IB QoS的带宽分配机制研究

郑 霄 陈忠平 吴志兵
(江南计算技术研究所 无锡 214083)

摘 要 InfiniBand(简称IB)网络的质量服务策略(简称IB QoS)能有效地分离和控制在IB网络上并发传输的不同应用负载的带宽,从而为基于InfiniBand网络互连的高性能计算(简称HPC)平台和数据中心平台用户提供网络服务的质量保障。但在IB QoS机制下,高优先级权值、低优先级权值和QoS因子3个关键参数究竟如何影响带宽分配,"IBA规范"并没有明确规定,目前也缺少相关的文档说明。首先在基于当前主流IB产品的实验平台上进行大量的IB QoS带宽测试实验,总结出IB QoS带宽分配机制的重要特性:当QoS因子和高优先级权值取值确定时,随着低优先级权值的增长,高优先级VL和低优先级VL的带宽比呈锯齿形增长,锯齿的偏幅逐渐减小,且周期性达到某一定值;接着根据测试数据推测出带宽分配与高优先级权值、低优先级权值和QoS因子等3个关键变量之间的量化关系式,并通过随机抽样检测的方式验证了该关系式的正确性;最后从量化关系式出发,用数学证明的方式为前面总结的带宽分配特性找到了理论依据。

关键词 InfiniBand网络,质量服务,虚通道,带宽分配,优先级权值
中图法分类号 TP393.03 **文献标识码** A

Research on Bandwidth Allocation Mechanism of IB QoS

ZHENG Xiao CHEN Zhong-ping WU Zhi-bing
(Jiangnan Institute of Computing Technology, Wuxi 214083, China)

Abstract The quality of service mechanism supported by the InfiniBand network is an important technique to effectively separate and control the bandwidth allocation between the multiple data-flows on the network, which provides a guaranteed transportation service to applications concurrently running in the HPC or datacenters interconnected by IB network. But the "InfiniBand architecture specification" does not propose demand for the IB device manufacturers how to implement IB QoS in detail, especially for the three key components(High-Priority, Low-Priority and Limit of High-Priority) how to numerically control the bandwidth allocation between Virtual Lanes(VL for short) in different priority-ranked queues and with different priority values, and also no documents have closed expatiation on it by now. In this paper, a quantitative formula to compute the ratio of bandwidth between high-and low-priority VLs according the three key components was presented. It was brought forward based on abundant results of bandwidth allocation experiments done on the most popular IB devices and proved to be true by random sample verification. We also found that the bandwidth ratio would change in zigzag shape with the increase of the Low-Priority value and periodically reach a stable value on condition that both High-Priority and Limit of High-Priority are certain. These characteristics were verified by mathematical reasoning method.

Keywords InfiniBand network, Quality of service, Virtual Lane, Bandwidth allocation, Priority value

1 概述

InfiniBand(简称IB)网络因其高带宽、低延迟等特性而被广泛用于高性能计算和数据中心的网络互连。随着应用业务的拓展,越来越多的应用都能得到这两大平台的很好支持,但随之而来的一个问题就是,多负载的并发传输需要共享网络资源,而不同类型的负载对网络的带宽与延迟又有不同的要求,有的对带宽要求很高,而有的对延迟更为敏感,因此常规的处理方法(例如均分带宽)并不能充分利用网络资源,有效地解决I/O整合问题。为此,《InfiniBand规范》[1]从1.2.1版开始引入了QoS特性(作为可选项而非强制项),此后出现的很多IB产品都选择支持QoS特性,例如Mellanox的ConnectX2 HCA卡[2]、MTS3600交换机[3]等。

但是,《InfiniBand规范》1.2.1版主要从原理和框架上对IB设备(包括HCA卡、IB交换机、IB路由器等)要支持的QoS特性需要具备哪些条件做出了规范,至于IB设备厂商应该如何去具体地实现QoS机制,并未给出详细的指导。在目前已公开发表的相关文档中,文献[13]研究了IB QoS的管理

