****

ATM Forum Document Number: ATM Forum/97-0422

****

Title: Performance analysis of ABR point-to-multipoint connections for bursty and non-bursty traffic with and without VBR background

****

Abstract:
A number of frameworks have been developed for extending ABR traffic management algorithms to support point-to-multipoint connections. This contribution provides a comprehensive analysis of the performance of ABR traffic management for point-to-multipoint connections under a variety of conditions. The study demonstrates that the extension frameworks preserve the efficiency and fairness properties of the original point-to-point switch scheme employed. The major problem specific to point-to-multipoint connections is the consolidation noise problem, which occurs when there are distant bottlenecks on a branch of the tree. In addition, many links on the branches of a point-to-multipoint connection may be underutilized because bottlenecks exist on other branches of the tree. Resolving the consolidation noise versus slow transient response problem, and correctly setting the ABR source parameters are essential for correct operation of ABR point-to-multipoint.

****

Source:
Sonia Fahmy, Raj Jain, Shivkumar Kalyanaraman, Rohit Goyal, Bobby Vandalore, and Xiangrong Cai
The Ohio State University (and NASA)
Department of Computer and Information Science

Raj Jain is now at Washington University in Saint Louis, jain@cse.wustl.edu http://www.cse.wustl.edu/~jain/

Seong-Cheol Kim
Principal Engineer, Network Research Group
Communication Systems R&D Center
Samsung Electronics Co. Ltd., Chung-Ang Newspaper Bldg.
8-2, Karak-Dong, Songpa-Ku, Seoul, Korea 138-160
Email: kimsc@metro.telecom.samsung.co.kr

****

Date: April 1997

****

Distribution: ATM Forum Technical Working Group Members (AF-TM)

****

Notice:
This contribution has been prepared to assist the ATM Forum. It is offered to the Forum as a basis for discussion and is not a binding proposal on the part of any of the contributing organizations. The statements are subject to change in form and content after further study. Specifically, the contributors reserve the right to add to, amend or modify the statements contained herein.

****

1 Introduction

ATM multicast capabilities are essential for the support of numerous applications, such as LAN emulation and IP multicasting. UNI 3.1 and 4.0 define point-to-multipoint connections, and a number of frameworks were developed for extending point-to-point traffic management to point-to-multipoint scenarios.

In August 1995, Hunt [4] pointed out that an existence proof is necessary for a multicast traffic management mechanism suitable for the expected range of traffic patterns, number of VCs, bandwidth bottlenecks, and round trip times (RTTs). Bursty traffic sources, as well as a wide potential range of RTTs from the source to the various leaves (RTT difference should be a couple of orders of magnitude) should be examined, but a worst case analysis needs to be provided. Cases to be tested include (1) bursty traffic models (2) dynamic CBR and VBR in the background on bottleneck links (3) many ABR VCs, especially point-to-multipoint VCs (4) several orders of magnitude variation in the the available bandwidth at the bottlenecks (5) changes in dynamic capacity at transient periods, as well as in steady state, and (6) a large RTT ratio from the nearest to the farthest leaf [4].

This contribution aims at providing a most comprehensive test of the performance of point-to-multipoint mechanisms, including the cases proposed in [4]. Because we use simulations, the number of leaves of a multicast connection is limited to a small number. For a large number of leaves, analytical techniques can be employed. Therefore, this contribution is limited to configurations with a small number of leaves.

The rest of the contribution is organized as follows. The next section briefly explains ABR flow control, then a survey of traffic management of point-to-multipoint connections is provided. This is followed by a brief description of the point-to-multipoint ERICA switch algorithm used in the simulations. The main portion of the contribution is devoted to a detailed discussion of the simulation results and their significance, and a comparison to point-to-point cases. Finally, the main results are summarized.

2 ABR Flow Control

The ABR service periodically advises sources about the rate at which they should be transmitting. The switches compute the available bandwidth and divide it fairly among active VCs. The feedback from the switches to the sources is indicated in Resource Management (RM) cells. The RM cells are periodically sent by the sources and turned around by the destinations.

The RM cells contain the current cell rate (CCR) of the source, in addition to several fields that can be used by the switches to provide feedback to the sources. One of those fields, the Explicit Rate (ER) field, indicates the rate that the network can support at that time. Each switch on the path of the VC reduces the ER field to the maximum rate it can support. The

sources examine the returning RM cells and adjust their transmission rates.

The RM cells flowing from the source to the destination are called Forward RM cells (FRMs) while those returning from the destination to the source are called backward RM cells (BRMs). When a source receives a BRM, it computes its allowed cell rate (ACR) using its current ACR, two of the flags in the RM cell, and the ER field of the RM cell.

3 Survey of Previous Work on Point-to-Multipoint ABR Flow Control

The ATM Forum traffic management specification currently provides some guidelines on traffic management of point-to-multipoint connections, but does not enforce nor suggest a specific strategy [3]. Congestion control strategies for multipoint-to-point, and multipoint-to-multipoint connections are still under study. This section surveys the work that has been done on point-to-multipoint traffic management.

The traffic management problem for point-to-multipoint connections is an extension to traffic management for unicast connections. However, some additional problems arise in the point-to-multipoint case. In particular, the consolidation of feedback information from different leaves of the tree is necessary for point-to-multipoint connections. This is because of the "feedback implosion" problem (feedback information provided to the sender should not increase proportional to the number of leaves in the connection). Scalability becomes a major concern.

Many general frameworks have been suggested that convert any unicast congestion control switch algorithm to work for point to multipoint connections [9, 12, 8]. In addition, several issues pertaining to source end system parameters for point-to-multipoint connections have been discussed [11, 4]. This section highlights the major issues in this area.

3.1 Source Parameters

ABR source parameters were briefly examined in [11], [2] and [4]. The main factor that complicates the setting of these parameters for point-to-multipoint connections is the possible existence of largely varying round trip times from the source to the different receivers, and the possible existence of a bottleneck on a distant branch. Thus a more conservative approach in the setting of the parameters is preferred.

Point-to-multipoint connections may suffer from initial overallocation until feedback is received from all the distant leaves. Initial overallocation can be overcome by correct setting of the Cells in Flight (CIF) parameter and the correct calculation of the Initial Cell Rate (ICR) parameter. In [11], a formula is proposed to calculate the optimal value of ICR, and an approximation is proved to achieve an approximately equivalent performance. The value

of ICR is a function of the CIF value, the longest RTT value, and the Rate Increase Factor (RIF).

Note that if the calculation of ICR depends on the round-trip time (RTT), a problem arises: should ICR change when nodes join or leave the group to account for the longest RTT for all destinations? The farthest leaf's RTT reduces ICR according to the proposed formula. What happens when that farthest leaf leaves the multicast group? [4]. The root should not need to be notified every time a leaf joins or leaves the group.

In addition, low Rate Increase Factor (RIF) values are used for point-to-multipoint connections to avoid transient queues in cases of distant bottlenecks that are multiple branch points away [11].

3.2 Consolidation of Feedback Information

One of the common goals for point-to-multipoint ABR is to ensure that all destinations receive all cells from the source. This requires that the source be controlled to the minimum rate supported by all the destinations. This is especially useful for LANE and non-real-time data services [10].

The source in a point-to-multipoint VC sends at the minimum of the rates allowed by all the destination nodes. The minimum rate is the technique most compatible with the typical data requirements where no data should be lost and the network can take whatever time it requires to deliver the data intact. A register MER is set to min (MER, ER in BRM cell) whenever a BRM cell is received from one of the branches. When an FRM cell is received, it is multicast to all branches, and a BRM is sent with the MER value as the ER. MER is then set to the ER value in the FRM cell (typically PCR). Thus the minimum of all the rates supported by any branch is selected and returned to the source [9].

Observe that at each switch node with N branches, an additional cycle of N cells is required in order to accumulate the information from the branches. Thus if a multicast tree has 5 levels of branching, then the information from the lowest branches will take 5N cell times to propagate back to the source (as opposed to N cell times in the point-to-point case). As a result of this additional delay, the *responsiveness of the multicast algorithm will be worse than that for the point-to-point VCs*. Thus, buffer allocations for the multicast queues will have to be somewhat higher since it takes longer for congestion information from one branch to reach the source. If the multicast traffic is small, the extra logic and buffer memory required is quite small and multicast ABR operation can be achieved very easily [9, 2].

At the ATM forum, the consolidation algorithm was proposed to be implementation-specific with the previously explained algorithm given as an example consolidation. The source and destination behaviors are unaltered, and the consolidation algorithm at the switches is optional. Sources may also need do the consolidation in that case [1].

However, an additional problem may arise. Since each BRM cell is generated from a branch

point-to-the root when an FRM arrives, and the BRM contains the consolidated information from the branches that provided feedback after the last BRM was sent, the BRM, in general, does not capture feedback information from all branches. This introduces noise called *"consolidation noise"* [4].

In [12], the multicast extension was applied to an ABR rate allocation algorithm, and the results prove that the extended algorithm is efficient and max-min fair if the point-to-point algorithm is max-min fair. Several variations on the previously described algorithm were proposed in [8]. They employ other approaches to consolidate the feedback information in the multicast tree. Some of the new schemes are simpler to implement than the previous proposal that required the branch point to generate a returning RM cell for every forward RM cell, while others attempt to achieve better performance.

The first modification proposed tries to alleviate the "consolidation noise" problem. The early proposal suffers from consolidation noise, where BRM generated by switch may not consolidate feedback from all tree branches. In fact, if a BRM generated by a switch does not get accumulate feedback from any branch, the feedback can erroneously be given as the peak cell rate. A simple enhancement to avoid this problem is to maintain a flag, and only turn around the RM cell if a BRM has been received from a branch since the last BRM was sent by the split point [8].

Another idea reduces the complexity of the algorithm as follows. The backward RM cells are generated solely by the destinations and NOT by the switches, which is similar to the case of unicast [7]. The motivation behind this modification is as follows. If switches turn around RM cells, the implementation has a high cost and complexity. In the previously mentioned algorithms, the number of BRMs generated by switches per forward RM cell from the source is proportional to the number of branch points in a multicast tree. The new algorithm does not generate BRMs at split points whenever FRMs are received, but simply sets a flag indicating the receipt of the FRM and broadcasts it to all leaves. When a BRM is received from a branch, it is passed back the source (after using the minimum allocation), only if the previously mentioned flag was set. The flag is then reset, as well as the MER value [8].

It is natural to extend this idea to only send back the BRM when BRMs from *all* branches are received. This can be easily implemented by maintaining a separate bit for each branch that indicates if a BRM has been received since the last BRM was sent. Clearly this method incurs additional complexity, compared to the previous one. Moreover, it has to deal with the problems of failure of one of the branches by implementing timeouts. The four variations of the algorithm were compared in [8]. While consolidation noise was less with the last method, the additional complexity might not be worth the slight benefits, especially that the method exhibits a slow transient response.

In summary, the different variations developed exhibit a tradeoff between complexity and minimization of consolidation noise.

4 The Point-to-Multipoint ERICA Switch Algorithm

The ERICA algorithm is used in this performance study. This section highlights the main idea of the algorithm, and its point-to-multipoint extensions.

The ERICA algorithm requires monitoring the available capacity and the current demand on the resources. The algorithm is applied to each output port (or link). In this section, we briefly describe the algorithm. For a complete description of the ERICA algorithm, including pseudocode and flowcharts, see [6].

4.1 The Original Algorithm

The switch periodically monitors the load on each link and determines a load factor, z, the available capacity, and the number of currently active VCs.

The load factor is calculated as the ratio of the measured ABR input rate at the port to the target ABR capacity of the output link.

$$\text{Load Factor}(z) \leftarrow \frac{\text{ABR Input Rate}}{\text{Target ABR Capacity}}$$

where:

$$\text{Target ABR Capacity} \leftarrow \text{Target Utilization} \times \text{Link Bandwidth}$$

The Input Rate is measured over an interval called the switch averaging interval. The above steps are executed at the end of the switch averaging interval.

Target utilization is a parameter which is set to a fraction (close to, but less than 100%) of the available capacity.

The fair share of each VC, $FairShare$, is also computed as follows:

$$\text{FairShare} \leftarrow \frac{\text{ABR Capacity}}{\text{Number of Active Sources}}$$

The switch allows each source sending at a rate below the $FairShare$ to rise to $FairShare$ every time it sends a feedback to the source. If the source does not use all of its $FairShare$, then the switch allocates the remaining capacity to the sources which can use it. For this purpose, the switch calculates the quantity $VCShare$ as follows:

$$\text{VCShare} \leftarrow \frac{CCR}{z}$$

Hence, $VCShare$ aims at bringing the system to an efficient operating point, which may not necessarily be fair, and $FairShare$ allocation aims at achieving fairness, possibly leading to overload (inefficient operation). A combination of these two quantities is used to rapidly reach optimal operation as follows:

ER Calculated←Max (FairShare, VCShare)

The calculated ER value cannot be greater than the ABR Capacity which has been measured earlier. Hence, we have:

ER Calculated←Min (ER Calculated, ABR Capacity)

To ensure that the bottleneck ER reaches the source, each switch computes the minimum of the ER it has calculated and the ER value in the RM cell.

4.2 Achieving Max-Min Fairness

To achieve max-min fairness, the basic ERICA algorithm is extended by remembering the highest allocation given during an averaging interval and ensuring that all sources are given this high allocation. We add a variable $MaxAllocPrevious$ which stores the maximum allocation given in the previous interval.

For $z > 1 + \delta$, where δ is a small fraction, we use the basic ERICA algorithm and allocate the source Max (FairShare, VCShare). But, for $z \leq 1 + \delta$, we attempt to make all the rate allocations equal, by giving the allocation Max (FairShare, VCShare, MaxAllocPrevious). This method equalizes the allocations of the VCs during underload.

4.3 Fairshare First to Avoid Transient Overloads

The inter-RM cell time is a factor in determining the transient response time when load conditions change. With the basic ERICA scheme, it is possible that a source which receives feedback first can keep getting rate increase indications, purely because it sends more RM cells before competing sources can receive feedback. This results in unnecessary spikes (sudden increases) in rates and queues with the basic ERICA scheme.

This problem can be alleviated by incorporating the following change to the ERICA algorithm. When the calculated ER is greater than the fair share value, and the source is increasing from a CCR below $FairShare$, we limit its increase to $FairShare$.

