IBA 的管理架构*)

侯宗浩[1,2] 董小社[1] 郑守淇[1] 黄泳翔[1] 乔 楠[1]

(西安交通大学新型计算机研究所 西安710048)[1] (第四军医大学西京医院信息科 西安710032)[2]

Management Framework for IBA

HOU Zong-Hao[1,2] DONG Xiao-She[1] ZHENG Shou-Qi[1] HUANG Yong-Xiang[1] QIAO Nan[1]

(NeoComputing Research Institution, Xi'an JiaoTong University, Xi'an 710049)[1]

(Information office of Xijing Hospital, the Forth Military Medical University, Xi'an 710032)[2]

Abstract InfiniBand is new channel-based, switched-fabric technology that will be built into the next generation of servers or IDCs and replace today's shared-bus I/O standards, such as PCI. The InfiniBand Specification defines a management infrastructure that is the foundation for achieving multi-vendor interoperability in InfiniBand networks. This paper mainly depicts the management model and fundamental concepts of IBA, subnet management state machine, format and usage of the Management Datagram.

Keywords Infiniband architecture, Cluster management, Active network

1 引言

InfiniBand 体系结构是一个新的工业标准,它定义了一个交换式结构(fabric)来连接处理节点和 I/O 节点,以形成系统域网 SAN(system area network)。Fabric 由主机通道适配器 HCA(Host Channel Adapter)、目标通道适配器 TCA(Target Channel Adapter)和可以堆叠的交换机 Switch 组合而成。HCA 负责处理节点的通讯;TCA 负责 I/O 一侧的通信,以满足外部控制器的特定需求,外部控制器代表了处理 IO 事务的软硬件,如 SCSI 接口控制器、LAN 控制器等。这种结构保证了 I/O 不必局限在系统内部,从而有效地克服了现有共享总线(如 PCI)的缺陷。

IBA 独立于主机操作系统和处理平台。由于采用了 VIA 的通讯机制,精简了协议栈,通讯功能由专门的适配器(HCA/TCA)来实现,避免了协议处理时的上下文切换,获得了更快的数据吞吐率和更低的延迟。新的服务器/存储器的体系结构既实现了 I/O 设备的共享,也保证了更高的 RAS (reliability, availability, scalability)、更好的热插拔功能、更好的安全性和服务质量。

管理模型是 IBA 的一部分,一些重要的 RAS 机制得益于此。它贯穿于 IBA 的各个层次,功能包括物理拓扑的发现、配置、通信和容错等;支持来自多供应商的管理组件和网络的升级,确保了各供应商不同时期产品的互操作性,以及与数据中心的企业级管理管理工具的集成。

2 IBA 管理机制模型

IBA 的子网管理实体分为子网管理和通用服务管理两大类,分别由管理器、管理代理和管理接口组成。管理器负责子网的控制和检查,代理负责设置和查询各端节点的管理参数。对子网中的每一类管理器来说,Fabric 上的每一个节点都有相应的管理代理。管理器和管理代理通过管理接口进行通信,管理接口与底层的 IBA 栈进行通信。IBA 定义的管理器和管理代理有:

子网管理器 SM (Subnet Manager)。进行拓扑发现、节点初始化及配置(包括各 switch 内路由表的配置)、子网的维护。

子网管理代理 SMA(Subnet Management Agent)。负责维护与 SM 有关的节点和端口的管理参数。

通用服务管理器 GSM(General Services Manager)。完成各种与性能、通信和 I/O 设备等有关的管理功能。

通用服务代理 GSA(General Services Agent)。维护与 GSM 有关的节点和端口的参数。

在 IB 子网中,每个管理类通常只有少量的管理器,但每个节点(通道适配器、交换机和路由器)都有一个代理与之相对应。因此,每一个节点都有一个 SMA 和多个 GSA。管理器和代理之间传送的是不可靠的数据报 MAD (Management Datagram)。管理器及其管理代理在 IBA 中的分布如图1所示。

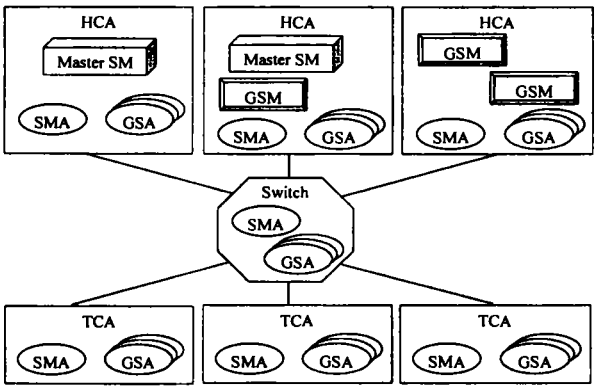

图1 管理实体的分布

IBA 中有两种 MAD:

子网管理包 SMP (Subnet Management Packet)。用于 SM 和 SMA 之间的通信,SMP 数据流在 VL15上传送。

通用服务数据包 GMP (General Management Packet)。用于 GSM 和 GSA 之间的通讯。GMP 的数据流可选择 VL0

*)项目资助:"863"重点项目,新型网络服务器系统(2001AA111120)。侯宗浩 博士生,研究方向为计算机体系结构;董小社 副教授,研究方向为网格计算。郑守淇 博士生导师,研究方向为计算机体系结构。黄泳翔、乔楠 硕士生,研究方向为并行文件系统。

一VL14进行传送。

子网管理接口 SMI(Subnet Management Interface)是 SM 与 SMA 之间的通信接口，通用服务接口 GSI(General Services Interface)是GSM 和GSA 的通信接口。一般情况下，SMI 使用 QP0(Queue Pair 0)，而 GSI 使用 QP1(Queue Pair 1)。如果 QP1出现瓶颈，GSI 也可以重定向到其它的 QP 或同一节点的其它端口，如图2所示。

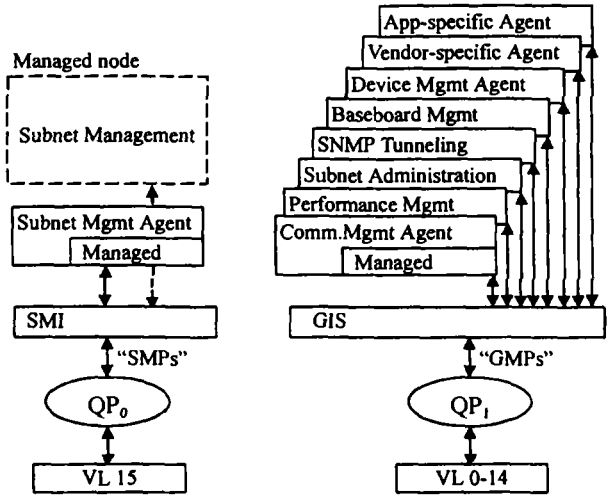

图2 MAD 接口

2.1 子网管理机制

IBA 的子网管理机制由子网管理器 SM、子网管理代理 SMA 和管理接口 SMI 组成。

子网管理器 SM 主要负责配置和管理整个 IB 网络(fabric，包括路由器、交换机和通道适配器等)。每个子网中可以包含多个 SM，但只有一个 SM 会被激活，成为主 SM(Master Subnet Manager)，其它的 SM 则处于待机状态，为主 SM 做备份。SM 可以驻留在端节点、交换机和路由器的任何一个端口，由软件或者硬件来实现。

SM 通过与位于每一个节点的 SMA 通信来发现子网的拓扑结构，为子网上的每一个节点指定子网标识符和全局标识符(LID & GID)等参数，建立端节点穿过子网的路径，通过定期的扫描(sweeping)识别拓扑结构的变化。另外，SM 也要负责维护交换机上的转发表(forwarding tables)和每个节点上的 VL 仲裁表。

SMA 存在于子网中的每一个节点，SM 使用 QP0并通过 SMI 与 SMA 进行通信。

SMI 一般采用 SMP (Subnet Management Packet)进行消息传送。SMP 分为直接路由包(direct routing packets)和 LID 路由包。直接路由包一般在端点设备和子网的初始化时使用，采用存储转发的路由方式。LID 路由包在子网初始化后使用，通过 switch 中的 LID 路由表进行路由。

2.2 通用服务管理机制

IBA 的通用服务管理机制由通用服务管理器(GSM)和通用服务代理(GSA)组成。IBA 通用服务的类型包括：

子网管理服务类 SubnAdm(Subnet Administration)。该类的管理器简称 SA(Subnet administration Agency)，在子网中充当类管理的角色，是 SM 的一部分，但它属于 GSM 类，如图3。它主要用于存储子网的管理信息，所有节点通过它来发现其它的节点和服务，从而决定与其它节点进行通信的路径和路径参数并建立通道。SA 也向其它节点提供一种注册自身的服务和了解其它已注册的服务的手段。

