

电力通信中基于动态阈值的流量控制机制研究

刘金锁¹, 孙信军¹, 李 洋¹, 冯 宝¹, 高凯强²

(1. 南瑞集团公司(国网电力科学研究院), 江苏 南京 211000;

2. 南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘 要: InfiniBand 架构(IBA)是一种基于交换机的互连技术,拥有高带宽和低时延的特点。InfiniBand 网络很适合构建高速网络集群系统并且已经被批准为 I/O 技术和进程间通信标准。提出了一种适用于电力广域高性能计算网络的基于动态工作阈值的有效流量控制方法。与以往的静态阈值方法不同,该方法根据一个周期内对链路中业务流量的监听,记录下突发大流量,最后根据所记录的流量特点来动态设定工作阈值。交换机对业务流量评估是一个动态连续的过程,对每一个业务流的最大突发值进行评估,动态调整仲裁表来设置恰当的阈值门限,使每个信道都能适应当前链路的流量特点,提高网络突发大流量的处理能力、带宽利用率和传输效率,降低网络能耗、丢包率和时延。采用最大熵原理(Maximum Entropy, ME)来分析所提出的流量控制机制,并用广义指数分布 GE-Type 模拟业务流到达率和业务服务时间。仿真结果表明,所提出的机制可以实现电力通信网络中广域高性能计算网络流量的有效控制。

关键词: InfiniBand 技术;动态阈值;流量控制;GE-Type;最大熵原理

中图分类号: TP39

文献标识码: A

文章编号: 1673-629X(2017)08-0187-05

doi:10.3969/j.issn.1673-629X.2017.08.039

Research on a Flow Control Mechanism Based on Dynamic Threshold in Power Communication

LIU Jin-suo¹, SUN Xin-jun¹, LI Yang¹, FENG Bao¹, GAO Kai-qiang²

(1. NARI Group Corporation (State Grid Electric Power Research Institute), Nanjing 211000, China;

2. College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: InfiniBand Architecture (IBA) is defined as a switch-based interconnection technology with high bandwidth and low-latency, which is suitable for constructing high-speed networks for cluster systems and has been ratified as a new industry standard for the server I/O and inter-processor communication. An efficient flow control mechanism with dynamic job threshold for InfiniBand networks is proposed. Unlike the existing static threshold methods, it dynamically sets threshold according to the recording traffic in a cycle of traffic monitoring on the link. Switch to traffic assessment is a dynamic and continuous process. The maximum burst value of each business flow to assess, dynamic adjustment of the table to set the appropriate threshold. It can improve the processing capacity, bandwidth utilization, transmission efficiency of network burst traffic and reduce the delay, blocking probability, mean queue length. The principle of Maximum Entropy (ME) is adopted as an effective methodology to analyze the new mechanism with the generalized exponential distribution (GE-Type) for modeling the inter-arrival times and service times of the input traffic. The simulation results show that it can achieve the effective control of traffic flow in the high performance computing network in power communication.

Key words: InfiniBand; dynamic threshold; flow control; GE-Type; maximum entropy principle

1 概 述

随着智能电网的飞速发展,电力通信网络结构日益复杂,承载业务日趋多元化。电力通信部门需要通过可靠、有效的技术手段对网络业务流量进行监控,以

降低能耗、丢包率和网络时延。然而,由于交换机网络技术的固有缺陷,如网络风暴、网络拥塞、流量管理和控制等,在一定程度上影响了站内系统的安全与性能。为避免上述风险,必须做好变电站通信网络的规划设

收稿日期: 2016-08-26

修回日期: 2016-11-29

网络出版时间: 2017-07-05

基金项目: 国家自然科学基金资助项目(61302100, 61471203); 教育部博士点基金资助项目(20133223120002); 国家电网公司 2016 年科技项目

作者简介: 刘金锁(1980-), 男, 硕士研究生, 高级工程师, 研究方向为电力系统通信及信息安全防护技术。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20170705.1650.042.html>

