An Oracle White Paper

May 2012

Oracle Big Data Appliance: Datacenter Network Integration

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Table of Contents

Executive Summary 1

Oracle Big Data Appliance in the Datacenter Network 1

Oracle Big Data Appliance System I/O 2

- Big Data Appliance Data Center Service Network Connectivity 3
- Network Services in the Oracle Big Data Appliance 3
- Ethernet Gateway for the Oracle Big Data Appliance 3
- Default Ethernet Configuration 5

Integrating the Oracle Big Data Appliance with the Datacenter LAN 6

- Datacenters with Existing 10 Gb L2 Infrastructure 6
- Datacenters without an Existing 10 Gb L2 Infrastructure or with Unique Connectivity Requirements 7
- Connecting Multi-Cabinet Deployments with InfiniBand 9

LAN Connectivity and Network Isolation 10

- Server Connectivity 10

Conclusion 12

Executive Summary

The Oracle Big Data Appliance (BDA) utilizes InfiniBand as a converged I/O network fabric to provide all I/O services. Applications residing within the Big Data Appliance system can access all network services provided within the datacenter network through InfiniBand. The Oracle Big Data Appliance system can be easily integrated into an existing datacenter's network infrastructure, if the physical connectivity and network services provided within the network infrastructure are understood during the deployment planning phase.

This white paper outlines the physical connectivity solutions supported by the Oracle Big Data Appliance and provides an overview of network service delivery in the Oracle Big Data Appliance.

Oracle Big Data Appliance in the Datacenter Network

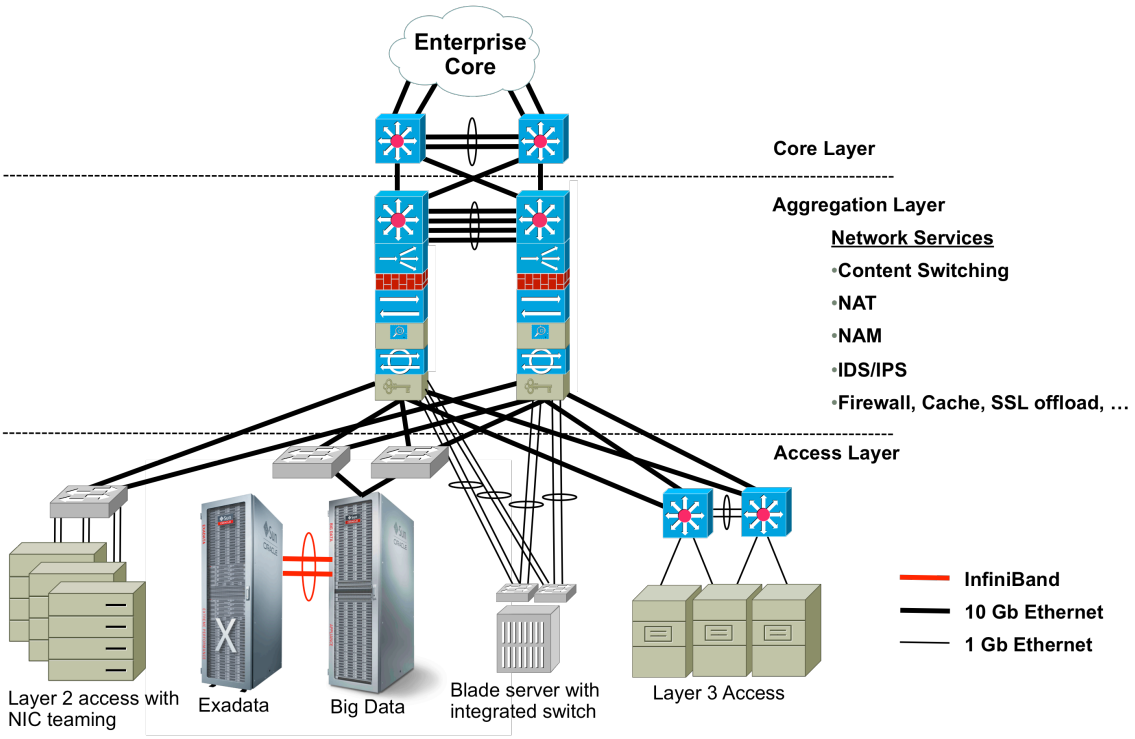

Figure 1: The Datacenter Network

From the viewpoint of the datacenter LAN, BDA appears as of a collection of 10 Gb Ethernet attached servers. Connecting BDA into the LAN requires that the system be connected to a layer-2 switch. A BDA system may connect to additional BDA, Exalogic, Exadata, SPARC

SuperCluster, Exalytics, or ZFS Storage Appliance systems via InfiniBand, but the primary interface to the system for access to the data center's service network is through the 10 Gb Ethernet interface.

