The need for timeout in ILT and AN

Comments #131, 184, 543, 545

Jeff Slavick

Supporters

The need for timeout in ILT and AN

In D1.3 Clause 73 the link_inhibit_timer for the new 802.3dj rates is set as TBD.

In D1.3 ILT process in 178B states there is no timeout for this process.

Comments 131, 184, 543, 545 are all related to addressing the TBD.

Cl **73** *SC* **73.10.2** *P* **130** *L* **16** # 131

Slavick, Jeff — Broadcom

Comment Type **TR** *Comment Status* **X**

TBD needs to be filled in.

SuggestedRemedy

Set link fail inhibit timer to be 15 to 15.1s

Cl **73** *SC* **73.10.2** *P* **130** *L* **16** # 184

Brown, Matt — Alphawave Semi

Comment Type **T** *Comment Status* **X**

Value for link_fail_inhibit_timer is TBD. Need value.

SuggestedRemedy

Expect a contribution with proposals.

Cl **73** *SC* **73.10.2** *P* **130** *L* **15** # 545

Dawe, Piers — Nvidia

Comment Type **TR** *Comment Status* **X**

According to 178B.14.2.1, there should be no time limit

SuggestedRemedy

Change the two "TBD" to infinity

Cl **178B** *SC* **178B.14.2.1** *P* **783** *L* **22** # 543

Dawe, Piers — Nvidia

Comment Type **TR** *Comment Status* **X**

This says "There is no specified time limit for the ILT protocol", which is misleading because it seems the Clause 73 link_fail_inhibit_timer will override it.

SuggestedRemedy

As it seems the intention is that there should be no time limit, and this is unlike e.g. 50GBASE-CR and 100GBASE-CR1, refer to Table 73-7 in 73.10.2 and say that link_fail_inhibit_timer does not apply at 200G/lane. In Table 73-7 in 73.10.2, set link_fail_inhibit_timer to infinite.

Link Training (LT) and Auto Negotiation (AN)

Clause 72 defined the NRZ Link training process and Clause 73 defines Electrical Auto-Neg (AN). Both contain a timeout timer that limits/watches out for failed init processes. If the path is not up before AN times out, then the AN process begins all over again.

Clause 136 defines the PAM4 link training process for 50/100Gbps serial for electrical link segments and included a timeout timer as well.

Prior to 802.3dj, the link training process is ONLY run between the PMDs and any AUI between the MAC and the PMD is left to the system provider to initialize prior to/during the LT execution across the medium.

Optical PMDs don’t perform either AN or Link Training prior to 802.3dj

Inter-sublayer Link Training (ILT) being added in 802.3dj

ILT is leveraged off the Clause 136 PAM4 link training process, but has removed the timeout timer.

ILT adds a mode of operation over Optical link segments.

The intent of ILT is to support executing link training across all inter-sublayer links (ISL) between the MACs, including both electrical and optical. This should help facilitate end-to-end initialization processes.

The Scenarios

There are several types of paths that can occur in systems. Ones with:

1. Optical PHYs that do not run AN (usually mix of electrical & optical ISLs)
2. Electrical PHYs that do not run AN (with 1 or more electrical ISLs)
3. Electrical PHYs that run AN (with 1 or more electrical ISLs)

We need to decide if a timeout function(s) is necessary for each of these situations.

Paths with Optical PHYs

These paths have not specified a timeout in the past. System vendors have dealt with the start-up situations for how to establish the path and determined their own behaviors on how/when to start and restart the initialization sequence.

Paths with Electrical PHYs not running AN

Many systems are fixed rate operation or support forcing a particular rate.

At NRZ rates these paths typically bypass both AN and LT as there was not a method to align the LT timeout timer on both sides of the link segment.

At PAM4 rates the LT timeout timer starts when both sides indicate they have achieved frame lock during the LT process. This allows the use of LT outside of the AN for the PMDs at these rates. As a result the allowed adaptation time bound the startup timing in this mode.

How the AUI links are initialized is system dependent, but had to occur prior to PMD startup to ensure PMD couldn’t finish first.

So for NRZ rates no hardware timer timeout existed, but for PAM4 the time to path up was bounded by the LT timeout.

