digital™

Networks Marketing

Link State Routing Timers

Factors affecting the reconfiguration timers of routers using Link State routing

Document Number: NIPG-006-01

December, 1993

Abstract

This paper sets out the factors which affect the time taken for a router network to reconfigure following some disruption such as a router or link failure. The specific examples describe the detailed operation of the IS-IS protocol (ISO/IEC 10589), however, the general principles apply equally to any Link State Routing protocol, such as OSPF (RFC 1247). The failure detection principles also apply to Routing Vector protocols.

The information in this document is based on a paper written by Mike Shand, a senior routing architect in the Networks Interconnect Product Group within Networks Engineering.

Introduction

Routing timers affect the time taken for a router network to reconfigure following some disruption such as a router or link failure. Reconfiguration requires a number of operations to be completed, each of which has its own time constraints. At the highest level these are:

- Detection of failure
- Generation of new local Link State Packets (LSPs)
- Propagation of LSPs throughout the network
- Running of the decision process and generating new forwarding tables

Detection of Failure

The failure of a link or an adjacent router can be detected either by the underlying datalink layer or by the network layer. In some cases it is possible to detect the failure at the physical layer.

Detection at the Data Link Layer

Connection oriented datalinks, such as HDLC, X.25 or DDCMP operate an protocol which requires each data transmission to be acknowledged. When an acknowledgment has not been received for a certain period of time (commonly known as T1), the data is retransmitted. If N consecutive transmissions are unacknowledged, the data link connection is considered to have failed and the information is signaled to the user of the service. The network layer can use this information to indicate that the adjacency has failed.

Thus failure is detected after a period of N*T1 has expired since the last attempt to transmit data. Typical values for N and T1 are 10 and 3 seconds respectively, giving a detection time of 30 seconds. However, it is possible to reduce these values.

In some cases it is possible to detect failure as a result of the loss of modem signals, which may give a more rapid failure indication, but there are many failure conditions which would not cause a physical level error to be indicated.

Datagram datalinks

Connectionless datalinks, such as CSMA CD and FDDI LANs, do not provide any capability to detect failure of the datalink. In some cases it may be possible to detect that the local connection to the LAN has been broken, for example by removal of the adapter cable.

Detection by the network layer

In all cases, an adjacency to a neighboring system is maintained by means of hello messages. These are transmitted every T seconds and contain a holding time for which they are valid. If another hello has not been received before the holding time elapses, the adjacency is considered to have failed. The holding time is normally expressed as a multiplier of the hello time (m). Thus the holding time is mT seconds.

The hello time and multiplier vary according to the the type of datalink and in the case of a LAN, whether or not the router is the designated router. The holding multipliers for point to point subnetworks and LANs are presets (holding Multiplier1 and ISIS holding Multiplier) with default values of 3 and 10 respectively. The default values are summarized in the following table.

Circuit Type	System	Timer Name	Default Value	Holding Multiplier	Default Time	Minimum Holding Time
Point to point	Any router	Hello Timer	10 s	3 s	30 s	3 s
LAN	Non designated router	IS-IS Hello Timer	3 s	10 s	30 s	10 s
LAN		DR IS-IS Hello	1 s	10 s	10 s	10 s

In V2.3 of the DECNIS multiprotocol router software it is possible to change the value of the Holding Multiplier preset in the range [2..63].

Generation of new local LSPs

When the adjacency failure has been detected it is necessary to generate a new local LSP (or LSPs) to reflect that information. There is a hold down (Minimum LSP Generation Interval2, which is a preset with a default value of 30 seconds) on the generation of new LSPs, which means that if that particular LSP had recently been regenerated (for example as a result of a periodic generation, or

because some other adjacency change had necessitated it), it would not be possible for it to be regenerated until the timer had expired.

For most failures this would normally not be a problem unless it happened to coincide with an unrelated event, but in cases where a new adjacency is deliberately enabled on failure of a previous one (for example by some intelligent management application), the good news of the new link would always be delayed by the previously issued link failure information, unless the new adjacency happened to be reported in a different LSP.

Propagation of LSPs

So that other routers can regenerate their routing tables new LSP information must be propagated throughout the entire routing domain. In the case of level 1 routing information, this is throughout the entire area; in the case of level 2, this is throughout the entire level 2 routing domain. This propagation is accomplished on a hop by hop basis by the Update Process.

The total propagation delay is equal to the sum of the propagation delays across each hop. In many cases however, where the disruption to the network is localized, the connectivity of the network will be restored when some subset of the routers have received and processed the new information (although it is still necessary for the complete propagation to be acheived eventually), and therefore the reconfiguration time may be less than the maximum domain wide propagation delay.