Within a hierarchical network model, the Oracle Big Data Appliance resides within the access layer. BDA connects into the network through layer-2 switches and "higher level" network services such as general firewall capabilities, content switching, etc., are provided through the aggregation (or distribution) layer.

Oracle Big Data Appliance System I/O

The Oracle Big Data Appliance utilizes QDR InfiniBand as a foundation for BDA's I/O subsystem. All BDA nodes within a BDA system are redundantly connected to the Big Data Appliance's InfiniBand fabric, which functions as a highly available, high-performance system backplane for the Oracle Big Data Appliance. The InfiniBand fabric is constructed of **two** InfiniBand Gateways, which act both as InfiniBand "leaf" switches and InfiniBand to 10 Gb Ethernet Gateways, as well as **one** 36-port "spine" switch, which enables the expansion of the Big Data Appliance system. This expansion can be achieved by attaching up to eight Big Data Appliance, Exalogic, Exadata, SPARC SuperCluster, or ZFS Storage Appliance systems to the fabric without adding additional InfiniBand switches. Solutions can expand beyond eight cabinets with additional Oracle InfiniBand switches.

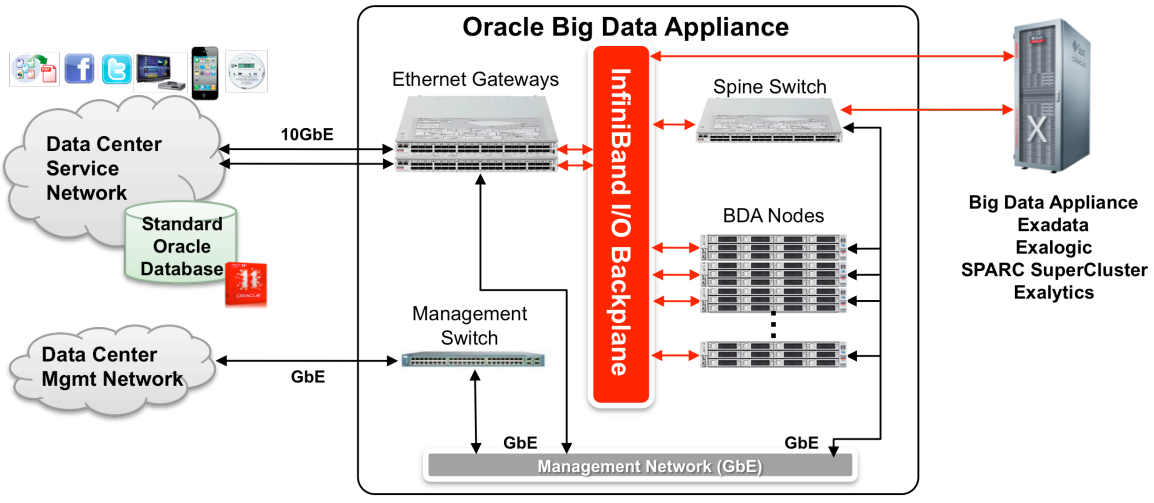

Figure 2: Big Data Appliance Internal Network Connectivity

Big Data Appliance Data Center Service Network Connectivity

The Oracle Big Data Appliance utilizes InfiniBand as a converged network. As such, Ethernet connections are carried over BDA's InfiniBand fabric. BDA server nodes are not individually connected to the data center's 10 Gb Ethernet infrastructure, but rather share up to sixteen 10 GbE connections provided through the Ethernet ports on the InfiniBand Gateways.

Network Services in the Oracle Big Data Appliance

All network services, which are provided in an enterprise data center, can be made available to applications running within the Oracle Big Data Appliance system. For source-destination connections that lie within BDA, the Big Data Appliance I/O infrastructure provides hardware-based layer-2 services for InfiniBand as well as software-based services for Ethernet and IP. Table 1 outlines the network services provided through BDA's I/O subsystem.

