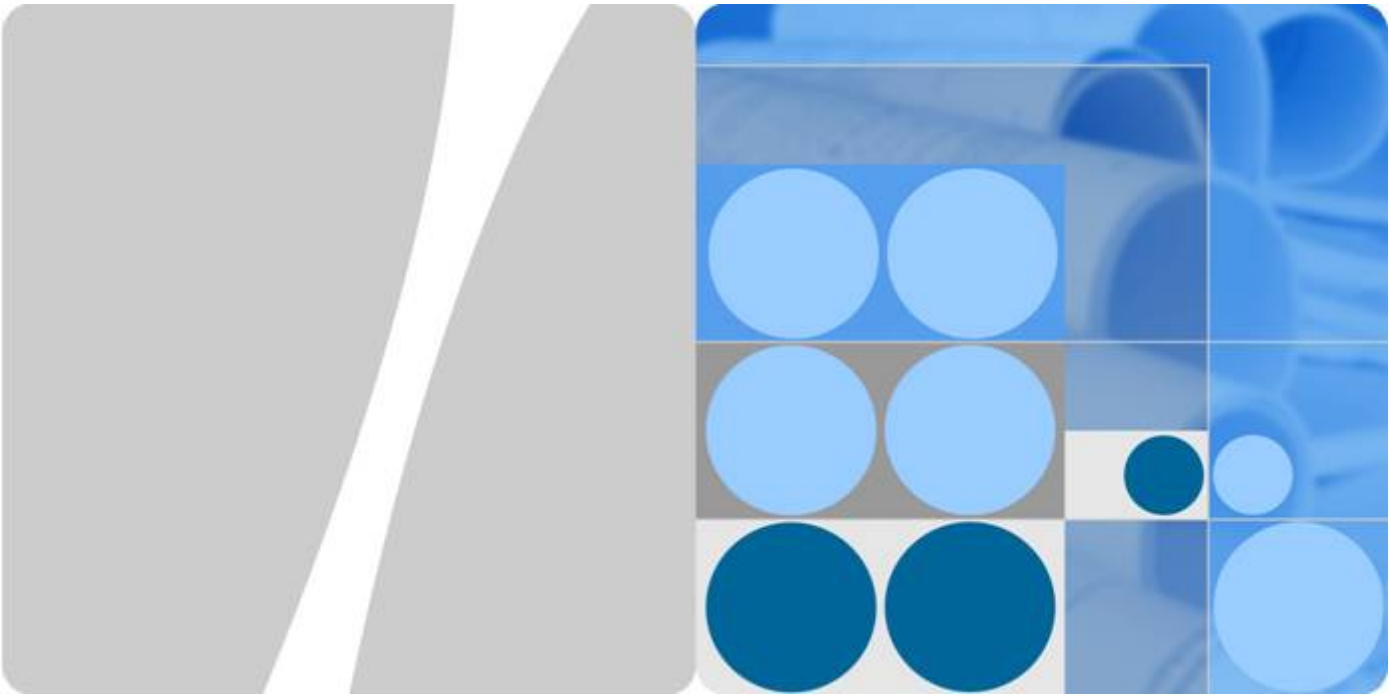


Huawei MZ710 NIC

V100R001

White Paper

Issue 08

Date 2016-11-21

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://e.huawei.com

About This Document

Purpose

This document describes the MZ710 in terms of its functions, appearance, features, applications, and technical specifications. You can obtain comprehensive information about the MZ710 by reading this document.

Intended Audience

This document is intended for:

- Huawei presales engineers
- Channel partner presales engineers
- Enterprise presales engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
 DANGER	Alerts you to a high risk hazard that could, if not avoided, result in serious injury or death.
 WARNING	Alerts you to a medium or low risk hazard that could, if not avoided, result in moderate or minor injury.
 CAUTION	Alerts you to a potentially hazardous situation that could, if not avoided, result in equipment damage, data loss, performance deterioration, or unanticipated results.
 NOTE	Provides additional information to emphasize or supplement important points in the main text.

