue when congestion occurs. In the RALM protocol, it is decided that the AF12 queue's drop thesholds must be shorter than that of the AF11 queue and its maximum drop probability must be higher than the corresponding parameter in the AF11 queue. The net result is AF12 packets are dropped sooner and faster than AF11 packets when the router starts to experience queue buildup due to congestion. Fig. 3 shows the queue thresholds and maximum drop probabilities that may be configured in the AF virtual queues.

RED Parameters	AF11 Queue	AF12 Queue
Minimum queue threshold $\min_{th}$	10	5
Maximum queue threshold $\max_{th}$	30	15
Maximum drop probability $\max_{p}$	0.1	0.2

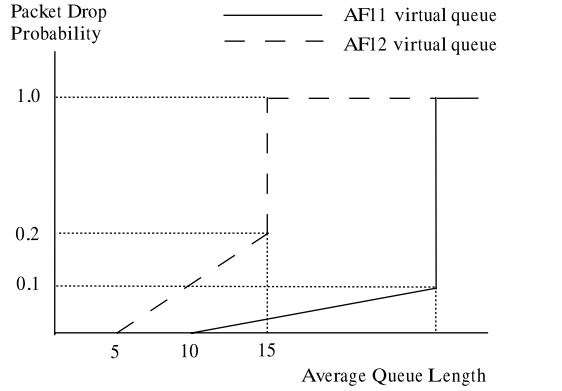

Fig. 3. RED parameter values of AF virtual queues superimposed on a single graph.

2.2. Receiver-driven congestion control algorithm

The optimal subscription level of an RALM session along a branch of its multicast tree is achieved when two conditions are satisfied:

(1) there are no excessive losses in the AF11 marked protected layers such that there is a need to decrease the current subscription level, and
(2) there are sufficient losses in the AF12 marked sacrificial layers to avoid subscribing more than one sacrificial layer.

For a given bottleneck link shared by a number of sessions, the optimal subscription level of each session is achieved when they are able to share the bandwidth in a max–min fair manner [1]. Once the layer markings in an RALM session match its optimal subscription level, the receivers will be able to reach this level by adding layers until they receive the first sacrificial layer.

Basically, the receiver joins an RALM session by subscribing the minimal two layers and increases its subscription level by using its current highest subscribed layer as a probe layer. The loss rate sampled in the probe layer is compared to the layer add loss threshold:

$$T_{\text{add}} = \max_{\text{add}} -\delta_{\text{add}} l, \tag{1}$$

where $\max_{\text{add}}$ is the maximum loss rate beyond which no new layer is added, δ_{add} is the separation between adjacent subscription levels' T_{add}, and l is the current subscription level of the receiver. If the sampled loss rate is lower than this threshold, a new layer is added unless all layers in the session are subscribed. The aggregated loss rate sampled in the lower layers is compared to the layer drop loss threshold:

$$T_{\text{drop}} = \max_{\text{drop}} -\delta_{\text{drop}} l, \tag{2}$$

where $\max_{drop}$ is the maximum loss rate beyond which all layers are dropped, δ_{drop} is the separation between adjacent subscription levels' T_{drop}, and l is the current subscription level of the receiver. If the sampled loss rate is higher than this threshold, the probe layer will be dropped unless the minimum subscription level of two layers is reached.

A receiver signals its desire to drop a layer by sending an IGMP leave message to its leaf router. However, the leaf router does not stop forwarding the dropped layer immediately. Instead, the router must query its downstream nodes about their multicast membership to ensure that no other receiver is requesting the dropped layer. This membership query process usually lasts for a few seconds before the multicast leave procedure is executed. As a result, an RALM receiver samples the loss rates in current subscribed layers using a 10 s measurement window to account for the long IGMP leave latency.

When a receiver's subcription level is below the optimal level, all its subscribed layers are protected from congestion due to their AF11 marking. The loss rate sampled in its probe layer will be lower than the current level's T_{add}, which is a good indication that there is sufficient bandwidth for the receiver to increase its subscription level. To avoid a contradicting layer drop indication by the aggregated loss rate in the lower layers, T_{add} is set below T_{drop} of each subscription level as shown in Fig. 4. Thus, the receiver adds a new layer to its current subscription. This is continuously done until the receiver adds a sacrificial layer. Since the optimal subscription level is near the bottleneck bandwidth, adding the sacrificial layer causes the probe layer's loss rate to rise drastically beyond the current level's T_{add}. However, the subscribed protected layers' aggregated loss rate is still below the current level's T_{drop}. Now, the receiver has converged on the optimal subscription level, and will neither increase nor decrease its subscription level. Conversely, if the receiver's subscription level is above the optimal subscription level, part of the aggregated loss rate will be sampled from sacrificial layers. Thus, the excessive losses in these layers will cause the receiver to drop them as its current subscription level's T_{drop} is exceeded.

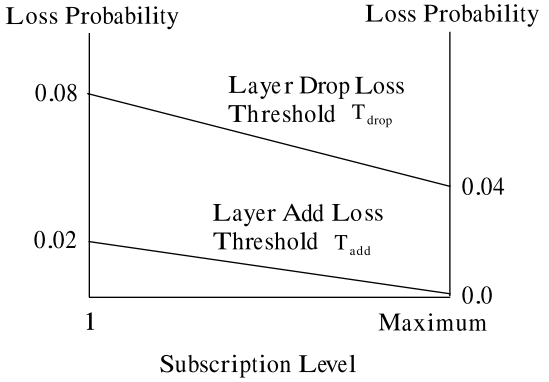

Fig. 4. The range of values that T_{add} and T_{drop} in an RALM session can take.

An RALM receiver converges on the optimal subscription level by oversubscribing a sacrificial layer. The condition of convergence requires the probe layer to suffer excessive losses while layers within the optimal subscription level are protected from the ongoing congestion. Thus, the loss rate performance of the RALM protocol is measured from the layers constituting the optimal subscription level and excluding the probe layer. Whether to utilise the data received in the probe layer is decided by individual receivers.

