面向 InfiniBand 数据中心的区别化传输层带宽划分机制

张子文 吕高峰 孙志刚 王琨珺 卢锡城

(国防科学技术大学计算机学院 长沙 410073)

摘 要 数据中心已经成为提供资源和服务的重要平台,在虚拟化基础上部署的多种应用对网络性能提出了更加严格和多样化的需求,从而促使改变以传统 TCP 为中心的传输层结构.文中从数据中心环境的特性出发,归纳得到改进数据中心网络传输层设计的3个方向.在深入剖析 InfiniBand(IB)互联传输优化相关技术的基础上,文中提出了一种更好的支持网络可感知、应用可区分、策略可管理的数据中心网络新型传输控制模型.文中对 IB 传输控制核心的 IB 拥塞控制算法及其参数进行了深入的分析,证明了在流级通过区别化的拥塞控制参数配置能够实现成比例的带宽划分.最后,在真实系统中的实验验证了通过参数的动态配置能够保持租户级的公平带宽共享,更好支持云数据中心的带宽隔离需求.

关键词 数据中心网络; InfiniBand; 拥塞控制; 带宽隔离; 传输策略

中图法分类号 TP393

DOI号: 10.3724/SP.J.1016.2012.02505

Differentiated Transport-Layer Network Bandwidth Allocation for InfiniBand Datacenter

ZHANG Zi-Wen LV Gao-Feng SUN Zhi-Gang WANG Cheng-Jun LU Xi-Cheng

(School of Computer, National University of Defense Technology, Changsha 410073)

Abstract Datacenters have become the important infrastructure to provide resource and service. The different applications based on virtualization require more rigid and diverse network performance guarantee which motivates us to change traditional TCP-based transmission mechanism in datacenter networks (DCN). After introducing the inherent characteristic of DCN, we concluded three kinds of insight of the improvement to DCN transport layer. Based on the elaboration of key techniques in IB (InfiniBand) transmission optimization we propose the new-type transmission control model for DCN with the advantage of observable network, differentiated application services as well as administrable policies. This paper analyze IB Congestion Control (CC) algorithm as the core transmission control in IB and its related parameters. The proportional bandwidth allocation at the flow level is proved based on the differentiated CC parameter settings. In the end, real IB platform is established to verify that tenant-level fair bandwidth share can be maintained based on the dynamic parameter configuration. It obviously supports the bandwidth isolation demand in cloud datacenter.

Keywords datacenter networks; InfiniBand; congestion control; bandwidth isolation; transmission policy

收稿日期:2012-05-07;最终修改稿收到日期:2012-07-13. 本课题得到国家“九七三”重点基础研究发展规划项目基金(2009CB320503)资助. 张子文,男,1984年生,博士研究生,主要研究方向为数据中心网络、高性能互联. E-mail: ziwen@nudt.edu.cn. 吕高峰,男,1980年生,博士,助理研究员,主要研究方向为网络安全、SDN. 孙志刚,男,1973年生,博士,研究员,主要研究领域为下一代互联网体系结构、高性能互联. 王琨珺,男,1983年生,博士研究生,主要研究方向为拥塞控制算法、DTN网络. 卢锡城,男,1946年生,教授,中国工程院院士,主要研究领域为计算机网络和并行分布处理.

1 引言

数据中心通过把计算、存储和网络设备整合成虚拟资源池为用户提供集成服务环境. 国内外服务提供商建立的云平台发展迅猛, 云数据中心规模日益扩大, 已经达到 10 K~100 K 台服务器的量级^[1]. 数据中心网络主要设计目标为高带宽、低延时和高可靠性. 其它如扩展性、安全性、管理配置、故障诊断等方面也面临着诸多亟待解决的问题.