Point to Point Subnetworks

Over point to point subnetworks, such as HDLC, DDCMP, PPP and X.25, the update process operates an acknowledged protocol. When an LSP is either generated or received from another system and which needs to be propagated over a particular link, the LSP is transmitted and a flag is set on that LSP. The flag is checked every minimum LSP Transmission Intervall, a preset with a default value of 5 seconds. If no acknowledgment has been received in this time, the flag is cleared and the LSP will be retransmitted. This process continues until an acknowledgment has been received. Thus on an error prone link, the LSP transmission may be delayed a multiple of minimum LSP Transmission Interval.

In the DECNIS software V2.3, it is possible to change the value. In addition, even when the LSP is `transmitted' by the network layer, some delay may be introduced by the data link layer. The LSP may have to wait for the transmission of a previous data or control packet to complete. On a slow link such as a 64K

link, a 1500 byte packet will take up to 200 ms to transmit. Assuming that the control traffic is given priority over data traffic, which is the case with the DECNIS router, there may still be other control packets such as IIH PDUs, other LSPs or SNPs in the queue. Each of these will take up to 200 ms to transmit. If the link is subject to errors, then further delay can be introduced while waiting for the acknowledge timer and then retransmitting the packet.

In addition to the above, the standard requires that the maximum propagation delay through the router from a packet being received by to being marked for transmission is one second. Most routers easily meet this requirement.

Broadcast subnetworks

Over a LAN, such as Ethernet or FDDI, the update process operates an unacknowledged protocol. When an LSP is identified for transmission over a particular link, a flag (SRMflag) is set on the LSP, but it is not transmitted immediately. A process runs every broadcast LSP Transmission Interval, a preset with a default value of 33 milliseconds, which chooses at random an LSP with SRMflag set and multicasts it on the LAN to the appropriate multicast address depending on the level of the LSP.

A delay is introduced to limit the rate at which LSPs can be transmitted when a topology change causes a large number to be marked for transmission. Since the broadcast medium cannot provide back pressure, failure to take this precaution could result in the receivers being over-run. The value of 33 milliseconds represents the maximum rate at which a typical router can receive LSPs. While the DECNIS router is able to receive at a rate faster than this, it is best not to decrease this value.

Note that while this mechanism reduces the probability of over-run it is still possible for this to occur, especially when a large number of routers on the LAN are simultaneously attempting to transmit multiple LSPs. For router implementations which are unable to process the maximum theoretical rate of packet reception and, if necessary, discard data packets in preference to control packets, the presence of a high rate of data packets will also have an adverse effect. This is actually particularly likely during a period of massive topological change, since temporary routing loops can arise, thus potentially multiplying the data rate by a factor of half the maximum lifetime field (say 64). This is not a problem on the DECNIS implementation, since it can analyse packets at the maximum reception rate.

The algorithm used is actually a slight optimization of the above. If there are LSPs waiting on the receive queue of a router circuit , the transmission of LSPs over that circuit is limited to 1 per second. This has the effect of speeding up the convergence time of the algorithm since it reduces the probability of multiple routers transmitting duplicate copies of the same LSP. This is because the router is more likely to detect that a particular LSP has already been transmitted before transmitting it itself.

Detection of lost update LSPs for whatever reason is achieved by the periodic transmission by the Designated Router (DR) of a Complete Sequence Numbers Packet (CSNP). This lists all the LSPs which the DR has in its database. If another router on the LAN has an LSP not mentioned by the DR, it multicasts it on the LAN. If it does not have an LSP which the DR claims to have, it multicasts a Partial Sequence Numbers Packet (PSNP), to which the DR responds by multicasting the LSP requested.

Thus, the periodic CSNP multicast serves to detect and correct any mis-synchronization of the LSP databases. The CSNP is transmitted every complete SNP interval (a preset with a default value of 10) seconds. When the receipt of a CSNP requires the transmission of a PSNP, a flag (SSNflag) is set which indicates that this LSP (header) should be included in the next PSNP to be multicast. PSNPs are multicast every Partial SNP Interval, a preset with a default value of 2 seconds, if required. The worst case delay for a single recovery cycle is 12 seconds using default values, plus any transmission and processing delays.

The size of CSNPs is a significant factor in determining the minimum rate at which they can be transmitted. A single maximum sized CSNP (1492 octets) will contain information about 91 LSPs. For a large network with say 500 LSPs, this requires the transmission of 6 CSNPs at each transmission interval. In fact, the standard permits them to be spread over the interval rather than occurring in a burst.

