UNIVERSITY of BRADFORD Library

University of Bradford eThesis

This thesis is hosted in Bradford Scholars – The University of Bradford Open Access repository. Visit the repository for full metadata or to contact the repository team

© University of Bradford. This work is licenced for reuse under a Creative Commons Licence.

Performance Analysis and Improvement of InfiniBand Networks

Shihang Yan

Ph.D

2012

Performance Analysis and Improvement of InfiniBand Networks

Modelling and Effective Quality-of-Service Mechanisms for

Interconnection Networks in Cluster Computing Systems

Shihang Yan

A thesis submitted for the degree of

Doctor of Philosophy

Department of Computing

School of Computing, Informatics and Media

University of Bradford

2012

Abstract

The InfiniBand Architecture (IBA) network has been proposed as a new industrial standard with high-bandwidth and low-latency suitable for constructing high-performance interconnected cluster computing systems. This architecture replaces the traditional bus-based interconnection with a switch-based network for the server Input-Output (I/O) and inter-processor communications. The efficient Quality-of-Service (QoS) mechanism is fundamental to ensure the import at QoS metrics, such as maximum throughput and minimum latency, leaving aside other aspects like guarantee to reduce the delay, blocking probability, and mean queue length, etc.

Performance modelling and analysis has been and continues to be of great theoretical and practical importance in the design and development of communication networks. This thesis aims to investigate efficient and cost-effective QoS mechanisms for performance analysis and improvement of InfiniBand networks in cluster-based computing systems.

Firstly, a rate-based source-response link-by-link admission and congestion control function with improved Explicit Congestion Notification (ECN) packet marking scheme is developed. This function adopts the rate control to reduce congestion of multiple-class traffic. Secondly, a credit-based flow control scheme is presented to reduce the mean queue length, throughput and response time of the

system. In order to evaluate the performance of this scheme, a new queueing network model is developed. Theoretical analysis and simulation experiments show that these two schemes are quite effective and suitable for InfiniBand networks. Finally, to obtain a thorough and deep understanding of the performance attributes of InfiniBand Architecture network, two efficient threshold function flow control mechanisms are proposed to enhance the QoS of InfiniBand networks; one is Entry Threshold that sets the threshold for each entry in the arbitration table, and other is Arrival Job Threshold that sets the threshold based on the number of jobs in each Virtual Lane. Furthermore, the principle of Maximum Entropy is adopted to analyse these two new mechanisms with the Generalized Exponential (GE)-Type distribution for modelling the inter-arrival times and service times of the input traffic. Extensive simulation experiments are conducted to validate the accuracy of the analytical models.

Keywords: Performance Analysis, Modelling, QoS, InfiniBand, Interconnection Networks, Cluster Computing Systems.

Acknowledgements

First of all, I would express my sincere gratitude to my supervisor Dr. Geyong Min, for his continuous guidance and the supportive way he has helped me during the course of my PhD. His rich expertise, deep insight, helpful suggestions, and constructive feedback guided me through the work on this thesis. He also gave a lot of friendly and warm helps on my life. One could not wish for a better or friendlier supervisor.

I would like to thank my second supervisor, Prof. Irfan Awan, his broad knowledge and invaluable advice provided many insightful comments and suggestions on my research study. I also would like to thank my friends Jia Hu, Yuelei Wu, Lan Wang, Sha Sha, Hatem, and Noushiu Najjari for their emotional support.

Special thanks to my parents for understanding, love and encouragement throughout all my studies in UK. Without their unconditional support, it would have been impossible to finish my study.

Last but not least, I would like to take this opportunity to thank the Department of Computing, School of Computing, Informatics and Media, University of Bradford.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACK	acknowledgment packet
AIMD	Additive Increase Multiplicative Decrease
AJT	Arrival Job Threshold
ATM	Asynchronous Transfer Mode
BE	Best-Effort
CC	Credit-Counter
CPP	Compound Poisson Process
DB	Dedicated Bandwidth
DBTS	Dedicated Bandwidth Time Sensitive
ECN	Explicit Congestion Notification
ET	Entry Threshold
FCFS	First Come First Served
FIMD	Fast Increase Multiplicative Decrease
GE	Generalized Exponential
GID	Global Identifier
GUID	Globally Unique Identifier
HCA	Host Channel Adapter
HPC	High Performance Computing
IBA	InfiniBand Architecture
IBTA	InfiniBand^{SM} Trade Association
I/O	Input/Output
LID	Local Identifier
ME	Maximum Entropy
MPLS	Multi-Protocol Label Switching
P2P	Point-to-Point
PBE	Preferential Best-Effort
PCI	Peripheral Component Interconnect

PDF	Probability Distribution Function
pdf	probability density function
PIPD	Power Increase and Power Decrease
QoS	Quality of Service
RAID	Redundant Array of Inexpensive Disks
SAN	System Area Network
SCV	Squared Coefficient of Variation
SLs	Service Levels
SMP	Symmetric Multi-Processor
TCA	Target Channel Adapter
TCP	Transmission Control Protocol
VC	Virtual Circuit
VLAribitration Table	VirtualLane Arbitration Table
VLs	Virtual Lanes
WAN	Wide Area Networking

Publications

Journal Publications

[1] S.H. Yan, G.Y. Min, and I. Awan, "Effective admission and congestion control for interconnection networks of cluster computing systems," *International Journal of High Performance Computing and Networking,* vol. 5,no.4, pp. 374-380, 2008.

[2] S.H. Yan, G.Y. Min, and I. Awan, "Performance analysis of credit-based flow control in Infiniband interconnection networks," *Journal of Interconnection Networks,* vol. 7, no.4, pp. 535-548, 2006.

Conference Publications

Some papers in preparation for IEEE GLOBECOM 2012

[1] S.H. Yan, I. Awan, and G.Y. Min, "Performance analysis of an efficient flow control mechanism with job threshold in InfiniBand networks," *presented at the 22nd Advanced Information Networking and Applications (AINA'2008)*, GinoWan, Okinawa, Japan, 25-28 March, 2008, pp.249-256.

[2] S.H. Yan, I. Awan, and G.Y. Min, "Performance analysis of an active flow control mechanism with entry threshold in InfiniBand network," *presented at the 21st Advanced Information Networking and Applications (AINA'2007)*, Niagara Falls, Canada, 21-23 May, 2007, pp.118-125.

[3] S.H. Yan, G.Y. Min, and I. Awan, "An enhanced congestion control mechanism in InfiniBand networks for high performance computing systems," *presented at the 20th Advanced Information Networking and Applications (AINA'2006)*, Vienna, Austria, 18-20 April, 2006, pp.845-850.

[4] S.H. Yan, G.Y. Min, and I. Awan, "An enhanced congestion notification mechanism in InfiniBand networks," *presented at the Postgraduate*

Symposium on Convergence of Telecommunications, Networking and Broadcasting (PGNet'2005), Liverpool, UK, 27-28 June, 2005, pp.385-390.

[5] S.H. Yan, G.Y. Min, and I. Awan, "Quality of service mechanisms for InfiniBand architecture networks," *presented at the Postgraduate Research Conference in Electronics, Photonics, Communications & Networks, and Computing Science (PREP'2005)*, Lancaster UK, 30 March -1 April, 2005, pp.209-210.

Chapter 1

Introduction

The IBA is a new industry-standard architecture for server I/O and inter-server communications [9, 49, 103]. It was developed by the InfiniBand^{SM} Trade Association (IBTA) to provide the levels of reliability, availability, performance, and scalability necessary for present and future server systems much higher than those can be achieved with bus-oriented I/O structures [96]. It defines a System Area Network (SAN) environment where multiple processor nodes and I/O devices are interconnected using a switched Point-to-Point (P2P) network [19, 80, 85, 117].

The rest of this chapter is organized as follows: the motivations and challenges of this research are pointed out in Section 1.1. The research aims and major contributions of this thesis are then introduced in Section 1.2. Finally, the outline of this thesis is presented in Section 1.3.

1.1 Motivations and Challenges

High Performance Computing (HPC) Systems combine multiple Symmetric Multi-Processor (SMP) computer systems together with high-speed interconnections to achieve the raw-computing power of supercomputers [92, 93, 111]. These systems work in tandem to complete a single request by dividing the work among the server nodes, reassemble the results and present them to the client as if a

single-system did the work. Clustered, standards-based, building block architectures can now challenge traditional multiple processor, tightly-coupled, vertically-scaled computing designs. HPC clusters have been widely used for solving problems in various application domains. These applications range from high-end, floating-point intensive scientific and engineering problems to commercial data-intensive tasks. Nowadays, InfiniBand is quickly becoming the choice of interconnection networks for HPC systems [101].

I/O buses have become a bottleneck for disk access, especially in HPC. Because the bus-based I/O systems do not use their underlying electrical technology well enough to provide high data transfer bandwidth from the system to devices. Today, most popular I/O bus offers limited bandwidth and this limitation is unacceptable for a large number of current applications and server systems. InfiniBand is a standard for communication between processing nodes and I/O devices as well as for interprocessor communication. Instead of directly replacing the Peripheral Component Interconnect (PCI) bus [34, 98], IBA uses a switch-based interconnection to access I/O device. These devices are attached to a Host Channel Adapter (HCA), which is connected to the PCI bus. So the HCA affords the desired reliability, concurrency, and security.

On the other hand, most of the current networking protocols, such as TCP/IP, are not suitable for most HPC cluster applications because of its high CPU overhead and high latency. IBA has been proposed as a new industrial standard that provides high-bandwidth and low-latency data transfers for high-speed I/O inter-processor

communication. It is designed around a switch-based interconnection technology with high-speed point-to-point links. The point-to-point interconnection means that in the IBA networks every link has exactly one device connected at each end of the link, thus providing the better performance than traditional bus-shared architecture [83, 123]. It is designed not only to enable large-scale server clusters but also to provide the ability to those clusters into Grid Computing environments.

An IBA network is divided into subnets interconnected by routers, each subnet consisting of one or more switches, processing nodes and I/O devices. Processing nodes can include CPUs and memory modules, and they use the HCAs to connect to the fabric I/O devices. I/O devices can have any structure, from a simple console to a RAID subsystem and these devices use target channel adapters (TCAs). Each IBA device has a globally unique identifier (GUID). [45, 46, 115, 116].

1.2 Research Aims and Contributions

The research work in this thesis is focused on developing cost effective analytical models for performance evaluation of QoS of the InfiniBand networks. All intermediate objectives of this thesis and the steps to achieve the main research aims are outlined below:

- To develop a stochastic, reliable and efficient mechanism for rate-based congestion control in the high performance cluster computing systems with the InfiniBand Networks environment.

- To develop a new analytical Markovian network model for performance evaluation of enhanced credit-based flow control mechanism.
- To propose the analytical models for performance analysis in Inifniband Networks with active flow control scheme under Generalized Exponential (GE) distribution.

The original contributions of this research are outlined as follows:

- An Explicit Congestion Notification (ECN) packet marking scheme is used. To this end, an effective source response function – Power Increase and Power Decrease (PIPD) that adopts the rate-based flow control with a window limit to reduce congestion of multiple-class traffic in the InfiniBand networks is designed.
- An enhanced credit-based flow control scheme is proposed. Credit-based flow control that can be used to support both end-to-end and link-level flow control is becoming increasingly popular in high performance computing systems, e.g. InfiniBand networks. A Markov chain is used to model the queuing system of this new credit-based scheme.
- Two efficient flow control mechanisms are presented for InfiniBand networks in order to reduce the delay, blocking probability, and mean queue length as well as improve the system utilization. The new mechanisms are all based on the setting threshold science the threshold function is one of the congestion control methods to improve the QoS of

the queueing system. One mechanism is to set the Entry Threshold (ET) for each entry in VLArbitration table based on the number of the entry for flow control, and the VLArbitration table is stored in the InfiniBand switches. The other mechanism is to set an Arrival Job Threshold (AJT) for virtual lanes in the VLArbitration table, which is based on the number of jobs in each virtual lane.

1.3 Outline of the Thesis

The remainder of this thesis is organized as follows.

Chapter 2 firstly introduces the architecture of the InfiniBand Networks and the traffic classification for the different traffic flows. Moreover, this chapter gives an overview of QoS for the InfiniBand networks, including Service Levels (SL), Virtual Lanes (VL) and VirtualLane Arbitration Table (VLArbitration Table). Finally, the Generalized Exponential (GE) distribution and Maximum Entropy (ME) formalism are introduced. The GE-Type queueing system is an effective methodology for modelling the inter-arrival times and service times of the input traffic.

Chapter 3 studies the rate-based flown control in the InfiniBand networks. Firstly, an improved link-by-link admission control mechanism is presented. Then congestion spreading is explained to present the congestion control in the InfiniBand networks. Following this, a new response function is proposed, which is referred to

as Power Increase and Power Decrease (PIPD) based on the packet marking scheme.

Chapter 4 proposes an enhanced credit-based flow control mechanism for InfiniBand networks. By virtue of such a scheme, the downstream node sends credits to the upstream node indicating the availability of buffer spaces. Upon receiving credits, the upstream node injects packets into the networks. A queueing network model is developed to evaluate the performance of this scheme.

Chapter 5 proposes two efficient flow control mechanisms based on the threshold function: Entry Threshold (ET) and Arrival Job Threshold (AJT). The approximate analytical solution is obtained using information theoretic principle of ME with traffic modelled by GE-type distribution.

Finally, Chapter 6 concludes the thesis and highlights the future research work.

Chapter 2

Background and Literature Review

IBA has been proposed as a new industry standard for high-speed I/O inter-processor communication. It is designed around a switch-based interconnection technology with high-speed point-to-point links [15, 59, 88, 108]. The point-to-point interconnection means that in the IBA networks every link has exactly one device connected at each end of the link, thus providing greater performance than the traditional share bus architecture. It is designed not only to enable large-scale server clusters but also to provide the ability to those clusters into Grid Computing environments [87].

In the area of performance modelling and evaluation, cost-effective algorithms for queueing and network models under various traffic handling schemes are widely used in distributed systems, transformation networks, flexible manufacturing system, and communication networks with QoS guarantees [8, 70, 73].

Many existing analytical models for performance analysis in interconnection networks are based on the non-bursty Poisson arrival process. More recently, an analytical mode, Maximum Entropy (ME), has been proposed for Queueing Networks Models (QNMs) [3, 84], which provides a self-consistent method of inference for characterising an unknown but true probability distribution, subject to known (or known to exist) mean value constraints. The ME solution can be

expressed in terms of normalizing constant and product of Lagrangian coefficients corresponding to the constraints. The traffic generated by the source nodes is moddled by the Compound Poisson Process (CPP) with geometrically distributed batch size of, equivalently, Generalised Exponential (GE) distributed inter-arrival time to capture the properties of the bursty and batch arrivals. ME analytic solutions are subject to appropriate GE-type queueing and delay theoretic mean value constraints and some closed-form expressions are determined for the state and blocking probability distributions. In Chapter 5, the principle of ME is adopted as an effective methodology to analyse the new mechanisms with GE-Type distribution for modelling the inter-arrival times and service times of the input traffic [74].

The rest of this chapter is organized as follows. The overview of InfiniBand network architecture is introduced in Section 2.1. The classification for the different traffic flows in InfiniBand networks is given in Section 2.2. A detailed literature review on QoS in IBA is then presented in Section 2.3. Section 2.4 is the introduction of Generalised Exponential (GE) distribution and the solution of GE-Type queueing models approximated by ME analysis. Finally, Section 2.5 concludes the chapter.

2.1 InfiniBand Network Architecture Overview

An IBA network is divided into subnets interconnected by routers, each subnet consisting of one or more switches, processing nodes and I/O devices. Processing nodes include CPUs and memory modules and use the HCAs to connect to the fabric.

I/O devices, simple as RAID subsystem is connected to the fabric by the TCAs. IBA subnet is structured by the fabric [12, 44]. Routing in IBA subnets is distributed, based on forwarding tables stored in each switch. IBA supports any topology defined by the user, including irregular noes, in order to provide flexibility and incremental expansion capability. IBA is an industry-standard architecture with two basic characteristics[9]:

- Point-to-Point connections: all data transfer in IBA is point-to-point, not bused. This avoids arbitration issues, provides fault isolation, and allows scaling to large size by the use of switched networks.
- Channel (message) semantics: commands and data are transferred between hosts and devices not as memory operations but as messages.

The smallest complete IBA unit is a subnet, as illustrated in Fig. 2.1 Multiple subnets can be joined by routers to create large IBA networks. The elements of a subnet are processor node, switches, links and a subnet manager. Processor nodes, such as hosts and devices, send messages over links to other processor; the messages are routed by switches. Routing is defined by the Subnet Manager. Channel Adapters connect processor to links.

Processor nodes are directly attached to a switch through a HCA and I/O devices can be attached to a switch through a TCA. HCA has a collection of features that are defined to be available to host programs; and a TCA has no defined software interface. The position of an HCA in a host system is vendor-specific [13, 126].

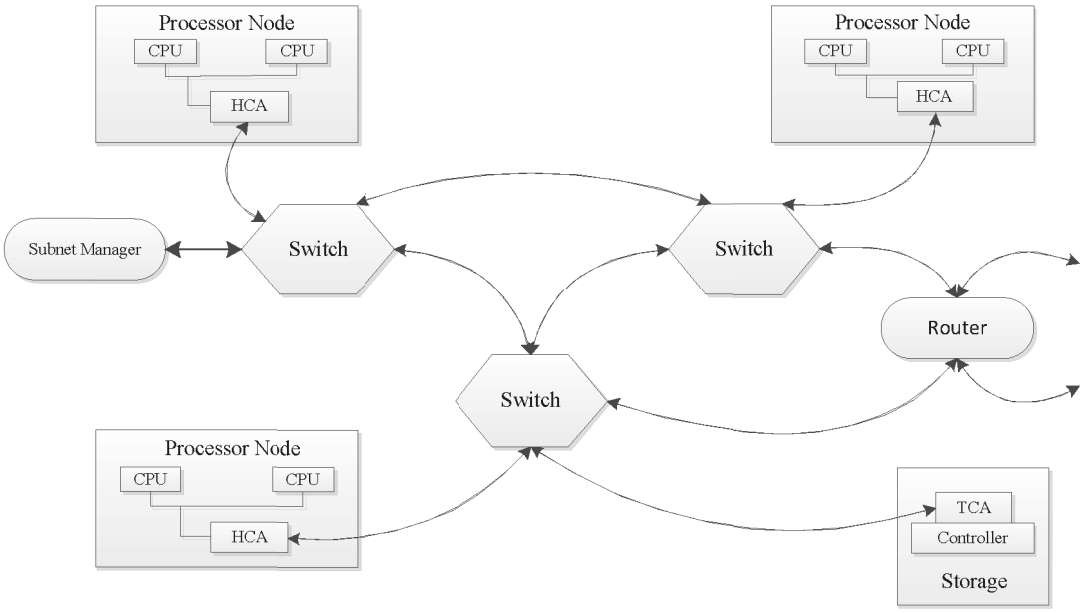


Fig. 2. 1: IBA subnet: each subnet includes a set of switches and point-to-point links

It is expected that initial implementations will provide HCAs as cards attached to a standard I/O bus in order to quickly provide platforms for software development and evaluation. Channel adapters source the several communication service types of IBA, using queues to hold requests for work to be done and for completion. IBA links are bidirectional point-to-point communication channels, and may be copper cable, optical fiber or printed circuit on a backplane.

IBA switches route messages from source to destination based on forwarding tables. The forwarding tables are programmed with forwarding information initialization and after network modification. Messages are segmented into packets for transmission on links and through switches. Switch size, represented by the number of ports, is vendor-specific, as is the link bandwidth supported [65, 66, 78]. It is anticipated that a wide variety of switch implementations will be available with very different capabilities and price points. The maximum switch size supported in

IBA is 256 ports, and switches can be cascaded to form large networks. Switch also optionally supports multicast. Packets sent to a multicast address and then are replicated in the switch, sent out multiple ports, as defined by separate multicast forwarding tables. Switches also support multiple Virtual Lanes (VL_s) through a mechanism called Service Levels (SL_s); this will be discussed in Section 2.3. Routing between different subnets is carried out on the basis of a Global Identifier (GID) and the addressing used by switches is with Local Identifiers (LID).

IBA management is defined in terms of managers and agents. Managers are active entities; agents are passive entities that respond to messages from managers. Every IBA subnet must contain a single master subnet manager, residing on a processor node or a switch. It discovers and initializes the network, assigning LIDs to all elements [118].

2.2 Traffic Classification

QoS refers to the capability of a network to provide the better service to network traffic or applications over various networking or interconnect technologies. Such networking and interconnect technologies include Asynchronous Transfer Mode (ATM), Ethernet, SONET, InfiniBand and others [17, 18, 79, 81, 97, 124]. The primary goal of QoS is to provide priority to selected traffic including dedicated bandwidth and controlled latency. However, depending on the technology and application, QoS offers other benefits as well. For example, for networking technologies such as Ethernet, QoS provides better handling of packet loss or packet drop

characteristics and improves upon best-effort service delivery. For Wide Area Networking (WAN) and Internet topologies, QoS related protocols like Virtual Circuits (VCs) for ATM networks and Multi-Protocol Label Switching (MPLS) for Ethernet networks provide tunnelling and security services as well [1, 108].

InfiniBand has been traditionally used in high-performance computing and clustering applications. This application focuses the inherent high bandwidth and low latency characteristics available in InfiniBand networks [28, 62] that are important because inefficiencies in those parameters are the reason why congestion occurs in the first place and lead to creation of mechanisms for congestion control and QoS [4, 37]:

- **Bandwidth:** The latest available InfiniBand HCA and switch solutions can deliver up to 20Gbps and 60Gbps bandwidth respectively using high-performance processors and PCI Express-based motherboards on end nodes.

- **Latency:** Through the use of proven and mature RDMA, zero copy, kernel bypass and transport offload solutions, HCAs in end nodes can sustain very low latencies. The combination of low-latency HCA features with the cut-through forwarding mechanism available in InfiniBand switches results in very low end-to-end latency for applications – less than 3 microseconds.

At the heart of any QoS implementation is the concept of traffic classes or flows. A combination of source and destination addresses, source and destination socket numbers, or a session identifier may be used to define a flow or a traffic class. Or more broadly, any packet from a certain application, from an incoming interface, or from a certain user or user group can be defined as a flow or a traffic class. Pelissier proposed in [91] a traffic classification for the different traffic flows. This classification is based on the QoS requirements of the applications.

- **DBTS** (Dedicated Bandwidth Time Sensitive): This type of traffic requires a given minimum bandwidth and must be delivered within a given latency in order for the data to be useful. Video-conference of is as an example of DBTS of traffic.