Change History

Issue 08 (2016-11-21)

This issue is the eighth official release.

Issue 07 (2016-08-17)

This issue is the seventh official release.

Issue 06 (2016-05-15)

This issue is the sixth official release.

Type	Change Description
Modify	The document content is optimized.

Issue 05 (2016-03-31)

This issue is the fifth official release.

Issue 04 (2015-11-09)

This issue is the fourth official release.

Issue 03 (2015-05-30)

This issue is the third official release.

Issue 04 (2016-03-31)

This issue is the fourth official release.

Issue 03 (2015-05-30)

This issue is the third official release.

Issue 02 (2015-03-20)

This issue is the second official release.

Issue 01 (2014-12-02)

This issue is the first official release.

Contents

Figures

Tables

1 Overview

About This Chapter

1.1 Functions

The MZ710 is an Ethernet network interface card (NIC). It is used for E9000 compute nodes and provides 40GE ports for compute nodes to connect to switch modules in the chassis.

The MZ710 uses the Mellanox ConnectX-3 Pro (CX3 Pro) chip and supports NIC applications. The 40GE ports on the MZ710 support 40G/10G auto-negotiation, and 40GE and 10GE port applications. The MZ710 supports the Remote Direct Memory Access over Converged Ethernet (RoCE) feature to address low-latency network applications.

1.2 Appearance

The MZ710 can be installed in slot Mezzanine1 (Mezz1 for short) or Mezzanine2 (Mezz2 for short) on a half-width E9000 compute node or in slot Mezz1, Mezz2, Mezzanine3 (Mezz3 for short), or Mezzanine4 (Mezz4 for short) on a full-width E9000 compute node.

The MZ710 provides network ports for connecting to switch modules:

- When the MZ710 is installed in slot Mezz1 or Mezz3, its two 40GE ports connect to switch modules in slots 2X and 3X.
- When the MZ710 is installed in slot Mezz2 or Mezz4, its two 40GE ports connect to switch modules in slots 1E and 4E.