到稿日期:2011-07-29 返修日期:2011-09-15 本文受国家科技支撑计划项目(2011BAH04B03)资助。
郑 霄(1976-),男,博士,工程师,主要研究方向为高性能计算网络,E-mail:uu88zheng@126.com;陈忠平(1975-),男,硕士,工程师,主要研究方向为高性能计算网络;吴志兵(1967-),女,硕士,工程师,主要研究方向为高性能计算网络。

模型与关键技术，并描述了启用 IB QoS 机制的配置方法和实验过程。文献[4]比较了 IB 网络、以太网络和 ASI 网络(Advanced Switching Interconnect)等 3 种网络的 QoS 机制的实现原理。文献[6]在 IB 集群上通过为不同负载配置不同 QoS 服务级别来评估 QoS 的能力。文献[5,7]通过改善 VL 仲裁表机制来优化集群系统的 IB QoS 性能的研究；默认情况下，MPI 库仅允许每个作业使用一个 QP 和一个 VL 进行通信，这样所有的负载(包括对延迟敏感的通信)都在一个队列中排队等待传输。文献[8]尝试了允许每个 MPI 作业使用多个 QP 和多个 VL 进行通信，从而实现了针对 MPI 应用的性能优化。文献[9]研究了 IB QoS 策略对 SDP(Socket Direct Protocol)和 IPoIB 两种上层协议的不同影响。但到目前为止，尚未出现过关于高优先级权值、低优先级权值和 QoS 因子等 3 个关键参数与带宽分配的量化关系的研究。

本文在一定的试验平台基础上开展了以下几方面工作：

1)测试实验：首先进行多应用程序并发运行时的 IB QoS 实验，获取大量带宽分配的实测数据，总结 IB QoS 的带宽分配的规律特性；

2)公式推测：然后根据实测数据，推测出高优先级权值、低优先级权值以及 QoS 因子等 3 个关键变量与 VL 带宽分配的关系式；

3)公式验证：接着采用随机抽样检测的方式来验证公式的正确性，即通过任意选取高优先级权值、低优先级权值和 QoS 因子的变量值进行实测，然后比较此数据是否与通过分配公式计算的结果相符合；

4)公式推理与证明：最后，在验证公式的正确性以后，从理论上推理与证明第 1)步根据实测数据总结的 IB QoS 带宽分配的规律特性是有理论依据的。

本文第 2 节简述 QoS 的实现机制；第 3 节是 IB QoS 的带宽测试实验，给出我们的部分测试数据，并总结 IB QoS 带宽分配的规律特性；第 4 节首先根据测试结果推测带宽分配公式，然后任意选取参数值进行实测，并通过比较实测数据与公式计算结果的吻合度来验证公式的正确性；第 5 节首先将前面总结的规律特性转化成 3 个等价的数学命题，然后依次给予证明，从而为前面总结的带宽分配特性找到理论依据；最后是结束语。

2 IB QoS 的实现机制

IB QoS 的流控机制[1]本质上是一个排队系统，如图 1 所示[13]：在应用上层，启用 QoS 机制的 IB 网络为不同的应用程序贴上不同的服务级别(Service Level，简称 SL)标签，该标签的取值可为 0 到 15 之间的任意整数；在硬件底层，通信链路提供了 16 个带有固定编号的虚通道(Virtual Lane，简称 VL)队列，除了 VL15 专门用于传输管理包之外，VL0－VL14 为数据虚通道，均可用于应用程序的数据包发送。IB QoS 最重要的任务主要有两个：一是建立 SL 与 VL 的映射关系，这是通过 SL2VL 映射表来实现的，它将决定每个标有不同 SL 的应用程序的数据包会通过哪个 VL 进行收发，未与任何服务级别建立映射关系的 VL 将不会得到数据包收发服务，而未分配 SL，或者 SL 未与任何 VL 映射的应用程序也无法得到数据包收发服务；二是仲裁某一时刻物理链路该为哪个 VL 队列的数据包提供服务，这是通过虚通道仲裁表(VL-Arbitration Table)来完成的。

