Thoughts on optical automatic link configuration

Matt Brown, Alphawave Semi
Ali Ghiasi, Ghiasi Quantum
Mike Dudek, Marvell
Kent Lusted, Intel

Contributors

- Adam Healey, Broadcom

Supporters

- Chris Cole, Quintessent
- Tony Chan Carusone, Alphawave Semi
- Roberto Rodes, Coherent

Introduction

- There has been interest expressed in providing a mechanism to automate the selection of PHY type and/or PHY mode.
- Auto-negotiation methodology developed for Ethernet electrical backplane and copper cable would be a good candidate as basis.
- Although we adopted no optical PHY baselines, the proposal encompasses any of the proposals discussed so far.
- Complementary presentation ghiasi_3dj_01_2311 proposes a method for optical transmitter adaptation.

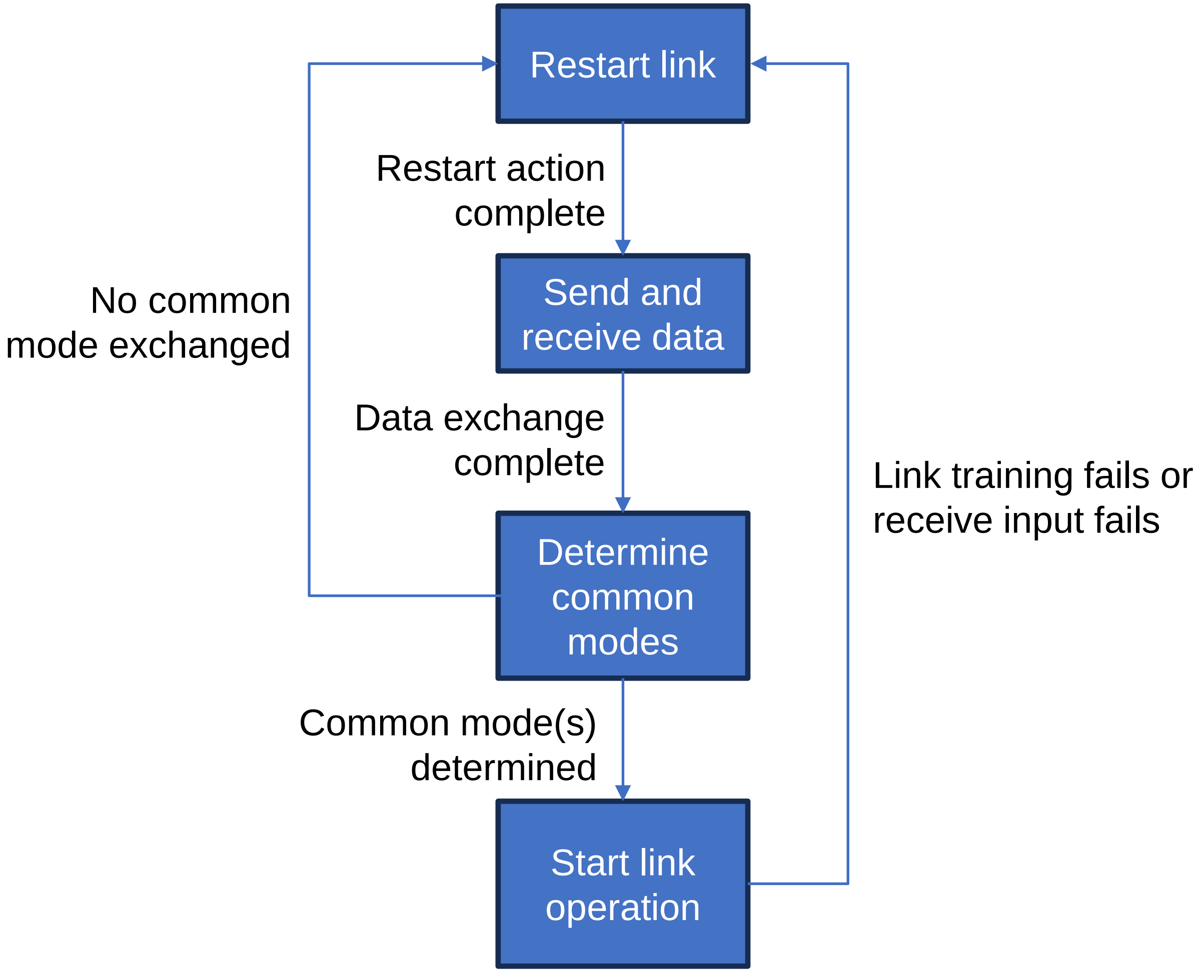
Purpose

- Establish common configuration (or lack thereof) for link partners at each end of an optical fiber.
- Examples…
  - Determine common PHY type
    - both ends use PHY with inner FEC or both use PHY without inner FEC
  - Common inner FEC mode
    - both ends use inner FEC or both ends disable/bypass inner FEC
  - For inner FEC case
    - both ends select with convolution interleaver or both ends do not use

Useful features

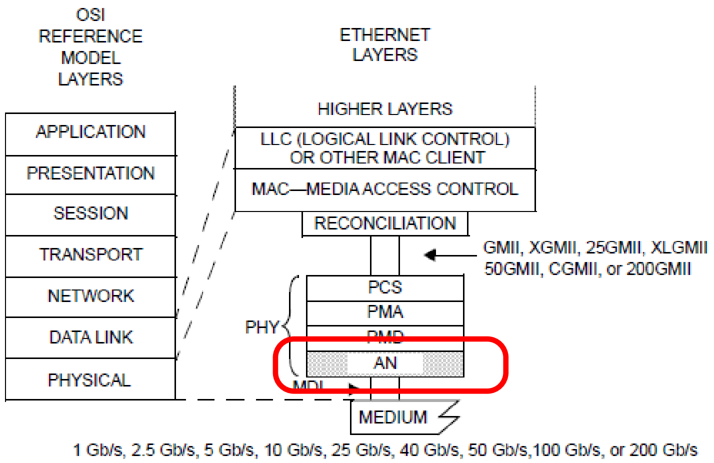
- Means to exchange information between link partners
  - e.g., signaling and data structure
- Means to initiate automatic configuration
- Means to select which technologies to permit
- Means to select a common mode of operation
- Means to transition from automatic configuration to data mode
- State machines to coordinate the above features

High level state diagram



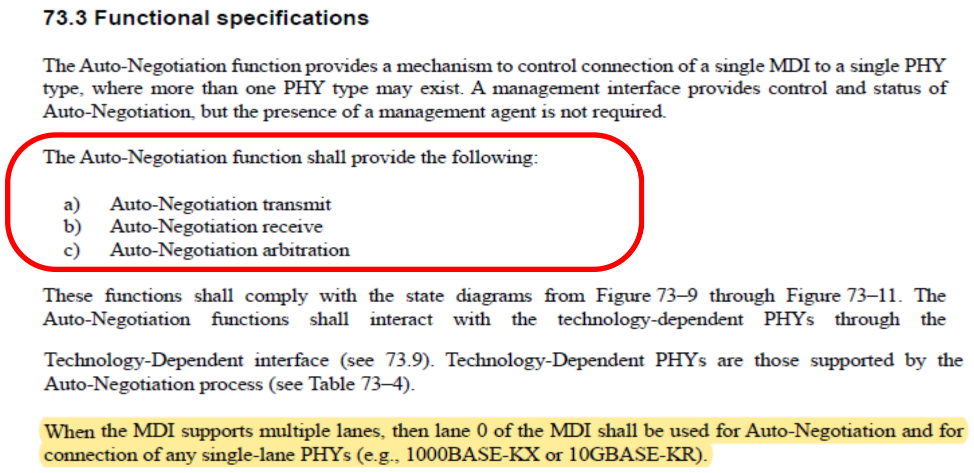
Summary of auto-negotiation defined for backplane and copper cable PHYs in Clause 73 of IEEE 802.3-2022 and IEEE Std 802.3ck-2022.