4.4 ABR Operation with VBR and CBR in the Background

ATM links will be used by constant bit rate (CBR) and variable bit rate (VBR) traffic along with ABR traffic. CBR and VBR have a higher priority. Only the capacity left unused by VBR and CBR is allocated to ABR sources. Hence, we need to measure the CBR and VBR usage along with the input rate. The ABR capacity is then calculated as follows:

$$\text{Target ABR Capacity} \leftarrow \text{Target Utilization} \times \text{Link Bandwidth} - \text{VBR Usage} - \text{CBR Usage}$$

4.5 Point-to-Multipoint ERICA

The framework proposed in [9] was used to extend ERICA for point-to-multipoint operation. The ERICA algorithm itself is employed immediately before sending a BRM. At a branch point, when an RM cell is to be turned around or passed, the explicit rate is calculated by ERICA for each branch, and the minimum of all these rates, and the explicit rate field stored at the switch, is indicated in the BRM cell to be sent. The algorithm at a branch point operates as follows.

Upon the receipt of an FRM cell:

1. Multicast this cell to all participating branches
2. Let MXR = ER from FRM cell
3. Let MER = min (MER, ER calculated by ERICA for all branches)
4. Return a BRM with ER = MER to the source
5. Let MER = MXR

Upon the receipt of a BRM cell:

1. Let MER = min (MER, ER from BRM cell)
2. Discard the BRM cell

The congestion indication (CI) and no increase (NI) fields can be handled similar to the ER field, but instead of the minimum operation, an OR operation is performed.

As previously discussed, a number of variations of this algorithm were developed in [8] to improve the performance and reduce complexity. The algorithm can be improved by maintaining a flag that denotes that a BRM cell has been received since the last BRM cell has been sent, and then sending a new BRM only if that flag is set. A further improvement involves maintaining a similar flag for each branch and only turning around the RM cell if

BRM cells have been received from all branches. In addition, the complexity of the branch point algorithm can be reduced by avoiding turning around the RM cells, and simply passing one of the BRMs returned by the destinations. These variations were explained in the survey section.

5 Performance Analysis Objectives and Metrics

Point-to-multipoint ERICA has been tested for a variety of networking configurations using several performance metrics. Its performance in the presence of various background traffic patterns and various source models has also been examined. We present simulation results for several configurations, which have been specifically selected to demonstrate particular aspects of the scheme. We prefer to use simple configurations when applicable because they are more instructive in finding problems [5]. Note that a very large number of configurations was simulated, but only the more interesting subset is presented here.

The results are presented in the form of four graphs for each configuration:

- (a) Graph of allowed cell rate (ACR) in Mbps over time for each source
- (b) Graph of ABR queue lengths in cells over time at each switch
- (c) Graph of link utilization (as a percentage) over time for each link
- (d) Graph of number of cells received at the destination over time for each destination

Finally, each figure also includes a schematic description of the configuration simulated.

We examine the efficiency, fairness, transient and steady state performance of the scheme, and its adaptation to variable capacity and various source traffic models. The experiments are also selected such that they have varying round trip times, feedback delays and number of connections. We also examine the consolidation noise.

6 Performance Results

This section discusses the simulation configurations, parameters and results.

6.1 Parameter Settings

Throughout our experiments, the following parameter values are used:

1. All links have a bandwidth of 155.52 Mbps.

2. Except where indicated, all LAN links are 1 km and all WAN links are 1000 kms.

3. All point-to-multipoint traffic flows from the root to the leaves of the tree. No traffic flows from the leaves to the root, except for RM cells. All other VCs are also unidirectional.

4. Except for the experiments that aim at examining the effect of the source parameter Rate Increase Factor (RIF), RIF is set to one, to allow immediate use of the full explicit rate indicated in the returning RM cells at the source. ICR is also set to very high values (almost PCR) in most experiments.

5. The source parameter Transient Buffer Exposure (TBE) is set to large values to prevent rate decreases due to the triggering of the source open-loop congestion control mechanism. This was done to isolate the rate reductions due to the switch congestion control from the rate reductions due to TBE.

6. The switch target utilization parameter was set at 95% for LAN simulations and at 90% for WAN simulations.

7. The switch averaging interval was set to the minimum of the time to receive 50 cells and 1 ms for LAN simulations, and to the minimum of the time to receive 100 cells and 1 ms for WAN simulations.

8. All sources are deterministic, i.e., their start/stop times and their transmission rates are known. The bursty traffic sources send data in bursts, where each burst starts after a request has been received from the client.

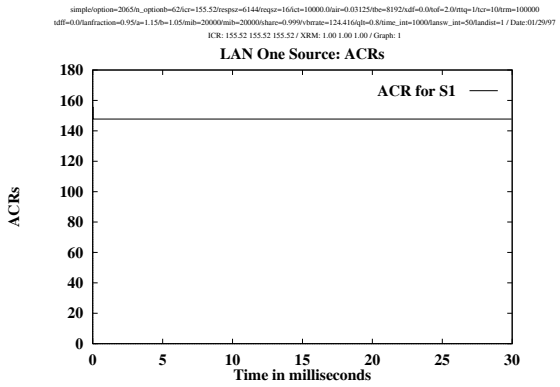
9. In WAN cases, VBR sources are on for 20 ms and off for 20 ms, and in LAN cases the on/off period is 3 ms. Experiments were also conducted with self-similar VBR background, but the performance did not illustrate any new results.

10. Simulation time is 30 ms from LAN simulations and 200 ms for WAN simulations. Some WAN simulations were run for 400 ms to allow the system enough time to reach steady state.

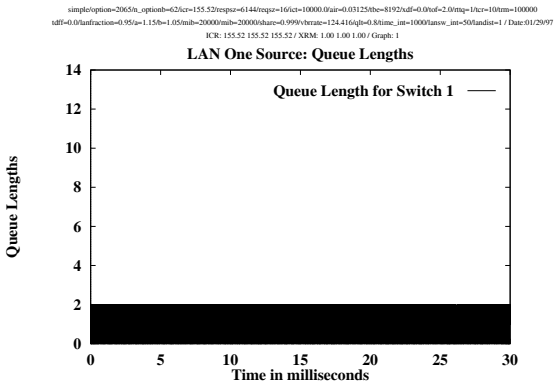
6.2 Simulations Results and Analysis

This section explains the results obtained. It proceeds from the simpler configurations to the more complex ones.

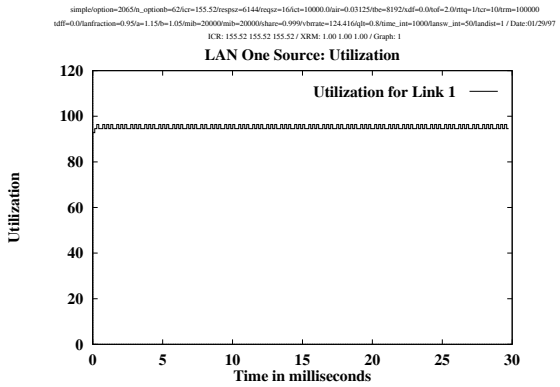
- Configuration 1 [One source and 2 destinations]: This is a LAN configuration (all links are 1 km). One source is sending to two destinations with same RTT (figure 1(e)). As seen in figure 1, the source ACR is close to PCR, the queues are small, the link is optimally utilized and the throughput is high.



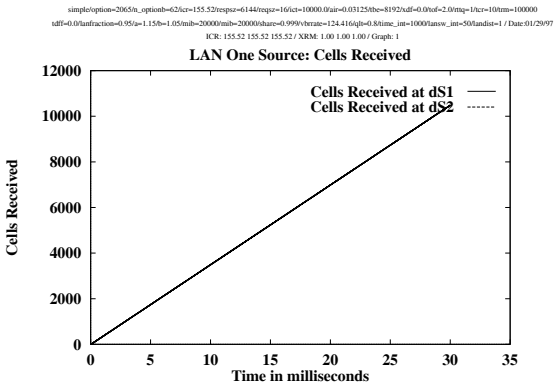
(a) Transmitted Cell Rate



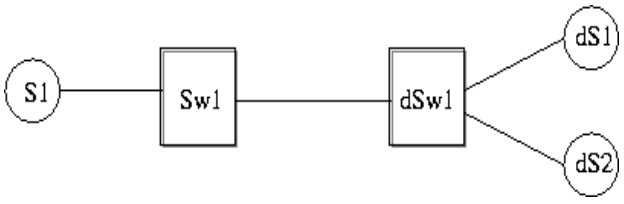
(b) Queue Length



(c) Link Utilization



(d) Cells Received



(e) One source and 2 destinations configuration

Figure 1: Results for a one source and 2 destination (same RTT) configuration in a LAN

- Configuration 2 [Two source]: This is also a LAN configuration. Two sources are each sending to two destinations with same RTT (figure 2(e)). As seen in figure 2, initially there are small queues because of the high ICR (ICR = 150 Mbps), but the queues quickly drain. Max-min fairness is seen from the rate graph, and utilization and throughput are high.

- Configuration 2b [Transient]: This is the same as configuration 2 (figure 3(e)), with two differences. First, it is a WAN configuration, and second, one of the sources is a transient source. The transient source comes on after about one third of the simulation time, stays on for about a third, and then goes away. Figure 3 shows the large transient queue till source 1 drops in steps when it detects the presence of source 2. There is a drop in utilization when source 2 goes away, until source 1 detects that. However, the transient response of the scheme is not too slow.

- Configuration 3 [Parking lot]: Here one source is sending to three destinations with different RTTs (figure 4(e)). The configuration is a LAN configuration. As seen in figure 4, the ACR is close to PCR, queues are small, and utilization and throughput are high.

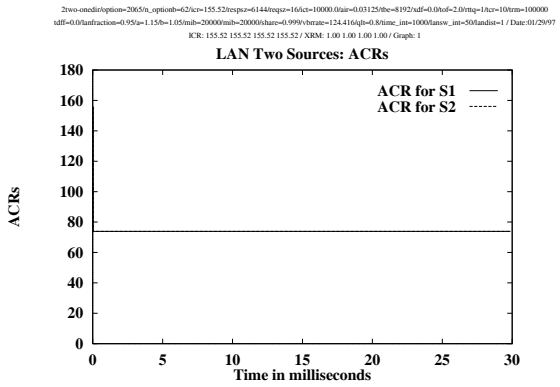
- Configuration 3b [Parking lot]: This configuration is identical to the previous one (figure 5(e), but it is a WAN. Figure 5 shows that the queues are extremely small, the utilization and throughput are high, and the ACR close to PCR after an initial delay.

- Configuration 4 [Parking lot and point-to-point]: This configuration is identical to configuration 3, except that there is a unicast VC sharing a branch of the multicast tree, as shown in figure 6(e). The VC does not share any portion of the path from the source to one of the destinations (dS3). As seen in figure 6, the bandwidth is equally divided among the two sources, and max-min fairness is preserved. The initial queues are caused by the high ICR setting.

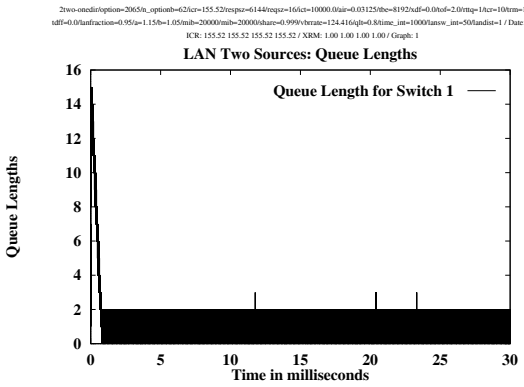
- Configuration 4b [Parking lot and transient]: This is the same as configuration 4 (figure 7(e)), but the unicast connection is transient. A transient queue builds up when the point-to-point connection starts (see figure 7), but the response of the scheme is fast. The response is also fast when the transient connection terminates. Utilization is high.

- Configuration 5 [Parking lot and point-to-point with long feedback delay]: This is an interesting configuration. The configuration is same as configuration 4 (figure 8(e)), but link 1 is much longer (5000 km) than the others (50 km). Figure 8 illustrates that ACR drops in steps to the correct value. There are some slight oscillations seen in the rate graph. Queues are bounded, and utilization of the bottleneck link is high. The step effect is due to the interaction of the *consolidation noise* with the ERICA algorithm.

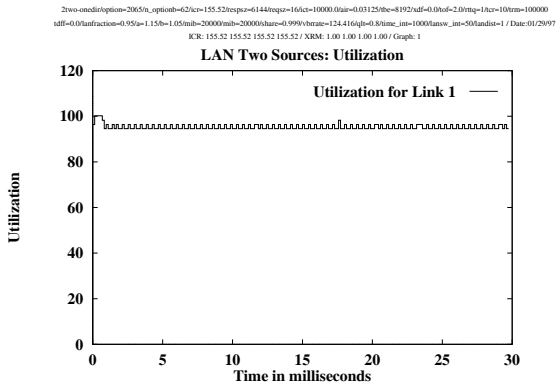
- Configuration 6 [Parking lot and point-to-point with highly varying RTTs]: This is also the same as configuration 4 (figure 9(e)), but here link 2 is much longer (5000 km) than the others (50 km). As seen in figure 9, ACR drops to the correct value. Queues