The degree of congestion in a bottleneck router is reflected in the loss rate curves sampled by all downstream receivers. The implicit sharing of loss rate knowledge can help receivers in different RALM sessions to cooperate in the discovery of their optimal subscription levels that share the bottleneck bandwidth fairly. This cooperation is realised by scaling the T_{drop} and T_{add} with the subscription level. Thus, for two RALM sessions that stream layers of the same granularity, the session at a lower subscription level has a higher loss tolerance than the other session at a higher subscription level. Through the judicious use of layer markings, a new session can force an incumbent session to drop its subscription level by causing a lowering of the optimal subscription level that the competing sessions should converge on. This new optimal subscription level should represent the max–min fair share of the bottleneck bandwidth between the RALM sessions.

3. Layer marking discovery protocol

The RALM protocol only provides the loss differentiated congestion control algorithm to guide receivers to the optimal subscription levels. However, it does not describe how these optimal subscription levels are discovered so that packets of the session's layers are marked appropriately. The LMDP protocol is designed to assist the RALM session in the discovery of suitable packet markers in the DiffServ network, and facilitates the cooperation between markers and receivers in the discovery of the optimal subscription levels.

3.1. Discovery and maintenance of the marking tree

In an RALM session, a source generates cumulative layers of the same data rate and transmits each layer to a separate multicast group. The multicast trees constructed for the layers overlap, with the base layer multicast tree covering the entire scope of the session. Basically, the multicast trees can be seen as a single tree with branches of different thickness that match the optimal subscription levels achieved by the receivers. In the RALM protocol, receivers start their subscriptions from the base layer. To discover the markers of an RALM session, the source transmits pathfinder packets down the multicast tree of the base layer only. In this way, all receivers learn the locations of their upstream markers regardless of their subscription levels.

A pathfinder packet is an IP packet with an unique protocol number. This protocol number signals packet markers like DiffServ border routers to process it. Other routers that do not understand this protocol number just forward the packet downstream like a normal IP packet. The pathfinder packet contains a 8-bit field that indicates the number of layers available in the RALM session and a 32-bit field that contains the IP address of the previous packet marker. The layer number field is set by the source and is not changed in the network. The previous marker address field is initially set to either the source's address if it can mark packets or zero if it cannot. If the source is not capable of packet marking, its access router must be a packet marker and inserts its address in the pathfinder packets. The previous marker address field is updated by every marker along the multicast tree before it forwards the pathfinder packet.

Like the source path messages in the pragmatic general multicast (PGM) protocol [16], the pathfinder packets install and maintain the RALM session information in a marker like a DiffServ border router. The marker executes the following sequence of actions when it receives a pathfinder packet:

(1) Identify the RALM session by extracting the pathfinder packet's source IP address and destination multicast address. The concatenation of these two addresses adequately identifies the RALM session because the layered data is transmitted to contiguous multicast addresses beginning with the base layer's.
(2) Search its session information table for an entry of the identified session. If the entry does not exist, create one by recording the source and destination addresses, the number of layers and the address of the previous marker in the table. A timer is then started for the entry. If the entry exists, reset its timer.
(3) Update the previous marker address field with its own address before forwarding the pathfinder packet down its branch of the multicast tree.

This sequence of actions is repeated at every marker along the multicast tree of the RALM session. When the pathfinder packet reaches a receiver, the receiver records the address of the previous marker which is the target of its feedbacks for the session. The description of pathfinder packet processing is illustrated in Fig. 5. Once the pathfinder packets have terminated at all receivers in an RALM session, the entire hierarchy of markers known henceforth as the marking tree would have been discovered. Feedbacks of bottleneck bandwidth estimates can be sent up the tree so that the markers can mark the session's layers to the optimal subscription levels discovered by the receivers.

Like the flow-state information in the RSVP protocol [12], the session information in the

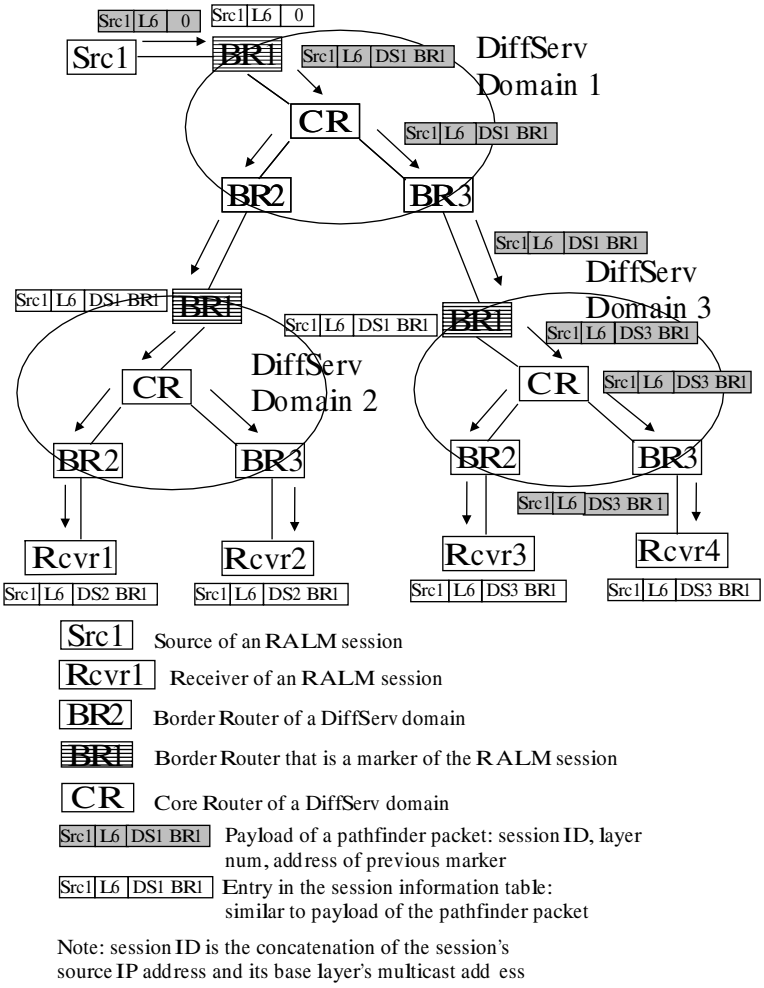

Fig. 5. Discovering markers along a path of an RALM session.