Clause 73 – AN general


OSI REFERENCE MODEL LAYERS
APPLICATION
PRESENTATION
SESSION
TRANSPORT
NETWORK
DATA LINK
PHYSICAL
ETHERNET LAYERS
HIGHER LAYERS
LLC (LOGICAL LINK CONTROL) OR OTHER MAC CLIENT
MAC—MEDIA ACCESS CONTROL
RECONCILIATION
GMII, XGMII, 25GMII, XLGMII 50GMII, CGMII, or 200GMII
PHY
PCS
PMA
PMD
AN
MDI
MEDIUM
1 Gb/s, 2.5 Gb/s, 5 Gb/s, 10 Gb/s, 25 Gb/s, 40 Gb/s, 50 Gb/s,100 Gb/s, or 200 Gb/s

200GMII = 200 Gb/s MEDIA INDEPENDENT INTERFACE
25GMII = 25 GIGABIT MEDIA INDEPENDENT INTERFACE
50GMII = 50 Gb/s MEDIA INDEPENDENT INTERFACE
AN = AUTO-NEGOTIATION
CGMII = 100 Gb/s MEDIA INDEPENDENT INTERFACE
GMII = GIGABIT MEDIA INDEPENDENT INTERFACE
MDI = MEDIUM DEPENDENT INTERFACE
MDI = MEDIUM DEPENDENT INTERFACE
PCS = PHYSICAL CODING SUBLAYER
PHY = PHYSICAL LAYER DEVICE
PMA = PHYSICAL MEDIUM ATTACHMENT
PMD = PHYSICAL MEDIUM DEPENDENT
XGMII = 10 GIGABIT MEDIA INDEPENDENT INTERFACE
XLGMII = 40 Gb/s MEDIA INDEPENDENT INTERFACE

Figure 73–1—Location of Auto-Negotiation function within the ISO/IEC OSI reference model



73.3 Functional specifications

The Auto-Negotiation function provides a mechanism to control connection of a single MDI to a single PHY type, where more than one PHY type may exist. A management interface provides control and status of Auto-Negotiation, but the presence of a management agent is not required.

The Auto-Negotiation function shall provide the following:

a) Auto-Negotiation transmit
b) Auto-Negotiation receive
c) Auto-Negotiation arbitration

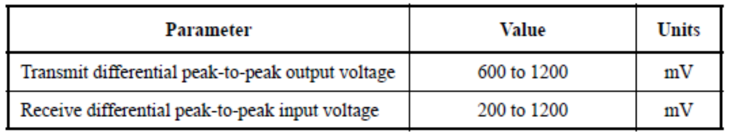
These functions shall comply with the state diagrams from Figure 73–9 through Figure 73–11. The Auto-Negotiation functions shall interact with the technology-dependent PHYs through the

Technology-Dependent interface (see 73.9). Technology-Dependent PHYs are those supported by the Auto-Negotiation process (see Table 73–4).

When the MDI supports multiple lanes, then lane 0 of the MDI shall be used for Auto-Negotiation and for connection of any single-lane PHYs (e.g., 1000BASE-KX or 10GBASE-KR).

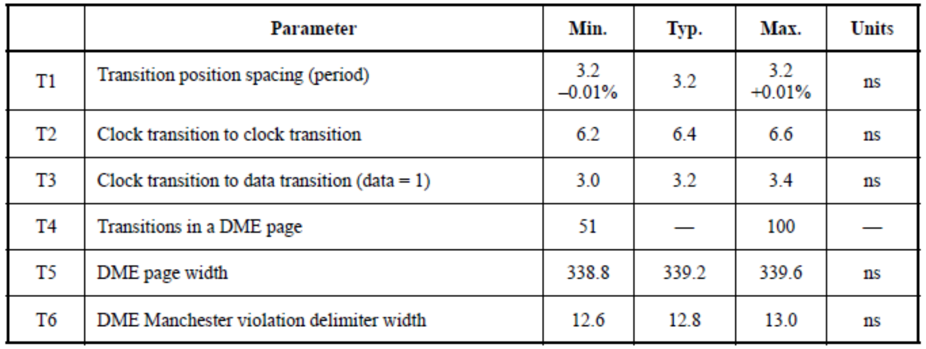
Clause 73 – AN signaling

Table 73–1—DME electrical characteristics



Parameter	Value	Units
Transmit differential peak-to-peak output voltage	600 to 1200	mV
Receive differential peak-to-peak input voltage	200 to 1200	mV

Table 73–2— DME page timing summary



	Parameter	Min.	Typ.	Max.	Units
T1	Transition position spacing (period)	3.2 –0.01%	3.2	3.2 +0.01%	ns
T2	Clock transition to clock transition	6.2	6.4	6.6	ns
T3	Clock transition to data transition (data = 1)	3.0	3.2	3.4	ns
T4	Transitions in a DME page	51	—	100	—
T5	DME page width	338.8	339.2	339.6	ns
T6	DME Manchester violation delimiter width	12.6	12.8	13.0	ns

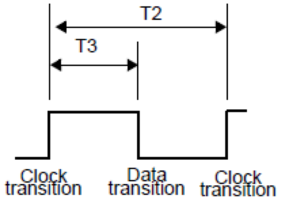


Figure 73–4—DME page transition timing

Given 3.2 ns width (T3) this is equivalent to NRZ signaling rate 312.5 MBd

The encoding of data using DME bits in an DME page is illustrated in Figure 73–3.

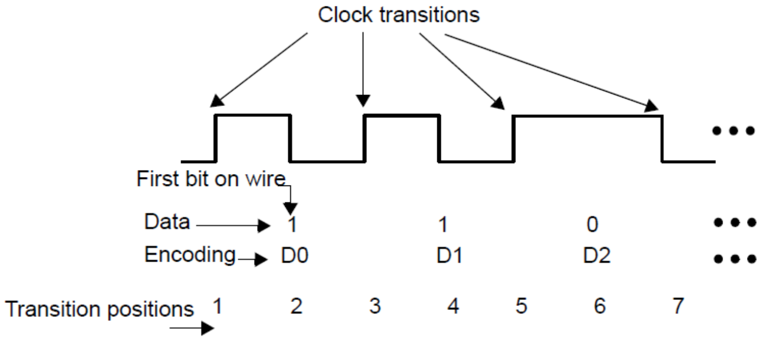


Figure 73–3—Data bit encoding within DME pages

73.5.3.1 Manchester violation delimiter

A violation is signaled as shown in Figure 73–5.