Note also that in worst case conditions it may require multiple CSNP cycles to ensure synchronization of the database, since the retransmitted packets or indeed the CSNPs themselves may be lost. Analysis shows that up to 7 or 8 CSNP cycles may be required to fully synchronize a 1000 LSP database, where all 1000 LSP are required to be transmitted across the LAN.

Under heavy load conditions a further delay can be introduced by the acquisition time for the broadcast medium. For example on Ethernet, an attempted transmission which collides with another station will invoke the backoff algorithm.

Subsequent collisions will increase the backoff factor up to a maximum of 300 ms. On FDDI, a number of relatively rare conditions, such as token claim can result in transmission delays in excess of 128 seconds!

Running the Decision Process

Arrival of a new LSP triggers the running of the decision process3, which uses a modified form of Dijkstra's SPF (Shortest Path First) algorithm to compute the shortest path to each destination. The standard requires that this can be run to completion in a maximum of 5 seconds. Potentially, it will take up to 5 seconds before the information in a newly received LSP is made available to the forwarding process.

However, if the decision process is already running when the new LSP arrives it may be necessary to wait until the run in progress completes before the newly arrived LSP can be processed. In fact it is possible for a new LSP to be processed if the old copy of the LSP hasn't yet been processed by this run of the decision process, but not all implementations take advantage of this optimization. In any case, if the new LSP was close to the root of the tree (i.e. it was from a nearby system), then it is quite likely that it would have already been processed.

The time taken for the decision process is only constrained to be a maximum of 5 seconds. For simple topologies the time taken could be much less, but in order to guarantee to make progress with other task (such as forwarding in a single processor router, or processing of management commands) a typical implementation will allow the Decision process to be interrupted provided it doesn't compromise the 5 second limit.

To optimize the running of the decision process, some implementations (such as DECNIS) implement a hold down on the running of the decision process, such that it is not triggered by the first new LSP to arrive. There is a possibility that it will require running again to process subsequent LSPs. The value of this hold down is typically of the order of 5 seconds4.

Impact of Other Routing Timers

There are a number of other timers associated with the operation of the IS-IS protocol which have little or no impact on the convergence time of the network. However, for the sake of completeness, their use is described here. In fact the running of the Decision Process is triggered by the arrival, generation or deletion of an LSP.

Lifetime of LSPs

Each LSP has a lifetime of Maximum Age, a preset with a default value of 1200 seconds. This means that the LSP will persist within the LSP database of all the routers in the routing domain for that period unless it is purged for some reason. Normally the LSP will be regenerated with a new sequence number at least every Maximum LSP. In the DECNIS V2.3 software, it is possible to change the value of the Maximum LSP Generation Interval in the range [60..900] and the Maximum Age in the range [1..65535]

If the router responsible for generating the LSP crashes, or if the routing domain is partitioned such that some parts cannot be reached, the LSP lifetime will expire. At this point the variable part of the LSP is deleted leaving only the short header, and the header is flooded by the update process to synchronize the expiration of the LSP with all other copies of the same LSP. Different clock rates and propagation times can result in the LSP lifetime expiry potentially occurring at different times in different routers. The header is retained for a maximum of Zero Age Lifetime, a preset with a default value of 60 seconds, after which it is completely deleted.

Clearly the values of Maximum Age and Maximum LSP Generation Interval must be coordinated at any particular router (i.e. the former must be sufficiently larger than the latter such that there is no possibility of an LSP being erroneously timed out). There is no requirement that they should have the same values at different routers, although there is little to be gained by having them different.

The only reason to increase the values would be to reduce the background traffic generated by the periodic refreshing of the LSPs. The only reason to reduce the value would be to ensure that "stale" LSPs generated by routers which are no longer present are removed from the database as quickly as possible and hence don't take up memory. There is no need to ensure removal of "stale" LSPs to ensure correct route computation by the decision process, since LSP generated by routers to which there are no adjacencies from other routers will not be referenced by the decision process.

If router memory were no object it might be thought that a very large value for the lifetime would suffice with a corresponding saving in refresh traffic bandwidth. However, the requirement to refresh LSPs every Maximum LSP Generation Interval puts an upper bound on the time for which corrupted routing information can persist since it will be refreshed from the source within that time.

It is unwise to set this value to a value larger than the default. Zero Age Lifetime must be sufficiently long to ensure a high probability that the purge header has been propagated to all copies of the LSP within the routing domain. If it is set too short, it is possible that a copy of the complete LSP to be purged could still exist somewhere in the routing domain and would then be flooded back into all the routers, thus wasting memory until its full lifetime expires.

