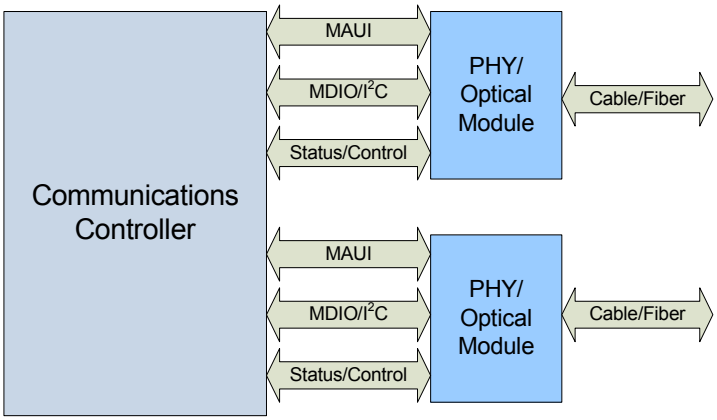


Figure 3-21 Basic PHY MAC Connectivity

The MDIO interface is a simple 2-wire serial interface between MAC and PHY and is used to access Control and Status registers inside the PHY. The interface is implemented using two LVTTL I/Os:

1. MDC — MDIO-interface clock signal driven by an external MAC (STA) device.
2. MDIO — Read/write data between an external MAC and PHY.

3.7.6.1 MDIO Timing Relationship to MDC

The MDC clock toggles during a read/write operation at a frequency of 24 MHz, 2.4 MHz or 240 KHz depending on the link speed and register bit HLREG0.MDCSPD as listed in Table 3-19.

Table 3-19 MDC Frequency as Function of Link Speed and MDC Speed Bit

Link Speed	MDCSPD=1b	MDCSPD=0b
10 Gb/s	24 MHz	2.4 MHz
1 Gb/s	2.4 MHz	240 KHz
100 Mb/s	240 MHz	240 KHz

MDIO is a bidirectional signal that can be sourced by the Station Management Entity (STA) or the PHY. When the STA sources the MDIO signal, the STA must provide a minimum of 10 ns of setup time and a minimum of 10 ns of hold time referenced to the rising edge of MDC, as shown in Figure 3-22 (measured at the MII connector).

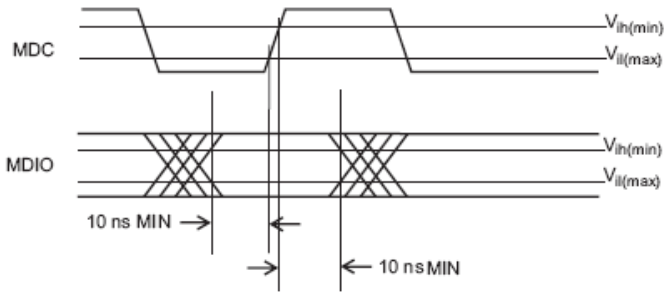


Figure 3-22 MDIO Timing Sourced by the MAC

When the MDIO signal is sourced by the PHY, it is sampled by the MAC (STA) synchronously with respect to the rising edge of MDC. The clock to output delay from the PHY, as measured at the MII connector, must be a minimum of 0 ns, and a maximum of 300 ns, as shown in Figure 3-23.

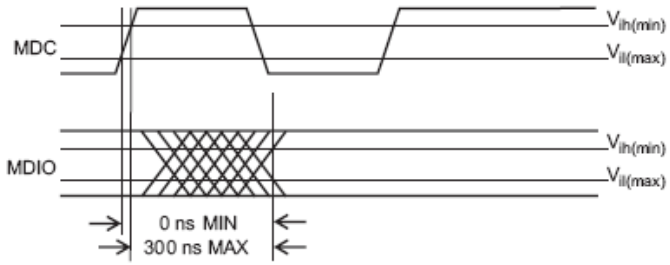


Figure 3-23 MDIO Timing Sourced by the PHY

3.7.6.2 IEEE802.3 Clause 22 and Clause 45 Differences

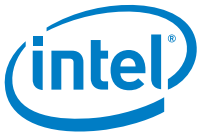
IEEE802.3 clause 45 provides the ability to access additional device registers while still retaining logical compatibility with interface defined in Clause 22. Clause 22 specifies the MDIO frame format and uses an ST code of 01 to access registers. In clause 45, additional registers are added to the address space by defining MDIO frames that use a ST code of 00.

Clause 45 (MDIO interface) major concepts:

a. Preserve management frame structure defined in IEEE 802.3 Clause 22.

b. Define mechanism to address more registers than specified in IEEE802.3 Clause 22.

c. Define ST and OP codes to identify and control the extended access functions.



3.7.6.3 MDIO Management Frame Structure

The MDIO interface frame structure defined in IEEE802.3 clause 22 and Clause 45 are compatible so that the two systems supporting different formats can co-exist on the same MDIO bus. The 82599 supports both frame structures to enable interfacing PHYs that support either protocol.

The basic frame format as defined in IEEE802.3 clause 22 can optionally be used for accessing legacy PHY registers is listed in Table 3-20.

Table 3-20 Clause 22 Basic MDIO Frame Format

	Management Frame Fields							
Frame	**Pre**	**ST**	**OP**	**PRTAD**	**REGAD**	**TA**	**Data**	**Idle**
Read	1...1	01	10	PPPPP	RRRRR	Z0	DDDDDDDDDDDDDDDD	Z
Write	1...1	01	01	PPPPP	RRRRR	10	DDDDDDDDDDDDDDDD	Z

The MDIO interface defined in clause 45 uses indirect addressing to create an extended address space enabling access to a large number of registers within each MDIO Managed Device (MMD). The MDIO management frame format is listed in Table 3-21.

Table 3-21 Clause 45 Indirect Addressing MDIO Frame Format

	Management Frame Fields							
Frame	**Pre**	**ST**	**OP**	**PRTAD**	**DEVAD**	**TA**	**Address / Data**	**Idle**
Address	1...1	00	00	PPPPP	EEEEE	10	AAAAAAAAAAAAAAAA	Z
Write	1...1	00	01	PPPPP	EEEEE	10	DDDDDDDDDDDDDDDD	Z
Read	1...1	00	11	PPPPP	EEEEE	Z0	DDDDDDDDDDDDDDDD	Z
Post-Read Increment Address	1...1	00	10	PPPPP	EEEEE	Z0	DDDDDDDDDDDDDDDD	Z

To support clause 45 indirect addressing each MMD (PHY — MDIO managed device) implements a 16-bit address register that stores the address of the register to be accessed by data transaction frames. The address register must be overwritten by address frames. At power up or device reset, the contents of the address register are undefined. Write, read, and post-read-increment-address frames must access the register whose address is stored in the address register. Write and read frames must not modify the contents of the address register. Upon receiving a post-read-increment-address frame and having completed the read operation, the MMD increments the Address register by one (up to a value of 0xFFFF). Each MMD supported implements a separate address register, so that the MMD's address registers operate independently of one another.