OSI Layer		Services
2	InfiniBand	- Switching/forwarding of packets destined for local subnet* - Forwarding of packets to router function (IB ROUTING NOT CURRENTLY SUPPORTED) - IB multi-cast within an IB partition
	Ethernet	- "Software-based" switching/forwarding of frames destined for IB connected hosts in local subnet* (EoIB) - Ethernet routing (inter-VLAN) needs to be provided via external network
3	IP	- "Software-based" switching/forwarding of packets destined for IB connected hosts in local subnet* (IPoIB) - IP multi-cast for IB connected hosts in local subnet (IPoIB) - Any other layer 3 services need to be provided via external network
4 - 7		- Other services need to be provided via external network

Table 1 – Network services provided through the I/O subsystem within the Oracle Big Data Appliance

Ethernet Gateway for the Oracle Big Data Appliance

The Gateway function in the Oracle Big Data Appliance is provided by the *Sun Network QDR InfiniBand Gateway Switch*. The *Sun Network QDR InfiniBand Gateway Switch* enables hosts on the InfiniBand fabric to share one or more 10 GbE links to an external LAN and allows the hosts sharing a 10 GbE link to communicate between each other as well as with nodes on the external LAN as if they all had a private NIC connected to the external LAN. The Gateway is implemented as a shared NIC Model and is presented as a multi-address endpoint on both IB and Ethernet fabrics. Hosts on the IB fabric communicate with nodes on the Ethernet LAN via the Gateway, but with other hosts on the IB fabric directly. However, logically the hosts sharing an external Ethernet port are all part of the same L2 Ethernet subnet seen from the external LAN.

Through the InfiniBand Gateways, each OS instance residing within BDA can be provisioned with virtualized Ethernet NICs. An InfiniBand Gateway provides a total of eight physical ports of 10 Gb Ethernet. Each physical port supports up to 1,024 virtual NICs, each with it's own MAC address and (optional) VLAN ID.

The *Sun Network QDR InfiniBand Gateway Switch* utilizes QSFP connectors to provide both Ethernet and InfiniBand connectivity. Two QSFP ports on the Gateway (the two upper right-most ports, as shown in Figure 3) are for Ethernet connectivity. The QSFP connector aggregates

four 10 Gb Ethernet connections to a single connector. The Gateway's Ethernet ports support 10G Base-SR multi-mode fiber. Within the Oracle Big Data Appliance system, the 10G Base-SR QSFP transceivers are pre-installed in the Gateway Ethernet ports. Passive fiber cables are available in two formats:

- "Splitter" cables which provide an MTP/MPO termination on one end of the cable and four separate male LC-terminated "pig tails" on the other end of the cable. The splitter cable enables connectivity to four 10 Gb Ethernet switch ports through 10G Base-SR transceivers.
- "Straight" cables which provides MTP/MPO termination on both ends of the cable. The straight cable enables connectivity to a single port on either a 10 Gb Ethernet switch that supports QSFP connectors (4x10 GbE) or to a patch panel supporting MTP-MPO.

