Minutes of the IEEE 802.3ad (Link Aggregation) Task Force Interim Meeting

September 1-2, 1998

Austin Marriott at the Capitol

Austin, TX

The meeting was called to order by the Steve Haddock, Task Force Chair, a few minutes after 9 am, 9/1.

Introductions

Presentation of the agenda

Pointers to reflector and web site

the email exploder is stds-802-3-trunking@ieee.org

to subscribe, send email to

majordomo@majordomo.ieee.org

containing the line

subscribe stds-802-3-trunking@ieee.org <your email address>

the website is at

http://grouper.ieee.org/groups/802/3/ad/index.html

Timeline for the development of the standard

New proposal cutoff is at the end 11/98 meeting.

No current indication that any additional proposals for the standard will be forth coming.

Presentations

Link Aggregation Control Protocol Update - Tony Jeffree

An update to the presentation given in July, 98. Based on the best portions of the presentations of Finn/Wakerly/Fine and Jeffree and incorporates comments from the July, 98 meeting

The flush protocol is not yet addressed.

Discussion/comments from the floor during the presentation.

Objection by Jeff Lynch to the objective of very low probability of misdelivery. Why should not the objective be zero probability of misdelivery? Response: nothing is perfect.

Question as to whether a special key is desirable that says that a link can not aggregate? Possible, will consider later.

Discussion between Michael Fine, Norm Finn, Mick Seaman and others about the need for the InSync/Out of Sync bit.

The terms "Desirable" and "Auto" are considered confusing by some and change is requested. The terms "Nervous" and "Cool" are also considered a problem.

Question about the need/benefit for/of "auto" mode relative to its added complexity. Reason is to support plug and play.

Rich Siefert questioned the complexity added for unreliable links. Protocol may be running where the link quality signal is not visible. And it may be operating in some ways outside the specified conditions. This lead to a discussion of whether standard should worry about operation outside the defined domain.

Suggestion that interaction with spanning tree be explicitly discussed in the slides/presentation.

Discussion of the relationships between AgPorts and PhyPorts. What is presented in the slides is assumptions or guide lines, not normative requirements. What is necessary for correct operation and what goes into the standard is not yet determined.

The goal of determinism needs to be explicitly decided as it can cause some rearrangement if a port is removed from an aggregation and connected to another system. Determinism is defined as not being history dependent. Non-determinism solves a few problems, but introduces a number of others.

More discussion on the rules for associating PhyPorts with AgPorts and their side effects.

More discussion of the InSync bit and its use and whether it is needed. The SNMP limitation of not being able to handle a port than can receive, but not transmit, the fact that for some MACs the distributor and collector can not be turned on separately and the fact that the software generally sees a port as totally up or down suggests that another approach in stead of the use of the InSync and Collector bits may be desirable.

Question as to whether the distribution algorithm is to be standardized. The group has to decided not to standardize the algorithm, with the exception that the collector must be required to not reorder the packets.

More discussion on initial state. Three possible initial states "I am desirable", "I am auto" and my partner is "auto", I am "auto" and my partner is "desirable".

For discussion tomorrow

Determinism
Initial State
State Names

IEEE 802.3 Link Aggregation, Flush Requirements and Operation - Jeff Lynch

---

Discussion/comments from the floor during the presentation.

Question raised as to why a flush is needed when a link is removed from a group?

Discussion of when the flush response is issued. Response must not be issued until the message reaches the collector. If there are multiple priority queues, the flush message must go on the lowest priority queue that may contain traffic from any of the flows being flushed.

Suggestion that there is no need to retransmit a flush and just set the timer to around one second and on timeout just go back to transmit in the Flush Transmit state diagram.

Suggestion that packets be dropped rather than attempting to hold them during a flush.

The flush packet must have a priority as low as any on the flows that it is flushing, but a high "do not drop" priority.

Since there are no priorities in 802.3, it was suggested that it might be desirable to not even mention priorities.

The decision of how the distributor responds to a flush response is local to the distributor.

Discussion of how often to allow transmission of flush frames, primarily to prevent poor designs. No specific conclusion reached.

Discussion of how many flush frames can be outstanding. No conclusion.

Discussion of whether there should be any specification on the flush receiver. Sentiment was to not place specifications on the receiver.