Idle Condition (IDLE) — The IDLE condition on MDIO is a high-impedance state. All three state drivers must be disabled and the PHY's pull-up resistor pulls the MDIO line to a logic one.

Preamble (PRE) — At the beginning of each transaction, the station management entity must send a sequence of 32 contiguous consecutive one bits on MDIO with 32 corresponding cycles on MDC to provide the PHY with a pattern that it can use to establish synchronization. A PHY must observe a sequence of 32 contiguous consecutive one bits on MDIO with 32 corresponding cycles on MDC before it responds to any transaction.

Start of Frame (ST) — The ST is indicated by:

- <00> pattern for clause 45 compatible frames for indirect access cycles.
- <01> pattern for clause 22 compatible frames for direct access cycles.

These patterns ensure a transition from the default value of one on the MDIO signal, and identifies the start of frame.

Operation Code (OP) — The *OP* field indicates the type of transaction being performed by the frame.

For Clause 45 compatible frames:

- A <00> pattern indicates that the frame payload contains the address of the register to access.
- A <01> pattern indicates that the frame payload contains data to be written to the register whose address was provided in the previous address frame.
- A <11> pattern indicates that the frame is an indirect read operation.
- A <10> pattern indicates that the frame is an indirect post-read-increment-address operation.

For Clause 22 compatible frames:

- A <10> pattern indicates a direct read transaction from a register.
- A <01> pattern indicates a direct write transaction to a register.

Port Address (PRTAD) — The PRTAD is five bits, allowing 32 unique PHY port addresses. The first *PRTAD* bit to be transmitted and received is the MSB of the address. A station management entity must have prior knowledge of the appropriate port address for each port to which it is attached, whether connected to a single port or to multiple ports.

Device Address (DEVAD) — The DEVAD is five bits, allowing 32 unique MMDs per port. The first *DEVAD* bit transmitted and received is the MSB of the address. This field is relevant only in clause 45 compatible frames (ST=<00>).

Register Address (REGAD) — The REGAD is five bits, allowing 32 individual registers to be addressed within each PHY. The first *REGAD* bit transmitted and received is the MSB of the address. This field is relevant only in clause 22 compatible frames (ST=<01>).

Turnaround (TA) — The TA time is a 2-bit time spacing between the *DEVAD* field and the *Data* field of a management frame. This is to avoid contention during a read transaction. For a read or post-read-increment-address transaction, both the STA and the PHY must remain in a high-impedance state for the first bit time of the TA. The PHY must drive a zero bit during the second bit time of the TA of a read or postread-increment-address transaction. During a write or address transaction, the STA must drive a one bit for the first bit time of the TA and a zero bit for the second bit time of the TA. Figure 3-24 shows the behavior of the MDIO signal during the *TA* field of a read transaction.

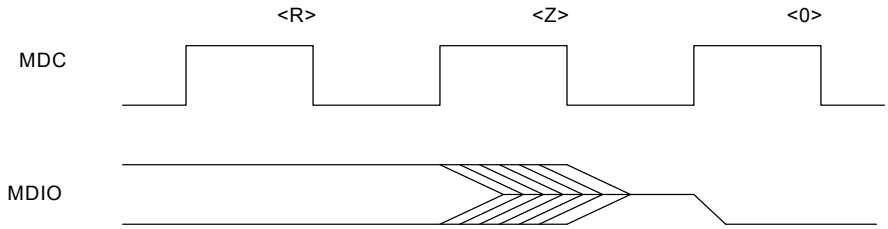


Figure 3-24 Behavior of MDIO During TA Field of a Read Transaction

- Clause 45 compatible frames have 16-bit address/data fields. For an auto-negation address cycle, it contains the address of the register to be accessed on the next cycle. For the data cycle of a write frame, the field contains the data to be written to the register. For a read or post-read-increment-address frame, the field contains the contents of the register. The first bit transmitted and received must be bit 15.
- Clause 22 compatible frames have 16-bit data fields. The first data bit transmitted and received must be bit 15 of the register being addressed.

3.7.6.4 MDIO Direct Access

The MDI is accessed through registers MSCA and MSRWD. A single management frame is sent by setting bit MSCA.MDICMD to 1b after programming the appropriate fields in the MSCA and MSRWD registers. The MSCA.MDICMD bit is auto cleared after the read or write transaction completes. To execute clause 22 format write operations, the following steps should be done:

1. Data to be written is programmed in field MSRWD.MDIWRDATA.
2. Register MSCA is initialized with the appropriate control information (start, code, etc.) with bit MSCA.MDICMD set to 1b.
3. Wait for bit MSCA.MDICMD to reset to 0b when indicating that the transaction on the MDIO interface is complete.

The steps for clause 22 format read operations are identical to the write operation except that the data in field MSRWD.MDIWRDATA is ignored and the data read from the external device is stored in register field MSRWD.MDIRDDATA bits. Clause 45 format read/write operations must be performed in two steps. The address portion of the pair of frames is sent by setting register field MSCA.MDIADD to the desired address, field MSCA.STCODE to 00b (start code that identifies clause 45 format), and register field MSCA.OPCODE to 00b (clause 45 address register write operation). A second data frame must be sent after the address frame completes. This second frame executes the write or read operation to the address specified in the PHY address register.

3.7.7 Ethernet Flow Control (FC)

The 82599 supports flow control as defined in 802.3x, as well as the specific operation of asymmetrical flow control defined by 802.3z. The 82599 also supports Priority Flow Control (PFC), sometimes referred to as Class Based Flow Control or (CBFC), as part of the DCB architecture.

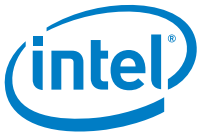
Note: The 82599 can either be configured to receive regular flow control packets or Priority Flow Control (PFC) packets. The 82599 does not support the reception of both types of packets simultaneously.

Flow control is implemented to reduce receive buffer overflows, which result in the dropping of received packets. Flow control also allows for local controlling of network congestion levels. This can be accomplished by sending an indication to a transmitting station of a nearly full receive buffer condition at a receiving station.

The implementation of asymmetric flow control allows for one link partner to send flow control packets while being allowed to ignore their reception (for example, not required to respond to PAUSE frames).

The following registers are defined for the implementation of flow control. In DCB mode, some of the registers are duplicated per Traffic Class (TC), up to eight duplicate copies of the registers. If DCB is disabled, index [0] of each register is used.

