



ENET Switch Fabric Interposer with IXP2400

Optimized switching for scalability and QoS

Erlang Technology
345 Marshall Avenue
Suite 300
St. Louis, MO 63126
Phone: (314) 336-5900
Fax: (314) 336-5902
Web: <http://www.erlangtech.com>
e-mail: info@erlangtech.com

Copyright © 2001 by Erlang Technology, Inc.

All Rights Reserved.

No part of this document may be reproduced, stored, retrieved, or transmitted by any means, electronic, mechanical, photocopying, recording, otherwise, without written prior permission of Erlang Technology, Inc.,

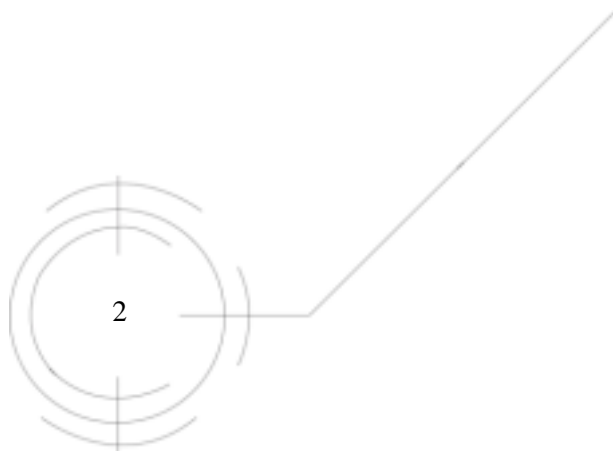
St. Louis, Missouri, USA.

All terms mentioned in this document that are known to be trademarks or service marks have been appropriately capitalized. Erlang Technology, Inc. cannot attest to the accuracy of this information. Use of a term in this document should not be regarded as affecting the validity of any trademark or service mark. Trademarks are property of their respective owners.

Erlang Technology, Inc. reserve the right to change, amend, or reissue this document, or any part thereof, at its sole discretion and without further advice to any party. Any specification contained herein is subject to change without notice.

Table of Contents

1	INTRODUCTION.....	3
2	Intel IXP2400.....	3
3	Erlang’s ENET and IXP2400 Configuration	3
3.1	Intel IXP2400 with XeI.....	3
3.1.1	NP Interface Pins.....	4
3.1.2	Maximum Transmission Unit (MTU).....	6
3.2	CSIX-L1 Mode	6
3.2.1	Conformity to the NP Forum Specification	6
3.2.2	Supported CFrames Types	7
3.2.3	Parity Checking.....	7
3.2.4	Flow Control	7
3.2.5	Ingress (NP-to-Fabric) Flow Control.....	8
3.2.6	Egress (Fabric-to-NP) Flow Control.....	8



1 INTRODUCTION

The ENET products have various interfaces to network processors including CSIX_L1, POS-PHY level 3 (SPI-3), and Agere APC format. The Intel's IXP2400 network processor depends on CSIX_L1 interface to switch fabrics or TMs. The XeI and the SeI-CSIX could be used as an interposer to connect to the IXP2400.

2 Intel IXP2400

The IXP2400 supports the rate up to 2.5Gbps of user traffic. Apart from Intel's earlier IXP12** product, IXP2400 supports Utopia/SPI-3 for Media interface and CSIX_L1 for Fabric interface. A single set (ingress/egress) of user configurable Media/Switch Fabric interface is equipped in a single IXP2400 device. The ability of the traffic handling rate of IXP2400 makes it ideal for a wide variety of high-performance applications such as Wide Area Networking (WAN) multi-service switches, DSL access multiplexers (DSLAMs), Cable Modem Termination System (CMTS) equipment, 2.5G and 3G wireless access network elements and core network elements, and Layer 4-7 switches. Refer to Intel (<http://www.intel.com>) for detailed applications of IXP2400.

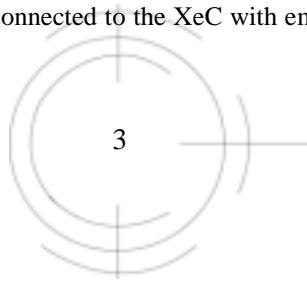
3 Erlang's ENET and IXP2400 Configuration

The ENET devices for IXP2400 are the SeI-CSIX and the XeI. The SeI-CSIX has two sets of CSIX network processor ports which support up to two sets of IXP2400. The XeI has four CSIX network processor ports which support up to four sets of IXP2400. The detailed design examples described here are just a subset of the all the possible designs. Focus is on showing some tangible circuit configurations in real environments.