计,应使用具备风暴抑制和流量管理功能的专用交换机,并采用合适的网络架构和流量管控措施^[1]。但现有方式无法完全实现流量的有效控制问题。

高性能计算网络集群已被广泛地应用于不同的领域去解决富有挑战性的问题^[2]。从高端的浮点密集型科学和工程计算问题到商业数据密集任务,很多现行的产品已经实现了吞吐量最大化和延时最小化,但在带宽保证、有限的数据包投递时间和有限的到达时延等方面仍存在不少问题。Infiniband 网络由于其可扩展性成为高性能计算网络的首选。Infiniband 架构是一种新的工业架构标准,可以使 Infiniband 网络支持时延约束和多种 QoS 服务要求的应用。Infiniband 提供了一系列机制,例如:服务级别(Service Levels, SL)、虚拟链路(Virtual Lanes, VL)、虚拟链路仲裁表(VLArbitration Table),通过一定配置可以提供满足不同业务需求的 QoS 服务^[3]。这些机制包括不同的业务类型和不同输出端口的仲裁。仲裁表存储在 Infiniband 网络的交换机中,可以根据严格的 QoS 要求配置数据包的优先权^[4]。现有技术中有一种固定工作阈值的流量控制机制^[5-6],该方法虽在一定程度上提高了链路吞吐量,降低了时延,但是由于部分电力业务具有突发大流量的特点,仅设置一个静态阈值难以满足各类电力业务的通信需求。静态阈值过低,交换机会频繁切换,导致网络延时和能耗增加。相反地,静态阈值设置过高,交换队列的长度就会增加,导致网络延时增加、带宽利用率下降,当突发流量到达时会导致数据包的丢失,可见静态阈值不是有效的配置方法。

文中提出了一种有效地应用于 Infiniband 网络中的动态流量控制机制。该机制的基本思想是在虚拟链路仲裁表中给虚拟链路引入一个动态的工作阈值,这些阈值通过严格的 QoS 限制有效控制了不同业务的带宽划分,从而提高了系统的总体性能。采用广义指数分布对外部的通信量进行建模,捕获网络突发数据流量,利用信息论中的最大熵原理可以得出近似的分析结果,实现对通信网络的简单、可靠、高效的分析和预测。

2 一种基于动态阈值的流量控制机制

2.1 Infiniband 网络

Infiniband 技术规范描述了一个系统区域网络连接了多重独立的处理器平台、I/O 平台和 I/O 设备等。SAN 是一个通信管理设备,支持单个和多个计算机系统的 I/O 流和处理器间通信。IBA 的设计是基于交换机的高速点到点链路互连技术。一个 IBA 网络可以划分为由路由器互连的多个子网,每个子网由一个或多个交换机、处理节点和 I/O 设备组成。在 IBA 中,消息

是通信的基本单元,数据被分成数据包在链路上传输。每一个数据包包括数据头信息和实体数据,每一个包的长度为 256 字节到 4 096 字节^[7]。IBA 有三种机制支持 QoS:服务级别,虚拟链路和虚拟链路仲裁表^[8]。IBA 中规定了最大 16 个服务级别,它取决于管理员如何在不同服务级别之间分派不同的流量类型,并提供了一个字段用来标记服务的级别。根据不同的需求处理不同的业务,可以在一条物理链路上创建多条虚拟链路连接机制。

在 Infiniband 网络中,每一个节点最小有两个最大 16 个服务级别($VL_0, VL_1, \dots, VL_{15}$)。 VL_{15} 是为子网管理预留的,所有端口都支持并有最高的数据业务级别。因为交换机支持不同的服务级别,子网管理器通过端口的使用数量来配置服务级别的数量。当有超过两个服务级别执行时,仲裁机制将允许一个输出节点选择虚拟链路进行传输。由于 VL_{15} 用来进行流量控制,并且拥有最高的优先权,因此仅对数据的服务级别进行仲裁。虚拟链路仲裁表定义了数据通道的优先级别,如图 1 所示。