- MAC Flow Control (MFLCN) register — Enables flow control and passing of control packets to the host.
- Flow Control Configuration (FCCFG) — Determines mode for Tx flow control (no FC vs. link based versus priority based). Note that if Tx flow control is enabled then Tx CRC by hardware should be enabled as well (HLREG0.TXCRCEN = 1b).
- Flow Control Address Low, High (RAL[0], RAH[0]) — 6-byte flow control multicast address.
- Priority Flow Control Type Opcode (PFCTOP) — Contains the type and opcode values for priority FC.
- Flow Control Receive Threshold High (FCRTH[7:0]) — A set of 13 bit high watermarks indicating receive buffer fullness. A single watermark is used in link FC mode and up to eight watermarks are used in priority FC mode.
- Flow Control Receive Threshold Low (FCRTL[7:0]) — A set of 13 bit low watermarks indicating receive buffer emptiness. A single watermark is used in link FC mode and up to eight watermarks are used in priority FC mode.
- Flow Control Transmit Timer Value (FCTTV[3:0]) — a set of 16 bit timer values to include in transmitted PAUSE frame. A single timer is used in link FC mode and up to eight timers are used in priority FC mode.
- Flow Control Refresh Threshold Value (FCRTV) — 16-bit PAUSE refresh threshold value (in legacy FC FCRTV[0] must be smaller than FCTTV[0]).



3.7.7.1 MAC Control Frames and Reception of Flow Control Packets

3.7.7.1.1 MAC Control Frame — Other than FC

IEEE reserved the Ethertype value of 0x8808 for MAC control frames as listed in Table 3-22.

Table 3-22 MAC Control Frame Format

DA	The *Destination Address* field can be an individual or multicast (including broadcast) address. Permitted values for the *Destination Address* field can be specified separately for a specific control opcode such as FC packets.
SA	Port Ethernet MAC address (6 bytes).
Type	0x8808 (2 bytes).
Opcode	The MAC control opcode indicates the MAC control function.
Parameters	The *MAC Control Parameters* field must contain MAC control opcode-specific parameters. This field can contain none, one, or more parameters up to a maximum of minFrameSize =20 bytes.
Reserved field = 0x00	The *Reserved* field is used when the MAC control parameters do not fill the fixed length MAC control frame.
CRC	4 bytes.

3.7.7.1.2 Structure of 802.3X FC Packets

802.3X FC packets are defined by the following three fields (see Table 3-23):

1. A match on the six-byte multicast address for MAC control frames or a match to the station address of the device (Receive Address Register 0). The 802.3x standard defines the MAC control frame multicast address as 01-80-C2-00-00-01.
2. A match on the *Type* field. The *Type* field in the FC packet is compared against an IEEE reserved value of 0x8808.
3. A match of the *MAC Control Opcode* field has a value of 0x0001.

Frame based flow control differentiates XOFF from XON based on the value of the PAUSE *Timer* field. Non-zero values constitute XOFF frames while a value of zero constitutes an XON frame. Values in the *Timer* field are in units of pause quanta (slot time). A pause quanta lasts 64 byte times, which is converted in to an absolute time duration according to the line speed.

Note: XON frame signals the cancellation of the pause from that was initiated by an XOFF frame pause for zero pause quanta).

Table 3-23 802.3X Packet Format

DA	01_80_C2_00_00_01 (6 bytes).
SA	Port Ethernet MAC address (6 bytes).
Type	0x8808 (2 bytes).
Opcode	0x0001 (2 bytes).
Time	XXXX (2 bytes).
Pad	42 bytes.
CRC	4 bytes.

3.7.7.1.3 PFC

DCB introduces support for multiple traffic classes assigning different priorities and bandwidth per TC. Link level Flow Control (PAUSE) stops all the traffic classes. PFC or CBFC allows more granular flow control on the Ethernet link in an DCB environment as opposed to the PAUSE mechanism defined in 802.3X.

PFC is implemented to prevent the possibility of receive packet buffers overflow. Receive packet buffers overflow results in the dropping of received packets for a specific TC. Board designers can implement PFC by sending a timer indication to the transmitting station traffic class (XOFF) of a nearly full receive buffer condition at the 82599. At this point the transmitter would stop transmitting packets for that TC until the XOFF timer expires or a XON message is received for the stopped TC.

Similarly, once the 82599 receives a priority-based XOFF it stops transmitting packets for that specific TC until the XOFF timer expires or XON packet for that TC is received.

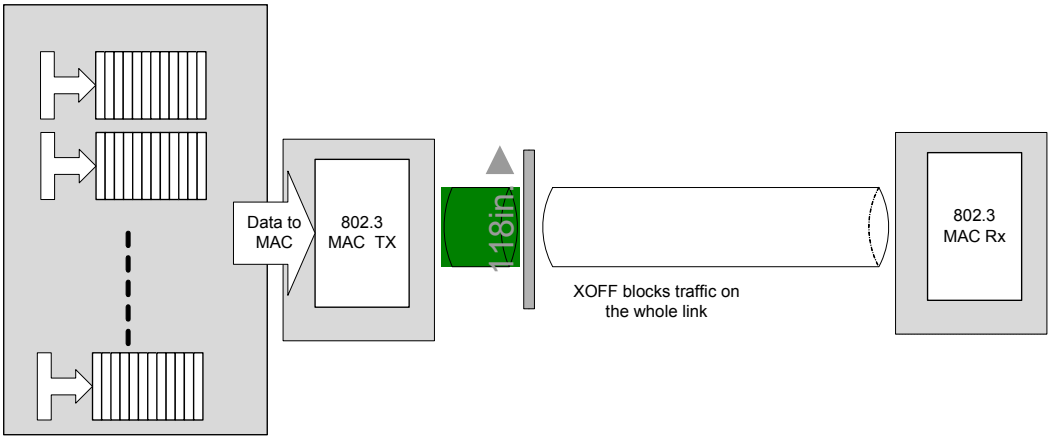


Figure 3-25 802.3X Link Flow Control (PAUSE)

Link flow control (802.3X) causes all traffic to be stopped on the link. DCB uses the same mechanism of flow control but provides the ability to do PFC on TCs as shown in Figure 3-26.