- **DB** (Dedicated Bandwidth): This category includes that type of traffic only requiring a guarantee referring to the minimum bandwidth. DB traffic is not very sensitive to the latency. So, it is not necessary to provide it with any guarantee in this sense. Video visualization from a server is an example of DB traffic.

- **PBE** (Preferential Best-Effort): This type of traffic dose not has any QoS guarantees, but it will have priority over the usual best effort traffic.

- **BE** (Best-Effort): This traffic tends to be bursty in nature and is largely insensitive to both bandwidth and latency. BE traffic includes that generated by the transfer of files, email, printing services, etc.

This classification is used to split up the traffic with different priority into the different Service Levels (SL_s). We used this classification to devote the high-priority arbitration table of InfiniBand to the DBTS traffic and the low-priority arbitration table to the remaining categories.

2.3 InfiniBand Networks Support for QoS

InfiniBand has been traditionally used in HPC and clustering applications. The provisioning of QoS in data communication networks is currently the centre of much discussion and research in the industry. When discussing QoS in interconnection networks, there are three properties of significant importance: bandwidth, latency, and packet loss. The granularities of the object on which these metrics are applied are single data streams, classes of traffic, or all-network traffic. InfiniBand networks enables QoS support through a rich set of mechanisms to segregate traffic flows into traffic classes and to provide hop level forwarding control over these individual classes. It provides high throughput and low latency for efficient I/O and cluster communication. InfiniBand has three mechanisms to support QoS: Service Levels (SL_s), Virtual Lanes (VL_s), and Virtual Lane Arbitration (VLArbitration) for transmission over links [5].

2.3.1 Service Levels (SL_s)

InfiniBand defines a maximum of 16 Service Levels (SL_s). It depends on the implementation or on the administrator how to distribute the different existing traffic

type among the SL_s. By allowing the traffic to segregate by category, different treatment can be given.

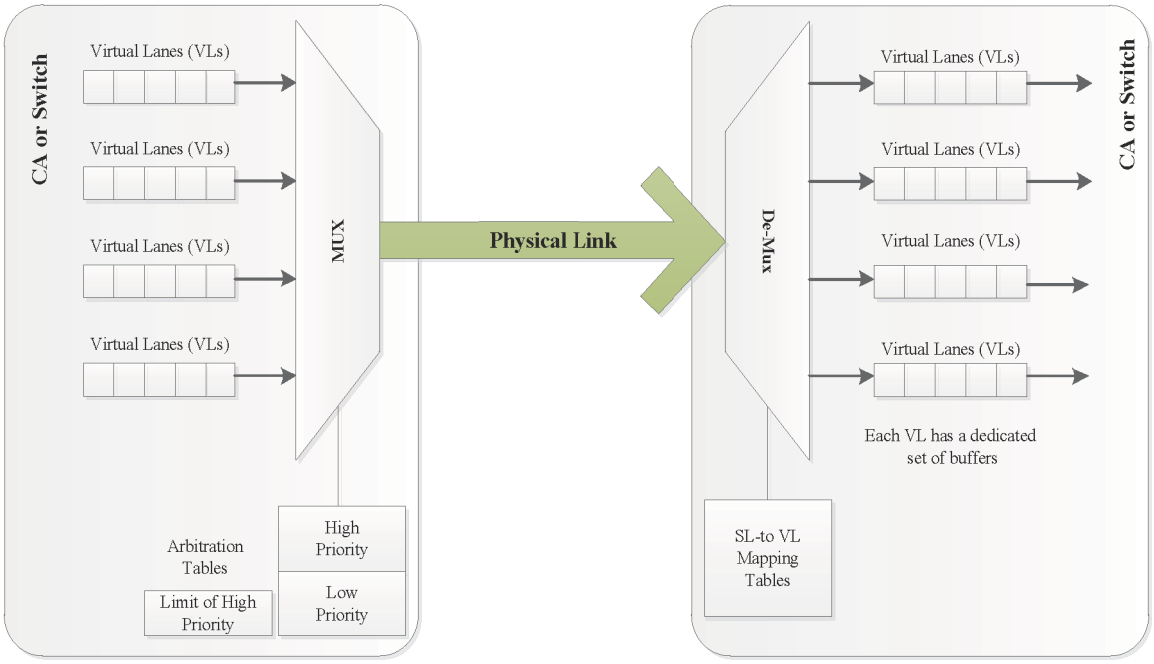


Fig. 2. 2: Operation of Virtual Lanes in a physical link

2.3.2 Virtual Lanes (VL_s)

VLs provide a mechanism for creating multiple virtual links within a single physical link (shown in Fig 2.2) in InfiniBand network ports.

Each VL must be an independent resource for flow control purposes. InfiniBand ports have to support a minimum of 2 and a maximum of 16 VL_s (VL_0 … … VL_{15}). All ports support VL_{15} that is reserved for subnet management, and must always have priority over data traffic in the other VL_s. The subnet manager configures the number of VL_s used by a port. Because systems can be constructed with switches supporting different numbers of VL_s, packets are marked with a SL and each VL must be an independent resource for flow control purposes [9].

2.3.3 Virtual Lane Arbitration (VLArbitration)

When more than two VLs are implemented, an arbitration mechanism is used to allow an output port to select which VL to transmit. VLArbitration defines the priorities of the data lanes. This arbitration is only for data VL_s because VL_{15} transports control traffic and always has priority over other VL_s.

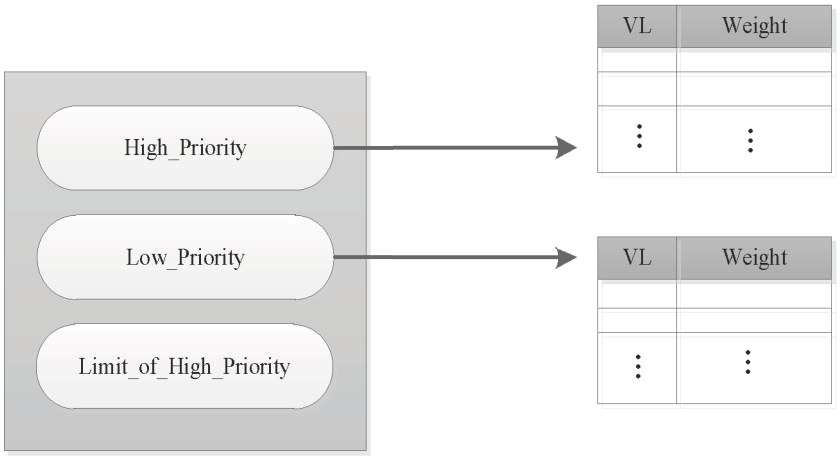


Fig. 2. 3: VLAribitration Table structure

The structure of the VLArbitration is shown in Fig. 2.3. There are two tables in each VLArbitration, one for delivering packets from high-priority VL_s and another one for low-priority VL_s. The arbitration tables implement weighted round-robin arbitration within each priority level. Up to 64 table entries are cycled through and the weight must be in the range of 0 to 255, which is always rounded up as a whole packet. A Limit_of_High_Priority value specifies the maximum number of high-priority packets that can be sent before a low-priority packet is sent [7].

2.4 GE Distribution and ME Formalism

This section describes GE distribution and the solution of GE-Type queueing models approximated by ME analysis.

2.4.1 Generalised Exponential (GE) Distribution

Generalised Exponential (GE) distribution is often used to model bursty and batch arrival network traffic. The measurements of actual inter-arrival or service rate with GE distribution may be generally limited and so only few parameters need to be computed reliably [68].

Let x denote a radom variable represented by a GE distribution with mean arrival rate, $1/\lambda$, and squared coefficient of variation (SCV) of x, C^2. The probability density function (pdf) $f(t)$ is [69]:

$$f(t) = (1-\tau)\lambda_0(t) + \tau^2 \lambda e^{-\tau\lambda t}, t \geq 0 \qquad (2.1)$$

where $\tau = \dfrac{2}{C^2+1}$ and $\lambda(t) = \begin{cases} \infty, & \text{if t} = 0 \\ 0, & \text{if t} \neq 0 \end{cases}$.

In this context, the probability distribution function (PDF), is given by

$$F(t) = P(x \leq t) = 1 - \tau\, e^{-\tau\lambda t}, \qquad t \geq 0, \qquad (2.2)$$

For $C^2 > 1$, the GE distribution can be interpreted as a bulk-type distribution. Moreover, the underlying counting process of the GE distribution is a Compound Poisson Process (CPP) that has geometrically distributed batch sizes with mean batch inter-arrival time $1/\partial$ with $\sigma = \tau\lambda = \dfrac{2\lambda}{C^2+1}$.

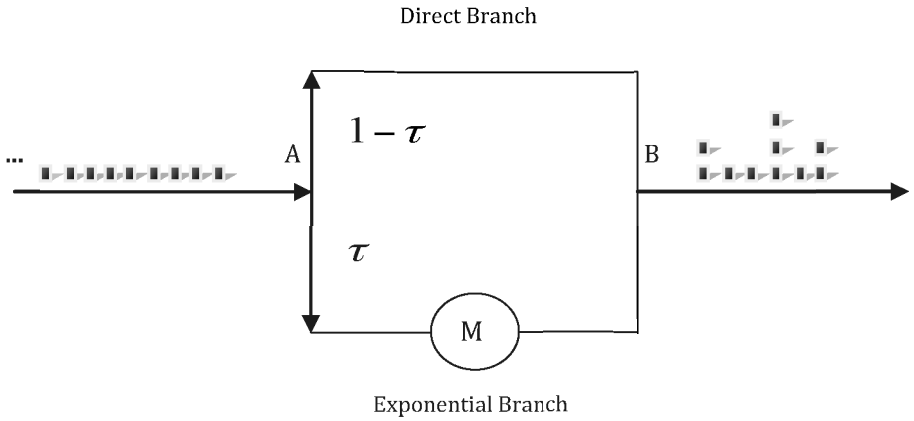


Fig. 2. 4: The GE distribution with parameters λ and C^2

The arrival process of the random variable with a GE distribution is depicted in Fig. 2.4. Assume that there are always packets waiting at Point A to enter into the system. Each packet has to select either the up-branch (i.e., direct branch) or the down-branch (i.e., exponential branch) to reach the departing Point B. The selection criterion per packet constitutes a Bernoulli trial process. Specifically, a packet chooses the down-branch with probability, τ, and then receives an exponential service with mean time $1/\sigma$. The packet can also choose the up-branch with probability, $1-\tau$, and reaches Point B directly without service. The bulk of packets consist of a head which comes from the exponential branch together with a number of successive packets arriving through the direct branch before the appearance of a new packet coming from the exponential branch. The GE distribution is versatile with a simple form of cumulative distribution function and

the pseudo-memoryless properties, making the solution of many queueing systems and networks with GE distribution analytically tractable [69].

The mean and variance are two most important parameters of a random variable. The GE distribution is versatile and possesses a pseudo-memoryless property which makes the solution of many GE-Type queueing systems analytically tractable. The burstiness of the arrival process is characterized by the SCV of the inter-arrival time in the GE distribution. Moreover, the number of packet arrivals at exponentially distributed intervals is in geometrically sized batches; this can model traffic burstiness. The GE distribution of processing times describes a similar batched behavior for service completions. In this thesis, GE-Type distribution will be used to model the bursty traffic in the InfiniBand networks. Analytic expressions for the utilization, mean queue length and blocking probability are determined [72].

2.4.2 ME Analysis for Queueing System

The principle of maximum entropy (ME) provides a self-consistent method of inference for characterizing, under general conditions, an unknown but true probability distribution, subject to known (or, known to exist) mean value constraints [121].

Since the 1960s, classical queueing theory cannot easily handle by itself complex QNMs with many interacting elements. As a consequence, alternative tools had been developed in the queueing theory, such as ME, which provides an analytical solution approximate to that of queueing system and networks with

stochastic and operational analysis.

In a series of publications, ME is proposed as a new analytic framework which can be used to derive minimally biased approximations of performance distributions for queueing networks, subject to mean value constraints which depend on the mean rate and SCV of the inter-arrival and service time distribution. In [67], analysis of queueing system, the subject to constraints.

a) Normalisation

$$\sum_{n=0}^{N} P_N(n) = 1 \tag{2.3}$$

b) Utilization, $U, 0 < U < 1$

$$\sum_{n=0}^{N} h(n) P_N(n) = 1 \tag{2.4}$$

where $h(n) = \begin{cases} 0, & \text{n} = 0 \\ 1, & \text{n} \neq 0 \end{cases}$

c) The Mean Queue Length

$$\sum_{n=1}^{\infty} n P_N(n) = \overline{L} \tag{2.5}$$

d) The full buffer state provability $P_N = \phi, 0 < \phi < 1$, written as

$$\sum_{n=0}^{N} f(n) P_N(n) = \phi \tag{2.6}$$

where $f(n) = \begin{cases} 0, & \text{n} < \text{N} \\ 1, & \text{n} = \text{N} \end{cases}$ and satisfying the flow balance condition

$$\lambda(1-\tau) = \mu U .$$

A general ME solution $P(n) = n - 1, 2, ...,$ N that maximises the system's entropy function

$$H(P) = -\sum_{n=0}^{\infty} P(n)\ln(P(n)) \tag{2.7}$$

is given by $P(n) = \frac{1}{Z} g^{h(n)} x^n y^{f(n)}$

where $Z = 1/P(0)$.

The Largrangian coefficients x, g, and y are given by

$$x = \frac{\bar{L} - \rho}{\bar{L}} \tag{2.8}$$

$$g = \frac{\rho(1-x)}{x(1-\rho)} \tag{2.9}$$

$$y = \frac{1-\rho}{1-x} \tag{2.10}$$

where $\rho = \lambda/\mu$, and the mean queue length distribution can be seen in (2.11), namely

$$\bar{L} = \frac{\rho}{2}\left\{1 + \frac{1+\rho C_s^2}{1-\rho}\right\} \tag{2.11}$$

The ME solution can be expressed in terms of a normalizing constant and a product of Lagrangian coefficients corresponding to the mean value constraints. In an information theoretic context, the ME solution corresponds to the maximum

disorder of system states and, thus, is considered to be the least biased distribution estimate of all solutions that satisfy the system's constraints.

2.5 Summary

In this chapter, we introduced the overview of Infiniband networks architecture and explained the traffic classification for the different traffic flows. These unique characteristics make QoS in IBA unique. Then GE distribution and Maximum Entropy (ME) formalism are introduced. The GE-Type queueing system can be modelled by the ME solutions, which is an effective methodology for modelling the inter-arrival times and service times of the input traffic to analyse the new mechanisms that proposed in chapter 5.

Chapter 3

Rate-Based Flow Control in the InfiniBand Networks

3.1 Introduction

The characteristics of IBA make the flow control mechanism specifically challenging [30, 55, 56, 100, 102]. Firstly, unlike traditional networks, InfiniBand switch cannot drop packets at switches to deal with congestion. Link level flow control [31, 32, 75, 112] is used in InfiniBand switches, which prevents a switch from transmitting a packet when the downstream switch lacks sufficient buffering to receive it. The well-known congestion collapse scenario of traditional networks is avoided by preventing packet dropping at switches but an undesired effect known as congestion spreading may be caused, which will be discussed in detail in Section 3.3.

As a result, packet losses cannot be used in the congestion mechanism for the InfiniBand networks. Secondly, InfiniBand switches are single-chip devices with small packet buffers [104, 120]. A typical InfiniBand switch can hold 4 packets of 2KB per port. So there are only a few packets in transit at any time. Thirdly, the switch and end-device processing latency is very low. This characteristic leads to

small bandwidth-delay and the flow can use all the available bandwidth on its network path [110, 114].

So, in this environment, a traditional window control mechanism as used by Transmission Control Protocol (TCP) is inadequate for controlling flow rates [51, 59]. For end-to-end congestion control in traditional networks, flow sources use packet dropping or changes in network latency as a signal of congestion [32].

In this chapter, we use the Explicit Congestion Notification (ECN) to detect congestion and notify the flow endpoint. ECN has been widely used in ATM networks and the Internet with TCP [22, 51, 90]. We adopt the ECN mechanism for InifniBand networks because the configuration of the switches with the input buffer is the same as that with output buffer in ATM networks [43]. A source adjusts its packet injection rate after receiving congestion information. We propose a new source response function to improve the performance of InfiniBand.

The main contributions of this chapter are summarized as follows:

- An admission control solution suited to InfiniBand networks: we extend a traditional admission control mechanism to suit the InfiniBand network architecture. Considering the different connections with various SLs in the networks, this mechanism can provide the multiple-class connections with different priorities.

- A new rate-based flow control source response function: we propose a new and effective source response function that is able to provide the higher bandwidth utilization and ensure the fairness.

The rest of this chapter is organized as follows: Section 3.2 presents an improved link-by-link admission control mechanism. Section 3.3 describes the congestion control and the source response function methodology. We then propose a new response function, referred to as Power Increase and Power Decrease (PIPD). Section 3.4 motivates the need for congestion control in the InfiniBand networks by showing the congestion spreading in a simple scenario, and uses simulation experiments to investigate the performance of the improved admission control mechanism and the new response function subject to multiple-class traffic with different priorities. Section 3.5 summarizes this chapter.

3.2 Admission Control

An admission control [23, 36, 63] algorithm determines whether a new connection request should be accepted or rejected. To conduct admission control in InfiniBand networks, traffic source sends a probe packet to the destination before starting data transmission. Then the admission control check is performed at the corresponding HCA which connects the processor node and the switch. If accepted, the solicited bandwidth is added to the total currently used bandwidth of the physical link; then the probe packet is forwarded to the connected switch/router. If rejected, an information packet is sent back to the traffic source without changing the used

bandwidth of the physical link. When the data transmission is completed, the source inserts the information of the amount of the bandwidth used to setup this connection into the header of the final data packet. The HCA and switches/routers then release the reserved resources after this packet passes them [50, 82].

Admission control in packet-switched networks has been a rich research area. There are two broad categories of admission control algorithms: deterministic and statistical admission control [125]. For real-time services that need a hard or absolute bound on packet delay, a deterministic admission is used [38]. For such deterministic services, the admission control algorithm calculates the worst-case behaviour of the existing flows in addition to the incoming flow before deciding whether the new flow should be admitted. Many of the new applications, e.g., media streams do not need hard performance guarantees and can tolerate a small violation in performance bounds. A statistical admission control scheme can be used for such applications [10, 47, 57, 105].

In this section we present an improved link-by-link admission control scheme that is suitable for the InfiniBand architecture. With the link-by-link approach a bandwidth broker records the load of every link and consults the availability of bandwidth before accepting or rejecting a new connection requirement. We adopt the simple sum approach described in [53] which ensures the sum of requested resources does not exceed the link capacity. Let bw represent the total link bandwidth, s the sum of bandwidth of the already admitted connections, and p the bandwidth requested by the new connection. This algorithm accepts the new

connection under the condition $p + s < bw$. In the InfiniBand architecture, connections with different SLs have different bandwidth requirements. Let s_{sli} be the sum of bandwidth of admitted connections with the SL_{sli}. The connection is accepted if $p + s_{sli} < bw_{sli}$, where bw_{sli} is the effective bandwidth available to SL_{sli}. The InfiniBand architecture defines a maximum of 16 SLs.

3.3 Power Increase Power Decrease (PIPD)

This section describes the congestion control mechanism for InfiniBand using an ECN packet marking scheme and a new source response function. A switch detects and identifies packets, which are contributing to congestion. A single bit ECN field in the header of an identified packet is set by switch to indicate the occurrence of congestion to the destination. ECN value is returned by the destination in the acknowledgment packet and the source uses this information to adjust its packet injection rate.

3.3.1 Packet Marking

A full buffer propagates congestion since it blocks an upstream switch from transmitting packets. Therefore, a straightforward way to detect and indicate the occurrence of congestion would be to mark the packets in a buffer whenever it becomes full. Due to the small buffers currently adopted in SAN switches, using a lower buffer occupancy threshold for marking packets is likely to reduce the overall link utilization as this causes the buffer to become empty more frequently. If the

buffers become larger, using a threshold close to the maximum capacity might be beneficial by preventing congestion spreading while preserving high utilization [99]. We use an ECN [40] packet marking mechanism for input-buffered switches, where switches detect incipient congestion and notify flow endpoints, for example, by marking packets when the occupancy of a switch buffer exceeds a desired operating point. There is a single bit ECN field in the header of an identified packet which indicates the occurrence of congestion [89]. The destination returns an acknowledgment packet (ACK) which includes the ECN value and the source will use this information to control the packet injection rate. We used the packet marking mechanism proposed in [13, 109].

In this mechanism, two counters are needed for each output link. The first counter cnt_1 records the current number of packets in the switch that are waiting for the output link. The second counter cnt_2 records the number of subsequent packets that need to be marked when being transmitted on the output ink. Counter cnt_2 is initialized to zero. Whenever the buffer becomes full, the value of counter cnt_1 is copied to counter cnt_2. Then, the output port starts marking the next transmitted packets, decrementing cnt_2 at each transmission until it reaches zero again.

3.3.2 Source Response Function

The flow rate is adjusted in response to network congestion feedback. Because the feedback is received through the ACK packet, the flow injection rate will be

adjusted whenever an ACK packet is received. If the receipt is an unmarked ACK, the source response must increase the injection rate. On the other hand, if the receipt is a marked ACK, the source response must decrease the injection rate.

In the proposed algorithm, we use increasing function $f_{inc}(r)$ and decreasing function $f_{dec}(r)$ to control the source injection rate r.

In order to design $f_{inc}(r)$ and $f_{dec}(r)$, we use the conditions defined in [108] as follows:

Condition 1. Avoiding Congested State

$$f_{inc}(f_{dec}(r)) \le r \tag{3.1}$$

The flows experience the same (or higher) degree of congestion after recovery.

Condition 2. Fairness

$$T_{rec}(r_1) \le T_{rec}(r_2) \text{ for } r_1 \le r_2 \tag{3.2}$$

If the recovery time $T_{rec}(r)$ for lower rate flows does not exceed that of higher flows, the fairness is guaranteed.

Condition 3. Efficiency

$$T_{rec}(r_1) = \frac{1}{R_{\min}} \quad \text{for} \quad f_{dec}^{-1}(R_{\min}) \le r \le R_{\max} \tag{3.3}$$

In order to ensure the source response function to be efficient, the flows should recover the injection rate quickly to maximize the bandwidth utilization.

In [108], we also find some conclusions based on these conditions

$$F_{inc}(t) = f_{dec}(F_{inc}(t + T_{rec})) \quad (3.4)$$

After an adjustment to rate r, the next ACK is received in a time interval $1/r$. Thus, we have

$$f_{inc}(r) = \min(F_{inc}^{r}(1/r), R_{\max}) \quad (3.5)$$

In summary, to obtain an increase function $f_{inc}(r)$ we need to find a function $F_{inc}(t)$ that satisfies Equation (3.4). In the next section, we show how to obtain $f_{inc}(r)$ for a specific response function.

