A Configuration Management Model on the High-speed Networks

Jie Huang, Lin Chen

[1] School of Computer, National University of Defense Technology,
Chang Sha, China
huangjie@nudt.edu.cn

Abstract. The infiniband-based high-performance cluster system can not only provide the large bandwidth and the low latency, but also achieve CPU bypass operations during the data exchange. Being the part of the IBA (Infiniband Architecture), the management model runs through each layer, and brings some important RAS (Reliability, Availability and Scalability) mechanisms to IBA. The IBA network configuration management, which is the most important part of the IBA management model, is the basic interconnection technique of the infiniband network. In this paper, the IBA network configuration management has been discussed. We propose a network configuration management model in the infiniband network named NCM-IB. Based on that, we describe the selection mechanism of master NCM, the topology discovery mechanism, the routing algorithm and the distribution mechanism. The NCM-IB model can effectively maintain the normal operation of the infiniband network in the practical application.

Keywords: Infiniband Architecture; Network Configuration Management; Topology Discovery; Distribution Mechanism

1 Introduction

Infiniband-based high-performance cluster system can not only provide large bandwidth and low latency, but also achieve CPU bypass operations during data exchange[1,2]. It can also implement zero-copy based on infiniband, which can maximize the CPU performance and improve the overall performance of cluster systems significantly[3]. Being the part of IBA (Infiniband Architecture), management model runs through each layer, and brings some of important RAS (Reliability, Availability and Scalability) mechanisms to IBA[4,5]. The features of IBA management model, which include topology discovery, configuration, communication and fault-tolerant and so on, support the management components of the multi-suppliers and the network upgrade, which ensures both the interoperability of different version products from various suppliers and the integration with enterprise-level management tools of data center[6-8]. IBA network configuration management, which is the most important part of IBA management model, is the basic interconnection technique of infiniband.

In this paper, IBA network configuration management has been discussed. We propose the network configuration management model in infiniband network named NCM-IB. Based on that, we present the selection mechanism of the master network configuration manager (NCM), the topology discovery mechanism, and the up/down routing algorithm and the distribution mechanism. The NCM-IB model can effectively maintain the normal operation of the infiniband network in practical application.

2 The Configuration Management Model

The infiniband networks need the correct configuration before the normal communication. The NCM-IB model is built on top of user layer verbs source language and user layer MAD (Management Datagram), which transmits information and configures network through HCA (Host Channel Adaptor) hardware and the network [9-11].

By following the object-oriented framework, The NCM-IB model denotes physical component of infiniband network as the class in object-oriented program, and the relationship between classes reflects the association between physical components which include nodes, switches, ports, network configuration manager and so on. For example, each node object contains several port objects, and each port object point to another port object, which belongs to some switch object. All device objects are organized according to certain relationship in the infiniband network, which is composed of the infiniband network objects, the NCM objects, SA objects and some other auxiliary objects.

3 The Management Mechanism

The NCM-IB is built on top of user layer verbs source language and user layer MAD, which transmits information through some base hardware including HCA, switches in infiniband network and the other hosts. Basing on the NCM-IB model, we describe the selection mechanism of the master NCM, the topology discovery method, the routing algorithm and the distribution mechanism.

3.1 The selection mechanism of the master NCM

The NCM runs on the port of HCA. However, if there are more than one NCM in the network, only one of them can be the master NCM, and the rest of them will become backup NCM. Meanwhile, each port can only have one NCM. The selection of the master NCM is a part of the initialization procedure, and is the key factors of the successful initialization and the configuration in the infiniband network. The main effect of master NCM is as follows:

1) Discovery of physical topology in the current network.
2) Allocation of local identifiers (LID) for each node, switch and router.

3) To determine a reasonable path between each pair of nodes.

4) To scan the network, and discover the change of network topology, then handle node joining and leaving.

In the NCM model, there are four states including discovering, standby, master and not active. In the initialization of the infiniband network, the NCM node will turn itself into discovering state, and begin to scan the network. While discovering a prior NCM or the master NCM in the network, the current NCM node will turn itself into standby state and inquiry the master NCM continuing from time to time, and if the master NCM do not respond, which means it has expired, the current NCM node turn itself back to the discovering state. If the current NCM node hasn't found any prior or the master NCM at the end of its discovering state, it will turn itself into the master state, and begin to initialize the network. While being the standby state, a NCM isn't going to inquiry the master NCM, it can turn itself into not active state. The master NCM scans the network at the predetermined time, and if it discovers a prior NCM which is in the standby state, it will transfer its master identity to this prior NCM by sending management message, and turn itself into the standby state.

