Xilinx Embedded RDMA Enabled NIC v1.0

PB059 (v1.0) February 6, 2019 **LogiCORE IP Product Brief**

Introduction

The Xilinx® Embedded RDMA Enabled NIC LogiCORE IP is an implementation of RDMA over Converged Ethernet (RoCE v2) enabled NIC functionality. This parameterizable soft IP core can work with a wide variety of Xilinx hard and soft MAC IP implementations providing a high through-put, low latency, and completely hardware offloaded reliable data transfer solution over standard Ethernet. The ERNIC IP allows simultaneous connections to multiple remote hosts running RoCE v2 traffic.

Additional Documentation

A full product guide is available for this core. Access to this material can be requested by clicking on this registration link: https://www.xilinx.com/member/rnic.html.

Features

- Support for RDMA functionality
  - RoCE v2
  - Packet re-transmission on errors in hardware
- 100 Gb/s data path
- Support for hardware based reliable connection
- Hardware handshake on user interface
- Supports incoming, and outgoing RDMA SEND, RDMA READ, and RDMA WRITE
- Does not support incoming atomic operations
- Designed to scale up to 255 RDMA Queue pairs
- Support for IPv4 and IPv6 packets
- Support for Explicit Congestion Notification (ECN)
- Supports memory Protection Domains as described in Section 3.5.5 of InfiniBandTM architecture Specification Volume 1 Release 1.2.1

IP Facts

LogiCORE™ IP Facts Table	
Core Specifics	
Supported Device Family[1]	Kintex® UltraScale+™, Virtex® UltraScale™, Virtex UltraScale+, and Zynq® UltraScale+
Supported User Interfaces	AXI4-Lite, AXI4-Stream, and AXI4-Stream
Resources	Performance and Resource Use web page (registration required)
Provided with Core	
Design Files	Encrypted RTL
Example Design	Verilog
Test Bench	Not Provided
Constraints File	Xilinx Constraints File
Simulation Model	Not Provided
Supported S/W Driver[1]	N/A
Tested Design Flows[2]	
Design Entry	Vivado® Design Suite and Vivado IP Integration
Simulation	For supported simulators, see the Xilinx Design Tools: Release Notes Guide.
Synthesis	Vivado Synthesis
Support	
Provided by Xilinx at the Xilinx Support web page	

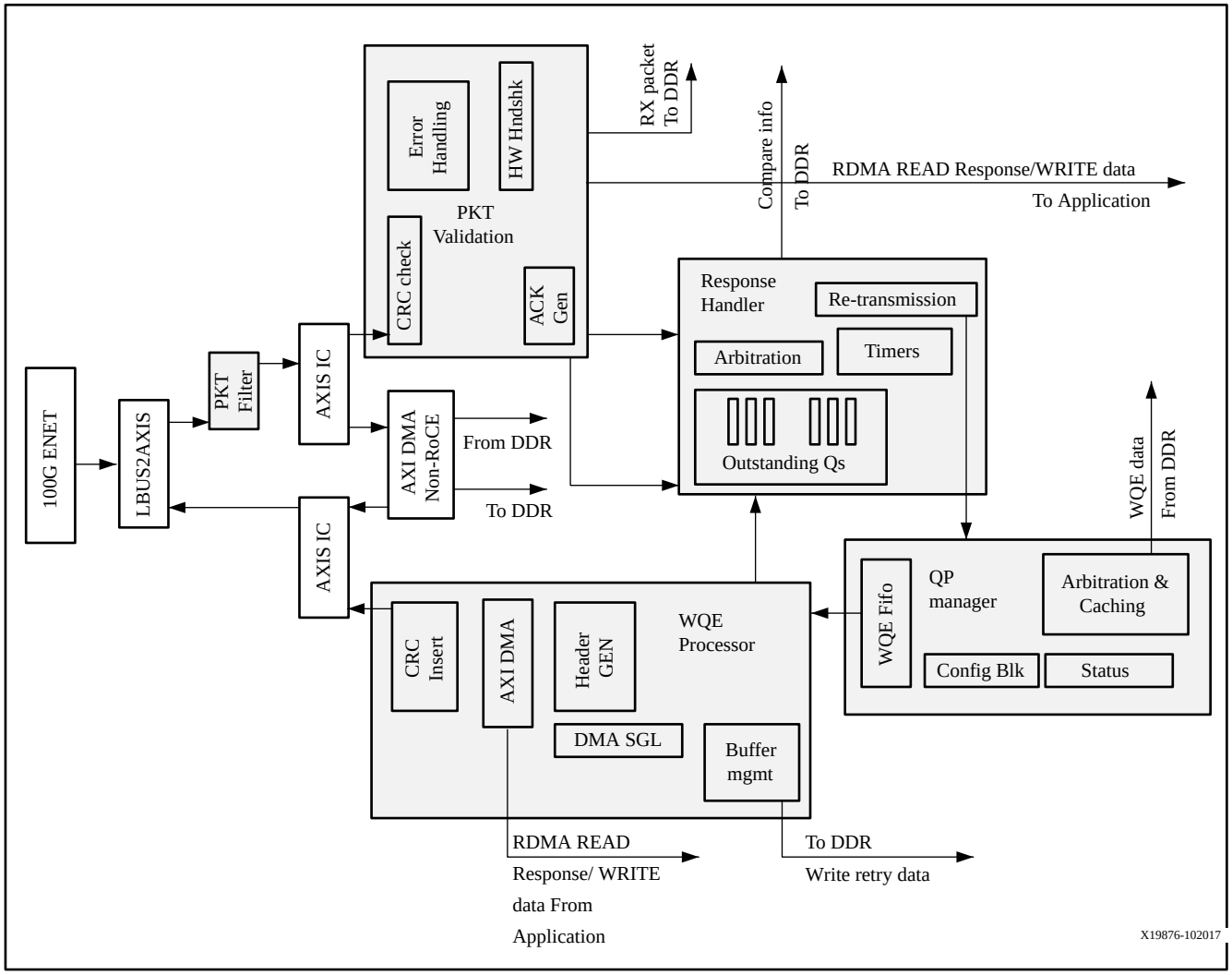
Notes:

1. For a complete list of supported devices, see the Vivado® IP catalog.
2. For the supported versions of the tools, see the Xilinx Design Tools: Release Notes Guide.
3. For UltraScale devices, only 40G data path is supported.

Overview

This chapter contains an overview of the core and details of applications, licensing, and standards. ERNIC is a soft IP implementing RDMA over a Converged Ethernet (RoCE v2) protocol for embedded target or initiator devices. This implementation is based on the specifications described in InfiniBand™ Architecture Specification Volume 1, Annex A16 RoCE and Annex 17 RoCE V2. The following figure shows the ERNIC IP and its connections to the other IPs in the system.

Figure 1: **ERNIC IP and Subsystem**



***Note*:** The user logic or target IP that connects to ERNIC is referred to as application and the direction of the arrows is from master to slave.

Apart from the ERNIC IP, the ERNIC subsystem includes the Xilinx Ethernet IP, AXI DMA, and AXI Interconnect among other IPs. On the user application front, the ERNIC IP exposes side band interfaces to allow efficient doorbell exchanges without going through the interconnect. Each queue is identified with a set of read and write pointers called the Producer Index (write pointer) and Consumer Index (read pointer). The register address locations for these pointers are termed as doorbells in this document. A doorbell exchange or doorbell ringing indicates that the corresponding register location is updated.

Feature Summary

The ERNIC IP interfaces with any Ethernet mac IP using an AXI Streaming interface. Access to DDR or any other memory region is necessary for reading and writing various data structures for RDMA packet processing. This connection is achieved using multiple AXI4 Full interfaces. The IP works on a 512 bit internal datapath that is completely hardware accelerated and does not require any software intervention for data transfer. All recoverable faults like re-transmission due to packet drops are also handled entirely using hardware.

The ERNIC IP implements embedded RNIC functionality. As a result, the following subset of RoCE v2 functionality is implemented (an Embedded RDMA NIC and not a general purpose NIC):

- Support for RDMA SEND, RDMA READ and RDMA WRITE for incoming and outgoing packets. Atomic operations are not supported. RDMA SEND with immediate and RDMA WRITE with immediate opcodes are not supported.
- Support for RDMA READ, RDMA WRITE, and RDMA SEND work requests
- Support for up to 254 connections
- Scalable design of up to 255 RDMA Queue pairs

***Note*:** Default Vivado strategies allow for the timing to pass up to 127 queue pairs. To match the timing for 255 queue pairs, use the Vivado strategy — Performance_refinePlacement.

- Supports dynamic memory registration with proprietary APIs
- Hardware handshake mechanism for efficient doorbell exchange with the user application

***Note*:** When switching in the handshake mode, the upper layer should not have any traffic on that particular QP.

ERNIC Modules

The ERNIC IP consists of four main modules that are explained in this section.

- QP Manager
- WQE Processor Engine
- RX PKT Handler
- Response Handler

- **QP Manager:** The QP Manager module houses the configurations for all the QPs and provides an AXI Lite interface to the processor. It also arbitrates across the various SEND Queues and caches the SEND Work Queue Entries (WQEs). These WQEs are then provided to the WQE processor module for further processing. This module also handles the QP pointer updates in the event of re-transmission.

- **WQE Processor Engine:** The WQE Processor Engine reads the cached WQEs from the QP Manager module and handles the following tasks:

  - Validates the incoming WQE packets for any invalid opcode
  - Creates the header for the packets based on the Payload Max Transfer Unit (PMTU) and programs the Scatter Gather Lists (SGLs) for the internal DMA engine

- Triggers the DMA to start the outgoing packet transfers

The WQE Processor Engine is also responsible for sending outgoing acknowledgment packets for the incoming RDMA SEND/WRITE requests and read responses for incoming RDMA READ requests.