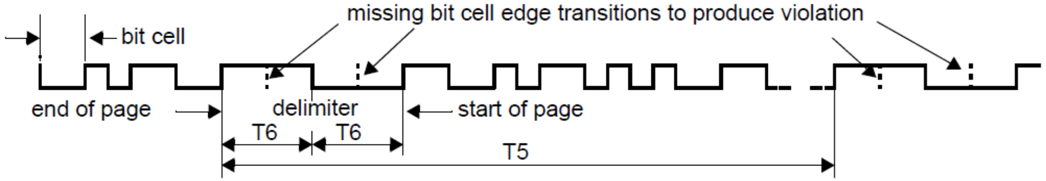
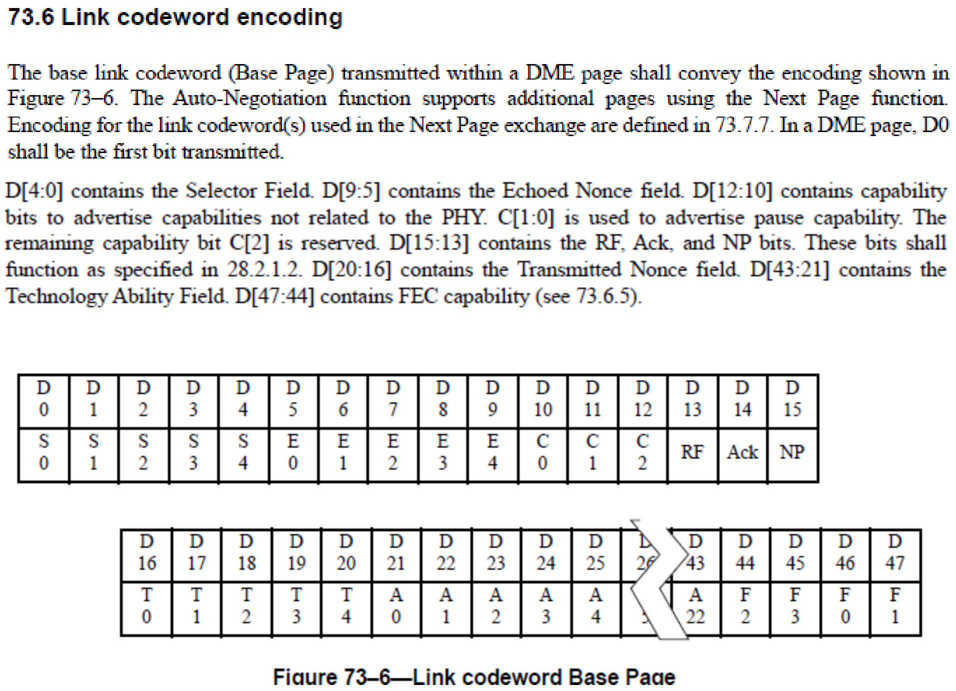


Figure 73–5—Manchester violation

Clause 73 – link codeword— encoding



73.6 Link codeword encoding

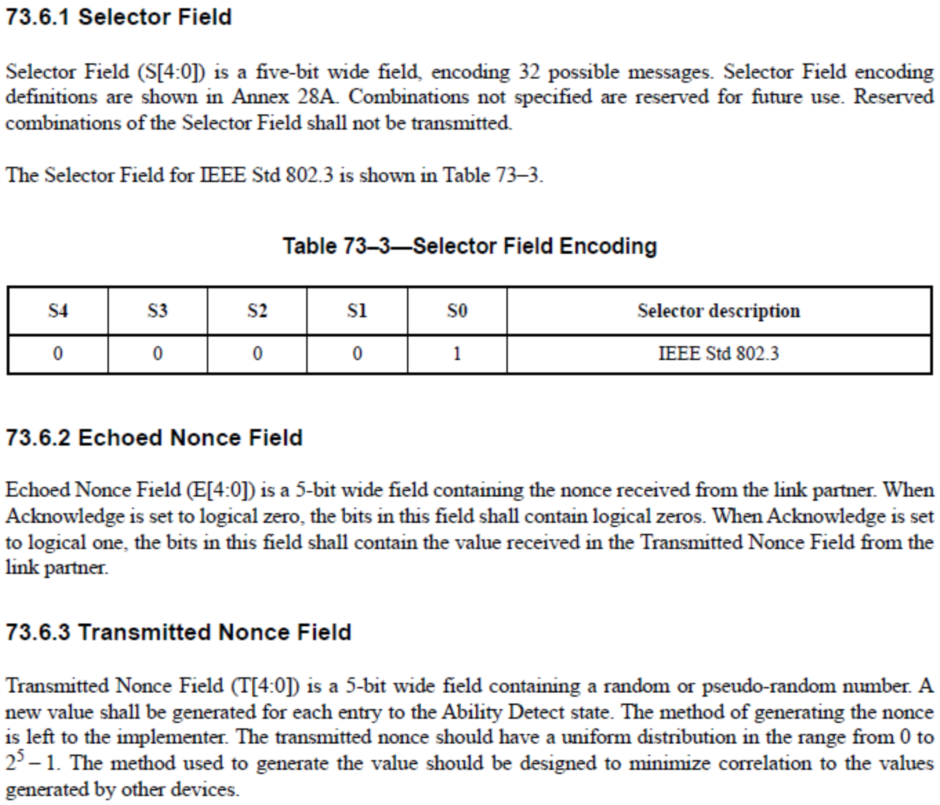
The base link codeword (Base Page) transmitted within a DME page shall convey the encoding shown in Figure 73–6. The Auto-Negotiation function supports additional pages using the Next Page function. Encoding for the link codeword(s) used in the Next Page exchange are defined in 73.7.7. In a DME page, D0 shall be the first bit transmitted.

D[4:0] contains the Selector Field. D[9:5] contains the Echoed Nonce field. D[12:10] contains capability bits to advertise capabilities not related to the PHY. C[1:0] is used to advertise pause capability. The remaining capability bit C[2] is reserved. D[15:13] contains the RF, Ack, and NP bits. These bits shall function as specified in 28.2.1.2. D[20:16] contains the Transmitted Nonce field. D[43:21] contains the Technology Ability Field. D[47:44] contains FEC capability (see 73.6.5).

D 0	D 1	D 2	D 3	D 4	D 5	D 6	D 7	D 8	D 9	D 10	D 11	D 12	D 13	D 14	D 15
S 0	S 1	S 2	S 3	S 4	E 0	E 1	E 2	E 3	E 4	C 0	C 1	C 2	RF	Ack	NP

D 16	D 17	D 18	D 19	D 20	D 21	D 22	D 23	D 24	D 25	D 26	D 43	D 44	D 45	D 46	D 47
T 0	T 1	T 2	T 3	T 4	A 0	A 1	A 2	A 3	A 4	5	A 22	F 2	F 3	F 0	F 1

Figure 73–6—Link codeword Base Page



73.6.1 Selector Field

Selector Field (S[4:0]) is a five-bit wide field, encoding 32 possible messages. Selector Field encoding definitions are shown in Annex 28A. Combinations not specified are reserved for future use. Reserved combinations of the Selector Field shall not be transmitted.

The Selector Field for IEEE Std 802.3 is shown in Table 73–3.