传统 TCP 协议在解决低延时数据中心网络拥塞控制、满足多样化用户需求、实现可控可管的传输等方面仍存在较多问题, 改进传输控制必须考虑数据中心环境的内在特性和需求. 首先, 传统 TCP 只能实现流级的公平性, 而数据中心内应用规模相对有限, 需要以较细的粒度将各自的 QoS 需求传达给网络管理者, 从而实现相应的传输控制策略和灵活的网络资源划分, 同时促进网络定价机制的发展. 第二, 不同于 TCP 基于端到端 RTT 和丢包实现网络探测, 由于网络和服务端处在同一管理域, 数据中心能够直接将网络信息通告给传输层, 如交换机接口的统计计数器^[2]等, 网络对传输层的接口能够优化传输控制. 第三, TCP 应用于自组织的 Internet 环境, 集中管理的数据中心需要更好的可控性, 网络管理者能够为用户配置不同的策略, 同时监控传输质量, 从而启动流量调度等传输优化机制.

IB 是诞生于 1999 年的一种高带宽低延时互连交换技术^[2], 其与以太网相比具有更强的异构互联能力, 在延时、带宽、虚拟化和管理配置等方面其性能都有显著的提高. 2011 年底最新的 top500 高性能计算机排名中, 全球最快的 500 台计算机中已经有 42.6% 使用 IB 互连. 近年来, IB 在数据中心网络中的应用逐渐得到关注. 2010 年, Oracle 通过在其数据仓库和数据处理平台 Exadata 中使用 IB 交换, 处理性能提升 10 倍^[3]. 2010 年, Google 在其研究工作中采用 IB 和蝶形网络拓扑构建数据中心, 网络功耗节约 85%^[4]. 2011 年, Microsoft 基于 IB 组建高性能数据中心支持其 Bing Maps 应用, 与传统基于以太网的架构相比节约 80% 的功耗和 50% 的成本^[5]. Stanford 课题组通过 IB 互联分布式 DRAM 实现高性能数据中心存储系统^[6]. 2012 年初, Intel 通过收购试图将 IB 技术与其至强处理器绑定, 进一步扩大在服务器市场的优势地位, 推动其在数据中心集群的应用. IB 的传输控制较之传统 TCP 加以以太网模式更加适应数据中心环境, 本文对基于 IB 的

数据中心传输层设计展开深入讨论.

本文的创新点在于: (1) 将目前数据中心网络的主要研究点归纳映射为如何扩展和改进传输层功能以满足数据中心网络传输的设计需求; (2) 基于 IB 提供的传输优化技术, 创新性地提出新型传输控制机制, 通过定义各功能模块和接口实现网络可感知、应用可区分、策略可管理的数据中心网络传输; (3) 基于 IB 的拥塞控制机制, 通过对其相关参数的研究和真实环境的实验证明通过不同的参数配置能够达到区别化的端到端网络传输能力, 从而实现管理者的传输控制策略.

本文第 2 节综述目前改进传输层设计的相关研究; 第 3 节提出数据中心网络新型传输控制模型; 第 4 节深入研究 IB 拥塞控制算法及其参数配置对流传输能力的影响; 第 5 节通过实验验证 IB 拥塞控制的公平性及区别化带宽划分的有效性.

2 数据中心网络传输层设计相关研究

早期数据中心网络研究普遍关注于从体系结构的拓扑、编址、路由等角度解决面临的问题, 而近期的很多工作都是围绕着如何设计传输控制机制来满足数据中心传输层的新应用需求. 针对数据中心的应用环境, 其传输层设计的主要需求包括:

(1) 新型拥塞控制算法

完全基于端到端的 TCP 方法面临着在低延时数据中心环境下对突发拥塞响应不及时、窗口限速能力不足^[7]等问题. 目前数据中心应用具有很强的流量突发性, 如大数据领域的 Hadoop 会产生 all-to-all 的流量模式^[1], many-to-one 的同步流量场景会造成 Incast^[8-9]问题. 出于成本的考虑, 数据中心网络通常采用超额认购 (Oversubscription)^[1]的方式, 从而导致核心网络成为瓶颈. 交换网络的报文缓冲区比较小也导致更容易出现拥塞^[10]. 目前的很多研究通过基于网络拥塞信息加源限速的方式进行传输拥塞控制. DCTCP^[10]针对数据中心的需求修改 ECN (Explicit Congestion Notification) 机制, 采用一个多比特的反馈, 同时改进 TCP 发送接收端的 AIMD 速率调整算法, 从而保证高的突发容忍性和短流的低延时, 同时能够解决 Incast 问题. 以太网通过 IEEE 802.1Qau 的标准实现了 QCN (Quantized CN), Jiang 等人^[11]对数据中心内 QCN 的稳定性和收敛性进行建模分析, 结果显示 QCN 能够改进吞吐量和延时, 同时得到比例公平性, 但是网络抖动问题比较严重. Alizadeh 等人^[12]采用了一种当