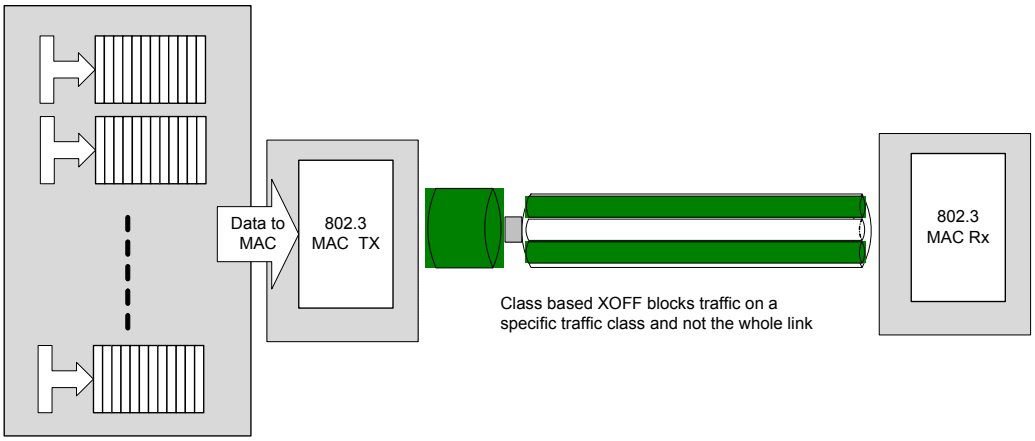


Figure 3-26 Priority Flow Control

Table 3-24 Packet Format for Priority Flow Control

DA	01_80_C2_00_00_01 (6 bytes).
SA	Port Ethernet MAC Address (6 bytes).
Type	0x8808 (2 bytes).
Opcode	0x0101 (2 bytes).
Priority Enable Vector	0x00XX (2 bytes).
Timer 0	XXXX (2 bytes).
Timer 1	XXXX (2 bytes).
Timer 2	XXXX (2 bytes).
Timer 3	XXXX (2 bytes)
Timer 4	XXXX (2 bytes).
Timer 5	XXXX (2 bytes).
Timer 6	XXXX (2 bytes).
Timer 7	XXXX (2 bytes).
Pad	26 bytes.
CRC	4 bytes.

Table 3-25 Format of Priority Enable Vector

	ms octet	ls octet
Priority Enable vector definition	0	e[7]...e[n]...e[0]
e[n] =1 => time (n) valid e[n] =0 => time (n) invalid		

The Priority Flow Control Type Opcode (PFCTOP) register contains the type and opcode values for PFC. These values are compared against the respective fields in the received packet.

Each of the eight timers refers to a specific User Priority (UP). For example, Timer 0 refers to UP 0, etc. The 82599 binds a UP (and therefore the timer) to one of its TCs according to the UP-to-TC binding tables. Refer to the RTTUP2TC register for the binding of received PFC frames to Tx TCs, and to the RTRUP2TC register for the binding of transmitted PFC frames to Rx TCs.

Tx manageability traffic is bound to one the TCs via the MNGTXMAP register, and should thus be paused according to RTTUP2TC mapping whenever receiving PFC frames.

When a PFC frame is formatted by the 82599, the same values are replicated into every *Timer* field and priority enable vector bit of all the UPs bound to the concerned TC. These values as configured in the RTRUP2TC register.

The following rule is applicable for the case of multiple UPs that share the same TC (as configured in the RTTUP2TC register). When PFC frames are received with different timer values for the previous UPs, the traffic on the associated TC must be paused by the highest XOFF timer's value.

3.7.7.1.4 Operation and Rules

The 82599 operates in either link FC or in PFC mode. Enabling both modes concurrently is not allowed:

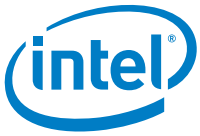
- Link FC is enabled by the *RFCE* bit in the MFLCN register.
- PFC is enabled by the *RPFCE* bit in the MFLCN register.

Note: Link flow control capability must be negotiated between link partners via the auto-negotiation process. PFC capability is negotiated via some higher level protocol and the resolution is usually provided to the driver by the DCB management agent. It is the driver's responsibility to reconfigure the link flow control settings (including RFCE and PRFCE) after the auto-negotiation process was resolved.

Receiving a link FC frame while in PFC mode might be ignored or might pause TCs in an unpredictable manner. Receiving a PFC frame while in link FC mode is ignored.

Once the receiver has validated the reception of an XOFF, or PAUSE frame, the device performs the following:

- Increments the appropriate statistics register(s)
- Initialize the pause timer based on the packet's PAUSE *Timer* field (overwriting any current timer's value)



  - In case of PFC, this is done per TC. If several UPs are associated with a TC, then the device sets the timer to the maximum value among all enabled timer fields associated with the TC.

- Disable packet transmission or schedule the disabling of transmission after the current packet completes.
  - In case of PFC, this is done per paused TC
  - Tx manageability traffic is bound to a specific TC as defined in the MNGTXMAP register, and is thus paused when its TC is paused

Resumption of transmission can occur under the following conditions:

- Expiration of the PAUSE timer
  - In case of PFC, this is done per TC
- Reception of an XON frame (a frame with its PAUSE timer set to 0b)
  - In case of PFC, this is done per TC

Both conditions set the relevant TC_XON status bits in the Transmit Flow Control Status (TFCS) register and transmission can resume. Hardware records the number of received XON frames.

3.7.7.1.5 Timing Considerations

When operated at 10 Gb/s line speed, the 82599 must not begin to transmit a (new) frame more than 60 pause quanta after receiving a valid Link XOFF frame, as measured at the wires (a pause quantum is 512 bit times). When connected to an external 10GBASE-KR PHY with FEC or to an external 10GBASE-T PHY, the response time requirement decreases to 74 pause quanta, because of extra delays consumed by these external PHYs.

When operating at 1 Gb/s line speed, the 82599 must not begin to transmit a (new) frame more than 2 pause quanta after receiving a valid Link XOFF frame, as measured at the wires.

The 802.1Qbb draft 1.0, proposes that the tolerated response time for Priority XOFF frames are the same as Link XOFF frames with extra budget of 19072 bit times if MACSec is used, or of 2 pause quanta otherwise. This extra budget is aimed to compensate the fact that decision to stop new transmissions from a specific TC must be taken earlier in the transmit data path than for the Link Flow Control case.

3.7.7.2 PAUSE and MAC Control Frames Forwarding

Two bits in the Receive Control register control transfer of PAUSE and MAC control frames to the host. These bits are Discard PAUSE Frames (DPF) and Pass MAC Control Frames (PMCF). Note also that any packet must pass the L2 filters as well.

- The *DPF* bit controls transfer of PAUSE packets to the host. The same policy applies to both link FC and priority FC packets as listed in Table 3-26. Note that any packet must pass the L2 filters as well.

- The *PMCF* bit controls transfer of non-PAUSE packets to the host. Note that when link FC frames are not enabled (RFCE = 0b) then link FC frames are considered as MAC Control (MC) frames for this matter. Similarly, when PFC frames are not enabled (RPFCE = 0b) then PFC frames are considered as MC frames as well.

Note: When virtualization is enabled, forwarded control packets are queued according to the regular switching procedure defined in Section 7.10.3.4.

Table 3-26 Transfer of PAUSE Packet to Host (DPF Bit)

RFCE	RPFCE	DPF	Link FC handling	Priority FC handling
0b	0b	X	Treat as MC (according to PMCF setting).	Treat as MC (according to PMCF setting).
1b	0b	0b	Accept.	Treat as MC (according to PMCF setting).
1b	0b	1b	Reject.	Treat as MC (according to PMCF setting).
0b	1b	0b	Treat as MC (according to PMCF setting).	Accept.
0b	1b	1b	Treat as MC (according to PMCF setting).	Reject.
1b	1b	X	Unsupported setting.	Unsupported setting.