Table 73–3—Selector Field Encoding

S4	S3	S2	S1	S0	Selector description
0	0	0	0	1	IEEE Std 802.3

73.6.2 Echoed Nonce Field

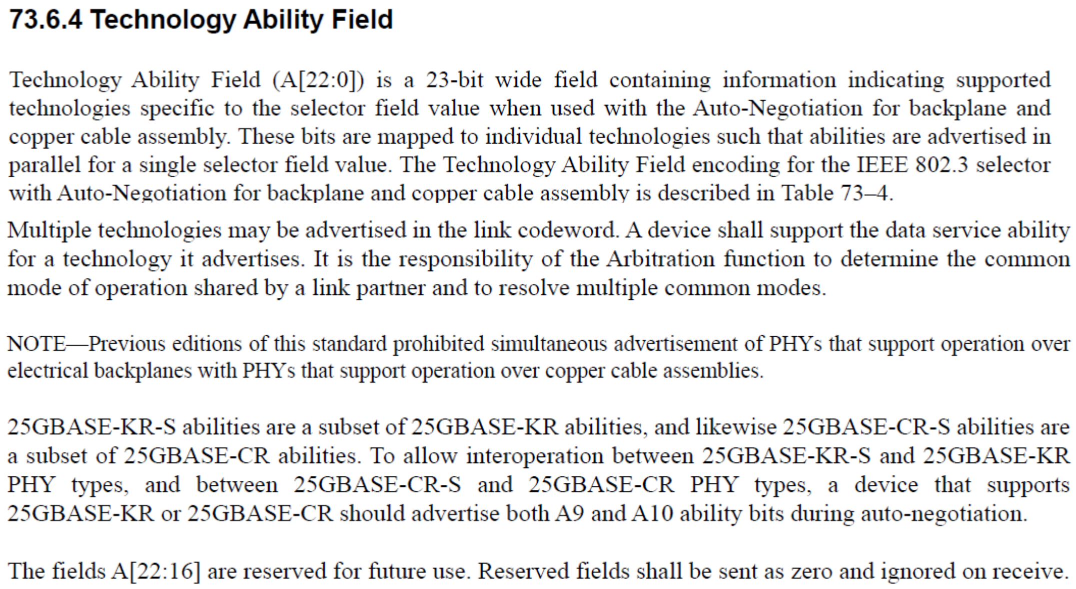
Echoed Nonce Field (E[4:0]) is a 5-bit wide field containing the nonce received from the link partner. When Acknowledge is set to logical zero, the bits in this field shall contain logical zeros. When Acknowledge is set to logical one, the bits in this field shall contain the value received in the Transmitted Nonce Field from the link partner.

73.6.3 Transmitted Nonce Field

Transmitted Nonce Field (T[4:0]) is a 5-bit wide field containing a random or pseudo-random number. A new value shall be generated for each entry to the Ability Detect state. The method of generating the nonce is left to the implementer. The transmitted nonce should have a uniform distribution in the range from 0 to $2^5 - 1$. The method used to generate the value should be designed to minimize correlation to the values generated by other devices.

Clause 73 – link codeword – technology abilities field

From 802.3-2022…



73.6.4 Technology Ability Field

Technology Ability Field (A[22:0]) is a 23-bit wide field containing information indicating supported technologies specific to the selector field value when used with the Auto-Negotiation for backplane and copper cable assembly. These bits are mapped to individual technologies such that abilities are advertised in parallel for a single selector field value. The Technology Ability Field encoding for the IEEE 802.3 selector with Auto-Negotiation for backplane and copper cable assembly is described in Table 73–4.

Multiple technologies may be advertised in the link codeword. A device shall support the data service ability for a technology it advertises. It is the responsibility of the Arbitration function to determine the common mode of operation shared by a link partner and to resolve multiple common modes.

NOTE—Previous editions of this standard prohibited simultaneous advertisement of PHYs that support operation over electrical backplanes with PHYs that support operation over copper cable assemblies.

25GBASE-KR-S abilities are a subset of 25GBASE-KR abilities, and likewise 25GBASE-CR-S abilities are a subset of 25GBASE-CR abilities. To allow interoperation between 25GBASE-KR-S and 25GBASE-KR PHY types, and between 25GBASE-CR-S and 25GBASE-CR PHY types, a device that supports 25GBASE-KR or 25GBASE-CR should advertise both A9 and A10 ability bits during auto-negotiation.

The fields A[22:16] are reserved for future use. Reserved fields shall be sent as zero and ignored on receive.

From 802.3-2022…

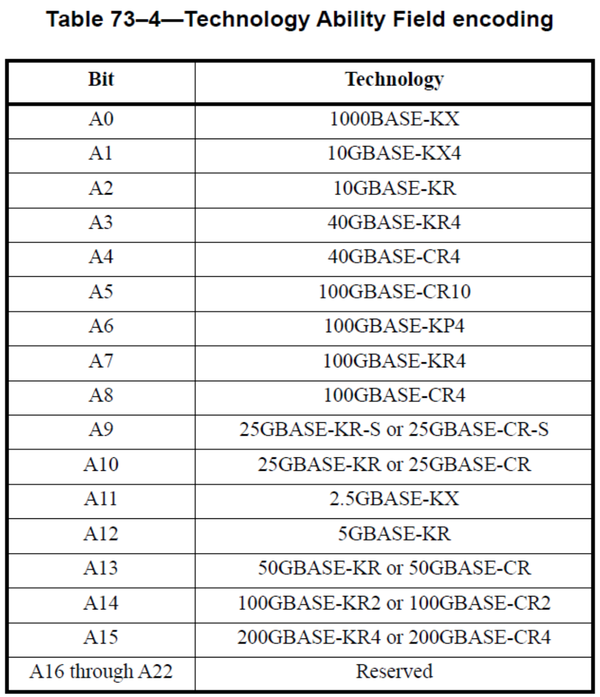


Table 73–4—Technology Ability Field encoding

Bit	Technology
A0	1000BASE-KX
A1	10GBASE-KX4
A2	10GBASE-KR
A3	40GBASE-KR4
A4	40GBASE-CR4
A5	100GBASE-CR10
A6	100GBASE-KP4
A7	100GBASE-KR4
A8	100GBASE-CR4
A9	25GBASE-KR-S or 25GBASE-CR-S
A10	25GBASE-KR or 25GBASE-CR
A11	2.5GBASE-KX
A12	5GBASE-KR
A13	50GBASE-KR or 50GBASE-CR
A14	100GBASE-KR2 or 100GBASE-CR2
A15	200GBASE-KR4 or 200GBASE-CR4
A16 through A22	Reserved