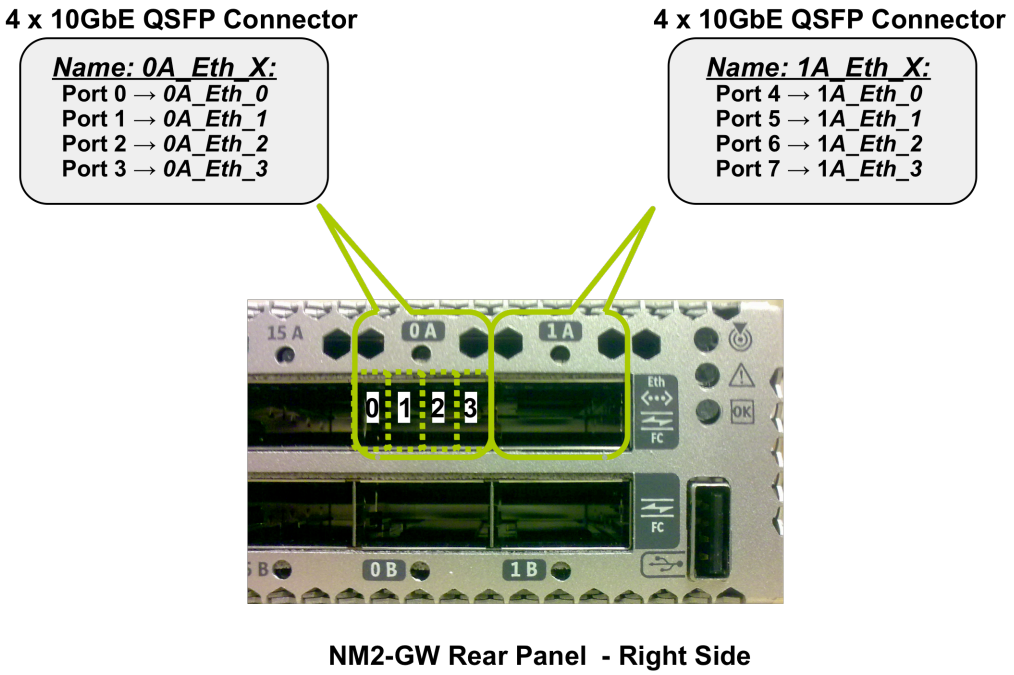

Figure 3 - Locating the Ethernet ports on the InfiniBand Gateway

BDA System Ethernet Provisioning

The Oracle Big Data Appliance is capable of providing up to sixteen 10 Gb Ethernet links to the data center LAN. Up to eight 10 GbE ports are provided by each InfiniBand Gateway, yielding a total available Ethernet bandwidth of up to 160 Gbps through both of the Gateways.

	Minimum Required	Maximum Available
BDA System 10 GbE Connections	2	16
Ethernet Bandwidth (Half Duplex)		
Total BDA System (Gbps)	20	160
Per BDA Node (Gbps)	1.1	8.9

Table 2 - BDA Ethernet Provisioning

Table 2 shows the available and minimum required 10 Gb Ethernet connectivity for an Oracle Big Data Appliance system. The Ethernet bandwidth required by a BDA system deployment is dependent upon the anticipated peak network load for the specific deployment. The peak network load can be estimated from the desired system "ingestion" rate used to size the BDA deployment.

Minimum Required Ethernet Provisioning

At a minimum, two 10 Gb Ethernet links **must** be connected to the BDA system. One 10 GbE connection **must** be made to each of the InfiniBand Gateways to provide HA network connectivity.

A functional client network **must** be established prior to software installation to allow proper configuration of the software. **The Hadoop software installation process requires that at least one 10 Gb Ethernet connection be made to each InfiniBand Gateway and configured prior to initiating the installation process.** If no 10 GbE connections are available, the software installation process will not proceed.

Default Ethernet Configuration

The I/O infrastructure within the Oracle Big Data Appliance is extremely flexible and can support provisioning BDA nodes with multiple virtualized NICs (vNICs), which can be attached to separate LANs or VLANs. The default configuration of Ethernet connections for BDA nodes is as follows:

- Two virtualized NICs (vNICs) are created per BDA node (eth8 and eth9).
- Each of the two vNICs is associated with a physical 10 Gb Ethernet port on one of the two InfiniBand Gateways installed within the BDA system. The vNICs are attached to different Gateways to provide highly available Ethernet connectivity.
- The vNIC associations for the 18 BDA nodes are distributed among the "active" physical 10 Gb Ethernet ports on the InfiniBand Gateways in round-robin fashion. For example, if each gateway switch has the following 6 active 10GigE ports: 0A-ETH-1, 0A-ETH-2, 0A-ETH-3, 0A-ETH-4, 1A-ETH-1 and 1A-ETH-2 then server 1 will be mapped to 0A-ETH-1, server 2 to 0A-ETH-2, …, server 7 back to 0A-ETH-1 and so on.

- The two vNICs are bonded as an active-passive bond (bondeth0). The "active" and "passive" ports of the bonds from each BDA node are distributed between the two InfiniBand Gateways such that the network load is balanced between the two Gateways.
- The vNICs are not configured to utilize VLAN tagging.
- This default configuration can be changed to fit the specific network configuration needs of the BDA system deployment. To change the configuration, the vNICs should be deleted and re-created with the desired properties.