出现延时增加时能够使得拥塞控制环回稳定的算法 Averaging Principle, 证明该策略在 QCN 以太网能够达到高稳定性。

(2) 传输层安全和性能隔离

由于传统 TCP 协议能够保证流级的公平性, 故数据中心内的应用和租户能够通过建立大量 TCP 连接抢占有限的带宽资源。虚拟机提供带宽设限机制解决性能隔离, 当速率超过分配的带宽时丢弃报文, 从而导致 TCP 的降速, 但存在的问题是固定的带宽分配阈值导致带宽浪费, 同时应用需要假设租户使用的是合理行为的 TCP。SecondNet^[13] 通过集中式的 VDC 分配算法确定虚拟机到物理机的匹配、路由以及服务器管理程序对带宽的预留状态, 从而在数据中心网络保证了虚拟机对之间的流量。Seawall^[14] 根据每台虚拟机在当前链路发送数据的权重, 通过控制边缘网络服务器虚拟层的分布式限速器为每条链路提供带宽划分, 可扩展性好。但是这两种机制提供虚拟机对之间的带宽更有利于虚拟机对数多的租户; 且数据中心通信模式动态性强, 很难指定虚拟机对之间的带宽需求。NetShare^[15] 实现了各个租户之间的加权层次带宽划分, 将问题简化为端点的服务器链路在各个共享的虚拟机之间的分配, 但是集中的带宽分配器难适应大规模的数据中心、处理快速的速率变化和虚拟机的迁移。

(3) 传输服务的应用区分能力

在传输层面除了传统的公平带宽性能隔离, 目前越来越多的研究关注于如何为应用提供更细粒度的差异化带宽划分, 尤其是在资源竞争的场景下。近两年来的典型研究包括, AF-QCN^[16] 在维持 QCN 简单性、实时响应能力和稳定性特点的基础上, 实现了灵活的流或者流量类的可编程带宽划分。Fang 等人^[17] 提出数据中心网络的 ECN 主动队列管理和基于 AIMD 的限速器, 通过区别的 AIMD 参数设置能够实现不同优先级的拥塞控制区分, 从而达到区别的带宽利用率控制。Wilson 等人^[18] 认为流的公平性会损害应用的性能, 提出了一种在数据中心环境下 deadline 感知的控制协议 D³。D³ 使用显式速率控制为尽快能够结束的任务提供更高的带宽, 从而最终优化在 deadline 之前完成的总任务数。Vattikonda 等人^[19] 采用时分多路复用 (TDMA) 来解决区分服务, 通过在商业交换机硬件实现的 MAC 层, 以 TDMA 的方式切分流的链路带宽和交换机缓存空间, 使得主机能够不需要 TCP 的可靠性和拥塞控制就能够满足带宽、延时等不同的应用需求。

此外, 传统的 TCP 协议无法利用数据中心网络

多路径所提供的负载均衡能力, 且应用传输对于网络状态的变化缺乏感知能力。

在 IB 网络传输控制方面, Santos 等人^[20] 最早对其 ECN 机制进行研究, 提出异步的报文标记行为能够改进公平性, 放宽的速率增加条件能够改进静态和动态流量场景下的吞吐量。Pfister 等人^[21] 通过模拟验证 IB 拥塞控制如何在胖树结构下解决热点链路的拥塞问题。Yan 等人^[22] 提出在 IB 中改进 ECN 报文标记机制以及新的源响应幂增幂减函数, 通过带窗口限制的速率控制减少 IB 网络中有多种优先级类流量的拥塞。Yan 等人^[23] 又提出基于作业阈值的拥塞控制方法, 一方面能够减少延时、阻塞概率和平均队列长度, 同时提高 IB 网络的 QoS。Lugones 等人^[24] 通过自适应路由算法将流量分配到 IB 中的多条可选路径以解决拥塞问题, 实验验证了该机制同时改进了延时和吞吐量。Gran 等人^[7] 对硬件实现的 IB 拥塞控制进行分析实验, 通过调整参数能够有效地解决拥塞, 提高 benchmark 性能, 并通过解决停车位问题改进公平性。Gran 等人^[25] 还研究了在不丢包网络中拥塞控制、交换机仲裁和公平性之间的关系。Escudero-Sahuquillo 等人^[26] 通过限制流量注入和隔离拥塞流这两种方法的结合实现对拥塞的快速反应, 该机制可扩展性好且对拥塞流公平。