Figure 1-1 MZ710 appearance

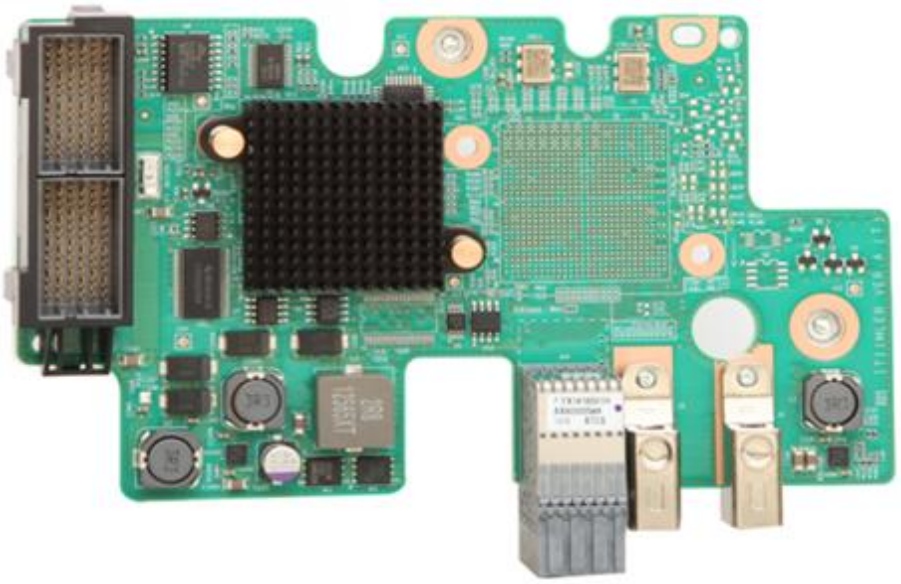


Figure 1-2 MZ710 installation positions on a half-width compute node

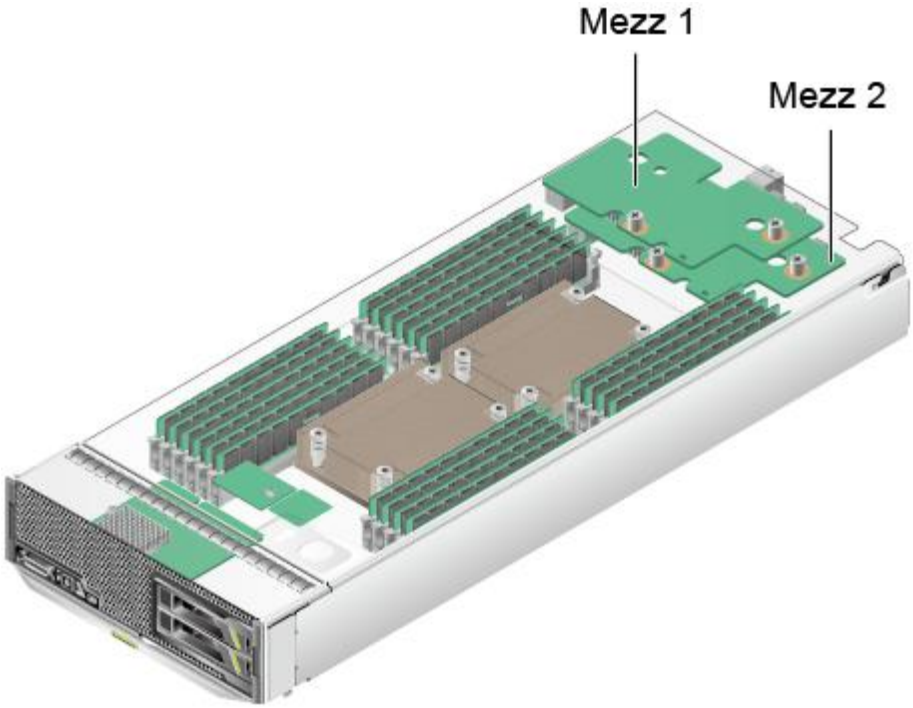
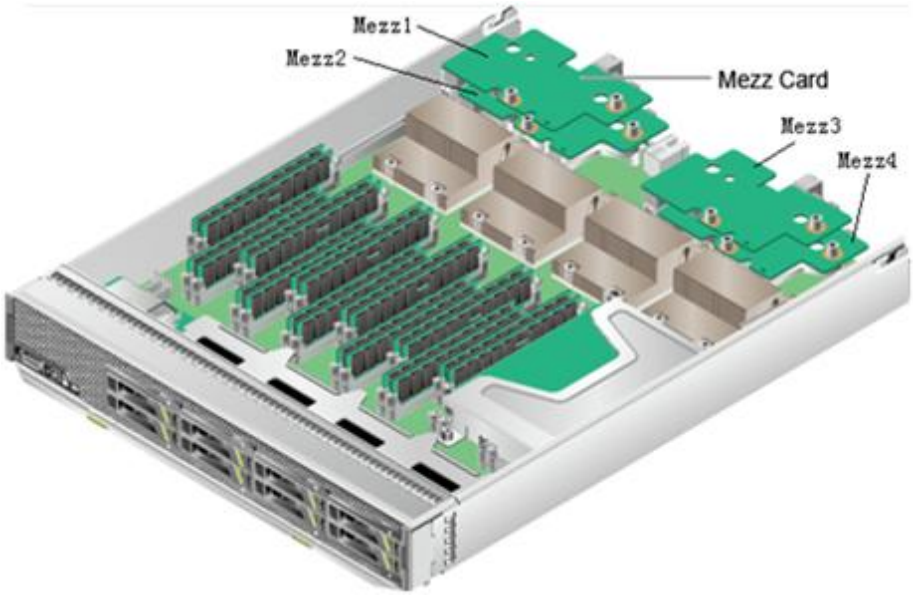


