****

ATM Forum Document Number: ATM Forum/95-1343

****

Title: Straw-Vote comments on TM 4.0 R8

****

Abstract:

Several problems with Xrm and ICR computation using RTT and
use of rule 5 are pointed out.

****

Source:

Raj Jain, Shiv Kalyanaraman, Sonia Fahmy, and Fang Lu
The Ohio State University
Department of CIS

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

Saragur Srinidhi
NASA Lewis Research Center and Sterling Software
Scientific Systems Division
21000 Brookpark Road, MS 141-1
Cleveland,OH 44135
Phone: 216-433-8987, Fax: 216-433-8000

The presentation of this contribution at the ATM Forum is sponsored by
NASA.

****

Date: October , 1995, Honolulu

****

Distribution: ATM Forum Technical Working Group Members
(Traffic Management)

****

****

Here are our comments on the TM 4.0 95-0013R8 specification sent
out for straw vote:

1. Major Comment: XRM Range

In section 5.10.3.1, the parameter Xrm has a specified range from
0-255. And a note says a 24 bit implementation may be preferable
for large delay bandwidth product situations.

In August meeting, after 3 months of simulation analysis we made a presentation as to why XRM of 256 limits extensibility of ATM networks to high speed networks or long delay networks. The particular note was discussed and not voted. Instead, with an almost unanimous vote, the group passed a motion "XRM is an integer. Its size is implementation dependent."

The next day, somehow the note was slipped in again with a motion without proper study or justification.

As it stands right now, Xrm is not signalled and is an internal parameter of the NIC. There is no need for the standard to specify its width. There are numerous other quantities whose width will be decided by the implementors. We should not justify one particular vendor's choice by a note.

The note also introduces inconsistency in the spec. Xrm is computed as follows:

Xrm = min (CIF/Nrm, PCR*RTT/Nrm)

Since the maximum value of CIF is 2**24 and minimum value of Nrm is 2, it is possible that for some values of CIF and Nrm, an 8-bit XRM will not be able to hold the result of the above equation. Therefore, by justifying that this value is sufficient we are automatically assuming that certain ranges of Nrm and CIF will not be used.

This is a major issue for NASA and all companies trying to support ATM over satellite links.

2. Major Comment: Xrm and ICR are not properly controllable

In 5.10.3.2.1, pg 53, Xrm and ICR are calculated from CIF and RTT. There are three problems with this approach. First, RTT is a highly random value.

Its value at the time of connection setup affects the performance of the VC for its entire life. Unless we come up with a renegotiation mechanism whereby as RTT changes, XRM and ICR can be readjusted, the use of one instance of RTT (or its percentiles at that instant) makes the scheme "non-dynamic" and "random".

The second problem with this formula is that both Xrm and ICR are correlated. ICR determines the rate at which idle sources can start transmitting and Xrm determines the cells that they can send during the first round trip. A switch may want to give high or low ICR depending upon the number of active VCs and totally independently give high or low Xrm depending upon its buffer size. Currently, this is not possible. If RTT is high both at the time of connection setup, both ICR and XRM are low. If RTT is low, both ICR and XRM are high.

The third problem with this formula is that it often gives values that are not what one would use for proper operation. For example, for LANs with RTTs of a few microseconds, we found that XRM value comes out to be less than 1. Rounding it up to 1 means that XRM is triggered on almost every RM cell event.

There is a slight dependence of CIF and ICR on path length. For LANs, we need small values of CIF, for WANs we need medium values of CIF, and for GANs (Global Area Networks), we need large values of CIF. Thus, qualifying a path as one of the three (or four) possible categories would help decide the proper value of CIF. By

having this formula where CIF is a continuous function of RTT, we have stretched the relationship too far. We have made it an "precise" function of a "random" quantity. Thus, the final result is "precisely random." It is not precise.

