



Cortina Systems® IXF1104 4-Port Gigabit Ethernet MAC Flow Control

Application Note

22 December 2006

Document Number 305304

Revision 2.0

*Other names and brands may be claimed as the property of others.

© Cortina Systems, Inc. 2006

Contents

1.0	Introduction	3
1.1	IEEE 802.3x* Flow Control Mechanism	3
2.0	RX FIFO Flow Control	4
2.1	RX FIFO Threshold Determination	4
2.2	RX FIFO Depth Determination Factors to Avoid Overflow (Setting High Watermark).....	6
2.3	RX FIFO Depth Determination Factors to Avoid Underflow (Setting Low Watermark).....	7
2.4	RX FIFO Flow Control Registers	8
3.0	TX FIFO Flow Control	9
3.1	TX FIFO Threshold Determination.....	9
3.2	Setting TX FIFO High Watermarks	10
3.3	Setting TX FIFO Depth Setting Low Watermark	10
3.4	TX FIFO Flow Control Registers.....	10
4.0	External Pause Interface	11

Figures

1	Optimum RX FIFO High and Low Watermark Thresholds	5
2	Optimum TX FIFO High and Low Watermark Thresholds.....	9
3	External Pause Interface	11

Tables

1	RX FIFO Depth Determination Factors - High Watermark	6
2	RX FIFO Depth Calculations To Avoid Overflow (2000 Meter Fiber)	7
3	RX FIFO Depth Determination Factors (Low Watermark).....	7
4	RX FIFO Depth Calculations To Avoid Underflow (2000 Meter Fiber)	8
5	RX FIFO Flow Control Registers	8
6	TX FIFO Flow Control Registers	10

Revision History

Date	Revision	Description
22 December 2006	2.0	First release of this document from Cortina Systems, Inc.
25 February 2005	001	Initial release

1.0 Introduction

This document describes flow control implementation in the Cortina Systems® <Product/ProductFamily_Full> (hereafter called the IXF1104 MAC).

Flow control is the method used to throttle throughput to minimize data loss. This document focuses on flow control methods, including those implemented by the IXF1104 MAC and related system implications. The following topics are discussed:

- [Section 1.1, IEEE 802.3x* Flow Control Mechanism, on page 3](#)
- [Section 2.0, RX FIFO Flow Control, on page 4](#), including RX FIFO flow control features in the IXF1104 MAC
- [Section 3.0, TX FIFO Flow Control, on page 9](#), including TX FIFO flow control features in the IXF1104 MAC
- [Section 4.0, External Pause Interface, on page 11](#), including system implementation

The IXF1104 MAC sublayer performs the following two main functions:

- Data Encapsulation
- Media Access Management

First-In, First-Out (FIFO) queues perform IXF1104 MAC data encapsulation and media access management functions. FIFOs store frame data. Ideally, data is never lost due to limitations in network performance or FIFO sizing; however, data loss can occur when packet data is received faster than it is transmitted. If this occurs, the FIFOs might fill and overflow.

The IXF1104 MAC either overwrites data in the FIFOs (and loss of the oldest data occurs) or the IXF1104 MAC stops writing to the FIFOs (and loss of the newest data occurs). To avoid data loss, the IXF1104 MAC slows down the receive-data stream until the transmit stream catches up. Flow control throttles the receive data stream to keep the FIFOs from filling and overflowing. To manage flow control, it is necessary to balance the amount of throughput loss due to flow control versus the amount of data loss without flow control.

1.1 IEEE 802.3x* Flow Control Mechanism

The IEEE 802.3x flow control mechanism is accomplished within the IXF1104 MAC sublayer. The FIFO begins to fill as packets are received. Once the FIFO has reached a pre-programmed threshold, the IXF1104 MAC control sublayer signals an internal state machine to transmit a PAUSE frame. This signal (XOFF) informs the link partner to halt transmission for a specified length of time.

The IXF1104 MAC continues to transmit PAUSE frames with the programmed idle time, as long as the threshold is exceeded. If the FIFO level falls below the threshold prior to the idle time expiration, another PAUSE frame is sent with a zero time specified. This PAUSE frame (XON) informs the link partner to resume transmissions.

