Timing Problems with Connection-Oriented Protocols on Local Area Networks

Document Number TR 29.0943

Kathleen W. Cowart

LAN Microcode
Research Triangle Park, N.C.

IBM Internal Use Only

IBM Internal Use Only

ABSTRACT

The Logical Link Control protocols used on Local Area Networks are based on protocols developed for point-to-point connections that use a response timer for error discovery. For the protocol to operate correctly the response timer must accurately reflect the maximum round trip time for frame delivery and acknowledgement. This value was easily established in a point-to-point situation. With the development of multi-access media, and of networks consisting of interconnected media with widely varying characteristics, it is becoming increasingly difficult. This paper describes some of the problems that arise when the timing requirements are not met.

This paper requires some familiarity on the part of the reader with Local Area Networks (LANs), and with Logical Link Control (LLC) protocols. Some knowledge of the operation of HDLC or SDLC type protocols would be helpful. A glossary and list of publications are provided at the end of the paper which may assist those without such knowledge.

ITIRC KEYWORDS

Logical Link Control

Local Area Networks

IEEE 802.2

Type 2 LLC

INTRODUCTION

A Local Area Network consists of communicating devices connected by one or more of several different kinds of media. The IEEE has issued standards governing the physical media, and the access protocols used to communicate on them, as well as a standard governing the type of Logical Link Control (LLC) protocols used to handle data delivery. In terms of the OSI reference model, LLC is the upper half of layer 2 (Data Link Layer), the Media Access Control (MAC) protocols form the lower half. The LLC standard, IEEE 802.2 (now ISO 8802/2), describes three types of service:

- Type 1 - Connectionless Service. There is no guarantee of frame delivery, and no retry at layer 2.
- Type 2 - Connection Oriented Service. This service provides guaranteed delivery of sequentially numbered frames, with error recovery, after a connection is established between the communicating stations.
- Type 3 - Acknowledged Connectionless. This service provides guaranteed delivery and retry on a frame-by-frame basis, without the need to establish a connection.

This paper is concerned with the operation of the Type 2 service, which is based on earlier data link control procedures such as High-Level Data Link Control (HDLC) and Synchronous Data Link Control (SDLC).

The IBM Token-Ring Network is one type of Local Area Network. It conforms to the IEEE 802.5 standard at the MAC layer, and to IEEE 802.2 at the LLC layer. The state tables published in the IBM Token-Ring Network Architecture Reference Manual are not identical to those in IEEE 802.2, but are essentially the same. The information in this paper is based on experience with the IBM Token-Ring, but there is no reason to assume that a LAN using IEEE 802.3 CSMA/CD protocols, for example, would escape these problems.

CONNECTION ORIENTED PROTOCOLS

This paper addresses the operation of Connection Oriented protocols at the LLC sublayer (hereinafter referred to as Type 2 protocols). Connection Oriented protocols may also be implemented at higher layers of the protocol stack, such as the Transport Layer, either instead of, or as well as, at the LLC layer. Higher-layer implementations may less often encounter the problems described, principally due to the longer timer values used, but they are not immune.

There are two significant differences between Type 1 and Type 2 protocols:

- Connection establishment
- Guaranteed delivery

A Type 1 frame is simply sent to a destination address, with no prior determination (at the LLC layer) as to whether the target exists, and with no subsequent check to see whether the frame was received and accepted. Before data frames are sent using Type 2 protocols the two parties involved exchange frames confirming that they exist, and that they are willing to communicate. All data frames sent after the connection is established carry sequence numbers, and the sender expects to receive confirmation of receipt from the target. If confirmation is not received after a certain period of time, the sender will start recovery procedures (enter the link state known as Checkpointing in the IBM state tables, Await in the IEEE state tables). Recovery involves sending a command-poll frame to which the receiver is obligated to reply with a response-final frame. The command-poll frame is also timed, and is resent a certain number of times if no reply is received. If no reply is received after the retry count is exhausted, the connection has been lost: if a reply is received the sender determines which data frames (if any) need to be resent.