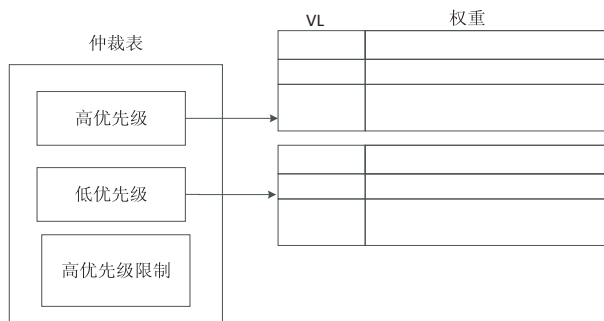


图 1 仲裁表结构

2.2 流量分类

Pelissier 基于目前应用的 QoS 提出了五种业务级别^[9]:专用带宽时间敏感业务、专用带宽业务、尽力优先服务(Preferential Best Effort, PBE)、尽力服务和富有挑战性的流量业务。每一类使用一个不同的 SL,因此可以达到所要求的 QoS。下一步是找到合适的方法填写仲裁表。Pelissier 提出对于 DBTS 业务使用高服务级别,其他业务使用低服务级别。Alfaro 等提出了一种填补仲裁表权重的策略^[5]。上述两种方法在广域高性能网络中处理突发流量时都存在时延过大和丢包率的问题,因此文中提出一种新的有效的流量控制机制。利用动态阈值以有效填写虚拟仲裁表,提高了网络突发大流量的处理能力、带宽利用率和传输效率,降低了网络的能耗、丢包率和时延。

2.3 过程分析

如图 2 所示,本地通信代理根据每个交换机的本地信息来决定接受或拒绝连接请求。这些信息包括输出链路的状态以及它们已预留的带宽。当一个连接被

接受,代理根据连接请求更改虚拟链路仲裁表,而且为每一个虚拟链路设定初始工作阈值,阈值函数是一个增加排队系统利用率的拥塞控制函数,并且是否到达工作阈值取决于每一个虚拟链路的流量大小。交换机动态地记录下链路中的突发大数据流量。每一个虚拟链路保持各自的动态阈值。动态阈值 (Dynamic Threshold, DT) 的计算公式为:

$$DT = L_{cap} - LB_{total} \tag{1}$$

其中, L_{cap} 为虚拟链路的容量;中间变量 LB_{total} 计算如下:

$$LB_{total} = \sum_{i=1}^n LB_{max_i} \tag{2}$$

其中, LB_{max_i} 表示第 i 条虚拟链路观察到的最大流量。

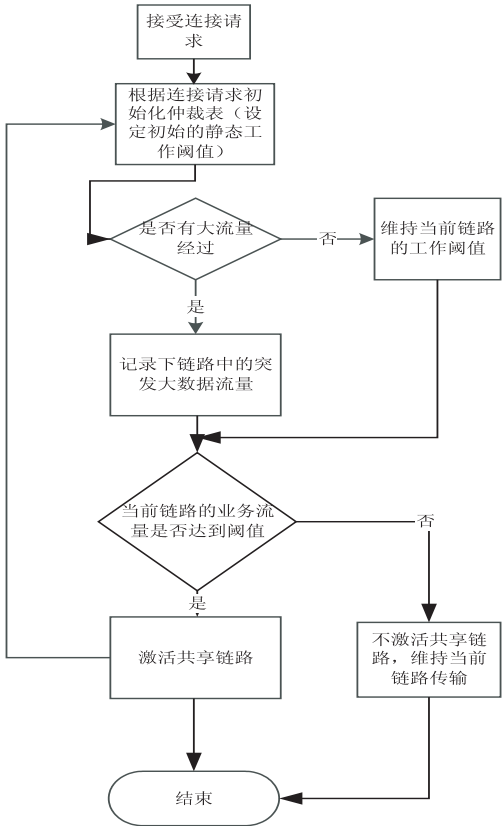


图 2 系统流程图

重点关注拥有高优先权的 VL_i 并且按照服务时间分布,到达工作阈值,先到先服务 (FCFS) 的原则,把它模拟成一个业务到达间隔服从广义指数分布 GE/GE/1/N/ET/FCFS 的排队系统^[10]。业务级别 VL_s 可分为两个部分:常态 VL_s 和共享 VL_s ,如图 3 所示。