markers is soft-state in nature, i.e. the session entry is deleted if no pathfinder packet of the session is received before its timer timeouts. Making the session information soft-state has the advantage that the marking tree can adapt to changes in the multicast tree as the session membership changes. When a branch of the multicast tree is added or pruned, the paths taken by the pathfinder packets change together with that of data packets. With the timeout of the session entry timer, old markers drop out from the marking tree without the need of explicit signaling. In the LMDP protocol, a pathfinder packet is sent every 5 min and the session entry timer is set to 15 min in order to balance between the need for low pathfinding traffic and fast adaptation to multicast tree changes.

3.2. Optimal subscription level estimation algorithm

An RALM receiver estimates its bottleneck bandwidth by tracking its subscription level. Essentially, the subscription level achieved by a receiver is indicative of the amount of session traffic that can traverse the bottleneck. By tracking the subscription level changes over a time period, the receiver can estimate an optimal subscription level for its branch of the multicast tree that it thinks matches its bottleneck bandwidth. The estimate is

sent to the marking tree in feedback packets and the markers mark the session's layers to reflect the estimated optimal subscription level. Then, the receiver tracks its subscription level over a new time period before sending feedbacks of its current estimate to the marking tree to update their marking levels. The objective of this close-loop control system is to achieve an unchanging estimate of the optimal subscription level that implicitly matches the bottleneck bandwidth seen by the receiver.

Statistics of the subscription levels achieved by an RALM receiver is kept for an estimation period of 90–100 s. Based on the 10 s measurement window in the RALM protocol, 9–10 subscription level samples are obtained within the estimation period to form an analysable set of data. As the RALM protocol drives the receiver to increase its subscription level as high as possible, the distribution of the subscription levels seen in an estimation period is biased towards the optimal subscription level that matches the bottleneck bandwidth. Based on this observation, the optimal subscription level for an estimation period is determined in the following manner:

(1) If there is only one subscription level seen in the estimation period, it means that the RALM receiver has converged on a stable subscription level. The optimal subscription level is thus one level lower than this value to account for the sacrificial layer subscribed in stable state.
(2) If there are two subscription levels seen in the estimation period, the optimal subscription level is estimated to be the lower level. The reason for selecting the lower level is the fluctuating subscription levels have shown that the bottleneck can carry the lower level but not the higher level. Clearly, the layers comprising the lower subscription level should be protected by being marked with low drop precedence.
(3) If more than two subscription levels are seen in the estimation period, count the occurrences of every subscription level seen. If the two most common subscription levels make up more than two third of the subscription levels seen, it means that these two levels are close to the optimal subscription level and the selection method for two observed subscription levels can be used. Otherwise, the most common subscription level seen is likely to be close to the optimal subscription level and is selected as the estimated optimal subscription level for this estimation period.

Once the optimal subscription level is estimated, the receiver formats a feedback packet that contains this estimate and sends it to the marker whose address is recorded in the receiver. The statistics of achieved subscription levels is reset to zero before a new estimation period begins.

3.3. Feedback-driven layer marking mechanism

Although there is only one entry for an RALM session in the session information table of a marker, the session may have multiple layer marking levels in the marker. This is because a marker with multiple output links may be a branching node of the session's multicast tree. In other words, it may mark for different bottlenecks estimated on different branches of the multicast tree forking from it. The optimal subscription levels and consequently the marking levels for different output links may be different. Henceforth, the layer marking operations described in the marker are performed on every interface that transports a branch of the multicast tree as shown in Fig. 6.

A timer of 300 s is set for an interface so that the marker can adjust its marking levels of traversing sessions and send aggregated feedbacks upstream based on the optimal subscription level estimates received during the timer period. The reception of every feedback in an RALM session at the interface triggers the marker to determine the highest and lowest subscription level estimates seen in the session since the last timeout of the interface timer. In other words, the subscription level estimates of all traversing sessions are reset after the expiration of the interface timer so that the marker can determine the estimates anew for the next interface timer period. The lowest subscription level estimate may potentially be the current marking level of the session's layers for the

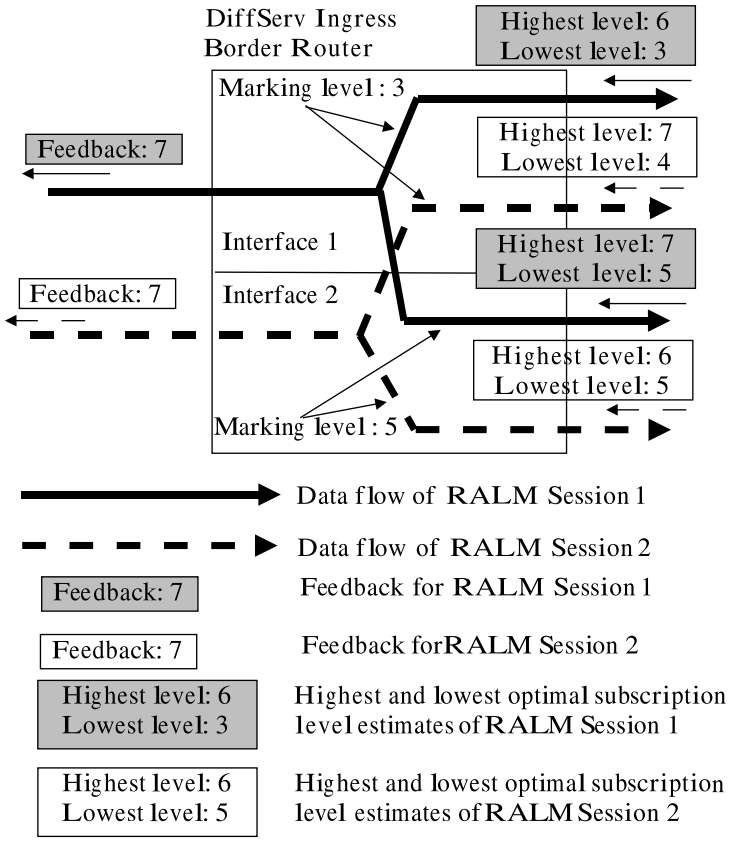

Fig. 6. Layer marking and feedback aggregation in a packet marker.

interface. And the highest subscription level estimate is the aggregated estimate that may be sent upstream as feedbacks for the downstream receivers.