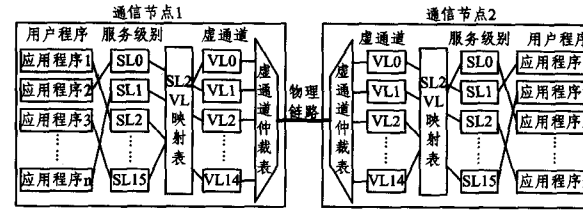

图 1 IB QoS 的实现机制

虚通道仲裁表使用双重方案来仲裁数据 VL 间包的发送顺序[3]：一是抢占式调度(preemptive scheduling)，也就是将 VL 分成两大类：高优先级 VL 和低优先级 VL，高优先级 VL 可以依据系统设定的 QoS 因子参数，在适当的时机抢占使用物理端口进行数据包传输；二是加权公平方案(a weighted fair scheme)，即为每个数据 VL 设定一个优先级权值，属于同一优先级的数据 VL 将依赖自身的优先级权值来决定自己使用物理端口收发数据包的能力。权值为 0 的 VL 也将无法得到数据包收发服务。在高低优先级 VL 队列或低优先级 VL 队列内，端口采用 weighed Round Robin 模式，根据各 VL 所具有的优先级权值(高优先级权值(High-Priority)或低优先级权值(low-Priority))来为各 VL 提供服务；在高低优先级 VL 队列与低优先级 VL 队列之间，端口则根据 QoS 因子(Limit of High-Priority)来决定为哪个 VL 服务。虚通道仲裁表的结构如图 2 所示[10]。

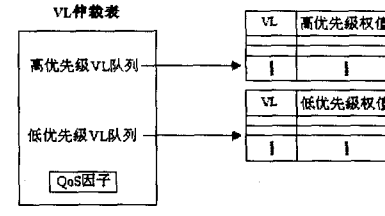

图 2 虚通道仲裁表的结构

简言之，QoS 因子、高优先级权值和低优先级权值这 3 个参数将决定每个 VL 所能分得的带宽。对于它们之间到底有何种量化关系，带宽分配有何规律可循，目前的公开资料尚且缺乏相关的内容，这在一定程度上阻碍了用户对 IB 网络的 QoS 机制的使用，这也是本文主要的研究动机。

3 IB QoS 测试实验

鉴于 IB QoS 的具体实现策略是由各个 IB 设备厂商自行决定的，而 Mellanox[11]是 IB 设备的最大制造商，因此本实验将在 Mellanox 的产品基础上开展。

3.1 测试环境

IB QoS 实验的测试环境是由 3 台 HP 服务器(Intel Xeon16 核 CPU，单核主频 2.80GHz，36GB 主存)和一台 MTS3600 交换机互连而成的，如图 3 所示。每台服务器配置有一张 ConnectX2 HCA 卡(型号为 MT26428)，该卡的理论带宽是 40Gb/s，实测带宽约为 3.4～3.5GB/s。在这 3 个节点中，1 号节点专门用于启动子网管理器(Subnet Manager，简称 SM)服务，以维护 IB 网络的正常运行；2 号节点和 3 号节点分别用作通信测试的服务端和客户端。之所以将 SM 与应用负载分开，是为了尽量降低其它开销对通信测试节点的影响，从而最大程度地保证测试结果的可信度。