The ENET 2nd generation switch fabric device set, Xe, gives much more benefits especially when used in bigger system configurations. Due to the embedded Serdes inside Xe, the system design could be easier with less device counts in addition to the performance benefits of Xe.

3.1 Intel IXP2400 with XeI

The XeI is an interposer for the XeC. The XeI provides four sets of IXP2400 network processors interfaces. The interface is shown in Figure 3-1, where no glue logic is required between the IXP2400 and the XeI. For a 10Gbps configuration, the XeI does not need the XeC. A single XeI could be a single 10Gbps switch. In Figure 3.1, the XeI functions as a single device 10Gbps switch along with four sets of IXP2400 network processors. The interface for the XeC should be disabled to save the power, in this configuration. To form a bigger capacity system, the XeI should be connected to the XeC with embedded Serdes which could run over the backplane.



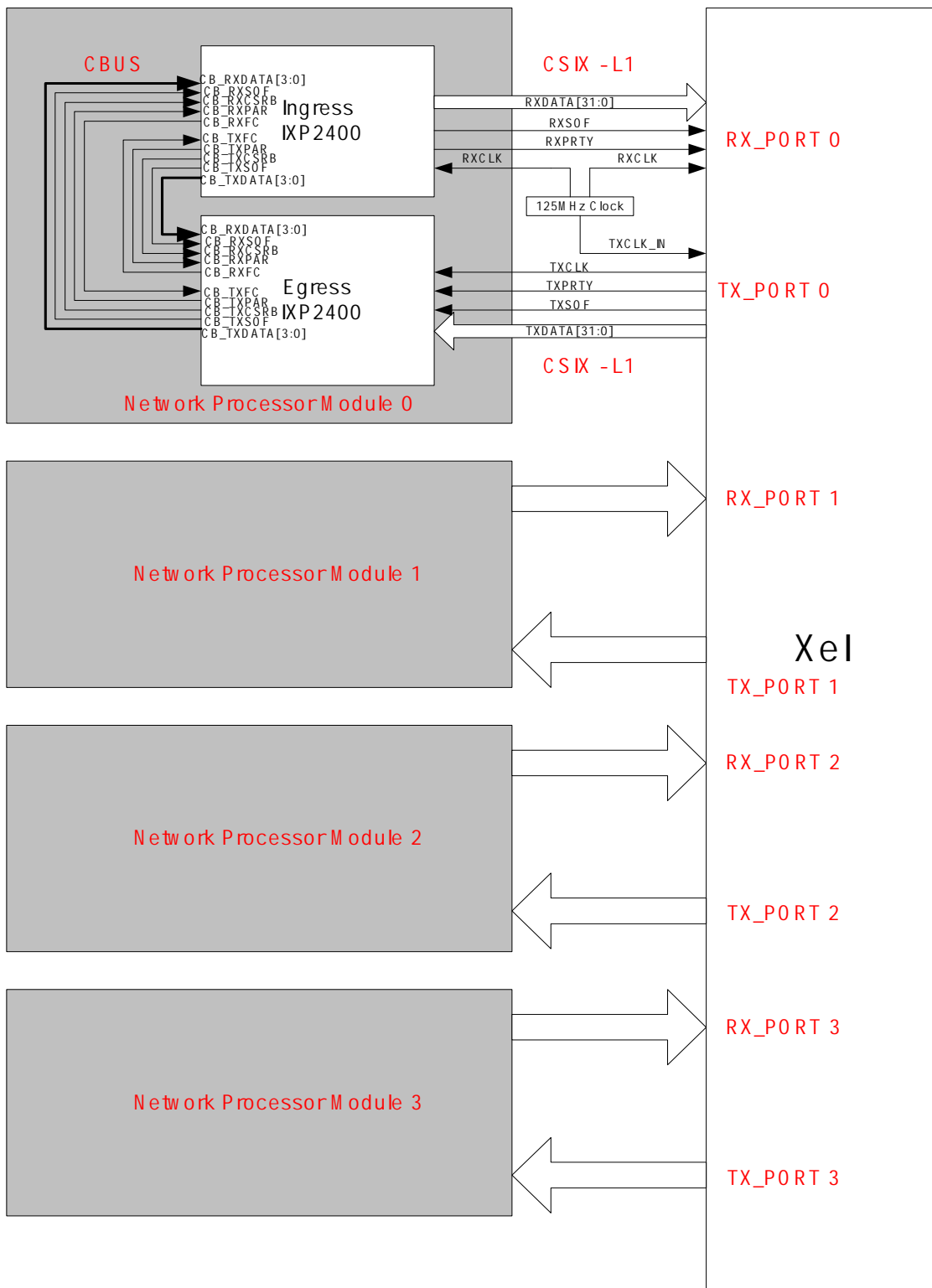


Figure 3-1 CSIX-L1 connectivity between the XeI and the Network Processor

3.1.1 NP Interface Pins

The NP interface pins of XeI assume different functions depending upon the configuration of the NP port.

