Contribution to the April '98 Link Aggregation Interim Meeting
Tony Jeffree
April 22, 1998

Strawman for Changes/Additions to 802.3 required in order to specify Link Aggregation

Prepared by Tony Jeffree as a contribution to the work of the Link Aggregation Study Group at its April 1998 interim meeting. **Its contents shall not be assumed to reflect any opinions other than those of the author.**

<<Comments on this document may be sent to the author:

Tony Jeffree
11A Poplar Grove
Sale
Cheshire
M33 3AX
UK
+44 161 282 3824 (Tel)
+44 161 973 6534 (Fax)
Email: tony@jeffree.co.uk

>>

<< Author's Notes

This Draft has been generated by the author as a contribution to the work of the Link Aggregation Study Group. The contents of the document reflects some of the material presented at the Seattle and Irvine meetings of the study group. However, as the authors of those presentations have had no part in the generation of this document, its contents cannot be assumed to reflect any opinion other than my own.

The document extends the previously presented material, in particular in the area of the operation of the Link Aggregation Controller and the operation of its associated protocol. However, the contents of the document must only be regarded as "Work In Progress" (and in some areas, very rough work at that!); the intent is to start to collect and consolidate ideas, and not to present a complete, coherent and polished solution.

The document is presented in the rough format of a standards draft, in order to accelerate the process of structuring the standard, and to get ideas & concepts documented in a form that is reasonably close to that needed for the final text.

Tony Jeffree
22 April 1998 >>

Contents

Figures

Tables

Strawman for Changes and Additions to 802.3 required in order to specify Link Aggregation

1.3 *Changes to* References

<<Author's Note: References to be added. These would consist of a set of changes/additions to existing 802.3 section 1.3 & Annex A.>>

1.4 *Changes to* Definitions[1]

<<Author's Note: The following Definitions to be added. These would form a set of changes/additions to 802.3 section 1.4.>>

1.4.1 Bridge Port

A point of attachment to a LAN through which a MAC Bridge transmits and receives MAC frames.

NOTE—See ISO/IEC 15802-3 7.2, 7.2.3.

1.4.2 Bridged LAN

A concatenation of individual IEEE 802 Local Area Networks interconnected by MAC Bridges.

NOTE—This is identical to the definition in ISO/IEC 15802-3.

1.4.3 End station

A system attached to a LAN that is an initial source or a final destination of MAC frames transmitted across that LAN.

<<Author's Note: The original definition was for "Host"; however, 802-speak for Host is End station.>>

1.4.4 Conversation

A set of MAC frames exchanged between a pair of end stations, where all of the MAC frames form a part of an ordered sequence, and where there exists a requirement for ordering to be maintained among the set of MAC frames exchanged. A conversation may be uni-directional (i.e., a monologue), or bi-directional (i.e., a dialogue).

There may be more than one conversation in progress between a given pair of end stations at any one time; similarly, a given end station may take part in conversations with more than one end station at any one time.

<<Author's Note: The intent of this definition is to encompass the concept of a "flow", without attempting to tie it to particular layer attributes, such as MAC addresses, IP addresses, protocols...etc. It was noted at the Seattle meeting that using "flow" as the name would have potential of confusion with "flow control", hence the change of name.>>

[1]These definitions are based on the set of definitions presented by Floyd Backes during the Seattle interim meeting, Feb 98. I have included all the definitions from that presentation that might be relevant to the standard, regardless of whether they are actually used in this particular draft. Any that are not needed can easily be excised at a later date. Some have been modified in the light of the discussion that took place in Seattle.

1.4.5 Aggregate Conversation

A set of conversations, treated as if they are all part of a single conversation. The particular set of conversations that is aggregated to form a given aggregate conversation is determined by means of a convesation aggregation rule.

NOTE—The terms Conversation and Conversation aggregation rule are defined in 1.4.4 and 1.4.6.

1.4.6 Conversation Aggregation Rule

A rule that specifies how individual conversations (1.4.4) are allocated to aggregate conversations (1.4.5).

NOTE—There are potentially many such aggregation rules; for example, a rule might specify aggregation on the basis of source/destination address hashing, VLAN ID, IP subnet, protocol type, etc. The terms Conversation and Aggregate conversation are defined in 1.4.4 and 1.4.5.

1.4.7 Link Aggregation Group

A grouping of Link Segments, of the same medium type and speed, that are treated as if they are all part of a single Link Segment. The MDIs associated with each Link Segment in a Link Aggregation Group are associated with the same pair of devices. For the purposes of this definition, a device is a MAC Bridge, an end station or a repeater.

Traffic is allocated to the individual Link Segments in a Link Aggregation Group on the basis of one or more Conversation Aggregation Rules; i.e., one or more Aggregate Conversations are associated with each Link Segment that is part of a Link Aggregation Group.

1.4.8 LAN Aggregation Group

A grouping of LANs or Bridged LANs, of the same or dissimilar medium access method, medium type or speed, that form parallel paths between a pair of connected end stations, and that are treated as if they form a single LAN between those end stations.

Traffic is allocated to the individual LANs in a LAN Aggregation Group on the basis of one or more Conversation Aggregation Rules; i.e., one or more Aggregate Conversations are associated with each LAN that is part of a LAN Aggregation Group.

<<Author's Note: This definition is probably only useful insofar as it will allow us to identify stuff that is outside the scope of this standard.>>

1.5 *Changes to* Abbreviations

<<Author's Note: Abbreviations to be added. These would form a set of changes/additions to 802.3 section 1.5.>>

91. Link Aggregation

<<Author's Note: Final chapter numbers to be determined once the project has been formally launched.>>

91.1 Overview

This supplement to ISO/IEC 8802-3 defines an optional Link Aggregation sublayer for use with CSMA/CD MACs. The sublayer allows one or more individual Link Segments to be aggregated together to form a Link

Aggregation Group (1.4.7), such that the MAC Client is able to treat the Link Aggregation Group as if it were a single Link Segment.

Figure 91-1[2] shows the positioning of the Link Aggregation sublayer in the CSMA/CD layer architecture, and the relationship of that architecture with the Data Link and Physical layers of the OSI Reference Model. The figure also shows the ability of the Link Aggregation sublayer to aggregate a number of individual Link Segments in order to present a single MAC interface to the MAC Client.

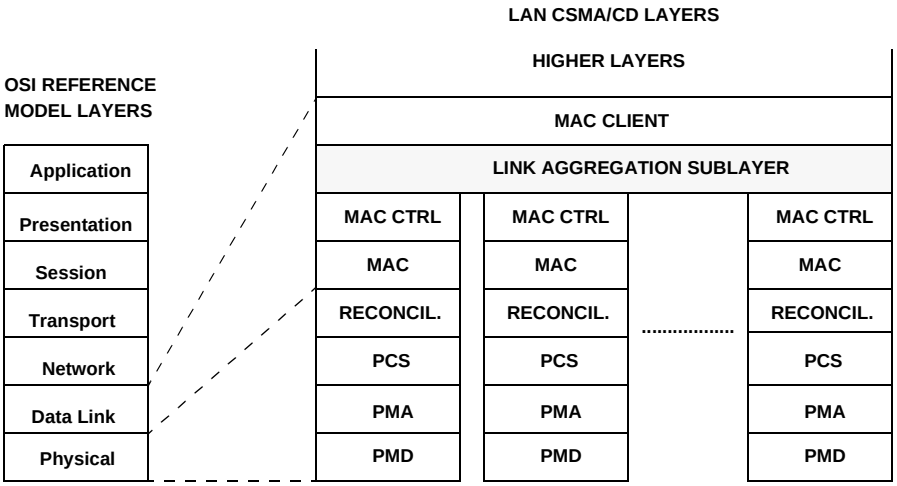

Figure 91-1—Link Aggregation Reference Model

Figure 91-2[3] shows the individual components that form the Link Aggregation Sublayer, and their interrelationships.