The IEEE 802.3x flow control mechanism can be implemented in three variants:

- Synchronous: transmits out PAUSE frames and responds to PAUSE frames
- Asynchronous: transmits out PAUSE frames but does not respond to received PAUSE frames
- Asynchronous: responds to received PAUSE frames but does not transmit out PAUSE frames.

The IXF1104 MAC supports all three options for enabling IEEE 802.3x flow control.

Note: Due to a known erratum (see the Cortina Systems® *IXF1104 4-Port Gigabit Ethernet Media Access Controller Specification Update*, document #278756), when operating in SerDes mode, the transmit MAC might lock-up when responding to PAUSE frames. This lock-up condition does not occur when operating in GMII or RGMII mode.

2.0 RX FIFO Flow Control

The IXF1104 MAC supports the IEEE802.3x flow control mechanism. The IXF1104 MAC RX FIFO flow control thresholds are fully programmable, along with the pause time (transmitted as data in the PAUSE frame).

- In full-duplex mode, flow control is implemented by transmitting PAUSE frames of a specified length to the link partner.
- In half-duplex mode, flow control is performed by deliberately forcing collisions on the line.

2.1 RX FIFO Threshold Determination

The following programmable RX FIFO thresholds (per port) are defined in the IXF1104 MAC:

- **The High Watermark Threshold:** The threshold above which flow control is implemented.
- **The Low Watermark Threshold:** The threshold below which the flow control is terminated.

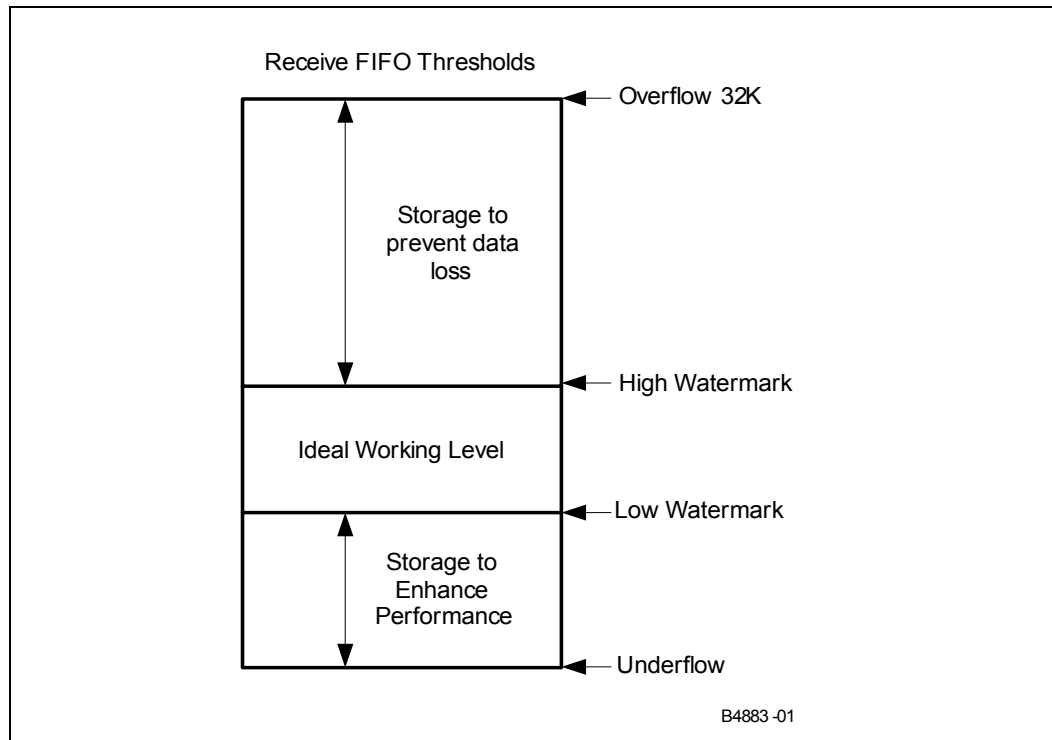
Proper FIFO threshold selection determines the effectiveness of the flow control implemented by the IXF1104 MAC.

To prevent loss of data, set the High Watermark Threshold low enough to ensure that the maximum amount of data that can be received is stored prior to flow control taking effect.