When a new receiver joins an RALM session through an interface of a marker, no feedback has been received by the marker and hence no estimate on the optimal subscription level for this branch of the multicast tree is provided. The marker sets the marking level of the session to the base layer so that it can be protected from any packet losses induced during the process of discovering the optimal subscription level. The marking level of the session will be subsequently adjusted based on the estimates received at the interface for all RALM sessions that have receivers connected through this interface.

The 300 s interface timer permits the marker to receive at least 3 feedbacks from downstream receivers before adjusting the marking level of a session. This minimum number of feedbacks is necessary so that the marker can determine the highest and lowest subscription level estimates even if there is only one receiver on this branch of the multicast tree. At the expiration of the timer, the marker searches for the lowest subscription level estimate among all sessions that have receivers connected through this interface. The marking levels of the sessions are then set to this estimate. As an example, interface 1 in Fig. 6 has two lowest subscription level estimates for sessions 1 and 2. Since session 1's estimate of 3 layers is lower than that of session 2's estimate of 4 layers, both sessions' layers are marked to 3 layers.

The expiration of the interface timer also triggers the marker to send aggregated feedbacks to the respective upstream marker of every active session. The aggregated feedback of a session at the marker is the highest subscription level estimated by the session's downstream receivers regardless of the interfaces they are connected through. In Fig. 6, two highest subscription level estimates are found on the two interfaces for session 1. Since interface 2's estimate of 7 layers is higher than interface 1's estimate of 6 layers, the aggregated feedback of session 1 is 7 layers.

By locking the marking levels of RALM sessions sharing an output interface to a single value, inter-session subscription coordination to achieve max–min fair share of the downstream bottleneck bandwidth can be facilitated. This is because the marking levels of competing sessions are lowered to the base layers when a new session is transported by the interface. Then, the sessions can cooperate to discover the new optimal subscription levels that share the output link fairly. The scalability of the LMDP protocol is good because the probability of feedback implosion is minimised by feedback aggregation up the marker tree.

4. Intra-session interactions

A single RALM session is simulated on topology T1 in the scenario described below:

Scenario S1: The source transmits to four receivers that are connected through different bottleneck links as shown in Fig. 7. Receiver R1 has no bottleneck since the narrowest link it sees is 512 Kbps which is higher than the 400 Kbps maximum rate of the RALM session. Receivers R2 and R3 share a bottleneck link of 256 Kbps at router N3. Likewise, the path to receiver R4 has a bottleneck link of 128 Kbps at router N3. Note that the two bottleneck links are connected to two different output interfaces of router N3. The receivers are started in the following sequence: (1) R1 at time 50 s; (2) R2 at time 600 s; (3) R4 at time 1500 s; and (4) R3 at time 2000 s.

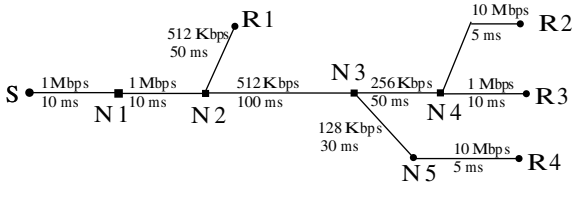

Fig. 7. Simulation topology T1 used to study the LDMCC intra-session interactions.

To achieve stable state in the RALM session where the interface timers are operating at 300 s intervals, the simulation is run for 5000 s. The objective of scenario S1 is to show that the LDMCC architecture is designed to mark the layers to the lowest optimal subscription level in an RALM session.

4.1. Multibottlenecks problem

A major problem with the LDMCC architecture is a session cannot find an optimal subscription level for all receivers when there are multiple bottlenecks within a DiffServ network as is the case in scenario S1. To satisfy the diverse bandwidth requirements of the receivers, the LDMCC architecture is designed to mark the layers to reflect the lowest optimal subscription level estimated in the session. This is illustrated in Fig. 8(a) and (b). Receiver R2 converges on the optimal subscription level of 6 layers (close to its bottleneck bandwidth of 256 Kbps) in the first 1500 s of the simulation because it is the lowest optimal subscription level estimated in the session. Once receiver R4 starts, the marker in router N1 discovers that the lowest optimal subscription level estimated in the session has been lowered to 3 layers (close to R4's bottleneck bandwidth of 128 Kbps) due to the narrower bottleneck on the path to R4. Thus, the session's layers are marked to the new optimal subscription level. R4 converges on the new optimal subscription level at the expense of R2 whose subscription level fluctuates between layers 6 and 7 because the new level underestimates its bottleneck bandwidth. This situation is not changed when receiver R3 is added to the wider bottleneck.