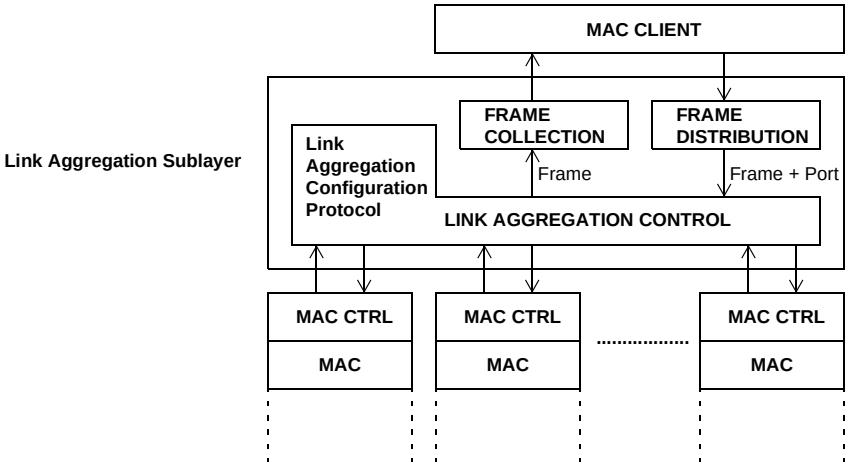

Figure 91-2—Link Aggregation Sublayer

[2]Figure 91-1 is based on Paul Bottorff's architectural diagram as presented at the Irvine meeting, simplified as agreed in the meeting. Other material from that presentation has been used as the basis for the subsequent description of the components of the sublayer.

[3]Figure 91-2 is based on the internal structure of the Link Aggregation Sublayer shown in Paul Congdon's presentation at the Irvine meeting. Other parts of that presentation have been used as the basis for the subsequent description of the components of the sublayer.

91.1.1 Frame Collection

Frame Collection is responsible for receiving incoming frames from the set of individual Link Segments that form the Link Aggregation Group with which the collection function is associated. Frames received are delivered to the MAC Client. Frames received from a given Link Segment are delivered to the MAC Client in the order that they are received by Frame Collection. As the Frame Distribution function is responsible for meeting any frame ordering constraints, there is no requirement for Frame Collection to perform any re-ordering of received frames across multiple Link Segments.

A detailed description of Frame Collection can be found in clause 91.6.

91.1.2 Frame Distribution

Frame Distribution is responsible for receiving outgoing frames from the MAC Client, and for transmitting them on the set of Link Segments that form the Link Aggregation Group with which the distribution function is associated. The distribution function is responsible for making all decisions related to load balancing among the Link Segments in the Link Aggregation Group. This supplement to ISO/IEC 8802-3 does not standardize the details of any load balancing algorithms that may be used to perform this function; however, any load balancing algorithm is required to ensure that:

a) The algorithm does not cause re-ordering of frames that are part of any given *conversation* (1.4.4);
b) The algorithm does not cause duplication of frames.

The former condition is met by ensuring that all frames that are part of a given conversation are transmitted on a single Link Segment, in the order that they are received from the MAC Client. Hence, the requirement not to misorder frames does not involve the addition of (or modification of) any information to the MAC frame, or any processing on the part of the corresponding collection function in order to re-order frames. This approach to the operation of the distribution function permits a wide variety of distribution and load balancing algorithms to be used, while also ensuring interoperability between devices that adopt differing algorithms.

A detailed description of Frame Distribution can be found in clause 91.7.

91.1.3 Link Aggregation Control

Link Aggregation Control is responsible for performing the configuration and control functions of the Link Aggregation Sublayer. These functions are performed on the basis of:

a) Static configuration information, local to the control function;
b) Dynamic configuration information, acquired and exchanged by means of the Link Aggregation Configuration Protocol.

Link Aggregation Control ensures that configuration (and re-configuration) of Link Aggregation Groups occurs automatically, within any constraints imposed by the static configuration information. In particular, it ensures that:

c) The configuration achieved is deterministic; that is to say, for a given static configuration and physical topology, the allocation of Link Segments to Link Aggregation Groups does not depend upon the order in which those segments are activated;
d) If a given Link Segment can be included in a given Link Aggregation Group, then it is included in that aggregation;
e) A given Link Segment cannot be included in more than one Link Aggregation Group at any one time.

A detailed description of Link Aggregation Control can be found in clause 92.3.

91.1.4 Link Aggregation Multiplexer

A Link Aggregation Multiplexer, or Mux, consists of an instance of the Frame Collection function and an instance of the Frame Distribution function. A single Link Aggregation Multiplexer is associated with each Link Aggregation Group. A Link Aggregation Multiplexer offers a MAC service to its associated MAC Client.

A detailed description of Link Aggregation Multiplexer can be found in clause 91.8.

91.1.5 Addressing

Associated with each Link Aggregation Multiplexer is a single, individual MAC address.

<<Author's Note: As observed in the Irvine minutes, there is much discussion to be had on addressing, the implications with respect to re-configuration, and the implications with respect to Bridge operation, before we're done.>>

A detailed description of Addressing can be found in clause 91.9.

91.2 Scope[4]

The purpose of this supplement to ISO/IEC 8802-3 is to increase link availability and bandwidth between DTEs by specifying the necessary mechanisms for parallel Link Segment aggregation. To this end, it specifies the establishment of DTE to DTE logical links, which consist of N parallel instances of an 802.3 Link Segment, all of which are full duplex point-to-point links of the same speed. A logical link so established will support existing ISO/IEC 8802.3 MAC Clients.

In particular, the following are specified:

a) The architectural model that establishes the relationship between Link Aggregation and the existing ISO/IEC 8802-3 architecture and standards;
b) The procedures involved in the establishment, configuration and removal of logical links;
c) The management functionality provided in order to allow static configuration of logical links;
d) The protocols required in order to allow dynamic configuration of logical links.

91.3 Conformance

91.3.1 Static conformance requirements

<<Author's Note: Static conformance requirements to be added.>>

91.3.2 Options

<<Author's Note: Options to be added.>>

91.4 Recommendations

<<Author's Note: To be added, if needed.>>

[4]This Scope is based on the text contained in Scope and Purpose in the proposed Link Aggregation PAR.

91.5 Relationship of Link Aggregation to other standards

<<Author's Note: To be added.>>

91.6 Frame Collection

<<Author's Note: Detailed description/definition of Frame Collection to be added.>>

91.7 Frame Distribution

<<Author's Note: Detailed description/definition of Frame Distribution to be added.>>

91.8 Link Aggregation Multiplexer

<<Author's Note: Detailed description/definition of Mux to be added.>>

91.9 Addressing

<<Author's Note: Detailed description/definition of Addressing to be added.>>

91.10 Protocol Implementation Conformance Statement

The supplier of an implementation that is claimed to conform to clause 91 of this standard shall complete a copy of the PICS proforma provided below and shall provide the information necessary to identify both the supplier and the implementation.