To obtain maximum throughput, set the Low Watermark Threshold high enough to keep the FIFO populated prior to flow control being released, and to limit the percentage of time that flow control is activated.

Figure 1 illustrates the relative levels for setting optimum RX FIFO High and Low Watermark Thresholds.

Figure 1 Optimum RX FIFO High and Low Watermark Thresholds



Maintaining the FIFO thresholds within the ideal working level avoids overflow and underflow conditions. To ensure that the optimal FIFO threshold is configured, the following system constraints must be considered:

- Packet Size
- Duplex Mode
- Link Speed
- Media Link Segment Length
- Media Link Segment Type
- MAC-PHY Latency

When selecting the RX FIFO thresholds for optimal throughput at a given loss rate, overflow and underflow conditions must also be considered. [Section 2.2](#) and [Section 2.3](#) contain examples of RX FIFO depth calculations that assume the possibility of Jumbo packets (packets that are more than 1500 bytes) across 2000 meters of fiber.

2.2 RX FIFO Depth Determination Factors to Avoid Overflow (Setting High Watermark)

To avoid an overflow condition, the FIFO must be able to store the amount of data that can be received prior to flow control taking effect. This amount of data is quantified as a combination of the factors in [Table 1](#).

Table 1 RX FIFO Depth Determination Factors - High Watermark

Factor	Description
Media Delay 1	The amount of time (in bytes) that the data transmitted by the link partner reaches the IXF1104 MAC receiver after the FIFO threshold is exceeded. This delay corresponds to data that was transmitted but not received (data currently traveling along the media).
Time Delay	The time delay (in bytes) caused by the length and type of media. All media have inherent time delays, and the length of the delay is dependent upon the type and length of the media. This delay can translate into large numbers of bytes, especially at fast speeds.
MAC Latency to Respond to an Over-Threshold Condition	The preparation time (in bytes) required to send a PAUSE frame.
Wait-to-Transmit	The elapsed time (in bytes) before beginning a new transmission. This is dependent upon the duplex mode and supported packet size. <ul style="list-style-type: none"> In half-duplex mode, the Wait-to-Transmission time is associated with receiving the remainder of the current packet before forcing a collision. In full-duplex mode, Wait-to-Transmission time is the length of time for the current transmission to end. The supported packet sizes can also impact the wait-to-transmit time. Large-sized packets might have longer wait times.
Inter Packet Gap (IPG)	The time (in bytes) to wait between transmissions.
Pause Packet Transmission	The amount of time (in bytes) to transmit the pause packet. The pause packet transmission time must be accounted for.
Media Delay 2	The amount of time (in bytes) for the pause packet to reach the link partner.
MAC-PHY Latency	The amount of time (in bytes) to transfer the pause packet through the PHY in route to the MAC.
IXF1104 MAC Reaction Time	The amount of time (in bytes) for the receiving IXF1104 MAC to react to the PAUSE frame. IEEE standards specify the maximum reaction time (refer to Table 4 on page 8 for the MAC reaction time value).
Packet Delay	The amount of time (in bytes) for the receiving IXF1104 MAC to complete the transmission of an existing packet before flow control takes effect. The receiving IXF1104 MAC might have started transmission of another packet; this transmission must be completed before flow control can take effect.

Table 2 is an example of FIFO depth calculations for a system receiving the maximum length, standard Ethernet frames at 1000 Mbps across 2000 meters of fiber while transmitting a Jumbo packet.

Table 2 RX FIFO Depth Calculations To Avoid Overflow (2000 Meter Fiber)

Delay Factor	Delay in Bytes
Media Delay that has already transmitted	1250
MAC Latency responding to threshold	10
Wait-To-Transmit for current transmission (Jumbo packet)	9843
IPG (IEEE specified)	12
Pause Packet (IEEE specified)	72
Media Delay (the delay for pause to reach the link partner)	1250
MAC-PHY Latency (the time through the PHY in route to the MAC)	4
MAC Reaction Time (IEEE specified)	128
Packet Delay when first starting to transmit	1530
Total - Overflow	14099 Bytes

The example in Table 2 estimates that the FIFO receives approximately 14 Kbytes of data after the High Watermark Threshold has been crossed. The RX FIFO of the IXF1104 MAC is 32 Kbytes in size; therefore, program the RX FIFO HIGH WATERMARK register for 18 Kbytes or lower, to prevent an overflow condition while transmitting a Jumbo packet. Adjustments to the calculations can be made for variations to line lengths, packet sizes, and rates.