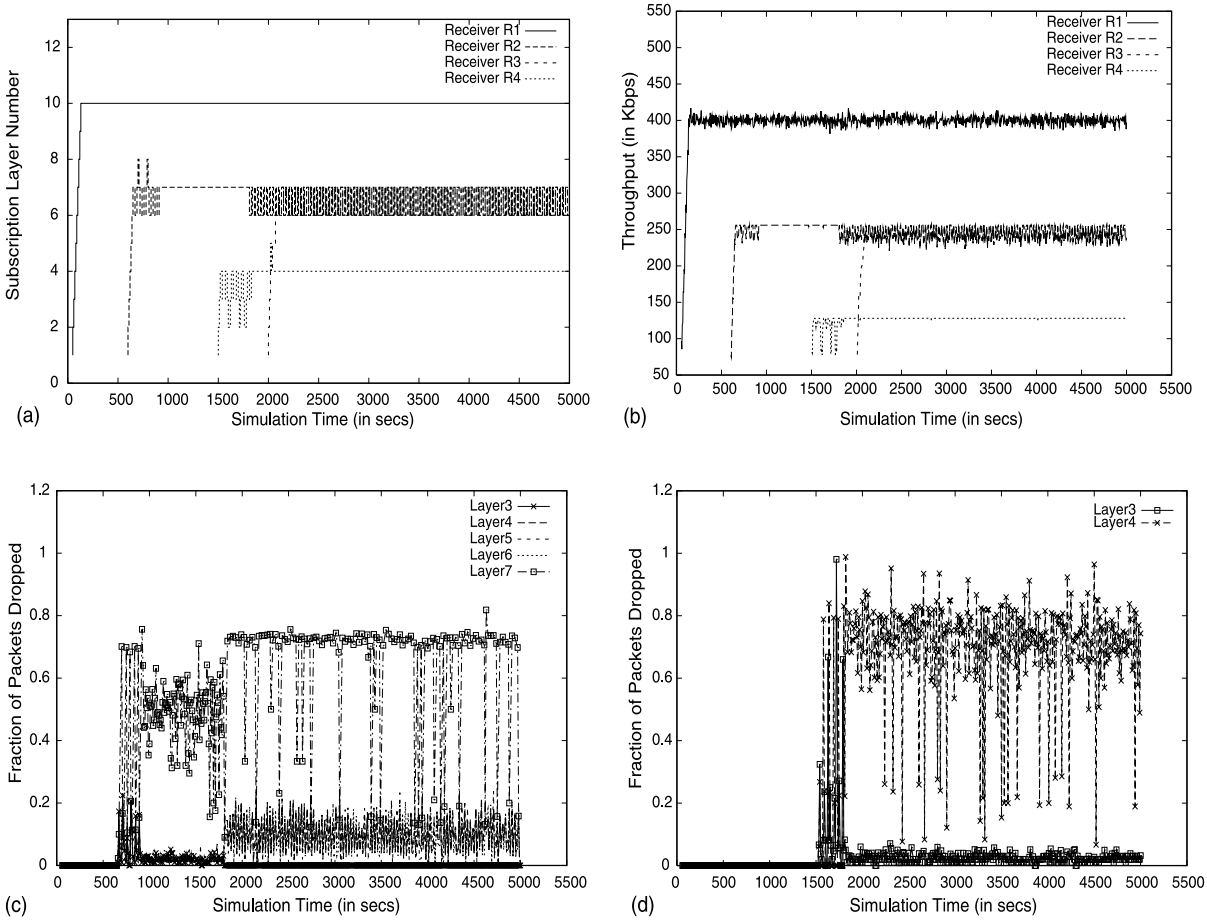

Fig. 8. The LDMCC architecture discovers the lowest optimal subscription level in a multibottlenecks RALM session: (a) layer subscriptions, (b) throughput curves, (c) loss rates experienced by receiver R2, (d) loss rates experienced by receiver R4.

Fig. 8(c) and (d) illustrate an important advantage of marking layers to the lowest optimal subscription level discovered in an RALM session. By marking the base layers commonly subscribed by all receivers in an RALM session with low drop precedence, the loss rates experienced in these layers by receivers behind wider bottlenecks will be low despite the underestimation of their bottleneck bandwidths by the layer markings. Most of the packet losses induced by the oversubscription of these receivers are absorbed by their sacrificial layers. Thus, the receivers are able to fluctuate close to their optimal subscription level of 6 layers.

The multibottlenecks problem in scenario S1 can be solved by placing an additional marker at router N3. The two bottlenecks are connected to N3 through different output interfaces. Thus, the marker can track feedbacks for the two bottlenecks separately and discover their optimal subscription levels. Then, the marker marks the session's layers to the different optimal subscription levels at their corresponding output interfaces. In this way, the layer markings along the two multicast tree branches forked at N3 reflect the number of layers that can be carried by their bottlenecks. As a result, the receivers downstream of the bottlenecks can converge on stable

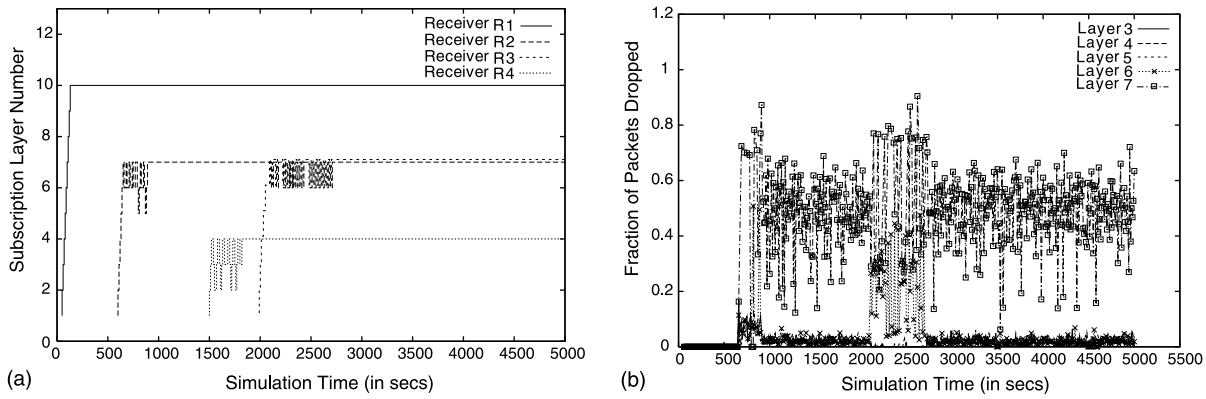

Fig. 9. The LDMCC architecture discovers all optimal subscription levels in a multibottlenecks RALM session if the marker tracks feedbacks for the bottlenecks separately: (a) layer subscriptions, (b) loss rates experienced by receiver R2.

subscription levels as shown in Fig. 9(a). Furthermore, the loss rates in the protected layers 4, 5 and 6 of receiver R2 are lowered in stable state as shown in Fig. 9(b).