<<Author's Note: PICS Proforma to be added.>>

92. Link Aggregation Control

This section describes the operation of Link Aggregation Control, and a protocol that is capable of automatically exchanging the information necessary in order for Link Aggregation Control to operate in the manner described.

92.1 Conformance

92.1.1 Static conformance requirements

<<Author's Note: Static conformance requirements to be added.>>

92.1.2 Options

<<Author's Note: Options to be added.>>

92.2 Recommendations

<<Author's Note: To be added, if needed.>>

92.3 Link Aggregation Control[5]

92.3.1 Scope of Link Aggregation Control

The scope of Link Aggregation Control includes:

a) Maintenance of configuration information for Link Segments and Link Aggregation Groups;
b) Exchange of configuration information with other systems, in order to determine the requirements for (re-)configuration of Link Aggregation Groups;
c) Creation and destruction of Link Aggregation Groups and their associated Link Aggregation Multiplexers;
d) Addition and removal of links from Link Aggregation Groups;
e) Communication of link state information to the Frame Collection and Frame Distribution functions.

92.3.2 Objectives of Link Aggregation Control

The operation of the Link Aggregation Control function meets the following objectives:

a) **Automatic configuration.** In the absence of manual override controls, an appropriate set of Link Aggregation Groups is automatically configured, and individual Link Segments are allocated to those groups. In other words, if it can aggregate, it will aggregate.
b) **Continuous operation.** Manual intervention, or initialization events, are not a requirement for correct operation. The configuration mechanism continuously monitors for changes in state that require re-configuration.
c) **Low protocol overhead.** The overhead involved in external communication of configuration information between devices will be small.
d) **Low risk of duplication or re-ordering.** The operation of the (re-)configuration functions minimizes the risk of frame duplication and frame re-ordering.

[5]This section incorporates some of the concepts and approaches introduced in the Irvine presentations by Norm Finn and Mick Seaman.

e) **Detect aggregation possibility on initial link startup.** Aggregation capable links are tested to determine whether they are connected to aggregation aware devices on startup, to avoid enabling such links un-aggregated and subsequently re-configuring them as members of an aggregation.
f) **Rapid convergence.** The configuration will resolve rapidly to a stable configuration, in the face of conflicting demands from each end of a link. Convergence will be achieved within at most a very few seconds, and will allow for more rapid convergence where supported.
g) **Deterministic convergence.** The configuration will resolve a deterministic configuration; i.e., the configuration achieved will not be dependent upon the order in which events occur, but will be completely determined by the combination of the capabilities of the individual links and their physical connectivity.
h) **Low failover delay.** Re-configuration on link failure occurs rapidly.
i) **Low risk of mis-configuration.** The configuration functions detect and correct mis-configurations, by performing re-configuration and/or by taking mis-configured links out of service.
j) **Integration of Aggregation-unaware devices.** Links that cannot take part in link aggregation, either because of inherent capabilities or of the capabilities of the devices to which they attach, operate as normal 802.3 links.
k) **Accommodate differing capabilities/constraints.** The configuration capabilities will allow devices with differing hardware and software constraints on link aggregation to be accommodated.

92.3.3 Overview

Link Aggregation Control is responsible for controlling the creation and maintenance of Link Aggregation Groups and their associated Link Aggregation Multiplexers. Its operation makes use of information from the following sources:

a) The inherent properties of the set of individual Link Segments that are visible to the controller;
b) Statically configured parameter values associated with those links;
c) Dynamic information exchanged with other Link Aggregation Controllers reachable via those links, exchanged by means of the Link Aggregation Control Protocol;
d) Dynamic information gleaned from the operation of other protocols;
e) The properties associated with any existing Link Aggregation Groups.

The operation of the Link Aggregation Controller involves the following activities:

f) Identification of links that are candidates for aggregation (92.3.3.1);
g) Checking that candidate links can actually be aggregated (92.3.3.2);
h) Controlling the addition of a link to a Link Aggregation Group, and the creation of the group and associated multiplexer if necessary (92.3.3.3);
i) Monitoring the status of aggregated links to ensure that the aggregation is still valid (92.3.3.4);
j) Removal of a link from a Link Aggregation Group if its membership is no longer valid, and removal of the group and its associated multiplexer if the group no longer has any member Links (92.3.3.5).

92.3.3.1 Identifying links that are candidates for aggregation

The operation of the Link Aggregation Controller is such that, if a given link is a suitable candidate for aggregation, then that link will be included in a suitable Link Aggregation Group. A link is a candidate for aggregation if the following are all true:

a) It is a point-to-point link; and
b) It is a full duplex link; and
c) It connects the same pair of systems; and
d) Both systems are capable of performing Link Aggregation; and
e) Its static configuration permits aggregation; and
f) It is active.

92.3.3.2 Checking that a candidate link can be added to a Link Aggregation Group

Before a link can be added to a Link Aggregation Group, it is necessary to check that the information on which the Link Aggregation Controller decided that the link is a candidate link is still valid, and that all necessary parameters are known. The Link Aggregation Control Protocol is used to validate any existing knowledge related to the link, and to determine the characteristics of the link as understood by the Link Aggregation Controller attached to its far end.

The result of this checking process is that the Link Aggregation Controller now understands:

a) The identity of the pair of systems that the link is connected between;
b) The set of characteristics that each of those systems has associated with that link;
c) Whether both systems understand the same information regarding the link.

92.3.3.3 Adding a link to a Link Aggregation Group

If a link is both a candidate for aggregation, and the link parameters have been successfully checked, then the Link Aggregation Controller will add it to an existing, compatible, Link Aggregation Group. If no compatible group exists, then the Link Aggregation Controller will create a new Link Aggregation Group and its associated Link Aggregation Multiplexer. As part of the process of establishing a new Group, the distribution algorithm that will be employed for that group is also determined.

NOTES:
1—A consequence of the approach described here is that a Link Aggregation Group can exist with only a single active Link Segment in the Group; in fact, all groups start out in life with a single link. This seems to be a simpler approach than treating the case of 2 compatible links as a special case. Hence, the creation of a group of 2 links involves the necessary pre-cursor step of creating a group containing one of the links, and then adding the second.
2—The definition of the distribution algorithms themselves is outside the scope of this standard.

The addition of a link to Link Aggregation Group is achieved in a manner that ensures preservation of frame ordering, and prevents frame duplication. If the link concerned is active, then it must be de-activated before it is added to the group, and sufficient time allowed for any frames that are in transit to be flushed. Activation of the link as part of the group involves signalling to Frame Collection that the link is active, and then ensuring that the corresponding Frame Collection function at the other end of the link is active before signalling to Frame Distribution that the link is active.

NOTE—

Link Segments that are not successful candidates for aggregation (e.g., links that are attached to other devices that cannot perform aggregation) are enabled to operate as normal 802.3 links. For consistency of modeling, these links are regarded as belonging to a "null Link Aggregation Group".