Electrical PHYs running AN

One of the features AN brings is its aligning when the startup process begins on both ends of the path. Since both sides start their initialization processes for the negotiated rate “at the same time” we are able to bound how long to wait for a link-up before declaring “something” went wrong. Specifying this time provides benefits for interop, system testing, design specifications, etc.

Note: The management system on each side of the link has to configure all parts of the path within the AN timeout duration to avoid a failed link-up process.

Why specify a timeout?

Bounds design choices implementers must follow for things like:

- Link adaptation time
- PCS frame lock durations
- Management access bandwidth
- System initialization times

If we remove this guidance implementers will have to “come up” their own specifications, which could have interop implications.

What timeouts should we specify

Electrical paths running AN:

AN synchronizes the startup time of both sides. AN requires management to be involved to swap the configuration of the systems to the HCD rate. ILT reduces the management overhead needed to initialize multiple ISLs. Maintaining a timeout for the path to be established makes sense to specify as there is a known point in time on both sides of the path when this timer begins. By specifying an AN timeout this means we should also specify a TRAINING duration timeout so there is time allocated for different parts of the startup sequence.

Paths not using AN:

Since each side of the path is often independently controlled and there is no mechanism to synchronize the startup time between ISLs, it makes sense to NOT have a timeout for achieving path up. However, specifying a TRAINING duration limit during ILT should be done to appropriately constrain implementations.

Solution A – no FSM change

- In 178.8.9, 179.8.9, 176C.4.3.1, 176D.7.6, 180.5.12, 181.5.12, 182.5.12, 183.5.12
  - Add sentence “To support reasonable link-up time durations, the time from exiting SEND_TRAINING to entering ISL_READY should be no more than X seconds (see Figure 178B-8).”

- Set the AN link_inhibit_timer to be 3s to 3.1s longer than the expected time local_rx_ready assertion for 200Gbps lanes.