3.7.7.3 Transmitting PAUSE Frames

The 82599 generates PAUSE packets to insure there is enough space in its receive packet buffers to avoid packet drop. The 82599 monitors the fullness of its receive FIFOs and compares it with the contents of a programmable threshold. When the threshold is reached, the 82599 sends a PAUSE frame. The 82599 supports both link flow control and PFC — but not both concurrently. When DCB is enabled, it sends only PFC, and when DCB is disabled, it send only link flow control.

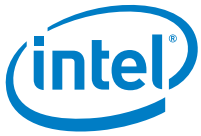
Note: Similar to the reception of flow control packets previously mentioned, software can enable flow control transmission by setting the FCCFG.TFCE field only after it is negotiated between the link partners (possibly by auto-negotiation).

3.7.7.3.1 Priority Flow Control

The same flow control mechanism is used for PFC and for 802.3X flow control to determine when to send XOFF and XON packets. When PFC is used in the receive path, Priority PAUSE packets are sent instead of 802.3X PAUSE packets. The format of priority PAUSE packets is described in Section 3.7.7.1.3.

Specific considerations for generating PFC packets:

- When a PFC packet is sent, the packet sets all the UPs that are associated with the relevant TC (UP-to-TC association in receive is defined in RTRUP2TC register).



3.7.7.3.2 Operation and Rules

The *TFCE* field in the Flow Control Configuration (FCCFG) register enables transmission of PAUSE packets as well as selects between the link flow control mode and the PFC mode.

The content of the Flow Control Receive Threshold High (FCRTH) register determines at what point the 82599 transmits the first PAUSE frame. The 82599 monitors the fullness of the receive FIFO and compares it with the contents of FCRTH. When the threshold is reached, the 82599 sends a PAUSE frame with its pause time field equal to FCTTV.

At this time, the 82599 starts counting an internal shadow counter (reflecting the pause time-out counter at the partner end). When the counter reaches the value indicated in FCRTV register, then, if the PAUSE condition is still valid (meaning that the buffer fullness is still above the low watermark), an XOFF message is sent again.

Once the receive buffer fullness reaches the low water mark, the 82599 sends an XON message (a PAUSE frame with a timer value of zero). Software enables this capability with the XONE field of the FCRTL.

The 82599 sends a PAUSE frame if it has previously sent one and the FIFO overflows. This is intended to minimize the amount of packets dropped if the first PAUSE frame did not reach its target.

3.7.7.3.3 Flow Control High Threshold — FCRTH

The 82599 sends a PAUSE frame when a Rx packet buffer is full above the high threshold. The threshold should be large enough to overcome the worst case latency from the time that crossing the threshold is sensed until packets are not received from the link partner. This latency is composed of the following elements:

- Threshold Cross to XOFF Transmission + Round-trip Latency + XOFF Reception to Link Partner Response, where:

Latency Parameter	Affected by. . .	Value at 10 GbE with Jumbo
Trigger to XOFF transmission.	Max packet size at all TCs.	9.5 KB (example).
Link partner XOFF to transmission hold.	Max packet size on the specific TC.	9.5 KB (example).
Round-trip Latency.	The latencies on the wire and the LAN devices at both sides of the wire.	8 KB (see the calculation that follows).

- Round-trip Latency Calculation:
  - Pause Quanta (PQ) = 512 bit time (bt)
  - Round trip for 10 GbE MAC + XAUI + 10 GbE PHY = 16+8+50 PQ ≅ 4.7 KB (using another PHY a lower latency can be taken)
  - Round trip capable (2x100 m) = 200 m x 50 bt/m = 10000 bt ≅ 1.25 KB (at other known topologies lower latency can be taken)
  - Plus 2 KB for some guard-band and processing latency of transmission and reception pause frames

The internal architecture of the Rx packet buffer is as follows:

1. Any packet starts at 32 byte aligned address.
2. Any packet has an internal status of 32 bytes. As a result, the Rx packet buffer is used at worst conditions when the Rx packet includes 65 bytes that are posted to the host memory. Assuming that the CRC bytes are not posted to host memory then in the worst case the Rx packet buffer can be filled at 1.44 higher rate than the wire speed (69-byte packet including CRC + 8-byte preamble + 12-byte back-to-back IFS consumes 4 x 32 bytes = 128 bytes on the Rx packet buffer).

Translating the latencies to possible consumed Rx packet buffer at worst case is:

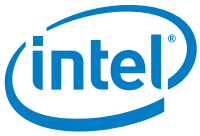
Latency Parameter	Value	Consumed Rx Packet Buffer
Trigger to XOFF transmission	9.5 KB	1.44 x 9.5 KB ≅ 14 KB
Link partner XOFF to transmission hold	9.5 KB	9.5 KB
Round-trip latency	8 KB	1.44 x 8 KB ≅ 11.5 KB

The FCRTH should be set to the size of the Rx packet buffer minus (14 + 9.5 + 11.5 = 35 KB). As previously indicated, these numbers are valid if jumbo frames are enabled in all traffic classes. When it is required to avoid packet lost, software must follow this requirement and enable flow control functionality.

When Tx to Rx switching is enabled, packets can be received to the Rx packet buffer by local VM-to-VM traffic. Once the Rx packet buffer gets full and is above the high threshold it might receive up to one additional packet from a local VM. Therefore, FCRTH should be set to the size of the Rx packet buffer minus (the size previously explained plus one additional max packet size).

3.7.7.3.4 Flow Control Low Threshold — FCRTL

The low threshold value is aimed to protect against wasted available host bandwidth. There is some latency from the time that the low threshold is crossed until the XON frame is sent and packets are received from the link partner. The low threshold can be set high enough so that the Rx packet buffer does not get empty before new whole packets are received from the link partner. When considering data movement from the Rx packet buffer to host memory, then large packets represent the worst. Assuming the host bandwidth is about as twice the bandwidth on the wire (when only a single port is active at a given time). Therefore, on 10 GbE network with jumbo packets a threshold that guarantee that the Rx packet buffer is not emptied should be set larger than: 2 x (2 x 9.5 KB + 8 KB) ≅ 54 KB. Setting the FCRTL to lower values than expressed by the previous equation is permitted. It might simply result with potential sub-optimal use of the PCIe bus once bandwidth is available.



3.7.7.3.5 Packet Buffer Size

When flow control is enabled, the total size of a packet buffer must be large enough for the low and high thresholds. In order to avoid constant transmission of XOFF and XON frames it is recommended to add some space for hysteresis type of behavior. The difference between the two thresholds is recommended to be at least one frame size (when 9.5 KB (9728-byte) jumbo frames are enabled) and larger than a few frames in other cases. If the available Rx packet is large enough, it is recommended to increase as much as possible the hysteresis budget. If the available Rx packet is not large enough it might be required to cut both the low threshold as well as the hysteresis budget. The following table lists a few examples while it is recommended to validate the values for a given use case.