3.3.3 PIPD Function

Based on the unique characteristics, the source response function of InfiniBand networks should be different from those used in other high-speed networks, such as the Additive Increase Multiplicative Decrease (AIMD) [29], which has been shown to converge to fairness under an assumption of synchronized feedback to flow sources. AIMD has been used for both window control and rate control [58]. It is known that the rate increase using the traditional functions is linear [14, 41]. Therefore, the injection rate cannot reach a high speed quickly. As a result, the majority of network bandwidth is under-utilized for a long period. To overcome this problem, we consider using the power increase of rate with a new function.

To design the increase function, we need to define a decrease function firstly, which uses the power function.

$$f_{dec}^{pipd}(r) = \max(r^{1/m}, R_{\min}) \quad (3.6)$$

where $m > 1$ is constant.

In order to avoid congested state from Equation (3.1), $F_{inc}(t)$ must satisfy the following condition

$$F_{inc}(t + T_{rec}) = (F_{inc}(t))^m \tag{3.7}$$

This equation shows that after each time interval T_{rec} the function is powered by m. From this equation, we can find a solution of the required increase function by setting $F_{inc}(0) = R_{\min}$ and $T_{rec} = 1 \times T_{rec}, 2 \times T_{rec}, 3 \times T_{rec} \ldots$. The new increase function can be given as follows

$$F_{inc}(t) = R_{\min}{}^{m^{t/T_{rec}}} \tag{3.8}$$

For any rate r, there exists a t' for which $r = F_{inc}(t') = R_{\min}{}^{m^{(t'/T_{rec})}}$. The notation $F_{inc}^{r}(t)$ represents that the injection rate increases to r at time t. We use $F_{inc}^{r_2}(t) = F_{inc}^{r_1}(t + t')$ for $R_{\min} \le r_1 < r_2 \le R_{\max}$ to represent that the injection rate r_1 increases to r_2 after time t'.

Therefore,

$$F_{inc}^{r}(t) = F_{inc}(t + t') = R_{\min}{}^{m^{(t'/T_{rec} + t/T_{rec})}} = r^{(t/T_{rec})} \tag{3.9}$$

As previously stated, we choose to adjust the rate at the reception of each unmarked ACK. After an adjustment to rate r, the next ACK is nominally received in a time interval $1/r$. Thus we define $f_{inc}^{pipd}(r) = \min(F_{inc}^{r}(1/r), R_{\max})$

In order to get the increase function $f_{inc}(r)$ from Equation (3.4), (3.6), (3.8), and (3.9)

$$f_{inc}^{pipd}(r) = \min(F_{inc}^{r}(1/r), R_{\max})$$

$$= \min(r^{R_{\min}^{(1/r*T_{rec})}}, R_{\max})$$

$$= \min(r^{m^{R_{\min}/r}}, R_{\max}) \qquad (3.10)$$

From the increase function $f_{inc}(r)$ of Power Increase and PIPD, we can see that the change of the injection rate is based on the value of $R_{\min}/r$. Therefore, the PIPD mechanism can classify different traffic priorities by giving different values of $R_{\min}$. The increase function of PIPD increases the injection rate by power. Unlike the traditional linear increase functions, the injection rate can increase exponentially and reach to a high value in a short time.

Next we will analyze how the PIPD function modifies the transmission rate in InfiniBand networks. At the beginning, injection rate r is set to $R_{\min}$ and the value of $R_{\min}/r$ is equal to 1. The injection rate will increase very quickly because it is powered by m. This can improve the utilization of the bandwidth. Then the injection rate becomes bigger than $R_{\min}$ and the value of $R_{\min}/r$ is less than 1, so the injection rate will not increase as fast as that at the beginning. The purpose of this modification is to avoid congestion condition in the networks. From our analysis, the increase function of PIPD will not only control the rate well to suit the InfiniBand networks but also can improve the bandwidth utilization.

The decrease function also suits the InfiniBand network environment. Because of small buffer size and the low network latency, the multiplicative decrease function is not able to deal with the congestion in the InfiniBand networks rapidly. The power decrease can reduce the injection rate quickly enough to make the traffic flows relieve the congestion condition.

3.4 Simulation Experiment and Performance Results

Firstly we conduct a series of simulation experiments based on the scenario that is shown in Fig. 3.1 and Table 3.1 shows the parameters used in our simulation. in order to understand the problem of congestion spreading and the relationship between a root source and a victim of congestion,

The topology structure of our simulation includes two switches: switch A and switch B, connected by a single physical link. The traffic flow generated at endpoint B_1 and destined to endpoint B_c. For clarity, let us call it as local flow because it is connected to the same switch of the receiver. The traffic flow generated at endpoint A_1 and destined to endpoint B_c; we call it remote flow because its packets need to be forwarded by switch A to switch B through an inter switch link. A victim flow generated at endpoint A_v and destined to endpoint B_v through the inter-switch link between the two switches. Victim flow is destined to a non-congested endpoint B_v and suffers from congestion spreading [108].

Assume the local flow is sending packets from B_1 to B_c so that endpoint B_c can receive at 100% of its capability. At the same time the remote flow is also

transmitting packets from A_1 to B_c through the inter-switch link. Because all the flows are greedy, they try to use network bandwidth as much as they can. At this point congestion spreading originates at the oversubscribed link connecting switch B to endpoint B_c.

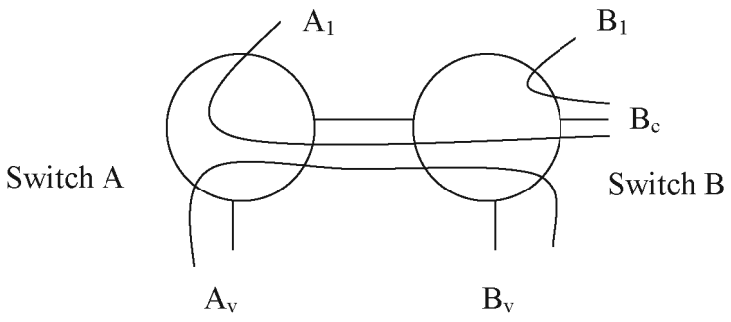


Fig. 3. 1: Simulation Scenario

For example, endpoint B_c would be the "root" congestion if it were receiving packets faster than it is able to forward them. The congestion in switch B has spread so that switch A is now congested. If the congestion spreading occurred, the approach is to limit the injection rate of flows, which are the root cause of the congestion (local flows and the remote flows), and the other flows are not affected (victim flows).

Table 3. 1: Simulation Parameters

Parameters	Value
Link bandwidth	1GB/sec (InfiniBand 4* link)
Packet header	20 bytes (InfiniBand Local Header)
Data packet size	20 + 2048 = 2068 bytes
Packet transmission time	2.068 μs
Buffer size	8 packets

Then, we have conducted a series of simulation experiments based on the scenario illustrated in Fig. 3.1 to evaluate the performance of the improved link-by-link admission control mechanism and the new source resource function PIPD with the simulation tool OPNET.

OPNET Modeler provides support to model communication networks and distributed systems. We use this tool to design and analyze communication equipment and protocols of Infiniband networks.

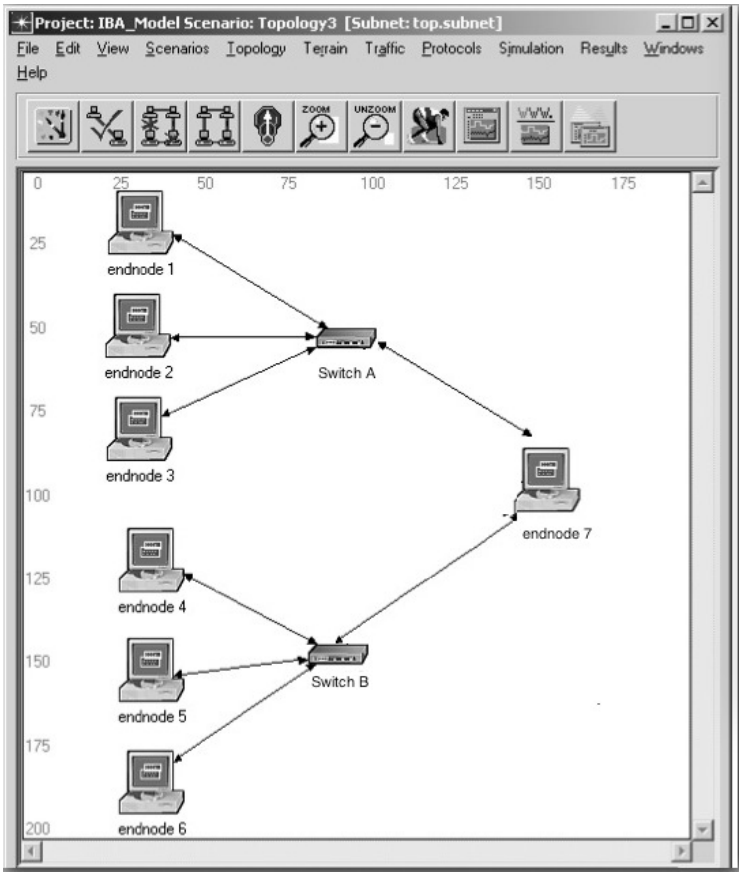


Fig. 3. 2: OPNET Project editor

OPNET defines three work domains: network, node, and process domain. The node level specifies objects belonging to the network level, and the process level specifies objects belonging to the node level. Network models consist of nodes, links,

and subnets. Nodes represent network devices. Links represent point-to-point and bus connections. Subnets organize network components into a single object. Network models are developed using the project editor. This editor allows rapid construction and test of various possible network configurations.

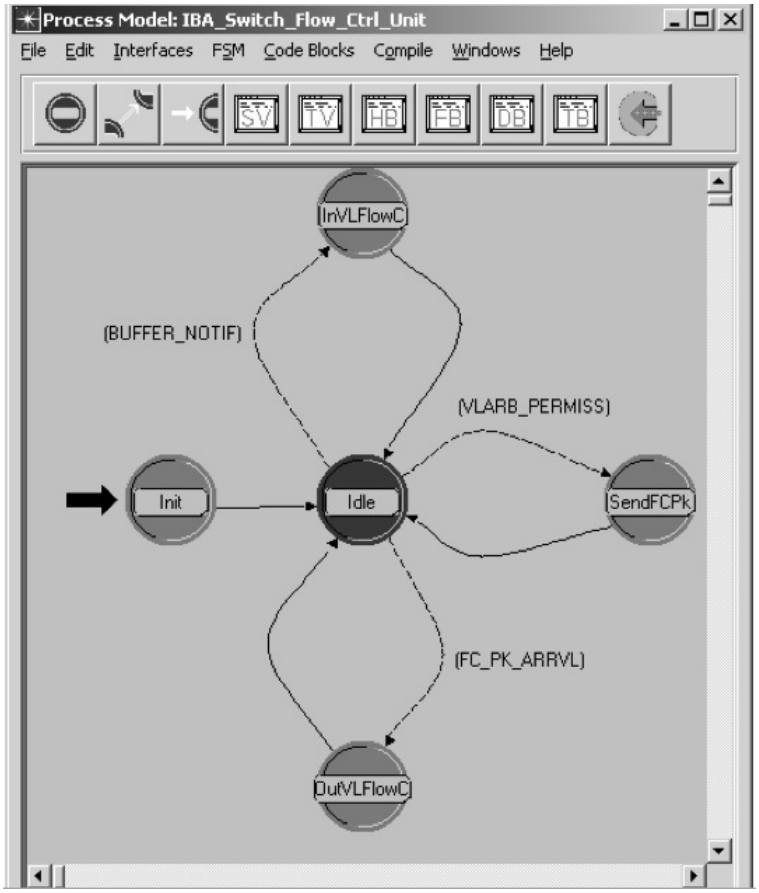


Fig. 3. 3: OPNET Processor editor

In Fig 3.2, we used the OPNET project editor to build the simulation scenario. Node models are built using basic OPNET modules and specifying the connections between them. he behavior of processors and queues are programmable via their process models. They consist of finite state machines containing blocks of C/C++ code and kernel procedures. State machines respond to interrupts generated by the OPNET simulation kernel. The user code is executed when the machine enters or

leaves a state, and it can also be associated to a transition. Kernel procedures provide the way to perform common tasks, such as to manipulate packets, collect statistics, operate with queues, or program future interrupts. Fig. 3.3 shows our process model

The topology structure of our simulation includes two switches: switch A and switch B, which are connected by a single physical link. We refer to the traffic flow generated at endpoint B_1 and destined to endpoint B_c as local flow because it is connected to the same switch. The traffic flow generated at endpoint A_1 and destined to endpoint B_c is referred to as remote flow because the packets need to be forwarded from switch A to switch B through an inter switch link. A victim flow is generated at endpoint A_v and destined to endpoint B_v through the inter-switch link between the two switches. The victim flow is destined to a non-congested endpoint B_v and suffers from congestion spreading [108]. Table. 3.1 lists the parameters used in the simulation including link bandwidth, packet header size, data packet size, packet transmission time, and buffer size.

We considered two different classes of traffic flows injected into each link with different SLs priorities. There are one remote flow and one local flow in the network. In the admission control stage, we set the bandwidth required by the high priority SL connection as two times of that required by the low priority SL connection. In the congestion stage, we set $R_{\min-high} = R_{\max}/128$ for the high priority flow r_2 and $R_{\min-low} = R_{\max}/256$ for the low priority flow r_1 with $m = 4$. Fig. 3.2 and Fig. 3.3 show how the flow injection rates oscillate in the two switches using our admission control mechanism and the PIPD function.

From these two figures we can notice that the injection rate of the high priority SL connection is almost two times of that of the low priority SL connection. The comparison of these two figures reveals that the utilization of the root link is better than the utilization of the inter-switch link. Because of the congestion spreading, switch A will be blocked if switch B is under the congestion condition. We can also find that the flow with the high priority gets the higher transmission rate than the flow with the lower priority. Table. 3.2 shows the average injection rate of each flow based on the proposed PIPD function. We can see that the high priority flow r_2 easily achieves the very high injection rate.

In order to compare the performance of the PIPD function to the existing Fast Increase Multiplicative Decrease (FIMD) function, we simulate FIMD using the source response functions that were defined in [108]. These functions are given as follows:

$$f_{dec}^{fimd}(r) = \max(\frac{r}{n}, R_{\min}) \tag{3.11}$$

$$f_{inc}^{fimd}(f) = \min(r * n^{R_{\min}/r}, R_{\max}) \quad \text{where} \quad n > 1 \,\text{constant} \tag{3.12}$$

We set factor $n = 4$ and the minimum rate $R_{\min} = R_{\max}/256$. The high priority flow r_2 uses the PIPD function and the low priority flow r_1 uses the FIMD function. The oscillation of injection rates in each switch with two functions is depicted in Fig. 3.6 and Fig. 3.7, respectively. Table. 3.3 shows the average injection rate of each flow with two functions. The performance results have demonstrated that the utilization of the root link is better than the utilization of the

inter-switch link. The comparison of rate r_1 using FIMD function (c.f., Fig. 3.6 and Fig. 3.7 and using PIPD function (c.f., Fig. 3.4 and Fig. 3.5) shows that the maximum rate for both functions are very close. But rate r_1 for PIPD function is most likely within the range $[1\times10^5,\ 2\times10^5]$, where rate r_1 for FIMD function is almost a constant value of $R_{\min}$.

Table 3. 2: Average Injection Rate in PIPD

	Average Injection Rate
Switch-A-r1	47234.08637836817
Switch-A-r2	149789.83324444026
Switch-B-r1	76440.10564262095
Switch-B-r2	259756.99827139667

Table 3. 3: Average Injection Rate in PIPD and FIMD

	Average Injection Rate
Switch-A-r1	18537.107824746185
Switch-A-r2	118861.56808060875
Switch-B-r1	20147.849538670973
Switch-B-r2	194463.99403664135

More specifically, Fig. 3.7 depicts the difference of the average injection rate of two flows in Switches A and B obtained from Table 3.2 and Table 3.3. It is clear that the PIPD function achieves the higher transmission rate. The simulation results show that the proposed PIPD function can improve the performance of congestion control in InfiniBand networks.

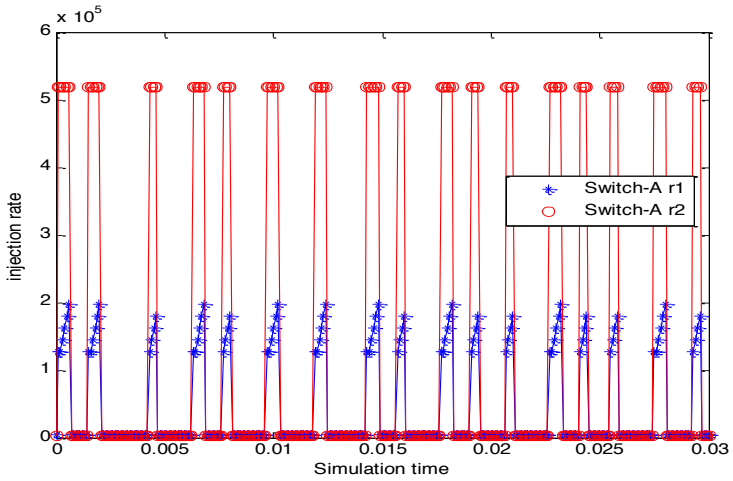


Fig. 3. 4: Result of Switch-A with PIPD

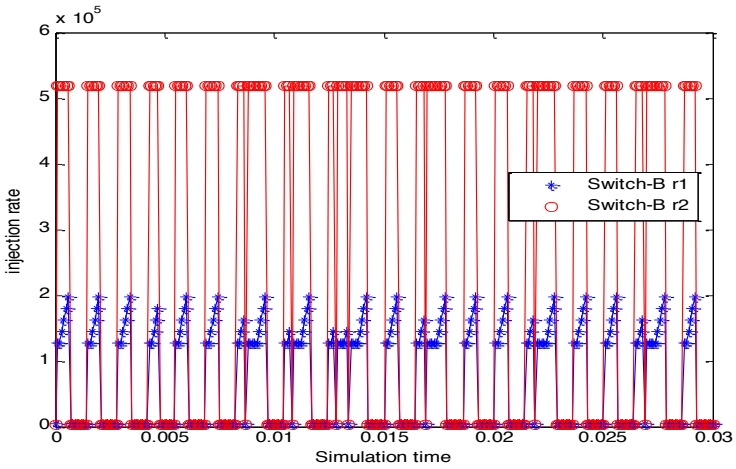


Fig. 3. 5: Result of Switch-B with PIPD

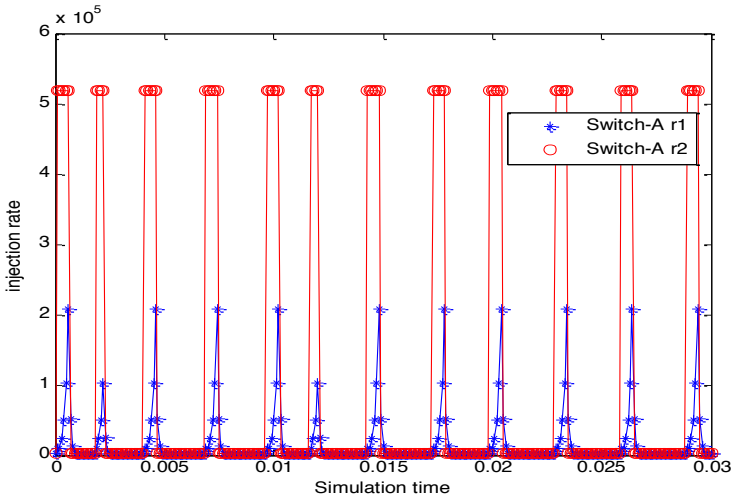


Fig. 3. 6: Result of Switch-A with two functions

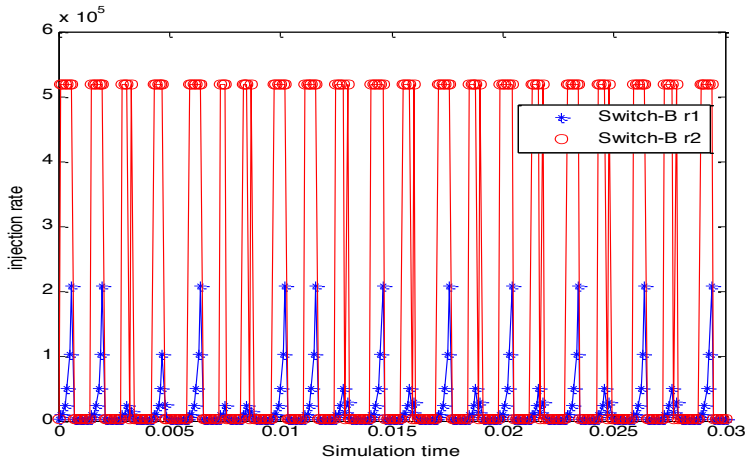


Fig. 3. 7: Result of Switch-B with two functions

3.5 Summary

In this chapter we have proposed an improved admission control mechanism and a new congestion control function for multiple-class traffic in InfiniBand networks. The performance of the admission control algorithm has been enhanced in order to suit the constraints of InfiniBand, to achieve no packet dropping, small buffer size, and low packet latency. The development of the proposed congestion scheme comprises of two parts: a simple ECN packet marking mechanism and a new Power Increase and Power Decrease (PIPD) source response mechanism that combines rate control. The experimental results have demonstrated that the proposed scheme can improve the performance of congestion control in InfiniBand networks, compared to existing schemes.

Chapter 4

Credit-Based Flow Control in the InfiniBand Networks

4.1 Introduction

The goal of flow control in communication networks is to prevent and resolve traffic congestion while ensuring high utilization and fairness among different applications. This task is achieved by controlling both traffic flows that enter the networks and those inside the networks. The design and analysis of efficient flow control schemes for provisioning of desired QoS is particularly difficult and challenging under heavy network loads if the service demands cannot be predicted in advance [33]. This is why congestion control, although only a part of the flow control issue, is most important for of flow control [2, 52].

Many different flow control schemes have been proposed over the past decade [21, 26, 27, 94, 113, 126]. Among these schemes, two well-known classes of flow control are the rate-based scheme and the credit-based scheme [76]. The rate-based flow control is an end-to-end scheme where the source node usually reduces the traffic rate at the injection points of the network to counteract congestion. Periodically, the source broadcasts a packet that makes a round trip to the destination and collects on its way back the minimum rate over all switches/routers

along the link. The source then uses the information carried by this packet to adjust the traffic injection rate [21]. Unlike the rate-based flow control, the credit-based flow control is a link-by-link mechanism as it uses separate window flow control for each link. The downstream switch of every link manages a buffer for each session passing through the link. Whenever a packet is sent downstream out of the buffer, a credit is sent upstream to inform the upstream switch that it can forward one more packet of this session. Every switch that transmits packet out of the session's buffers has to send both data and credits, in a round-robin manner (see Fig. 4.1) [76].