SA 的另外一个功能是注册 trap。当 SM 收到一个 trap，SA 将向所有注册过的节点转发这个 trap。

SA 一般与主 SM 位于同一个节点。因此，寻找 SA 的节点，只需要向主 SM 所在的节点发送 SubnAdm 类型的 MAD 即可。

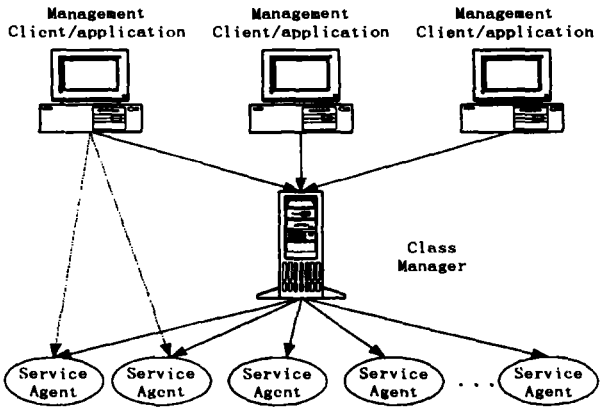

图3 类管理

其它的管理类还有：

性能管理 Perf (Performance Manager)。其功能是从性能代理中收集性能信息，完成 IB 组件中性能的统计和出错信息的检索，监视和管理性能监视器。

设备管理 DevConfMgt(Device Management)。配置和管理 I/O 控制器，提供 TCA 后面的 I/O 设备的管理。

基板管理器 BM(Baseboard Management)。监视和控制基板的属性，包括电源控制和温度传感器的报告等。

通信管理 ConMgt(Communication Management)。提供了两个端节点之间连接通道的建立、终止和迁移机制，提供了基本的服务 ID 到 QP 的解决方案。

SNMP 隧道(SNMP)。通过定义收发 SNMP 信息的方法提供 SNMP 的功能。

供应商专用管理 Vendor(Vendor Specific)。定义了某些由硬件供应商定义的服务，允许供应商远程配置和管理设备。

2.3 主机节点和 I/O 节点通信机制

I/O 单元包含一个 SMA 和多个 GSA，SMA 用于响应从 QP0上接收的 SMP，GSA(至少含 ConMgt 和 DevMgt)用于响应来自 GSI (QP1)的 GMP。

每一个 I/O 控制器由 GSA(一般为 DevMgt 类)来注册，从而可以使该 GSA 用控制器的专有信息去响应此类 GMP。

DevMgt 类 GMP 通常由处理器节点上的 I/O 资源管理器发出，以发现控制器的属性(即资源管理器配置相应 I/O 驱动程序所需的一切信息)。I/O 资源管理器安装这个驱动程序，由驱动程序创建相应的通信端口，并调用处理器节点的通信管理器来建立与 I/O 控制器的连接。

连接建立后，I/O 驱动程序则在此连接之上交换数据和控制信息，驱动程序所用端口的数目视具体情况而定。I/O 驱动程序可使用任何可用的服务类(包括可靠连接、不可靠连接、可靠数据报和不可靠数据报)，不同端口的服务类型也可以不同。

2.4 保护

IBA 采用各种密钥实现隔离和保护。密钥是一个专用的值，一般由类管理器设置和维护，并以多种方式在消息中使

用。

主子网管理器在每一个节点中放置一个不能被其它节点所读的密钥 M_Key(Management Key),接收者通过 M_Key 来确定发送者的身份,从而防止无此密钥的节点修改节点的设置信息。如有必要的话,SM 可以和对等的 SM 共享该密钥。IBA 也提供了一个租用过期机制(lease expiration),一旦该主 SM 失效,保证后继 SM 节点可以建立新的 M_Key。

IBA 中类似功能的密钥还有:基板管理密钥 B_Key (Baseboard Management Key)、分区密钥 P_Key (Partition Key)、队列密钥 Q_Key(Queue Key)、内存密钥 L_Key 和 R_Key (Memory Keys)等。