2.3 RX FIFO Depth Determination Factors to Avoid Underflow (Setting Low Watermark)

To avoid an underflow condition, the FIFO must have enough stored data to continue transmitting during the flow control transmission. The amount of stored data must exceed a combination of factors (see Table 3):

Table 3 RX FIFO Depth Determination Factors (Low Watermark) (Sheet 1 of 2)

Factor	Description
MAC Latency to Respond to Under Threshold	The preparation time to send the new PAUSE zero time frame.
Wait-to-Transmit	The elapsed time (in bytes) before beginning a new transmission. This is dependent upon duplex mode and supported packet size: <ul style="list-style-type: none"> In half-duplex mode, transmission is nearly instantaneous; flow control is performed by deliberately forcing collisions on the line. In full-duplex mode, the amount of time (in bytes) to wait for the current transmission to end. In either case, the supported packet sizes impact the time. Large-sized packets might have longer wait times.
IPG	The time (in bytes) to wait between transmissions.
Pause Packet Transmission	The amount of time (in bytes) to transmit the pause packet; the pause packet transmission time must be accounted for.
Media Delay 1	The amount of time (in bytes) for the pause packet to reach the link partner.

Table 3 RX FIFO Depth Determination Factors (Low Watermark) (Sheet 2 of 2)

Factor	Description
MAC-PHY Latency	The amount of time (in bytes) to transfer the pause packet through the PHY in route to the MAC.
MAC Reaction Time	The amount of time (in bytes) for the receiving IXF1104 MAC to react to the PAUSE frame. The maximum is specified by IEEE.
Media Delay 2	The amount of time (in bytes) between sending data from one end and receipt of data at the other end. Once the link partner has begun retransmission, the delay through the media must be accounted for before the data reaches the other end.

Table 4 is an example of RX FIFO depth calculations for a system transmitting a Jumbo packet at a 1000 Mbps data rate across 2000 meters of fiber.

Table 4 RX FIFO Depth Calculations To Avoid Underflow (2000 Meter Fiber)

Delay Factor	Delay in Bytes
MAC Latency response to threshold	2
Wait-to-Transmit for the current Jumbo packer transmission	9843
IPG (IEEE specified)	12
Pause Packet (IEEE specified)	72
Media Delay 1 (delay for pause to reach the link partner)	1250
MAC-PHY Latency (time through the PHY in route to the MAC)	4
MAC Reaction Time (IEEE specified)	128
Media Delay 2 (time for new data to reach the IXF1104 MAC)	1530
Total Underflow	12839 Bytes

The FIFO transmit total in Table 4 calculates to 12839 bytes of transmitted data after the threshold is crossed. Program the RX FIFO Low Watermark Register for that amount or higher. This leaves enough memory to ensure that an underflow flow condition does not occur when transmitting a Jumbo packet. Adjustments to the calculations can be made for variations to line lengths, packet sizes, and rates.

2.4 RX FIFO Flow Control Registers

Table 5 defines the RX FIFO Flow Control Registers.

Table 5 RX FIFO Flow Control Registers

Register Name	Definition
RX FIFO High Watermark	FIFO level to initiate PAUSE frame transmission.
RX FIFO Low Watermark	FIFO level to end PAUSE frame transmission.
RX FIFO Transfer Threshold	Maximum packet size treated as store-and-forward. Packets above the values stored in this register are transferred before EOP is received.
FC TX Timer Value	The value inserted into the flow control frame.
PAUSE Threshold	Controls the time between PAUSE frame transmissions when more than one PAUSE frame is needed to keep the link partner paused.

3.0 TX FIFO Flow Control

TX Flow Control across the SPI3 interface is controlled by Transmit Packet Available (xTPA) signals as defined by the OIF-SPI3-01.0 specification.

The IXF1104 MAC implementation of xTPA signals uses a byte-level mode only. This signal is asserted when the TX FIFO occupancy is above the High Watermark, and de-asserted when the occupancy level drops below the Low Watermark. The assertion of the xTPA signals to the Link Layer device signifies that the TX FIFO is almost full (sometimes referred to as back pressure).