4. Finally, there is a typo on Pg 53, section 5.10.3.2.1 :

ICR = Min(ICR, a*CIF/RTT)

should read

ICR = Min(PCR, a*CIF/RTT)

5. In 5.10.3.2.1, pg 53, in the formula, Xrm = Min (CIF/Nrm, PCR*RTT/Nrm) it is not clear whether to round up or truncate the real value obtained. Observe that Xrm = 0 is also an acceptable value.

Note: We observe that Xrm is like a timeout and in conventional network design, timeout implied a serious network situation like loss. On execution of timeout, a strong decision is taken (like bringing window sizes to 1)... Xrm policy should also be viewed in similar light}

6. Section 5.10.3.2, Pg 52 : "It is recommended that the queueing delay be estimated as the 95%ile of the delay distribution"

To get this "random" result we have to keep track of 95-percentile delay which is not a trivial task. Is all this complexity worth the final random result?

7. In section 5.10.4, Rule 5 is simply broken. It does not achieve its original intended function of "ACR Retention."

Source Rule 5a has a sharp slope even if we choose low values of TDFF as recommended in the base vectors. For long links we observe drops to ICR as the common case even for persistant source simulations. This coupled with a cascade of feedback requesting increase in rate, causes undesirable and persistant oscillations in source rates. We observe that this behavior is because the timeout uses a current (possibly transient) value of ACR and does not differentiate between a low rate, idle, network-forced idle and ACR retentive source. The problem is not restricted to long delay links - it is merely linked to a possibly low, static value of ICR and using transient ACR values. We note that unnecessary oscillations can affect the ACR policing mechanisms.

The performance of current rule 5a, which requires log and exponential computation in the NIC is as good as that of replacing it simply with ACR <- ICR. We are not recommending that this be done but want to point out that the complexity does not always means "better."

The original purpose of rule 5a was a "timeout" but it was changed soon to handle "ACR Retentions." Recall that the ACR Retention problem is that of a source not using its ACR for quite some time and then suddenly jumping to use the ACR. For example, a source may transmit at 10 Mbps while its ACR is 100 Mbps. Some switch mechanisms are sensitive to this behaviour and would result in underutilization. To fix this problem, it was suggested that the ACR should be reduced to at most two times (TOF times) the actual rate. Rule 5a is triggered whenever source transmits slower than 1/TOF of its ACR. One way to avoid ACR retention would be to adjust ACR to ACR/TOF, i.e., ACR = Max(ICR, ACR/TOF). The current rule 5a does reduce the rate (mostly to ICR) and does not solve the ACR retention problem.

The calculation of TDF is not specified in the source behavior (section 5.10.4,pg 55-56), though it is specified in the pseudo code (I.1, pg. 86).

The document has been changing continuously and will probably be changing for some time. It would be very helpful to have change bars.

8. The PNI implementation in the source behavior {section 5.10.4} as well as the pseudo code {section I.1} is erraneous. The right rule to replace line 2, pg 87 : "else if NI = 0 and ACR_ok" is "else if (NI = 0 and (ACR_ok or PNI=1))"

We observe however, that the dominant effect is that of 5a.

9. We have found that rules that automatically trigger reducts (rule 5, 6, etc) are often triggered incorrectly for low rate sources. Inter-RM cell gap degenerates as the RM cell passes through the network. At low rates, when a feedback path is established, the inter RM gap determines the responsivity of the network to transients. The inter RM cell time is determined not only by the queues in the network, but also by the minimum of the forward and reverse rates. This can cause triggering of Xrm, Trm, Tcr etc.

10. In section I.5.2, pg 93, it should be noted that overload factor can also be used as a load indicator. Further, in section I.5.2.2, we note that "rate of change of queue length" is not the only load indicator. "Overload factor" is used as load indicator in ERICA.

[1] R. Jain, S. Kalyanaraman, S. Fahmy, F. Lu, " Out-of-Rate RM Cell Issues and Effect of Trm, TOF, and TCR," ATM Forum Contribution 95-973R1, August 1995

Note: All our contributions and slides are available through our web site: http://www.cse.wustl.edu/~jain/atmforum.htm