Integrating the Oracle Big Data Appliance with the Datacenter LAN

As previously stated, The Oracle Big Data Appliance presents Ethernet NICs to the data center LAN through the Ethernet ports on BDA's InfiniBand Gateways. The Oracle Big Data Appliance contains two InfiniBand Gateways and thus can provide up to sixteen 10 GbE connections to the data center L2 infrastructure.

Datacenters with Existing 10 Gb L2 Infrastructure

In data centers with an existing 10 Gb Ethernet infrastructure, Big Data Appliance can be connected directly into the L2 infrastructure by connecting "splitter" cables to the InfiniBand Gateways and attaching 10G Base SR transceivers on the cable's four LC male terminated "pig tails". SFP+ SR transceivers are commonly supported by most switch vendors. It is recommended that the switch vendor's transceiver be utilized on the LC male terminated "pig tails" to ensure transceiver-switch interoperability.

In order to connect directly to an existing 10 Gb L2 infrastructure, the following must be supported:

- 10G Base-SR transceivers for connection to the L2 switch.
- Transceivers must support connection to LC male terminated optical cables
- Multi-mode fiber cable plant

Figure 4 identifies the components required to connect an Oracle Big Data Appliance into an existing 10 Gb L2 infrastructure. Table 3 identifies the component counts (along with associated part numbers) required to connect BDA to an existing 10 Gb L2 infrastructure.

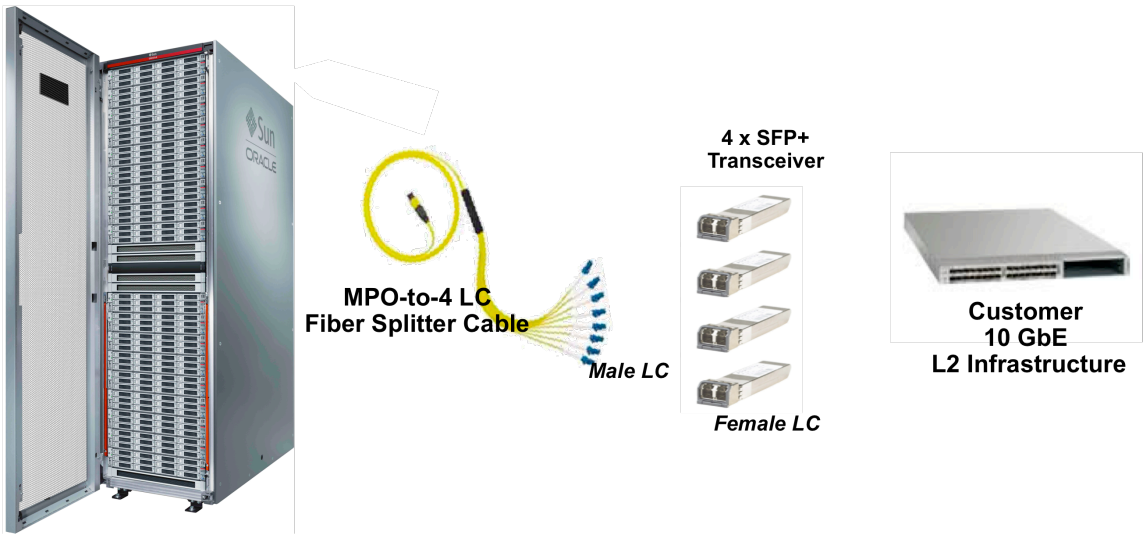

Figure 4 – Components required to connect the Oracle Big Data Appliance into an Existing 10 Gb L2 infrastructure

Quantity	Part Number	Description
4	X2127A-10m X2127A-20m X2127A-50m	MTO-to-4xLC "Splitter" Cable, 10 meters MTO-to-4xLC "Splitter" Cable, 20 meters MTO-to-4xLC "Splitter" Cable, 50 meters
16	Customer Supplied	SFP+ Transceivers Compatible with Customer's Ethernet Switch

Table 3 - Components required to connect the Oracle Big Data Appliance to an existing 10 Gb L2 infrastructure

Datacenters without an Existing 10 Gb L2 Infrastructure or with Unique Connectivity Requirements