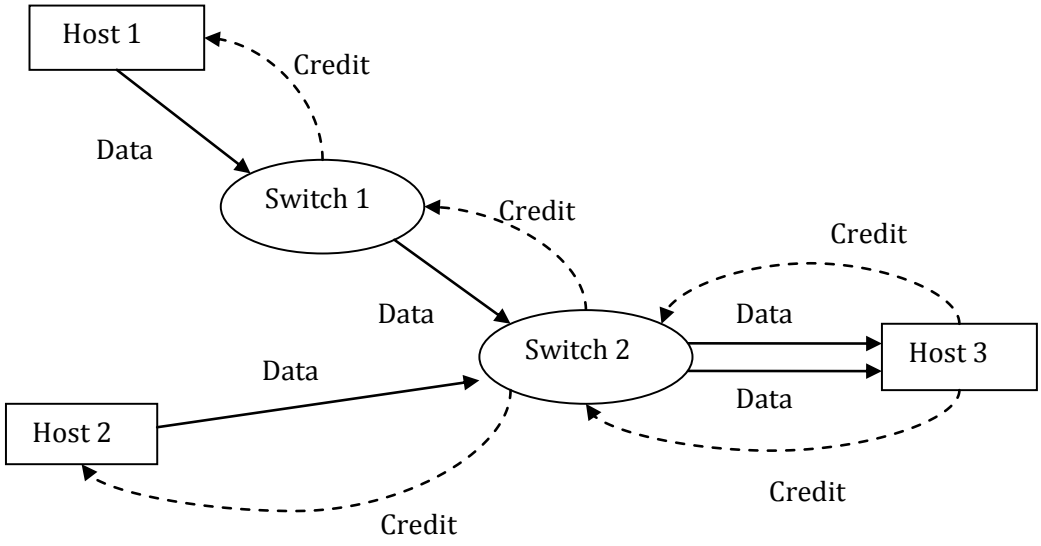


Fig. 4. 1: Credit-based Flow Control

The main advantage of the credit-based scheme [11, 16, 64, 76, 127] is zero packet loss, while the main disadvantage is its complexity of implementation. More specifically, a separate queue for each link is required. On the other side, the main advantage of the rate-based scheme is its simple switch architecture, while its main

disadvantage is the higher packet loss rate assuming that the same amount of buffer space is used in both schemes [25, 86].

Credit-based flow control scheme that can be used to support both end-to-end and link-level flow control is becoming increasingly popular in high speed system area networks (SAN), e.g. InfiniBand networks where multiple processor nodes and I/O devices are interconnected using switched point-to-point links [24, 89, 95, 110]. IBA was designed around a switch-based interconnection technology with high-speed point-to-point links in InfiniBand Trade Association (see Fig. 4.2). The point-to-point interconnection means that every link in the IBA networks has exactly one device connected at each end of the link, thus providing superior performance over the traditional shared-bus architecture [54].

In InfiniBand networks, the switches cannot drop packets to deal with the congestion since packet dropping in such networks extremely complex[35]. Under ideal conditions, the credit-based approach can guarantee zero packet loss regardless of traffic patterns, the number of connections, buffer sizes, numbers of nodes, range of link bandwidths, and range of propagation and queueing delays. Even under extreme overloads, the queue length cannot grow beyond the credits granted [27]. Under extreme overloads, it is possible for queues to grow large resulting in buffer overflow and packet loss.

The credit-based approach requires switches to keep a separate queue for each Virtual Circuit (VC). The congested VCs cannot block other non-congested VCs.

But with the rate-based flow congestion scheme, congested VCs may block other VCs unless random access queue scheme is implemented. The credit-based flow control algorithm is simple and requires no additional hardware, so it has been proposed for link-level and end-to-end flow control in InfiniBand interconnection networks.

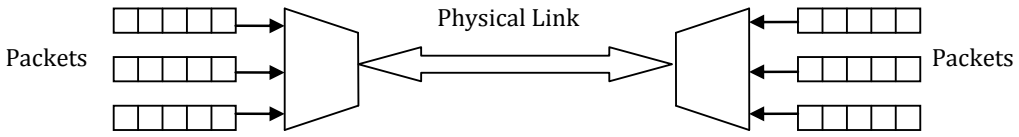


Fig. 4. 2: Multiple VLs in one physical link

With the credit-based flow control management approach, each downstream node in IBA networks forwards the data packet received from its upstream node without packet loss. Packets cannot be sent out until the upstream node is notified that data transfer via the buffer of the downstream node is possible [24, 60].

Credit-based flow control has been widely analyzed in ATM networks over the past decade [76, 87, 119]. Several studies have recently proposed to adopt credit-based flow control mechanisms to enhance the performance of InfiniBand networks. El-Taha and Heath [35] presented an analytical model of credit-based flow control in InfiniBand, but this model is based on the fork-join queue system and requires an extra queue to store the credit packets. In this paper, we develop a new and cost-efficient analytical model to evaluate the performance of credit-based flow control in InfiniBand networks. The important feature of the proposed model is its ability to fit with the characters and requirements of InfiniBand.

The rest of this chapter is organized as follow. In Section 4.2 we present a queuing network model for the credit-based flow control scheme in InfiniBand networks and analyze its performance using such a queuing network model. The validation of the analytical model and performance results are shown in Section 4.3. Finally, Section 4.4 concludes the study.

4.2 Queuing Analysis of Credit-Based Flow Control

The credit-based flow control scheme operates as follows. The upstream node (i.e., Node 1 in Fig. 4.3) does not transmit packets before receiving credits from the downstream node (i.e., Node 2 in Fig. 4.3) that are used to indicate the availability of buffer space at the downstream node.

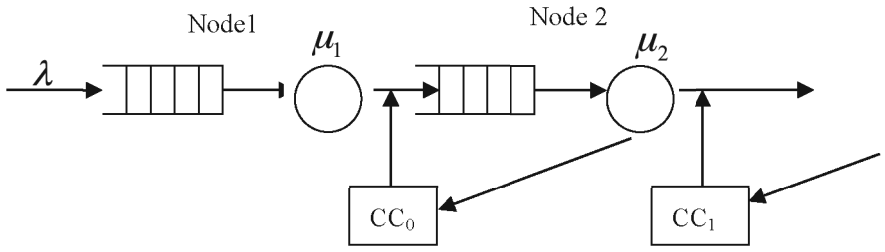


Fig. 4. 3: Queuing Network Model of Credit-Based Flow Control

Upon receiving credits from the downstream node, the upstream node will transmit the same number of packets as the value of credits it holds. The upstream node uses the Credit-Counter (CC) to keep tracking the number of credits it possesses. Following each packet being transmitted, the credit-counter decreases one. The upstream node can continue transmitting packets to the downstream node as long as the credit-counter is greater than zero, but must stop when it becomes zero.

The downstream node sends a credit to the upstream node when a buffer becomes available again [38].

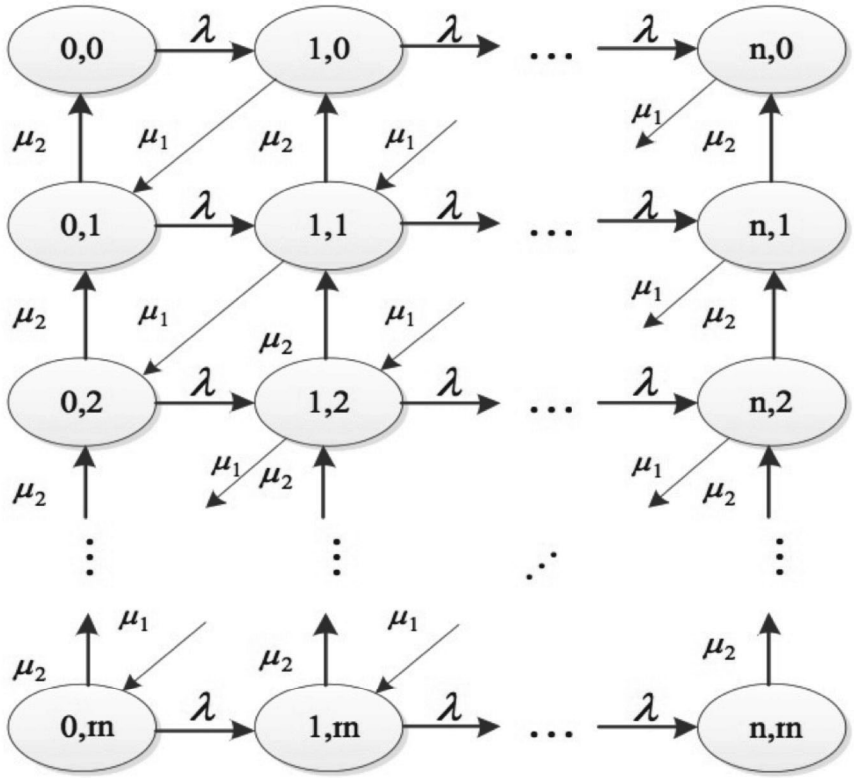


Fig. 4. 4: State Transition Diagram

The structure of the queuing network model of the credit-based flow control in InfiniBand networks is shown in Fig. 4.3. The model consists of two nodes: the upstream node (Node 1) and the downstream node (Node 2). Packets arrive at Node 1 according to a Poisson process with the mean arrival rate λ. The service time at Node 1 and Node 2 follows exponential distributions with mean μ_1 and μ_2.

The state transition diagram of the queuing network model is shown in Fig. 4.4. The queue capacity of Node 1 and Node 2 is denoted by n and m, respectively. The states of the model are represented by the pair (i, j), where i $(0 \le i \le n)$

and j $(0 \le j \le m)$ represent the number of packets (including the packets in the buffer and the packet being processed) in Nodes 1 and 2, respectively.

As packets are injected into Node 1 with rate λ, the average transition rate from (i, j) to $(i+1, j)$, where $(0 \le i < n, 0 \le j \le m)$, is λ. If there are credits available in Node 1, the packet will be forwarded to Node 2 with rate μ_1, so the state transfers from (i, j) to $(i-1, j+1)$ where $(1 \le i < n, 0 \le j \le m-1)$ with rate μ_1. After the packet departs form Node 2, it will send a credit back to Node 1. The state transfers form (i, j) to $(i, j-1)$ with rate μ_2, where $(1 \le i < n, 0 \le j \le m-1)$.

Let $P_{i,j}$ denote the joint probability that there are i packets in Node 1 and j packets in Node 2. Following the principle of the transition equilibrium between in-coming and out-coming flows of each state, we can find the following steady-state balance equations:

$$(\lambda + \mu_1 + \mu_2)P_{i,j} = \lambda P_{i-1,j} + \mu_1 P_{i+1,j-1} + \mu_2 P_{i,j+1}, \quad (0 < i < n, 0 < j < m) \qquad (4.1)$$

We can also find the balance equations below for the boundary sates (i.e., $i = 0, n$ and $j = 0, m$)

$$\begin{cases} \lambda P_{0,0} = \mu_2 P_{0,1} \\ (\lambda + \mu_1) P_{i,0} = \mu_2 P_{i,1} + \lambda P_{i-1,0} \\ (\lambda + \mu_2) P_{0,j} = \mu_1 P_{1,j-1} + \mu_2 P_{0,j+1} \\ \mu_2 P_{m,n} = \lambda P_{m-1,n} \\ (\mu_1 + \mu_2) P_{n,j} = \mu_2 P_{n,j+1} + \lambda P_{n-1,j} \\ (\lambda + \mu_2) P_{i,m} = \lambda P_{i-1,m} + \mu_1 P_{i+1,m-1} \\ \mu_1 P_{m,0} = \mu_2 P_{m,1} + \lambda P_{m-1,0} \\ (\lambda + \mu_2) P_{0,n} = \lambda P_{m+1,n-1} \end{cases} \tag{4.2}$$

In addition, the sum of the probability of all states is 1. So we have

$$\sum_{i=0}^{n} \sum_{j=0}^{m} P_{i,j} = 1. \tag{4.3}$$

We use a matrix analytical method to calculate the joint probability $P_{i,j}$. Based on the above analysis, the element, $G_{(i,j)\to(i',j')}$, of the transition probability matrix, G, of the queuing network model can be given by

$$\begin{cases} G_{(i,j)\to(i+1,j)} = \lambda & 0 \le i \le n\text{-}1 \text{ and } 0 \le j \le m \\ G_{(i,j)\to(i-1,j+1)} = \mu_1 & 1 \le i \le n \text{ and } 0 \le j \le m \\ G_{(i,j+1)\to(i,j)} = \mu_2 & 0 \le i \le n \text{ and } 0 \le j \le m \\ G_{(i,j)\to(i,j)} = -\sum_{i'=0}^{n} \sum_{j'=0}^{m} G_{(i,j)\to(i',j')} & 0 \le i,i' \le n,\ 0 \le j,j' \le m \text{ and } i \ne i',\ j \ne j' \\ G_{(i,j)\to(i',j')} = 0 & \text{otherwise} \end{cases}$$

(4.4)

The steady-state probability vector $P = (P_{i.j})$, $0 \le i \le n$ and $0 \le j \le m$, of the Markov chain satisfies the following equations:

$$PG = 0 \text{ and } Pe = 1. \tag{4.5}$$

Solving these equations yields the steady-state vector as [39]

$$P = B(I - \aleph + eB)^{-1} \tag{4.6}$$

where I is an $(n \times m) \times (n \times m)$ identity matrix, $e = (1,1,...,1)^T$ is a unit column vector of length $(n \times m)$. $\aleph = I + G / \min\{G_{(i,i)}\}$. B is an arbitrary row vector of P.

Mean queue length, mean response time and system throughput are the most important metrics of performance estimation for InfiniBand networks. Let $\bar{L}_1$, $\bar{L}_2$, and $\bar{L}$ denote the mean number of packets in Node 1, Node 2, and in the whole queueing network, respectively. $\bar{L}_1$, $\bar{L}_2$, and $\bar{L}$ can be written as

$$\bar{L}_1 = \sum_{i=0}^{n} \sum_{j=0}^{m} P_{i,j} \times i \tag{4.7}$$

$$\bar{L}_2 = \sum_{i=0}^{n} \sum_{j=0}^{m} P_{i,j} \times j \tag{4.8}$$

$$\bar{L} = \sum_{i=0}^{n} \sum_{j=0}^{m} P_{i,j} \times (i + j) \tag{4.9}$$

The system throughput T is the number of packets passing through the queuing network per time unit and can be given by

$$T = \mu_2 \times (1 - \sum_{i=0}^{n} P_{i,0}) = \mu_1 \times (1 - \sum_{i=0}^{n} P_{i,0}) = \lambda \times (1 - \sum_{j=0}^{m} P_{n,j}) \tag{4.10}$$

The mean response time $\bar{R}$ is defined as the average time elapsed between the arrival of packets at Node 1 and the departure at Node 2. Using Little Law [20], the expression for the mean response time can be derived as follows

$$\bar{R} = \frac{\bar{L}}{T} \tag{4.11}$$

The utility ρ of a link is defined as the ratio of the packet arrival rate λ to the service rate μ. The utility ρ of all links must satisfy $\rho \leq 1$ in order to achieve a steady state. So the rates and service rates of traffic at Nodes 1 and 2 in the queuing network should satisfy $\rho_1 \leq 1, \rho_2 \leq 1$, where $\rho_1 = \lambda / \mu_1, \rho_2 = \mu_1 / \mu_2$.

4.3 Model Validation and Performance Analysis

This section uses the analytical model we have developed to investigate the performance of the credit-based flow control scheme. The traffic parameters in the simulation are $\lambda = 5.0$, $\mu_1 = 6.0$, and $\mu_2 = 8.0$. To evaluate the effect of buffer sizes at the downstream node, Figs. 4.5- 4.8 depict the curves of the mean queue length, response time, and throughput as the buffer capacity at the downstream node changes from 2 to 11 and the buffer capacity at the upstream node is fixed at 12.

We can see from Fig. 4.6 that with the increase of buffer size at the downstream node (Node 2), the mean queue length ($\bar{L}_2$) at this node increases because the buffer can allocate more packets. Figs. 4.6 and 4.8 reveal that both the mean queue length ($\bar{L}_1$) at Node 1 and the system response time decrease as the buffer capacity at Node 2 increases. Moreover, Fig. 4.8 shows that the system throughput increases as well. The phenomena found in Fig. 4.7 and Fig. 4.8 are due to more credits available for the upstream node to transmit packets with the increase of the buffer capacity at the downstream node.

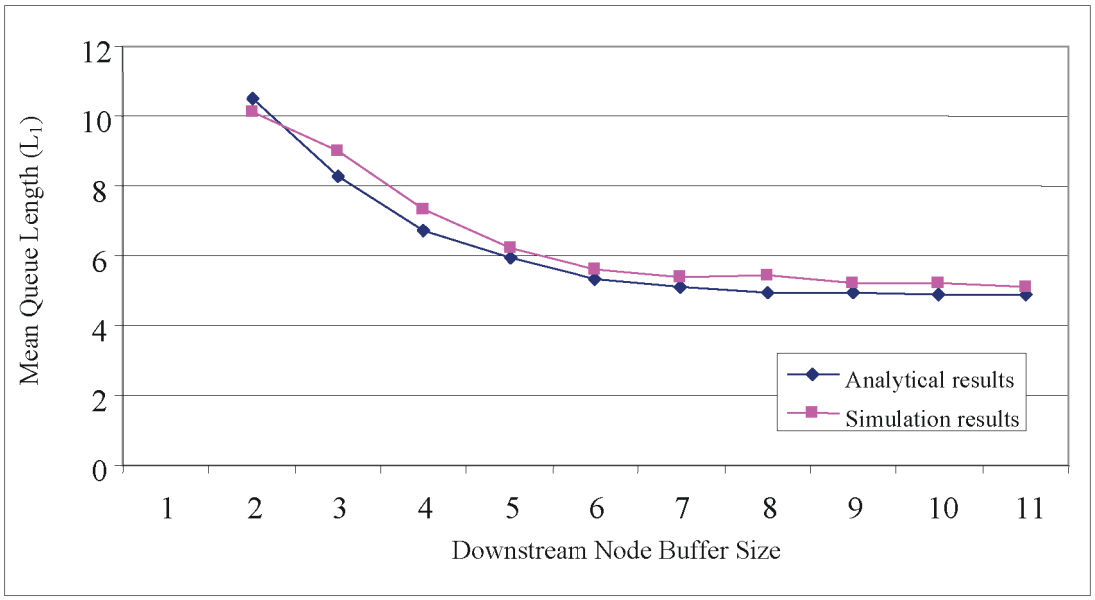


Fig. 4. 5: Mean queue length (L_1) versus the buffer capacity at the downstream node

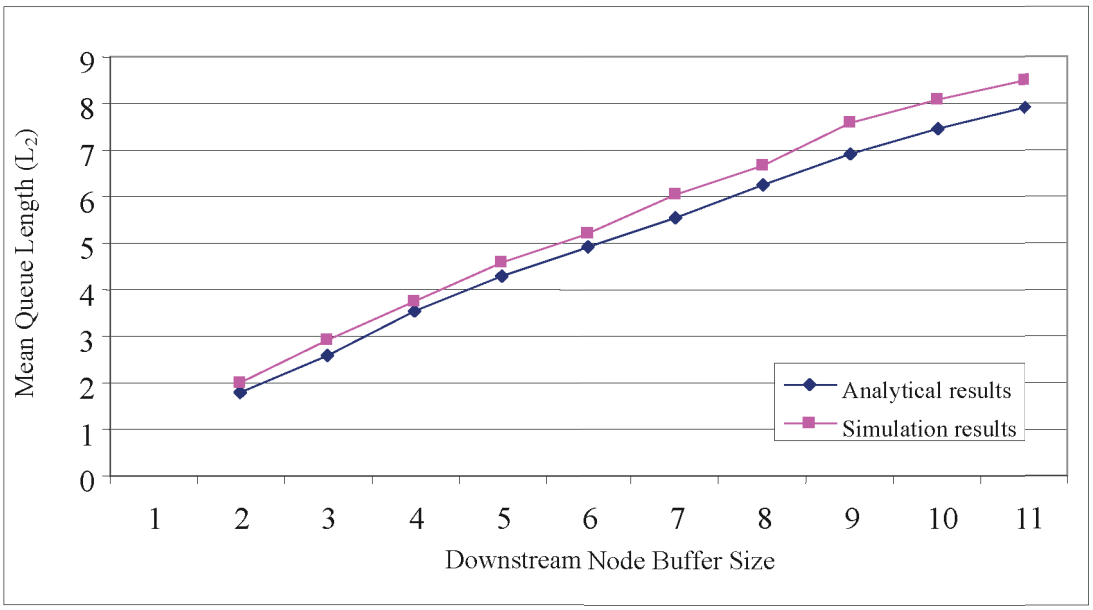


Fig. 4. 6: Mean queue length (L_2) versus the buffer capacity at the downstream node

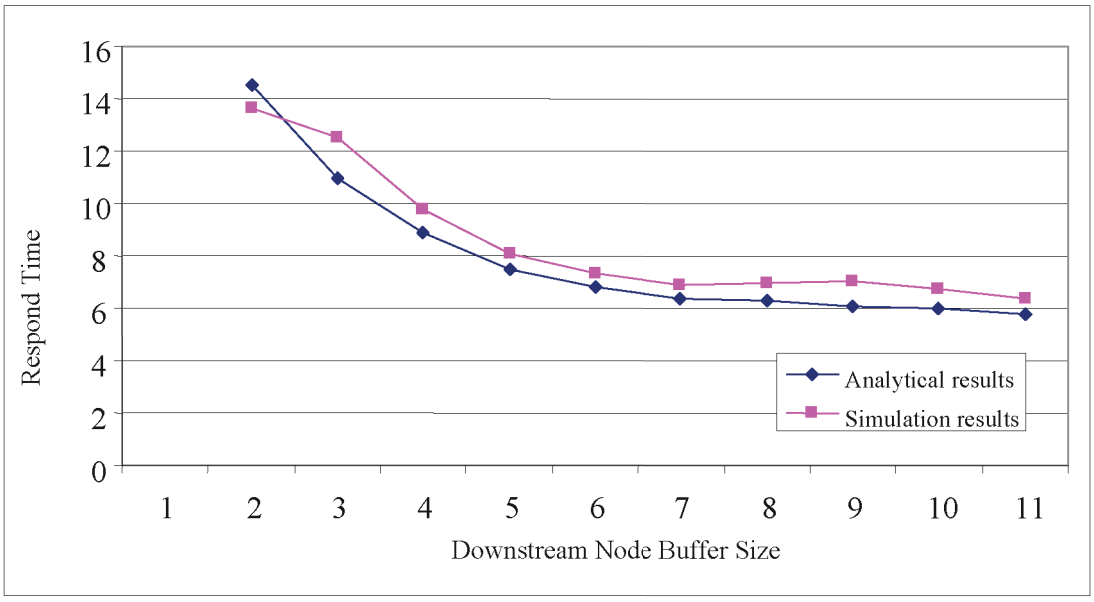


Fig. 4. 7: System response time versus buffer capacity at the downstream node

In order to investigate the impact of the buffer capacity at both the upstream and downstream nodes, Fig. 4.9- Fig. 4.12 reveal the mean queue length, response time, and throughput when the buffer capacities at the downstream and upstream nodes are the same and change from 4 to 10, simultaneously.