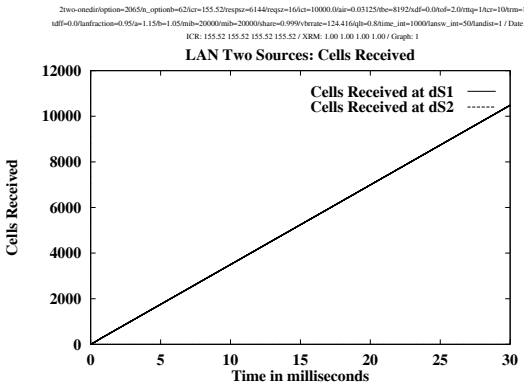
(a) Transmitted Cell Rate



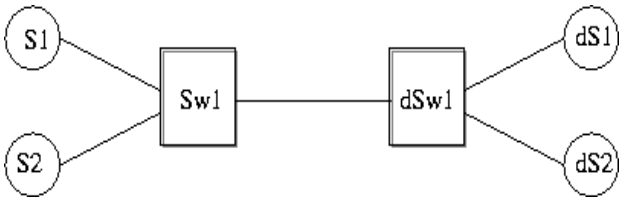
(b) Queue Length



(c) Link Utilization

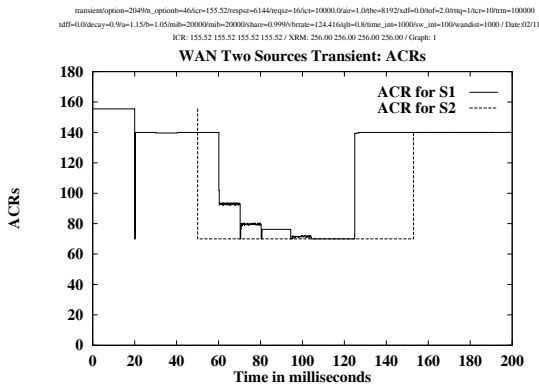


(d) Cells Received

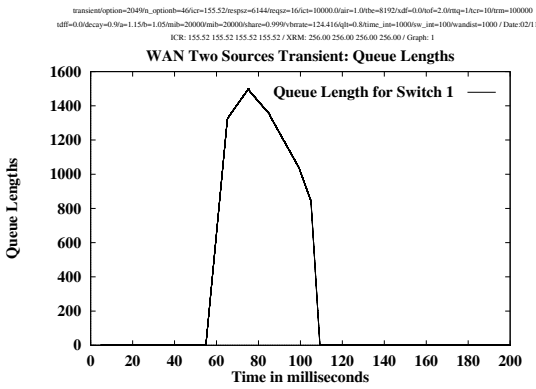


(e) Two source configuration

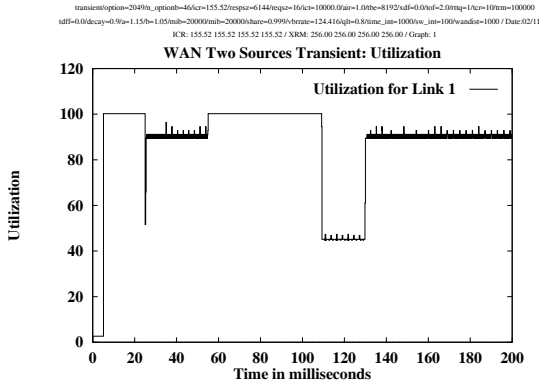
Figure 2: Results for a two source configuration in a LAN (2 destinations have same RTT)



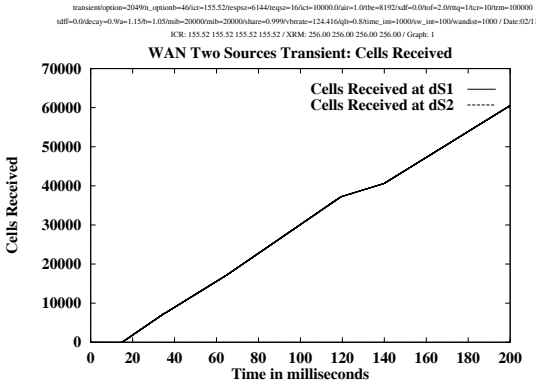
(a) Transmitted Cell Rate



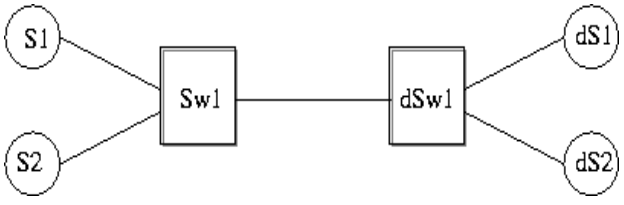
(b) Queue Length



(c) Link Utilization

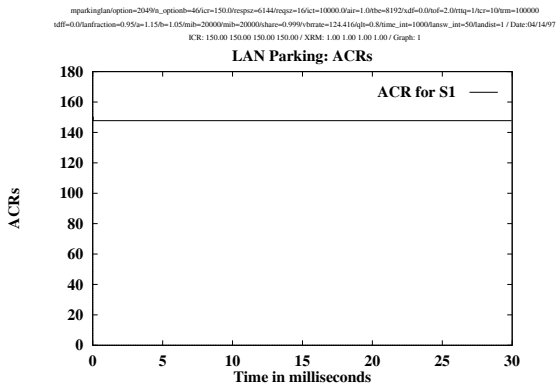


(d) Cells Received

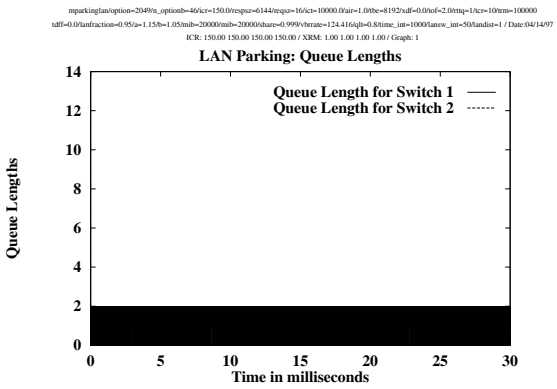


(e) Two source configuration

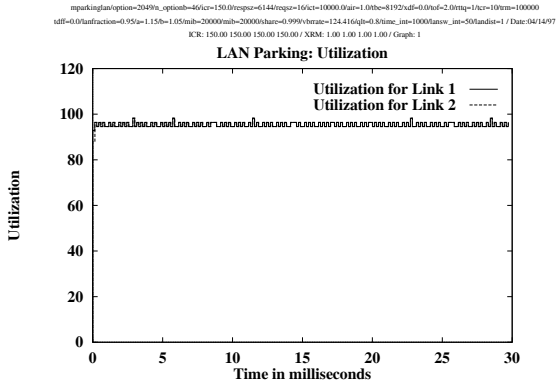
Figure 3: Results for a transient configuration



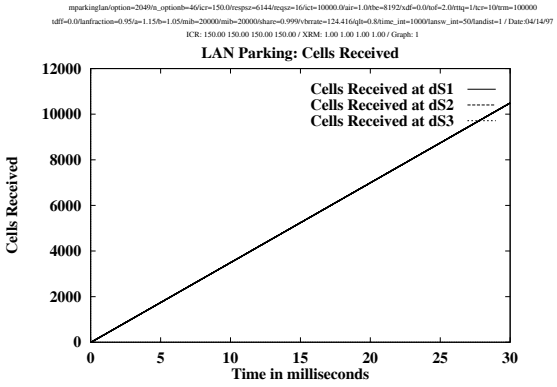
(a) Transmitted Cell Rate



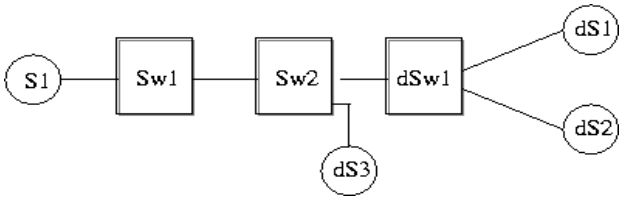
(b) Queue Length



(c) Link Utilization

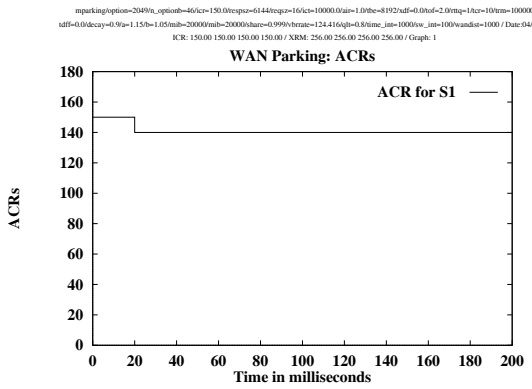


(d) Cells Received

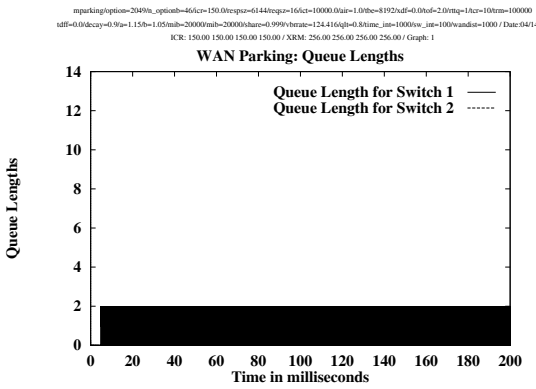


(e) Parking lot configuration

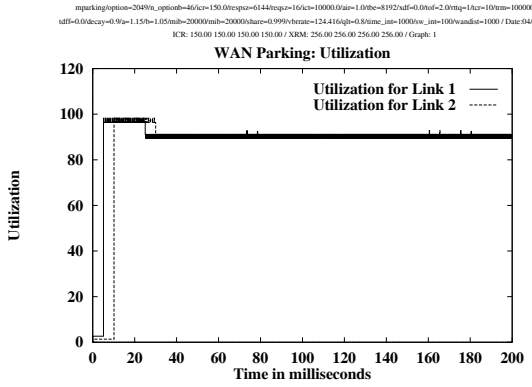
Figure 4: Results for a parking lot configuration in a LAN



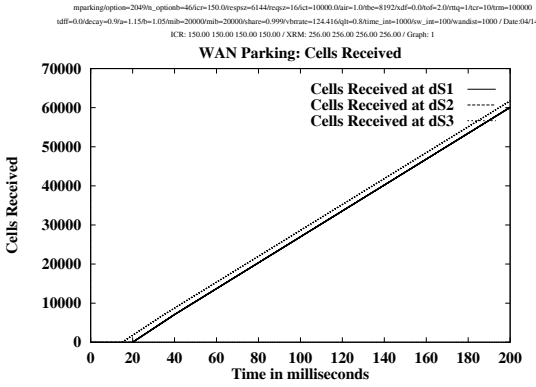
(a) Transmitted Cell Rate



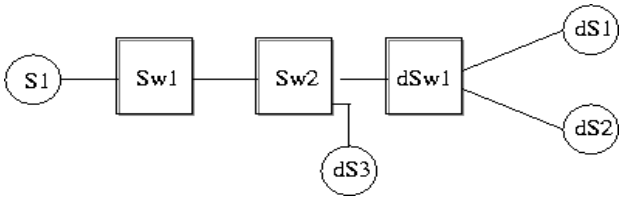
(b) Queue Length



(c) Link Utilization

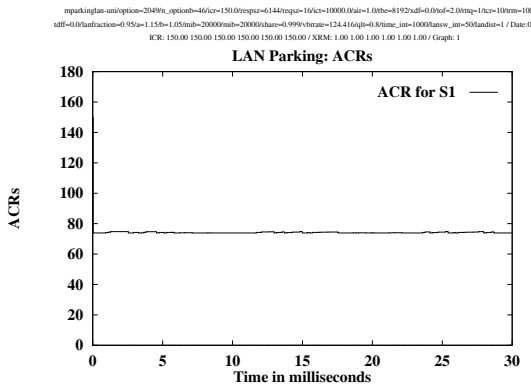


(d) Cells Received

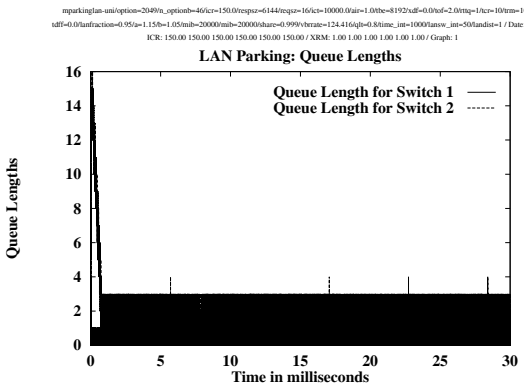


(e) Parking lot configuration

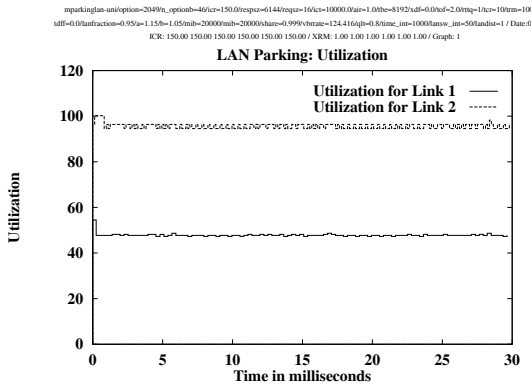
Figure 5: Results for a parking lot configuration in a WAN



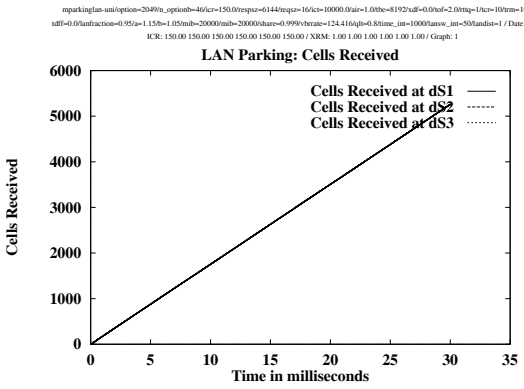
(a) Transmitted Cell Rate



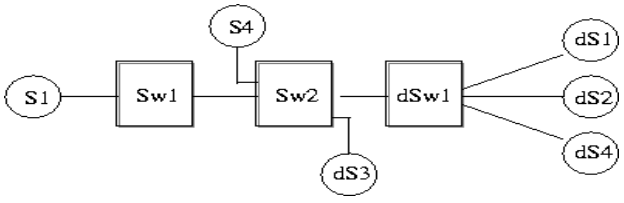
(b) Queue Length



(c) Link Utilization

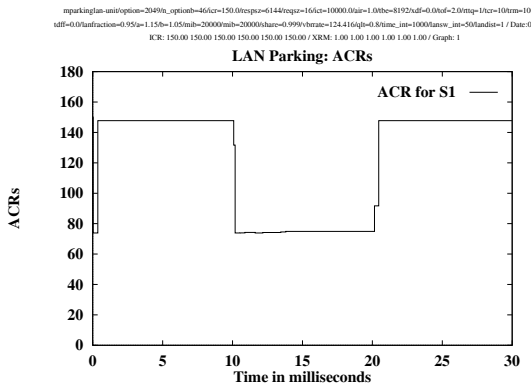


(d) Cells Received

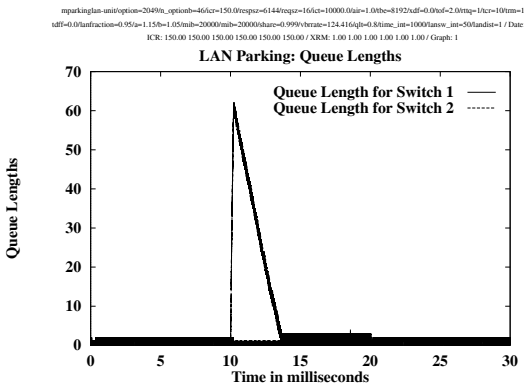


(e) Parking lot and point-to-point configuration

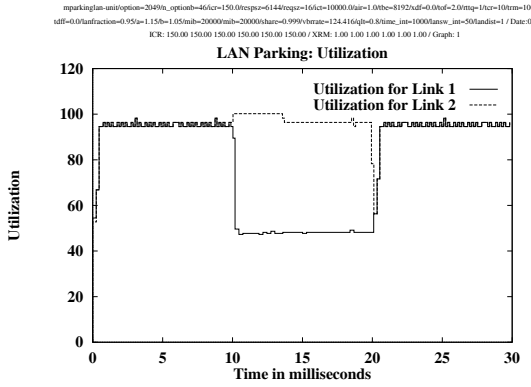
Figure 6: Results for a parking lot and point-to-point configuration in a LAN