当某一特定 VL_i 的业务量到达阈值时,系统就会分派空闲共享 VL_s 传输数据包,共享过程如图 4 所示。

假设虚拟链路 A 的服务速率为 u ,容量为 N ,当高优先权的 VL 业务量达到其阈值 (L_1) 时, VL_A 将会占用一个空闲数据共享虚拟链路 VL_B 。服务率将从 u_1

变为 u_2 ,而容量从 N 增加到 $2N$ 。另一方面,当业务量小于阈值时, VL_A 不能使用共享虚拟链路。同时,若 VL_B 的业务量达到了其阈值 L_2 ,则来自高优先级的虚拟链路将会占用第二条共享的虚拟链路^[11]。

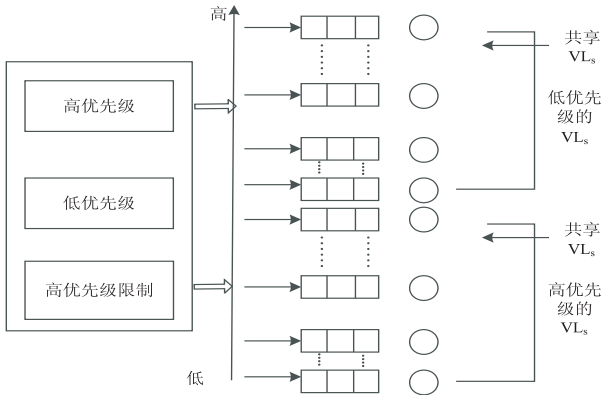


图 3 激活共享链路过程

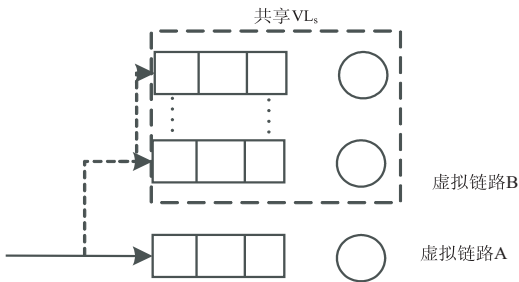


图 4 虚拟链路共享过程

3 算法分析

3.1 GE-Type 分布

GE-Type 分布常被用来模拟到达时间间隔和服务时间^[12-13],形式如下:

$$F(t) = P(W \leq t) = 1 - \tau e^{-\sigma t}, t \geq 0 \tag{3}$$

$$\tau = 2/(C^2 + 1) \tag{4}$$

$$\sigma = \tau v \tag{5}$$

其中, W 为随机变量; $1/v$ 和 C^2 分别为均值和方差。

均值和方差是随机变量的两个重要参数,GE-Type 分布具有很强的通用性和无记忆性,使得很多基于 GE 排队模型系统的分析很方便。突发性到达过程由方差和到达时间间隔来表征,这可以模拟突发性流量。文中用 GE-Type 分布来模拟无线带宽网络中的突发业务,推导出平均队列长度和阻塞概率。提出的流量控制机制基于如下假设:

(1) 当某一虚拟链路的流量达到阈值时,系统就会尝试使用空闲的共享信道传输数据包。如果所有的共享信道被占用,那么这些业务就会在虚拟链路仲裁表中等待。

(2) 一旦虚拟链路 j 接受来自 VL_i 的数据包,那么

在 VL_i 数据传输完成前,不能接受其他 VL 的数据。

(3) 当 VL_i 完成所有的数据包传输时,虚拟链路 i 会被立即释放。

为了清晰起见,每个符号的意义如表 1 所示。

表 1 参数表

参量	定义
n_1	高优先权的 VL 业务量
n_i	第 i 个共享信道的业务量
λ_i	平均到达率
μ_1	高优先级 VL 阈值下的服务完成率
μ_2	到达高优先级 VL 阈值时的服务完成率
μ_i	第 i 个 VL 的业务达到阈值时的服务完成率
T_1	高优先级别的 VL 阈值
T_i	第 i 个共享 VL 的阈值
$\rho_1 = \frac{\lambda}{\mu_1}$	业务量小于 T_1 时的流量强度
$\rho_2 = \frac{\lambda}{\mu_2}$	业务量大于 T_1 且小于 T_2 时的流量强度
$\rho_i = \frac{\lambda}{\mu_i}$	第 i 个阈值以后的流量强度
C_a^2	到达时间间隔分布的方差(SCV)
C_s^2	服务时间分布的方差(SCV)
$P(n)$	队列状态概率

在接下来的分析中,假定 $\lambda_i, \mu_1, \mu_2, \mu_i, C_a^2$ 和 C_s^2 形成一个基本集合,根据先前的知识得出队列分布概率。设 $P(n)$ 表示状态概率,其中 n 取:

$$n = \begin{cases} 0, 1, \dots, N(n_1 < T_1) \\ 0, 1, \dots, iN(T_{i-1} \leq n_i \& n_i < T_i), i=2, 3, \dots, n \\ 0, 1, \dots, (i+1)N(T_i \leq n_i) \end{cases} \quad (6)$$

式(6)归一化得:

$$\sum_{n=0}^{iN} P(n) = 1 \quad (7)$$

平均队列长度为:

$$\sum_{n=1}^{\infty} nP(n) = \bar{L} \quad (8)$$

3.2 最大熵原理

状态概率分布 $P(n)$ ($n=1, 2, \dots, iN$) 可由最大熵函数表征:

$$H(p) = - \sum_{n=0}^{iN} P(n) \log P(n) \quad (9)$$

系统模型的最大熵的状态概率分布由下式给出:

$$P(n) = \frac{1}{Z} g_1^{h_1(n)} g_2^{h_2(n)} \dots g_i^{h_i(n)} x^n y^{f(n)}, 0 \leq n \leq iN \quad (10)$$

其中:

$$Z = \sum_{n=0}^{iN} g_1^{h_1(n)} g_2^{h_2(n)} \dots g_i^{h_i(n)} x^n y^{f(n)}, 0 \leq n \leq iN \quad (11)$$

利用归一化限制,可以推导出 $P(0)$:

$$P(0) = \frac{1}{Z} =$$

$$\frac{1}{1 + g_1 \frac{x_1 - x_1^{T_1}}{1 - x_1} + g_2 \frac{x_2 - x_2^{T_2}}{1 - x_2} + \dots + g_i \frac{x_i - x_i^{iN - T_i}}{1 - x_i}} \quad (12)$$

由式(10)和式(12)可得队列长度的概率分布为:

$$P(n) = \begin{cases} P(0) g_1 x_1^n, n_i < T_1 \\ P(0) g_1 x_1^{T_1} x_2^{T_2} \dots x_i^{iN - T_i}, T_{i-1} \leq n_i \& n_i < T_i \\ P(0) g_1 x_1^{T_1} x_2^{T_2} \dots x_i^{iN - T_i}, T_i \leq n_i \end{cases} \quad (13)$$

式(13)表示在平均队列长度的约束下门限值为 T 的 GE/GE/1 队列的最大熵^[14],所以很容易得出拉格朗日系数 g_i 和 x_i ($i=1, 2$) 为:

$$g_1 = \frac{\rho_1(1 - x_1)}{x_1(1 - \rho_1)} = \frac{\rho_1^2}{(\bar{L} - \rho_1)(1 - \rho_1)} \quad (14)$$

$$x_1 = \frac{\bar{L}_1 - \rho_1}{\bar{L}_1} \quad (15)$$

其中

$$\rho_1 = \frac{\lambda}{\mu_1} \quad (16)$$

$$\bar{L}_1 = \frac{\rho_1}{2} (1 + \frac{C_a^2 + \rho_1 C_s^2}{1 - \rho_1}) \quad (17)$$