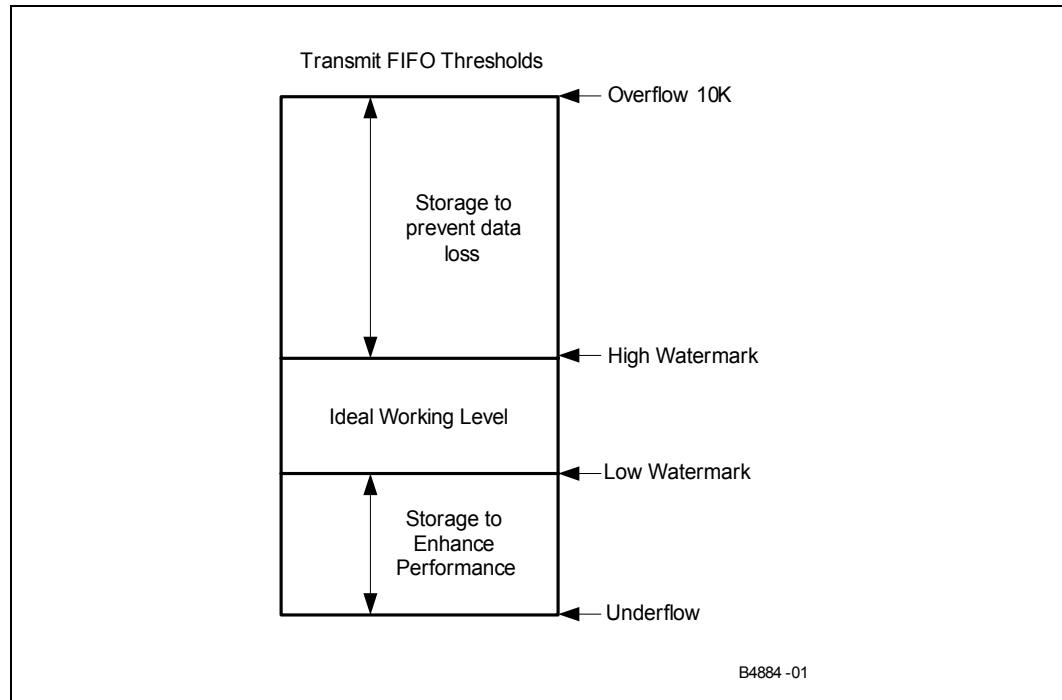
3.1 TX FIFO Threshold Determination

To prevent loss of data, set the High Watermark threshold low enough to ensure that the maximum amount of data that can be received across the SPI3 interface is stored prior to flow control taking effect.

To ensure maximum throughput, set the Low Watermark threshold high enough to keep the FIFO populated prior to the flow control being released, and to limit the percentage of time that flow control is activated.

Figure 2 illustrates the relative levels for setting optimum TX FIFO High and Low Watermark thresholds.

Figure 2 Optimum TX FIFO High and Low Watermark Thresholds



3.2 Setting TX FIFO High Watermarks

When the TX FIFO High Watermark threshold is crossed, back pressure is applied using sideband xTPA signals. Typically, these xTPA signals are not subject to delay times associated with transmitting or receiving packets. The xTPA signal is transmitted within a few byte times after the High Watermark threshold is crossed.

If the xTPA signals are accessed using a polled method (used with PTPA signals), the polling time must be considered. The maximum time required using the polled method is four clock cycles of the TFCLK on the SPI3 interface. This maximum time can result in a potential transfer of four bytes of data in SPHY mode or 16 bytes of data in MPHY mode.

When setting the TX FIFO High Watermark, include the latency to respond to back pressure, which depends on the System SPI3 link partner.

3.3 Setting TX FIFO Depth Setting Low Watermark

Back pressure applied to the SPI3 by the xTPA signal is de-asserted when the TX FIFO Low Watermark is crossed. To avoid a TX FIFO underflow condition, the FIFO must have enough stored data to continue transmitting during the time required for the System SPI3 link partner to respond to the xTPA signal. Include polling and system latency times.