Figure 1-3 MZ710 installation positions on a full-width compute node



2 Features

About This Chapter

2.1 Feature List

The MZ710 supports the following features and performance specifications:

- RoCE and RoCE v2 (R-RoCE over UDP)
- Single Root-I/O Virtualization (SR-IOV), supporting a maximum of 126 virtual functions (VFs)
- NetQueue
- 802.1Q VLAN, supporting a maximum of 4094 VLANs
- RoCE Offload, VxLAN Offload, and NVGRE Offload
- TCP Checksum Offload, UDP Checksum Offload, Receive Side Scaling (RSS), and Large Send Offload (LSO)
- Jumbo frames of 9 KB
- Data Plane Development Kit (DPDK)

NOTE

Certain functions are related to OS and switch features. For details, contact Huawei technical support.

2.2 Feature Description

802.1Q VLAN

The MZ710 supports a maximum of 4094 VLANs. Each 40GE port supports a maximum of 4094 VLANs. The VLAN IDs are integers ranging from 1 to 4094.

The MZ710 does not tag or untag packets, but transparently transmits them. VLAN IDs are specified by the operating system (OS) on an E9000 compute node.

SR-IOV

The MZ710 supports the SR-IOV feature. The NIC supports a maximum of 126 VFs, and each physical 40GE port on the NIC supports a maximum of 63 VFs.

The VFs derived from each physical function (PF) can be assigned to virtual machines (VMs). In this way, the mapping between VFs and VMs is set up.

PXE

The 40GE ports on the MZ710 support the Preboot Execution Environment (PXE) feature. PXE is used for remote boot over an Ethernet or IP network. It enables users to remotely connect to the PXE server to load OSs.

RoCE

The MZ710 supports the RoCE feature. This feature uses the Remote Direct Memory Access (RDMA) technology to reduce the packet processing and forwarding latency of the NIC, to reduce the CPU usage, and to implement low-latency data transmission over the data center network. With the RoCE feature, the end-to-end read and write delay of the NIC can reach 4 us when the packet length is 128 bytes. The MZ710 supports 16M I/O channels (equivalent to

IB QPs), and provides priority-based scheduling and flow control to support low-latency, high-bandwidth network transmission. The MZ710 supports OpenFabrics Enterprise Distribution for Linux (OFED for Linux) and Mellanox OFED for Windows (WinOF).

VxLAN and NVGRE

The MZ710 supports the VxLAN and NVGRE network virtualization overlay technologies.

VxLAN is put forwarded by VMware, and NVGRE is put forwarded by Microsoft. VxLAN and NVGRE encapsulate packets on the layer-2 data center network into layer-3 network packets for transmission over the layer-3 network. In this way, the layer-3 network provides transmission tunnels and routes for the isolated layer-2 data center network. VxLAN and NVGRE support deployment of the data center network across the IP network, and support virtual machine (VM) deployment and migration across the IP network for virtualization applications.

VxLAN and NVGRE add special packet headers (identifiers) to layer-2 Ethernet packets to identify packet tenants (hosts). The MZ710 supports hardware-based offload of VxLAN and NVGRE packet headers to reduce CPU usage and packet processing latency.

2.3 Standards Compliance

Table 2-1 lists the standards and protocols that the MZ710 complies with.

Table 2-1 Standards compliance

Standard	Protocol
IEEE 802.3x	Flow Control and Back Pressure
IEEE 802.3ba	40GBASE-KR4
IEEE 802.3ap	10GBASE-KR
IEEE 802.1Q	VLAN Tagging
IEEE 802.1ab	Station and Media Access Control Connectivity Discovery, also known as Link Layer Discovery Protocol (LLDP)
IEEE 802.3ad	Link Aggregation Control Protocol (LACP)

3 Applications

About This Chapter

3.1 Compatible Compute Nodes

The MZ710 can be installed in slot Mezz1 or Mezz2 on a half-width compute node or in slot Mezz1, Mezz2, Mezz3, or Mezz4 on a full-width compute node. **Table 3-1** lists the compute nodes that support the MZ710 and its installation positions on them.