5. Inter-session interactions

Inter-session interactions are investigated by simulating three RALM sessions on topologies T2 and T3 in two scenarios S2 and S3. The simulation scenarios are described below:

Scenario S2: 3 RALM sessions and a background best-effort CBR source-sink are simulated over the network shown in Fig. 10(a). The four sessions share a single bottleneck link of 700 Kbps between routers N1 and N2. In session 1, source S1 transmits to a single receiver R1. Source S2 transmits to 3 receivers {R2,R3,R4} in session 2. Source S3 transmits to 2 receivers {R5,R6} in session 3. The receivers of different RALM sessions are started at different times: (1) R1 starts at time 50 s; (2) R2 and R3 start at time 600 s; (3) R4 and R5 start at time 2000 s; (4) The background CBR source generates 400 Kbps of 500-bytes packets to congest the bottleneck link from 5000 to 6000 s; and (5) R6 starts at time 7000 s.

Scenario S3: 3 RALM sources are connected to their receivers through the network shown in Fig. 10(b). Two bottlenecks can be identified in the network. The wider bottleneck is the 700 Kbps link between routers N2 and N3 which is seen by receivers R1, R2 and R3. The narrower bottleneck is the 150 Kbps link between routers N2 and N5 seen only by receiver R4. Notice that the bottleneck links are connected to two different output interfaces of router N2. Receivers R1 and R4 subscribe to source S1 in session 1 while in sessions

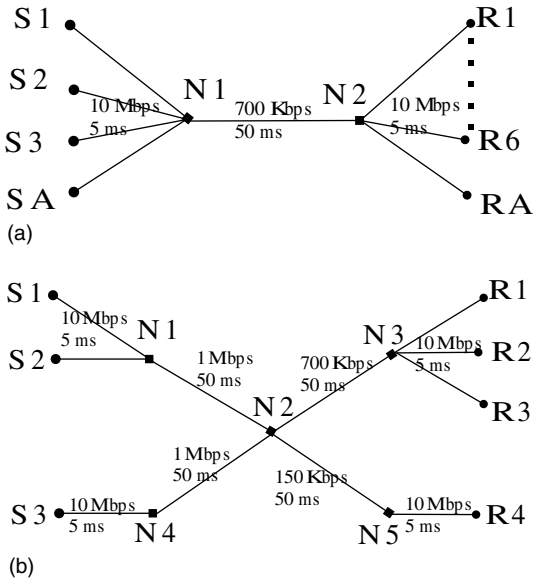

Fig. 10. Network topologies used to study the LDMCC inter-session interactions: (a) simulation topology T2, (b) simulation topology T3.

2 and 3, receivers R2 and R3 subscribe to sources S2 and S3 respectively. Thus, session 1 has two bottlenecks but sessions 2 and 3 have only one each. The simulations are conducted as follows: 1. R1 and R4 are started at time 50 s; 2. R2 starts at time 2000 s; and 3. R3 starts at time 5000 s.

To achieve stable state in the RALM sessions, the simulations are run for 10 000 s. The objective of scenario S2 is to show that the LDMCC architecture is designed to discover optimal subscription levels that provide multirate max–min fairness in the network even in the presence of unresponsive traffic. And scenario S3 shows that the LDMCC architecture loses its ability to provide multirate max–min fairness if competing sessions do not share a common marker.

5.1. Multirate max–min fairness

The network topology simulated in scenario S2 is simple, consisting only a bottleneck link and many access links. Thus, multirate max–min fairness in the network is achieved when the RALM sessions share the bottleneck bandwidth equally. A max–min fair share of the 700 Kbps bottleneck link among 3 RALM sessions is 233 Kbps or equivalently 5 layers per session. When the background CBR traffic is active, the bottleneck bandwidth is divided into two equal portions of 350 Kbps due to the round robin scheduler serving the AF and best-effort queues in router N1. In this case, multirate max–min fairness is applied solely to half of the bottleneck bandwidth which is allocated to the RALM traffic. Its max–min fair bandwidth share is 116 Kbps or equivalently 2 layers per session.

Fig. 11 shows the RALM sessions converging on the max–min fair optimal subscription level in the first 5000 s of the simulation regardless of the number of sessions in the network. When session 1 is started, its receiver R1 reaches the maximum subscription level of 10 layers rapidly since there is no bottleneck in the network. When session 2 is started at 600 s, the bottleneck link capacity is divided between the two sessions. Each session grabs 350 Kbps or 8 layers which is the max–min fair bandwidth share in this situation. When session 3 is started at 2000 s, the receivers converge on the max–min fair optimal subscription level of 5 layers after a marking level adjustment period of a few interface timer timeouts or approximately 1000 s. During the active period of the background CBR traffic, the RALM sessions swiftly drop layers to converge on the max–min fair optimal subscription level of 2 layers. After the end of the background CBR traffic, the RALM sessions recover their original max–min fair optimal subscription level of 5 layers over a long period of time.