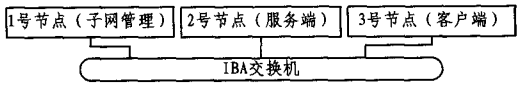

图 3　IB QoS 实验的测试环境

另外，测试程序采用 OFED 自带的点对点测试工具 Qperf。该工具可以从它的官方网站[12]免费下载。

3.2　测试方法

在进行本实验前，为了澄清其它因素（包括 QoS 策略是否启用、消息包大小以及 VL 数量等）对带宽测试的影响，我们事先进行了另外 3 组实验：第一组实验是 SM 分别带 QoS 策略启动和不带 QoS 策略启动，结果发现这两种情况下单应用程序的通信带宽没有变化；第二组实验是设定不同消息包大小进行测试，发现当消息包大小在 64KB～8MB 之间时，单应用程序的带宽基本稳定在 3.4～3.5GB/s；第三组实验是打开多个 VL，每个 VL 任意设定为高优先级或低优先级，结果发现多应用程序下两节点间的总带宽基本上保持不变。基于这 3 个结论可以确定：两通信节点间的通信总带宽与是否启用 QoS 策略无关，也与开通的 VL 数量无关。因此，为了便于量化高优先级权值、低优先级权值以及 QoS 因子 3 个变量与 VL 带宽分配的关系，本实验限定：每个 HCA 卡仅打开两个虚通道（VL0 和 VL1），两者分属高优先级队列和低优先级队列，并将通信的包大小设定为 64KB，两个通信测试程序（App1 和 App2）的服务级别分别设定为 SL0 和 SL1，并且分别对应于 VL0 和 VL1，如图 4 所示。

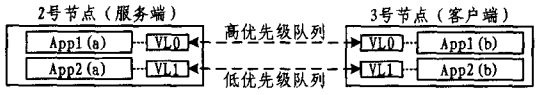

图 4　IB QoS 实验的测试方法

本文目的在于研究高优先级权值、低优先级权值和 QoS 因子这 3 个关键参数与带宽分配的量化关系，因此测试过程就是依次改变这 3 个关键变量的值，以观察分别位于高优先级虚通道队列的 VL0 和低优先级虚通道队列的 VL1 各自所分得的带宽变化。这一过程可用下面的伪代码来描述（见图 5）。

```
for(QoS 因子取值从 0 到 255)
  for(VL0 的优先级权值取值从 0 到 255)
    for(VL1 的优先级权值取值从 0 到 255)
    {
  分别测出 VL0 和 VL1 进行通信的应用程序所分得的带宽;
    }
```

图 5　IB QoS 实验的测试过程伪代码

这 3 个关键变量的值通过 1 号节点的 SM 配置文件进行修改，相关配置内容如图 6 所示。

图 6　带 QoS 策略的 SM 配置示例

其中，qos_max_vls 表示 HCA 卡能用于数据传输的最大 VL 数量，qos_high_limit 是 QoS 因子，而 qos_vlarb_high 和 qos_vlarb_low 分别列出了高、低优先级队列中所有 VL 的优先级权值，qos_sl2vl 依次列出了从 SL0 到 SL15 所映射到的 VL 号。在图 6 的示例中，SM 限定每个 HCA 卡最多支持 8 个数据 VL，QoS 因子取值为 1，VL0 位于高优先级队列，权值设为 16，VL1 位于低优先级队列，权值设为 64，而其余 6 个 VL（VL2－VL7）均未开放使用，因此优先级权值均设为 0；服务级别 SL0－SL7 依次分别对应于 VL0－VL7，SL8－SL15 均对应于 VL15，但 VL15 是管理 VL，不能用于数据传输，因此实际上 SL8－SL15 未被使用。

3.3　测试结果

按照图 5 的测试方法，我们获得了大量的测试数据，图 7 显示了其中的几组测试结果：图 7 中的(a)－(f)分别描述了 QoS 因子依次取值为 2,4,6,8,10,14 时高优先级 VL0 与低优先级 VL1 的带宽比随高优先级权值和低优先级权值的变化关系。图中的横坐标表示低优先级权值，纵坐标表示高优先级 VL0 与低优先级 VL1 的带宽比，而不同的高优先级权值采用不同颜色的数据线来表示。