92.3.3.4 Monitoring the membership of a Link Aggregation Group

Each link is monitored in order to confirm that the Link Aggregation Controllers at each end of the link still agree on the configuration information for that link. If the monitoring process detects a change in configuration that materially affects the link's membership of its current Link Aggregation Group, then it will be necessary to remove the link from the group, prior to considering its potential for membership of another group.

92.3.3.5 Removal of a link from a Link Aggregation Group

Removal of a link from a Link Aggregation Group is achieved in a manner that ensures preservation of frame ordering, and prevents frame duplication. The Frame Distribution function is informed that the link is no longer part of the group, the changed configuration information is communicated to the other end of the

link, and then the Frame Collection function is informed that the link is no longer part of the group. The link can then be considered for membership of another Link Aggregation Group.

92.3.3.6 Configuration and administrative control of link aggregation.

Administrative configuration capabilities allow a degree of control to be exerted over the way that links may be aggregated. In particular, administrative configuration allows:

a) The capabilities associated with a link to be identified or modified;
b) Links to be identified as being incapable of aggregation.

In addition to these administrative control capabilities, the configuration functions will provide the capability to signal changes in the aggregation configuration (e.g., creation/deletion of Link Aggregation Groups).

92.3.4 Interfaces

Figure 92-1 illustrates the components involved in the operation of Link Aggregation, and the interfaces between them.

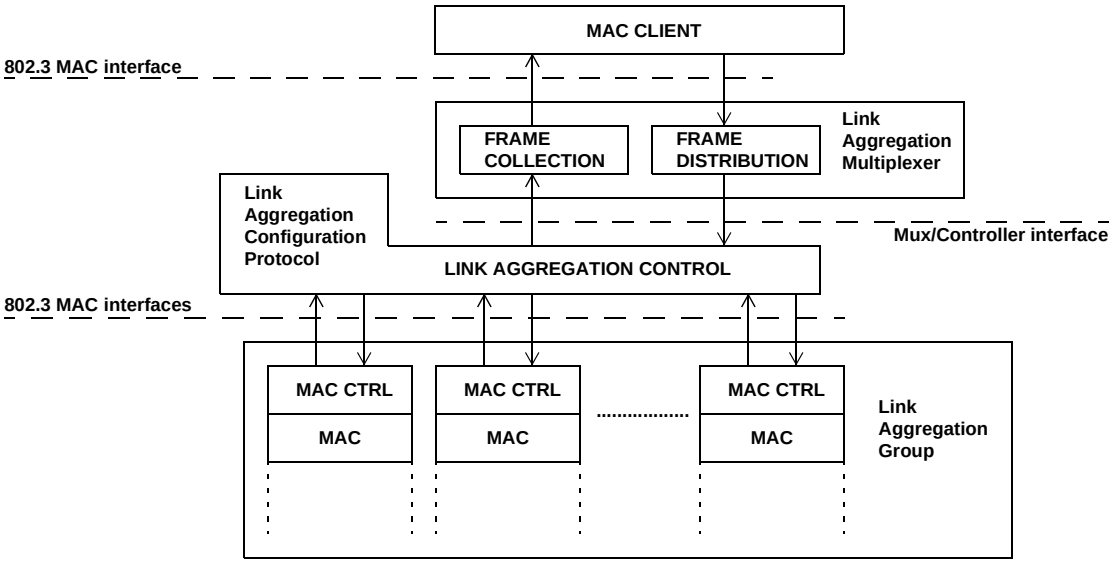

Figure 92-1—Link Aggregation components and interfaces

92.3.4.1 Interface between Link Aggregation Control and each Link

Each link that is part of a Link Aggregation Group presents an 802.3 MAC interface to Link Aggregation Control; this interface is used:

a) To allow Link Aggregation Control to exchange Link Aggregation Control Protocol Data Units (LACPDUs) with any other Link Aggregation Control instance(s) attached to the links;
b) To allow the Link Aggregation Multiplexer to transmit frames from, and receive frames destined for, its associated MAC Client.

The Link Aggregation Control function maintains the following information with respect to each link:

c) The identifier of the Link Aggregation Group to which it currently belongs;
d) The identifier of the Link Aggregation Multiplexer associated with that Link Aggregation Group;

e) The status of interaction between the Frame Collection function of the multiplexer and the link (Collection Enabled, or Collection Disabled);
f) The status of interaction between the Frame Distribution function of the multiplexer and the link (Distribution Enabled, or Distribution Disabled).

92.3.4.2 Interface between Link Aggregation Control and Link Aggregation Multiplexer

This interface is used by Link Aggregation Control to:

a) Inform the Link Aggregation Multiplexer as to the identity of its associated Link Aggregation Group;
b) Inform the Link Aggregation Multiplexer as to the Collection and Distribution status of each link in the Link Aggregation Group.

Data transfer interactions between the Link Aggregation Multiplexer and the individual links that form its Link Aggregation Group are controlled by the status information communicated to the Link Aggregation Multiplexer by the Link Aggregation Controller.

<<Author's Note: Although the currently agreed architecture diagram implies that data transfer between the Mux and the links passes via the Controller, the above description seems to be rather more rational; i.e., the Controller establishes the state information that determines which mux talks to which links, and the actual data exchange takes place directly between the mux & the designated links in its aggregation group, under control of the state information maintained by the controller. Perhaps a little re-drawing might help here.>>

92.3.4.3 Interface between Link Aggregation Multiplexer and MAC Client

The Link Aggregation Multiplexer presents an 802.3 MAC interface to the MAC Client. The Link Aggregation Controller maintains the following information with respect to the interface:

a) The status of interaction between the Frame Collection function of the multiplexer and the MAC Client (Receive Enabled, or Receive Disabled);
b) The status of interaction between the Frame Distribution function of the multiplexer and the MAC Client (Transmit Enabled, or Transmit Disabled).

These status values are exactly equivalent to the logical OR of the status of the Collection and Distribution status of the individual links; in other words, if one or more links in the Link Aggregation Group are Collection Enabled, then the multiplexer is Receive Enabled, and if one or more links are Distribution Enabled, then the multiplexer is Transmit Enabled.

The Transmit/Receive status of the Multiplexer effectively governs the point at which the Multiplexer becomes available to the MAC Client, or conversely, the point at which it ceases to be available.

92.3.5 System, aggregation, link and compatibility identification

In order to allow the Link Aggregation Controller to determine whether a set of links connect to the same system, and to determine whether those links are compatible from the point of view of aggregation, it is necessary to be able to establish:

a) A globally unique identifier for each system that is to participate in Link Aggregation;
b) A means of identifying the set of capabilities that are associated with each link, as understood by a given system;
c) A means of identifying a Link Aggregation Group and its associated Link Aggregation Multiplexer.

<<Author's Note: May also prove necessary to use global labels for the ends of individual links; if so, their individual MAC addresses would be used - see 92.3.6.>>

92.3.5.1 System identification

The globally unique identifier used to identify a system will be an individual MAC address.

NOTE—The MAC address chosen to identify a system may be the individual MAC address associated with one of its links.

92.3.5.2 Capability identification

A number of factors can determine the capabilities of a given link with respect to its ability to aggregate:

d) Its physical characteristics as determined by the ISO/IEC 8802.3 standard, such as speed, whether full duplex or not, point-to-point or shared medium, etc.;
e) Configuration constraints established by the network administrator;
f) Factors related to higher layer use of the link;
g) The characteristics or limitations of the implementation itself.