Latency Parameter	Flow Control High Threshold	Flow Control Low Threshold	Total Packet Buffer Size
9.5 KB (9728-byte) jumbo enabled with no DCB with flow control.	477 KB	54 KB	512 KB
9.5 KB (9728-byte) jumbo enabled x 8 TCs with flow control.	29 KB	19.5 KB	64 KB
9.5 KB (9728-byte) jumbo enabled x 8 TCs with flow control and flow director table enabled with 128 KB.	13 KB	9 KB	48 KB
9.5 KB (9728-byte) jumbo enabled x 4 TCs with flow control and 1500-byte (no jumbo) x 4 TCs with flow control and flow director table enabled with 128 KB.	21 KB 14 KB	11.5 KB 9 KB	56 KB (jumbo) 40 KB (1.5 KB)
9.5 KB (9728-byte) jumbo enabled x 4 TCs WITHOUT flow control and 1500-byte (no jumbo) x 4 TCs WITH flow control and flow director table enabled with 128 KB.	N/A 30 KB	N/A 20 KB	40 KB (jumbo) 56 KB (1.5 KB)

When Tx-to-Rx switching is enabled (in virtualization mode) the high threshold should take into account potential VM-to-VM reception. As a result, the Rx packet buffer's sizes should be increased, respectively.

3.7.7.4 Link FC in DCB Mode

When operating in DCB mode, PFC is the preferred method of getting the best use of the link for all TCs. When connecting to switches that do not support (or enable) PFC, the 82599 throttles the traffic using link FC. Following is the required device setting and functionality:

- The 82599 should be set to legacy link FC by setting MFLCN.RFCE.
- Reception of XOFF pauses transmission in all TCs.
- Crossing the Rx buffer high threshold on any TC generates XOFF transmission. Each TC can have its own threshold configured by the FCRTH[n] registers.
- Crossing the Rx buffer low threshold on any TC generates XON transmission. This behavior is undesired. Therefore, software should not enable XON in this mode by clearing FCRTL[n].XONE bits in all TC.
- The Flow Control Transmit Timer Value of all TCs must be set to the same value.

3.7.8 Inter Packet Gap (IPG) Control and Pacing

The 82599 supports transmission pacing by extending the IPG (the gap between consecutive packets). The pacing mode allows the average data rate to be slowed in systems that cannot support the full link rate (10 Gb/s, 1Gb/s or 100 Mb/s). As listed in Table 3-27, the pacing modes work by stretching the IPG in proportion to the data sent. In this case the data sent is measured from the end of preamble to the last byte of the packet. No allowance is made for the preamble or default IPG when using pacing mode.

Example 1:

Consider an example of a 64-byte frame. To achieve a 1 Gb/s data rate when link rate is 10 Gb/s and packet length is 64 bytes (16 Dwords), programmers need to add an additional IPG of 144 Dwords (nine times the packet size to reach 1 Gb/s). Which when added to the default IPG gives an IPG of 147 Dwords.

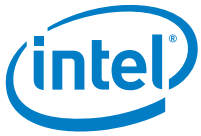
Example 2:

Consider an example of a 65-byte frame. To achieve a 1 Gb/s data rate when link rate is 10 Gb/s and packet length is 65 bytes (17 Dwords when rounded up) programmers need to add an additional IPG of 153 Dwords (nine times the packet duration in Dwords). Which when added to the default IPG gives an IPG of 156 Dwords. Note that in these case, where the packet length counted in Dwords is not an integer, programmers need to count any fraction of a Dword as a whole Dword for computing the additional IPG.

Table 3-27 lists the pacing configurations supported by the 82599 at link rates of 10 Gb/s. When operating at lower link speeds the pacing speed is proportional to the link speed.

Table 3-27 Pacing Speeds at 10 Gb/s Link Speed

Pacing Speeds (Gb/s)	Delay Inserted into IPG	Register Value
10 (LAN)	None	0000b
9.294196 (WAN)	1 byte for 13 transmitted	1111b
9.0	1 Dword for 9 transmitted	1001b
8.0	1 Dword for 4 transmitted	1000b
7.0	3 Dwords for 7 transmitted	0111b
6.0	2 Dwords for 3 transmitted	0110b
5.0	1 Dwords for 1 transmitted	0101b
4.0	3 Dwords for 2 transmitted	0100b
3.0	7 Dwords for 3 transmitted	0011b
2.0	4 Dwords for 1 transmitted	0010b
1.0	9 Dwords for 1 transmitted	0001b
10	None	Default



Pacing is configured in the *PACE* field of the Pause and Pace (PAP) register.

Note: The IPG pacing feature is a parallel feature to the Tx rate scheduler where IPG pacing is applied to the entire Tx data flow while the Tx rate scheduler is applied separately to each Tx queue. Therefore, if a single queue is used, either feature can be used to limit the Tx data rate; however, if multiple queues are used, the IPG pacing feature is a better choice for a homogeneous Tx data rate limitation.

3.7.9 MAC Speed Change at Different Power Modes

Normal speed negotiation drives to establish a link at the Highest Common Denominator (HCD) link speed. The 82599 supports an additional mode of operation, where the MAC establishes a link at the Lowest Common Denominator (LCD) link speed. The link-up process enables a link to come up at any possible speed in cases where power is more important than performance. Different behavior is defined for the D0 state and non-D0 states as a function of the AUTOC.D10GMP, AUTOC.RATD and MMNGC.MNG_VETO register bits.

The 82599 can initiate auto-negotiation without direct driver command in the following cases:

- When the state of MAIN_PWR_OK pin changes.
- When the MNG_VETO bit value changes.
- On a transition from D0a state to a non-D0a state, or from a non-D0a state to D0a state.

Figure 3-27 shows the 82599 behavior when entering low power mode and Figure 3-28 shows the 82599 behavior when going to power-up mode.