From 802.3ck-2022 (additions)…

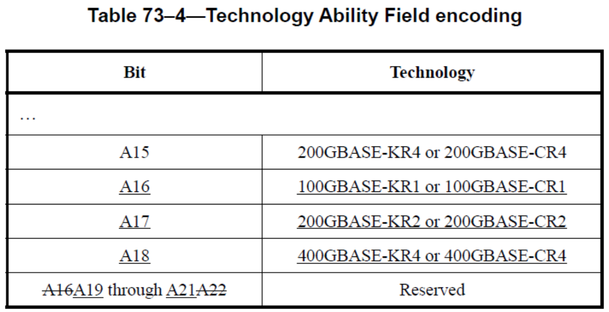
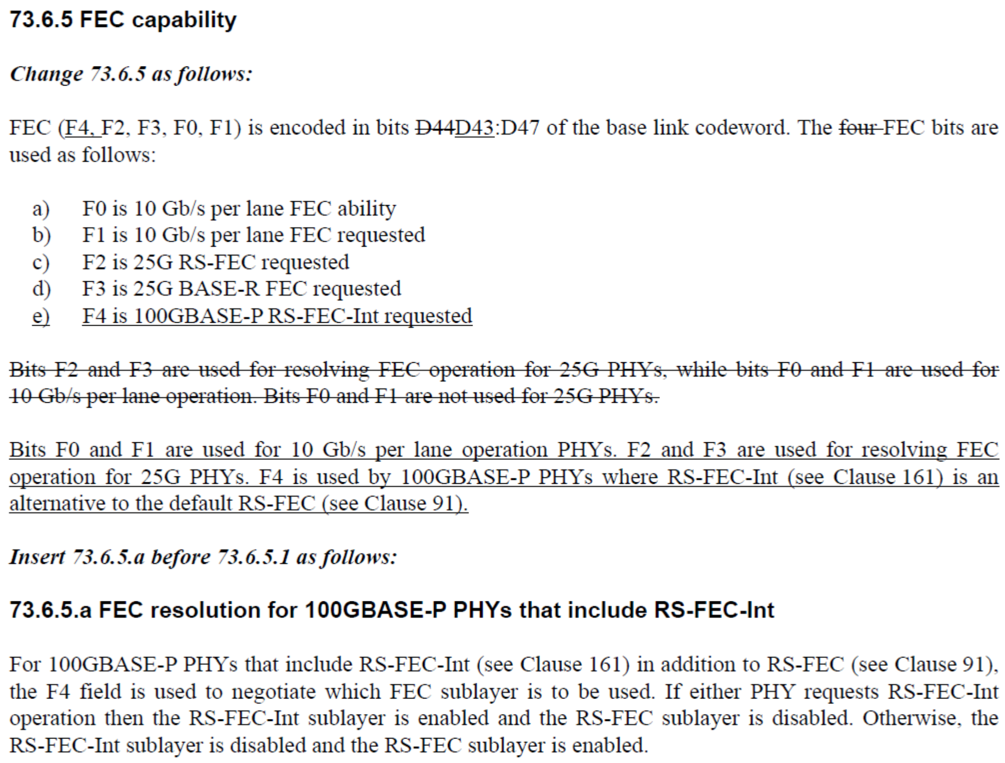


Table 73–4—Technology Ability Field encoding

Bit	Technology
…	
A15	200GBASE-KR4 or 200GBASE-CR4
A16	100GBASE-KR1 or 100GBASE-CR1
A17	200GBASE-KR2 or 200GBASE-CR2
A18	400GBASE-KR4 or 400GBASE-CR4
~~A16~~A19 through A21~~A22~~	Reserved

Clause 73 AN – Resolution of FEC type

From 802.3ck-2022…



73.6.5 FEC capability

Change 73.6.5 as follows:

FEC (F4, F2, F3, F0, F1) is encoded in bits ~~D44~~D43:D47 of the base link codeword. The ~~four~~ FEC bits are used as follows:

a) F0 is 10 Gb/s per lane FEC ability
b) F1 is 10 Gb/s per lane FEC requested
c) F2 is 25G RS-FEC requested
d) F3 is 25G BASE-R FEC requested
e) F4 is 100GBASE-P RS-FEC-Int requested

~~Bits F2 and F3 are used for resolving FEC operation for 25G PHYs, while bits F0 and F1 are used for 10 Gb/s per lane operation. Bits F0 and F1 are not used for 25G PHYs.~~

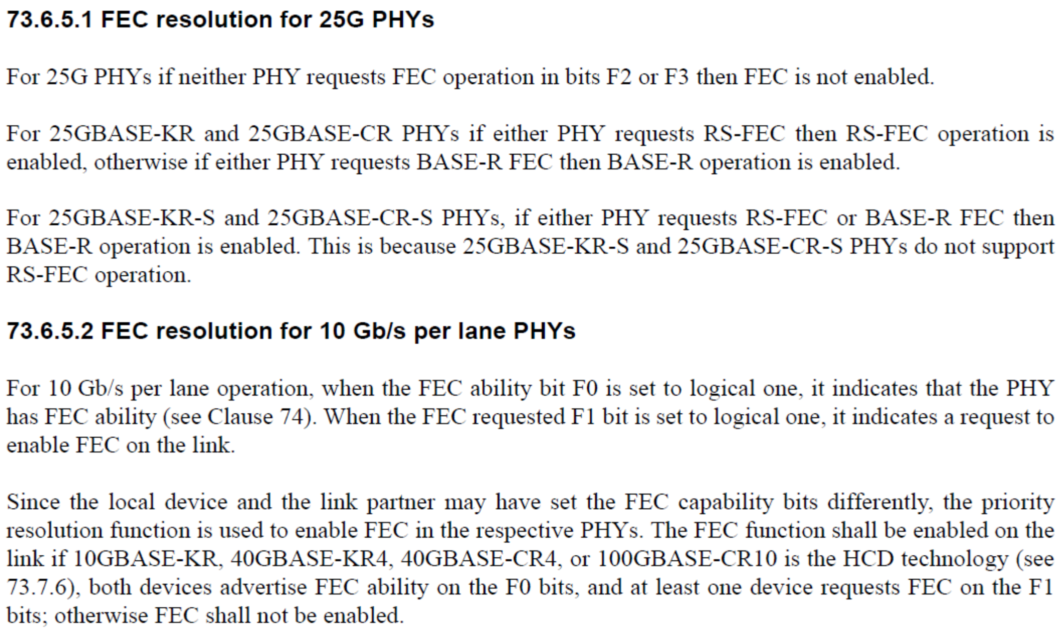
Bits F0 and F1 are used for 10 Gb/s per lane operation PHYs. F2 and F3 are used for resolving FEC operation for 25G PHYs. F4 is used by 100GBASE-P PHYs where RS-FEC-Int (see Clause 161) is an alternative to the default RS-FEC (see Clause 91).

Insert 73.6.5.a before 73.6.5.1 as follows:

73.6.5.a FEC resolution for 100GBASE-P PHYs that include RS-FEC-Int

For 100GBASE-P PHYs that include RS-FEC-Int (see Clause 161) in addition to RS-FEC (see Clause 91), the F4 field is used to negotiate which FEC sublayer is to be used. If either PHY requests RS-FEC-Int operation then the RS-FEC-Int sublayer is enabled and the RS-FEC sublayer is disabled. Otherwise, the RS-FEC-Int sublayer is disabled and the RS-FEC sublayer is enabled.

From 802.3-2022…



73.6.5.1 FEC resolution for 25G PHYs

For 25G PHYs if neither PHY requests FEC operation in bits F2 or F3 then FEC is not enabled.

For 25GBASE-KR and 25GBASE-CR PHYs if either PHY requests RS-FEC then RS-FEC operation is enabled, otherwise if either PHY requests BASE-R FEC then BASE-R operation is enabled.

For 25GBASE-KR-S and 25GBASE-CR-S PHYs, if either PHY requests RS-FEC or BASE-R FEC then BASE-R operation is enabled. This is because 25GBASE-KR-S and 25GBASE-CR-S PHYs do not support RS-FEC operation.

73.6.5.2 FEC resolution for 10 Gb/s per lane PHYs

For 10 Gb/s per lane operation, when the FEC ability bit F0 is set to logical one, it indicates that the PHY has FEC ability (see Clause 74). When the FEC requested F1 bit is set to logical one, it indicates a request to enable FEC on the link.