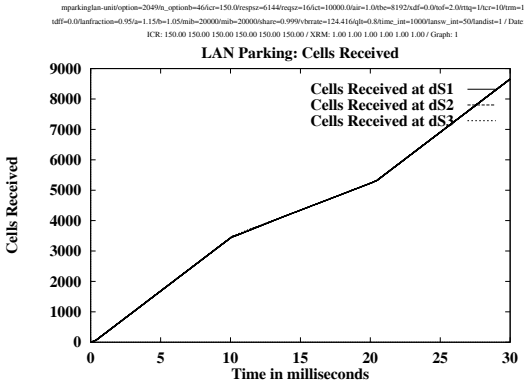
(a) Transmitted Cell Rate



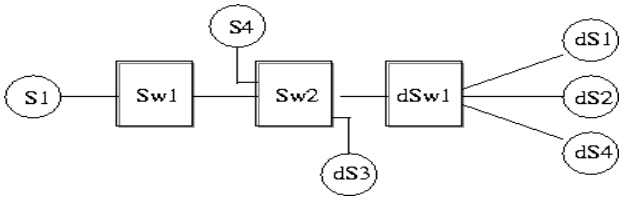
(b) Queue Length



(c) Link Utilization

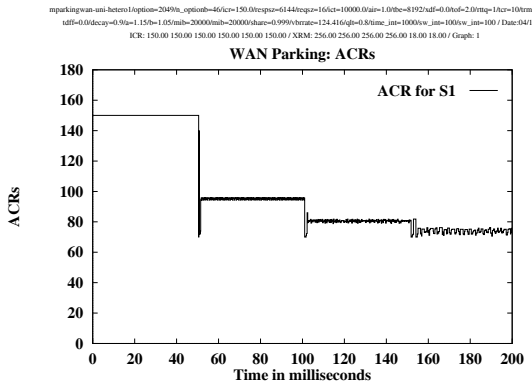


(d) Cells Received

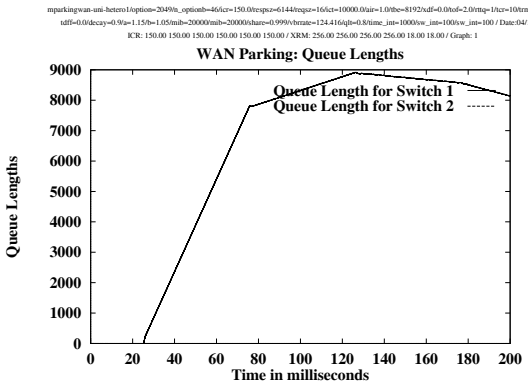


(e) Parking lot and point-to-point configuration

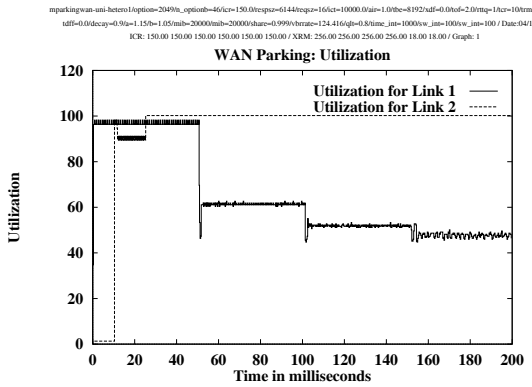
Figure 7: Results for a parking lot and transient point-to-point configuration in a LAN



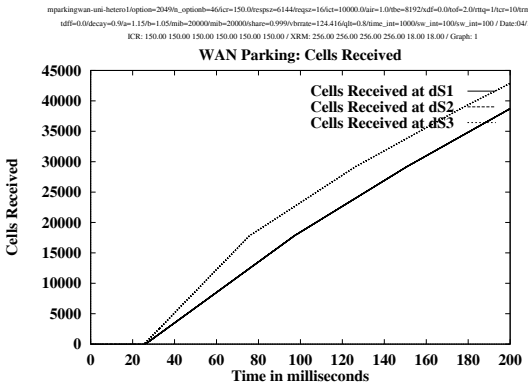
(a) Transmitted Cell Rate



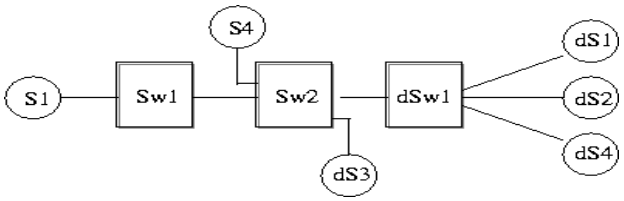
(b) Queue Length



(c) Link Utilization



(d) Cells Received



(e) Parking lot and point-to-point configuration

Figure 8: Results for a parking lot and point-to-point configuration in a WAN (long feedback delay)

are controlled, and little rate oscillations are observed. Compared to the previous configuration, it is clear that point-to-multipoint connections responded better here because *the bottleneck is not distant.*

- Configuration 7 [Parking lot and VBR point-to-point configuration]: As illustrated in figure 10(e), this configuration is the same as configuration 4 with all 1 km links, but the unicast VC is VBR. As seen in figure 10, the transient response is fast, and the queues are controlled. The first link is underutilized because the source cannot send at a high rate when VBR is on.

- Configuration 8 [Parking lot and 2 VBR]: This is a WAN configuration where all links are 1000 km. It is similar to configuration 7, but there is an additional unicast VBR VC sharing another portion of the path, as shown in figure 11(e). The two VBR VCs do not share any links and are set up such that when one is on, the other is off (one of them starts after 2 ms and the other starts after 22 ms, and both have a 20 ms on/off period).

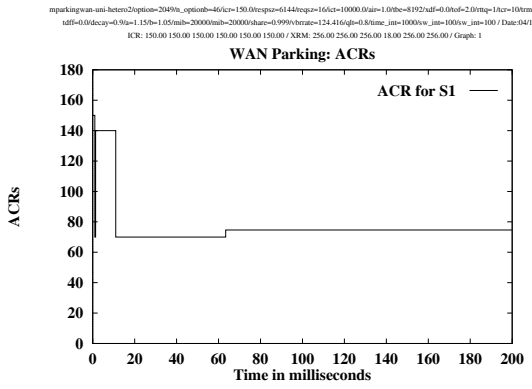
  The following values of ICR and RIF were used:
  RIF = 1, 0.03125
  ICR = 12 Mbps, 150 Mbps

  As seen in figures 11 through 14, *the utilizations of the links are low when the VBR connection on that link is off.* This is the disadvantage of taking the minimum of the rates indicated by all the leaves of the multicast tree. *If data applications can tolerate loss (such as stock market updates), better utilizations can be achieved by using more clever techniques.*

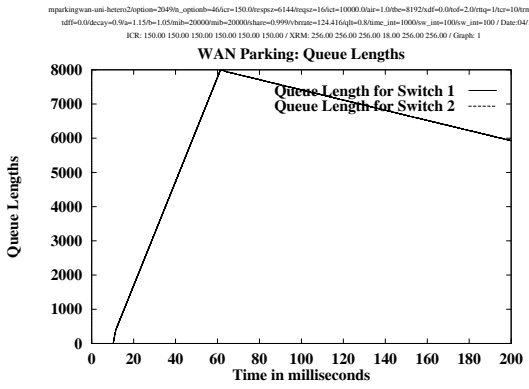
  For all values of RIF and ICR, we can see the low utilization when VBR is off, and the controlled queues. The effect of RIF and ICR is very clear in the rate graph and in the throughput values. *Small values of ICR are better to prevent overallocation till all feedback information is consolidated. However, a small RIF value can adversely affect the throughput and link utilization for schemes such as ERICA.*

- Configuration 9 [Parking lot and 2 VBR with long feedback delay]: This configuration is identical to the previous configuration (using high ICR and high RIF), but link 1 is much longer (5000 km) than the others (50 km). Thus the feedback delay is long when the second VBR is on (see figure 15(e)). Figure 15 illustrates the performance in this configuration. As seen in the figures, the high ICR and RIF values, together with the long feedback delay lead to large queues, especially initially. This is because VBR is not detected fast enough and ACR does not go down fast enough when VBR is on.

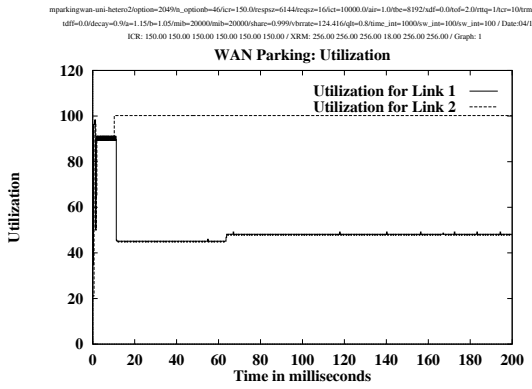
- Configuration 10 [Parking lot and 2 VBR with largely varying RTTs]: This configuration is identical to configuration 8 (with high ICR and high RIF), but link 2 is much longer (5000 km) than the others (50 km). Thus the RTT ratios are highly variant (see figure 16(e)). Figure 16 illustrates that queues are much smaller than the previous case, but utilization and throughput are also lower. There are small queue and rate spikes when the VBR source is not detected yet.



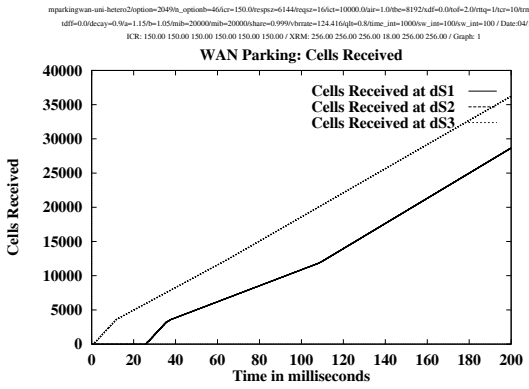
(a) Transmitted Cell Rate



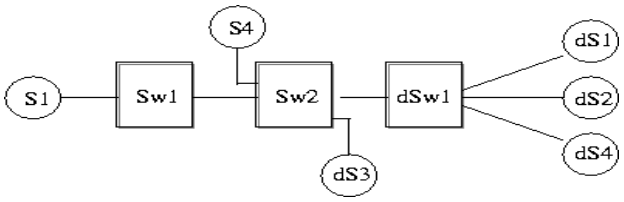
(b) Queue Length



(c) Link Utilization

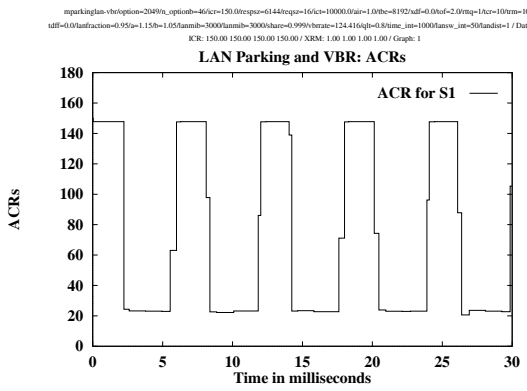


(d) Cells Received

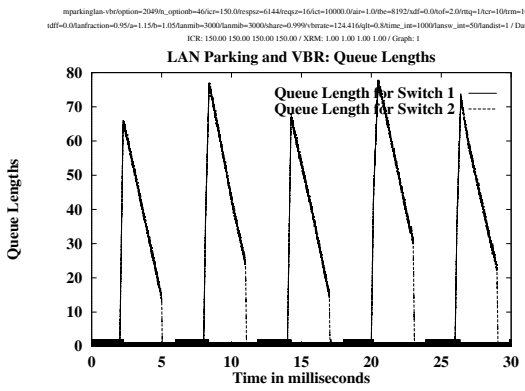


(e) Parking lot and point-to-point configuration

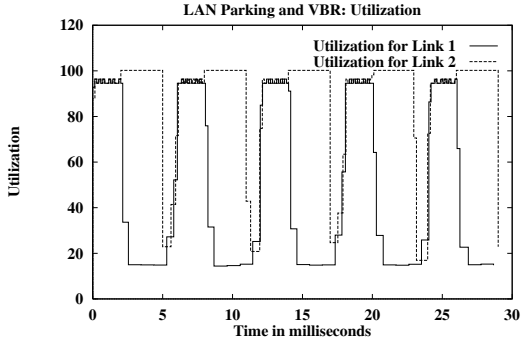
Figure 9: Results for a parking lot and point-to-point configuration in a WAN (largely varying RTTs)