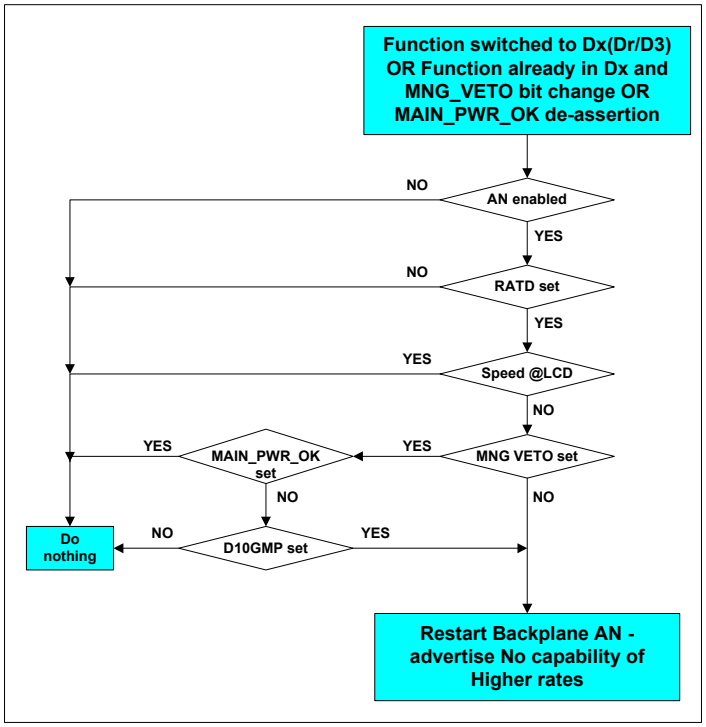


Figure 3-27 MAC Speed Change When Entering Power Down Mode

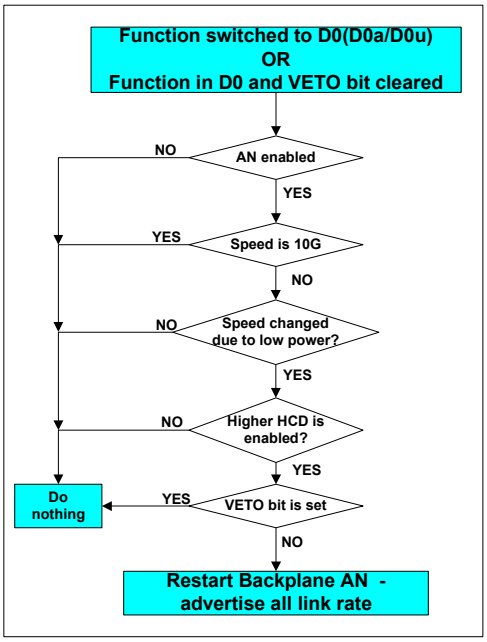


Figure 3-28 MAC Speed Change on Entering Power-up Mode

NOTE: ***This page intentionally left blank.***

4.0 Initialization

4.1 Power Up

4.1.1 Power-Up Sequence

The figure below shows the 82599 power-up sequence from power ramp up until the 82599 is ready to accept host commands.

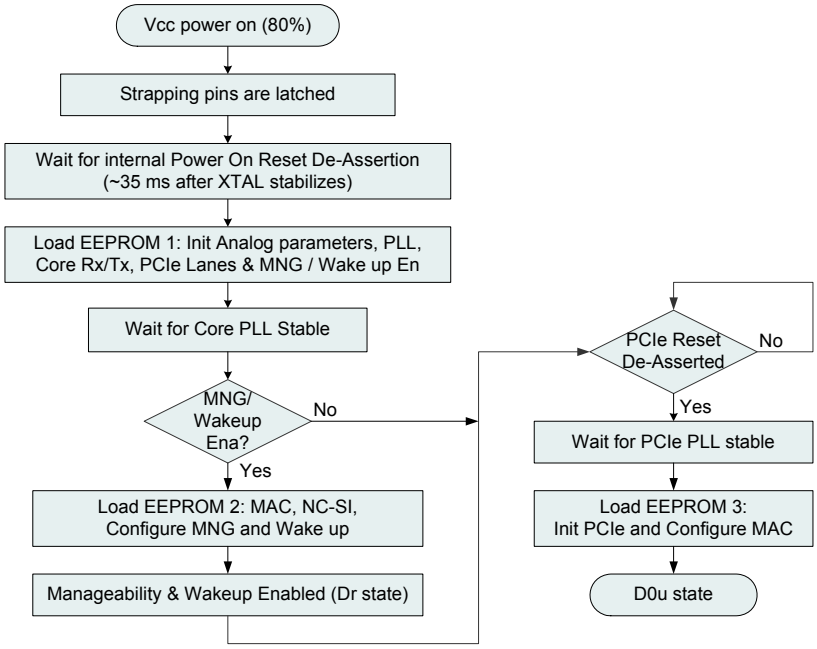


Figure 4-1 82599 Power-Up Sequence

4.1.2 Power-Up Timing Diagram

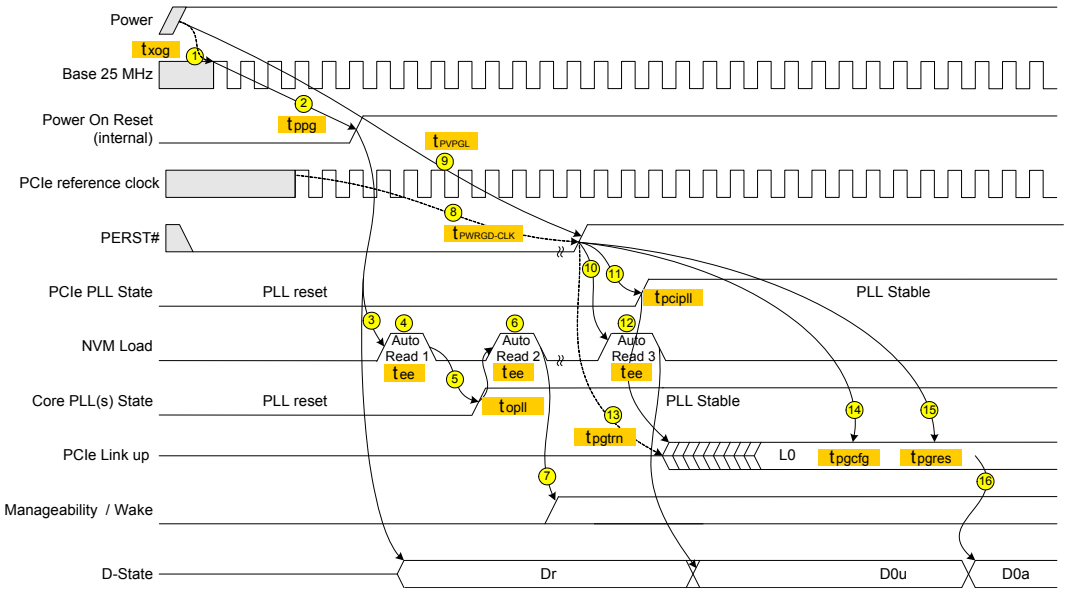


Figure 4-2 Power-Up Timing Diagram

Table 4-1 Notes for Power-Up Timing Diagram

Note	
1	Base 25 clock is stable t_{xog} after power is stable.
2	Internal Reset is released t_{ppg} after Base 25 is stable (also power supplies are good).
3	NVM read starts following the rising edge of the internal Power On Reset or external LAN Power Good.
4	EEPROM auto-load 1: EEPROM Init Section; PCIe Analog; Core Analog.
5	EEPROM auto-load 1 completion to Core PLL(s) stable — t_{opll}.
6	EEPROM auto-load 2: MAC module manageability and wake up (if manageability / wake up enabled).
7	APM wake up and/or manageability active, based on NVM contents (if enabled).
8	The PCIe reference clock is valid $t_{PWRGD-CLK}$ before the de-assertion of PCIe Reset (PCIe specification).
9	PCIe Reset is de-asserted t_{PVPGL} after power is stable (PCIe specification).
10	De-assertion of PCIe Reset invokes the EEPROM auto-load 3.
11	De-assertion of PCIe Reset to PCIe PLL stable t_{PCIPLL}.
12	EEPROM auto-load 3: PCIe General Configuration; PCIe Configuration Space; LAN Core Modules and MAC module if manageability is not enabled.
13	Link training starts after t_{pgtrn} from PCIe Reset de-assertion (PCIe specification).