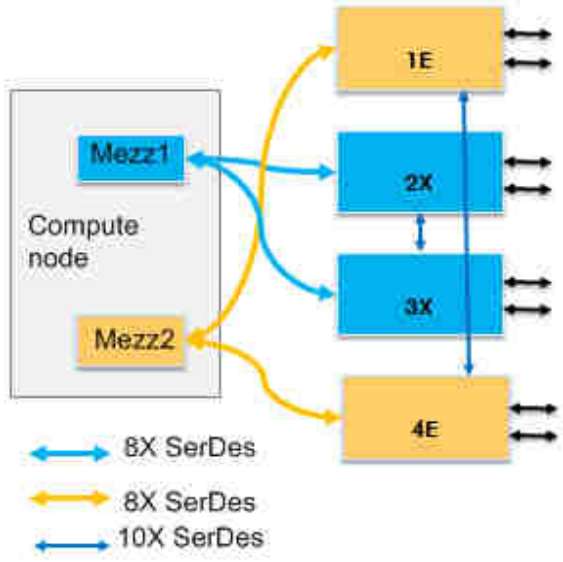
Table 3-1 Compute nodes that support the MZ710

Compute Node	Number of Mezz Module Slots	MZ710 Installation Position
CH121	2	Mezz1 and Mezz2
CH121 V3	2	Mezz1 and Mezz2
CH220	1	Mezz1
CH220 V3	2	Mezz2 and Mezz3
CH221	1	Mezz1
CH222	2	Mezz1 and Mezz2
CH222 V3	2	Mezz1 and Mezz2
CH226 V3	2	Mezz1 and Mezz2
CH240	2	Mezz1 and Mezz2
CH242 V3	4	Mezz1, Mezz2, Mezz3, and Mezz4
CH242 V3 DDR4	4	Mezz1, Mezz2, Mezz3, and Mezz4

3.2 Connected I/O Modules

MZ710s can connect to I/O modules (switch modules or interface boards). **Figure 3-1** shows the connections between the MZ710s on a half-width compute node and the ports on I/O modules.

Figure 3-1 Connections between the MZ710s on a half-width compute node and the ports on I/O modules



There are two or four groups of Serializer/Deserializer (SerDes, known as high-speed interconnect line) between each compute node and I/O module slots.

- Mezz1: 8X SerDes for connecting to I/O module slots 2X and 3X
- Mezz2: 8X SerDes for connecting to I/O module slots 1E and 4E
- Mezz3 (available only on a full-width compute node): 8X SerDes for connecting to I/O module slots 2X and 3X
- Mezz4 (available only on a full-width compute node): 8X SerDes for connecting to I/O module slots 1E and 4E

NOTE

The MZ710 provides two ports, and only 4X of each 8X SerDes is used.

Table 3-2 describes the I/O modules to which the MZ710 can connect.

Table 3-2 I/O modules to which the MZ710 can connect

I/O Module	I/O Module Slot	MZ710 (Mezz1)	MZ710 (Mezz2)	Typical Configuration	Remarks
CX710	2X/3X	√	X	Yes	-
	1E/4E	X	√	Yes	-
CX712	2X/3X	√	X	Yes	-
	1E/4E	X	√	Yes	-

3.3 MZ710 Networking

The MZ710 can connect to I/O modules (switch modules or interface boards) to provide Ethernet services.

The MZ710 can work with the CX710 switch module to provide 80 Gbit/s bandwidth with its two network ports, and connect to the Internet through 40GE ports on the CX710. See **Figure 3-2**.