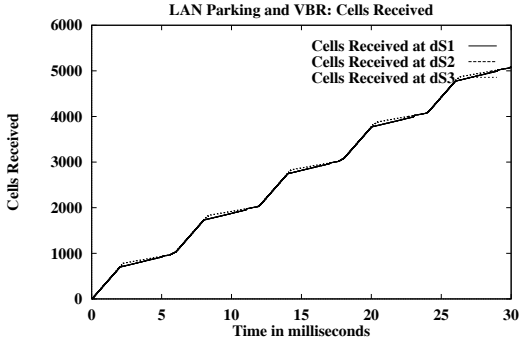
(a) Transmitted Cell Rate



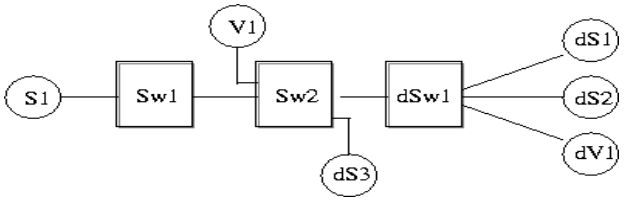
(b) Queue Length



(c) Link Utilization

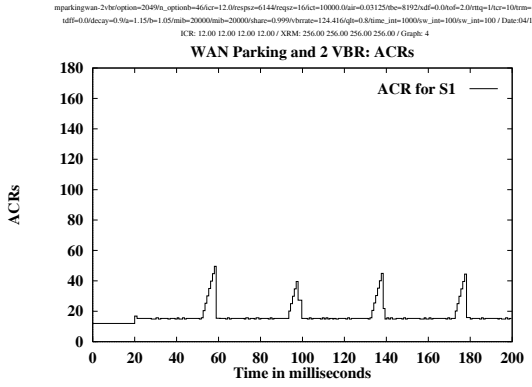


(d) Cells Received

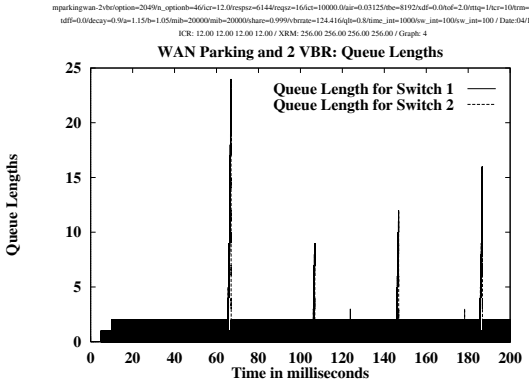


(e) Parking lot and VBR point-to-point configuration

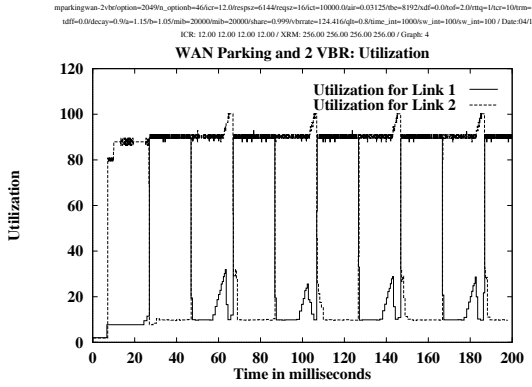
Figure 10: Results for a parking lot and VBR point-to-point configuration in a LAN



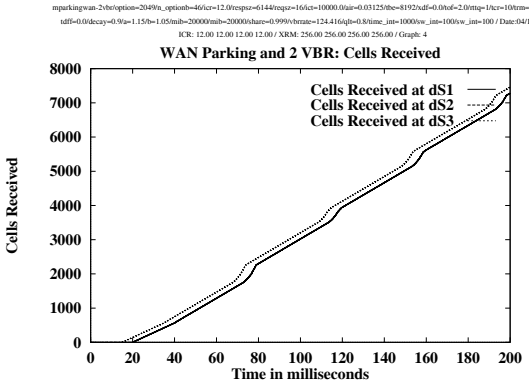
(a) Transmitted Cell Rate



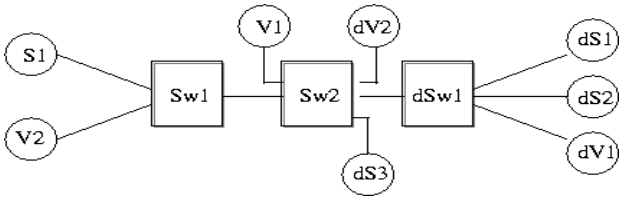
(b) Queue Length



(c) Link Utilization

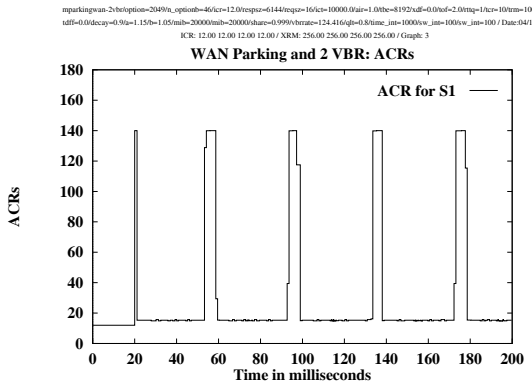


(d) Cells Received

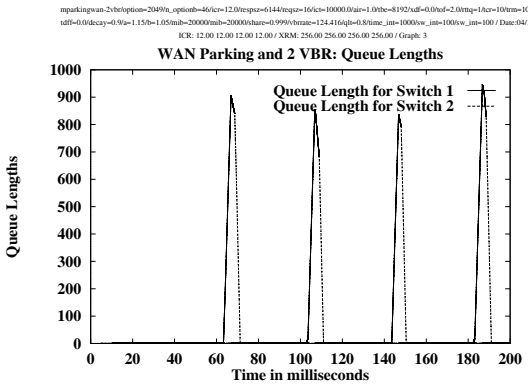


(e) Parking lot and 2 VBR configuration

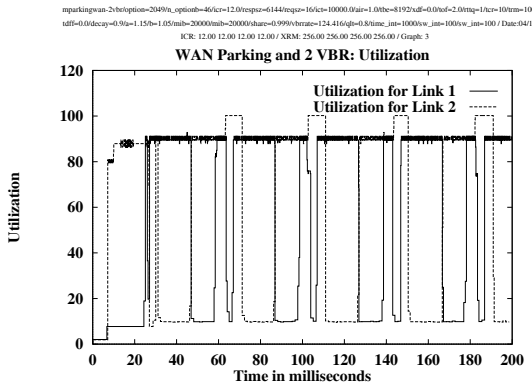
Figure 11: Results for a parking lot and 2 VBR configuration in a WAN (low ICR, low RIF)



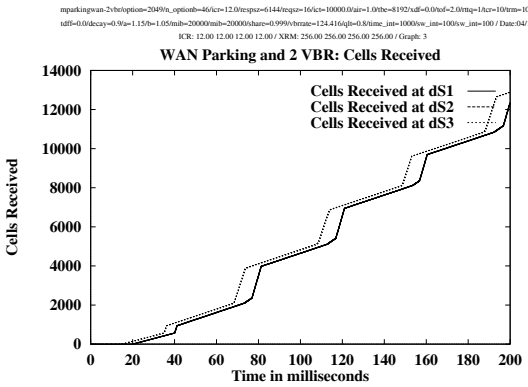
(a) Transmitted Cell Rate



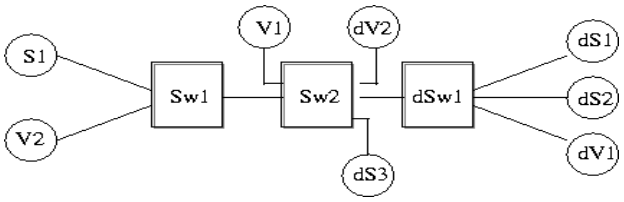
(b) Queue Length



(c) Link Utilization

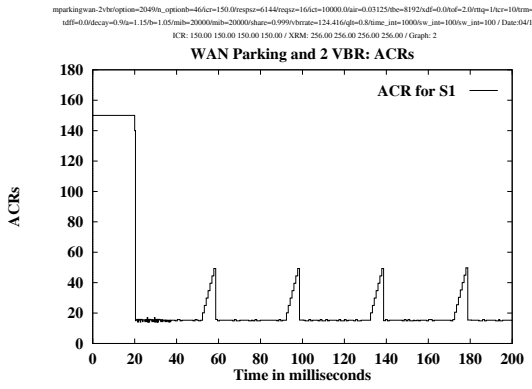


(d) Cells Received

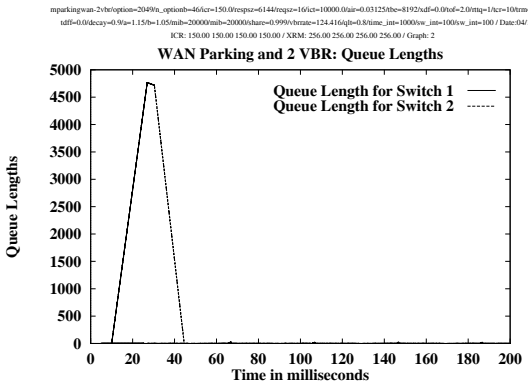


(e) Parking lot and 2 VBR configuration

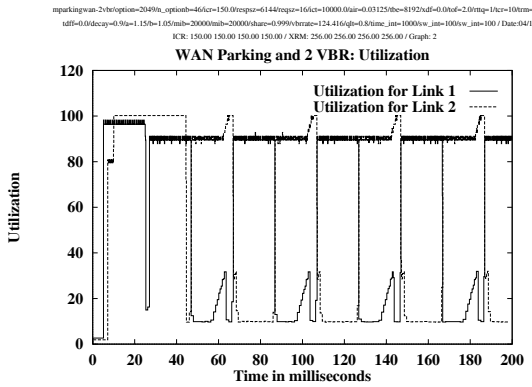
Figure 12: Results for a parking lot and 2 VBR configuration in a WAN (low ICR, high RIF)



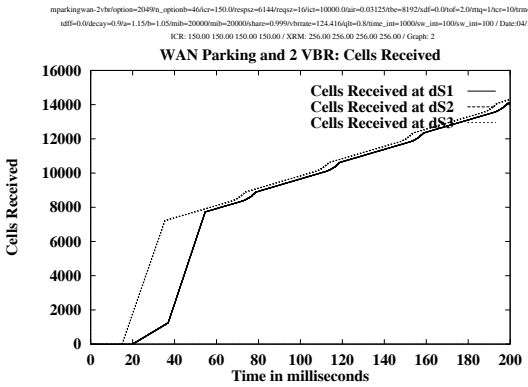
(a) Transmitted Cell Rate



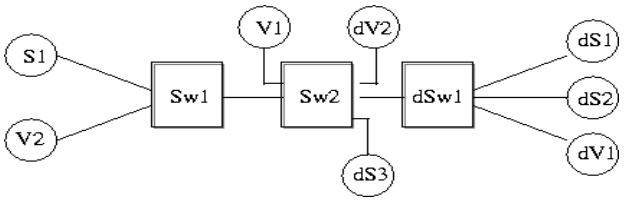
(b) Queue Length



(c) Link Utilization

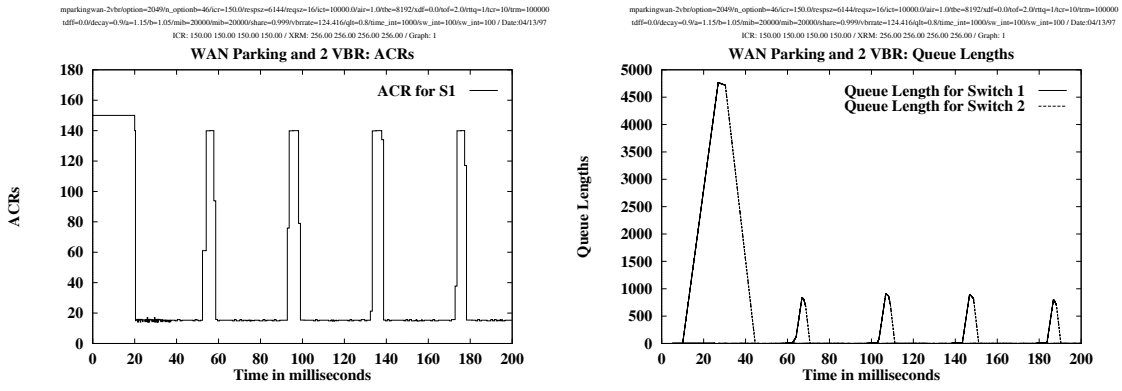


(d) Cells Received



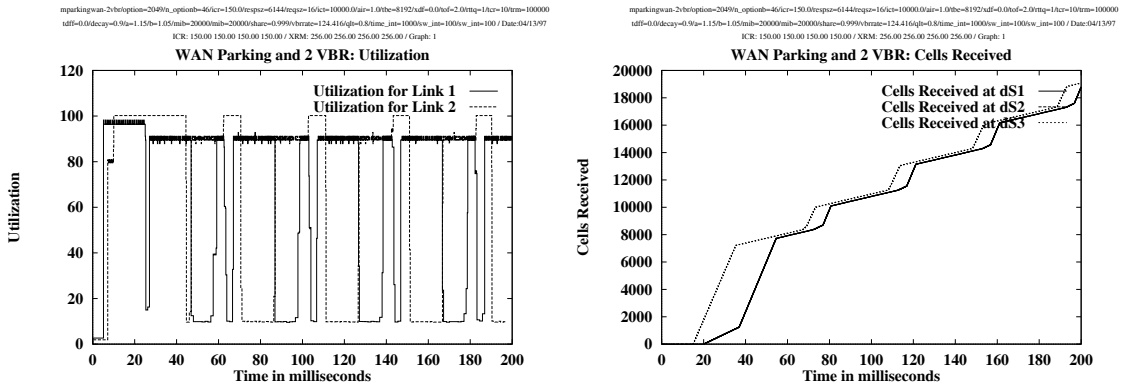
(e) Parking lot and 2 VBR configuration

Figure 13: Results for a parking lot and 2 VBR configuration in a WAN (high ICR, low RIF)