Since the local device and the link partner may have set the FEC capability bits differently, the priority resolution function is used to enable FEC in the respective PHYs. The FEC function shall be enabled on the link if 10GBASE-KR, 40GBASE-KR4, 40GBASE-CR4, or 100GBASE-CR10 is the HCD technology (see 73.7.6), both devices advertise FEC ability on the F0 bits, and at least one device requests FEC on the F1 bits; otherwise FEC shall not be enabled.

Clause 73 AN – Resolution of PHY type

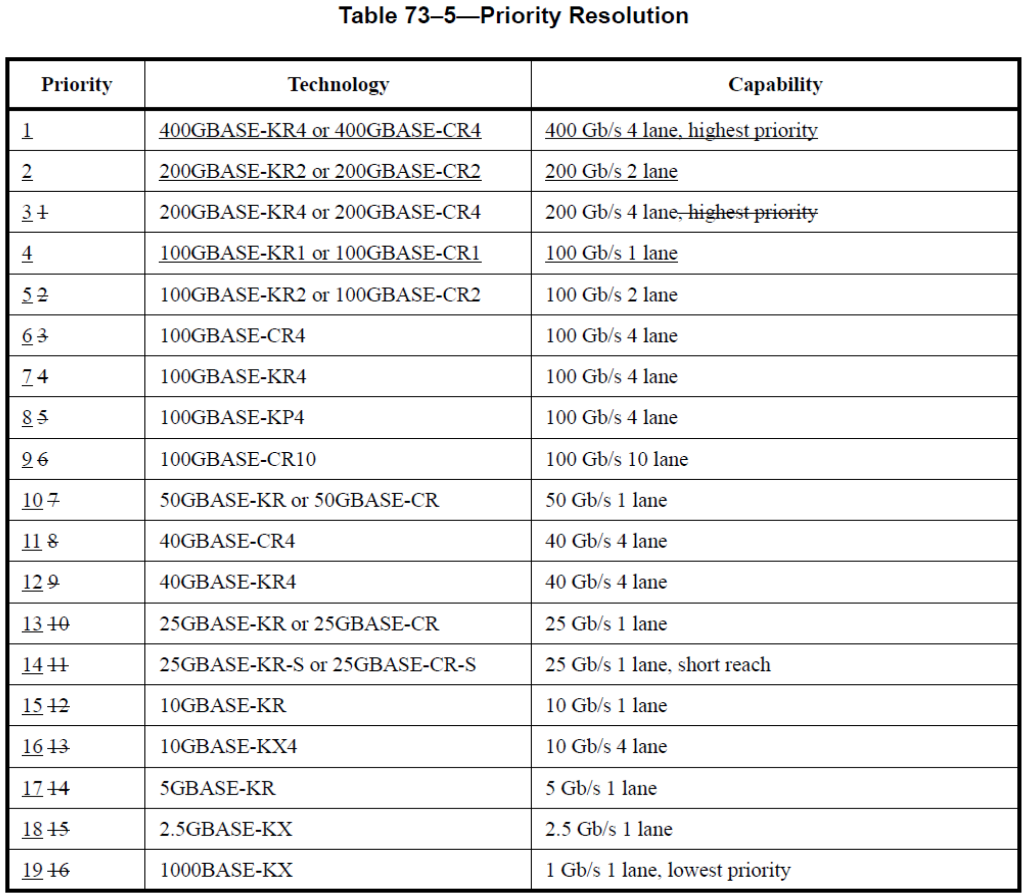
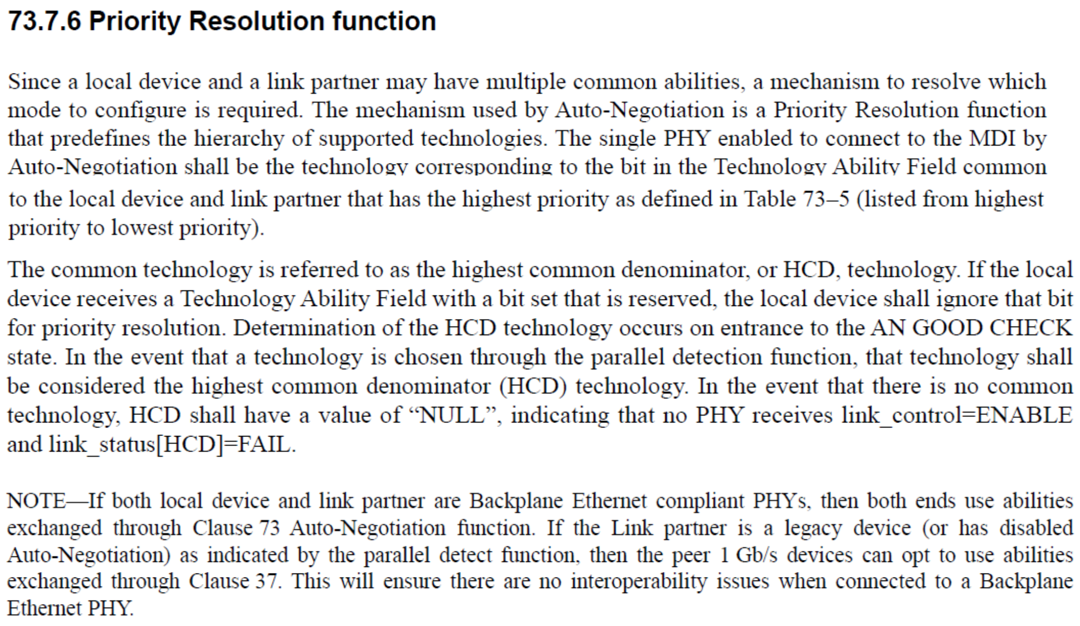


Table 73–5—Priority Resolution

Priority	Technology	Capability
1	400GBASE-KR4 or 400GBASE-CR4	400 Gb/s 4 lane, highest priority
2	200GBASE-KR2 or 200GBASE-CR2	200 Gb/s 2 lane
3 ~~1~~	200GBASE-KR4 or 200GBASE-CR4	200 Gb/s 4 lane~~, highest priority~~
4	100GBASE-KR1 or 100GBASE-CR1	100 Gb/s 1 lane
5 ~~2~~	100GBASE-KR2 or 100GBASE-CR2	100 Gb/s 2 lane
6 ~~3~~	100GBASE-CR4	100 Gb/s 4 lane
7 ~~4~~	100GBASE-KR4	100 Gb/s 4 lane
8 ~~5~~	100GBASE-KP4	100 Gb/s 4 lane
9 ~~6~~	100GBASE-CR10	100 Gb/s 10 lane
10 ~~7~~	50GBASE-KR or 50GBASE-CR	50 Gb/s 1 lane
11 ~~8~~	40GBASE-CR4	40 Gb/s 4 lane
12 ~~9~~	40GBASE-KR4	40 Gb/s 4 lane
13 ~~10~~	25GBASE-KR or 25GBASE-CR	25 Gb/s 1 lane
14 ~~11~~	25GBASE-KR-S or 25GBASE-CR-S	25 Gb/s 1 lane, short reach
15 ~~12~~	10GBASE-KR	10 Gb/s 1 lane
16 ~~13~~	10GBASE-KX4	10 Gb/s 4 lane
17 ~~14~~	5GBASE-KR	5 Gb/s 1 lane
18 ~~15~~	2.5GBASE-KX	2.5 Gb/s 1 lane
19 ~~16~~	1000BASE-KX	1 Gb/s 1 lane, lowest priority



73.7.6 Priority Resolution function

Since a local device and a link partner may have multiple common abilities, a mechanism to resolve which mode to configure is required. The mechanism used by Auto-Negotiation is a Priority Resolution function that predefines the hierarchy of supported technologies. The single PHY enabled to connect to the MDI by Auto-Negotiation shall be the technology corresponding to the bit in the Technology Ability Field common to the local device and link partner that has the highest priority as defined in Table 73–5 (listed from highest priority to lowest priority).