Data centers with a L2 infrastructure that does not meet the requirements outlined above should interface to a 10 Gb L2 switch, which does meet those requirements, such as the **Sun Network 10 GbE Switch 72p**. Additionally, many third party 10 Gb Ethernet switches can provide this capability. A pair (for high availability) of **Sun Network 10 GbE Switch 72p's** can interface up to eight Oracle Big Data Appliances to a mix of data center network speeds and media types including:

- 1 Gb Ethernet
  - Copper - 1000 Base-T
  - Fiber
    - Single Mode – 1000 Base-Lx
    - Multi Mode – 1000 Base-Sx
- 10 Gb Ethernet
  - Copper – SFP+ Direct Attach (aka 10GSFP+Cu or TwinAx)
  - Fiber
    - Single Mode – 10G Base-LR

Figure 5 depicts the components required to connect an Oracle Big Data Appliance system into a data center infrastructure. Table 4 identifies the component counts (along with associated part numbers) required to connect a BDA to the data center LAN.

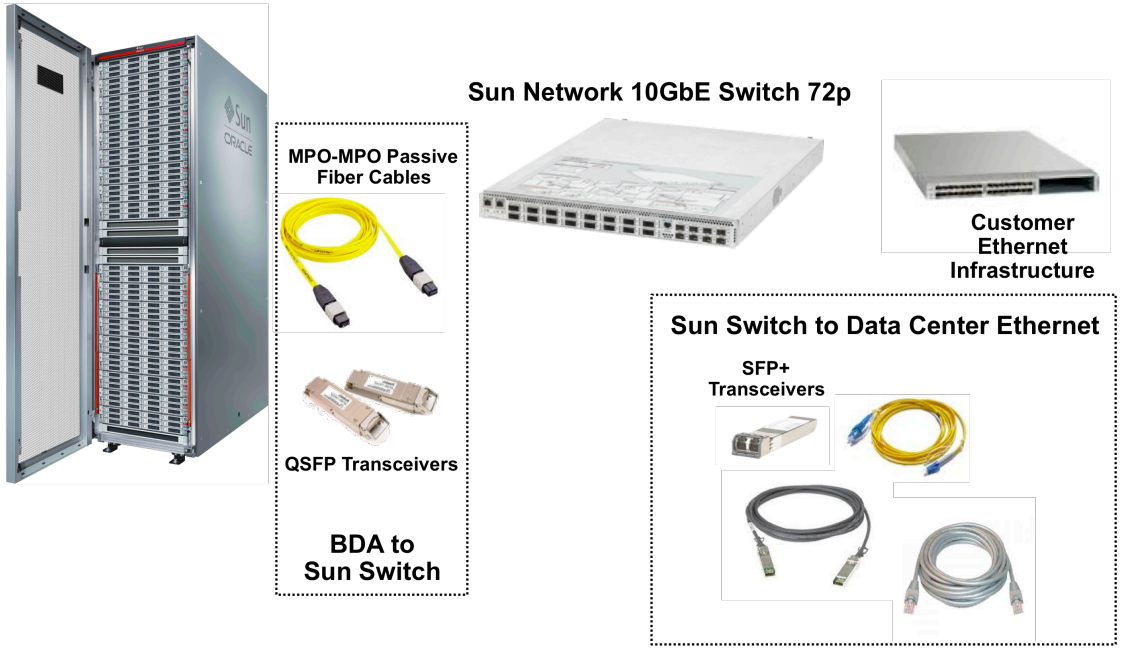

Figure 5 - Components required to connect the Oracle Big Data Appliance into a data center LAN