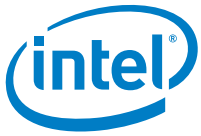


Table 4-1 Notes for Power-Up Timing Diagram (Continued)

Note	
14	A first PCIe configuration access might arrive after t_{pgcfg} from PCIe Reset de-assertion (PCIe specification).
15	A first PCI configuration response can be sent after t_{pgres} from PCIe Reset de-assertion (PCIe specification).
16	Setting the *Memory Access Enable* or *Bus Master Enable* bits in the PCI Command register transitions the 82599 from D0u to D0 state.

4.1.2.1 Timing Requirements

The 82599 requires the following start-up and power state transitions.

Table 4-2 Power-Up Timing Requirements

Parameter	Description	Min	Max.	Notes
t_{xog}	Base 25 MHz clock stable from power stable.		10 ms	
$t_{PWRGD-CLK}$	PCIe clock valid to PCIe power good.	100 µs	-	According to PCIe specification.
t_{PVPGL}	Power rails stable to PCIe Reset inactive.	100 ms	-	According to PCIe specification.
t_{pgcfg}	External PCIe Reset signal to first configuration cycle.	100 ms		According to PCIe specification.

Note: It is assumed that the external 25 MHz clock source is stable after the power is applied; the timing for that is part of txog.

4.1.2.2 Timing Guarantees

The 82599 guarantees the following start-up and power state transition related timing parameters.

Table 4-3 Power-Up Timing Guarantees

Parameter	Description	Min	Max.	Notes
t_{xog}	Xosc stable from power stable.		10 ms	
t_{ppg}	Internal power good delay from valid power rail.		35 ms	Use internal counter for external devices stabilization.
t_{ee}	EEPROM read duration.		20 ms	Actual time depends on the EEPROM content.
topll	EEPROM auto-load 1 completion to Core PLL(s) stable		10 ms	
tpciplI	De-assertion of PCIe Reset to PCIe PLL stable.		5 ms	
tpgtrn	PCIe Reset to start of link training.		20 ms	According to PCIe specification.
tpgres	PCIe Reset to first configuration cycle.	100 ms	1 sec	According to PCIe specification.

4.2 Reset Operation

4.2.1 Reset Sources

The 82599 reset sources are described in the sections that follow:

4.2.1.1 LAN_PWR_GOOD

The 82599 has an internal mechanism for sensing the power pins. Once the power is up and stable, the 82599 creates an internal reset, which acts as a master reset of the entire chip. It is level sensitive, and while it is 0b, all of the registers are held in reset. LAN_PWR_GOOD is interpreted to be an indication that device power supplies are all stable. LAN_PWR_GOOD changes state during system power up.

4.2.1.2 PE_RST_N (PCIe Reset)

The de-assertion of PCIe reset indicates that both the power and the PCIe clock sources are stable. This pin asserts an internal reset also after a D3cold exit. Most units are reset on the rising edge of PCIe reset. The only exception is the PCIe unit, which is kept in reset while PCIe reset is asserted (level).

4.2.1.3 In-band PCIe Reset

The 82599 generates an internal reset in response to a physical layer message from PCIe or when the PCIe link goes down (entry to polling or detect state). This reset is equivalent to PCI reset in previous (PCI) GbE controllers.

4.2.1.4 D3hot to D0 Transition

This is also known as ACPI reset. The 82599 generates an internal reset on the transition from D3hot power state to D0 (caused after configuration writes from D3 to D0 power state). Note that this reset is per function and resets only the function that transitioned from D3hot to D0.

4.2.1.5 Function Level Reset (FLR) Capability

The *FLR* bit is required for the Physical Function (PF) and per Virtual Function (VF). Setting of this bit for a VF resets only the part of the logic dedicated to the specific VF and does not influence the shared part of the port. Setting the PF *FLR* bit resets the entire function.

4.2.1.5.1 FLR in Non-IOV Mode

A FLR reset to a function is equivalent to a D0 –> D3 –> D0 transition with the exception that this reset doesn't require driver intervention in order to stop the master transactions of this function. FLR affects the device 1 parallel clock cycle from FLR assertion by default setting, or any other value defined by the *FLR Delay Disable* and *FLR Delay* fields in the PCIe Init Configuration 2 — Offset 0x02 word in the EEPROM.

4.2.1.5.2 Physical Function FLR (PFLR)

An FLR reset to the PF function in an IOV mode is equivalent to a FLR in non-IOV mode. All VFs in the PCIe function of the PF are affected.

The affected VFs are not notified of the reset in advance. The RSTD bit in the VFMailbox[n] is set following the reset (per VF) to indicate to the VFs that a PF FLR took place. Each VF is responsible to probe this bit (such as after a timeout).

4.2.1.5.3 Virtual Function FLR (VFLR)

A VF operating in an IOV mode can issue a FLR. The VFLR resets the resources allocated to the VF (such as disabling the queues and masking interrupts). It also clears the PCIe configuration for the VF. There is no impact on other VFs or on the PF.

Tx and Rx flows for the queues allocated to this VF are disabled. All pending read requests are dropped and PCIe read completions to this function can be completed as unsupported requests.

Note: Clearing of the *IOV Enable* bit in the IOV structure is equivalent to a VFLR to all the VFs in the same port.

Note: PF driver should clear the VF's VFMBMEM after a VFLR is detected.

4.2.1.6 Software Resets

4.2.1.6.1 Software Reset

Software reset is done by writing to the *Device Reset* bit of the Device Control register (CTRL.RST). The 82599 re-reads the per-function EEPROM fields after a software reset. Bits that are not normally read from the EEPROM are reset to their default hardware values.

Note: This reset is per function and resets only the function that received the software reset.

Fields controlled by the LED, SDP and Init3 words of the EEPROM are not reset and not re-read after a software reset.

PCI configuration space (configuration and mapping) of the device is unaffected. The MAC might or might not be reset (see Section 4.2.3).

Prior to issuing software reset, the driver needs to execute the master disable algorithm as defined in Section 5.2.5.3.2.