If a Polled back pressure signal (PTPA) is used, the time required to poll all active ports must be considered in addition to any other system latencies.

3.4 TX FIFO Flow Control Registers

Table 6 defines the TX FIFO Flow Control Registers.

Table 6 TX FIFO Flow Control Registers

Register	Definition
TX FIFO High Watermark	FIFO threshold to assert xTPA back pressure signals.
TX FIFO Low Watermark	FIFO threshold to de-assert xTPA back pressure signals.
RX FIFO Transfer Threshold	Maximum packet size treated as store-and-forward. Packets above the values stored in this register begin transferring before EOP is received.

4.0 External Pause Interface

In addition to flow control, the IXF1104 MAC implements a system-pause control mechanism to allow for higher-layer devices to signal the need to transmit a PAUSE frame.

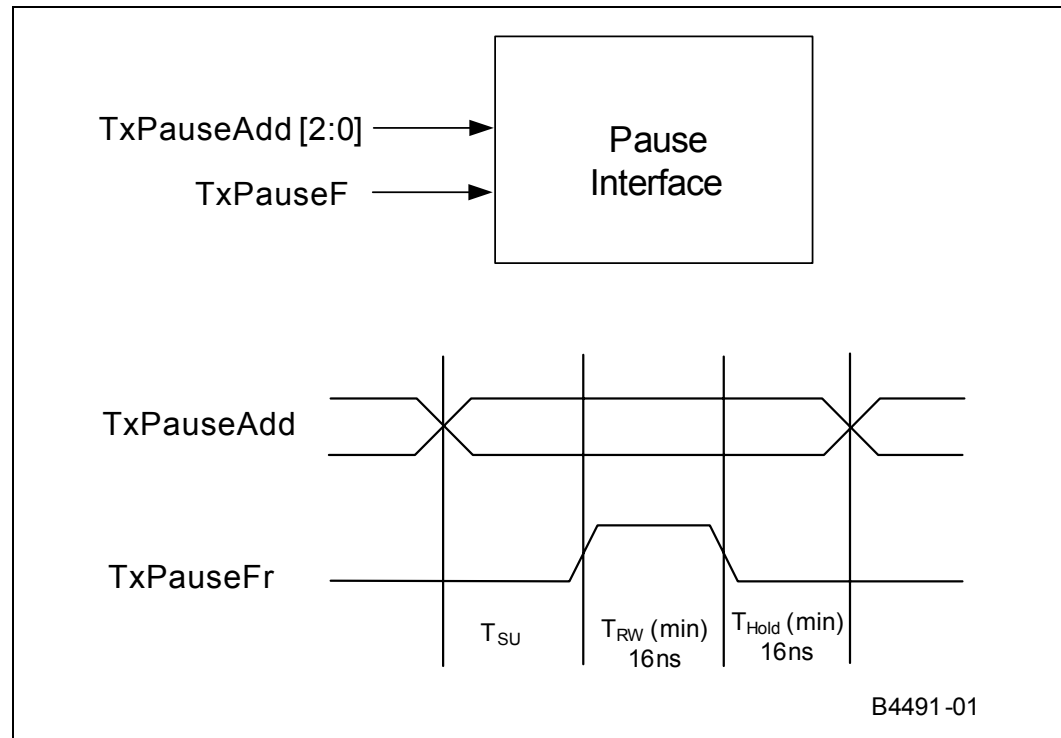
The system-pause control interface allows higher-layer devices to select the following:

- All ports send an XOFF PAUSE frame with a programmed pause time.
- All ports send an XON PAUSE frame with a zero time specified.
- Individual ports send XOFF PAUSE frames with the programmed pause time.

Implementing the External Pause interface forces the link partner to halt transmission for the specified pause time (see [Figure 3](#)). This mechanism is operational only when automatic flow control is enabled.

The External Pause interface is not used to respond to internal RX FIFO conditions, but can be implemented when data flow must be controlled for reasons other than internal RX FIFO occupancy levels. Each pause packet request using the External Pause interface initiates the transmission of a single PAUSE frame. Externally generated PAUSE frames use the same pause quanta value stored in the FC TX Timer Value.

Figure 3 External Pause Interface





For additional product and ordering information:

www.cortina-systems.com