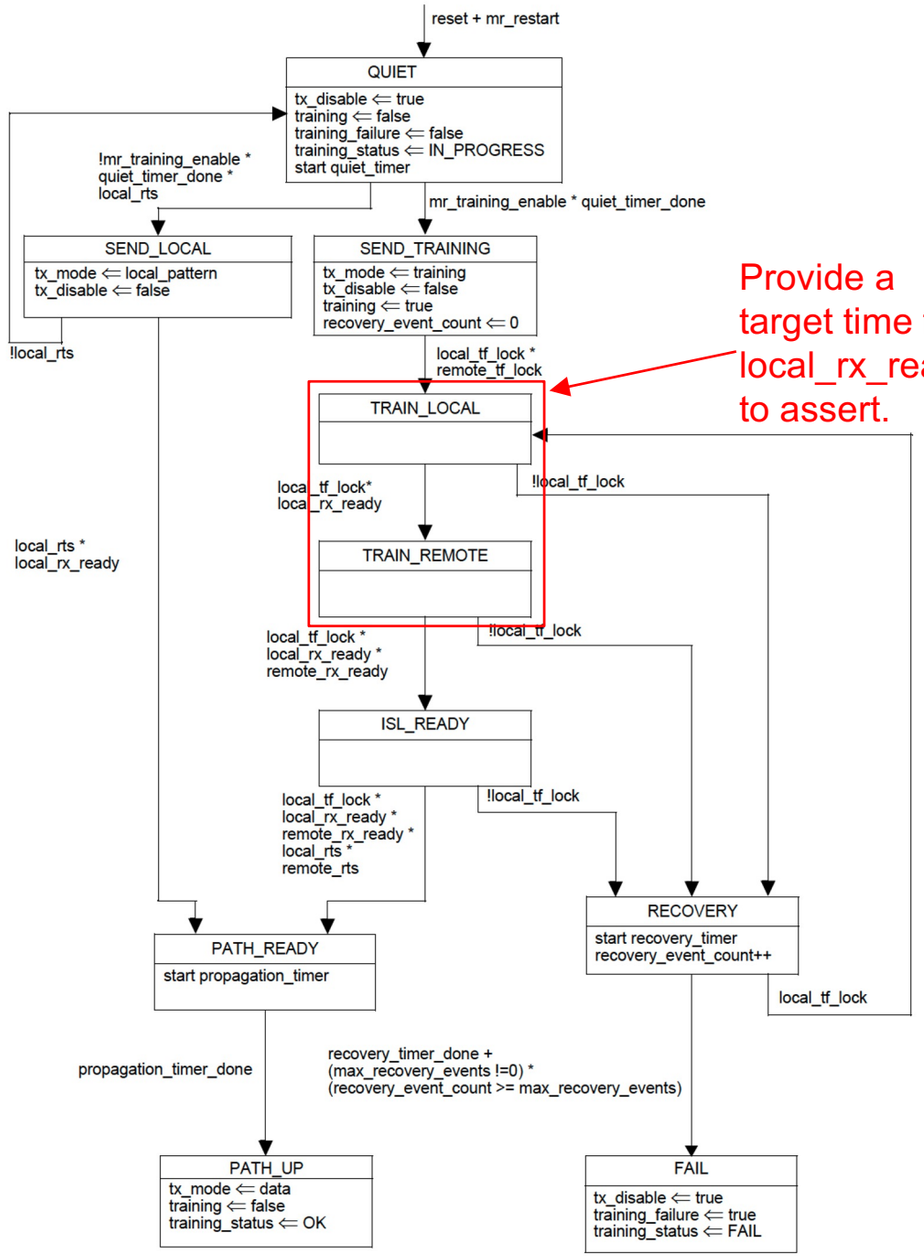


Figure 178B–8—Training control state diagram

Solution B – FSM update

- In 178B.14.3.3 define max_wait_timer
  - This timer sets the limit for how long the ILT startup protocol is allowed to adapt a link segment. If the timer expires before the ISL_READY state is reached, then a failure to train is indicated. The terminal count of max_wait_timer is set by the AUI component or PMD.
- Change 178B as shown.
- In 178.8.9, 179.8.9, 176C.4.3.1, 176D.7.6, 180.5.12, 181.5.12, 182.5.12, 183.5.12:
  - Add “The duration of the max_wait_timer is X seconds +/- X*1ms.”
- Set the AN link_inihbit_timer to be 3s to 3.1s longer than the max_wait_timer.

