Tuesday 19th January 1999 - Link Aggregation

Morning Session

Rich - Review of the 802.3ad Draft

Outlined changes to clause 1 and 2. Other changes are clause 43 (the balk), distribution rules plus the normative annexes for slow protocols.

Clause 1

From 1.4, some terms may be removed if they're not used.
Geoff T Some terms could be removed and used only if the concept is needed (example - bridged LAN).
Is a router an end station? Yes. Rich will add it.

"Port Thing" - MAC as seen by the aggregator. Better wording? Aggregator Port. Then have text which says "in the context of clause 43, Aggregator port is referred to as Port in this text, outside clause 43 Port means Port (as in the 802.3 context)".

"Link Thing" - Geoff suggests adding Full Duplex to the definition (even though the objectives say we'll only do full duplex). Dropped. Follow same tone as for "Port Thing".

"System" - Too vague currently. 22 definitions in the IEEE dictionary! Keep Rich's definition (23rd).

Clause 2

Rich outlined the new features; extras included: MA_DATA request now includes FCS and source address. The MA_DATA.indication now includes a FCS. The changes do not affect any other clauses (because Rich designed it to be backward compatible - if you don't fill in the source address, it gets filled in for you).

Do we need to go to every reference which uses this service interface and put an extra comma in to denote a null for the extra parameter? Resolution to drop the double comma which Rich used in clause 43 for nulls. (Making it consistently wrong!).

Clause 43

Comment on 43.1.1.H - "Low risk of duplication and misordering" - S Haddock. "link invariance are maintained" - what does this mean? Resolution to get rid of "invariant" sentence - it doesn't add any value.

Comment on 43.1.1.K - "Backwards compatibility with existing hardware" - S Haddock - is this a job of the standard? How can the "group" know? Change to a note - then the editor can remove before publication.

43.1.2 - Figure 43-2. Rich outlined how the diagram relates to the Link Agg flow of control and its relationship to the MAC Client. It would be more correct to have an aggregator Parser / Multiplexor connecting to each MAC rather than have the Control Parser / Multiplexor talk to it. Rich will figure out how to draw it.

Rich highlighted the fact that it is the responsibility of the distributor to maintain packet ordering. Also highlighted for discussion that the Mac address for the aggregator may be taken from one of the ports under the control of that aggregator. We should make the source and destination address rules explicit for better discussion next time (Mick's comment). Rich - summary - on rx you receive unicasts to the aggregators address; sending frames - the source address should be the port's for LACP and Flush; but the address for LACP data frames should be the one of the LA (which "may" be of one of the ports).

43.2.x Rich goes through the function of distributor and collector, including the state machines.
43.2.5+ Flush (Marker generator / receiver).
4.3.2.7 Strike note referring to state diagram 43.2.7.1.

Figure 43-5 Missing UCTs. Rich to fix. David Law points out that the DA, Length/Type and Type as vague, so suggestion is to change to "TLV Type". Subtypes should be globally defined to ensure other "slow protocols" don't reuse our subtypes. This will be qualified.

Joris - how do I know a marker comes after the data frames I'm flushing? Have two Das - one for priority Flush frames, and one for the other Aggregation Control frames. Discussion follows. General feeling that this would be hard work. This issue was dropped as nobody had a dying need to have the current text changed (Flush is optional after all). The model has no concept of queuing, so this is a nonissue as far as the standard is concerned.

Flush to be cl 43.2.11.2.3(a) - Change flush sentence to show that it refers to just the frames on a specific link (request from Joris).

43.2.11.2.5 Add sentence to say you need to flush for conversations you may move to a different link (request of Joris).

43.2.10 Addressing. Strike editor's note.

Change of speaker to Tony:

43.2.11 Link Aggregation Control. Cross reference to "Key" will be added (43.2.11.2.3) since it's the first use of this word. Also a typo to fix (the "2 two" bit - in fact, the whole sentence might need work).

If a collector is disabled, but you receive a frame, it gets "tossed". Frames sent to a distributor for which its ports are disabled, the frame is "tossed". Geoff T - do we need to add counters for "frames which are silently discarded". We will revisit this when we do the management stuff.

The same key value should, initially, be assigned to all ports which are capable of being aggregated. Tony to change the note in section 43.2.11.4.2 to reflect this.

S.Haddock - confusion over the types of collector (is it per port or what?). Tony to fix text in 43.2.11.5.1 to alleviate confusion.

Break for lunch 12:00

Afternoon Session:

43.3.1 Change note to normal text.

In the LACPDU, proposal to pad to make fields be on a 32-bit field. Resolution to make TLV values on 32-bit boundaries (as they're the interesting bits). 3 bytes of padding after Actor_State and 3 bytes of padding after Partner_State.

Value of “N” for the number of reserved bytes we need to come back to. General consensus that N be kept relatively small.