Some of these factors will be subject to standardization; others are potentially open for definition external to the scope of the standard.

In order to make it possible for link capabilities to be compared within a given system, and for capability information to be exchanged between systems, a *Link Capability Identifier,* or simply, a *Capability,* is associated with each link. A Link Capability Identifier is meaningful in the context of the system that allocates it. Hence, if a System S labels a set of links with Capability C, say, then it can be assumed that any subset of that set of links can potentially be aggregated together, should it prove to be the case that the subset all terminate in System T and that System T has labelled all of the links in the subset with Capability D. The set of links in a given system that share the same Capability value are said to be members of the same *Capability Group*.

The Capability is a simple 16-bit identifier; i.e., there is no sub-structuring within the value allocated that has any significance. Two special Capability identifier values are reserved, as indicated in Table 92-1:

h) The value 0 is the *Null Capability*; this signals the fact that a given link is incapable of aggregation.
i) The value 1 is the *Default Capability.* In the absence of any differences in capability between links (for any reason), all links are allocated this value, in which case all links are potentially able to aggregate together.

Table 92-1—Reserved Link Capability Identifier values

Value	Meaning
0	Null Capability
1	Default Capability

<<Author's Note: The above assumes a single Capability is assigned to each link. Does a given link ever need more than one capability identifier - i.e., given links A, B, C, is it ever the case that A+B is OK, B+C is OK but A+C is not OK?

Also, there may be some advantage to be gained in defining further default values, for example to allow default Capabilities to be defined according to link speed, for example (e.g., 1 = default for 10 megs, 2 = default for 100 megs, 3 = default for 1 gig...etc). However, this may not be ideal for devices that can autonegotiate among a number of speeds.>>

All other values of Capability are freely available for allocation, with locally significant meanings.

92.3.5.3 Link Aggregation Group identification

A Link Aggregation Group consists of a set of links that all share the same capability, and which terminate in the same pair of systems. A Link Aggregation Group Identifier (LAG ID) is therefore a compound identifier, consisting of:

a) The System Identifier associated with one end of the set of links, and the Capability assigned to the set of links by that system; and
b) The System Identifier associated with the other end of the set of links, and the Capability assigned to the set of links by that system.

Hence, if System S has allocated Capability C to a given Link Aggregation Group, and the same Link Aggregation Group terminates in System T with capability D, the (globally unique) identifier of that Link Aggregation Group is {SC, TD}.

If any of the elements {S, C, T, D} of a LAG ID are zero, then the LAG ID identifies a set of links that cannot be aggregated; either because one or other system believes the link to be non-aggregatable, or because the remote system ID is unknown. LAG IDs of this form are collectively identify the Null Link Aggregation Group; in other words, they collectively identify a set of links from which no two links can be successfully aggregated together.

NOTES:
1—There is no significance to the ordering of these system/capability pairs; hence {SC, TD} is the same as {TD, SC}, but {SC, TD} is not the same as {SD, TC}. A consequence of this formulation for the ID of a Link Aggregation Group is that, for a given {SC, TD} combination, only a single Link Aggregation Group can exist - see 92.3.6.
2—It may also prove to be convenient for some purposes to represent the {SC,TD} pair by a locally significant identifier.

92.3.5.4 Link Aggregation Multiplexer identification

A Link Aggregation Multiplexer Identifier (LAM ID) is a globally unique identifier consisting of an individual MAC address.

NOTE—This identifier may be the MAC address of one of the links in the associated Link Aggregation Group, or may be a distinct MAC address.

92.3.6 Configuration capabilities and restrictions

The formulation chosen for the Link Aggregation Group identifier (92.3.5.3) has the consequence that it is not possible to represent two or more Link Aggregation Groups that share the same combination of {SC, TD}. Hence, placing configuration restrictions on the size of an aggregation (e.g., for a Capability Group containing N members, restricting the size of any aggregation to subsets of N of no greater than M members) is only possible if it is also acceptable that only one Link Aggregation Group can be constructed from that Capability Group for a given {SC, TD}. In practice, this restriction can be somewhat alleviated by sub-dividing Capability Groups and allocating different Capabilities to each subdivision.

If restrictions on the size of Link Aggregation Groups are permitted by the standard, then, in order to maintain the objective of deterministic convergence of the configuration, it will be necessary to communicate link

identifiers between participating systems, and for the configuration to substitute links in the Group as they become available or unavailable.

<<Author's Note: Definitely desirable not to have such size restrictions if at all possible, as it complicates matters for the protocol; achieving a deterministic result with such restrictions would involve the establishment of a master/slave relationship between the systems in order to decide *whose* deterministic configuration to use. However, it is probably smart from a debugging perspective to have a link ID carried in the protocol anyway. The detailed operation and protocol sections below assume no such size restrictions; the mechanisms proposed therefore allow a peer relationship to exist between pairs of protocol partners.>>

92.3.7 Detailed operation of Link Aggregation Control

92.3.7.1 Parameters

Link Aggregation Control maintains parameter and state information as follows:

a) For each **Capability Group:**
   1) The Capability Identifier (92.3.5.2);
   2) The set of Link Identifiers of Links that share that Capability.

b) For each **Link Aggregation Group:**
   1) The Link Aggregation Group Identifier (92.3.5.3), consisting of the local System ID (92.3.5.1) and Capability and the remote System ID and Capability (i.e., an identifier of the form {SC, TD});
   2) Receive Enabled/Disabled state;
   3) Transmit Enabled/Disabled state;
   4) The set of Link Identifiers of Links that are members of the Link Aggregation Group;
   5) The identity of the Link Aggregation Multiplexer (92.3.5.4) associated with the Link Aggregation Group.

c) For each **Link:**
   1) The Link Aggregation Group Identifier that the local system believes the Link belongs to (Local LAG ID);
   2) The Link Aggregation Group Identifier that the remote system believes the Link belongs to (Remote LAG ID);
   3) The Local Collector Enabled/Disabled state that applies to the link;
   4) The Local Distributor Enabled/Disabled state that applies to the link;
   5) The Remote Collector Enabled/Disabled state.

92.3.7.2 Allocating a Link to a Link Aggregation Group

The operation of Link Aggregation Control will result in each Link either:

a) being allocated to an active Link Aggregation Group, or
b) being allocated to the Null Link Aggregation Group,

depending upon the configuration at either end of the link and the ability/inability of the remote system to engage in Link Aggregation behavior.

The allocation, and re-allocation, of links to Link Aggregation Groups is determined by the current values of the LAG ID parameters held for each link; these values also determine behavior with respect to whether or not Collection and Distribution are enabled. A key factor in this process is whether the Local and Remote systems agree on the value of LAG ID, or whether they disagree. The following possibilities exist:

c) Local LAG ID and Remote LAG ID differ, either because the local system has not received up-to-date information from the remote system, or vice versa. Attempts will be made (by means of the

Link Aggregation Control Protocol, or by other means) to reach a state where this information no longer differs; in which case, the situation becomes one of d) or e) below. However, if this situation persists regardless of attempts to update the LAG ID information, it can be assumed that the remote system cannot take part in Link Aggregation, and the link is therefore a member of the Null LAG; i.e., it can only be operated as a non-aggregated 802.3 link.

d) Local LAG ID and Remote LAG ID are the same, but the LAG IDs indicate that the Link is a member of the Null LAG; i.e., both systems can potentially take part in Link Aggregation, however, one or other system regards this link as not suitable for aggregation. The link is therefore a member of the Null LAG; i.e., it can only be operated as a non-aggregated 802.3 link.

e) Local LAG ID and Remote LAG ID are the same, and the LAG IDs indicate that the Link is a member of a Non-Null LAG; i.e., both systems can take part in Link Aggregation, and both systems regard this link as suitable for aggregation. The link can be enabled as a member of the LAG; i.e., it can be aggregated with other links that share the same LAG ID.