NON-DETERMINISTIC MEDIA

The basic assumption of the Type 2 protocols regarding the recovery timer is that when it expires the frame being timed has not been, and will not be, received by the target. Therefore the sender will never see a response to that frame. The response timer (T1 in the IBM state tables), should therefore be set to expire after the time required for all of the following events to occur:

1. the sending station transmits the frame

2. the frame traverses the media to the target

3. the target station receives and (to some extent) processes the frame

4. the target station prepares and transmits the reply

5. the reply traverses the media to the sender

6. the sender processes the reply

For the timer to be set correctly, the maximum duration of all of the above events must be determined. However, in all cases there are problems doing so.

Transmitting the frame

The simplest case of unbounded delay in the transmit process is the frame queue in the transmitting station. If multiple connections are supported, or Type 1 traffic is multiplexed with Type 2, there is no good reason to assume that issuing a transmit request to the MAC layer for a particular frame immediately results in an attempt by the MAC layer to transmit that frame. The length of the transmit queue is clearly variable.

The fact that a station is ready to transmit a frame (it reached the head of the transmit queue) does not mean that it is able to do so. It first must acquire the right to transmit on the media to which it is attached. Depending on the LAN this could mean that it receives a token (of the correct type and priority), or that it determines that no other station is currently transmitting. Some media (e.g. IEEE 802.3) are inherently non-deterministic in this respect. Others, (e.g. IEEE 802.5) are inherently deterministic for a given configuration, provided that all stations are using the same priority for all frames.

In addition, the media may be temporarily unavailable for data transfer regardless of how deterministic it may be in normal operation. The MAC protocols on the Token-Ring include elaborate error recovery procedures to restore normal operation when a problem occurs, but these take time to operate. There are also MAC protocols which allow for exceptional data transfer - the restricted token in FDDI can block all other asynchronous data traffic for an unlimited period of time.

There are two ways to alleviate these problems. One is to attempt to detect cases where there is significant delay in accessing the media, and either stop or extend the T1 timer accordingly. The T1 timer may be stopped for the duration of Beaconing on the Token-Ring, for example. It could also, conceivably, be stopped while restricted tokens were circulating on an FDDI ring. This solution only works, however, where both stations are on the same media segment - it does not work where the protocols are operated through a bridge or relay. If a higher-layer protocol is also timing frame transfers, the problem is propagated upwards.

The other alternative is to wait to start T1 until the MAC layer confirms that it has transmitted the frame. However, this results in significant complications to the LLC code, not covered by the state tables. By the time the MAC layer confirms transmission some other event - frame reception or timer expiration - may have caused a state change which renders the Start_T1 action unnecessary, or even harmful.

Traversing the media

Once a station has gained the right to transmit on the media, it would seem that the transit time is governed only by the media speed, and the distance to the target station. Unfortunately this is only true if the source and destination stations are both on the same media segment. If all media segments are of the same type, then it would appear that the transit time is governed by the distance between the source station and the first bridge/relay, the distance between each of the intermediate bridges, and the distance between the last bridge and the destination station, plus the delay in each bridge/relay. In a source routing network the source station will be able to determine the number of bridges involved by examining the routing information. However, each bridge has the same problems with respect to gaining access to the medium that the source station has.

There are also situations where the source station can make no determination as to the length and nature of the route to the target station - a network that uses transparent bridging for all or part of its connections deliberately isolates the end stations from knowledge of the route. The use of bridges to different media types in a source routing network can also introduce delays that are not known to the source station.

Receiving the frame

Once the frame arrives at the target station, there is no guarantee that it will be processed and acknowledged immediately by the LLC layer. Unless there is only one active link the frame will be queued behind frames previously received for other links, and it may also be queued behind a previously received frame for the same link. How soon it is processed may also depend on the implementation - if the LLC protocols are executed on a non-dedicated processor, for example, the preparation of the response frame may be delayed by the execution of another task, unrelated to the LAN communication.

Items 4-6 all encounter potential delays of the same nature as items 1-3.