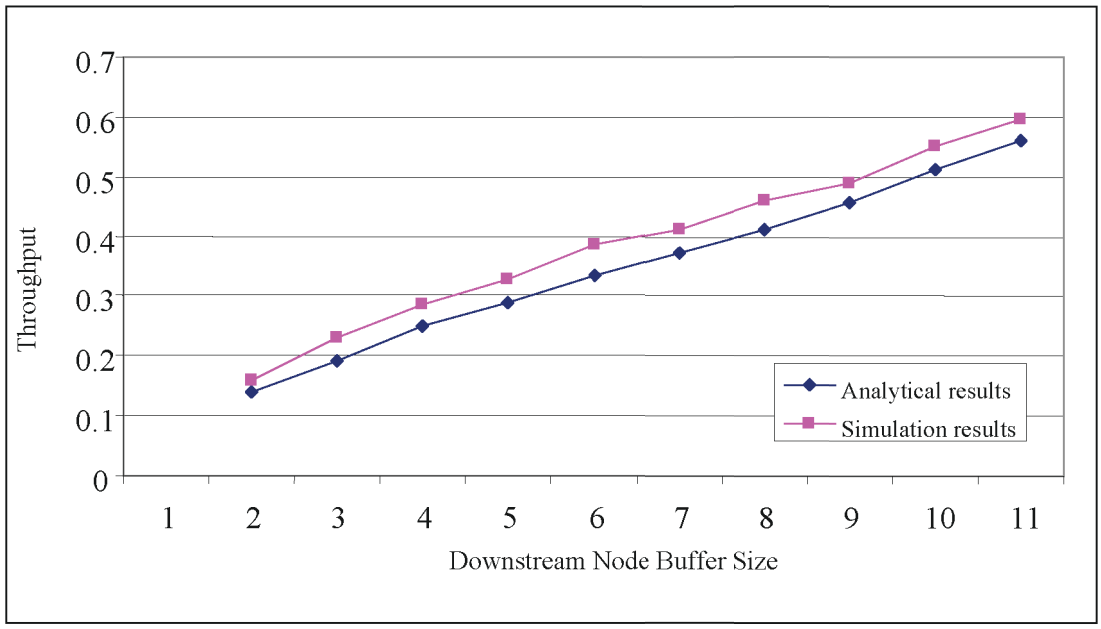


Fig. 4. 8: System throughput (*T*) versus the buffer capacity at the downstream node

As shown in Fig. 4.9, Fig. 4.10, and Fig. 4.12, the mean queue length at Nodes 1 and 2, and the system throughput increase significantly. However, from Fig. 4.11 we can see that the system response time decreases with the increase of buffer capacities. These results have demonstrated that the increase of the buffer capacity at the downstream and upstream nodes can improve the performance of the credit-based flow control in InfiniBand networks. For the purpose to validate the accuracy of the analytical model, performance results obtained from simulation experiments have been illustrated in these figures. We can see that the analytical results match those from simulation very well.

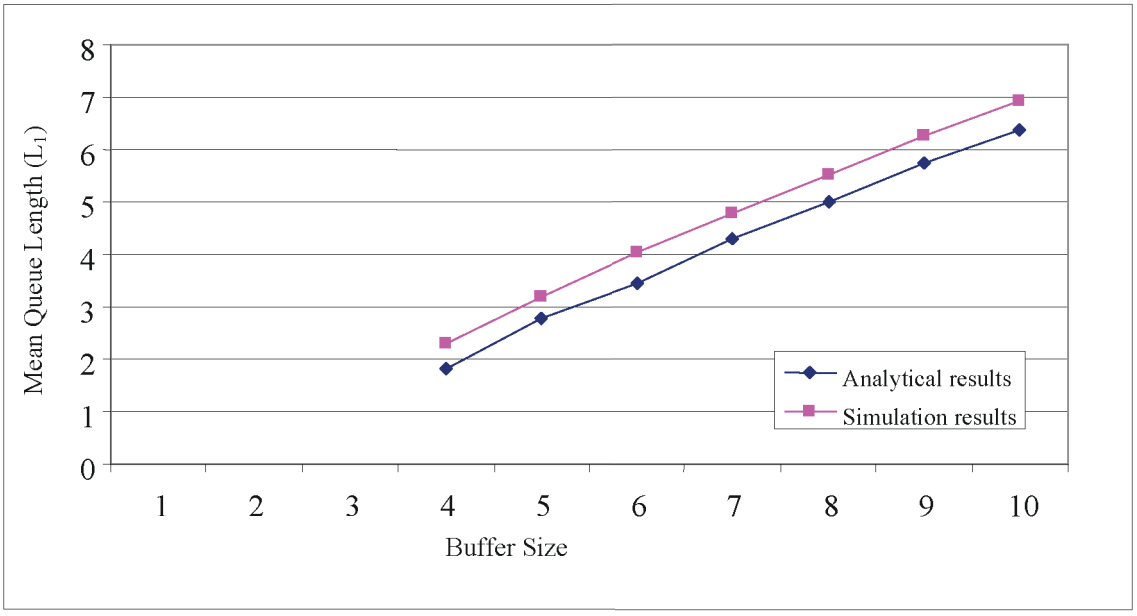
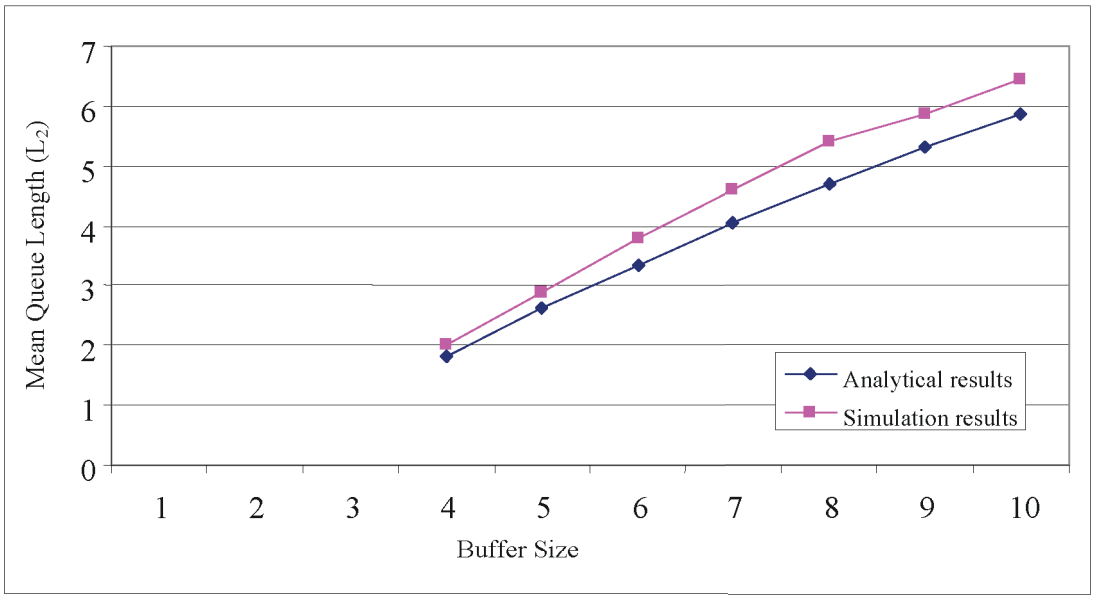


Fig. 4. 9: Mean queue length (L_1) versus the buffer capacity at the upstream and downstream nodes



Mean queue length (L_2) versus the buffer capacity at the upstream and downstream nodes

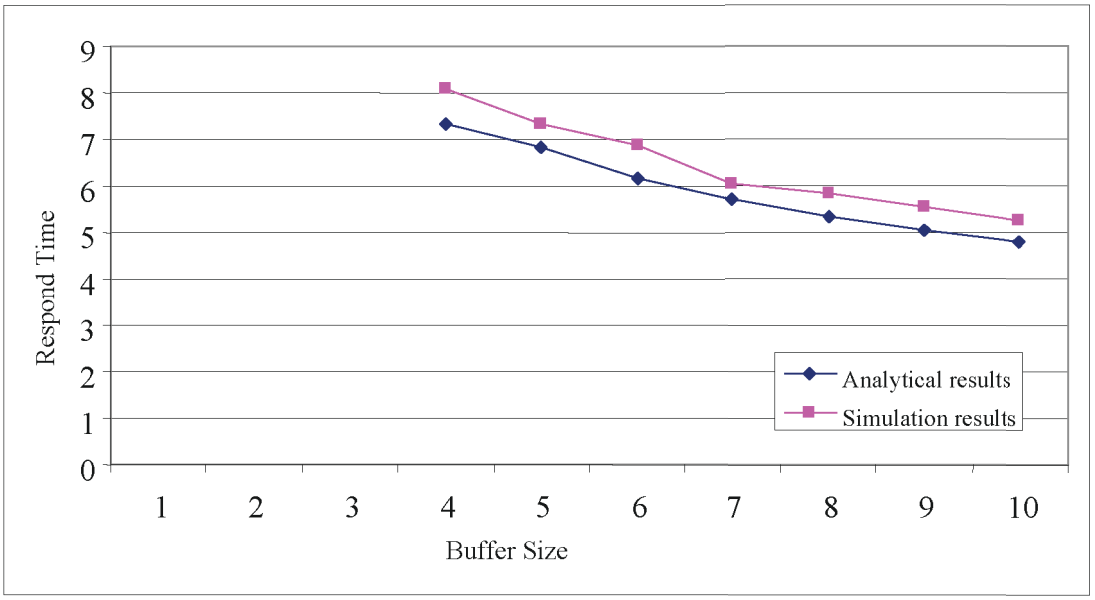


Fig. 4. 10 System response time versus the buffer capacity at the upstream and downstream nodes

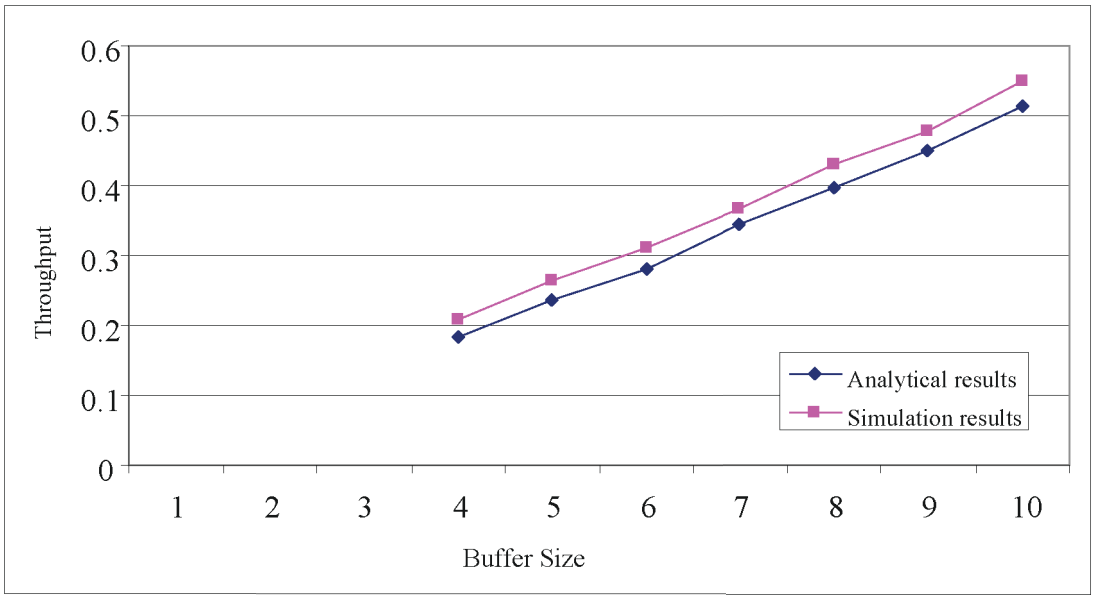


Fig. 4. 11 System throughput (T) versus the buffer capacity at the upstream and downstream nodes.

4.4 Summary

In this chapter, we have developed a queuing network analysis method for performance evaluation of the credit-based flow control in InfiniBand interconnection networks. The model is able to calculate the mean queue length, throughput and response time of the system. Through extensive analysis, we can conclude that the performance of credit-based flow control can be improved with the increase of the downstream node buffer size since the credits available for the upstream node increase. The simulation results have demonstrated that the queue network model is useful and accurate for performance evaluation of the credit-based flow control.

Chapter 5

Maximum Entropy (ME) Analysis for InfiniBand Networks

5.1 Introduction

High Performance Computing (HPC) Clusters have been widely used for solving challenging problems in various application domains ranging from high-end, floating-point intensive scientific and engineering computation problems to commercial data-intensive tasks. Many current networking systems have been dedicated to achieving maximum throughput and minimum latency, leaving aside other aspects like guaranteed bandwidth, bounded delivery deadline, bounded inter-arrival delays, etc. [122] . A large group of the most leading companies, such as Cisco, Dell, Compaq, HP, IBM, Intel, have recently joined together to develop a standard for communication between processing nodes and I/O devices as well as for inter-processor communication in the HPC clusters, known as InfiniBand [48].

IBA is a new industry-standard architecture and is important to enable InfiniBand to support the applications with latency constraints and also those with various QoS requirements. InfiniBand provides a series of mechanisms such as Service Levels (SL_s) Virtual Lanes (VL_s) and Virtual Lane Arbitration tables (VLArbitration Table), which are able to provide satisfied QoS when properly used.

These mechanisms include the segregation of traffic according to categories and the arbitration of the output ports based on the arbitration table. This table is stored in the InfiniBand switches and can be configured to assign high priority to the packets with strict QoS demands. Now InfiniBand is increasingly becoming the choice for HPC systems due to its scalability making more computation power available at a significantly lower price and expanding the number of organizations that can apply parallel computational techniques [6, 24, 107].

In this chapter, we propose two efficient flow control mechanism to be used in the InfiniBand Networks: Entry Threshold (ET) and Arrival Job Threshold (AJT). The idea of ET is to fill the VLArbitration table which is stored in the InfiniBand switches and set the entry threshold in this table for flow control. The system QoS can be improved under this scheme. AJT is based on introducing an arrival job threshold for virtual lanes in the VLArbitration table. These thresholds will effectively control the allocation of bandwidth for various streams with strict QoS constraints. Consequently, the overall system performance will be improved. In these two mechanisms, external traffics are modelled by the generalized exponential (GE) distribution which can capture the bursty property of network traffic. The approximate analytical solution is obtained using the information theoretic principle of Maximum Entropy (ME) which is a simple, reliable and cost-effective tool for network performance prediction and analysis.

The rest of this chapter is organized as follow: Section 5.2 and Section 5.3 present the new efficient flow control mechanisms, including GE analysis and

validation through simulation experiments. Finally, Section 5.4 summarises this study.

5.2 Entry Threshold (ET) in InfiniBand Networks

In the switche of InfiniBand networks, accepting or rejecting the connection requests are decided by the local communication management agents based on the local information available. The information includes the state of the output links and how much bandwidth they have already reserved. When a connection is accepted, the agent modifies the VLArbitration table based on the connection requirements [106].

The information includes the state of the output links and the reserved bandwidth. When a connection is accepted, the agent modifies the VLArbitration Table based on the connection requirements. According to the InfiniBand specifications, there are a maximum of 64 entries in each arbitration table and every entry has a maximum weight of 255. The bandwidth of each connection is proportional to the weight of its associated entry in the VLArbitration Table. The bandwidth of each connection is not simply equal to the particular value of the weight for a certain entry in the table, but the ratio between its weights. For example, if only two entries in the table are in use, they will consume all the bandwidth. If each entry has a weight of 20, each of them consumes 50 percent of the available bandwidth. If one entry has a weight of 30 and the other has a weight of 10, then the first one gets 75 percent of available bandwidth while the other gains the remaining

25 percent.

Therefore, we will compute the number of the entry and its weight based on the information from the VLArbitration Table. According to the InfiniBand specification, there are no more than 64 entries in each arbitration table. If each connection is devoted to a different entry, this would limit the total number of connections that can be accepted. Thus, a connection requiring very high bandwidth could also need more than one entry in the table. For this reason, we propose to group the connections with the same SL into a single entry of the table until completing the maximum weight of that entry before moving to another free entry. It is worth noting that the number of entries in the table is not a limitation for the acceptance of new connections.

Moreover, the weight of each entry is the sum of those of all connections in this entry, and this sum will be close to the maximum weight value 255. So we give the final weight value to each entry in the arbitration table as 255. This means each entry in the table has the same weight. The bandwidth occupied by the specific VL would release until all connections in this VL finish transferring the packets [42, 61, 77].

Moreover, we set an Entry Threshold (ET) for each entry in the arbitration table since the threshold function is one of the congestion control methods to increase the utilization of the queueing system and the entry threshold is based on the number of the entry.

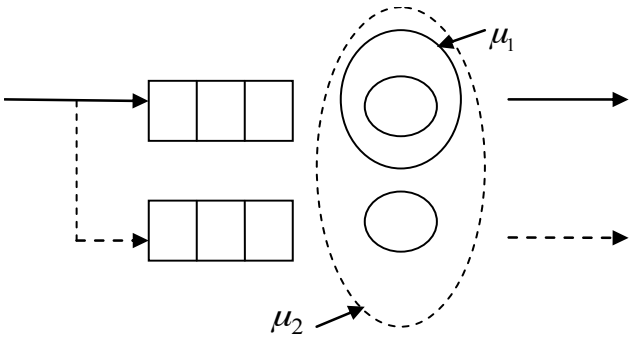


Fig. 5. 1: Architecture of Entry Threshold

We focus on the specific VL_i in the VLArbitration Table and model it as a GE/GE/1/N/ET/FCFS [71] queue system with GE-type inter-arrival and service time distributions, entry threshold (ET) and first-come-first-served (FCFS) service discipline. The bursty traffic used in this paper is modelled by a Compound Poisson Process (CPP) with geometrically distributed batches or, equivalently, GE-Type inter-arrival times.

When the number of entry (E) with the specific VL_i reaches threshold (ET), the system will acquire available empty VL_s to transfer the packets. As a result, more services will be provided over the threshold level. The VL_i comes into the system coupled by entry threshold with service rate μ_1, capacity N. The service rate will change from μ_1 to μ_2, if the weight value reaches threshold T.

On the other hand, when the number of entry (E) becomes less than threshold (ET), the server rate will reduce to μ_2 , as shown in Fig. 5.1 In this context, the improvement of service rate will reduce the delay time, mean queue length and blocking probability and the GE-Type queueing system with ET is one of the efficient models to increase network QoS.

5.2.1 Assumptions and Notation

The queueing system with GE/GE/1/N/ET/FCFS [71] is considered as either in an operational or stochastic context. The queue has an input and an output port, a service facility consisting of one server and a maximum capacity of N packets.

The model used to analyze the new flow control mechanism is based on the following assumptions.

a) When the value of the specific VL_i reaches the threshold, the system will try to use another empty VL to send packets. If all other VLs are occupied, these entries will wait in the VL arbitration tables.

b) Once a Virtual Lane j accepts the packet of an entry of VL_i, it must serve the remainder of the entry of VL_i before accepting serving any packet from another VL.

c) After VL_i completes transferring all the packets, this Virtual Lane will be released immediately and then check the VL arbitration tables. If any VL waits in the table, this Virtual Lane will accept one of them. Otherwise, it can accept other VL*s* in the VL arbitration tables whose entry value has reached the threshold.

For clarity, the following notations are adopted:

E	number of the entries of VL_i
λ_i	mean arrival rate of the specific VL_i

μ_1 service completion rate under threshold

μ_2 service completion rate when the number of entries reaches threshold

T threshold ET in the number of the entries with VL_i

t number of the messages in the VL_i when the number of entries reach the entry threshold

$\rho_1 = \frac{\lambda}{\mu_1}$ traffic intensity before threshold

$\rho_2 = \frac{\lambda}{\mu_2}$ traffic intensity after threshold

C_a^2 squared coefficient of variation for the inter-arrival time distribution

C_s^2 squared coefficient of variation for the overall service time distribution

$P(n)$ state probability of the queue

$\bar{L}$ mean queue length

The following analysis assumes that parameters λ_i, μ_1, μ_2, C_a^2, and C_s^2 form the basic set of a prior knowledge and presents the queue probability assignments subject to additional prior information.

5.2.2 Prior Information

The state probabilities, $P(n)$, have the following mean value constraints,

where $n = \begin{cases} 0, 1, 2, \ldots, N & (E < T) \\ 0, 1, 2, \ldots, 2N & (E \geq T) \end{cases}$ [67].

a) Normalization

$$\sum_{n=0}^{2N} P(n) = 1 \tag{5.1}$$

b) Utilization U_i, $0 < U_i < 1$, $i = 1, 2$

$$\sum_{n=0}^{N} h_1(n) P(n) = U_1 \tag{5.2}$$

where $h_1(n) = \begin{cases} 0 & (E = 0) \\ 1 & (0 < E \leq T) \\ 0 & (T < E \leq 2N) \end{cases}$

$$\sum_{n=0}^{2N} h_2(n) P(n) = U_2 \tag{5.3}$$

where $h_2(n) = \begin{cases} 0 & (E \leq T) \\ 1 & (\text{otherwise}) \end{cases}$

c) Mean Queue Length

$$\sum_{n=1}^{\infty} nP(n) = \bar{L} \tag{5.4}$$

d) Full buffer state probability, $P(n) = \phi$ $(0 < \phi < 1)$

$$\sum_{n=0}^{2N} f(n) P(n) = \phi \quad \text{where} \quad f(n) = \begin{cases} 0 & (n < 2N) \\ 1 & (n = 2N) \end{cases} \tag{5.5}$$

satisfying the flow balance condition, namely

$\lambda(1 - b) = \mu(1 - b)$,

where b is the blocking probability.