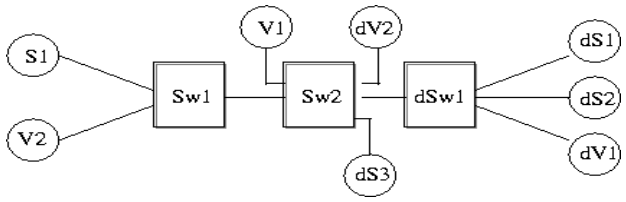
(a) Transmitted Cell Rate

(b) Queue Length



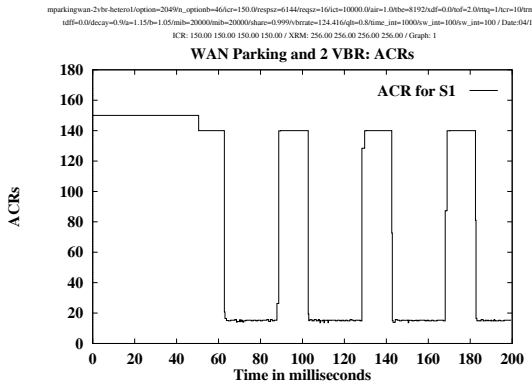
(c) Link Utilization

(d) Cells Received

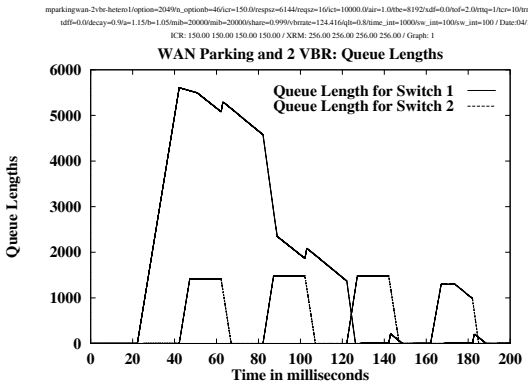


(e) Parking lot and 2 VBR configuration

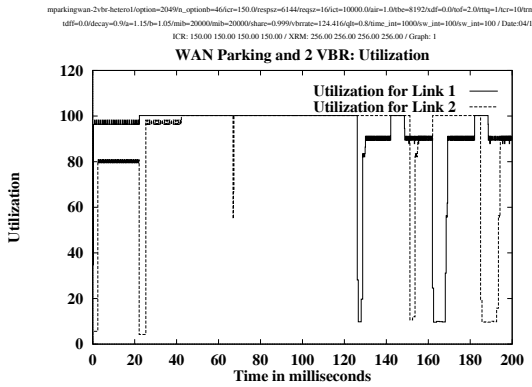
Figure 14: Results for a parking lot and 2 VBR configuration in a WAN (high ICR, high RIF)



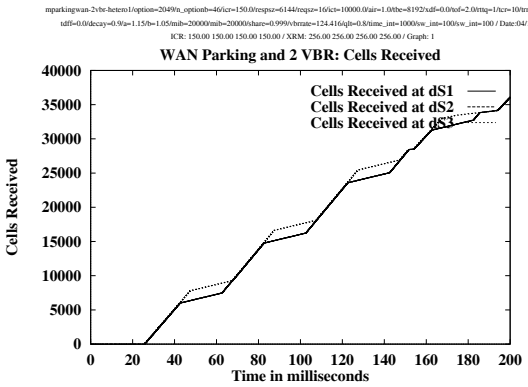
(a) Transmitted Cell Rate



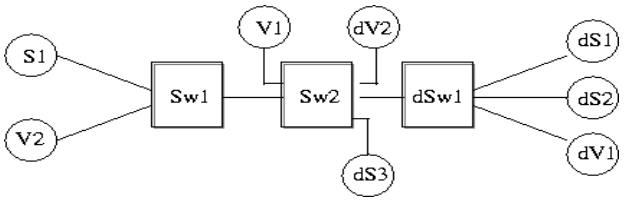
(b) Queue Length



(c) Link Utilization

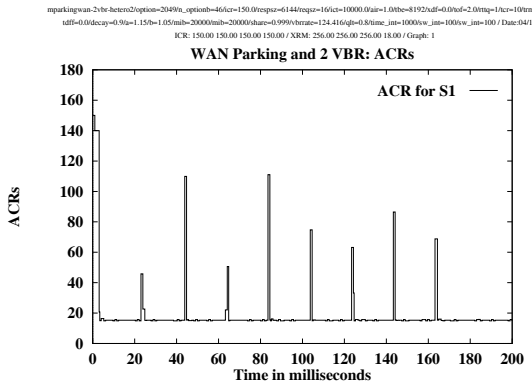


(d) Cells Received

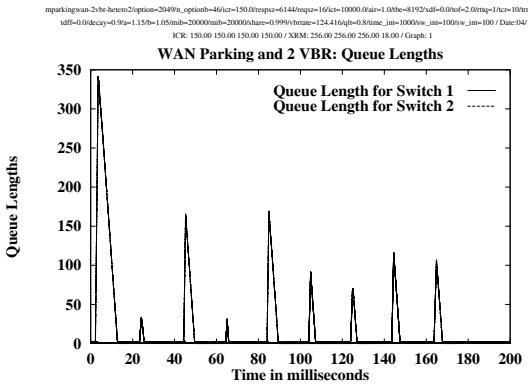


(e) Parking lot and 2 VBR configuration

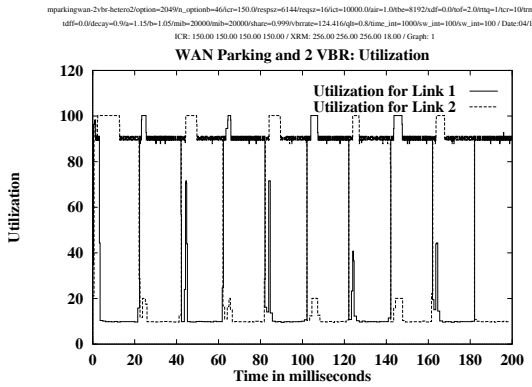
Figure 15: Results for a parking lot and 2 VBR configuration in a WAN (long feedback delay)



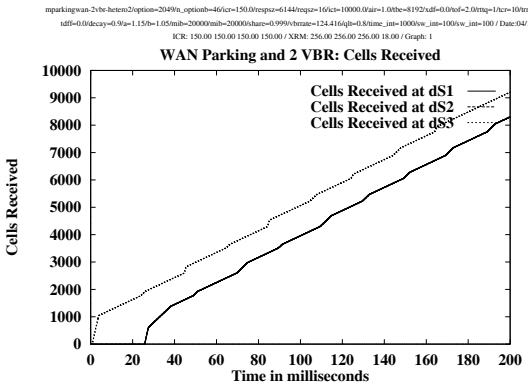
(a) Transmitted Cell Rate



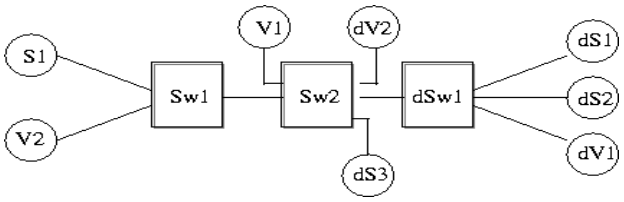
(b) Queue Length



(c) Link Utilization



(d) Cells Received



(e) Parking lot and 2 VBR configuration

Figure 16: Results for a parking lot and 2 VBR configuration in a WAN (largely varying RTTs)

- Configuration 11 [Bursty and infinite parking lot]: This configuration is similar to the parking lot configuration (configuration 3), except that the network is a WAN, and there are 2 traffic sources and six destinations, as illustrated in figure 17(e). One of the traffic source is bursty. It follows a closed-loop request-response type of connection with burst sizes of 6K. The other source, the infinite traffic source, shares link 1 and link 2 with the bursty source.

  As seen in figure 17, the infinite source rate drops in steps when bursty source sends. Due to the high ICR and the time it takes the infinite source to detect that the bursty source is sending, in addition to the lack timing accuracy of the large bursts with the feedback, queues are quite large and the link is almost always fully utilized. This configuration has the same problems with point-to-point connections, so the problems have simply carried over to the point-to-multipoint connections. ERICA+ solves those problems.

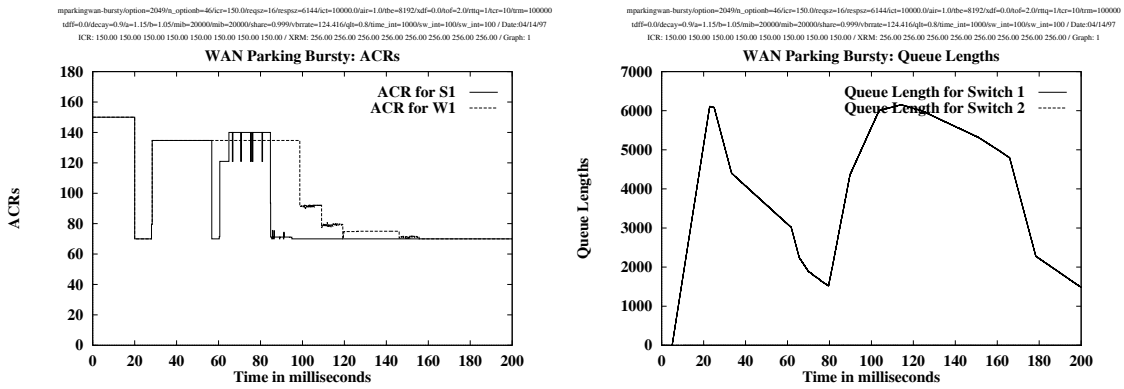
- Configuration 12 [Bursty and infinite parking lot with VBR]: As shown in figure 18(e), this configuration is similar to the previous one, but, in addition, a VBR connection shares link 2 with the bursty and infinite sources. Figure 18 illustrates that the performance is adversely affected by variation in both capacity and demand, but *the problems are the same for point-to-point connections.* Large queues build up.

- Configuration 13 [Bursty and infinite parking lot with VBR and long feedback delay]: This is identical to the previous configuration (figure 19(e)), but link 1 is much longer (5000 km) than the others (50 km), and hence the feedback delay is large. As illustrated in figure 19, the long feedback delay yields larger queues.

- Configuration 14 [Bursty and infinite parking lot with VBR and largely varying RTTs]: This configuration is also the same as configuration 12 (figure 20(e)), but link 2 is much longer (5000 km) than the others (50 km). The difference in RTTs of various branches also adversly affects the performance and yields larger queues, as shown in figure 20. As mentioned before, these problems are the same in point-to-point connections and are solved by switch schemes such as ERICA+.

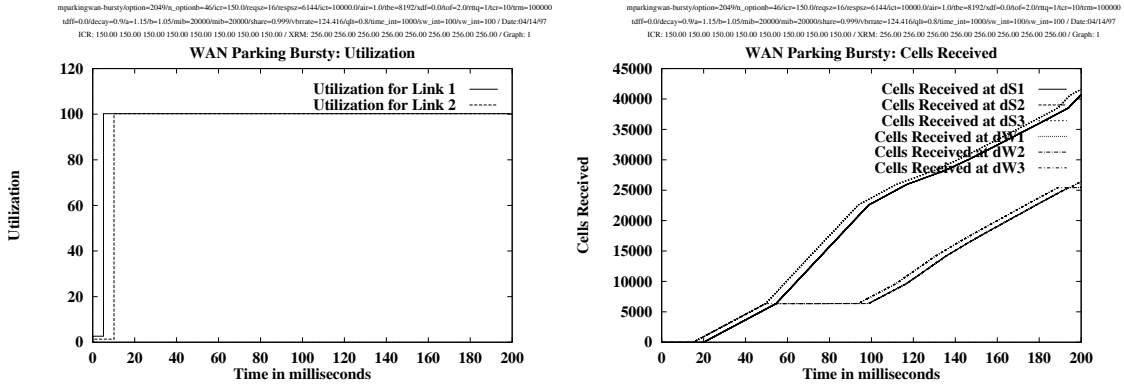
- Configuation 15 [Chain configuration]: As seen in figure 21(e), this configuration consists of 4 switches (a much more complex version of this configuration appears in [8]). The point-to-multipoint connection has 3 destinations at different RTTs (at each of the 3 distant switches). Destination 1 is the most distant destination. A unicast connection shares the link only with the connection to the last leaf. All links are 50 km, except link connecting second and third switches is 500 km and link connecting third and fourth switches is 5000 km.

  This is an excellent configuration to illustrate the *consolidation noise problem* because feedback from the distant leaves is unlikely to reach branch points at every iteration. Figure 21 illustrates that the rate and utilization oscillations due to the RM consolidation noise. *The rate oscillates from PCR (when no feedback from the most distant leaf arrives) to PCR/2 (when feedback from that leaf is received).* It is debatable whether



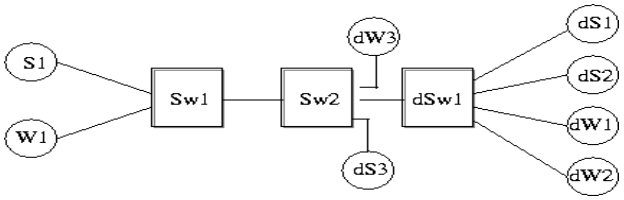
(a) Transmitted Cell Rate

(b) Queue Length



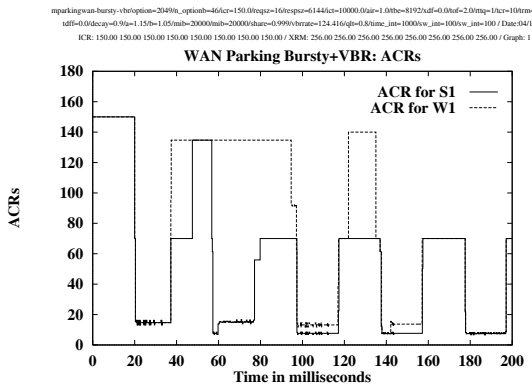
(c) Link Utilization

(d) Cells Received

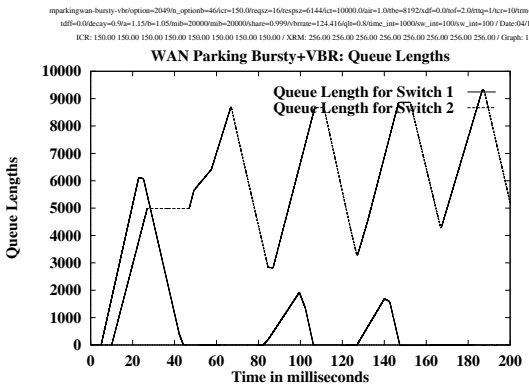


(e) Bursty and infinite parking lot configuration

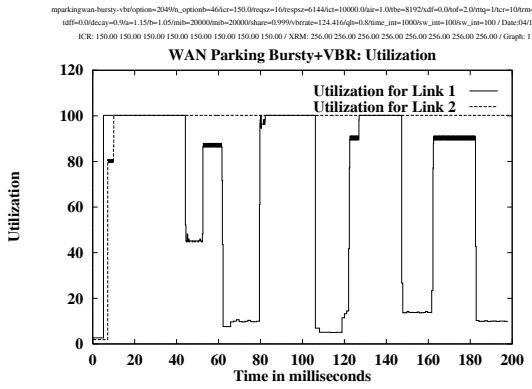
Figure 17: Results for a bursty and infinite parking lot configuration in a WAN