3 子网管理器的状态机模型

如图4,子网管理器有4种状态,分别是 Discovering、Standby、Master 和 Not-active。

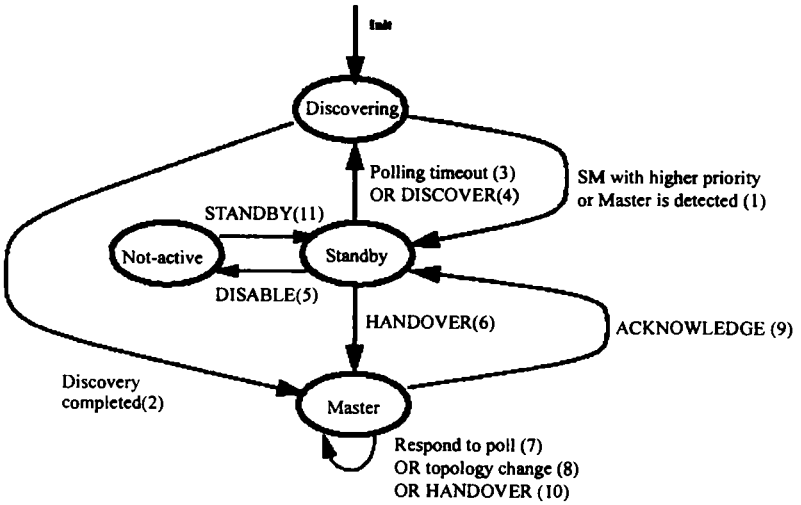

图4 子网管理器状态机模型

其过程可简单描述如下:

SM 节点在初始化时将进入 Discovering 状态,开始对子网进行扫描。当它发现子网中存在一个具有更高的优先级的子网管理器或子网中已存在主 SM 时,便进入 Standby 状态并不断地向主 SM 定时询问;如果主 SM 不响应询问,说明它已失效,该 SM 重新回到 Discovering 状态;如果在 Discovering 状态完成发现后没有找到优先级更高的节点和主 SM,则它被选为主 SM,并进入 Master 状态,开始初始化子网;在 Standby 时如不打算定时询问主 SM,则可进入 Not-active 状态;主 SM 定时扫描网络,如果发现一个处于 Standby 状态的具有更高的优先权的 SM 时,则通过一个发送消息将主 SM 身份移交过去,自身状态变为 Standby。

在发现和配置完子网拓扑结构以后,SM 将周期性地扫描(Sweeping)子网,以确定是否有变化产生,从而对拓扑结构进行升级维护。

4 管理数据报及其使用

在 IBA 中,标准的管理数据报 MAD 具有相同的包头格式,见表1,各字段属性见表2。

表1 MAD 的基本格式

Management Datagram					
0	BaseVersion	MgmtClass	ClassVersion	R	Method
4	Status		ClassSpecific		
8	TransactionID				
12					
16	AttributeID		Reserved		
20	AttributeModifier				
24	MAD Data				
…					
252					

表2 通用 MAD 的字段

字段名	长度	偏移	描述
BaseVersion	8	0	MAD 基本格式的版本..
MgmtClass	8	8	管理类型
ClassVersion	8	16	特定 MAD 类型格式的版本
R	1	24	响应位
Method	7	25	基于管理类的方法
Status	16	32	操作状态的指示
ClassSpecific	16	48	由 SM 指定和使用
TransactionID	64	64	事务 ID
AttributeID	16	128	属性 ID,定义了类的操作对象
Reserved	16	144	保留
AttributeModifier	32	160	属性修改器,提供了更进一步的属性范围
MAD Data	1856	192	数据域

一般 MAD 可分为两种类型:SMPs 管理消息和 GMPs,其中 SMPs 又可分为直接路由包和 LID 路由包。管理类详细分类见表3。

表3 管理类取值

管理类	取值	描述
Subn	0×01	LID 路由的子网管理类
Subn	0×81	直接路由的子网管理类
SubnAdm	0×03	Subnet Administration class
Perf	0×04	性能管理类
BM	0×05	基板管理类
DevMgt	0×06	设备管理类
CommMgt	0×07	通信管理类
SNMP	0×08	SNMP 隧道类
Vendor	0×09-0×0F	供应商所用类
Application	0×10-0×1F	专用应用类
0×00 0×20-0×80 0×82-0×FF		Reserved