Table 3-1 and Table 3-2 list the function of each pin under various configurations for one of the four NP ports. Therefore, there are four sets of such pins denoted with the prefix “NPn_”, where n takes the value of 0 to 3. Possible configurations of NP include CSIX-L1, and either the *Framer* or the *Link Layer* side of SPI-3 protocol.

Table 3-1: I/O pins of one Network Processor (NPn; n = 0, 1, 2 and 3) in the *ingress* flow.

Xel Pin Name	I/O	SPI-3 Mode (Framer)	SPI-3 Mode (Link Layer)	CSIX-L1 Mode
NPn_ICLK	I	NPn_TFCLK	NPn_RFCLK	NPn_RxCiK[0]
NPn_IERR	I	NPn_TERR	NPn_RERR	N/A
NPn_IDATEN	I	NPn_TENB	NPn_RVAL	N/A
NPn_IDAT[31:0]	I	NPn_TDAT[31:0]	NPn_RDAT[31:0]	NPn_RxData[31:0]
NPn_IPRTY	I	NPn_TPRTY	NPn_RPRTY	NPn_RxPar[0]
NPn_IMOD[1:0]	I	NPn_TMOD[1:0]	NPn_RMOD[1:0]	N/A
NPn_ISOP	I	NPn_TSOP	NPn_RSOP	NPn_RxSOF[0]
NPn_IEOP	I	NPn_TEOP	NPn_REOP	N/A
NPn_IPAUSE	O	NPn_TPA	NPn_RENB	N/A

As noted in Table 3-2, some non-standard clocking schemes are available as optional features in addition to the requirements of CSIX-L1 and SPI-3. For instance, the output clock (NPn_ECLKOUT) may be used in the SPI-3 mode to make the interface source-synchronous. Likewise, the input clock (NPn_ICLK) may be used in the CSIX-L1 mode to adopt a clocking scheme similar to the CSIX interface of Intel IXP2400.

Table 3-2: I/O pins of one Network Processor (NPn; n = 0, 1, 2 and 3) in the *egress* flow.

Xel Pin Name	I/O	SPI-3 Mode (Framer)	SPI-3 Mode (Link Layer)	CSIX-L1 Mode
NPn_ECLK	I	NPn_RFCLK	NPn_TFCLK	<i>NPn_TxCiK[0]</i> ¹
NPn_ECLKOUT	O	<i>NPn_RFCLKOUT</i> ²	<i>NPn_TFCLKOUT</i>	NPn_TxCiK[0]
NPn_EDATEN	O	NPn_RVAL	NPn_TENB	N/A
NPn_EDAT[31:0]	O	NPn_RDAT[31:0]	NPn_TDAT[31:0]	NPn_TxData[31:0]
NPn_EPRTY	O	NPn_RPRTY	NPn_TPRTY	NPn_TxPar[0]
NPn_EMOD[1:0]	O	NPn_RMOD[1:0]	NPn_TMOD[1:0]	N/A
NPn_ESOP	O	NPn_RSOP	NPn_TSOP	NPn_TxSOF[0]
NPn_EEOP	O	NPn_REOP	NPn_TEOP	N/A
NPn_EPAUSE	I	NPn_RENB	NPn_TPA	N/A

¹ NpnECLK is an optional *input clock* (“non-standard”) for CSIX.

² NpnECLKOUT is an optional *output clock* (“non-standard”) for SPI-3.

3.1.2 Maximum Transmission Unit (MTU)

The maximum transmission unit (maximum packet length) is 16 K-Byte in the SPI-3 mode. It may be necessary in some system configurations to limit the MTU size to a smaller value, in which case the user can set a register called *MaxPktLenInWords[13:0]* (Address 0x09) accordingly. As the name implies, the maximum packet length is specified in 4-Byte (32-bit) words, where the value 0 means one word (4 Bytes) and maximum value of 0x3FFF means 16K (16318) Bytes. Packets longer than *MaxPktLenInWords* will either be truncated or discarded entirely, depending on the settings of *PreservePktEn* bit of the main Control register (Address 0x01). The default value of zero for this bit instructs XeI to discard the entire packets whose length exceeds the limit.