3 IB 网络传输控制模型

我们认为除了网络本身性能的优势, IB 中已有的关键技术对数据中心网络扩展传输功能已经提供了各种机制上的支持, 满足数据中心网络传输对新型拥塞控制算法、传输性能隔离和传输服务的应用区分能力等要求。其中, QP (Queue Pair) 是传输层承上启下的核心, IB 面向每个单向连接分配发送/接收者对 QP 作为虚拟化的通信端口。与 TCP/IP 协议相比, 其传输层是基于硬件实现的, 且完全在用户模式下运行, 从而减少了通信协议本身和操作系统引入的开销。QP 在建立时和 QP 号、CA (Channel Adapters) 源/目的端口地址 LID (Local IDentifiers)、SL (Service Level) 和 MTU 等信息绑定, QP 在通信结束后销毁。与 TCP 连接不同, IB 采用基于目的 LID 的确定性路由, QP 在建立时事实上已经确定了一条源和目的节点之间的路径。

如图 1 所示, 基于 IB 实现的传输控制模型可以分为用户和网络管理两大部分, 用户根据应用需求建立、绑定和销毁 QP, 网络管理负责 QP 的参数及

其它区分化传输服务的配置接口,同时对 QP 传输质量进行监控,以动态地触发 QP 拥塞控制和应用的多 QP 流量调度.我们根据目前 IB 中已有的传输

机制提出了其传输控制模型,从网络、传输和应用三者的角度扩展原有机制,该模型提供对以下 3 个方面的支持:

图 1 IB 网络传输控制模型

(1) 感知多路径和网络状态的传输.

首先,IB 网络的每个端口能够虚拟化出 2^{LMC} (LID Mask Control) 个不同的 LID,在转发的时候对应 2^{LMC} 条不同的路径,如图 1 中 p_1, p_2 分别对应 A、B 之间数据传输两个 QP 连接的不同路径. IB 的子网建立过程,集中控制器发送探测数据报文,通过节点和端口逐步的自我链式反应发现网络拓扑并分配 LID 地址,这使得应用能够通过管理者的接口获得网络拓扑及路由信息,如应用 n 能够感知两条传输路径. 多路径提供了更加灵活的流量调度,每个应用能够利用端口的多个 QP 进行传输,如 p_1 发生拥塞时,ECN 反馈会使得 p_1 传输流量下降,而更多地切换到 p_2 上,通过同时多路径的动态负载均衡实现了带宽分配下的可靠网络传输.

显式拥塞通告 ECN 为端到端传输本身提供了感知网络拥塞的能力,通过对报文的标记和拥塞反馈使得源发送者能够了解拥塞的发生及程度,根据相应的机制降速以避免拥塞瓶颈点. 此外,IB 中还提供了大量关于网络状态的信息,如接口流量的报文计数器,主要是通过 SA(Subnet Administration) 建立的子网数据库存储子网的状态信息. SA 还通过路径记录的数据结构维护路径信息. 当端点试图建立一个 QP 时需向 SA 发送请求获得该路径信息,还可以通过 SA 与 QoS 管理器的交互得知路径的 QoS 是否能够满足需求.

(2) 新型传输模式和拥塞控制机制.

数据中心应用的性能要求越来越高,端节点开销已经成为通信的瓶颈,IB 通过无连接(数据报)/面向连接的传输和可靠/不可靠传输的 4 种组合定

义发送和接受 QP 这 4 种不同交互模式. 应用在建立 QP 之前首先选择适合的传输模式,IB 中硬件支持的传输模式与 TCP 和 UDP 相比能够提供更高的传输性能.