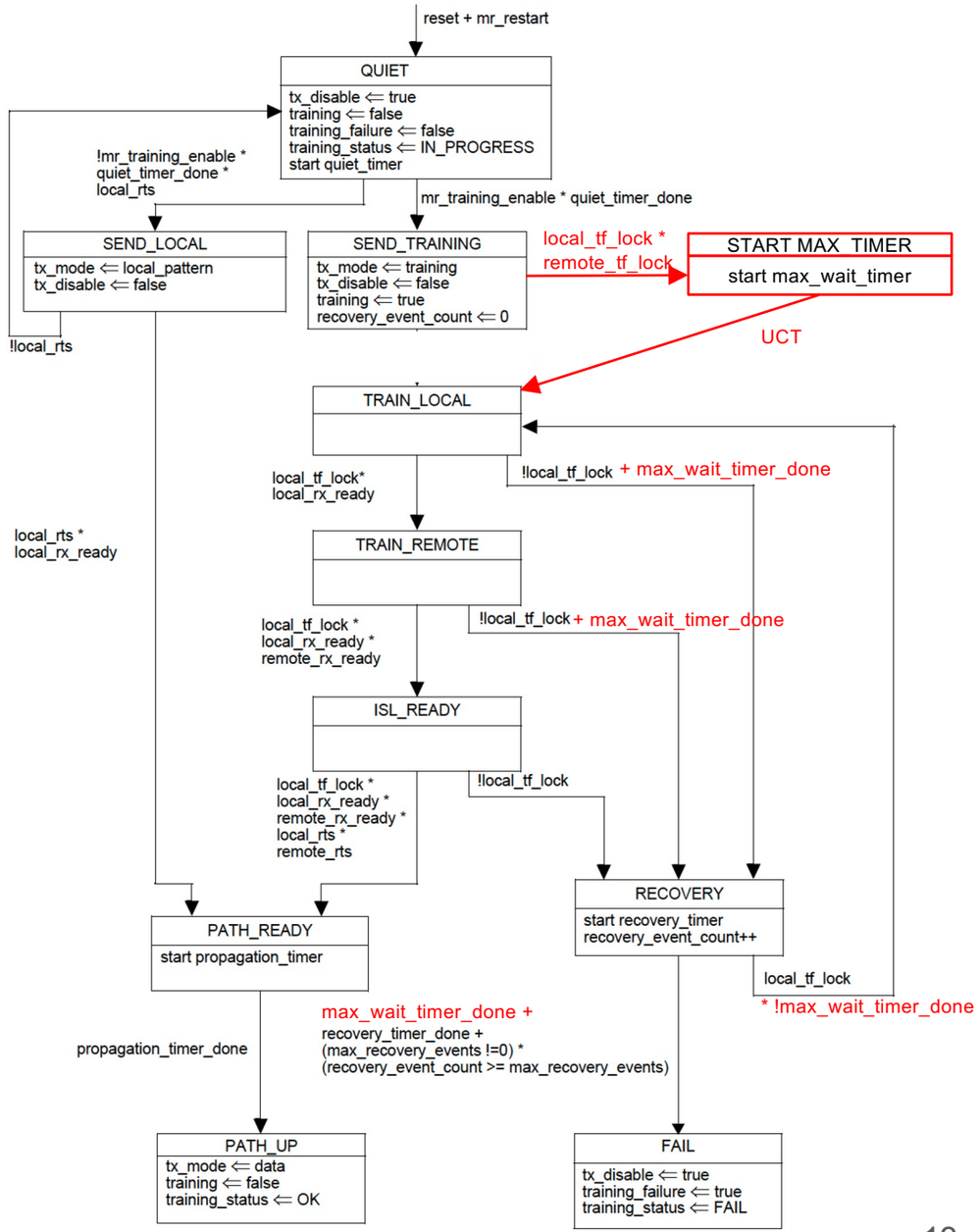


Figure 178B–8—Training control state diagram

Summary

Timeouts provide a way to communicate expected timelines behaviors of systems processes. These durations helps designers at the IP, chip, software and system levels to all be on a similar page for how long their piece of the puzzle can take and how much resources needs to be allocated to adhere to portion of timeline.

Modifications to 178B-8

Diagram adds a state between SEND_TRAINING and TRAIN_LOCAL to start the max-wait_timer.

Adds “+ max_wait_timer_done” to the RECOVERY transition from TRAIN_LOCAL and TRAIN_REMOTE (but not ISL_READY).

Adds “+ max_wait_timer_done” from RECOVERY to FAIL and “* !max_wait_timer_done” from RECOVERY to TRAIN_LOCAL

Also, quiet_timer_done needs to have a 0 time duration for AN mode of operation

Optical PHYs

Optical links have never specified a timeout for establishing the link. System vendors have dealt with the start-up situations for how to establish the links and determined their own behaviors on how/when to restart the links.

The typical flow being:

1. Detect Module existence
2. Query Modules speed
3. Configure Host to that speed
    a. including rate/module specific settings at host and possibly module
4. PCS traffic sent over AUI to module
5. Wait for Link up

Optical PHYs with ILT

A flow could be:

1. Detect Module existence
2. Query Modules speed
3. Configure Host to that speed
4. Start ILT on Host
5. Start ILT on Module
6. Wait for Link up

Non-AN electrical PHYs

Many systems operate in or support fixed rate startup. For NRZ rates LT was typically skipped for this mode of operation, but for PAM4 rates LT is often utilized. The Cl 136 PAM4 LT process synchronizes the timeout timers for the but only for each individual link segment, no coordination of cascaded segments.

The typical flow being:

1. Detect cable/bp connection is present
2. Configure the rate into all link segments
    a. Often programming any AUI with characterized settings for that rate
3. Start LT in the PMD
4. Wait for Link up

Non-AN electrical PHYs using ILT

A flow could be:

1. Detect cable/bp present
2. Configure rate into all link segments
3. Start ILT on all link segments
4. Wait for Link up

AN electrical PHY

Remember that AN is run out of the PMD. So steps 3 & 4 below get “tricky” if a PHY chip is used.

The typical flow being:

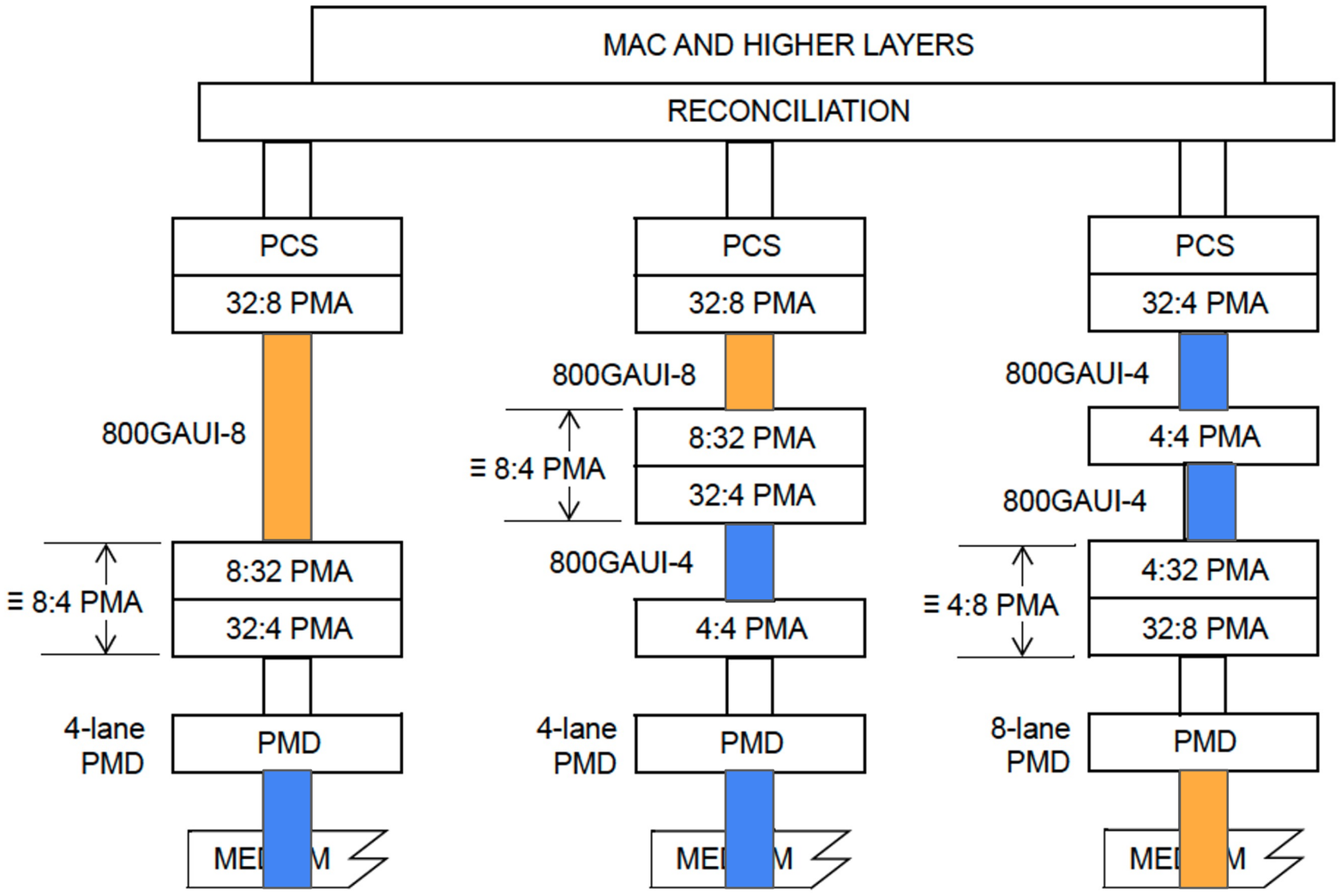
1. Configure Port for AN
2. Wait for AN result
3. Within 20ms of AN determining the HCD the management entity must configure negotiate rate into PMD SerDes and start LT for most rates
4. Prior to LT finishing management entity must configure any AUI components to negotiated rate and configure the PCS to the proper rate.
5. Wait for Link up

AN electrical PHY

A flow could be:

1. Configure port for AN
2. Wait for AN result
3. Configure each segment for the HCD rate and start ILT
4. Wait for LinkUp

ILT


MAC AND HIGHER LAYERS
RECONCILIATION
PCS
32:8 PMA
800GAUI-8
≡ 8:4 PMA
8:32 PMA
32:4 PMA
4-lane PMD
PMD
PCS
32:8 PMA
800GAUI-8
≡ 8:4 PMA
8:32 PMA
32:4 PMA
800GAUI-4
4:4 PMA
4-lane PMD
PMD
PCS
32:4 PMA
800GAUI-4
4:4 PMA
800GAUI-4
≡ 4:8 PMA
4:32 PMA
32:8 PMA
8-lane PMD
PMD