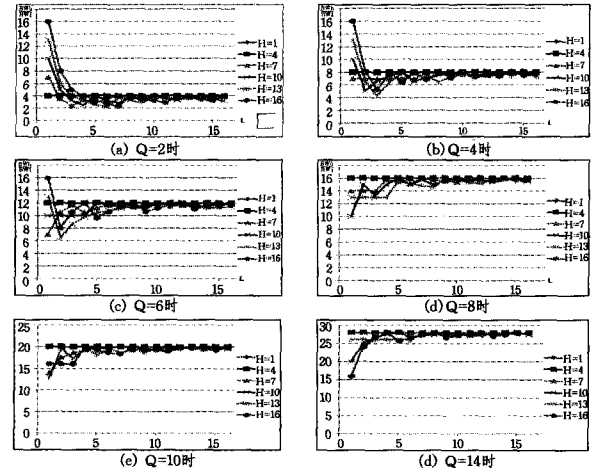

图 7　IB QoS 测试的部分结果

根据测试结果，可以总结出 IB QoS 带宽分配机制的一个重要特性：

特性 1　当 QoS 因子和高优先级权值取值确定时，随着低优先级权值的增长，高优先级 VL 和低优先级 VL 的带宽比呈锯齿形增长，锯齿的偏幅逐渐减小，且周期性达到某一恒定值。

特性 1 只是对已有数据的感性总结，尚且缺乏理论依据；下面首先推测出 IB QoS 的带宽分配公式，并对该分配公式的正确性进行验证，然后根据该公式，用数学证明的方式为这特性 1 寻找理论支撑点。

4　公式推测与验证

4.1　相关符号说明

为了便于以下的描述，表 1 给出了相关符号说明。

表 1　符号说明 1

变量符号	说明
Q	QoS 因子，其值可为[0,255]之间的任意整数。
H	高优先级权值变量，其值可为[1,255]之间的任意整数。
L	低优先级权值变量，其值可为[1,255]之间的任意整数。
BW_H	高优先级 VL 的带宽。
BW_L	低优先级 VL 的带宽。
y	高，低优先级 VL 的带宽之比，即：。

4.2 带宽分配公式推测

根据测试结果，我们推测出如下的 IB QoS 机制的带宽分配公式：

$$\frac{BWH}{BWL}\triangleq y=f(Q,H,L)=\begin{cases}\frac{H}{L}\times\left\lfloor\frac{L}{H}\triangleq\right\rfloor, & Q=0\text{ 时;(a)}\\ \frac{H}{L}\times\left\lfloor\frac{2Q\times L}{H}\right\rfloor, & Q>0\text{ 时;(b)}\end{cases}\tag{1}$$

这里需要对式(1)进行以下说明：

(1) H、L 取值为 [1,255]之间的任意整数，QoS 因子 Q 取值为 [0,255]之间的任意整数；

(2) "⌊⌋"为取整运算；特殊地，若⌊x⌋<1，则取值为 1。

根据说明(2)，式(1)中的情况(a)可视为情况(b)的一种特殊形式，因此后面的公式推论与证明都只需要针对式(1)b，即 $Q>0$ 的情况进行。

4.3 随机抽样验证

为验证式(1)的正确性，我们任意选择了几组参数值进行实测，发现测试结果与按照式(1)计算出来的结果非常吻合，误差不超过 1%，其结果如表 2 所列。

表 2 随机抽查验证的结果

参数值			带宽比		吻合度
Q	H	L	计算值	实测值（单位：GB/s）	
0	2	2	1:1	1.7:1.7	100%
0	3	4	3:4	1.46:1.94	99.7%
1	16	4	4:1	2.72:0.68	100%
1	16	25	48:25	2.28:1.19	99.8%
2	3	2	3:1	2.55:0.85	100%
3	16	12	16:3	2.9:0.54	99.3%
5	10	90	20:1	3.22:0.16	99.4%
6	8	40	12:1	3.6:0.3	100%
8	25	100	16:1	3.2:0.2	100%
16	5	96	1535:48	3.35:0.105	99.8%

由表 2 可以看出，当 3 个参数的取值能够让式(1)的计算值正好为整比时，实测的高优先级 VL 与低优先级 VL 的带宽比结果完全与计算结果相吻合，如(Q,H,L)分别取值为(0,2,2)，(1,16,4)，(2,3,2)，(5,10,90)，(6,8,40)和(8,25,100)时；当不符合这个条件时，实测的带宽比与计算结果有一定的误差，但这个值很小，完全可能是因为测试值仅取小数点后 2 位造成的，因此可以忽略。