By definition, the MTU size is limited to 256 Bytes in CSIX-L1 protocol, as there are only 8 bits in the CFrame to specify its length. Therefore, the ports that are in CSIX configuration ignore the *MaxPktLenInWords* register.

3.2 CSIX-L1 Mode

Each of the four NP ports of XeI can be configured independently as a bi-directional Common Switch Fabric Interface ([NP Forum CSIX-L1](#)). The CSIX implementation of XeI is fully compliant with the published spec. Additionally, some non-standard features that have been implemented by other companies are available as optional features. These features will be discussed **Error! Reference source not found.**

In the CSIX mode the NP and XeI exchange variable-length data and flow control frames. Figure 3-2 illustrates the connectivity of XeI and the Network processor in the CSIX-L1 protocol. As shown in Table 3-1 and Table 3-2, the CSIX interface uses a subset of the CMOS I/O pins to provide this connectivity.

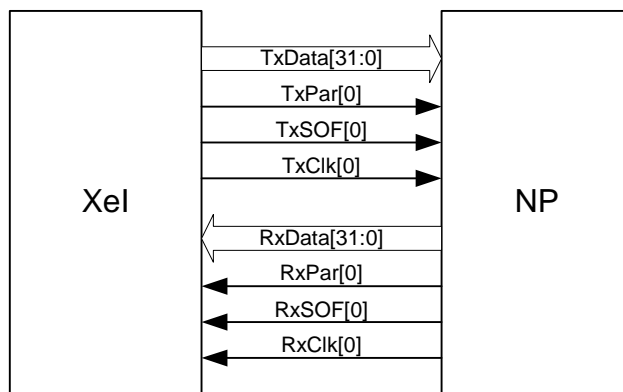


Figure 3-2 CSIX-L1 connectivity between the XeI and the Network Processor

3.2.1 Conformity to the NP Forum Specification

The XeI implementation meets the conformance requirements of the CSIX-L1 specification (Chapter 9).

The 32-bit interface uses 3.3V-Tolerant LVCMOS I/O pads, which can run at the 125 MHz. The interface can run at 133 MHz if the optional NPn_ECLK is used to provide an external input clock to the chip, which is the same as the clocking scheme of the CSIX interface of Intel IXP2400 network processor. However, in the standard mode, where XeI outputs the CSIX interface clock (NPn_ECLKOUT), the maximum frequency will be limited to 125 MHz, which is still compliant with the spec.

3.2.2 Supported CFrame Types

Support of multicast CFrame types is optional in the CSIX-L1 specification. The XeI does not support “Multicast ID” (CFrame Type 3), which requires a lookup process in the fabric to obtain the routing information. All the other types of CFrame are supported, which includes Idle, Unicast, Multicast Mask, Multicast Binary Copy, Broadcast and Flow Control CFrame. The format of various CFrame types is not described in this datasheet, because they are compatible with the [NP Forum CSIX-L1](#).

3.2.3 Parity Checking

The XeI supports both the horizontal and vertical parity, both of which can be disabled or optionally configured to use even parity for testing purposes. The *NOParityEn* (Bit 0) and *NOVertParityEn* (Bit 2) of the *NOConfig* register (Address 0x0E to 0x11) may be set to zero independently for each port to disable the horizontal or vertical parity respectively. For testing purposes, the user may set the *NOEvenParity* (Bit 1) and *NOEvenVertParity* (Bit 3) of the *NOConfig* registers to one to indicate that “even” parity must be used for the corresponding port.

If any horizontal or vertical parity error is detected in a CFrame, the entire frame will be discarded and the ready bits will be ignored.

3.2.4 Flow Control

The XeI implementation of CSIX Flow Control (FC) Frames is fully compliant with the CSIX-L1 spec at the link-level, as well at the higher level between the NP and the fabric.

The XeI uses a simple off/on scheme to control the traffic flow using binary Speed values. That is, the incoming 4-bit Speed variable will be interpreted as a pause command only when it is set to zero. Any non-zero value of Speed is interpreted as a resume command for the corresponding traffic flow. Likewise, the Speed value of zero is used by XeI to indicate that a flow must be paused, and 0xF will be used to resume the corresponding flow.