Quantity	Part Number	Description
Oracle Big Data Appliance to Oracle Switch		
4	7102869 7102870 7102871	10m MPO-MPO High Bandwidth optical cable 20m MPO-MPO High Bandwidth optical cable 50m MPO-MPO High Bandwidth optical cable
4	X2124A	QSFP Transceiver
2	X2074A-R X2074A-F	10GE Switch 72p (rear ventilation) 10GE Switch 72p (front ventilation)
Oracle Switch to Data Center Network		
10 GbE - Multi Mode Fiber (10G Base-SR)		
16	X2129A-N	SFP+ SR Transceiver
16	Customer Supplied	LC-LC Cables, MMF
16	Customer Supplied	Vendor Switch Compatible Transceiver
10 GbE - Single Mode Fiber (10G Base-LR)		
16	X5562A-Z	SFP+ LR Transceiver
16	Customer Supplied	LC-LC Cables, SMF
16	Customer Supplied	Vendor Switch Compatible Transceiver
10 GbE - TwinAx Copper (10GSFP+Cu)		
16	X2130A-1M-N X2130A-3M-N X2130A-5M-N	SFP+ Copper 1 Meter SFP+ Copper 3 Meter SFP+ Copper 5 Meter
1 GbE - Multi Mode Fiber (1000Base-Sx)		
16	X2129A-N	SFP+ SR Transceiver
16	Customer Supplied	LC-LC Cables, MMF
16	Customer Supplied	Vendor Switch Compatible Transceiver
1 GbE - Copper (1000Base-T)		
16	X2123A	SFP+ Transceiver Copper RJ45
16	Customer Supplied	Cat 5 / Cat 6 Cables

Table 4 - Components required to connect the Oracle Big Data Appliance to the data center LAN

Connecting Multi-Cabinet Deployments with InfiniBand

Oracle Big Data Appliance deployments can be scaled by interconnecting multiple BDA systems. Furthermore, the Oracle Big Data Appliance can be connected to other Oracle engineered systems such as Exadata using InfiniBand to interconnect the systems. Deployments of up to three adjacent cabinets can be interconnected without any additional hardware (spares kits contain sufficient cables). Deployments of up to eight cabinets can be constructed using InfiniBand optical cables. No additional switching components are required to interconnect up to eight cabinets.

Table 5 identifies the components required for a single InfiniBand optical connection. Consult the *Oracle Big Data Appliance Machine Owner's Guide* to identify the number of cables required to interconnect systems.

Quantity	Part Number	Description
For Each Optical Connection		
1	7102869 7102870 7102871	10m MPO-MPO High Bandwidth optical cable 20m MPO-MPO High Bandwidth optical cable 50m MPO-MPO High Bandwidth optical cable
2	X2124A	QSFP Transceiver

Table 5 - Components required for Optical InfiniBand Connections

LAN Connectivity and Network Isolation

Server Connectivity

The network interfaces on the BDA nodes utilize active-passive bonding for high availability. Each server is provisioned with a dual-ported InfiniBand HCA (Host Channel Adapter). The InfiniBand ports are active-passive bonded through the software stack and each port is connected to a separate InfiniBand Gateway with the "active" network load distributed amongst the InfiniBand Gateways in the system. (Note this distribution is statically established during system configuration.)

Ethernet traffic is carried over the InfiniBand fabric by encapsulating Ethernet frames within InfiniBand packets. The Ethernet over InfiniBand Protocol (EoIB) is a network interface implementation over InfiniBand. EoIB encapsulates Layer 2 datagrams over an InfiniBand Datagram (UD) transport service. The InfiniBand UD datagrams encapsulates the entire Ethernet L2 datagram and its payload. The EoIB Protocol also enables routing of packets from the InfiniBand fabric to a 1 or 10 Gb Ethernet subnet.

Ethernet provisioning is provided by instantiating bonded virtualized NICs (aka VNICs) within the OS instance. The active-passive bonding of the VNICs follows that of the associated InfiniBand HCA ports. Each VNIC is associated with a physical port on the InfiniBand Gateway and can be associated with a VLAN. The Oracle Big Data Appliance is configured such that each VNIC in a bonded pair is associated with physical port on a different InfiniBand Gateway for high availability.

Figure 6 depicts the server connections within the Oracle Big Data Appliance and the attachment of the system to a single LAN. Note each server has active-passive bonded InfiniBand HCA ports and VNICs.