Figure 43.7 - Bit encoding. Discussion on the bit ordering - it is compatible with 802.3. Suggestion to add an arrow or something to indicate the direction in which it is transmitted (i.e. which bit hits the wire first).

S.Haddock - 43.3.3 LACPDUs are processed “sequentially”. Can more than one machine run in parallel? Yes, this refers to a single state machine. Suggestion to strike the first sentence, and simply state the operation of the state machines.

Figure 43.8 - NTT can be emitted by any of the state machines; MUX, Periodic, Match and Selection - but the formal descriptions don’t show this. Tony to fix.

Peter S. suggests that the diagram (43.3) does not show management interaction. Tony to add a big arrow to indicate that management can change anything in this diagram.

43.3.5 Suggestion to put tolerances on the variables. As the timer values are not very sensitive, it is seen as unnecessary to do this.

David Law pointed out that there is no definition (other than in the notes) for the value of a timer tick. Tony to fix.

Figure 43.9 How does it signal new information? Info Expired is a shared variable and goes to TRUE at initialization and FALSE after the first PDU. Other state machines detect changes by seeing that the variables they’re looking at have changed.

Note: there are a number of variables in the list which are no longer used and will be removed.

Ben Brown - Figure 43.10 - the state machines are “ugly”. Action to tidy them up and make them “pretty” if Tony really must. Translation process - currently there are two timers, which will be rationalized to one in a future draft.

43.3.11 Match State machine - Tony will add a diagram showing the state machine. (There is an editor’s note to this effect).

43.3.12.1 Peter S. - LAGIDs match, but the port cannot be aggregated (indicated by the individual bit). However, 43.3.12.1 says that ports which cannot be aggregated will have different keys. So, what if the keys are the same for two ports, but the individual bits are set? How do you distinguish this on management? Tony to fix up the text to explain the use of Link Agg Ids and the individual bit.

Page 42 section 43.3.12.2 (a) New Information is signaled by the receive machine. No it doesn’t - it just sets the new information and its up to the other machines to read their variables and know that their values have changed.

Figure 43.12 - Action to tidy the state machine diagram.

Side note: Tony said that he has an intention to have a churn machine at the other end of the link too.

Change of speaker to Rich for the Flush Protocol

43.4.1 Add a pointer to the annex which recommends when it is a good idea to use Flush.

Marker protocol frames is already aligned to 12 bytes (32-bit aligned therefore), as nobody wants to change this.

The reserved N bytes will be defined at some later stage. It does not need to be the same as the Link Aggregation control frames. However, Rich stated that his Grandmother said that, as with chicken soup, it "couldn't hurt". To be resolved.

Figure 43A-1 S.Haddock - Terms such as Server are not defined in 802.3. However, as the text is informative, definitions are not required. May be worth adding something to say DA/SA distribution algorithm is useless for example D in this figure. Tony pointed out though that there is some text that explains this.

Requirements for Slow Protocols - discussion added as Annex 43B.

Martin B. - 43B.2 Protocol transmission characteristics. The text currently doesn't say whether this is per port or system or what? The text to be updated to say "per MAC".

Would you ever want to tag one of these frames? Discussion then moved to "Ingress and Egress rules for PDUs" and "what happens when you're blocked by STP". This is 802.3 and it doesn't care about higher layers, and no, the frames are not to be tagged. The frames are as defined.

Do we want to limit the frame size for Slow Protocols? Discussion deferred until we decide the value of the Ns.

Rich reiterated that the Ethertype for frames will be new, and the DA will be taken from the reserved list in 802.1. The PICS will specify what parts of the frame will have to be tested in order to determine that it is a LA control frame or a Flush.

Action item: Tony to make a formal request to 802.1 to get a reserved address.

Action item: Geoff T. to get an Ethertype for Link Aggregation.

Rich to add in 43B a note that the multicast address picked is for link constrained Slow Protocols and it is from the reserved 802.1 addresses and will not pass through a bridge.

There will be a PICS for Annex 43B and will be part of clause 43's PICS, subject to checking with IEEE protocol in doing these kind of things. Ben Brown has a precedent for this - the PICS is with the Annex.

The Value of N - Bytes per Slow Protocol and LACPDU Frames

128 bytes for LACP and Flush.
Slow Protocols - recommend as 128 in the clause 43 annex.

Wednesday 20th January - Link Aggregation

Tony Jeffree - Management for Link Aggregation

Presentation source material and document is on the IEEE LAG web page.

Summarizes changes to clause 30 of 802.3. To an extent, the managed objects are based on a proposal from Cisco from their own proprietary MIB.

Bob Grow - Precedent for using SNMP RFCs in 802? Rich pointed out some previous 802 work and Tony referred to some 802.1 work. There will be a clause 30 annex for GDMO and another for SNMP.

Tony showed the proposed object structure, and Rich pointed out that objects were needed for Flush too. Current objects include the Aggregator, Port, LACPStatistics and LACPDebugInformation. The names will change to fit in with 802.3.