3.2.4.1 Link-Level CSIX Flow Control

CSIX-L1 uses the two-bit Ready field of every CFrame to stop and resume the data and flow control between the switch fabric and the network processor. These two Ready bits of the CFrames that are sent from NP to XeI, which indicate whether the *Egress NP* is ready to receive traffic are referred hereinafter as *EReady[1:0]* bits. Similarly, the Ready bits in the CFrames that are sent from XeI to NP, which indicate whether the *Ingress FIFOs* of XeI can receive traffic are referred hereinafter as *IReady[1:0]* bits.

XeI will always be ready to receive flow control CFrames and will never attempt to stop the control traffic.

Therefore, it will always assert the Control bit *IReady[0]* of outgoing CFrames. The *IReady[1]* bit is used by XeI to stop the data traffic when the ingress FIFO levels exceeds a configurable threshold. The threshold may be changed in the *NOFifoBpThr* register (Address 0x18) for all four NP ports.

At the ingress, on the other hand, both EReady bits are extracted from the incoming CFrames to stop and resume the control and data traffic from XeI to the NP. The response time of XeI to deassertion of EReady bits is on the order of three to four CFrames.

3.2.4.2 Preventing Deassertion of Ready Bits on Errors

The CSIX-L1 requires both Ready bits to be deasserted if any error such as horizontal parity error is detected in any CFrame. This behavior may not be desirable in systems with redundant switch fabric (including interposers) for fault tolerance. The failure of the ingress flow, in such systems, will trigger the switchover to the redundant interposer. However, it may not be desirable to stop the egress data flow, which would prevent potentially valuable data from draining out of the shared buffer of the switch fabric with a faulty ingress port.

Therefore, the user is allowed to disable this feature for any given port independently by setting the *NOClrEReadyOnErr* field of *NOConfig* register (Address 0x0E to 0x11) to zero. When this bit is set to 0, the ready bits of CFrames with any errors will be disregarded, but the previous state of EReady bits is maintained until new error-free CFrames are received. If the *NOClrEReadyOnErr* bit is set to one, the ready bits of CFrames with any errors are disregarded and current state of EReady will be cleared to 0, *i.e.* “not ready”.

3.2.5 Ingress (NP-to-Fabric) Flow Control

The FC Entries are extracted from error-free FC frames coming to the fabric from the NP to control the flow of traffic to the NP. The XeI treats the FC entries contained in one frame as independent messages processed sequentially, which could possibly be contradictory. That is, if more than one FC entries specify different speed values for the same flow, the last FC entry will be honored and the previous ones will be ignored. Also, if XeI detects any error associated with an FC frame, it ignores all of the FC entries contained in that specific FC frame.

The XeI does not respond to traffic-type-specific FC Entries and can only control all types of traffic together for a given priority. The reason is that the XeI used priority-based output queues that are not type-specific. Therefore, only FC entries of Type “All” (type encoding value of 3) for a given class (or class wildcard) are considered and any other FC entries will be ignored.

3.2.6 Egress (Fabric-to-NP) Flow Control

The XeI creates and sends FC frames to each NP in order to control the flow of data traffic from the corresponding NP independently. The Class and Port wildcard types are also used to convey the information more efficiently. The conditions upon which different types of FC Frames are generated are configurable via the CPU interface and will be discussed in detail in the backpressure section.

Since there is a possibility for flow control messages being lost between XeI and the NP, the state backpressure in NP is constantly updated by sending some refresh messages at a low rate. These messages

are referred herein as the *refresh* messages. When a the status of a queue inside the fabric changes, a message is sent to NP to update its state accordingly. These message, which are indeed more urgent than the refresh messages, are referred herein as *urgent* messages.

The urgent messages are given a higher priority than the refresh messages in the flow control scheduler, such that no refresh messages will be served as long as there are some urgent messages to be sent to the NP. In the absence of urgent messages, a configurable portion of the bandwidth on the CSIX bus is dedicated to the refresh messages. The *CpCsixRefreshRate[7:0]* register (Address 0x0C) specifies the bandwidth allocated to the slow flow control refresh process in the absence of any urgent messages. The bandwidth is expressed as a fraction of 1024, with the default value of 16. The default value allocates 1.56% (16/1024) of the bandwidth on the bus for the refresh flow control messages. The desired bandwidth may be different for different aggregate capacities in the system and depending on the flow control latency requirements. Writing zero to this register effectively disables the refresh process, which is not recommendable.