5.2.3 Maximum Entropy (ME) Formalism

The form of the state probability distribution, $P(n)$, where $(n = 1, 2, ..., 2N)$ can be characterized by maximizing the entropy function [70]:

$$H(p) = -\sum_{n=0}^{2N} P(n)\log P(n) \tag{5.6}$$

Provided the mean value constraints for utilization, mean queue length and full buffer state about the state probability, $P(n)$ as known it is implied that after some manipulation the ME state probability distribution for the proposed model can be given by:

$$P(n) = \frac{1}{Z} g_1^{h_1(n)} g_2^{h_2(n)} x^n y^{f(n)} \quad (0 \le n \le 2N) \tag{5.7}$$

where the normalizing constant, Z, is clearly given by

$$Z = \sum_{n=0}^{2N} g_1^{h_1(n)} g_2^{h_2(n)} x^n y^{f(n)} \quad (0 \le n \le 2N) \tag{5.8}$$

Using the normalizing constraint (b), $P(0)$ can be derived as

$$P(0) = \frac{1}{Z} = \frac{1}{1 + g_1 \dfrac{x_1 - x_1^t}{1 - x_1} + g_2 \dfrac{x_2 - x_2^{2N-t}}{1 - x_2}} \tag{5.9}$$

Using Equation (5.7) and Equation (5.9), the probability distribution for the queue length can be obtained as below:

$$P(n) = \begin{cases} P(0)g_1 x_1^t & (0 \le E \le T) \\ P(0)g_2 x_2^{2N-t} & (T \le E \le 2N) \end{cases} \tag{5.10}$$

With $\rho_1 < 1$, these expressions represent the maximum entropy solutions of GE/GE/1 queue with threshold T, at equilibrium, subject to the utilization and mean queue length constraints. It easily follows that the Lagrangian coefficients g_i and x_i, $(i = 1,2)$, are given by [70]

$$g_1 = \frac{\rho_1(1-x_1)}{x_1(1-\rho_1)} = \frac{\rho_1^2}{(\bar{L}-\rho_1)(1-\rho_1)} \tag{5.11}$$

and $$x_1 = \frac{\bar{L}_1 - \rho_1}{\bar{L}_1} \tag{5.12}$$

where ρ_1 and $\bar{L}_1$ are

$$\rho_1 = \frac{\lambda}{\mu_1} \tag{5.13}$$

$$\bar{L}_1 = \frac{\rho_1}{2}\left(1 + \frac{C_a^2 + \rho_1 C_s^2}{1-\rho_1}\right) \tag{5.14}$$

Also we define ρ_2 as:

$$\rho_2 = \frac{\lambda}{\mu_2} \tag{5.15}$$

Then, we get g_2 and x_2:

$$g_2 = \frac{(\frac{\rho}{1-\rho} - X)(\rho - (1-\rho)X)}{\bar{L} - (1-\rho)\left(\frac{X}{1-x_1} - tx_1^{t+1} + \frac{t\rho}{1-\rho} - (t+1)X - 1\right)} \tag{5.16}$$

$$x_2 = \frac{\rho(1-X)-X}{(1-\rho)(g_2-X)+\rho} \tag{5.17}$$

where $X = \frac{g_1 x_1 (1-x_1^t)}{1-x_1}$

Using Equation (5.10) for $(n = 2N)$, and flow balance condition the Lagrangian coefficient y is given by

$$y = \frac{\rho\left(1+g_2\frac{x_2-x_2^{2N-t}}{1-x_2}\right)-\left(g_2\frac{x_2-x_2^{2N-t}}{1-x_2}\right)}{\rho g_2 x_2^{2N-t}} \tag{5.18}$$

Finally, ρ and the mean queue length $\bar{L}$ are

$$\rho = \frac{t\rho_1 + (2N-t)\rho_2}{2N} \tag{5.19}$$

$$\bar{L} = \bar{L}_1 + x_1\bar{L}_2 + tx_1^t \tag{5.20}$$

where

$$\bar{L}_1 = \frac{\rho_1}{2}\left(1+\frac{C_a^2+\rho_1 C_s^2}{1-\rho_1}\right) \tag{5.21}$$

$$\bar{L}_2 = \frac{\rho_2}{2}\left(1+\frac{C_a^2+\rho_2 C_s^2}{1-\rho_2}\right) \tag{5.22}$$

5.2.4 Model Validation and Performance Analysis

This section focuses on the validation of the ME analytical performance results of the proposed active flow control against simulation results. We consider the simulation scenario which has been widely used in the literature and is illustrated in Fig. 5.2.

Specifically, the simulation scenario has two nodes (channel adapter) and one switch. Each node includes 4 virtual lanes with the buffer capacity of 10 packets. When the number of the entry in the VLArbitration Table reaches threshold ET, the service rate jumps from μ_1 to μ_2 based on the entry threshold (ET) function.

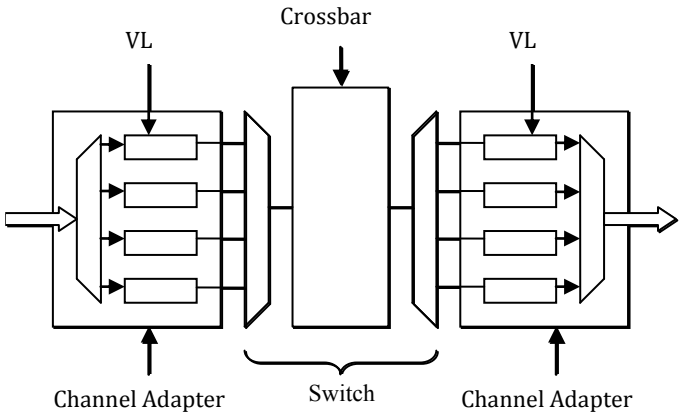


Fig. 5. 2: Simulation Scenario

On the other hand, the server rate goes back to μ_1 if the number of entries is less than the threshold. The traffic parameters in the simulation are $N = 8$, $\mu_1 = 8.0$, $\mu_2 = 16.0$, $C_a^2 = 5$, and $C_s^2 = 7$.

We use the error measures (EMs) [41] between simulation (Sim) results and ME algorithmic values as the criterion to test the accuracy of the analytical results. The absolute ratio of the marginal Mean Queue Length, $\bar{L}$, is defined by

$$EM(\bar{L}_j) = \left| \frac{Sim(\bar{L}_j) - ME(\bar{L}_j)}{Sim(\bar{L}_j)} \right| \quad (j = 1,2) \tag{6.23}$$

where $Sim(\bar{L}_j)$ is the result obtained from the simulation and $ME(\bar{L}_j)$ is the

result of the ME. The approximate solution is said to be within a tolerance ε ($\varepsilon > 0$) if the estimated EMs do not exceed ε.

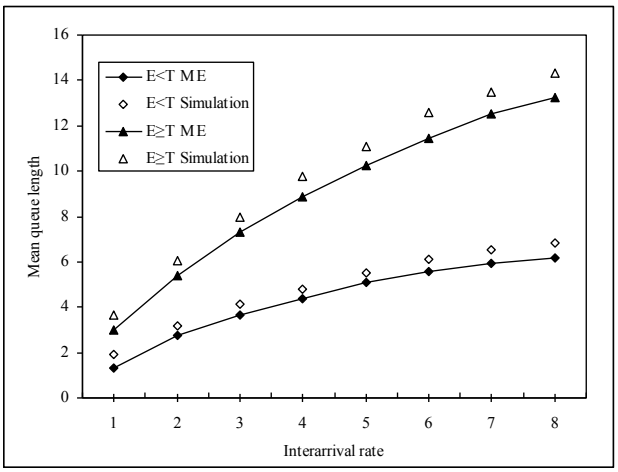


Fig. 5. 3: Effect of traffic variability on the mean queue length

Fig. 5.3 and Fig. 5.4 respectively depict the results of the mean queue length predicted by the ME analysis of the new mechanism plotted against those provided by the simulation results. The figures reveal that the simulation results match those predicted by the ME analytical results.

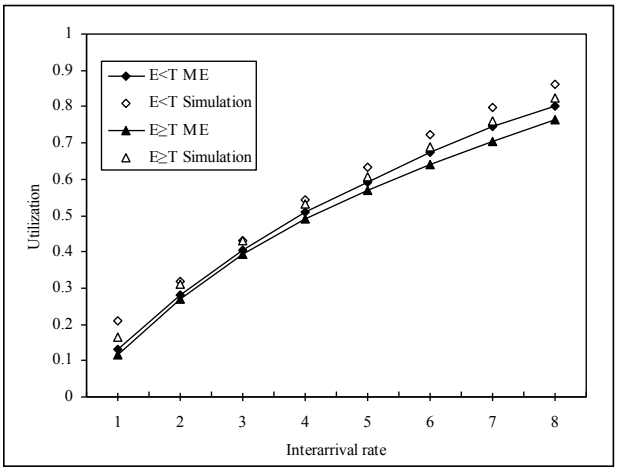


Fig. 5. 4: Effect of traffic variability on the utilization

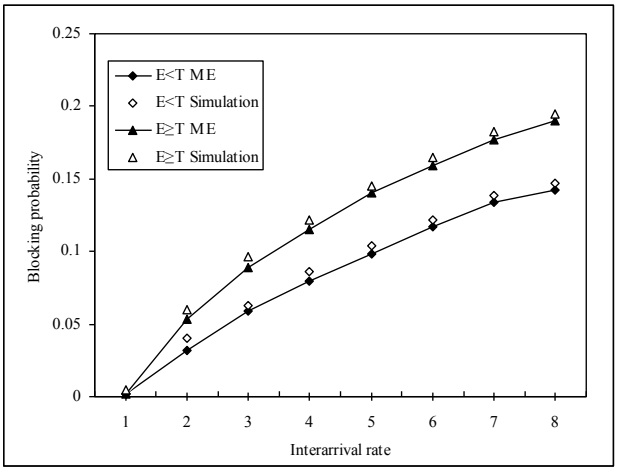


Fig. 5. 5: Effect of traffic variability on the blocking probability

We have computed the EMs and found that the average EM is around 5%~10%, demonstrating the accuracy of the analytical results. So the entry threshold function can be modelled by the ME solutions.

We also compare the blocking probability with and without the entry threshold function. The results are depicted in Fig. 5.5. It is clear that the entry threshold can effectively reduce the blocking probability and the ME analysis results match those obtained from simulation as well.

5.3 Arrival Job Threshold (AJT) in InfiniBand Networks

Accepting or rejecting the connection requests are decided by the local communication management agents based on the local information available in each InfiniBand networks switch. The information includes the state of the output links and how much bandwidth they have already reserved. When a connection is

accepted, the agent modifies the VLArbitration table based on the connection requirements.

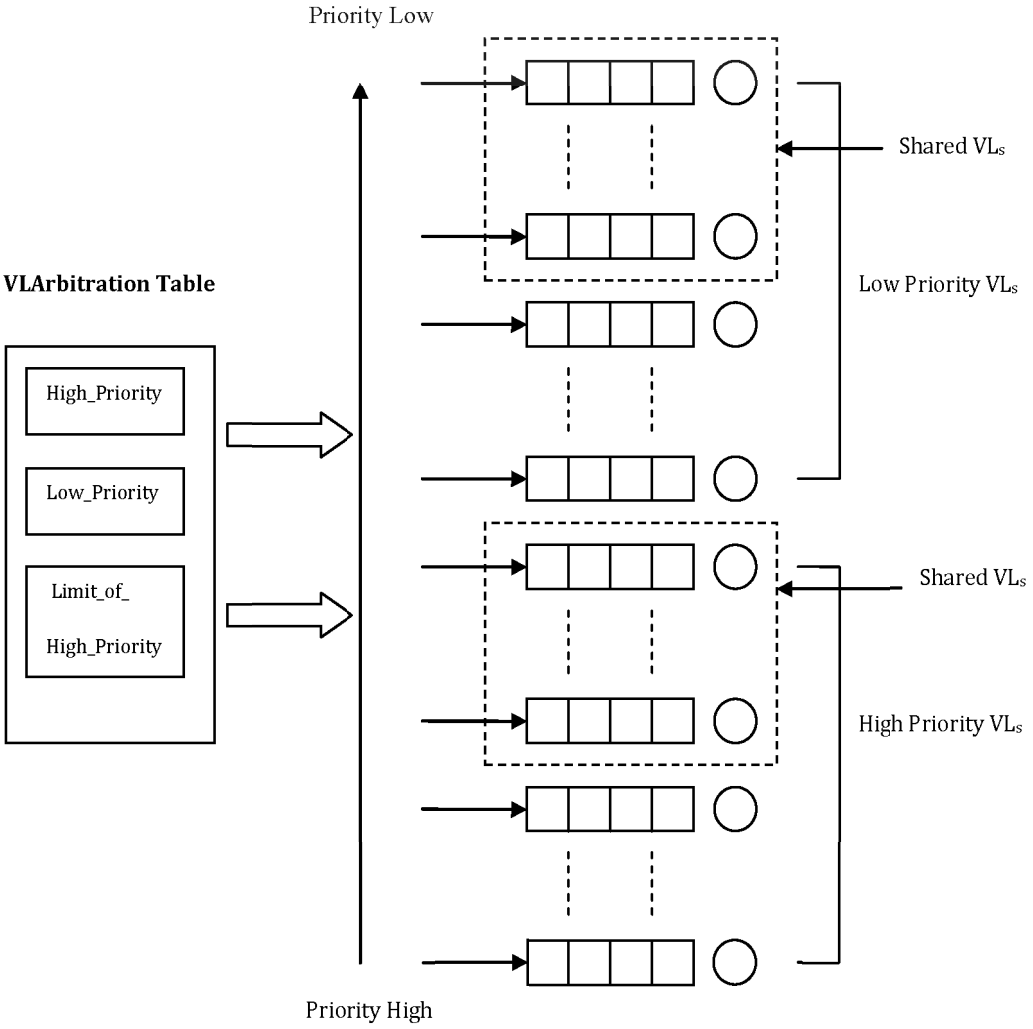


Fig. 5. 6: Architecture of Arrival Job Threshold

Moreover, we set an Arrival Job Threshold (AJT) for each VL, since the threshold function is one of the congestion control methods to increase the utilization of the queueing system and the arrival job threshold is based on the number of job in each virtual lane.

We focus on the specific VL_i with high priority and model it as a

GE/GE/1/N/ET/FCFS [71] queue system with GE-type interarrival and service time distribution, arrival job threshold (AJT) and first-come-first-served (FCFS) service discipline. The bursty traffic used in this paper is modeled by a Compound Poisson Process (CPP) with geometrically distributed batches or, equivalently, GE-Type interarrival times.

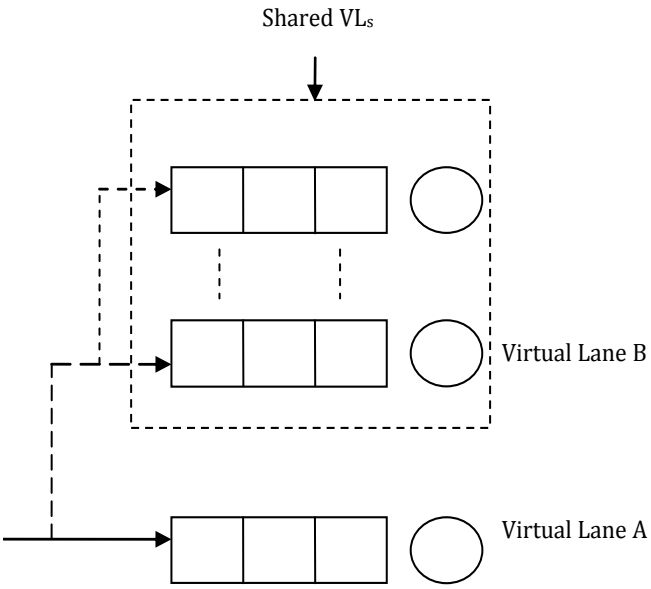


Fig. 5. 7: The process of sharing the virtual lanes

The high priority VLs and low priority VLs are divided into two parts: normal VLs and shared VLs, as shown in Fig. 5.6. When the number of jobs with the specific normal VL_i reaches threshold (AJT), the system will acquire available empty shared VLs to transfer the packets. As a result, more services will be provided over the threshold level. When the total number of the jobs in high priority Virtual Lane (Virtual Lane A with service rate μ_1, capacity N) reaches the Arrival Job Threshold (AJT) L_1, the Virtual Lane A will use one of the empty Shared

Virtual Lanes (Virtual Lane B). The service rate will change from μ_1 to μ_2 and the capacity changes from N to $2N$.

On the other hand, when the number of arrival jobs becomes less than threshold (AJT), the Virtual Lane A cannot use the Shared VLs any more. Also if the total number of the jobs in Virtual Lane B reaches Threshold L_2, the jobs from the high priority virtual lane (Virtual Lane A and Virtual Lane B) will use the second empty shared virtual lane, and so on, as shown in Fig. 5.7.

5.3.1 Assumptions and Notation

The queueing system with GE/GE/1/N/AJT/FCFS [71] is considered as either in an operational or stochastic context. The queue has an input and an output port, a service facility consisting of one server, maximum capacity of N packets and a threshold.

The model we proposed to analyze the new flow control mechanism is based on the following assumptions.

a) When the value of the specific VL_i reaches the threshold, the system will try to use an empty shared VL to send packets. If all shared VLs are occupied, these entries will wait in the VL arbitration tables.

b) Once a Virtual Lane j accepts the packet of an entry of VL_i, it must accept the remainder of the entry of VL_i before accepting any packet from another VL.

c) After the VL_i completes transferring all the packets, the Virtual Lane i will be released immediately. This Virtual Lane will then check the VL arbitration tables. If any VL_i waits in the table, this VL will accept one of them. Otherwise, it can accept other VLs in the VL arbitration tables whose value of the entries has reached threshold.

For clarity, the following notations are adopted:

n_1	number of the jobs in the high priority VL
n_i	number of the jobs in the *i*th shared VL
λ_i	mean arrival rate
μ_1	service completion rate under threshold of the high priority VL
μ_2	service completion rate when number of arrival job reaches threshold of the high priority VL
μ_i	service completion rate when number of jobs reaches threshold of the *i*th shared VL
T_1	threshold in the high priority VL
T_i	threshold in the *i*th shared VL
$\rho_1 = \frac{\lambda}{\mu_1}$	traffic intensity before threshold T_1
$\rho_2 = \frac{\lambda}{\mu_2}$	traffic intensity after threshold T_1 and before threshold T_2

$\rho_i = \frac{\lambda}{\mu_i}$ traffic intensity after ith threshold T_i

C_a^2 SCV for the inter-arrival time distribution

C_s^2 SCV for the overall service time distribution

$P(n)$ state probability of the queue.

The following analysis assumes that the parameters λ_i, μ_1, μ_2, μ_i, C_a^2, and C_s^2 form the basic set of a prior knowledge and presents the queue probability assignments subject to additional prior information.

5.3.2 Prior Information

Suppose all that is known about the state probabilities, $P(n)$, where

$$n = \begin{cases} 0, 1, 2, ..., N & (n_1 < T_1) \\ 0, 1, 2, ..., iN & (T_{i-1} \le n_i \ \& \ n_i < T_i), \ i = 2, 3, 4..., n \\ 0, 1, 2, ..., (i+1)N & (T_i \le n_i) \end{cases}$$

are the following mean value constraints [67].

a) Normalization

$$\sum_{n=0}^{iN} P(n) = 1 \tag{5.24}$$

b) Utilization U_i, $0 < U_i < 1$, $i = 2, 3, 4..., n$

$$\sum_{n=0}^{N} h_1(n) P(n) = U_1 \tag{5.25}$$

where $h_1(n) = \begin{cases} 0 & (n_1 = 0) \\ 1 & (0 < n_1 < T_1) \\ 0 & (\text{otherwise}) \end{cases}$

$$\sum_{n=0}^{iN} h_i(n)P(n) = U_i \tag{5.26}$$

where $h_i(n) = \begin{cases} 0 & (n_{i-1} < T_{i-1}) \\ 1 & (n_{i-1} \geq T_{i-1}) \\ 0 & (\text{otherwise}) \end{cases}$

$$\sum_{n=0}^{(i+1)N} h_{i+1}(n)P(n) = U_{i+1} \tag{5.27}$$

where $h_{i+1}(n) = \begin{cases} 0 & (n_i < T_i) \\ 1 & (n_i \geq T_i) \\ 0 & (\text{otherwise}) \end{cases}$

c) Mean Queue Length

$$\sum_{n=1}^{\infty} nP(n) = \bar{L} \tag{5.28}$$

d) Full buffer state probability, $P(n) = \phi \ (0 < \phi < 1)$, $i = 2, 3, 4..., n$

$$\sum_{n=0}^{iN} f(n)P(n) = \phi \tag{5.29}$$

where $f(n) = \begin{cases} 0 & (n < iN) \\ 1 & (n = iN) \end{cases}$

satisfies the flow balance condition, namely

$$\lambda(1-b) = \mu(1-b) \tag{5.30}$$

where b is the blocking probability.