In cases d), and also if case c) persists, the link is simply enabled as a normal 802.3 link.

In case e), the link is added to the Link Aggregation Group identified by the LAG ID; if the Link Aggregation Group does not exist, the Link Aggregation Controller will create the group, and create an instance of the Link Aggregation Multiplexer to service it.

Once the link has been added to a (non-Null) Link Aggregation Group, its Local Collector state can be switched to Enabled, thus preparing the link for reception of frames from the remote Frame Distribution function, and that information communicated to the remote Link Aggregation Controller. As this means that at least one link in the Link Aggregation Group has its Local Collector state Enabled, then the Receive state of the corresponding Link Aggregation Multiplexer will also be Enabled. Once the state information held for the Link also indicates that the Remote Collector state is enabled, the Link Aggregation Controller can set the Local Distributor state to Enabled, thus allowing the link to be used by the Frame Distributor function. As this means that at least one link in the Link Aggregation Group has its Local Distributor state Enabled, then the Transmit state of the corresponding Link Aggregation Multiplexer will also be Enabled.

92.3.7.3 Moving a Link to a new Link Aggregation Group

If either the Local or Remote LAG ID information changes, due to re-configuration of either end of the Link, then it will be necessary for the Link Aggregation Controller to move the link from its existing Link Aggregation Group to a new Link Aggregation Group. At the point where the change is detected, the Local Frame Collector and Local Frame Distributor states are set to Disabled. The link can then be removed from its current Link Aggregation Group, once it is certain that there are no more frames that are in transit on the link.

<<Author's Note: Ensuring that all frames are flushed from the link could involve the use of an explicit "flush" protocol, as suggested in Norm Finn's Irvine presentation; alternatively, the use of a suitable time delay could be sufficient here.>>

Once the link has been removed from its Link Aggregation Group, the situation has effectively been reduced to the one described in 92.3.7.2; the link can be allocated to its new Link Aggregation Group and re-enabled once agreement has been reached between the Local and Remote LAG IDs.

92.3.7.4 Link Aggregation Control State Machine

The operation of Link Aggregation Control with respect to an individual Link, as described informally in 92.3.7.2 and 92.3.7.3, is shown below. The abbreviations used to identify the states are as follows.

<<Author's Note: State machine description to be added.>>

92.4 Link Aggregation Control Protocol

The Link Aggregation Control Protocol provides a means whereby the necessary information can be exchanged between the two ends of a Link in order to allow two interacting Link Aggregation Control instances to reach agreement on the identity of the Link Aggregation Group to which the Link belongs.

NOTE—This is only one means whereby such agreement could be reached. Others include static configuration of the necessary parameters, or the use of information gleaned from other protocol mechanisms.

92.4.1 Protocol Design Principles

The following design principles were used in developing the protocol described in this section:

a) The protocol depends upon the transmission of information and state, rather than the transmission of commands. In other words, messages sent by the first party convey to the second party:
   1) What the first party knows, both about its own state and that of the second party;
   2) How confident the first party is in that knowledge.
b) The information conveyed in the protocol is sufficient to allow the recipient to determine what action to take next.
c) In the absence of a response from the other party indicating that information has been received, the sender can assume that if a message has been sent twice, then its content has been seen by the recipient (if one exists).

92.4.1.1 Informal Protocol Description

The Link Aggregation Control Protocol operates by each of the two Link Aggregation Control Protocol Entities attached to a Link declaring to the other what it currently knows of the state of that Link. The information exchanged consists of the following elements:

a) The System ID (92.3.5.1), S, of the local system;
b) The Capability (92.3.5.2), C, that the local system has assigned to the Link;
c) The System ID of the remote system, T, as currently known by the local system;
d) The Capability assigned to the link by the remote system, D, as currently known by the local system;
e) The state of the Local Collector (Enabled or Disabled) with respect to the Link;
f) A "Need to Know" flag, signalling whether the local system is confident of the information communicated, or not. (If asserted, this prompts the remote system to "need to tell"; i.e., to update the local system's information.);
g) The Link identifier assigned to the link by the local system.

Information is exchanged between the systems attached to the link in either of two circumstances:

h) When the sender needs to know something from the recipient: in other words, when what the sender knows is incomplete, appears to conflict with what the receiver knows, or is sufficiently old to be no longer reliable;
i) When the sender needs to tell something to the recipient: in other words, when the sender knows that changes in its own state mean that the information that the recipient has is no longer valid, or that the recipient's information may need to be updated.

The transmission of information is controlled by a single timer. The value used to restart the timer will depend upon the state that the link is in; for example, if the link is known to be connected to a system that cannot engage in the protocol, then long timer values are used, whereas short timer values are used when it is known that a configuration change has occurred and therefore that a new state of agreement needs to be reached.

The following diagrams illustrate the sequence of protocol exchanges that would occur in order to reach agreement in various circumstances. In these diagrams, the following notation is used:

SC, TD: Link Aggregation Group Identifier, in which:
- C: Capability ID for System 1. (0 = No capability or Unknown.)
- D: Capability ID for System 2. (0 = No capability or Unknown.)
- S: System ID for System 1. (0 = Unknown.)
- T: System ID for System 2. (0 = Unknown.)

Figure 92-2 illustrates the establishment of a newly enabled link as a member of a Link Aggregation Group.

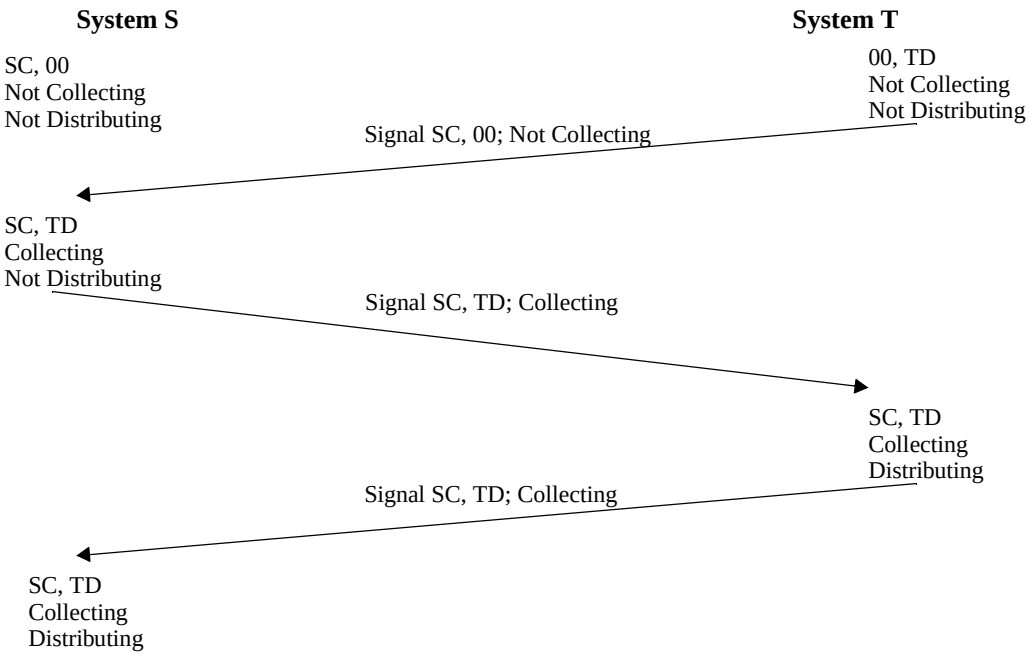

Figure 92-2—Protocol exchanges to enable a link

Figure 92-3 illustrates the effect of changing the capability in one of the two systems, for a link that is part of an active Link Aggregation Group.