SOME SPECIFIC PROBLEMS

Given that T1 cannot be set to an accurate value in all cases, there will be occasions when it will expire, and a response to the timed frame will still be received by the sender. Since a basic premise of the protocol is that this will never happen, there are obviously problems when it does happen. This section will describe three specific scenarios that have been observed and will address some possible ways around the problems. The various figures illustrate frame flows between stations, and use the following conventions:

- Station A is the transmitter and Station B the receiver
- The first field in a frame description defines the frame type and command code: I for I-format frames (data frames), RR, RNR, RJ, for supervisory frames, and FRMR for Frame Reject.
- The second field in a frame description indicates whether the frame was a command (C) or response (R) and the state of the poll (P or NP) or final (F or NF) bit.
- The remaining fields in I-format frames represent the N(r) and N(s) sequence numbers (in hexadecimal).
- The remaining field in supervisory frames represents the N(r) number.

The Unexpected Response-Final Problem

The simplest case is that of the unexpected, or spurious, response-final frame. Consider the following frame flows, first that seen by the receiver:

```
        Station A                         Station B

        RR CP 46     ------------->

                     <-------------     RR RF 3C

        RR CP 46     ------------->

                     <-------------     RR RF 3C
```

Figure 1. Receiver's frame flow for the response-final problem

Station B sees nothing unusual - it processed two RR command-poll frames. However, Station A sent the second RR command-poll frame because its T1 expired on the first. The flow it sees is in the second figure.

```
   Station A                Station B

     RR CP 46  ------------->
```

```
RR CP 46  ------------->
     <------------- RR RF 3C
     <------------- RR RF 3C
```

Figure 2. Transmitter's frame flow for the response-final problem

The second response-final frame, according to the initial version of the LAN FAP state tables, is a protocol violation - the final bit is unexpected as there is no unanswered poll bit outstanding. This should cause a Frame Reject (FRMR) to be sent by Station A, resulting in either termination of the connection , or an attempt to reset it. Even in the best case (reset), data transfer will be suspended for a considerable period.

The later versions of the LAN FAP state tables were updated to allow the second response-final to be accepted by Station A. If a frame is received in any of the information-transfer states with the final bit set, and no poll bit is outstanding, the final bit is simply ignored. The frame is treated as if it had been a response-not-final. This allows the link to remain up during a transient timing problem, such as beaconing on an intermediate ring.

The Frame Reject Problem

This problem was observed when running with substantial delays in the receiving station. Both stations operate according to the protocol, but the frame flow they observe is different. The transmitter believes that a checkpointing frame has been lost, while the receiver in fact receives and procoesses the frame. From that point on the stations see the frame flow differently.

```
Station A                          Station B          V(r)

RR CP   46     ------------->                         3C

               <-------------     RR RF   3C

RR CP   46     ------------->

               <-------------     RR RF   3C

I   CP   3C 46 ------------->                         3D

               <-------------     RR RF 3D

I   CP   3C 46 ------------->

               <------------      RJ RF   3D

I   CNP 3D 46  ------------->                         3E

               <------------      RR RNF 3E

FRMR RNF       ------------>
```

Figure 3. Receiver's frame flow for the FRMR problem

Station B has every right to be puzzled by the FRMR it has just received. It has answered incoming command-poll frames with response-final frames, it has acknowledged all I frames on receipt, and it has sent a Reject to a duplicate I frame - all according to the protocol. However, the view from Station A is a little different...

```
V(s)      Station A        Station B
3C        RR CP   46       ------------->
          RR CP   46       ------------->
                           <-------------       RR RF   3C
3D        I  CP   3C 46    ------------->
3C                         <-------------       RR RF   3C
3D        I  CP   3C 46    ------------->
                           <-------------       RR RF   3D
3E        I  CNP  3D 46    ------------->
3D                         <------------        RJ RF   3D
                           <------------        RR RNF  3E
          FRMR RNF         ------------>
```

Figure 4. Transmitter's frame flow for the FRMR problem

When Station A received RJ RF 3D it prepared to resend I frame 3D, and reset its sequence numbers accordingly. Thus, when frame RR RF 3E, the acknowledgement for I frame 3D, arrives Station A has forgotten that 3D was sent, and treats the acknowledgement of the unsent frame as a protocol violation, which causes the FRMR RNF to be sent.