5 分配特性的理论证明

5.1 相关说明

为了便于下面对 IB QoS 的带宽分配特性进行形式化描述与证明，这里采用图形的方式来直观描述低优先级 VL 的带宽与其权值的变化关系，如图 8 所示。

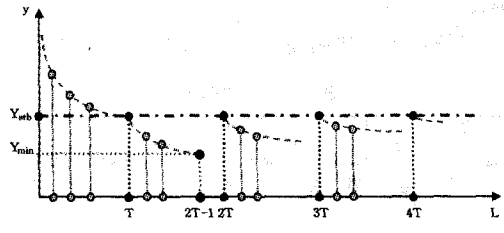

图 8 带宽比与低优先级权值的变化关系示意图

除了表 1 给出的符号以外，对图 8 及下面新出现的符号描述如表 3 所列。

表 3 符号说明 2

变量符号	说明
Y_{stb}	高、低优先级带宽比的稳定值。
Y_{min}	高、低优先级带宽比的最小值。
Y_L	当低优先级权值为 L 时高，低优先级 VL 的带宽比。
T	高、低优先级带宽比取得稳定值时的低优先级权值的周期。
L_n	高、低优先级带宽比第 n 次取得稳定值时的低优先级权值。

5.2 特性推理

特性 1 可以分解为以下 3 条性质：

性质 1 当 QoS 因子和高优先级权值确定时，随着低优先级权值的增长，高、低优先级 VL 的带宽比会周期性达到某个恒定值。

性质 2 当 QoS 因子和高优先级权值确定时，高、低优先级 VL 的带宽比会随着低优先级权值的增长而周期性地出现单调递减。

性质 3 当 QoS 因子和高优先级权值确定时，从第一个变化周期开始，每个周期内高、低优先级 VL 的带宽比的最小值越来越大。

基于式(1)b，这 3 条性质又分别等价于以下 3 个命题：

命题 1 $\forall Q\in[1,255),\forall H\in[1,255),\exists T\in\mathbf{Z}^+,\exists Y_{stb}$，使得当 $L=L_n\triangleq nT(n\in\mathbf{Z}^+)$时，$Y_L\equiv Y_{stb}$。

命题 2 $\forall Q\in[1,255),\forall H\in[1,255)$，则 $\exists T'>0$，当 $L\in[nT',(n+1)T')(n\in\mathbf{Z}^+)$时，函数 $y=f(Q,H,L)$单调递减。

命题 3 $\forall Q\in[1,255),\forall H\in[1,255)$，若 $L\in[T',\infty)$，则当 $L=2T'-1$ 时，有 $Y_L=Y_{min}$。

5.3 数学证明

下面根据式(1)b 对这 3 个命题给予证明。

命题 4 $\forall Q\in[1,255),\forall H\in[1,255),\exists T\in\mathbf{Z}^+,\exists Y_{stb}$，使得当 $L=L_n\triangleq nT(n\in Z^+)$时，$Y_L\equiv Y_{stb}$。

证明：

令 $\frac{2Q}{H}=\frac{\mu}{\lambda}$，$\frac{\mu}{\lambda}$为$\frac{2Q}{H}$的最简式，$\mu\in\mathbf{Z}^+,\lambda\in\mathbf{Z}^+$。

由式(1)b 有：

$$Y_L=f(Q,H,L)=\frac{H}{L}\times\left\lfloor\frac{2Q\times L}{H}\right\rfloor=\frac{H}{L}\times\left\lfloor\frac{\mu}{\lambda}\times I\right\rfloor\tag{2}$$

令 $T=\lambda$，则当 $L=L_n\triangleq nT(n\in\mathbf{Z}^+)$时，由式(2)得：

$$y=\frac{H}{nT}\times\left\lfloor\frac{2Q\times nT}{H}\right\rfloor=\frac{H}{n*\lambda}\times\left\lfloor\frac{1}{\lambda}\times n\lambda\right\rfloor=\frac{2Q}{n}\times n=2Q$$