3.2 The Topology Discovery Method

After being chosen as the master, the NCM has to scan the topology of the network in order to find out the connection relations of the network, and then allocate the local identifiers (LID) which will be used in the network communications for each node. The topology scan will be executed once during the initialization of the infiniband network, and will be executed according to the predetermined time or the time-trap after that, in order to detect the change of the network topology in time. The NCM scan the network by utilizing MAD in the infiniband network. The MAD transmit message using the direct routing, which ensure that MAD can begin to work before network configuration.

In the NCM-IB model, we introduce the network scan and the trap mechanism to detect network changes, in order to get the latest network topology. We combine the advantage of both complete discovery mechanism, which discards previously configured information in network discovery process and discover the latest topology information starting from scratch, and partial discovery mechanism, which scans only the area affected by topology change rather than the whole network, in order to reduce the number of MAD and computational overhead during network topology discovery.

During completing the discovery mechanism, the NCM object discards all collected information and starts to discover the complete network topology while being aware of the change of network topology. The discovery procedure has to use direct routing, which is relatively slower than LID routing management packets. In order to discover all the active network nodes, the NCM constantly sends corresponding MAD, which will be received by the destination node of each direct route, and induces the corresponding response. For the sake of simple, the topology discovery procedure is considered as spreading in order, which means the spread of scanning MAD is in an uncertain way. The NCM sends a new scanning MAD while receiving respond from other devices of the network. The algorithm of the complete discovery mechanism is shown in Fig.1, including not only respond handling MAD

and inquiry sending the MAD in scanning procedure, but also scan handling the MAD in the discovery procedure.

When the topology change, including adding or deleting nodes or links, affects only a small part of network, we can activate the partial discovery mechanism which scans only the affected part of network. Because the topology information collected before is fully utilized in the partial discovery mechanism, the number of MAD and the computational overhead during topology discovery can be greatly reduced.

```
if AttributeID = SwitchInfo then                    {sweeping MAD }
  if SwitchInfo.PortStateChange then                {change detected}
    delete the topology database
    send a Get(NodeInfo)                            {launch the exploration}
  endif
elseif AttributeID = NodeInfo then                  {discovery MAD}
  if sender not visited then
    add this node to the topology database
    for each port in sender do                      {explore sender ports}
      send a Get(PortInfo)
    endfor
  endif
elseif AttributeID = PortInfo then                  {discovery MAD }
  if management port then
    send a Set(PortInfo)                            {send the assigned LID}
  endif
  if PortInfo.PortState <> DOWN then                {active port}
  add this port to the sender ports list
  send a Get(NodeInfo)                              {discover a new device}
  endif
Endif
```

Fig. 1. The description of the discovery algorithm.

In this paper, we implement the partial discovery mechanism with both direct routing and LID routing, in order to reduce the transmission overhead. While a topology change occurs, the NCM can detect at least one switch port which has changed its state during the periodic network scanning. If a switch port becomes active, the NCM begins to discover new devices on the other side of this port, and scan unknown area as complete discovery mechanism. Scanning MAD can make use of a section of initial LID routing to reach a node, which fully utilize the information that already exists in forwarding table. We can reach a switch that has just changed its state by initial LID routing, but the path from the changing switch port to new device must be tracked by direct routing. That is because there is no useful configuration information in the initialization of network, and scanning MAD has to use direct routing.

If a switch port stops activity in changing list of the NCM, a part of the network will be no longer reachable. We put all unreachable devices into a collection U (Unreachable), and search all ports of reachable devices that directly connect to U according to the existing network topology information. By using these reachable ports, we can scan devices in U by direct routing, and denote devices that can not be scanned successfully as M (Missing).

3.3 The Routing Algorithm

The NCM assumes that the network topology is irregular, especially in the large scale network. It is easy to design a scalable, flexible routing system on the irregular network topology, however, the irregularity of network topology also makes packets routing and the deadlock avoidance mechanism too complicated.

The Each link in the network will be denoted as either up or down direction in up/down routing algorithm. The turn of message transmission from the down direction to the up direction will be forbidden, which avoids the circular dependency between the channels, and make the algorithm deadlock-free.

The routing algorithm is described as follows:

1) Select the node which has the smallest LID as the root node, and build a BFS (Breadth-First Search) spanning tree.

2) Assign the direction for each link. Considering a link that connects two switches, one direction of the link has been denoted as the up direction, and the other direction will be denoted as the down direction. The up direction of a link is defined as the closer direction to the root node of the BFS spanning tree.