The problem is caused by the fact that Station A sent two command-poll frames at the beginning of the sequence, and both were received by Station B. Since the definition of the response timer is that once it expires the frame being timed will never be seen by the intended receiver, Station A forgets that it ever sent the first frame in the sequence. The response-final frame that is actually sent in response to the first frame is treated by Station A as satisfying the outstanding poll bit in frame number two. From that point on the two stations are one frame out of synchronization, eventually leading to the FRMR transmission.

There are (at least) two possible changes that can be made to the protocols to handle this problem. The problem is partly caused by the fact that the transmitter is setting the poll bit on in the I frames. IEEE 802.2 permits setting the poll bit on at any time: the LAN FAP recommends setting the poll bit on in certain I frames sent when running the dynamic window algorithm used to reduce bridge congestion (see the IBM Token-Ring Architecture Reference Manual, pp 11-29ff). Since the dynamic window algorithm temporarily reduces the transmit window below the user-defined value, the poll bit is used to solicit acknowledgement from the receiving station, whose receive window (MAXIN) value is unchanged. When the dynamic window algorithm is first invoked the transmit window is set to one, therefore the first frame sent in the scenario under discussion is sent with the poll bit set. If the poll bit had not been set, the second RR RF 3C frame would have had no effect on Station A - the final bit would have been ignored as no poll bit was outstanding, and the N(r) value would not have caused retransmission.

One solution, therefore - provided unexpected response-final frames are accepted -

is to always send I frames command-not-poll.

An alternate fix is to recognize the fact that the RR RF 3E frame is not truly invalid - I frame 3D was in fact sent. This can be accomplished by using a new state variable, Saved_V(s), which holds the highest value used as V(s). If an incoming frame fails the check for valid N(r) value, the check is redone using Saved_V(s). If the second check is successful, the N(r) field in the frame is replaced with the last 'correct' N(r) value. In addition, the final bit is ignored. Thus the frame does not cause a FRMR, but neither does it affect the state of the protocol.

In the scenario above the frame RR RF 3E is thus converted to RR RNF 3D and is effectively ignored by Station A which goes ahead and completes retransmission of I frame 3D instead of sending the FRMR. Station B will then respond with RJ RF 3E, which Station A treats as an acknowledgement of I frame 3D.

The Link Lost Problem

The FRMR problem described in the previous section was observed when the receiving station was heavily loaded. The time required for frame delivery is also effected by the path between the communicating stations. With the development of Local Area Networks it becomes more and more likely that the stations are on different media segments, and indeed that the segments are not necessarily of the same media type. This clearly results in changes in the time required for frame delivery and acknowledgement, and can lead to a worsening of the FRMR scenario in which the fix described above becomes inadequate. In this case the FRMR is avoided by the fix, but eventually the link is lost due to the failure of the transmitting station to exit checkpointing.

The 'Frame Numbers' shown in the following figures refer to the sequence of frames on the ring in which Station B is inserted. All frames relating to the transmission of I frames by Station B have been omitted for clarity.

Frame No.	Station A		Station B
01	I CNP 4 45	------------->	
02		<-------------	RR RNF 5
03	RR CP 45	------------->	
04		<-------------	RR RF 5
05	RR CP 45	------------->	
06		<-------------	RR RF 5
07	I CNP 5 45	------------->	
08		<-------------	RR RNF 6
09	RR CP 45	------------->	
10		<-------------	RR RF 6
11	I CP 5 45	------------->	
12		<-------------	RJ RF 6
13	I CNP 6 46	------------->	
14		<-------------	RR RNF 7
15	I CNP 7 46	------------->	
16		<------------	RR RNF 8
17	I CP 6	------------->	
18		<------------	RJ RF 8

```
19      RR CP 46          ------------->
20                        <------------        RR RF 8
```

Figure 5. Receiver's frame flow for the link lost problem

Station B answers each frame as it is received, however receipt of the responses by Station A, which is inserted into a different ring, is delayed. The repetition of the checkpointing frame sent after I CNP 4 45 (frames 3 and 5) is an indication that there is a problem.