IBA 定义了可为各管理类型使用的方法,见表4和图5~10。

表4 通用管理的方法

方法的名称	类型	取值	描述
Get()	请求	0×01	读取 CA、switch 或 router 的属性的请求(图5).
GetResp()	响应	0×81	对 Get 或 Set 请求的响应
Set()	请求	0×02	写 CA、switch 或 router 属性的请求(图6)
Send()	消息	0×03	发送数据报但不要求响应(图7)
Trap()	消息	0×05	由 CA、switch 或 router 主动发出的数据报,表明有意外事件发生(图8)
TrapRepress()	消息	0×07	通知 trap 的发送者停止发送重复的 trap. (图9)
Report()	请求	0×06	转发 event/trap/notice (图10)
ReportResp()	响应	0×86	对 Report()的响应.
		0×00, ×04, 0×08-0×0F, 0×80, 0×82-0×85, 0×87-0×8F	保留
		0×10-0×7F, 0×90-0×FF.	具体类方法. 由该类定义

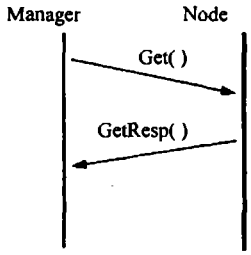

图5 Get

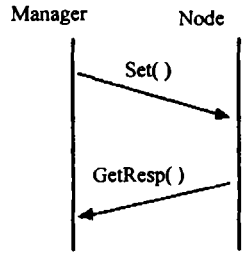

图6 Set

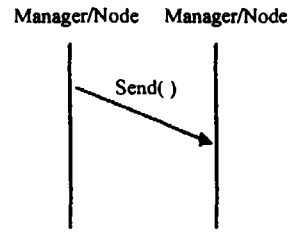

图7 Sent

5 相关工作

管理应用和管理实体之间的通信存在许多标准,如 SNMP,DMI 和 CIM 等.这些标准可以位于 IBA 管理模型之上,并通过模型中定义的服务或专用接口与 IBA 管理模型相连。换句话说,IBA 管理模型也为这类非 IBA 应用提供了一种获得子网拓扑和配置信息的手段。

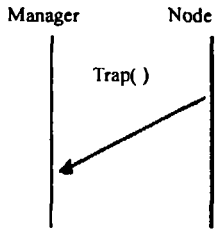

图8 Trap

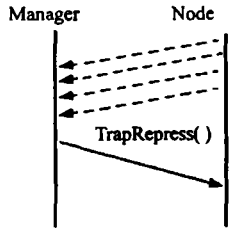

图9 TrapRepress

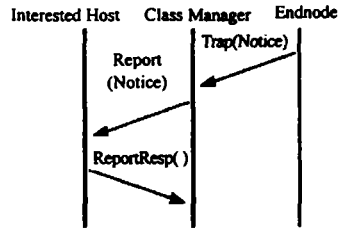

图10 Trap 的转发

作为一种新的体系结构和技术,IBA 有许多待开发的领域.开发管理组件主要面临以下挑战:

①SM 的协商和移交,需考虑到异构环境下多操作系统互操作性和多个子网的合并;

②主 SM 需要对来自于不同供应商的 SMA 统一管理;

③服务解析(即将服务名解析到提供该服务的节点和端口地址)、通信管理等需要和高级目录服务去集成,以适应各种应用和管理环境;

④多网络多管理环境数据中心集成了多种网络技术(Ethernet, Fibre Channel, Storage Area Networks),要求 IB 管理器能平滑地将它们集成在一起。

IBA 及其集群管理技术将对数据中心的构造产生深远的影响。广泛的包容性是 IBA 管理模型的主要特点,Fabric 管理软件和其它管理组件的集成、早期开发为 IBA 技术的成功提供了保证。

参 考 文 献

1 InfiniBand Architecture, Volume 1 - General Specifications, InfiniBand Trade Association, 2000

2 Baker M. University of Portsmouth, UK, Cluster Computing White Paper, Status-Final Release, Version 2.0,Dec. 2000

3 Acosta R D. Introduction to InfiniBand Management,Lane15 Software, Inc.., June 2001. http://www.infinibandta.org/data/press/mem_whitepapers/ intro_to_IB_mgmt_1101.pdf

4 Hartmann A. Using The VIEO INFINIBAND Channel Abstraction Layer (CAL), VIEO. Inc. http://www.vieo.com/infiniband/calapi.pdf

5 IBTA White Papers. http://www.infinibandta.org/newsroom/whitepapers/

6 http://www.lane15.com/

7 http://www.vieo.com

8 侯宗浩,等.IBA 体系结构的研究.计算机科学,2003,30(2)