Setting the value too short could allow some of the old LSPs to exist for longer than necessary and hence increase the possibility of causing an LSP database memory overload. However, setting the value too large will cause the headers to be retained for longer than necessary, which will also take up additional memory (though obviously not as much as the Complete LSPs themselves). The default value of 60 seconds is a reasonable compromise, but the value is not at all critical and there is probably little to be gained by changing it.

Note that when initiating a purge of an LSP whose lifetime is not yet zero (for example because a designated router change has made the set of pseudonode LSPs superfluous) the purge header is retained for Maximum Age rather than Zero Age Lifetime since it is better to pay the memory penalty of keeping just the headers for up to 20 minutes rather than risk the possibility that a complete LSP will be retained for up to 20 minutes.

Waiting time

When an LSP arrives for which there is insufficient memory it indicates that the LSP database at the router in question is definitely unsynchronized with the topology of the network. Therefore routes computed on the basis of that information are likely to be incorrect and could give rise to routing loops of a more permanent nature than the normal transient routing loops caused by temporary mis-synchronization of the LSP databases resulting from normal propagation delays.

For this reason the router in question is put into "waiting state", and signals this fact to other routers by generating a new version of its own LSP with a bit set to indicate that no routes are to be computed through this "overloaded" router. The LSP which couldn't be stored will continue to be retransmitted to the router using the usual mechanisms until such time as the is enough memory to store it (as a result of some stale LSPs being timed out or purged).

When a router fails to store an LSP it starts a timer for Waiting Time, a preset with a default value of 60 seconds, and enters waiting state. Each subsequent failure restarts the timer, so the timer will not expire until Waiting Time seconds

have elapsed since the last failure. On expiry of the timer, the router re-enters running state and regenerates its own LSP with the bit clear to indicate that routes can once more be computed through it.

Setting the timer value too short will result in oscillations between waiting and running states. Clearly a value which was less than the retransmission time over the circuit concerned (5 or 10 seconds) would be completely unacceptable. Setting the value too high results in the recovery from waiting state taking longer than necessary. However it should be noted that any occurrence of waiting state represents either a serious network design failure or misconfiguration. There are two reasons why this could occur

- Static overload - this occurs when the size of the LSP database is greater than the design limits of the router. This can be caused by the addition of a single system which just pushes the size over the limit, but for the reasons indicated below this is very unlikely.

  The most likely cause is when a large extra portion of network has accidentally been connected into the routing domain. For example, by two LANs running separate areas are accidentally connected by a bridge, or when a filtering bridge which was keeping two portions separate is accidentally reconfigured to remove the filtering. The condition will persist until the cause is identified and steps are taken to remedy it. Unless there is an isolated router with smaller than normal memory capacity, it is likely that all the routers will enter waiting state and the entire network will be down.

- Dynamic overload - this occurs when the LSP database size is temporarily larger than the steady state size. The most common reason for this is the failure of a designated router, where the old pseudonode LSPs and new pseudonode LSPs can coexist for a period of time, normally only a few seconds since the old LSPs will be purged.

In any reasonably designed network, allowance should be made for at least 2 copies of the largest set of pseudonode LSPs in the domain to exist simultaneously. A very conservative design would require that two copies of all sets of pseudonode LSP should be capable of coexisting, but this would permit the very unlikely event of all the designated routers on all the LANs in the domain changing during the same interval. A safety margin of 25% of the LSP database size is reasonable and allows for some level of future growth without the possibility of either dynamic or single system static overload occurring. For these reasons there is little to be gained from changing Waiting Time from its default value.

Impact on Transport Connections

In most cases, a delay in the reconfiguration of the router network, during which certain destinations are unreachable, is acceptable provided that existing transport connections are not broken. Both DECnet Phase IV NSP and OSI transport operate a similar mechanism for detecting failure of transport connections. The round trip delay is constantly monitored by the operation of the transport protocol, and a weighted average round trip delay time estimate is computed. The relative weights of the existing value and a newly measured value are affected by the Delay Weight attribute.

The round trip delay time estimate is multiplied by Delay Factor and the result added to "acknowledgment time", which is passed from the remote transport implementation during connection establishment, to produce the current value of the retransmission timer. Failure to receive an acknowledgment within this interval will result in a retransmission and after retransmit threshold, a value which is specified by the Transport service user on connection establishment, or which defaults to a value provided by management in the retransmit threshold characteristic of the local NSAP entity, retransmissions the confidence variable associated with the transport connection will be set false.