Straw poll on terminology taken very informally (a sheet of paper with the candidates was circulated though the attendees).

Winners:

Active/Passive
want fast/want slow

There were complaints about both sets of winners.

"Active/Passive" considered by some to be too common and what is active or passive is not specified .

"Want Fast/Want Slow" suggests symmetry which is not correct. "Want Fast/Slow OK" is more accurate.

After a number of suggestions and much discussion, "Active LAC/Passive LAC" was accepted by a majority on a voice vote and " Short Timeout/Long Timeout" was selected by unanimous voice vote.

The meeting was recessed for the day.

The meeting was reconvened about 9 am, 9/2.

Remaining agenda items -

two presentations
open discussion

Link Aggregation Control, Frame Format and Syntax - Rich Siefert

Discussion/comments from the floor during the presentation.

Suggestion that a LA frame be such that if one does escape, it will cause no harm.

Observation that some switches discard all frames of certain types making such bridges non-upgradable. The BPDU addresses are the issue.

There was a heated debate about whether 802.3x requires that 802.3x requires throwing away MAC control frames with unrecognized opcodes.

There was a heated debate about whether Link Aggregation (LA) should use a mechanism that breaks a lot of ports in the field. The specific issue is the possible use of 802.3x MAC control frames for LA. One vendor claims to have millions of ports in the field that throw away MAC 802.3x MAC control frames that are not defined.

Observation made that using MAC control frames that can leap over regular data frames for LA control introduces a bug that can result in stale data going to the wrong place.

Observation made that using MAC control frames for LA that flow in spite of flow control could cause high frame loss rate due to lack of buffers.

Observation made that to use 802.3x MAC control frames for communication between layers above the MAC control layer is a problem as 802.3x MAC control frames are not a transport mechanism for higher layers.

Concern expressed that using 802.3x MAC control frames would overload the management processor.

Follow on discussion

Comments by Shimon Muller - Sun

Does not support Rich's proposal

The cleaner/architecturally pure solution is to use frames that are sourced and sunk by the layers that use the frames.

802.3x is an extension of the MAC.

Discussed some options

Proposes using a different Ethertype instead of recycling the 802.3x MAC control Ethertype.

Follow on Discussion

Reiteration of concern about 802.3x MAC control frames moving ahead of data frames resulting in stale data going to the wrong place.

Observation that LA and LAC (Link Aggregation Control) are layers above the MAC and the MAC control layer and a desire to not change the interface between the MAC and the MAC client.

Observation made that we will not be able to define LA such that it is all previous/existing implementations and we can not ignore all previous /existing implementations.

LACP - Frame Types and Protocol Extensibility - Tony Jeffree

Discussion/comments from the floor during the presentation.

Observation made that 802.3x need not be changed unless MAC control frames are selected for use LACP.

Observation was also made that it would likely be a good thing to bring forward text clarifying 802.3x text with regard to the handling of MAC control frames with unrecognized opcodes.

Observation made that it is necessary to have a very strict interpretation of extensibility built in to ensure that forward extensibility works.

Observation made that adding text to standards to support future extensions is different from crisp text that details fault/error handling and that the former is not something 802.3 has generally done.

Question raised as to whether using a different reserved address will circumvent the problem that arises if the MAC control frame address is used. We have been consuming these addresses at a significant rate. And if we do use a new address, we need to make sure that its use is extensible.

Observation that the handling of 802.3x MAC control frames is generalized for any MAC control frame and not specific to flow control.

Observation that while designing an extensible protocol is highly desirable, it is very difficult to do (providing adequate extensibility for needs you do not know). And there is a schedule implication if we undertake defining an extensible protocol.

Observation that the IETF has extensive experience in developing extensible protocols. That does not mean that you can always succeed.

Observation made that if you really expect a version n+1, you must define how you handle

Observation that in defining a standard that "reserved" does not mean "don't care" and commonly, what something is reserved for is not yet known. Therefore, great care must be used in handling of "reserved" fields.

Observation that undefined MAC control frames opcodes are marked "reserved".

Several expressions of concern about reflecting fields that you do not understand.

Observation made that variable size control frames that are likely to be handled in hardware is a problem with respect to insuring adequate buffers.