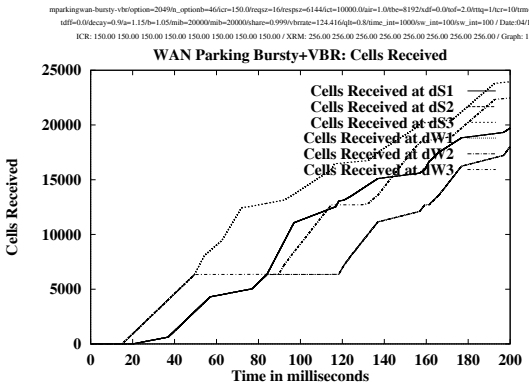
(a) Transmitted Cell Rate



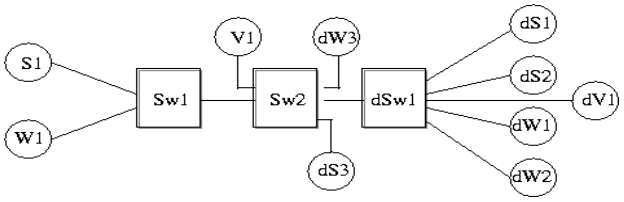
(b) Queue Length



(c) Link Utilization

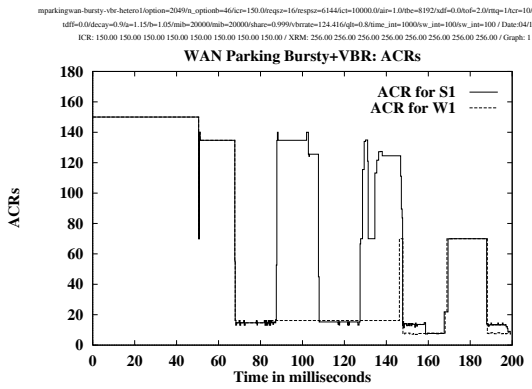


(d) Cells Received

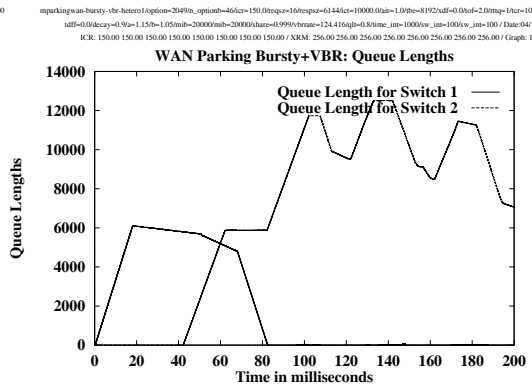


(e) Bursty and infinite parking lot configuration with VBR

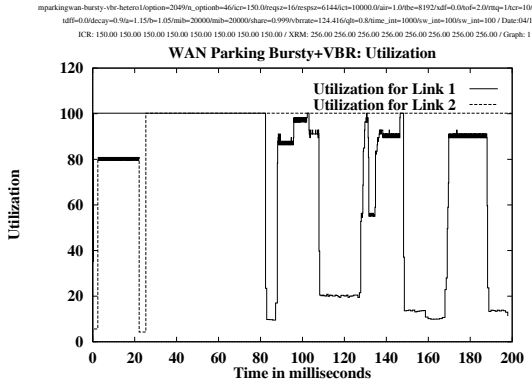
Figure 18: Results for a bursty and infinite parking lot configuration with a VBR connection in a WAN



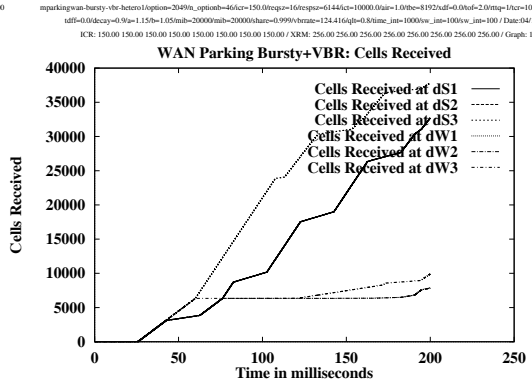
(a) Transmitted Cell Rate



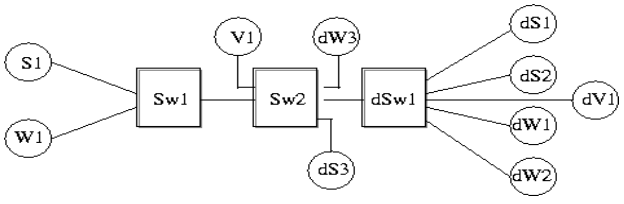
(b) Queue Length



(c) Link Utilization

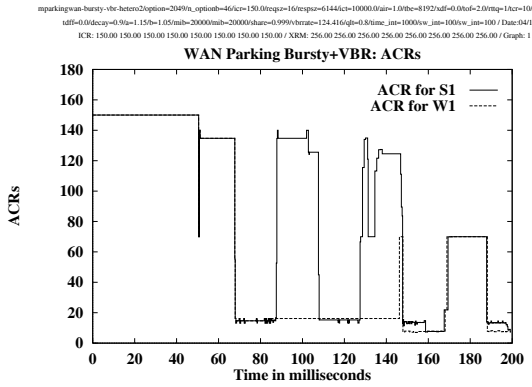


(d) Cells Received

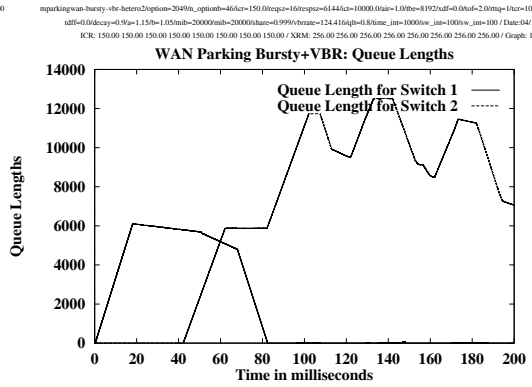


(e) Bursty and infinite parking lot configuration with VBR

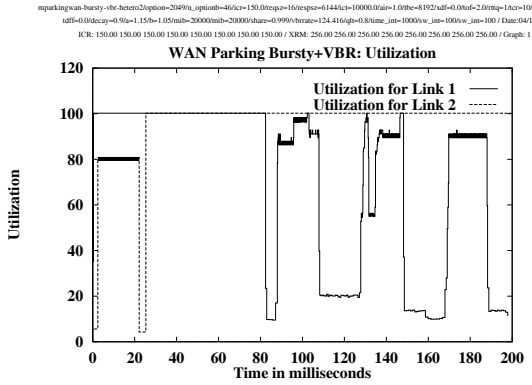
Figure 19: Results for a bursty and infinite parking lot configuration and a VBR connection in a WAN (long feedback delay)



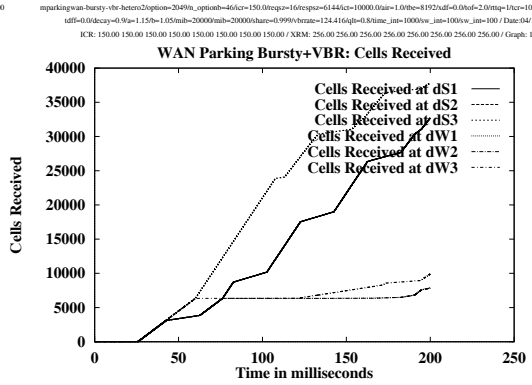
(a) Transmitted Cell Rate



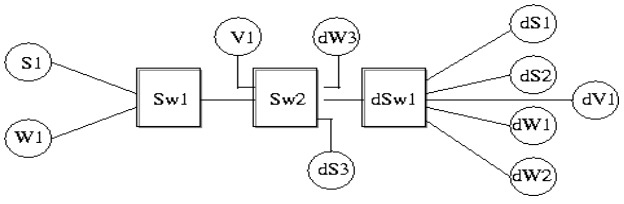
(b) Queue Length



(c) Link Utilization



(d) Cells Received



(e) Bursty and infinite parking lot configuration with VBR

Figure 20: Results for a bursty and infinite parking lot configuration and a VBR connection in a WAN (largely varying RTTs)

modifying the algorithm to give feedback only when all branches have sent their feedback is worth the additional complexity incurred.

- Configuration 16 [Fork configuration]: This configuration also consists of 4 switches (a much more complex version of this configuration appears in [8]). There are 2 multicast connections as shown in figure 22(e). Each of the multicast connections has a source connected at one of the 2 intermediate switches. The 3 destinations are connected at each of the other 3 switches (the connection forks at the first switch, and hence, the name of the configuration, fork). All links are 1000 km.

  Figure 22 illustrates the performance of the fork configuration. There are some initial queues due to the high ICR, but the queues quickly become very small when both sources detect the bottleneck links and send at the optimal rates. The bottleneck links are fully utilized, while other links are half utilized. Fairness among the connections is preserved.

6.3 Comparison with Multiple Point-to-Point Connections

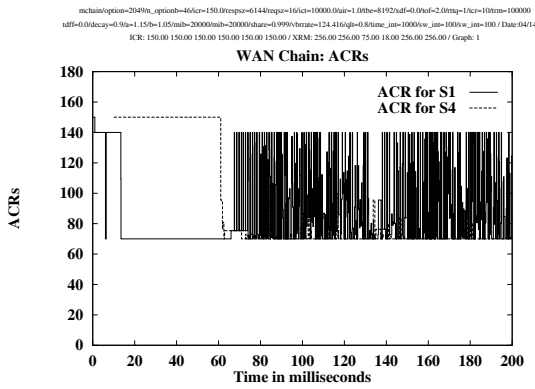
To compare a point-to-multipoint configuration with a similar point-to-point configuration, configuration 8 with high ICR and high RIF values (figure 14) was selected as the multipoint configuration to use in the comparison. Two possibilities exist to correspond to the mutlicast parking lot connection: we can have three sources sending to three destinations, or one source sending to one destination. Figures 23 and 24 illustrate the results for both configurations.

Figure 23 shows the performance of with three ABR unicast connections in a modified upstream configuration (with VBR connections). The configuration is illustrated in figure 23(e). Clearly the huge initial queues are caused by the high ICR value. In addition, there is slight overallocation when VBR is on. Destination 3 is unaffected by one of the VBR connections, because it does not share the link with the second VBR source. Running this simulation for 400 ms shows that queue only stabilizes after 400 ms. Note that the number of cells received is comparable to the multicast scenario, but the number of cells sent is triple.

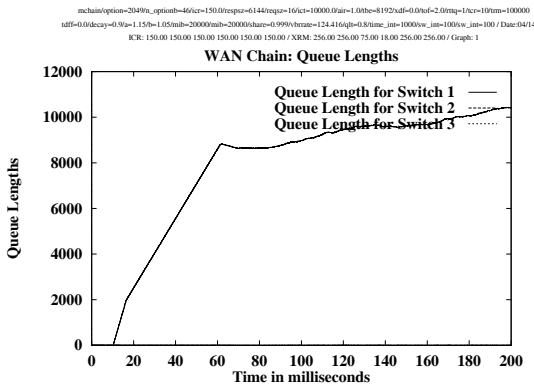
In a one source equivalent (see figure 24(e)), the load is alleviated. As seen in figure 24, the high ICR does not cause problems since there is only one source. It is actually this configuration that is more comparable to the multicast counterpart. The performance suffers from *the link underutilization problem* when VBR is on in another portion of the path, and slight overallocation when VBR is on.

7 Summary and Discussion

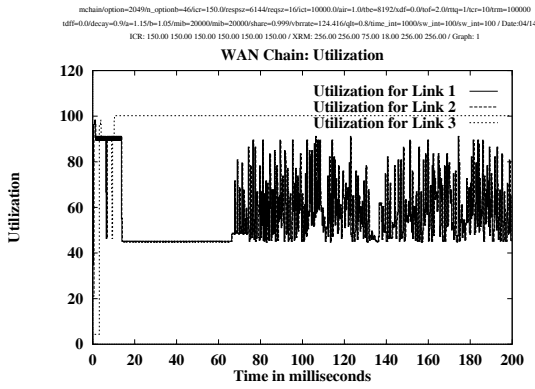
Table 1 summarizes the main results. The first column of the table shows the configuration number as given in the last section. The second column gives the maximum queue length of any of the switches. When that number is high only because of high ICR values, the steady



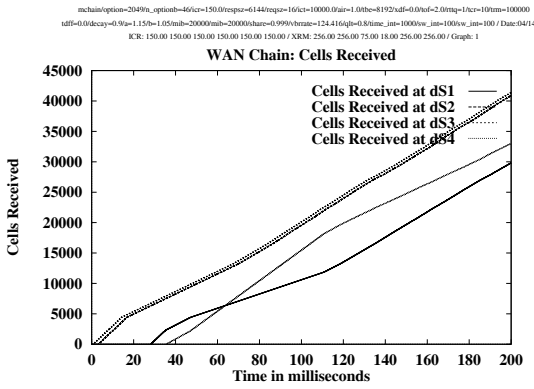
(a) Transmitted Cell Rate



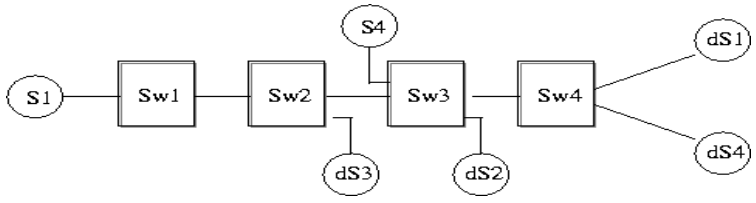
(b) Queue Length



(c) Link Utilization

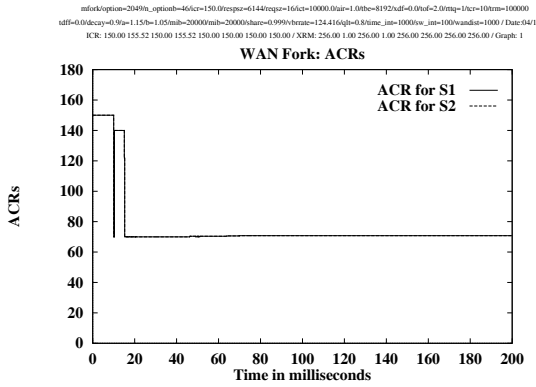


(d) Cells Received

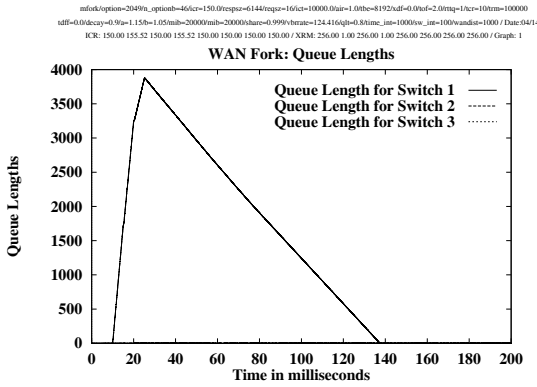


(e) Chain configuration

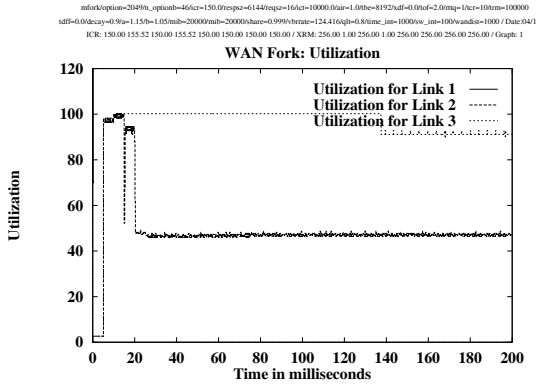
Figure 21: Results for a chain configuration in a WAN