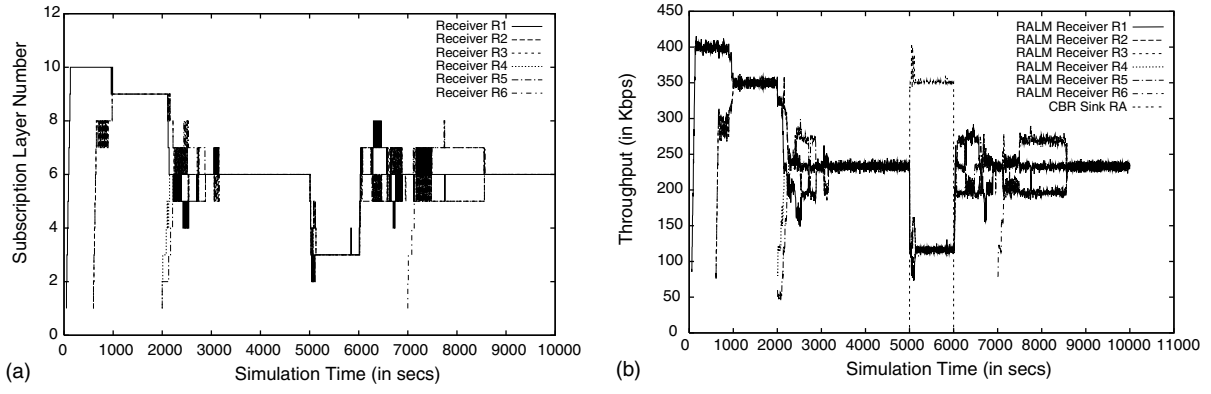

Fig. 11. The LMDP protocol ensures multirate max–min fair bandwidth share among multiple RALM sessions: (a) layer subscriptions, (b) throughput curves.

The convergence times of the RALM receivers are long if they are started at different times during the life of the session. This is because a new receiver starts its subscription from the base layer without being aware of the optimal subscription level already discovered by existing receivers in the session. Thus, the marker will reset the marking levels of all traversing sessions to the subscription level estimated by the new receiver. It takes time for the sessions to rediscover their original max–min fair optimal subscription levels. But once these levels are rediscovered, the receivers in all sessions can maintain a stable max–min fair share of the bottleneck bandwidth.

5.2. Shared-path-orthogonal-markers problem

The prerequisite for the achievement of multirate max–min fairness among RALM sessions traversing a network is the sessions sharing a common path must be marked by a marker along this path. This prerequisite is important because the inter-session subscription coordination mechanism in the LMDP protocol is local in significance. In other words, only a marker on a path shared by multiple RALM sessions receives feedbacks of optimal subscription level estimates from their receivers before it marks the sessions' layers. Without the presence of this marker, the sessions cannot coordinate their subscriptions to discover optimal subscription levels that provide max–min fairness along the common path, and by extension multirate max–min fairness cannot be achieved in the network. Therefore, the shared-path-orthogonal-markers problem is defined as the problem of not achieving multirate max–min fairness due to path sharing RALM sessions being marked by markers not on the common path.

The common marker requirement is easily fulfilled in scenario S2 because there is only one bottleneck link in topology T2 and the marker is at the head of this link. Thus, all sessions are marked by the same marker during the simulation. In contrast, the RALM sessions in scenario S3 enter the network through different access routers. Sessions 1 and 2 are marked by router N1 and session 3 is marked by router N4. Sessions 1, 2 and 3 share the bottleneck link between routers N2 and N3. Max–min fairness along this bottleneck link can be achieved only by sessions 1 and 2 when session 3 is inactive as illustrated in Fig. 12(a) from time 2000 to 5000 s. If session 3 is active, its layers are marked based solely on feedbacks from its receiver. Without explicit coordination with the other two sessions, the marking level for session 3 increases as long as receiver R3 provides higher subscription level estimates. In the long term, session 3 can grab more bandwidth than sessions 1 and 2. This is illustrated in Fig. 12(a) where receiver R3's optimal subscription level reaches layer 7 while the optimal subscription levels of receivers R1 and R2 are lowered to layer 4 at the end of the

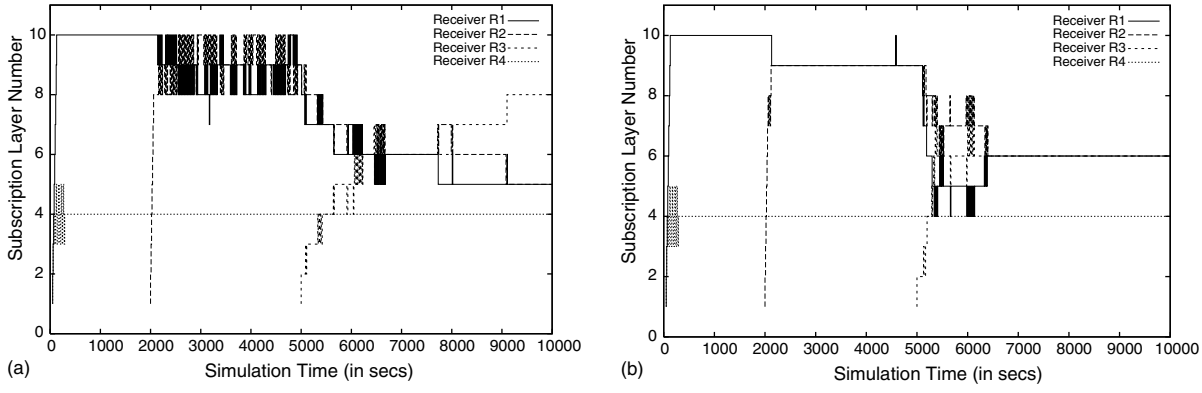

Fig. 12. Multirate max–min fairness cannot be assured unless markers are placed at all bottleneck links in a network: (a) layer subscriptions in a network that exhibits the shared-path-orthogonal-markers problem, (b) layer subscriptions in a network where competing sessions are marked by common markers.

simulation. Clearly, multirate max–min fairness does not exist in this network.

The shared-path-orthogonal-markers problem can be solved by placing an additional marker at router N2 from which the two bottleneck links originate. Since the marker at N2 is on the shared path of sessions 1, 2 and 3, it can coordinate the subscriptions of their receivers by locking the marking levels of the sessions to a common value. Thus, the sessions can discover the optimal subscription level of 5 layers that provides multirate max–min fairness in the network as shown in Fig. 12(b).