The common technology is referred to as the highest common denominator, or HCD, technology. If the local device receives a Technology Ability Field with a bit set that is reserved, the local device shall ignore that bit for priority resolution. Determination of the HCD technology occurs on entrance to the AN GOOD CHECK state. In the event that a technology is chosen through the parallel detection function, that technology shall be considered the highest common denominator (HCD) technology. In the event that there is no common technology, HCD shall have a value of "NULL", indicating that no PHY receives link_control=ENABLE and link_status[HCD]=FAIL.

NOTE—If both local device and link partner are Backplane Ethernet compliant PHYs, then both ends use abilities exchanged through Clause 73 Auto-Negotiation function. If the Link partner is a legacy device (or has disabled Auto-Negotiation) as indicated by the parallel detect function, then the peer 1 Gb/s devices can opt to use abilities exchanged through Clause 37. This will ensure there are no interoperability issues when connected to a Backplane Ethernet PHY.

Clause 73 AN – State diagrams – transmitter and receiver

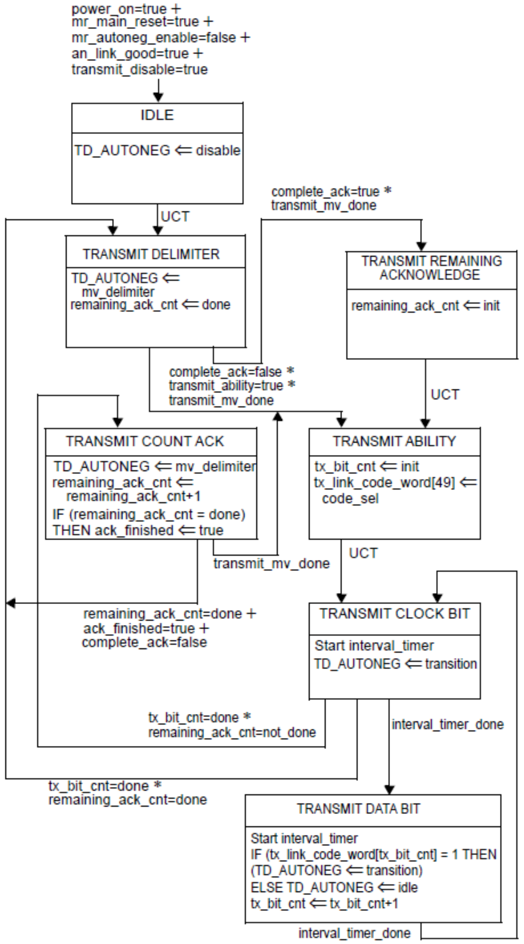


Figure 73–9—Transmit state diagram

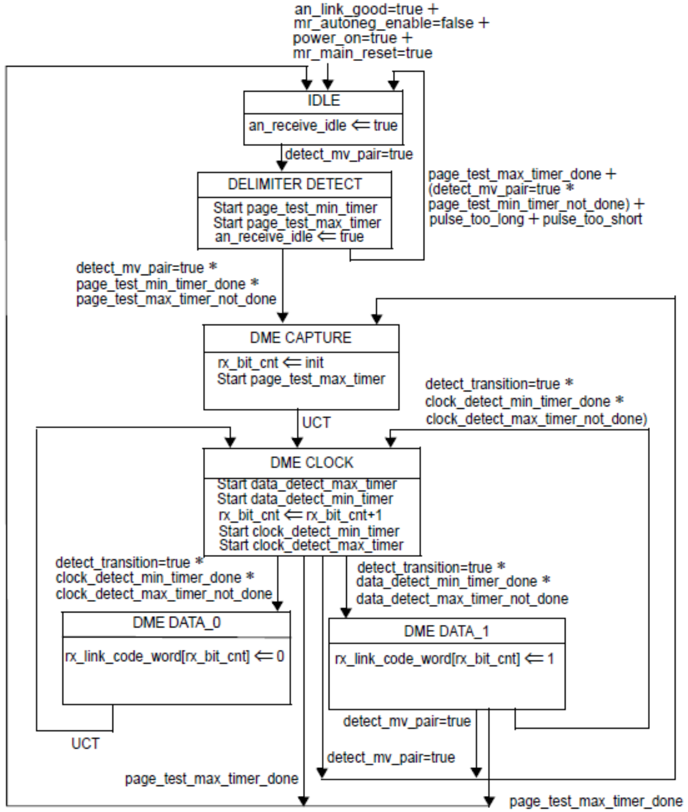

an_link_good=true +
mr_autoneg_enable=false +
power_on=true +
mr_main_reset=true
IDLE
an_receive_idle ⇐ true
detect_mv_pair=true
DELIMITER DETECT
Start page_test_min_timer
Start page_test_max_timer
an_receive_idle ⇐ true
page_test_max_timer_done +
(detect_mv_pair=true *
page_test_min_timer_not_done) +
pulse_too_long + pulse_too_short
detect_mv_pair=true *
page_test_min_timer_done *
page_test_max_timer_not_done
DME CAPTURE
rx_bit_cnt ⇐ init
Start page_test_max_timer
detect_transition=true *
clock_detect_min_timer_done *
clock_detect_max_timer_not_done)
UCT
DME CLOCK
Start data_detect_max_timer
Start data_detect_min_timer
rx_bit_cnt ⇐ rx_bit_cnt+1
Start clock_detect_min_timer
Start clock_detect_max_timer
detect_transition=true *
clock_detect_min_timer_done *
clock_detect_max_timer_not_done
detect_transition=true *
data_detect_min_timer_done *
data_detect_max_timer_not_done
DME DATA_0
rx_link_code_word[rx_bit_cnt] ⇐ 0
DME DATA_1
rx_link_code_word[rx_bit_cnt] ⇐ 1
detect_mv_pair=true
UCT
detect_mv_pair=true
page_test_max_timer_done
page_test_max_timer_done