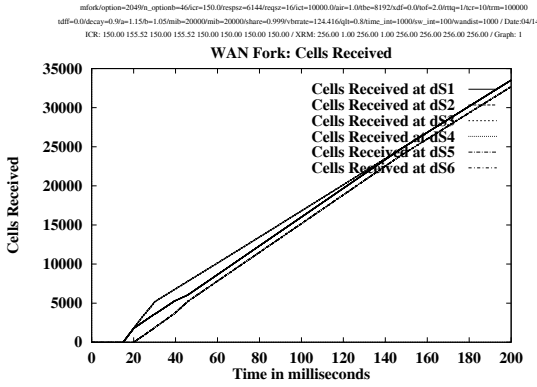
(a) Transmitted Cell Rate



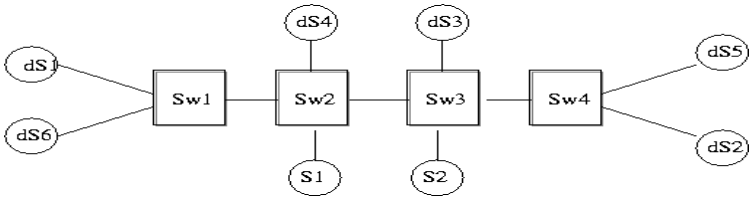
(b) Queue Length



(c) Link Utilization

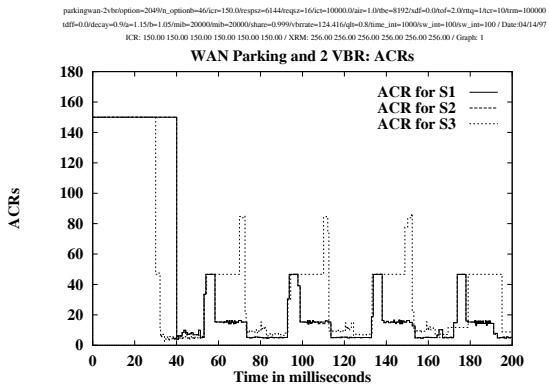


(d) Cells Received

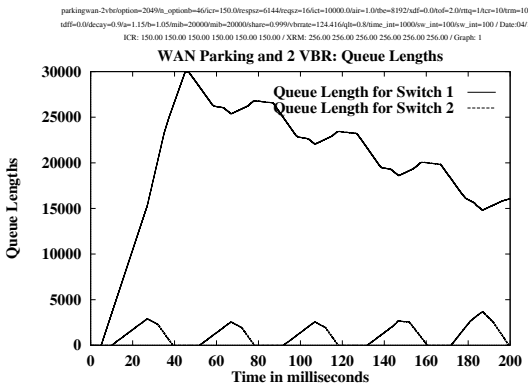


(e) Fork configuration

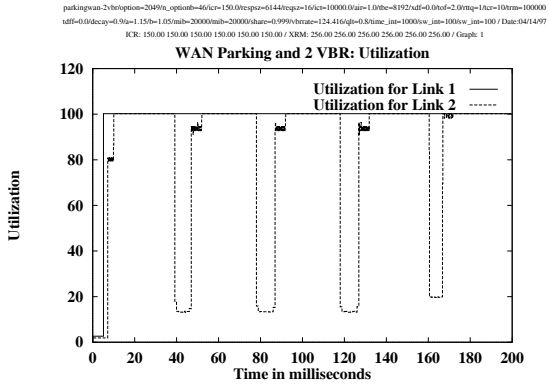
Figure 22: Results for a fork configuration in a WAN



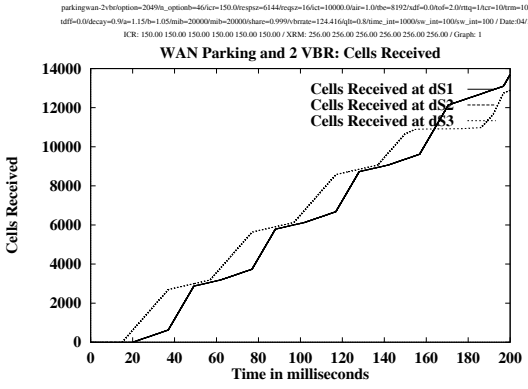
(a) Transmitted Cell Rate



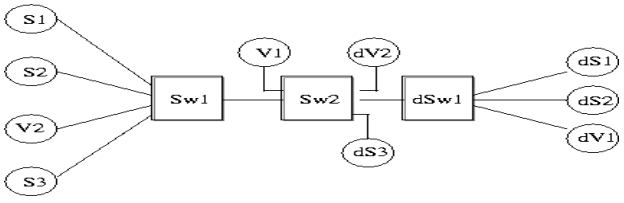
(b) Queue Length



(c) Link Utilization

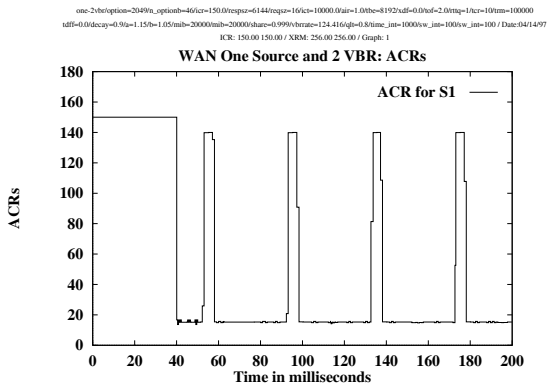


(d) Cells Received

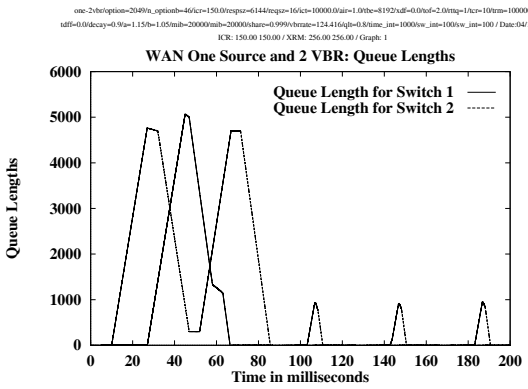


(e) Three source and 2 VBR configuration

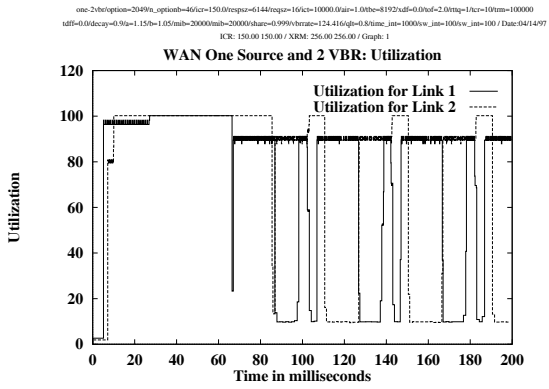
Figure 23: Results for a modified upstream configuration and 2 VBR connections in a WAN



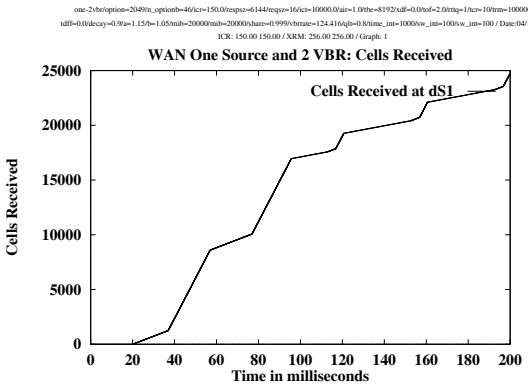
(a) Transmitted Cell Rate



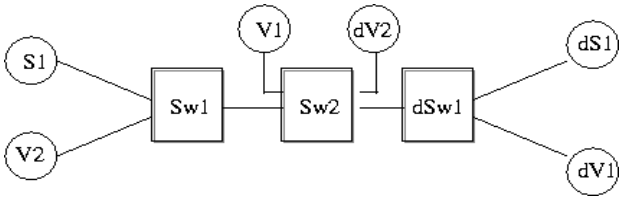
(b) Queue Length



(c) Link Utilization



(d) Cells Received



(e) One source and 2 VBR configuration

Figure 24: Results for a one source configuration and 2 VBR connections in a WAN

state maximum queue length is given. The third column gives the average link utilization for the bottleneck link throughout the simulation period. Finally, the last column shows the cells received at the destination, indicating the throughput achieved.

Table 1: Summary of Simulation Results

Configuration	Maximum Queue (cells)	Average Utilization	Cells Received
1	2	95	10000
2	15	95	10000
2b	1500	85	60000
3	2	95	10000
3b	2	91	60000
4	15	100	9000+other
4b	70	95	6000+other
5	9000	98	42000+other
6	8000	98	36000+other
7	80	80	5000+VBR
8, low ICR, low RIF	24	50	7000+VBR
8, low ICR, high RIF	900	65	12000+VBR
8, high ICR, low RIF	24	55	14000+VBR
8, high ICR, high RIF	900	70	19000+VBR
unicast 3 source	3000	70	14000 each +VBR
unicast 1 source	900	60	25000+VBR
9	1500	90	36000+VBR
10	170	50	9000+VBR
11	6000	98	40000+25000
12	9000	98	23000+20000+VBR
13	12000	98	35000+30000+VBR
14	13000	98	36000+31000+VBR
15	10400	98	40000
16	50	99	32000 each

The main conclusions that can be drawn from this performance analysis include:

- The definition of max-min fairness for point-to-multipoint connections is an extension of the definition for point-to-point connections
- Switch algorithm extension frameworks proposed for point-to-multipoint congestion control preserve the efficiency and fairness properties of the original point-to-point switch scheme employed in the framework

- All problems with a point-to-point scheme re-appear (and maybe accentuated) in the point-to-multipoint extension of this scheme
- The major problem specific to point-to-multipoint connections is the consolidation noise problem, which is clearly seen in many configurations. This problem occurs when there are distant bottlenecks, and feedback from those bottlenecks is not always received in a timely fashion, when the RM cells need to be returned by the branching point. Alleviating the consolidation noise problem (by ensuring feedback is received from all leaves) may create additional problems, such as slow transient response and additional complexity
- Point-to-multipoint connections may suffer from a slow response to distant bottlenecks. The source relies on partial feedback information from a subset of the branches, and hence may overload the network
- Many links on the branches of a point-to-multipoint connection may be underutilized because bottlenecks exist on other branches of the tree
- Transient queues can be avoided by setting the RIF ABR source parameter to a small value. Although conservative values are advisable for point-to-multipoint connections (where feedback response from distant bottlenecks is not always available), such small values have the adverse effect of slowing the rise to the optimal value
- Initial rate overallocation before feedback is received from all bottlenecks can be overcome by setting the ICR parameter to small values, but if ICR depends on the RTT, should it change when nodes join/leave the group?

References

[1] Raed Awdeh, Fred Kaudel, and Osama Aboul-Magd. Point-to-multipoint behavior of ABR. ATM-FORUM/95-0941, August 1995.

[2] Flavio Bonomi, Kerry Fendick, and Nanying Yin. ABR point-to-multipoint connections. ATM-FORUM/95-0974R1, August 1995.

[3] The ATM Forum. The ATM forum traffic management specification version 4.0. ftp://ftp.atmforum.com/pub/approved-specs/af-tm-0056.000.ps, April 1996.

[4] Doug Hunt. Open issues for ABR point-to-multipoint connections. ATM-FORUM/95-1034, August 1995.

[5] Raj Jain, Sonia Fahmy, Shivkumar Kalyanaraman, and Rohit Goyal. ABR switch algorithm testing: A case study with ERICA. ATM-FORUM/96-1267, October 1996.

[6] Raj Jain, Shivkumar Kalyanaraman, Rohit Goyal, Sonia Fahmy, and Ram Viswanathan. ERICA switch algorithm: A complete description. ATM-FORUM/96-1172, August 1996.

[7] Wenge Ren. Congestion control for data traffic over ATM networks. PhD proposal. Available through W. Ren's home page, 1996.

[8] Wenge Ren, Kai-Yeung Siu, and Hiroshi Suzuki. On the performance of congestion control algorithms for multicast ABR service in ATM. In *Proceedings of IEEE ATM'96 Workshop, San Francisco*, August 1996.

[9] Lawrence Roberts. Rate based algorithm for point to multipoint ABR service. ATM-FORUM/94-0772R1, November 1994.

[10] Lawrence Roberts. Addition to TM spec 4.0 on point-to-multipoint. ATM-FORUM/95-0339, April 1995.

[11] Lawrence Roberts. Point-to-multipoint ABR operation. ATM-FORUM/95-0834, August 1995.

[12] Kai-Yeung Siu and Hong-Yi Tzeng. Congestion control for multicast service in ATM networks. In *Proceedings of the IEEE GLOBECOM*, volume 1, pages 310–314, 1995.

All our papers and ATM Forum contributions are available through http://www.cis.ohio-state.edu/~jain/