Figure 3-2 Connection between the MZ710 and the CX710

NOTE

The CX3 Pro chip system on the MZ710 uses a PCIe 3.0 x8 port with a theoretical bandwidth of 64 Gbit/s. The two 40GE ports on the MZ710 can actually provide a total bandwidth of 56 Gbit/s instead of 80 Gbit/s because of protocol overhead and efficiency loss.

3.4 Supported OSs

Table 3-3 lists the OSs supported by the MZ710.

Table 3-3 OSs supported by the MZ710

OS	Version	Remarks
Windows	Windows Server 2008 Enterprise SP2	-
	Windows Server 2008 R2 Enterprise Edition x64	-
	Windows Server 2012 Enterprise Edition x64	-
Red Hat	Red Hat Enterprise Linux (RHEL) 6.1 x86	-
	RHEL 6.1 x86_64	-
	RHEL 6.3 x86	-
	RHEL 6.3 x86_64	-
SUSE	SUSE Linux Enterprise Server (SLES) 11 SP1 x86	-

OS	Version	Remarks
	SLES 11 SP1 x86_64	-
	SLES 11 SP2 x86	-
	SLES 11 SP2 x86_64	-
VMware	VMware ESXi 5.0	-
	VMware ESXi 5.1	-
	VMware ESXi 5.5	-
Oracle Linux	Oracle Linux 6.2 x86	-
	Oracle Linux 6.2 x86_64	-
	Oracle Linux 6.3 x86	-
	Oracle Linux 6.3 x86_64	-

For the latest versions of OSs, see the ***Huawei Server Compatibility Checker***.

4 Technical Specifications

Table 4-1 lists the technical specifications for the MZ710.

Table 4-1 Technical specifications

Item	Specifications
Dimensions (length x width)	148 mm x 85 mm (5.83 in. x 3.35 in.)
Power supply	12 V DC
Net weight	0.15 kg (0.33 lb)
Maximum power consumption	8 W
Temperature	Operating temperature: 5°C to 40°C (41°F to 104°F) (ASHRAE Class A3 compliant))
	Storage temperature: –40°C to +65°C (–40°F to +149°F)
Temperature change rate	15°C/h (27°F/h)
Humidity	Operating humidity: 5% RH to 85% RH (non-condensing)
	Storage humidity: 5% RH to 95% RH (non-condensing)
Altitude	● 40°C (104°F) at 1800 m (5905.44 ft) ● 30°C (86°F) at 3000 m (9842.40 ft) When the MZ710 is used at an altitude between 1800 m and 3000 m, the highest operating temperature decreases by 1°C (1.8°F) as the altitude increases by 120 m (393.70 ft).
PCIe port bandwidth	64 Gbit/s (PCIe 3.0 x8)
Port rate	10.3125 Gbit/s (10GE port) or 41.25 Gbit/s (40GE port)
Number of ports	2

Item	Specifications
Port type	Ethernet
Chip model/ manufacturer	ConnectX-3 Pro (CX3 Pro)/Mellanox

A Acronyms and Abbreviations

D	
DPDK	Data Plane Development Kit
L	
LACP	Link Aggregation Control Protocol
LLDP	Link Layer Discovery Protocol
LSO	Large Segmentation Offload
N	
NIC	network interface card
NVGRE	Network Virtualization using Generic Routing Encapsulation
O	
OFED	OpenFabrics Enterprise Distribution
OS	operating system
P	
PCIe	Peripheral Component Interconnect Express
PF	physical function
PG	priority group
PXE	Preboot Execution Environment
Q	
QP	queue pair

R	
RDMA	Remote Direct Memory Access
RoCE	RDMA over Converged Ethernet
RSS	Receive Side Scaling
S	
SR-IOV	Single Root I/O Virtualization
T	
TC	traffic class
TCP	Transmission Control Protocol
U	
UDP	User Datagram Protocol
V	
VF	virtual function
VLAN	virtual local access network
VM	virtual machine
VxLAN	Virtual Extensible LAN