Figure 73–10—Receive state diagram

Clause 73 AN – arbitration

Exchange advertised abilities

Restart link

“Parallel detection” for PHYs that do not support auto-negotiation

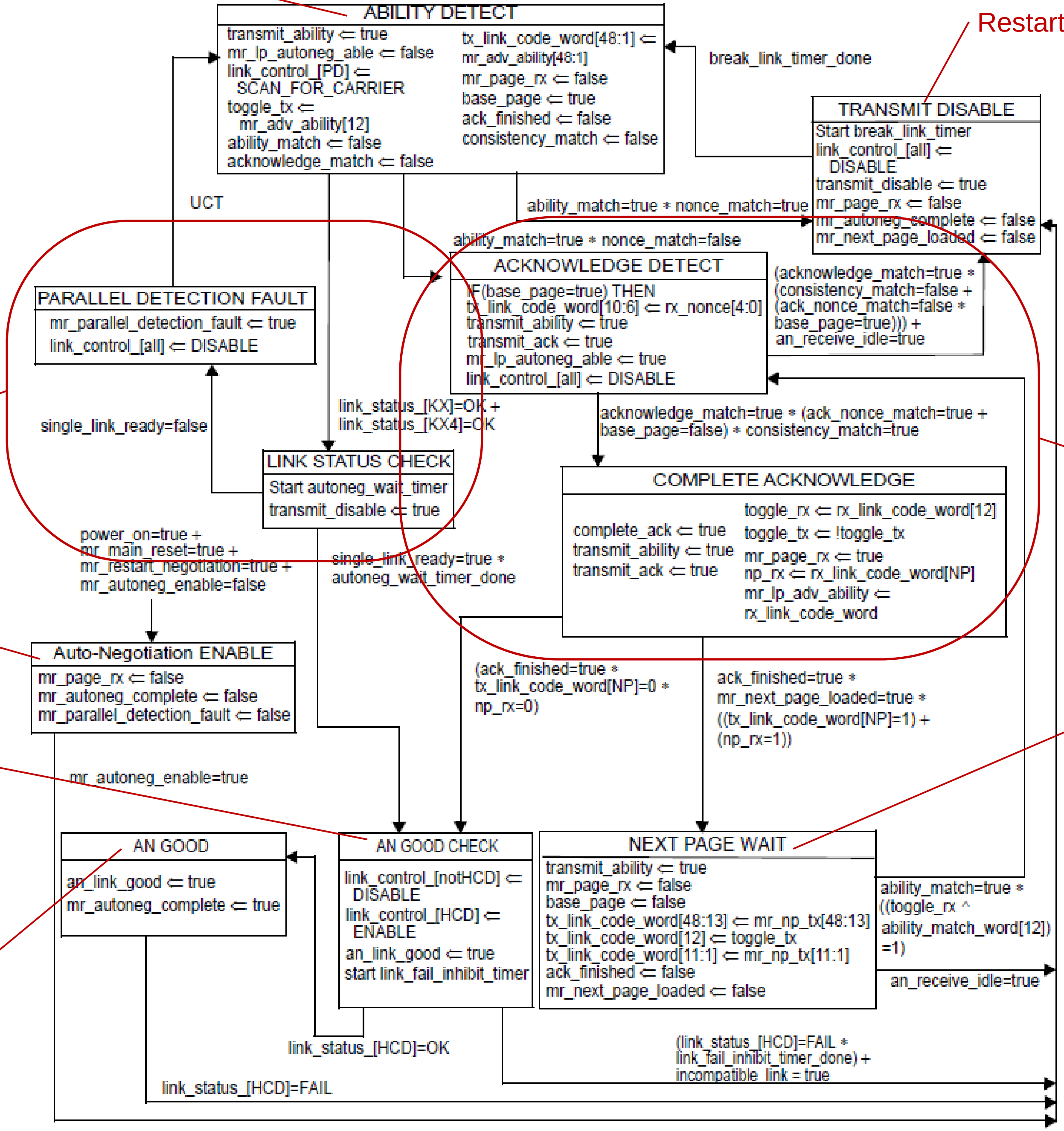
Initiate AN (start here)

Wait for negotiated link to start up successfully

Normal operation

Mutually acknowledge that common mode is available.

Wait for next page if available.



Optical link auto-configuration using Clause 73 auto-negotiation as basis

Reuse of Clause 73 auto-negotiation

- Include new auto-configuration sublayer below the PMD.
- Use signaling as defined in Clause 73, except…
  - edge spacing at 75.3 ps (53.125 GBd / 4 = 13.28125 GBd) rather than 3.2 ns (312.5 MBd)
    - Same as training frame control channel proposed in ghiasi_3dj_01_2311; 1/8 PAM4 symbol rate
  - Specify OMA requirements rather than electrical peak to peak swing as well as other optical parameters as necessary, inclusive of any anticipated PMD specifications
- Signaling over a specific lane TBD for multi-lane PHYs, other lanes disabled.
- Use page structure (base and next pages) and delineation as defined in Clause 73.
- Use state machines defined in Clause 73.
- Specify new allocation of PHY types and capabilities to base page.
- Specify new PHY prioritization table.
- Specify selection criteria for other capabilities, e.g., FEC, interleaving.
- Should it be mandatory to implement and optional to use as it is for Clause 73?
- AN is co-resident with the PMD on the module.
  - Host may need to confirm PCS status to module or PCS monitor on module. Some coordination between host and module is required regardless.

Example priority resolution table assuming two inner FEC modes (FECo/FECi) per PHY

Priority	Technology	Capability
1	1.6TBASE-DR8-2	1.6 Tb/s, 2 km, parallel
2	1.6TBASE-DR8	1.6 Tb/s, 500 m, parallel
3	800GBASE-LR4	800 Gb/s, 10 km
4	800GBASE-FR4	800 Gb/s, 2 km, duplex
5	800GBASE-DR4-2	800 Gb/s, 2 km, parallel
6	800GBASE-DR4	800 Gb/s, 500 m
7	400GBASE-DR2-2	400 Gb/s, 2 km, parallel
8	400GBASE-DR2	400 Gb/s, 500 m
9	200GBASE-FR1	200 Gb/s, 2 km, parallel
10	200GBASE-DR1	200 Gb/s, 500 m
Notes: - Separate FEC mode advertisement and priority handling is required.		

Example priority resolution table assuming single inner FEC mode (FECi/FECo) per PHY

Priority	Technology	Capability
1	1.6TBASE-DR8-2 (FECi)	1.6 Tb/s, inner FEC, 2 km, parallel
3	1.6TBASE-DR8 (FECi)	1.6 Tb/s, inner FEC, 500 m, parallel
4	1.6TBASE-?R8 (FECo place holder)	1.6 Tb/s, reach TBD, parallel
5	800GBASE-LR4 (assume always inner FEC)	800 Gb/s, 10 km
6	800GBASE-FR4 (FECi)	800 Gb/s, inner FEC, 2 km, duplex
7	800GBASE-?R4 (FECo place holder)	800 Gb/s, reach TBD, duplex
8	800GBASE-DR4-2 (FECi)	800 Gb/s, inner FEC, 2 km, parallel
10	800GBASE-DR4 (FECi)	800 Gb/s, inner FEC, 500 m
11	800GBASE-?R4 (FECo place holder)	800 Gb/s, reach TBD
12	400GBASE-DR2-2 (FECi)	400 Gb/s, inner FEC, 2 km, parallel
13	400GBASE-DR2 (FECi)	400 Gb/s, inner FEC, 500 m
14	400GBASE-?R2 (FECo place holder)	400 Gb/s, reach TBD
15	200GBASE-FR1 (FECi)	200 Gb/s, inner FEC, 2 km, parallel
16	200GBASE-DR1 (FECi)	200 Gb/s, inner FEC, 500 m
17	200GBASE-?R1 (FECo place holder)	200 Gb/s, reach TBD

Notes:
- ?Rn is for PMDs associated with new objectives proposed in lusted_3dj_05_2311

Things to think about

- For multilane PMDs, which lane should be used for signaling?
- Should AN be mandatory to implement and optional to use?
- Support for negotiation between lane rates, future proof.

Summary

- Purpose of optical automatic configuration along with helpful features discussed.
- Clause 73 auto-negotiation, used for electrical links, reviewed.
- Optical automatic configuration using Clause 73 as a starting point outlined.
  - This could be a candidate for optical link automatic configuration.

Thanks