The action taken depends on whether or not the "automatic disconnection" facility was specified by the user when the connection was established. If it was specified, the transport connection will be immediately broken. Otherwise the fact that the confidence variable has been set false is signaled to the application, and it is up to the application to determine whether or not to break the connection.

Most transport implementations have a minimum value of delay which is used for the acknowledge timer. Typical values of the transport parameters and network round trip delay times give transport connection lifetimes of the order of 30 to 45 seconds. Provided that the network layer reconfiguration is complete within this time, users will only experience a delay and no transport connections will be broken.

In reality, the operation of the transport protocols is more complex than this summary implies. In particular, the inactivity timer and window timer mechanisms allow the detection of a connection failure even in the absence of data traffic. Typical values for these timers allow detection of a failure that exceeds 240 seconds duration even in the absence of data traffic.

End System considerations

End System adjacencies to routers are maintained by a hello PDU mechanism similar to that used for router to router adjacencies. The attributes used are Hello Timer and Holding Multiplier, which have default values of 600 and 3 respectively, giving a holding time of 30 minutes.

This means that if an End System fails, it will take up to 30 minutes before the routers will be aware of the failure and stop sending data packets to its address. Other than the fact that during this period the routers will not return an address unreachable error and therefore any new connection attempts will have to time out rather than failing by error report, this is normally of no consequence as the End System is unreachable in any case.

However, in the case of an End System which has multiple connections to the network (for example via separate LANs), it is possible for one connection to fail while the other(s) remain operational. In this case, if the failed connection were the best route to the End System, or was one of a set of equally good routes, packets destined for the End System will continue to be forwarded by the failed route and communication will be lost or severely degraded in the case of equal cost paths.

In this case it is important for the failed adjacency to be detected as quickly as possible. To achieve this, when an End System is configured as a multi-circuit End System (for example as a DECnet Phase V cluster, which has similar properties), the value used for the hello timer is that held in Multicircuit ESHello Timer, which has a default value of 10 seconds. Hence a failure can be detected within 30 seconds (10 * Holding Multiplier).

It is possible for a similar problem to arise when the End System is not connected simultaneously to multiple circuits but when it is moved from one circuit to another. If this is done as a manual operation, then it is necessary to reduce the value of Hello Timer for the End System in question, perhaps just during the period of the move. Provided the End System is shut down graciously, rather than crashing or being unplugged, this will not be necessary since part of the shutdown procedure ensures that the holding time of the adjacency is zeroed, causing the router to close the adjacency.

However, there is one common case where an End System may appear to move from one circuit to another. This is when a bridged extended LAN partitions as a result of bridge failure or unplugging, but the two halves are still connected by routers. In this case what originally appeared as a single subnetwork will now

appear as two subnetworks and some End Systems will appear to have moved from one to the other. In order to deal with this and some similar cases, when the routers detect that the router topology on the LAN has changed, they initiate an End System poll. The holding time of any existing End System adjacencies are reduced to a minimum value (ISIS Hello Timer + PollESHelloRate) * Holding Multiplier. Using default values, this equates to 159 seconds.

IS hello PDUs are transmitted with a suggested configuration timer value of pollESHelloRate, a preset with a default value of 50 seconds. This causes the End Systems to temporarily increase the rate at which they send ES hellos, and hence the adjacencies which are still valid are maintained, but the ones which should no longer exist are allowed to time out. After transmitting Holding Multiplier (default 3) IS hellos with the short suggested configuration timer, it reverts to the normal value of default ESHello Timer, a preset with a default value of 600 seconds.

It would be impractical to operate a large LAN continuously with a short ES hello time. A large extended LAN may have up to 10,000 End Systems. When these are transmitting ES hellos every 50 seconds this represents an ES hello rate of 200 per second, whereas with the normal value of 600 seconds this is reduced by a factor of 10.

Jittering of Timers

When messages are transmitted as a result of timer expiry, there is a danger that the timers of individual systems may become synchronized. The result of this is that the traffic distribution will contain peaks. Where there are large numbers of synchronized systems this can cause overloading of both the transmision medium and the systems receiving the messages. To prevent this occuring, all periodic timers in the network layer which can cause the transmission of message have `jitter' applied to them. This is achieved by randomly choosing a timer interval between 75% and 100% of the specified value for the timer, each time the timer is set. Thus the observed intervals between generated PDUs will vary within this range.

Digital believes the information in this publication is accurate as of its publication date. Such information is subject to change without notice. Digital is not responsible for any errors.

The following are trademarks of Digital Equipment Corporation: Digital, DECnet, and DECNIS.