3) If the both switches the link connects located in the same level of spanning tree, we will take the side of switch that has smaller LID.

The messages from source node firstly go along with zero or more up direction links, and then go along with zero or more down direction links, and finally get to the destination node.

From the routing strategy of the routing algorithm, we can see that neither 'up direction' links nor 'down direction' links can form a routing loop. Each node in the spanning tree will eventually reach the root node by going along with the up direction links. The root node can reach any node in the spanning tree by going along with the down direction links. Therefore, for any pair of nodes in the tree, a node firstly go along with zero or more up direction links, and then go along with zero or more down direction links, and finally it will get to any other node. A path that connects any pair of nodes can be formed in this way.

In the path that messages pass, it could prevent from the deadlock happens that the turn of message transmission from the down direction to the up direction is forbidden. If the destination node is further from the root node or the LID of the destination node is larger, a node will choose the down direction links to forward its messages. If the destination node is closer from the root node or the LID of the destination node is smaller, a node will choose up direction links to forward its messages, and then its messages will go along with some down direction links to reach the destination node.

3.4 The Distribution Mechanism

Basing on the topology discovery method and the routing algorithm described above, we can construct a forwarding table for each switch in the network. In the forwarding table, the address field stands as the destination LID, and the port field stands for the port to be used. The forwarding table in the switch must include all legal LID in order to forward all packets correctly, so we have to generate one table item in the form of <LID，PORT> for any legal LID when constructing the forwarding table for each switch. Considering each legal LID in forwarding table for each switch, if this LID

stands for a switch, the forwarding port is corresponding port field value of the minimum LID in the LID matrix. If the LID stands for a HCA, we firstly find the LID of the switch that directly connects to this HCA, and the forwarding port is corresponding port field value of the minimum LID in the LID matrix.

Based on the forwarding table which has been constructed successfully, the management MAD of the NCM will configure the forwarding table to corresponding switches and activate the HCA which connects to the switch port, and finish the configuration of the infiniband network

5 Conclusion

In the paper, we propose the IBA network configuration management model named the NCM-IB model. Based on that, we describe the selection mechanism of the master NCM, the topology discovery mechanism, the routing algorithm and the distribution mechanism. The NCM-IB model can effectively maintain the normal operation of the infiniband network in practical application.

Acknowledgement. This work is supported by Program for Changjiang Scholars and Innovative Research Team in University (No.IRT1012), "Network technology" Aid program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province and Hunan Provincial Natural Science Foundation of China(11JJ7003).

References

1. International Business Machines (IBM) Corporation, “InfiniBlue Host Channel Adapter Access Application Programming Interfaces Programmer's Guide,” October 2002.
2. Mellanox Technologies Inc., “Mellanox IB-Verbs API (VAPI),” 2001.
3. Ron Brightwell and Arthur B. Maccabe, “Scalability Limitations of VIA-Based Technologies in supporting MPI,” Proceedings of Fourth MPI Developer's and User's Conference, March 2000.
4. Ron Brightwell, Arthur B. Maccabe, and Rolf Riesen, “The Portals 3.2 Message Passing Interface Revision 1.1,” Sandia National Laboratories, November 2002.
5. Vieo Inc., “Channel Abstraction Layer (CAL) API Reference Manual V2.0,” January 2002.
6. Intel Corporation, “Linux InfiniBand Project,” http://InfiniBand.sourceforge.net.
7. InfiniBand TM Architecture Specification Release 1. 2. 1 [EB/OL] . [200922218] . http :/ / www. infinibandta. org/ .
8. Bermúdez, Casado R, Quiles F J,et al. Evaluation of a Subnet Man-agement Mechanism for Infiniband Networks [C]//Proc of Int'l Conf on Parallel Processing, 2003:117-124.
9. Ranjit Noronha, Dhabaleswar. Designing High Performance DSM Systems Using InfiniBand Features.http://nowlab.cis.ohiostate.edu/publications/conf-papers/2004/noronha-cgrid04.pdf, 2004-05.
10. Jiesheng Wu, Pete Wyckoff, Dhabaleswar K.Panda.PVFS over InfiniBand: Design and Performance Evaluation. http://nowlab.cis.ohio-state.edu/publications/conf-papers/2003/wuj-icpp03.pdf, 2003-07.
11. IBTA.2000.Supplement to InfiniBand TM Architecture Specification, Volume 1. http://www.infinibandta.org/, 2004-12.