- **RX PKT Handler:** The RX PKT Handler module receives the incoming packets. The ERNIC IP handles the following types of incoming RoCE v2 packets:
  - RDMA SEND, RDMA WRITE, RDMA READ and response packets for RDMA READ (request sent from ERNIC)
  - Acknowledgment packets for RDMA WRITE/RDMA SEND (request sent from ERNIC)
  - Communication management (Management Datagram) packets to QP1

  The RX PKT Handler module is responsible for validating the incoming packets. It also triggers outgoing acknowledgment packets for incoming RDMA SEND and RDMA WRITE requests and pushes the packets that pass the validation to the corresponding memory location. The RDMA READ responses are channeled to the target application directly. The module handles the incoming RDMA READ requests and forwards the request to the Tx path.
- **Response Handler:** The Response Handler module manages the outstanding queues. These queues hold the information about all packets sent to the remote host but have not yet been acknowledged or responded to. In addition, this module triggers a re-transmission if the remote host sends a Negative Acknowledgment (NAK). If this module does not receive a response from the remote host within a specified time (timeout value), it triggers a timeout related re-transmission.

Applications

The ERNIC IP can be used in abroad range of applications that require reliable transfer of packets across the network fabric. A few such applications are listed here:

- Sensor Data Acquisition and transfer over RoCE V2
- Video/Image capture and transfer over RoCE V2
- Remote storage nodes over RoCE V2

Unsupported Features

The ERNIC IP does not support these operations:

- Incoming ATOMIC operations
- Outgoing ATOMIC operations and OFED Stack APIs
- Incoming/outgoing RDMA SEND packets of 0 length
- Incoming/outgoing RDMA SEND with invalidate (NA)
- Incoming/outgoing RDMA SEND with immediate
- Incoming/Outgoing RDMA WRITE with immediate
- Maximum RQ buffer size supported is 8 MB

- Incoming RDMA READ requests with DMA Length equal to or less than four
- Incoming RDMA READ/WRITE requests with virtual address not on 64 byte boundary

Technical Support

Xilinx provides technical support on the Xilinx Community Forums for this LogiCORE™ IP product when used as described in the product documentation. Xilinx cannot guarantee timing, functionality, or support if you do any of the following:

- Implement the solution in devices that are not defined in the documentation.
- Customize the solution beyond that allowed in the product documentation.
- Change any section of the design labeled DO NOT MODIFY.

To ask questions, navigate to the Xilinx Community Forums.

Licensing and Ordering

This Xilinx® LogiCORE™ IP module is provided under the terms of the Xilinx Core License Agreement. The module is shipped as part of the Vivado® Design Suite. For full access to all core functionalities in simulation and in hardware, you must purchase a license for the core. To generate a full license, visit the product licensing web page. Evaluation licenses and hardware timeout licenses might be available for this core. Contact your local Xilinx sales representative for information about pricing and availability.

For more information about this core, visit the Embedded RDMA Enabled NIC product lounge (registration required).

License Checkers

If the IP requires a license key, the key must be verified. The Vivado® design tools have several license checkpoints for gating licensed IP through the flow. If the license check succeeds, the IP can continue generation. Otherwise, generation halts with an error. License checkpoints are enforced by the following tools:

- Vivado Synthesis
- Vivado Implementation
- write_bitstream (Tcl command)

***Note*:** IP license level is ignored at checkpoints. The test confirms a valid license exists. It does not check IP license level.

Documentation Navigator and Design Hubs

Xilinx® Documentation Navigator (DocNav) provides access to Xilinx documents, videos, and support resources, which you can filter and search to find information. To open DocNav:

- From the Vivado® IDE, select **Help** → **Documentation and Tutorials**.

- At the Linux command prompt, enter `docnav`.

Xilinx Design Hubs provide links to documentation organized by design tasks and other topics, which you can use to learn key concepts and address frequently asked questions. To access the Design Hubs:

- In DocNav, click the **Design Hubs View** tab.
- On the Xilinx website, see the Design Hubs page.

***Note*:** For more information on DocNav, see the Documentation Navigator page on the Xilinx website.

Revision History

The following table shows the revision history for this document.

Section	Revision Summary
02/06/2019 Version 1.0	
Initial release	N/A

Please Read: Important Legal Notices

AUTOMOTIVE APPLICATIONS DISCLAIMER

AUTOMOTIVE PRODUCTS (IDENTIFIED AS "XA" IN THE PART NUMBER) ARE NOT WARRANTED FOR USE IN THE DEPLOYMENT OF AIRBAGS OR FOR USE IN APPLICATIONS THAT AFFECT CONTROL OF A VEHICLE ("SAFETY APPLICATION") UNLESS THERE IS A SAFETY CONCEPT OR REDUNDANCY FEATURE CONSISTENT WITH THE ISO 26262 AUTOMOTIVE SAFETY STANDARD ("SAFETY DESIGN"). CUSTOMER SHALL, PRIOR TO USING OR DISTRIBUTING ANY SYSTEMS THAT INCORPORATE PRODUCTS, THOROUGHLY TEST SUCH SYSTEMS FOR SAFETY PURPOSES. USE OF PRODUCTS IN A SAFETY APPLICATION WITHOUT A SAFETY DESIGN IS FULLY AT THE RISK OF CUSTOMER, SUBJECT ONLY TO APPLICABLE LAWS AND REGULATIONS GOVERNING LIMITATIONS ON PRODUCT LIABILITY.

Copyright