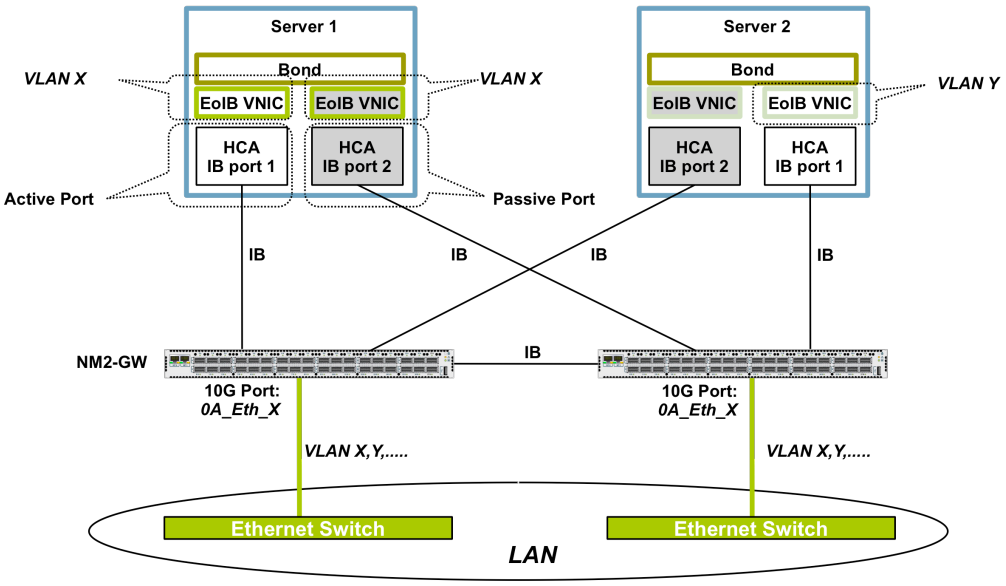

Figure 6 – Big Data Appliance Server Ethernet Connectivity, showing servers provisioned to support multiple VLANs connected to a single LAN.

Figure 7 depicts BDA nodes connected to multiple external LANs. Note that a VNIC is instantiated for each external connection.

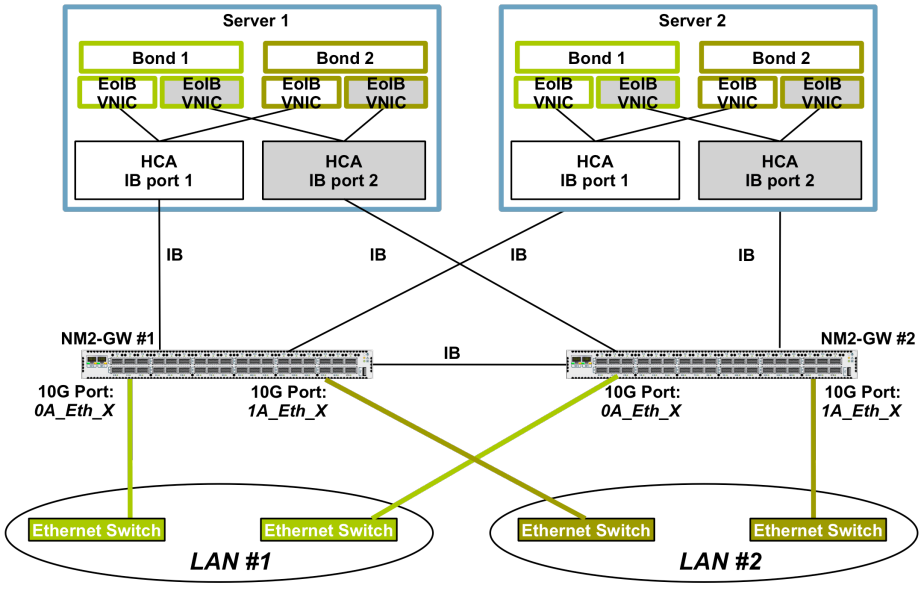

Figure 7 – Oracle Big Data Appliance server Ethernet connectivity, showing servers connected to multiple LANs

Conclusion

The I/O subsystem within the Oracle Big Data Appliance is capable of delivering a rich network environment to applications residing within BDA. That network environment can be dynamically configured to meet the demands of the application environment. The converged network infrastructure enabled by InfiniBand provides a flexible, secure, high performance solution for all application I/O.

With appropriate planning, the Oracle Big Data Appliance can be seamlessly integrated into a datacenter's network infrastructure.

Oracle Big Data Appliance:
Datacenter Network Integration
November, 2012
Authors: Steve Callahan, Ola Torudbakken

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

Oracle is committed to developing practices and products that help protect the environment

030

THIS PAGE INTENTIONALLY LEFT BLANK