Figure 92-4 illustrates the effect of changing the capability in one of the two systems, where that capability had hitherto indicated that the link could not participate in Link Aggregation.

Figure 92-5 illustrates the effect of removing the capability in one of the two systems, where that capability had hitherto indicated that the link could participate in Link Aggregation.

92.4.1.2 Link Aggregation Control Protocol State Machine

The state table contained in Table 92-2 gives a description of the Link Aggregation Control Protocol state machine, from the point of view of the state changes caused by message receive events and management events.

The following conventions are used in the table:

State identifiers consist of the following components (and in this order):

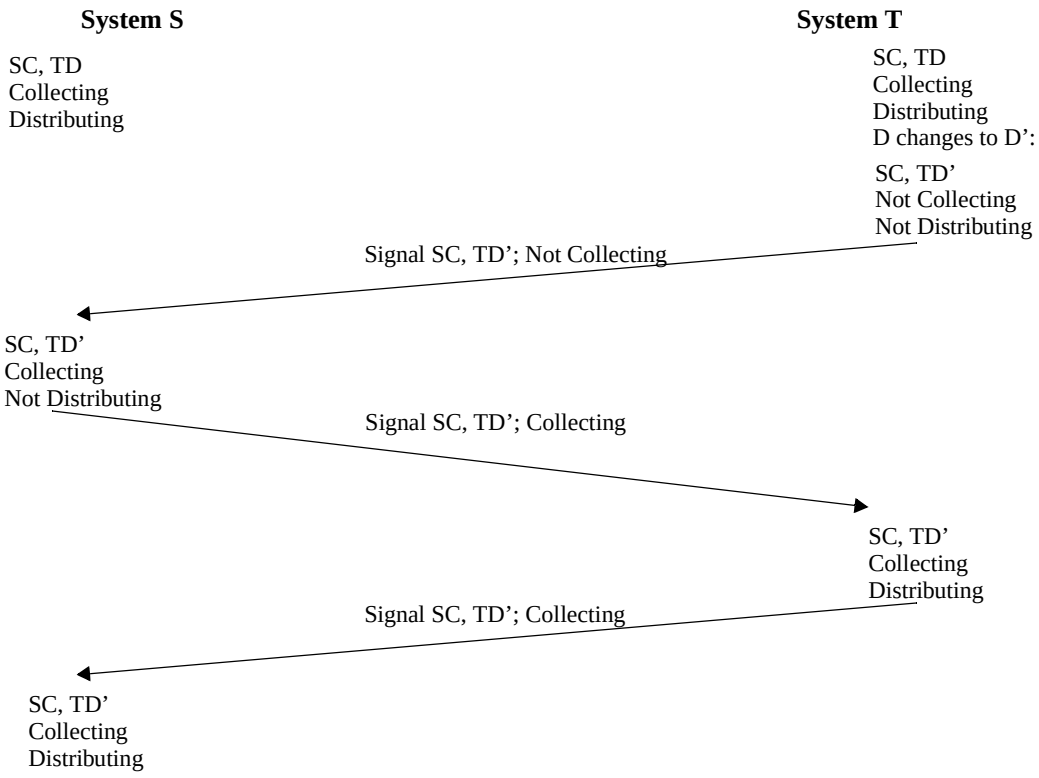

Figure 92-3—Protocol exchanges to change a Capability - link formerly aggregated

D: Capability of this system with respect to this link; 0 = Null Capability
C: Capability of remote system with respect to this link; 0 = Null Capability or Unknown
K/0: Collecting/Not collecting
J/0: Distributing/Not distributing

The following valid states exist:

0000: Neither system believes the link to be capable of aggregation.
D000: Local system believes the link to be capable of aggregation; remote does not.
0C00: Remote system believes the link to be capable of aggregation; Local does not.
DC00: Both systems believe the link to be capable of aggregation, but collection/distribution not enabled locally yet.
DCK0: Collection enabled, Distribution not yet enabled.
DCKJ: Both collection and distribution enabled. (Implies that the remote system has enabled collection.)

Messages received from the remote system are represented in the Events column as a sequence of letters {SCN, TD}, where:

S: System identifier of remote system
C: Capability assigned to link by remote system
E/0: Whether the remote system has enabled Collection for this link. E= Enabled, 0= Disabled
T: System identifier of local system, as understood by remote system
D: Capability assigned to link by local system, as understood by remote system

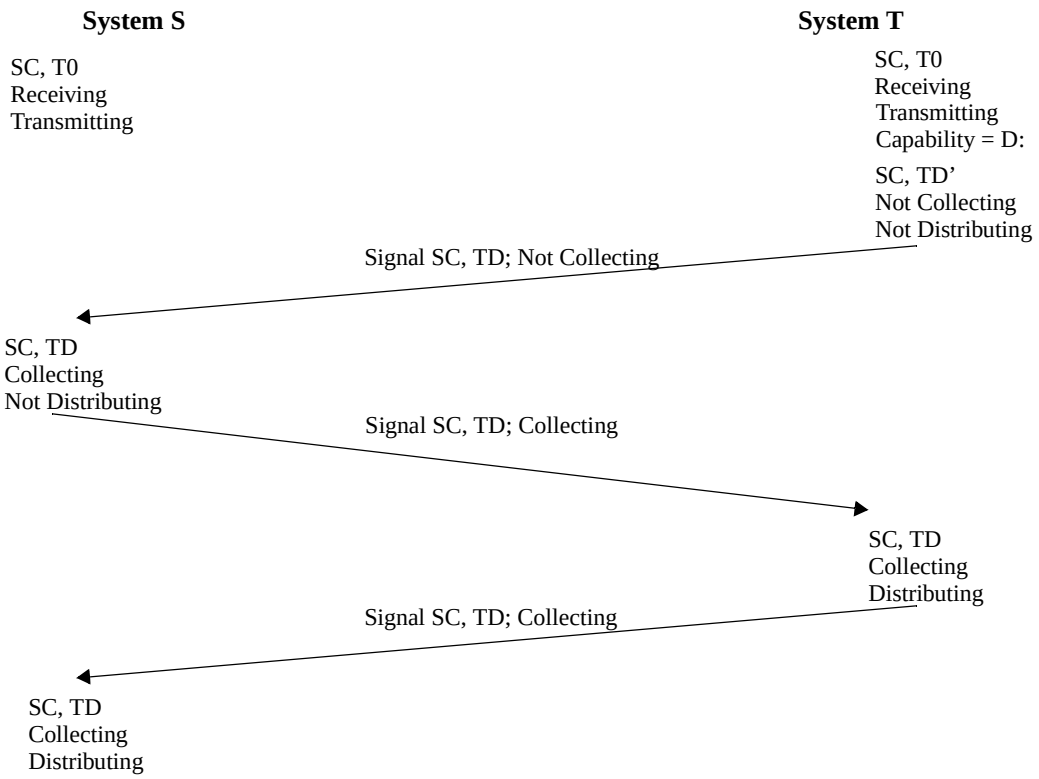

Figure 92-4—Protocol exchanges to add a Capability - link formerly not aggregated