6. Conclusion

In this paper, the LDMCC architecture has been described and evaluated through a series of simulations. The simulation results show that the LDMP protocol assists the receivers of RALM sessions to discover and maintain stable optimal subscription levels in a DiffServ network. However, a single optimal subscription level cannot be found in an RALM session if the marker receives feedbacks of different optimal subscription level estimates from its downstream receivers when they probe different bottleneck bandwidths. In this case, the marker marks the layers to the lowest optimal subscription level estimated in the session. Although the receivers behind wider bottlenecks cannot maintain a stable subscription level, the base layers in their subscriptions are protected from packet losses induced by their periodic oversubscription of the bottleneck bandwidths.

Apart from the multibottlenecks problem, two other major problems are identified in the LDMCC architecture. First, the receivers take at least 300 s to converge on their optimal subscription levels due to the fixed marking level adjustment periods in the markers. The already long convergence time is further lengthened if new receivers subscribe to the sessions at different times and upset the stable state that has been achieved by the RALM sessions. Second, multirate max–min fairness in a network is achievable only if all RALM sessions sharing common paths are marked by markers on these paths. Without the subscription coordination mechanism provided by these markers, the path sharing sessions cannot compete in a max–min fair manner.

A simple solution to the problems of shared-path-orthogonal-markers and multibottlenecks is to place markers on all bottleneck links in the network. By tracking receivers' feedbacks on a per-interface basis, a marker can adjust the marking levels of the traversing sessions to guide the receivers to the optimal subscription levels of individual bottlenecks. The net result is all receivers reach stable subscription levels that provide multirate max–min fairness in the network. However, the placement of markers on all bottleneck links potentially degenerates the network into a mesh of single hop domains. This is contrary to the DiffServ goal of pushing processing complexity to the network edge. But by leveraging the segregation of RALM traffic from other traffic types and through traffic engineering, the number of markers that must be deployed can be limited to an acceptable size by creating known bottleneck points in the multicast network topology.

References

[1] D. Bertsekas, R. Gallager, Data Networks, second ed., Prentice Hall, Engelwood Cliffs, NJ, 1992.
[2] S. Bajaj, L. Breslau, S. Shenker, Uniform versus priority dropping for layered video, in: Proceedings of ACM SIGCOMM'98, Vancouver, BC, Canada, 1998, pp. 131–143.
[3] S. Floyd, V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ACM Trans. Network. 1 (4) (1993) 397–413.
[4] D. Lin, R. Morris, Dynamics of random early detection, in: Proceedings of ACM SIGCOMM'97, Cannes, France, September 1997, pp. 127–137.
[5] S. McCanne, V. Jacobson, M. Vetterli, Receiver-driven layered multicast, in: Proceedings of ACM SIGCOMM'96, Stanford, CA, 1996, pp. 117–130.
[6] D. Rubenstein, J.F. Kurose, D.F. Towsley, The impact of multicast layering on network fairness, in: Proceedings of ACM SIGCOM'99, Cambridge, MA, September 1999, pp. 27–38.
[7] R. Gopalakrishnan, J. Griffioen, G. Hjsson, C.J. Sreenan, S. Wen, A simple loss differentiation approach to layered multicast, in: Proceedings of IEEE INFOCOM 2000, Tel Aviv, Israel, 2000, pp. 461–469.
[8] A. Legout, E. Biersack, Pathological behaviors for RLM and RLC, in: Proceedings of IEEE NOSSDAV 2000, Chapel Hill, NC, 2000, pp. 164–172.

[9] L. Vicisano, L. Rizzo, J. Crowcroft, TCP-like congestion control for layered multicast data transfer, in: Proceedings of IEEE INFOCOM'98, San Francisco, CA, 1998, pp. 996–1003.
[10] W. Almesberger, J.H. Salim, A. Kuznetsov, Differentiated Services on Linux, Work in Progress, IETF, June 1999.
[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An Architecture for Differentiated Services, RFC 2475, IETF, December 1998.
[12] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Resource ReSerVation Protocol (RSVP) Version 1 Functional Specification, RFC 2205, IETF, September 1997.
[13] W. Fenner, Internet Group Management Protocol (IGMP), Version 2, RFC 2236, IETF, November 1997.
[14] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, Assured Forwarding PHB Group, RFC 2597, IETF, June 1999.
[15] V. Jacobson, K. Nichols, K. Poduri, An Expedited Forwarding PHB, RFC 2598, IETF, June 1999.
[16] T. Speakman, J. Crowcroft, D. Farinacci, L. Rizzo, L. Vicisano, et al., PGM Reliable Transport Protocol Specification, Work in Progress, IETF, February 2001.
[17] S. Floyd, various notes on RED. Available from <http://www.aciri.org/floyd/red.html>.
[18] S. McCanne, S. Floyd, The *ns* Network Simulator. Available from <http://www.isi.edu/nsnam/ns/>.
[19] B. Vickers, C. Albuquerque, T. Suda, Source-adaptive multi-layered multicast algorithms for real-time video distribution, Technical Report ICS-TR 99-45, University of California, Irvine, 1999.

Yung-Sze Gan graduates from the National University of Singapore with a Bachelor and a Master of Engineering degree in 1998 and 2002 respectively. Currently, he is a research and development engineer in Siemens Pte Ltd Mobile Core Research and Development Department, Singapore. His current work is largely concentrated in the UMTS IP Multimedia Subsystem with some interests in the areas of policy-based network management and multicast networks.

Chen-Khong Tham is an Assistant Professor at the Department of Electrical and Computer Engineering (ECE) of the National University of Singapore (NUS). His research interests are in the areas of quality of service (QoS) in computer networks and application servers, and predictive and proactive network management. Dr. Tham is in-charge of the Computer Communication Networks (CCN) Laboratory at the ECE Department, NUS, and was the Manager for Advanced Applications at the Singapore Advanced Research & Education Network (SingAREN) from 1997 to 1999. Dr. Tham obtained his Ph.D. and M.A. degrees in Electrical and Information Sciences Engineering from the University of Cambridge, United Kingdom.