If we use a new MAC address, don't call it a LAC address as that conveys too narrow a meaning if we intend any extensibility.

Suggestion in the list of use extensibility principles, separate what to do with packets that are too long from what to do with packets that contain values in fields that the receiver believes are reserved.

Observation that MAC control frames were intended for speed and functions that do not require short response time should not be loaded onto MAC control frames.

Multiple comments about LACP frame length

Link Aggregation layer model - Geoff Thompson

MAC client

LA (distributor/collector)
LA-control (Add/Delete link)
MAC-Control
MAC
Physical

Difference between proposed versions of layer model is whether MAC-control and LA-control are mergered or separated.

Straw poll on what ethertype LAC frames use

mac control ethertype value- 0
new ethertype value - 35

straw poll on LAC frame addresses

first vote

multicast address in the 802.1 block - 39
multicast address outside the 802.1 block - 1

second vote

pause multicast address - 9
new multicast address from 802.1 block - 12
no opinion at this time - 24

Discussion

Argument that address outside the 802.1 block makes ensuring the non-propagation on LAC frames difficult and that we should use an address within the 802.1.

Some arguments for not closing the door at this point and wait until the tales of woe are in.

Observation that the tales of woe may be more important than the small theoretical benefits of 1 over 2 or vice versa.

Call for Patents by Geoff Thompson

The 802.3 patent policy was read.

The policy applies to the 802.3 attendee and his/her employer.

Letters responding to the patent call are presented to PatCom for review for compliance with IEEE requirements and then accepted and the information is a placed in an IEEE data base.

Letters for patents under application need only state that a patent application is pending that may apply to the standard under development and that if a patent is grated, it will be made available for license on a reasonable and non-discriminatory basis.

Other business

Minutes of July meeting will be approved during the November Plenary

Request that positions taken at this meeting be presented to 802.3 during the opening Plenary in preparation for an 802.3 vote during the closing 802.3 plenary.

Some discussion of whether to schedule a Monday morning meeting at the November Plenary.

802.3ad will have a meeting Monday of the November Plenary week beginning at 9 am.

Reminder to review the 802.3 minutes which are on the 802 web page.

Suggestion that we make a list of items that we have already reached consensus on. Resolution was to prepare the list and circulate it on the exploder before the next meeting.

Request that the LA protocol use as few timers as possible. The current proposal has three timers. Observation made that timers are expensive in both hardware and software.

Discussion on Determinism -

The proposed determinism rules an not invariant part of the current protocol.

Questions to ponder

> how much do we need/want to standardize? Do we specify only the simplest case or do we specify much more?

Observation that at least one view is that standard should be limited to a single aggregation port in which case the determinism issue goes away.

Opinion that absolute minimum be in the standard to insure interoperability.

Opinion that determinism is not a central issue. It is desirable that a given box do the same thing each time.

Observation that multiple agports is what allows the same standard to serve for routers, end stations, switches and servers.

Observation that the totally non-stop quality (move one wire after another and the system never goes down) may be more important to some than "determinism".

Observation that the end user would like the box to come up the same way each time and not have to look at the box to know which way it came up. Simplicity to the end user is important.

Opinion that we need a strong "determinism" so that all of these boxes look the same at least at powerup or reset.

Opinion that having the same result after boot/powerup as after reconfiguration is not as important as knowing what the configuration be after reconfiguration as non-stop operation is important to some/many.

Opinion that deterministism as proposed should be required, but perhaps allow other algorithms to support other goals such as non-stop.

Opinion that debugging will be much more difficult if configuration algorithm is not the same as on power up as all the reconfiguration may not be reproducible.

Opinion that pseudo-static assumption of networks should be applied here.

Steve's summary

aggregation rules are orthogonal to the general protocol

great variation of opinion as to the volume and specificity of determinism we want to require.

whether a common denominator is needed

Observation that determinism question may be answered based on whether it is based on physical connections or is not based on physical connection.

Concern expressed about allowing optional behaviors in additional to a specific behavior.

Observation that issues of which port to attach a bridge filter to has been ignored in much of the discussion.

The meeting was adjourned about 4 pm.

Bill Quackenbush
Secretary du Jour