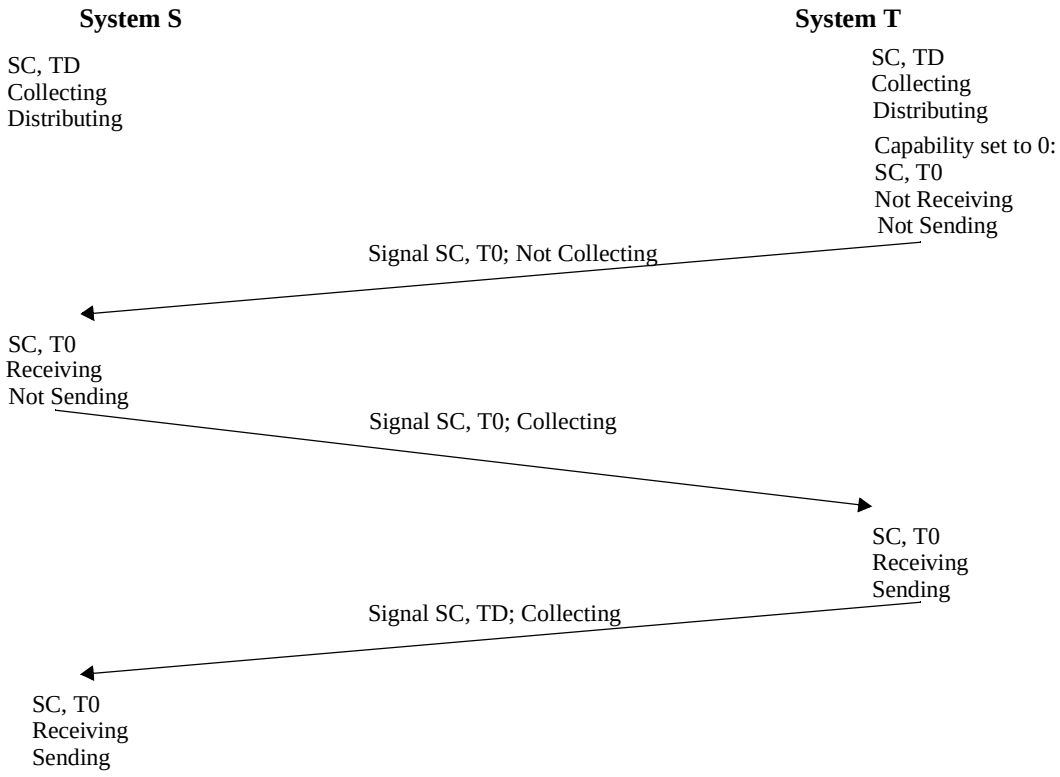

Figure 92-5—Protocol exchanges to remove a Capability - link formerly aggregated

Use of S, C, T or D followed by an apostrophe (e.g., S') indicates a value of that element that has changed with respect to the last value seen. Use of X indicates "don't care" values.

The entries in the cells of the table indicate the next state that will be reached on receipt of the event specified.

Table 92-2—Link Aggregation Control Protocol State Table

EVENT	STATE					
	0000	**D000**	**0C00**	**DC00**	**DCK0**	**DCKJ**
S0X, T'D' or s0x, TD	-	-	0000	D000	D000	D000
XCX, T'D'	0C00	DC00	-	-	DC00	DC00
SC0, TD	0C00	DCK0	-	DCK0	-	DCK0
SCE, TD	0C00	DCKJ	-	DCKJ	DCKJ	-
T0X, XX	-	-	0000	D000	D000	D000
TCX, XX	0C00	DC00	-	-	DC00	DC00
Timer expired, no message received since last expiry	-	-	0000	D000	D000	D000
Management makes link aggregatable (D <>0)	D000	-	DC00	-	-	-
Management makes link non-aggregatable (D=0)	-	0000	-	0C00	0C00	0C00

NOTES:
1—Messages TXX, XX are messages that indicate the existence of a loopback condition.
2—The refresh timer is restarted on receipt of any message.
3—DC00 is only entered as a result of loopback (or faulty communication).

The state machinery that determines when messages are sent by a system on a given link depends upon the following variables:

NK: True = Need to Know, False = Don't Need to Know.
NT: True = Need to Tell, False = Don't Need to Tell.
Retry: Simple counter, can carry values 1, 2 or 3.
Refresh timer.

The behavior with respect to these variables is as follows:

a) Expiry of the Refresh timer causes NK to be set to True, and the timer restarted.
b) Receipt of a message containing inconsistent information (i.e., information that differs from the state known locally), or if the NK flag in the message is asserted, causes NT to be set to True.
c) A change of local configuration state (i.e., D or J change state) causes NT to be set to True.
d) If a message transmission opportunity occurs, and Retry is less than 3, then a message is transmitted giving current known state, and NK value and the Retry count is incremented. If the Retry value is now 2, the Refresh timer is restarted with a short value. If the Retry count is now 3, then the Refresh timer is restarted with a long value, and NK and NT are set to False.
e) Receipt of any message causes NK to be set to False. If the message is consistent with current knowledge, and NK is not asserted in the message, then NT is set to False, the Retry count is set to 3 and the Refresh timer is restarted with a long value.

<<Author's Note: Need to add some definitions for timer sizes & the circumstances under which different sizes are used. Different timer values are needed for:
Link in no aggregation
Link in aggregation of 1
Link in aggregation of many.>>

92.4.2 Link Aggregation Control Protocol Data Unit structure and encoding

<<TBA>>

92.5 Management

<<Author's Note: This section T.B.A. Needs to specify:

- The management functionality that is available for control/monitoring of Link Aggregation;

- The MIB definition that realizes that functionality, using appropriate notation (SNMP MIB, GDMO defs, etc.) to fit in with the format of other 802.3 MIB definitions.

Management is mostly achieved by managing the Capability values associated with individual Links. May also be desirable to allow the configuration to be "forced", i.e., allow total manual control of allocation of Links to LAGs.

In addition to straight configuration capability, management "hooks" to allow requests for resynchronization, in order to enable the use of likely "hints" from other protocols that the configuration may be changing. This resync capability should be available both for individual links and for Link Aggregation Groups.>>

92.6 Protocol Implementation Conformance Statement

The supplier of an implementation that is claimed to conform to clause 92 of this standard shall complete a copy of the PICS proforma provided below and shall provide the information necessary to identify both the supplier and the implementation.

<<Author's Note: PICS Proforma to be added.>>