即 $\exists T=\lambda,\exists Y_{stb}=2Q$，使得当 $L=L_n\triangleq nT(n\in\mathbf{Z}^+)$时，$y\equiv Y_{stb}$。

其中，λ为$\frac{H}{2Q}$的最简式分子，Q为已知的 QoS 因子。

故命题 1 得证。

命题 5 $\forall Q\in[1,255),\forall H\in[1,255)$，则 $\exists T'>0$，当 $L\in[nT,(n+1)T')(n\in\mathbf{Z}^+)$时，函数 $y=f(Q,H,L)$单调递减。

证明：

令 $T'=\frac{H}{2Q}$，则由式(1)b 有：

$$Y_L=f(Q,H,L)=\frac{H}{L}\left\lfloor\frac{2Q\times L}{H}\right\rfloor=\frac{H}{L}\times\left\lfloor\frac{L}{T'}\right\rfloor\tag{3}$$

（下转第 134 页）

况具有极大的性能提升。还要进一步研究数据块大小选择的实际影响、多线程参数的选择以及节点失效情况等。

参 考 文 献

[1] 周文莉，雷振明. BitTorrent 文件共享系统的流量模型与文件评估方法[J]. 计算机工程，2006，32(13)：15-17

[2] 王杨，王汝传，徐小龙，等. 资源共享 P2P 网络的进化博弈激励模型[J]. 计算机工程，2011，37(11)

[3] Vazhkudai S. Enabling the co-allocation of grid data transfers [C]// Grid Computing，2003. 2003：44-51

[4] Vazhkudai S. Distributed downloads of bulk，replicated grid data [J]. Journal of Grid Computing，2004，2(1)：31-42

[5] Yang C T，Yang I H，Li K C，et al. A recursive-adjustment Co-allocation scheme in data grid environments[J]. Distributed and Parallel Computing，2005，3719：40-49

[6] Bhuvaneswaran R S，Katayama Y，Takahashi N. Dynamic co-allocation scheme for parallel data transfer in grid environment [C]// Semactics，Knowledge and Grid，2005. 2005：17

[7] Bhuvaneswaran R，Katayama Y，Takahashi N. Coordinated Co-allocator Model for Data Grid in Multi-sender Environment[C]// Service-Oriented Computing—ICSOC 2006. 2006：66-77

[8] Feng J，Cui L，Wasson G，et al. Toward seamless grid data access：Design and implementation of gridftp on. net[C] // Grid Computing，2005. 2005：8

[9] Chang R S，Chen P H. Complete and fragmented replica selection and retrieval in Data Grids[J]. Future Generation Computer Systems，2007，23(4)：536-546

[10] Ghemawat S，Gobioff H，Leung S T. The Google file system[C]// SOSP'03 Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles. 2003

[11] Yang C T，Yang I H，Li K C，et al. Improvements on dynamic adjustment mechanism in co-allocation data grid environments [J]. The Journal of Supercomputing，2007，40(3)：269-280

[12] 刘道群，孙庆和，刘君. 一种基于不同角色和反馈可信度的 P2P 信誉模型[J]. 重庆邮电大学学报：自然科学版，2010，22(6)：834-839

(上接第 130 页)

令 $g(L)=\left\lfloor \frac{L}{T'} \right\rfloor$，则由式(3)得：

$g(nT')=\left\lfloor \frac{oT'}{T'} \right\rfloor=n$

$\therefore g(L_{n+1})=n+1, g(L_{n+1})-g(L_n)=1$。

$\because L\in[L_n, L_{n+1})$，

$\therefore$ 可令 $L=L_n+x$，其中 $0\leqslant x<T'$。

按照取整函数的特性，有 $g(L)=g(L_n)=n$。

$\therefore f(Q,H,L)=\frac{H}{L_n+x}\times g(L)=\frac{nH}{L_n+x'}$。

由于参数 H、n 和 L_n 均为已知，因此该函数显然是 x 的单调递减函数。

即 $\exists T'=\frac{H}{2Q}$，当 $L\in[nT', (n+1)T')(n\in Z^+)$ 时，函数 $y=f(Q,H,L)$ 单调递减。