Frame No.	Station A		Station B
01	I CNP 4 45	------------->	
03	RR CP 45	------------->	
02		<-------------	RR RNF 5
05	RR CP 45	------------->	
04		<-------------	RR RF 5
07	I CNP 5 45	------------->	
09	RR CP 45	------------->	
06		<-------------	RR RF 5
11	I CP 5 45	------------->	
08		<-------------	RR RNF 6
10		<-------------	RR RF 6
13	I CNP 6 46	------------->	
15	I CNP 7 46	------------->	
12		<------------	RJ RF 6
17	I CP 6	------------->	
19	RR CP 46	------------->	
14		<------------	RR RNF 7
16		<------------	RR RNF 8
18		<------------	RJ RF 8
21	RR CP 46	------------->	
20		<------------	RR RF 8

Figure 6. Transmitter's frame flow for the link lost problem

When Station A receives frame number 12 it resets its send sequence number, and accepts the following frames from Station B (numbers 14, 16, 18, and 20) only on the basis of the protocol change described in the previous section. However, as described, the final bit in these frames is ignored, and Station A is unable to exit checkpointing, as this requires a response-final frame in answer to the command-poll. Eventually Station A exceeds its retry count and declares that the link has been lost.

Of course, the obvious solution is to extend the fix and recognize the final bit in the fixed frame. However, this seems to be stretching the protocol too far.

ACKNOWLEDGED CONNECTIONLESS PROTOCOLS

The most significant differences between Type 2 and Type 3 protocols are:

- Connection establishment
- Number of outstanding unacknowledged frames

Type 3 protocols do not require that a connection be established. However, a send state variable is maintained for a given combination of Destination Address, Source SAP, and priority. Only one frame can be outstanding (awaiting acknowledgement) for the combination at any one time, and the receiver is obligated to send a response frame. A response timer and retry count are associated with the frame, and the higher layer is informed if both expire without acknowledgement. However, since the next frame for the combination will be sent with the same sequence number as an unacknowledged frame, it is possible that a delayed response to the first frame would be thought to acknowledge the second frame.

The Type 3 protocol definition specifically states:

> If sequential delivery of Type 3 PDUs is required, the data link user must not queue a Type 3 request to LLC for transmission between a given SSAP and remote station ... at a given priority if a previous such request has not yet been confirmed by LLC. This restriction is necessary to allow higher layers to perform recovery operations before resuming normal data transmission in case LLC is unsuccessful in transmitting a PDU (after retries).

This puts the responsibility for maintaining correct frame sequences on the higher layer.

POSSIBLE SOLUTIONS

One solution to the problems discussed above is to keep setting T1 to higher and higher values until no problems are observed. This can result in a very long time being taken to detect that a frame really has been lost, and to retransmit it.

The next least disruptive procedure is to 'fix' the protocol. However, this can reach the point at which there seems little point in running the protocol at all.

Probably the most effective change would be to alter the protocol so that supervisory frames also carried sequence numbers: an acknowledgement could then be unambiguously associated with a particular command-poll frame. This is the method used by the TCP layer of TCP/IP. Since this would require a major change to an international standard, and obsolete existing code, it is effectively a non-starter.

Another option is to abandon the Type 2 protocols and run Type 1 link layer protocols with frame sequencing and error recovery at a higher layer. As with the first suggestion, however, this will result in significantly delaying recovery from a truly lost frame. It will also eliminate one method of pacing - the use of Receiver-Not-Ready frames at the LLC layer. Depending on the higher-layer protocol used, it could also result in simply moving the problem upwards.