进一步得出:

$$g_i = \frac{(1 - x) \rho_i}{(1 - \rho) x_i} \quad (18)$$

$$x_i = \frac{\bar{L}_i - \rho_i}{\bar{L}_i} \quad (19)$$

最后得出平均队列长度为:

$$\rho = \frac{t \rho_1 + (2N - t) \rho_2}{2N} \quad (20)$$

$$L = \bar{L}_1 + x_1^{T_1} \bar{L}_2 + x_2^{T_2} \bar{L}_3 + \dots + x_{i-1}^{T_{i-1}} \bar{L}_i + T_i x_1^{T_1} (\rho - (1 - \rho)) g_1 \frac{x_1 - x_1^{T_1}}{1 - x_1} - (1 - \rho) g_2 \frac{x_2 - x_2^{T_2}}{1 - x_2} - \dots - (1 - \rho) g_i \frac{x - x_i^{T_i}}{1 - x_i} \quad (21)$$

其中

$$\bar{L}_1 = \frac{1}{2} (1 + \frac{C_a^2 + \rho_1 C_s^2}{1 - \rho_1}) \quad (22)$$

$$\bar{L}_2 = \frac{1}{2} (1 + \frac{C_a^2 + \rho_2 C_s^2}{1 - \rho_2}) \quad (23)$$

$$\bar{L}_i = \frac{1}{2} (1 + \frac{C_a^2 + \rho_i C_s^2}{1 - \rho_i}) \quad (24)$$

4 仿真结果分析

文中仿真场景有两个节点(信道适配器)和一个

路由器,每一个节点包含四条虚拟链路,缓冲区容量为十个数据包。当高优先权的链路到达阈值时,服务速率从 u_1 到 u_2 再到 u_3 ,容量从 N 到 $2N$ 再到 $3N$ 等等。另一方面,如果到达业务的数据量小于阈值时,服务速率和容量将会减少。仿真中的参数设置为: $i = 3, u_1 = 8.0, u_2 = 16.0, u_3 = 12.0, C_a^2 = 5$ 。

用仿真结果和 ME 算法值进行误差评估,平均队列长度误差函数 (Error Measures, EM) 为:

$$EM(\bar{L}_i) = \left| \frac{\text{Sim}(\bar{L}_i) - \text{ME}(\bar{L}_i)}{\text{Sim}(\bar{L}_i)} \right|, i = 1, 2 \quad (25)$$

基于 ME 对文中提出的动态阈值模型和先前的静态阈值进行研究对比,如图 5 所示。

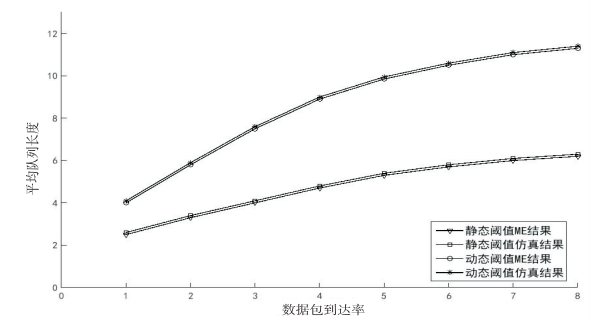


图 5 对平均队列长度的影响

图中,仿真结果和 ME 预测的结果在误差允许的范围内是一致的。通过计算发现,误差值在 0.05 ~ 0.1 之间,证明了结果分析的正确性。所以阈值函数可以由 ME 的解来计算,并且图中动态阈值的平均队列长度明显好于静态阈值,减少了队列的平均长度,降低了通信时延和丢包率,并且提高了带宽利用率和传输效率。