故得证。

命题 6 $\forall Q\in[1,255), \forall H\in[1,255)$，若 $L\in[T,\infty)$，则当 $L=2T'-1$ 时，有 $Y_L=Y_{\min}$。

证明：

由命题 2 可知，函数 $y=f(Q,H,L)$ 在每个 $[nT', (n+1)T')$ 区间内 $(n\in Z^+)$ 单调递减，即在每个 $[nT', (n+1)T')$ 区间内，当 $L=(n+1)T'-1$ 时，Y_L 获得区间最小值。下面来证明：当 $L=2T'-1$(即 $n=1$)时，Y_L 也是区间 $[L_n,\infty)$ 的最小值。

令 $L_x=(n+1)T'-1$，由命题 2 可知：

$$Y_{Lx}==\frac{H}{(n+1)T'-1}\times n=\frac{H}{T'+\frac{T'-1}{n}}$$

$\because$ 参数 H 和 T' 均已确定，

$\therefore Y_{Lx}$ 显然是 n 的单调递增函数。

也就是说，当 $n=1$，即 $L=2T'-1$ 时，有 $Y_L=Y_{\min}$。

故得证。

结束语 本文采用实验测试与理论推导相结合的办法来研究 IB QoS 机制的高优先级虚通道与高优先级虚通道的带宽分配量化关系，该项工作成果可用于更好地指导利用 IB QoS 机制进行多并发应用程序下的流量控制，从而更有效地发挥 HPC 平台上 IB 网络资源的利用率。虽然本文的结论只是基于 Mellanox 的 IB 设备的实验平台，但研究方法同样适用于研究其它厂商的设备特性。

参 考 文 献

[1] InfiniBand Trade Association. InfiniBand architecture specification volume 1[s]. Release 1. 2. 1，Nov. 2007

[2] Mellanox Technologies. Virtual Machine Migration Acceleration using Mellanox ConnectX®-2 EN 40Gb/s IO Adapter[R]. www. mellanox. com，Dec. 2010

[3] Mellanox Technologies，MTS3600 36-port InfiniBand Switch Product Development Platform[R]，www. mellanox. com，June 2008

[4] Reinemo S-A，Skeie T，Sodring T，et al. An Overview of QoS Capabilities in InfiniBand，Advanced Switching Interconnect，and Ethernet[J]. IEEE Communications Magazine，2006，44(7)：32-38

[5] Joslfaro F J S，Mendui M，Josuato J. A Formal Model to Manage the InfiniBand Arbitration Tables Providing QoS[J]. IEEE Trans. Computers，2007，56(8)

[6] Martz-Vicente A，Apostolopoulos G，Joslfaro F，et al. Efficient Deadline-Based QoS Algorithms for High-Performance Networks[J]. IEEE Trans. Computers，2008，57(7)

[7] Grant A A R R，Rashti M J. An Analysis of QoS Provisioning for Sockets Direct Protocol vs IPoIB over Modern InfiniBand Networks[C]// P2S2 Workshop，in conjunction with ICPP. 2008

[8] Subramoni H，Lai P，Panda D K. Designing QoS Aware MPI for InfiniBand[R]. Jan. 2009

[9] Grant R E，Rashti M J，Afsahi A. An Analysis of QoS Provisioning for Sockets Direct Protocol vs. IPoIB over Modern InfiniBand Networks[R]. June 2008

[10] Alfaro F J，S'anchez J L，Duato J. A Strategy to Compute the InfiniBand Arbitration Tables[R]. Jan. 2002

[11] www. mellanox. com

[12] www. openfabrics. org

[13] 吴志兵，陈忠平. InfiniBand 网络的 QoS 管理技术研究[J]. 高性能计算技术，2010(1)：33-37