A better option, where delays are caused by the use of bridges and dissimilar media, is to terminate the LLC protocols at an intermediate point, rather than running them end-to-end. Just how much of the layer 3 protocols must be run in such an intermediate box is arguable, and may be governed by the lower-level routing technique in use. In an all source-routing network all that may be needed is some relay code, producing a box that has been called an 'LLC switch'. Where translation between source-routing and transparent bridging is required, a layer 3 router could be used. Use of a layer 3 router could also be considered where Type 2 protocols are used on one media type which is interconnected to another using only Type 1 protocols.

CONCLUSION

The LLC Type 2 protocols were originally designed for a situation where frame delivery and response time could be determined with reasonable accuracy. It was indeed valid to assume that when a certain period of time had elapsed, no response to a command-poll frame would arrive, because the frame would never be delivered. The world of LANs is not that deterministic - even where the MAC protocols appear to be deterministic, there are other considerations. Trying to set the response timer to a value large enough to take into account all possible delays can result in a value too large to be useful for error recovery cases.

While adjusting the T1 value may be adequate for problems encountered between nodes on the same media segment, more complex solutions are required in more complex networks. The design and implementation of boxes that terminate Type 2 protocols at an intermediate point is becoming an important consideration as local area networks increase in size and complexity.

BIBLIOGRAPHY

IBM Token-Ring Network Architecture Reference, SC30-3374-01, August 1987

ISO/DIS 8802/2, Logical Link Control Standard for Local area networks, 1987-01-14

ISO/DIS 8802/3, CSMA/CD ANSI/IEEE Std. 802.3-1985

ISO/DIS 8802/4, Token-Passing Bus Access Method IEEE Std. 802.4-1985

ISO/DP 8802/5, Token-Ring Access Method IEEE Std. 802.5-1985

FDDI Token-Ring Media Access Control, X3T9/84-100

GLOSSARY

Bridge . A device that links networks that use the same logical link protocols (i.e. a MAC-level relay).

Carrier Sense Multiple Access with Collision Detect (CSMA/CD) . Technique used on IEEE 802.3 conformant LANs to allocate bandwidth. A station transmits if it detects no other transmissions, and ceases to transmit if it detects a collision.

Checkpointing . a link state in which one station is attempting to determine whether its partner is still communicating with it.

Command-poll . a frame sent between two stations to which the recipient must respond with a response-final frame. The bits in the LLC control field indicating command and poll are both set appropriately.

Final bit . A bit in a defined location in the LLC control field which is set in a response frame to match an incoming poll bit.

Frame Reject - FRMR . A frame sent to indicate that a protocol violation has been detected by the sending station and that data transfer has ceased. The protocol violation may be one or more of a number of incorrect characteristics in an in-bound frame, for example an unrecognized command code.

N(s) - Transmitter Send Sequence Number . Sequence number of an I-format frame. Set from V(s), which is then incremented.

N(r) - Transmitter Receive Sequence Number . Sequence number of the next expected I-format frame. Acknowledges all frames through N(r)-1. Set from V(r), which is set from the N(s) of the last in-sequence I-format frame.

Poll bit . A bit in a defined location in the LLC header of a command frame which is set on to solicit the transmission of a response-final frame.

Receiver Not Ready (RNR) . A supervisory frame indicating that the sender can temporarily not receive data frames. It may be a command-poll or not-poll, or a response-final or not-final frame. It is used to halt data transmission while congestion is cleared.

Receiver Ready (RR) . A supervisory frame indicating that the sender can receive data frames. It may be a command-poll or not-poll, or a response-final or not-final frame.

Reject (RJ or REJ) . A frame sent when a data frame arrives with too high a sequence number - indicating that an intermediate data frame has been lost and must be resent.

Response-final (RF) . The response to a command-poll frame.

V(s) - Send State Variable . The sequence number of the next in-sequence I-format frame to be transmitted on the associated link.

V(r) - Receive State Variable. The sequence number of the next in-sequence I-format frame expected to be received on the associeated link.