5.3.3 Maximum Entropy (ME) Formalism

The form of the state probability distribution, $P(n)$, where $(n = 1, 2, ..., iN)$

can be characterized by maximizing the entropy functional [70].

$$H(p) = -\sum_{n=0}^{iN} P(n)\log P(n) \tag{5.31}$$

Provided the mean value constraints for utilization, mean queue length and full buffer state about the state probability, $P(n)$ as known, it is implied that after some manipulation the ME state probability distribution for the proposed model can be given by:

$$P(n) = \frac{1}{Z} g_1^{h_1(n)} g_2^{h_2(n)} \ldots g_i^{h_i(n)} x^n y^{f(n)} \quad (0 \le n \le iN) \tag{5.32}$$

Where Z, the normalizing constant, is clearly given by

$$Z = \sum_{n=0}^{iN} g_1^{h_1(n)} g_2^{h_2(n)} \ldots g_i^{h_i(n)} x^n y^{f(n)} \quad (0 \le n \le iN) \tag{5.33}$$

Using the normalizing Constraint (b), $P(0)$ can be derived as

$$P(0) = \frac{1}{Z} = \frac{1}{1 + g_1 \dfrac{x_1 - x_1^{T_1}}{1 - x_1} + g_2 \dfrac{x_2 - x_2^{T_2}}{1 - x_2} + \ldots + g_i \dfrac{x_i - x_i^{iN - T_i}}{1 - x_i}} \tag{5.34}$$

Using Equations (5.32) and (5.34), the probability distribution for the queue length can be obtained as below:

$$P(n) = \begin{cases} P(0) g_1 x_1^{n} \ (n_i < T_1) \\ P(0) g_1 x_1^{T_1} x_2^{T_2} \ldots x_i^{iN-T_1} \ (T_{i-1} \le n_i \ \& \ n_i < T_i) \\ P(0) g_1 x_1^{T_1} x_2^{T_2} \ldots x_i^{iN-T_3} \ (T_i \le n_i) \end{cases} \tag{5.35}$$

With $\rho_1 < 1$, these expressions represent the maximum entropy solutions of GE/GE/1 queue with threshold T, at equilibrium, subject to utilization and mean queue length constraints. It easily follows that the Lagrangian coefficients g_i and

x_i, $(i = 1,2)$, are given by [70]

$$g_1 = \frac{\rho_1(1-x_1)}{x_1(1-\rho_1)} = \frac{\rho_1^2}{(\bar{L}-\rho_1)(1-\rho_1)} \tag{5.36}$$

and

$$x_1 = \frac{\bar{L}_1 - \rho_1}{\bar{L}_1} \tag{5.37}$$

where ρ_1 and $\bar{L}_1$ will be

$$\rho_1 = \frac{\lambda}{\mu_1} \tag{5.38}$$

$$\bar{L}_1 = \frac{\rho_1}{2}\left(1 + \frac{C_a^2 + \rho_1 C_s^2}{1-\rho_1}\right) \tag{5.39}$$

Also we define ρ_2 as following:

$$\rho_i = \frac{\lambda}{\mu_i} \tag{5.40}$$

Then, we get g_i and x_i:

$$g_i = \frac{(1-x)\rho_i}{(1-\rho)x_i} \tag{5.41}$$

$$x_i = \frac{\bar{L}_i - \rho_i}{\bar{L}} \tag{5.42}$$

Using Equation (5.35) for $(n = iN)$, and flow balance condition the Lagrangian coefficient y is given by

$$y = \frac{\rho + (\rho - 1)(g_1 \frac{x_1 - x_1^{T_1+1}}{1 - x_1} + g_2 x_1^{T_1} \frac{x_2 - x_2^{T_2+1}}{1 - x_2} + \ldots + g_i x_1^{T_1} \ldots x_{i-1}^{T_{i-1}} x_i^{T_i} \frac{x_i - x_i^{iN - T_3}}{1 - x_i})}{\rho g_3 x_1^{T_1} x_2^{T_2} \ldots x_i^{iN - T_3}} \tag{5.43}$$

Finally, ρ and the mean queue length $\bar{L}$ are

$$\rho = \frac{t\rho_1 + (2N - t)\rho_2}{2N} \tag{5.44}$$

$$\begin{aligned} \bar{L} &= \bar{L}_1 + x_1^{T_1} \bar{L}_2 + x_2^{T_2} \bar{L}_3 + \ldots + x_{i-1}^{T_{i-1}} \bar{L}_i \\ &+ T_i x_1^{T_1} (\rho - (1 - \rho) g_1 \frac{x_1 - x_1^{T_1}}{1 - x_1} - (1 - \rho) g_2 \frac{x_2 - x_2^{T_2}}{1 - x_2} - \ldots - (1 - \rho) g_i \frac{x_i - x_i^{T_i}}{1 - x_i}) \end{aligned} \tag{6.45}$$

where

$$\bar{L}_1 = \frac{1}{2}\left\{1 + \frac{C_a^2 + \rho_1 C_S^2}{1 - \rho_1}\right\}$$

$$\bar{L}_2 = \frac{1}{2}\left\{1 + \frac{C_a^2 + \rho_2 C_S^2}{1 - \rho_2}\right\}$$

$$\bar{L}_i = \frac{1}{2}\left\{1 + \frac{C_a^2 + \rho_i C_S^2}{1 - \rho_i}\right\}$$

5.3.4 Model Validation and Performance Analysis

This section focuses on the validation of the ME analytical performance results of the proposed active flow control against simulation results. We consider the simulation scenario which has been widely used in the literature [41] and is illustrated in Fig. 5.8.

More specifically, the simulation scenario has two nodes (channel adapter) and one switch. Each node includes 4 virtual lanes with the buffer capacity of 10 packets.

When the number of the arrival job in the high priority VL reaches threshold (AJT), the service rate jumps from μ_1 to μ_2 and rom μ_2 to μ_3 based on the Arrival Job Threshold (AJT) function and the capacity changes from N to $2N$ and from $2N$ to $3N$ as well. On the other hand, the server rate and capacity will reduce if the number of arrival job is less than the threshold. The traffic parameters in the simulation are $N = 8$, $i = 3$, $\mu_1 = 8.0$, $\mu_2 = 16.0$, $\mu_3 = 12.0, C_a^2 = 5$, and $C_s^2 = 7$.

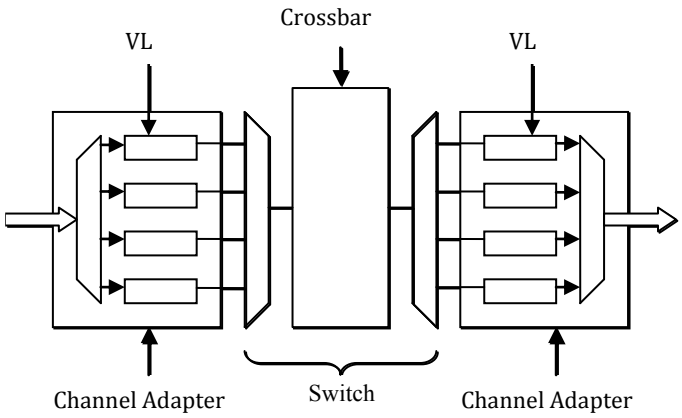


Fig. 5. 8: Simulation Scenario

We use the error measures (EMs) [52] between simulation (Sim) results and ME algorithmic values as the criterion to test the accuracy of the analytical results. The absolute ratio of marginal Mean Queue Length, $\bar{L}$, is defined by

$$EM(\bar{L}_j) = \left| \frac{Sim(\bar{L}_j) - ME(\bar{L}_j)}{Sim(\bar{L}_j)} \right| \quad (j = 1,2) \tag{5.46}$$

where $Sim(\bar{L}_j)$ is the result obtained from the simulation and $ME(\bar{L}_j)$ is the result of the ME.

The approximate solution is said to be within a tolerance ε ($\varepsilon > 0$) if the estimated EMs do not exceed ε.

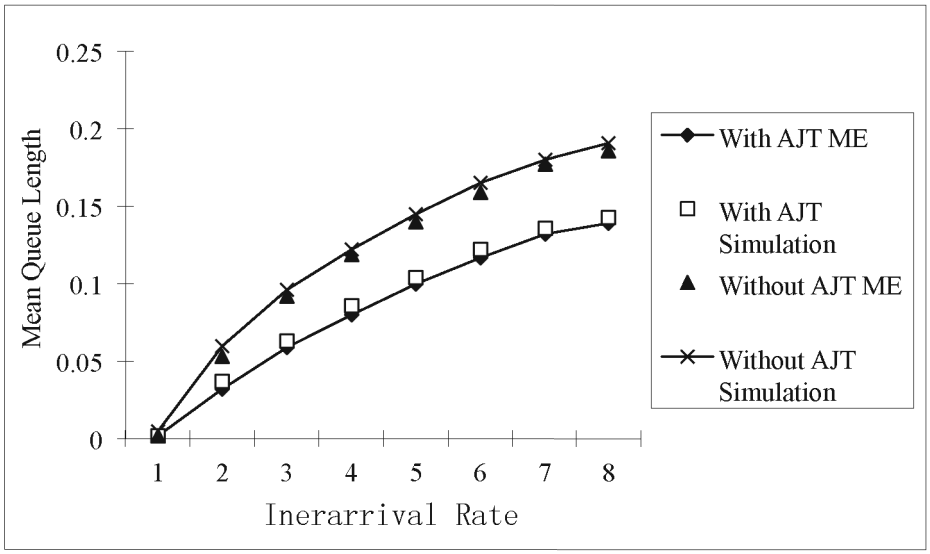


Fig. 5. 9: Effect of traffic variability on the mean queue length

Fig. 5.9 and Fig. 5.10 depict the results of the mean queue length predicted by the ME analysis of the new mechanism plotted against those provided by the simulation results, respectively. The figures reveal that the simulation results match with those predicted by the ME analytical results. We have computed the EMs and found that the average EM is around 0.05~0.1, demonstrating the accuracy of the analytical results. So the entry threshold function can be modelled by the ME solutions.

We have also compared the blocking probability with and without the entry threshold function. The results are depicted in Fig.5.11. It is clear that the Arrival Job Threshold can effectively reduce the blocking probability and the ME analysis results match the simulation as well.

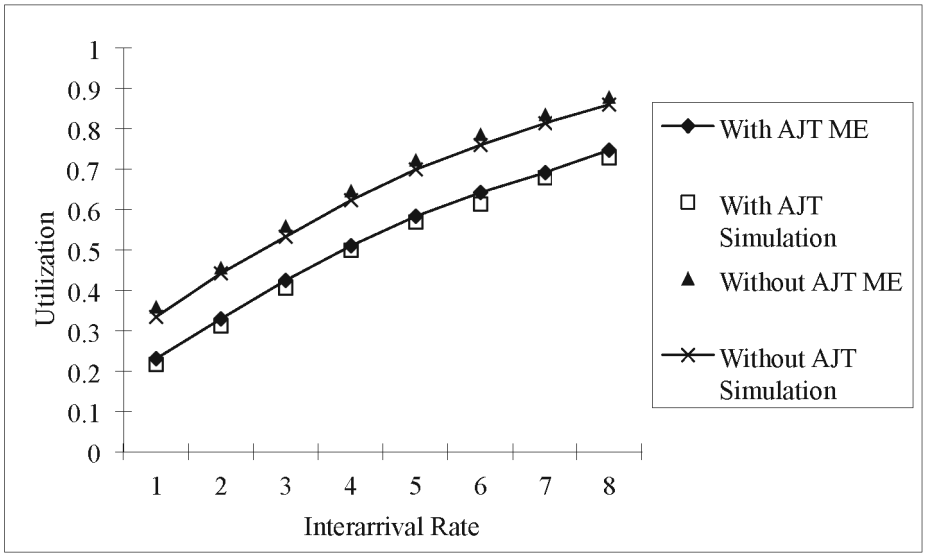


Fig. 5. 10: Effect of traffic variability on the utilization

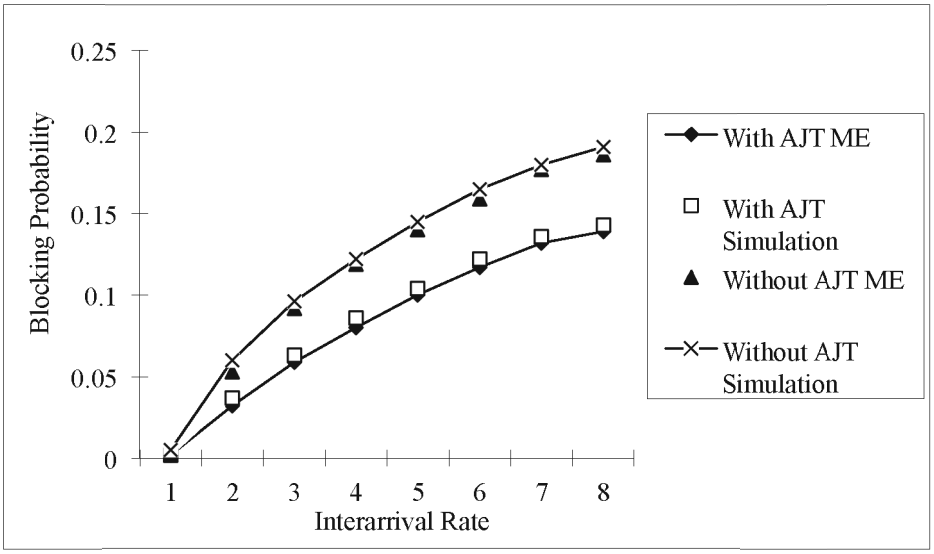


Fig. 5. 11: Effect of traffic variability on the blocking probability

5.4 Summary

This chapter has presented two active flow control mechanisms in the InfiniBand Networks architecture, Entry Threshold (ET) and Arrival Job

Threshold (AJT), in order to reduce the delay, blocking probability, and mean queue length as well as improve the system utilization. Consequently, these mechanisms can enhance the QoS of InfiniBand.

We have used the principle of ME, subject to GE-Type queueing theoretic constraints to develop the analytical queueing and delay mean value and provide simple and robust analytic building block tools for performance modelling including the inter-arrival times and service times of the input traffic. Typical numerical tests have been carried out to validate the credibility of the ME solutions against simulation results.

Chapter 6

Conclusions and Future Work

The primary focus of this thesis is the performance modelling of improving the QoS for the InfiniBand networks in cluster computing systems. In what follows, this chapter draws conclusions of the thesis and gives some suggestions for the future work.

6.1 Conclusions

The contributions of this thesis are concluded in this section as follow:

- An effective rate-based congestion control mechanism, named as Power Increase and Power Decrease (PIPD), for traffic in the InfiniBand network has been developed to provide the higher bandwidth utilization and ensure the fairness, which includes an ECN packet marking mechanism and a source response mechanism. Firstly, we analyzed problem of congestion spreading and the relationship between a root source and a victim of congestion in a simple InfiniBand network simulation scenario. Then we extend a traditional admission control mechanism to suit the InfiniBand network architecture. This mechanism can provide the multiple-class connections with different priorities.

- An improved credit-based flow control scheme has been proposed to guarantee zero packet loss. Because InifiniBand switches use link level flow control, and packet dropping may cause congestion spreading or tree saturation. A consequence of this characteristic is that InfiniBand network switch cannot use packet losses as indication of congestion. We have also developed a queuing network analysis method for performance evaluation of this new scheme. The performance metrics to be derived include mean queue length, throughput and response time of the system. Results obtained from the analytical model have showed that this model can effectively evaluate the performance of credit-based flow control in the InfiniBand networks.

- We presented two efficient flow control mechanisms in the InifniBand network architecture, and we used the principle of ME, subject to GE-Type queueing theoretic constraints to develop the analytical performance modelling. The idea of both mechanisms are based on the threshold science the threshold function is one of the congestion control methods to increase the utilization of the queueing system. One is Entry Threshold (ET) that is setting for each entry in the arbitration table and the entry threshold is based on the number of the entry. Another is Arrival Job Threshold (AJT) that is based on the number of job in each Virtual Lane (VL) in the InfiniBand networks. We modelled these two mechanisms as a GE/GE/1/N/FCFS queue system. The principle of

Maximum Entropy (ME) is adopted as an effective methodology to analyse these two mechanisms with the Generalized Exponential (GE)-Type distribution for modelling the inter-arrival times and service times of the input traffic in FCFS service discipline. ET and AJT mechanisms can reduce delay, blocking probability, and mean queue length as well as enhance the QoS of InifniBand networks.

6.2 Future Work

The thesis mainly investigates the modelling and effective QoS mechanisms for InfiniBand networks in high-performance cluster computing systems. Moving beyond the core of the present work, there are several interesting issues and open problems that require further investigation. The future work can be suggested to extend this research are briefly outlined below.

- The present research has focused on the analysis of proposed scheme subject to one or two classes of traffic. In practical networks, traffic generated based on kinds of application is classified into multiple categories. So we will extend the proposed modelling methods to handle the InfiniBand networks subject to multiple heterogeneous traffic classes.

- More and more measurement evidences have shown that the traffic generated in modern interconnected communication networks exhibit extremely bursty arrival nature over a wide range of time scales.

Self-similar or long-range-dependent processes have significantly different theoretical properties from those of the conventional Markovian non-memory arrival processes, which are adpoted to model fractal behavior of the packet arrivals. A more challenging extension of research work would be to develop the analytical model in presence of fractal self-similar traffic.

- Also we will focus on using some network simulation tool software, such as OPNET modeler to extend simulation scenarios in order to explore the flow control mechanisms with the richer traffic patterns and larger network topologies to describe the actual networks better.

Bibliography

[1] I. B. A. Gray, R. Kenway, L. Smith, M. Guest, C. Kitchen, P. Calleja, A. Korzynski, S. Rankin, M. Ashworth, A. Porter, I. Todorov, M. Plummer, E. Jones, L. Steenman-Clark, B. Ralston, C. Laughton, "Mapping application performance to HPC Architecture," *Computer Physics Communications,* vol. 183, pp. 520-529, 2012.

[2] Y. Afek, Y. Mansour, and Z. Ostfeld, "Convergence complexity of optimistic rate based flow control algorithms," in *28th Annual ACM Symposium on Theory of Computing (STOC'1996)*, Philadepia, Pennsylvania, USA, 22-24 May,1996, pp. 89-98.

[3] J. S. Alanazi and D. D. Kouvatsos, "A unified ME algorithm for arbitrary open QNMs with mixed blocking mechanisms," in *11th International Symposium on Applications and the Internet (SAINT'2011)*, Munich, Germany, July 18-21, 2011, pp. 781-816.

[4] F. J. Alfaro, J. Sánchez, L. Orozco, and J. Duato, "Performance evaluation of VBR traffic in InfiniBand," presented at the Canadian Conference on Electrical and Computer Engineering (CCECE'2002), Winnipeg, Manitoba, Canada, 12-15 May, 2002.

[5] F. J. Alfaro, J. L. Sánchez, and J. Duato, "QoS in InfiniBand subnetworks," *Parallel and Distributed Systems,* vol. 15, no.9, pp. 810-823, 2004.

[6] F. J. Alfaro, J. L. Sánchez, and J. Duato, "A strategy to manage time sensitive traffic in InfiniBand," in *16th International Parallel and Distributed Processing Symposium (IPDPS'2002)*, Florida, USA, 15-19 April, 2002, pp. 167-174.

[7] F. J. Alfaro, J. L. Sánchez, J. Duato, and C. R. Das, "A strategy to compute the InfiniBand arbitration tables," in *16th International Parallel and*

Distributed Processing Symposium (IPDPS'2002), Florida, USA, 15-19 April, 2002, pp. 6-11.

[8] N. Alzeidi, A. Khonsari, M. Ould-Khaoua, and L. Mackenzie, "A new approach to model virtual channels in interconnection networks," *Journal of Computer and System Sciences,* vol. 73, no.8, pp. 1121-1130, 2007.

[9] I. T. Association. *InfiniBand Architecture Specification Volume 1. Release 1.2. 1, Jan. 2008*. Available: http://www.infinibandta.org

[10] B. Awerbuch, R. Gawlick, T. Leighton, and Y. Rabani, "On-line admission control and circuit routing for high performance computing and communication," in *35th Foundations of Computer Science (FOCS'1994)*, Santa Fe, New Mexico, USA, 20-22 November, 1994, pp. 412-423.

[11] T. Bailey, M. M. Karp, U. S. O. o. Vocational, and A. Education, *Promoting college access and success: A review of credit-based transition programs*: US Dept. of Education, Office of Vocational and Adult Education, 2003.

[12] P. Balaji, S. Bhagvat, H. W. Jin, and D. K. Panda, "Asynchronous zero-copy communication for synchronous sockets in the Sockets Direct Protocol (SDP) over InfiniBand," in *6th Workshop on Communication Architecture for Clusters (CAC'2006)*, Rhodes Island, Greece, 25-29 April, 2006.

[13] P. Balaji, S. Bhagvat, D. Panda, R. Thakur, and W. Gropp, "Advanced flow-control mechanisms for the sockets direct protocol over infiniband," in *the IEEE International Conference on Parallel Processing (ICPP'2007)*, XiAn, China, 10-14 September, 2007, pp. 73-73.

[14] D. Bansal and H. Balakrishnan, "Binomial congestion control algorithms," in *20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2001)*, Anchorage, USA, 22-26 April, 2001, pp. 631-640

[15] M. Bencivenni, D. Bortolotti, A. Carbone, A. Cavalli, A. Chierici, S. Dal Pra, D. De Girolamo, M. Donatelli, A. Fella, and D. Galli, "Performance of 10 Gigabit Ethernet Using Commodity Hardware," *IEEE Transactions on Nuclear Science* vol. 57, no.2, pp. 630-641, 2010.

[16] B. Bensaou, D. H. K. Tsang, and K. T. Chan, "Credit-based fair queueing (CBFQ): a simple service-scheduling algorithm for packet-switched networks," *IEEE/ACM Transactions on Networking,* vol. 9, no.5, pp. 591-604, 2001.

[17] A. Bermudez, R. Casado, F. J. Quiles, and J. Duato, "Fast routing computation on InfiniBand networks," *IEEE Transactions on Parallel and Distributed Systems,* vol. 17, no.3, pp. 215-226, 2006.

[18] A. Bermudez, R. Casado, F. J. Quiles, and J. Duato, "Handling topology changes in InfiniBand," *IEEE Transactions on Parallel and Distributed Systems,* vol. 18, no.2, pp. 172-185, 2007.

[19] A. Bermúdez, R. Casado, F. J. Quiles, T. M. Pinkston, and J. Duato, "Modeling infiniband with OPNET," in *2nd Workshop on Novel Uses of System Area Networks at HPCA (SAN'2003)*, Anaheim, USA, 8-12 February, 2003.

[20] D. Bertsimas and D. Nakazato, "The distributional Little's law and its applications," *Operations Research,* vol. 43, no.2, pp. 298-310, 1995.

[21] F. Bonomi and K. W. Fendick, "The rate-based flow control framework for the available bit rate ATM service," *Network, IEEE,* vol. 9, no.5, pp. 25-39, 1995.

[22] L. S. Brakmo and L. L. Peterson, "TCP Vegas: End to end congestion avoidance on a global Internet," *Selected Areas in Communications,* vol. 13, no.8, pp. 1465-1480, 1995.

[23] L. Breslau, S. Jamin, and S. Shenker, "Comments on the performance of measurement-based admission control algorithms," in *19th 16th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2000)*, 26-30 March, 2000, pp. 1233-1242.

[24] O. Celebioglu, R. Rajagopalan, and R. Ali, "Exploring infiniband as an HPC cluster interconnect," *High-Performance Computing, Dell Power Solution,* 2004.

[25] H. L. Chao and W. Liao, "Credit-based fair scheduling in ad hoc wireless networks," in *IEEE Vehicular Technology Conference (VTC'2002)*, 24-28 September, 2002, pp. 1442-1446.

[26] A. Charny, "An algorithm for rate allocation in a packet-switching network with feedback," Technical Report MIT/LCS/TR-60, MIT Laboratory for Computer Science, 1994.

[27] A. Charny, D. D. Clark, and R. Jain, "Congestion control with explicit rate indication," in *2nd International Conference on Communications (ICC'1995)*, London, UK, 18-22 June, 1995, pp. 1954-1963

[28] A. Chien and J. Kim, "Approaches to quality of service in high-performance networks," *Parallel Computer Routing and Communication,* pp. 1-17, 1998.

[29] D. M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Computer Networks and ISDN Systems,* vol. 17, no.1, pp. 1-14, 1989.