Every aggregator will have an ifTable entry.

Counters - Geoff T - "What about resetting counters?". Feeling is, back by Allan Chambers (quoting IETF), that resetting is a bad idea because statistics between the aggregator and its members will be disjoint. Also moving a port from one aggregator to another may cause problems too. Tony says that the ports should feed deltas to the aggregator layer and not absolute numbers.

Some discussion on the support for link up/down traps. Views are that it should be optionally supported.

Table entries - actor operational and admin key values - this is because you may have a 10/100 port (for example) which has an admin key value, but of course, if it negotiated to a speed incompatible with the rest of the ports in the aggregation, the actual key given to the port may be changed.

Joris - "Is there a more efficient way of finding the ports given an aggregator? " Right now you have to dump the entire port table and look at the keys to work it out. Tony acknowledged this would be a good thing to add.

Aggregation Ids for some ports may be null; when a device has fewer aggregators than ports and all of the current aggregators are currently aggregating to other devices, for example.

Debate took place regarding Slow Protocols. Resolution to crisply define the behavior for link constrained unicast and multicast frames. This includes statistical behavior for correctly and incorrectly formed frames. Slow Protocol frames must be validated before action is taken on them (New Slow Protocols which are not supported on a specific box, for example). If a port gets a Slow Protocol frame but it does not have a client for the subtype, then the frame is thrown away.

LACP Debug "Last Rx Time" - this will be based on the SNMP sysUpTime. Paul Congdon pointed out that we need to specify the dependencies of the 802.3ad MIB on other RFCs. Annex F of 802.3 (1998) already has a definition for this, David Law found later.

Long discussion on MIBII interface stats. OutDiscards - at the aggregator level to count frames discarded by the distribution algorithm? Only views expressed were to leave it out.

Mick Seaman pointed out that there should be enough information for a "smart" user to figure out what's going on - and to a "dumb" user it appears to be as before. Tony to fix up the document to provide a guide on how to get this kind of information.

For debugging purposes - an extra Churn machine should be added to detect far end Churn.

What should be mandatory and optional? Stuff required for Link Aggregation configuration is mandatory. Counters and debug information relating to LACP will be optional.

Jeff (IBM) / Martin B - Should Aggregator Link Up/Down be optional or mandatory. After some discussion: Mandatory.

Joris - how do I find the members of a aggregator using management? This can be solved in SNMP using some kind of table. Tony to add a MIB to solve this. There is already a port-to-aggregation table, Tony's adding the aggregation to port table.

Jeff (IBM) / Joris - What should the speed value be in the ifTable? Currently the proposal says "Zero", but the new idea is just sum the speeds of number of active links in an aggregation.

Tom D - "Do we need a need a variable indicating how successful an aggregate was in configuring?" After some discussion it was decided that there was enough information already in the management objects proposed to do everything needed.

Afternoon Session

Change of speaker to Mick Seaman who presented his paper on Constrained Aggregations (or Avoiding Dynamic Keys). The paper is available from the LAG Web server.

The proposal includes the concept of port priority so that a device can express which ports are its preferred ports for placing into an aggregation. He also pointed out the possible need of a system priority. This feature determines which system makes the decision on which ports to include and overrides the default case of using the lowest System ID.

The group feeling is that this optimization is good and that it should be included.

It will be included in the next draft after S.Haddock did a straw poll and nobody objected. A motion to include it was deemed unnecessary because the draft is not yet adopted by the group.

Schedule

S.Haddock then outlined the timeline that is voted on below:

1. February 8th- send draft 1.0 for Task Force Ballot by email. Copy the document to Working Group with notification that we may request a WG Ballot following the March meeting.
2. February 26th - close of ballot.
3. March 3rd - notification to Working Group as to whether we anticipate a WG Ballot request in March.

Motion to create draft 1.0 moved by Bill Q., seconded Tom D. "That the task force instructs the editors to incorporate the changes agreed to at this meeting and other editorial changes as necessary to draft 0.1 and create 1.0."

Vote: Yes: 34 No:0 Abstain:1

Motion carried.

Motion to adopt the time line moved by: Tom D. Seconded: Tony J.

Vote: Yes:34 No:0 Abstain:1

Motion carried.

Next Meeting

Next meeting in March - request will be made for a meeting room at late on Sunday afternoon to resolve editorial comments.

Geoff T: Call for interest in 10Gb/s Ethernet at the next meeting. S.Haddock asked that the meetings do not overlap with 802.3ad. There has been no request to the chair of 802.3 to increase Ethernet frame size.

Patents (Geoff T):

Geoff read the IEEE text asking parties to submit applicable patent information to the IEEE.

Meeting Planning

Interim meeting suggested for August/September timeframe. Possibility it could be hosted by 3Com in the UK in York (North East of England). Proposed date in September some time.