图 6 为动态阈值函数和静态阈值函数的阻塞概率对比曲线。

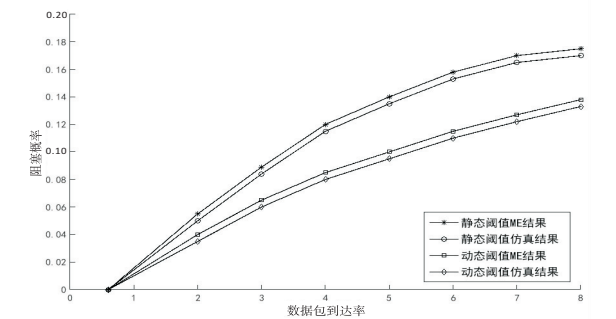


图 6 对阻塞概率的影响

从图 6 可知,动态阈值算法可以有效降低阻塞概率,提高网络突发大流量的处理能力。

5 结束语

文中提出了基于动态阈值的流量控制机制,该机制适用于电力通信系统的广域高性能

能计算网络中基于动态阈值的有效流量控制方法。首先本地代理根据连接请求更改虚拟链路仲裁表,并且为每一个虚拟链路设定初始的静态工作阈值,其次交换机动态地记录下链路中的突发大数据流量,最后仲裁表根据记录的突发数据流量信息,动态调整本链路中的工作阈值。仿真结果表明,该方法提高了网络突发大流量的处理能力、带宽利用率和传输效率,降低了能耗、丢包率和网络时延。

参考文献:

[1] 郑涪文. 交换机流量限制技术及其在智能变电站的技术应用分析[J]. 华东科技:学术版,2016(6):254.

[2] 黄建强,吴利,曹腾飞,等. 基于高性能计算平台和 WRF 环境实验的教学改革[J]. 实验室研究与探索,2016,35(2):94-97.

[3] 夏晓爽,刘轶,王允彬,等. 基于 InfiniBand 的多链路 mesh/torus 大规模并行系统互连网络[J]. 计算机研究与发展,2012,49(1):76-82.

[4] 徐迪威,余焯佳. InfiniBand 高速互连网络设计的研究[J]. 电脑与电信,2012(7):26-29.

[5] Alfaro F J, Nchez J, Duato J. A new strategy to manage the InfiniBand arbitration tables[J]. Journal of Parallel & Distributed Computing,2009,69(6):508-520.

[6] Gran E G, Reinemo S A, Lysne O, et al. Exploring the scope of the InfiniBand congestion control mechanism[C]//26th IEEE international symposium on parallel and distributed processing. [s. l.]:IEEE,2012:1131-1143.

[7] Kim E J, Yum K H, Das C R, et al. Performance enhancement techniques for InfiniBand architecture [C]//International symposium on high-performance computer architecture. [s. l.]:IEEE,2003.

[8] Alfaro F J, Sánchez J L, Duato J. A strategy to manage time sensitive traffic in InfiniBand [C]//Parallel and distributed processing symposium. [s. l.]:[s. n.],2001.

[9] Pelissier J. Providing quality of service over Infiniband architecture fabrics [C]//Symposium on hot interconnects. [s. l.]:[s. n.],2000:127-132.

[10] 熊方方. M/M/1/N→M/M/c/K 排队系统及其在锚地中的应用研究[D]. 武汉:武汉理工大学,2010.

[11] 王东洋. 基于虚拟设备的虚拟交换机设计[J]. 软件,2012,33(1):42-45.

[12] 周宗好. 通信网络中的排队模型研究[D]. 镇江:江苏大学,2011.

[13] Yahyaoui N, Sfina N, Lazzari J L, et al. Stark shift of the absorption spectra in Ge/Ge 1-x Sn x /Ge type-I single QW cell for mid-wavelength infra-red modulators[J]. Superlattices & Microstructures,2015,85:629-637.

[14] Kouvatsos D D. Entropy maximisation and queueing network models[J]. Annals of Operations Research,1994,48(1):63-126.