[30] A. Cohen, "A performance analysis of 4X InfiniBand data transfer operations," presented at the International Parallel and Distributed Processing Symposium (IPDPS'2003), Nice, France, 22-26 April, 2003.

[31] W. J. Dally, "Virtual-channel flow control," *Parallel and Distributed Systems, IEEE Transactions on,* vol. 3, no.2, pp. 194-205, 1992.

[32] J. Duato, I. Johnson, J. Flich, F. Naven, P. Garcia, and T. Nachiondo, "A new scalable and cost-effective congestion management strategy for lossless multistage interconnection networks," *IEEE Trans. on Computers,* vol. C-34, no.10, pp. 108-119, 2005.

[33] J. Duato, S. Yalamanchili, M. B. Caminero, D. Love, and F. J. Quiles, "MMR: A high-performance multimedia router-Architecture and design trade-offs," in *5th International High-Performance Computer Architecture (HPCA'1999)*, 9-12 January, 1999, pp. 300-309.

[34] J. Duato, S. Yalamanchili, and L. M. Ni, *Interconnection networks: An engineering approach*: Morgan Kaufmann, 2003.

[35] M. El-Taha and J. Heath, "Queueing network models of credit-based flow control," *Computers & Mathematics with Applications,* vol. 50, no.3-4, pp. 393-398, 2005.

[36] V. Elek, G. Karlsson, and R. Ronngren, "Admission control based on end-to-end measurements," in *19th annual Conference on Computer Communications (INFOCOM'2000)*, 26-30 March, 2000, pp. 623-630.

[37] Y. Fang, S. Iqbal, and A. Saify, "Designing High-Performance Computing Clusters," ed: Dell Power Solutions, 2005.

[38] D. Ferrari and D. C. Verma, "A scheme for real-time channel establishment in wide-area networks," *Selected Areas in Communications, IEEE Journal on,* vol. 8, no.3, pp. 368-379, 1990.

[39] W. Fischer Kathleen, "The Markov-modulated Poisson process (MMPP) cookbook," *Performance Evaluation,* vol. 18, no.2, pp. 149-171, 1993.

[40] S. Floyd, "TCP and explicit congestion notification," *ACM SIGCOMM Computer Communication Review,* vol. 24, no.5, pp. 10-23, 1994.

[41] S. Floyd, M. Handley, J. Padhye, and J. Widmer, *Equation-based congestion control for unicast applications* vol. 30: ACM, 2000.

[42] D. Goldenberg, M. Kagan, R. Ravid, and M. S. Tsirkin, "Zero copy sockets direct protocol over infiniband-preliminary implementation and performance analysis," in *13th High Performance Interconnects* 17-19 August, 2005, pp. 128-137.

[43] N. Golmie, Y. Saintillan, and D. Su, "ABR switch mechanisms: design issues and performance evaluation," *Computer Networks and ISDN Systems,* vol. 30, no.19, pp. 1749-1761, 1998.

[44] R. E. Grant, P. Balaji, and A. Afsahi, "A study of hardware assisted IP over InfiniBand and its impact on enterprise data center performance," in *IEEE International Symposium on Performance Analysis of Systems & Software (ISPASS'2010)*, New York, USA, 28-30 March, 2010, pp. 144-153.

[45] R. E. Grant, M. J. Rashti, and A. Afsahi, "An analysis of QoS provisioning for Sockets Direct Protocol vs. IPoIB over modern InfiniBand networks," 2008, pp. 79-86.

[46] R. E. Grant, M. J. Rashti, and A. Afsahi, "An analysis of QoS provisioning for Sockets Direct Protocol vs. IPoIB over modern InfiniBand networks," in *International Conference on Parallel Processing Workshop (ICPPW'2008)*, Portland, Oregon, USA, 8-12 Septmber, 2008, pp. 79-86.

[47] D. Gu and J. Zhang, "A new measurement-based admission control method for IEEE802. 11 wireless local area networks," in *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'2003)*, Beijing, China, 7-10 september, 2003, pp. 2009-2013

[48] M. Gusat, D. Craddock, W. Denzel, T. Engbersen, N. Ni, G. Pfister, W. Rooney, and J. Duato, "Congestion control in InfiniBand networks," in *13th Annual IEEE Symposium on High Performance Interconnects* Stanford, USA, 17-19 Augest, 2005, pp. 158-159.

[49] Q. He, Z. Li, H. Wang, and J. Sun, "Research on the conversion efficiency of InfiniBand," in *Multimedia Information Networking and Security (MINES'2011)*, Nanjing, China, 2-4 November, 2011, pp. 146-149.

[50] J. Hou, J. Yang, and S. Papavassiliou, "Integration of pricing with call admission control to meet QoS requirements in cellular networks," *Parallel and Distributed Systems, IEEE Transactions on,* vol. 13, no.9, pp. 898-910, 2002.

[51] V. Jacobson, "Congestion avoidance and control," in *Special Interest Group on Data Communication (SIGCOMM'1988)*, Standford, USA, August 16-18, 1988, pp. 314-329.

[52] R. Jain, "Congestion control and traffic management in ATM networks: Recent advances and a survey," *Computer Networks and ISDN Systems,* vol. 28, no.13, pp. 1723-1738, 1996.

[53] S. Jamin, S. J. Shenker, and P. B. Danzig, "Comparison of measurement-based admission control algorithms for controlled-load

service," in *16th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'1997)*, Kobe, Japan, 7-11 April, 1997, pp. 973-980

[54] W. Jiang, J. Liu, H. W. Jin, D. K. Panda, W. Gropp, and R. Thakur, "High performance MPI-2 one-sided communication over InfiniBand," in *Cluster Computing and the Grid (CCGrid'2004)*, Chicago, USA, 19-22 April, 2004, pp. 531-538.

[55] K. Kandalla, H. Subramoni, K. Tomko, D. Pekurovsky, S. Sur, and D. K. Panda, "High-performance and scalable non-blocking all-to-all with collective offload on InfiniBand clusters: a study with parallel 3D FFT," *Computer Science-Research and Development,* vol. 26, no.3-4, pp. 1-10, 2011.

[56] K. Kandalla, H. Subramoni, J. Vienne, S. P. Raikar, K. Tomko, S. Sur, and D. Panda, "Designing Non-blocking Broadcast with Collective Offload on InfiniBand Clusters: A Case Study with HPL," in *19th Annual IEEE Symposium on High Performance Interconnects (HOTI'2011)*, Santa Clara, USA, 24-26 Augest, 2011, pp. 27-34.

[57] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing systems,* vol. 9, no.1, pp. 5-15, 1991.

[58] J. Kenney, "Traffic Management Specification, Version 4.1," 1999.

[59] J. O. Kephart, "Research challenges of autonomic computing," in *27th International conference on Software Engineering (ICSE'2005)*, St Louis, USA, 15-21 May, 2005, pp. 15-22.

[60] E. J. Kim, K. H. Yum, and C. R. Das, "Performance analysis of a QoS capable cluster interconnect," *Performance Evaluation,* vol. 60, no.1, pp. 275-302, 2005.

[61] E. J. Kim, K. H. Yum, C. R. Das, M. Yousif, and J. Duato, "Performance enhancement techniques for InfiniBandTM Architecture," in *9th High-Performance Computer Architecture (HPCA'2003)*, Anaheim, USA, 12 February, 2003, pp. 253-262.

[62] J. H. Kim, "Bandwidth and latency guarantees in low-cost, high-performance networks," Department of Electrical Engineering, University of Illinois at Urbana-Champaign, Ph.D. thesis,1997.

[63] E. W. Knightly and N. B. Shroff, "Admission control for statistical QoS: Theory and practice," *IEEE Trans. on Network,* vol. 13, no.2, pp. 20-29, 1999.

[64] K. M. Kockelman and S. Kalmanje, "Credit-based congestion pricing: a policy proposal and the public's response," *Transportation Research Part A: Policy and Practice,* vol. 39, no.7-9, pp. 671-690, 2005.

[65] M. J. Koop, R. Kumar, and D. K. Panda, "Can software reliability outperform hardware reliability on high performance interconnects?: a case study with MPI over infiniband," in *22nd ACM International Conference on Supercomputing (ICS'2008)*, Cairo, Egypt, 20-24 October, 2008, pp. 145-154.

[66] M. J. Koop, P. Shamis, I. Rabinovitz, and D. Panda, "Designing high-performance and resilient message passing on InfiniBand," in *Parallel & Distributed Processing Workshops and Phd Forum (IPDPSW'2010)*, Atlanta, USA, 19-23 April, 2010, pp. 1-7.

[67] D. Kouvatsos, "Maximum entropy methods for general queueing networks," *Modelling Techniques and Tools for Performance Analysis, D. Poitier (Ed.),* pp. 589-608, 1985.

[68] D. Kouvatsos and S. Assi, *Generalised entropy maximisation and queues with bursty and/or heavy tails*. Network performance engineering: Springer, 2011.

[69] D. Kouvatsos and I. Awan, "Entropy maximisation and open queueing networks with priorities and blocking," *Performance Evaluation,* vol. 51, no.2, pp. 191-227, 2003.

[70] D. D. Kouvatsos, "Entropy maximisation and queueing network models," *Annals of Operations Research,* vol. 48, no.1, pp. 63-126, 1994.

[71] D. D. Kouvatsos, "A maximum entropy analysis of the G/G/1 queue at equilibrium," *Journal of the Operational Research Society,* vol. 39, no.2, pp. 183-200, 1988.

[72] D. D. Kouvatsos and J. Almond, "Maximum entropy two-station cyclic queues with multiple general servers," *Acta informatica,* vol. 26, no.3, pp. 241-267, 1988.

[73] D. D. Kouvatsos, P. Georgatsos, and N. M. Tabet-Aouel, "A universal maximum entropy algorithm for general multiple class open networks with mixed service disciplines," *Modelling Techniques and Tools for Computer Performance Evaluation, D. Poitier (Ed.),* vol. 43, pp. 397-419, 1989.

[74] D. D. Kouvatsos and N. P. Xenios, "MEM for arbitrary queueing networks with multiple general servers and repetitive-service blocking," *Performance Evaluation,* vol. 10, no.3, pp. 169-195, 1989.

[75] H. Kung, T. Blackwell, and A. Chapman, "Credit-based flow control for ATM networks: credit update protocol, adaptive credit allocation and statistical multiplexing," in *Special Interest Group on Data Communication (SIGCOMM'1994)*, London, UK, August 31 - September, 1994, pp. 101-114.

[76] N. Kung and R. Morris, "Credit-based flow control for ATM networks," *Network,* vol. 9, no.2, pp. 40-48, 1995.

[77] S. Liang, R. Noronha, and D. K. Panda, "Swapping to remote memory over infiniband: An approach using a high performance network block device," in *IEEE Annual International Conference on Cluster Computing (Cluster Computing'2005)*, Boston, USA, 26-30 September, 2005, pp. 1-10.

[78] J. Liu and D. K. Panda, "Implementing efficient and scalable flow control schemes in mpi over infiniband," in *Workshop on Communication Architecture for Cluster (CAC'2004)*, Santa Fe, New Mexico, USA, 26-30 April, 2004, p. 183.

[79] J. Liu, A. Vishnu, and D. K. Panda, "Building multirail infiniband clusters: Mpi-level design and performance evaluation," in *ACM/IEEE*

SuperComputer Confernece (SC'2004), Pittsburgh, USA, 6-12 November, 2004, p. 33.

[80] J. Liu, J. Wu, S. P. Kini, D. Buntinas, W. Yu, B. Chandrasekaran, R. Noronha, P. Wyckoff, D. K. Panda, and O. S. Center, "MPI over InfiniBand: Early experiences," *Network-Based Computing Laboratory Computer and Information Science, Ohio State University,* 2003.

[81] J. Liu, J. Wu, and D. K. Panda, "High performance RDMA-based MPI implementation over InfiniBand," *International Journal of Parallel Programming,* vol. 32, no.3, pp. 167-198, 2004.

[82] L. Massoulié and J. Roberts, "Arguments in favour of admission control for TCP flows," in *16th International Teletraffic Conference (ITC'1999)*, Edinburgh, UK, 7-11 June, 1999, pp. 33-44.

[83] C. Minkenberg, R. P. Luijten, F. Abel, W. Denzel, and M. Gusat, "Current issues in packet switch design," *ACM SIGCOMM Computer Communication Review,* vol. 33, no.1, pp. 119-124, 2003.

[84] H. Mouchos, A. Tsokanos, and D. Kouvatsos, "Performance modelling and traffic characterisation of optical networks," *Network performance engineering,* pp. 859-890, 2011.

[85] S. Narravula, H. Subramoni, P. Lai, R. Noronha, and D. K. Panda, "Performance of HPC middleware over InfiniBand WAN," in *37th International Conference on Parallel Processing (ICPP'2008)*, Portland, USA, 8-12 September, 2008, pp. 304-311.

[86] D. Pan and Y. Yang, "Credit based fair scheduling for packet switched networks," in *24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2005)*, Miami, USA, 13-17 March, 2005, pp. 843-854.

[87] D. Panda, S. Sur, and P. Balaji, "Designing High-End Computing Systems with InfiniBand and High-Speed Ethernet," in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC'2010)*, LA, USA, 13-19 November, 2010, pp. 125-127.

[88] I. Park and S. W. Kim, "Study of OpenMP applications on the InfiniBand-based software distributed shared-memory system," *Parallel Computing,* vol. 31, no.10-12, pp. 1099-1113, 2005.

[89] I. Park and S. Wook Kim, "The distributed virtual shared-memory system based on the InfiniBand architecture," *Journal of Parallel and Distributed Computing,* vol. 65, no.10, pp. 1271-1280, 2005.

[90] C. Parsa and J. Garcia-Luna-Aceves, "Improving TCP congestion control over internets with heterogeneous transmission media," in *7th International Conference on Network Protocols (ICNP'1999)*, Toronto, Canada, 31 October - 3 November, 1999, pp. 213-221.

[91] J. Pelissier, "Providing quality of service over Infiniband architecture fabrics," in *8th Symposium on High Performance Interconnects (Hot Interconnects'2010)*, 18-19 August, 2000.

[92] D. Pendery and J. Eunice, "InfiniBand Architecture: Bridge over troubled waters," *Research Note,* 2000.

[93] F. Petrini, E. Frachtenberg, A. Hoisie, and S. Coll, "Performance evaluation of the Quadrics interconnection network," *Cluster Computing,* vol. 6, no.2, pp. 125-142, 2003.

[94] W. Pfeiffer and N. J. Wright, "Modeling and predicting application performance on parallel computers using HPC challenge benchmarks," in *22nd IEEE International Parallel and Distributed Processing Symposium (IPDPS'2008)*, Shanghai, China, 21-25 May, 2008, pp. 1-12.

[95] G. Pfister, M. Gusat, W. Denzel, D. Craddock, N. Ni, W. Rooney, T. Engbersen, R. Luijten, R. Krishnamurthy, and J. Duato, "Solving hot spot contention using infiniband architecture congestion control," presented at the HP-IPC'2005, 2005.

[96] G. F. Pfister, *An introduction to the InfiniBand architecture*. High Performance Mass Storage and Parallel I/O: Technologies and Applications, 2001.

[97] G. F. Pfister, "An introduction to the InfiniBand architecture," *High Performance Mass Storage and Parallel I/O,* pp. 617-632, 2001.

[98] K. S. Puranik. (2003). *Advanced Switching Extends PCI Express*. Available: http://cdserv1.wbut.ac.in/81-312-0257-7/Xilinx/files/Xcell%20Journal%20Articles/xcell_47/xc_pcix47.pdf

[99] K. Ramakrishnan, S. Floyd, and D. Black, "The addition of explicit congestion notification (ECN) to IP," ed: RFC 3168, September, 2001.

[100] A. Ranadive, A. Gavrilovska, and K. Schwan, "Fares: Fair resource scheduling for vmm-bypass infiniband devices," in *10th Cluster, Cloud and Grid computing (CCGrid'2010)* victoria, Australia, 17-20 May, 2010, pp. 418-427.

[101] S. Reinemo, T. Skeie, T. Sodring, O. Lysne, and O. Torudbakken, "An overview of QoS capabilities in InfiniBand, advanced switching interconnect, and ethernet," *IEEE Communications Magazine,* vol. 44, no.7, p. 32, 2006.

[102] C. Reynolds, S. Winter, and Z. Lichtenberger, "Provisioning OpenCL capable infrastructure with infiniband verbs," presented at the 10th International Symposium on Parallel and Distributed Computing (ISPDC'2011), Cluk-Napoca, Romania, 6-8 July, 2011.

[103] S. Richling, H. Kredel, S. Hau, and H. G. Kruse, "A long-distance InfiniBand interconnection between two clusters in production use," in *High Performance Computing, Networking, Storage and Analysis (SC'2011)*, Washington, USA, 12-18 November, 2011, p. 15.

[104] G. Sachdeva, A. Patel, H. Kasim, and S. See, "Simulation of Infiniband Networks and Effciency Calculation on Non-blocking Fully Populated Fat-Tree Topology," in *6th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC'2011)*, Barcelona, Spain, 26-28 October, 2011, pp. 163-168.

[105] H. Saito and K. Shiomoto, "Dynamic call admission control in ATM networks," *Selected Areas in Communications, IEEE Journal on,* vol. 9, no.7, pp. 982-989, 1991.

[106] J. Sancho, A. Robles, J. Flich, P. Lopez, and J. Duato, "Effective methodology for deadlock-free minimal routing in InfiniBand networks," in *International Conference on Parallel Processing (ICPP'2002)*, Vancouver, Canada, 18-21 Auaust, 2002, pp. 409-418.

[107] J. C. Sancho, A. Robles, and J. Duato, "Effective strategy to compute forwarding tables for InfiniBand networks," in *International Conference on Parallel Processing (ICPP'2001)*, Valencia, Spain, 3-7 September, 2001, pp. 48-57.

[108] J. R. Santos, Y. Turner, and G. Janakiraman, "End-to-end congestion control for InfiniBand," in *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2003)*, San Franciso, USA, 30 March - 3 April, 2003, pp. 1123-1133 vol. 2.

[109] J. R. Santos, Y. Turner, and G. Janakiraman, "Evaluation of congestion detection mechanisms for InfiniBand switches," in *IEEE Global Telecommunications Conference (GLOBECOM'2002)*, Taibei, Taiwan, 17-21 November, 2002, pp. 2276-2280.

[110] U. Schwickerath and A. Heiss, "A First experience with the InfiniBand interconnect," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment,* vol. 534, no.1, pp. 130-134, 2004.

[111] T. Shanley and J. Winkles, *InfiniBand Network Architecture*: Addison-Wesley Professional, 2003.

[112] V. Shurbanov, D. Avresky, P. Mehra, and W. Watson, "Flow control in ServerNet® clusters," *The Journal of Supercomputing,* vol. 22, no.2, pp. 161-173, 2002.

[113] A. Snavely, L. Carrington, N. Wolter, J. Labarta, R. Badia, and A. Purkayastha, "A framework for performance modeling and prediction," in *ACM/IEEE conference on Supercomputing (Supercomputing'2002)*, Baltimore, USA, 16-22 November, 2002, pp. 21-21.

[114] H. Song, B. Kwon, and H. Yoon, "Throttle and preempt: A new flow control for real-time communications in wormhole networks," in *International Conference on Paralle Processing (ICPP'1997)*, Bloomington, USA, 11-15 August, 1997, pp. 198-202.

[115] H. Subramoni, P. Lai, M. Luo, and D. K. Panda, "RDMA over Ethernet—A preliminary study," in *Workshop on High Performance Interconnects for Distributed Computing (HPI-DC'2009)*, New Orleans, USA, 31 August, 2009, pp. 1-9.

[116] H. Subramoni, P. Lai, S. Sur, and D. K. Panda, "Improving Application Performance and Predictability using Multiple Virtual Lanes in Modern Multi-Core InfiniBand Clusters," in *39th International Conference on Parallel Processing (ICPP'2010)*, San Diego, USA, 13-16 September, 2010, pp. 462-471.

[117] S. Sur, M. J. Koop, L. Chai, and D. K. Panda, "Performance analysis and evaluation of Mellanox ConnectX InfiniBand architecture with multi-core platforms," in *15th Annual IEEE Sysposium on High-Performance Interconnects (HOT'2007)*, Stanford, USA, 22-24 Augest, 2007, pp. 125-134.

[118] S. Sur, Potluri, S., Kandalla, K.C., Subramoni, H., Panda, D.K., and Tomko, K., "Codesign for InfiniBand Clusters," *Computer* vol. 44, no.11, pp. 31-36, 2011.

[119] Y. Turner, J. R. Santos, and G. J. Janakiraman, "An approach for congestion control in InfiniBand," HP Laboratories Palo Alto, Internet Systems and Storage Laboratory2001.

[120] R. H. W.Baker, D.Sonnier, and W.Watson, "A Flexible Server Net-based Fault-Tolerant Architecture," presented at the 25th Flut-Tolerant Computing (FTCS'1995), Pasadena, CA, USA, June 27-30,1995.

[121] J. S. Wu, "Maximum entropy analysis of open queueing networks with group arrivals," *Journal of the Operational Research Society,* vol. 43, no.11, pp. 1063-1078, 1992.

[122] S. I. Y.C. Fang, A. Saify. (2005). *Designing High-PerformanceComputing Clusters*. Available: http://www.dell.com/downloads/global/power/ps2q05-20040181-Fang-OE.pdf

[123] W. Yu, N. S. V. Rao, and J. S. Vetter, "Experimental analysis of infiniband transport services on WAN," in *International Conference on Networking, Architecture, and Storage (NAS'2008)*, Chognqing, China, 12-14 June, 2008, pp. 233-240.

[124] K. H. Yum, E. J. Kim, and C. R. Das, "QoS provisioning in clusters: an investigation of router and NIC design," presented at the Proceedings. 28th Annual International Symposium on Computer Architecture, Goteborg, Sweden, 20 June - 04 July, 2001.

[125] K. H. Yum, E. J. Kim, C. R. Das, M. Yousif, and J. Duato, "Integrated admission and congestion control for QoS support in clusters," in *IEEE International Conference on Cluster Computing (CLUSTER'2002)*, Chicago, USA, 23-26 September, 2002, pp. 325-332.

[126] H. Zhang, W. Huang, J. Han, J. He, and L. Zhang, "A performance study of Java communication stacks over InfiniBand and giga-bit Ethernet," in *IFIP International Conference on Network and Parallel Computing Workshops (NPC'2007)*, Dialian, China, 18-21 September, 2007, pp. 602-607.

[127] S. Zhong, J. Chen, and Y. R. Yang, "Sprite: A simple, cheat-proof, credit-based system for mobile ad-hoc networks," in *22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'2003)*, San francisco, USA, 30 March - 3 April, 2003, pp. 1987-1997.