IB 拥塞控制采用显式拥塞标记的 ECN 机制, IB 交换机通过监测是否超过阈值感知拥塞,避免了传统 TCP 协议在 IB 这种极低延时网络中对延时反应不及时的问题. 当报文到达目的节点的路径上遇到拥塞(如图 1 IB 交换机 R),其报文 FECN(Forward ECN)位会进行标记,在接收端端口感知拥塞标记之后会通过 ACK 或者 CNP 报文中的 BECN (Backward ECN)位通知相应的源端,使其能够发现到前面发送的报文在该 VL 下遇到了拥塞. IB 采用 PFC(Priority Flow Control),在每个优先级,如果上游的邻居交换机没有接收缓冲区会发送控制帧停止本地交换机数据发送,从而避免丢包对于应用性能的严重影响. 但是链路层信用量的耗尽会造成头阻塞问题,进而造成拥塞传播,基于 ECN 的拥塞控制机制在数据中心环境下实现传统公平、收敛的拥塞控制同时能够避免拥塞传播对于无辜流的影响^[7,25].

(3) 基于 QoS 管理的区别化传输服务策略配置.

目前,IB 中尚未定义如何支持区分的应用服务策略及传输配置,IB 提供集中的子网管理结构,能够通过在其中增加资源分配功能实现高级的应用优化目标. 从数据中心应用自身流量特性和不同的 QoS 需求出发要求传输层协议提供区别的服务策略. 从用户之间的竞争关系出发,它们共享同一个数据中心网络资源,某些用户可能通过建立大量连接

等方式恶意侵占网络带宽,传输层需要实现用户之间的性能隔离.从应用的相对优先级出发,数据中心通过虚拟化为多个租户提供服务平台,区别收费需要传输层为租户提供不同质量的网络服务.

如图 1 所示,IB 协议中已经定义的 QoS 管理器实现了针对 SA 中定义的网络和路径监控信息的主动控制.当服务质量达不到时,QoS 管理调整配置和策略以满足所需要的服务,如通过与子网管理器的交互调整某个端口的 VL(Virtual Lanes)仲裁表.这种主动的调整将网络状态与资源分配结合起来,决定最终的带宽分配策略并配置到拥塞控制和流量调度模块,形成了一个闭环的动态控制系统.

IB 提供了两种传输质量的配置接口,首先在网络层面,每个物理链路上虚拟化多个 VL,代表链路级的发送接收缓冲区对,每个 VL 有各自的流控策略和加权优先级,采用高低两级抢占式调度,而同一优先级的 VL 能够进一步区分调度,从而获得不同的带宽性能.如图 1 所示,链路级的性能隔离是通过 SL 到 VL 的匹配策略以及 VL 基于不同加权值的仲裁调度来实现的.在用户节点的端到端传输层面,IB 的 QP 为应用实现了抽象的虚拟通信接口,是应用竞争网络资源以及进行拥塞控制的实体,每个 QP 的 CCT(Congestion Control Table)是拥塞控制的核心,如何通过 CCT 表相关参数的配置实现区分的应用服务策略是下一节的研究重点.

综上,IB 协议定义了比传统以太网加 TCP/IP 模式更加强大的传输层功能,对实现网络可感知、应用可区分、策略可管理的新型传输控制模型供了强大的支持.

4 基于拥塞控制的区别化 IB 带宽划分

4.1 IB 拥塞控制算法

源端的限速策略依赖于每个 QP 的 CCT 表,其中可以配置的相关参数如表 1 所示,CCT 的表项为 *IRD*(Injection Rate Delay),它定义了该 QP 向网络中注入报文的时间间隔.表项 *IRD* 在 CCT 表中按照由小至大进行排列,当发送端收到 BECN 时,增加 CCT 表的指针 *CCTI*(CCT Index)指向更大的 CCT 表项,从而降低报文发送速率.每次索引 *CCTI* 增加的单位值是 *CCTI_Increase*,该值事实上实现了类似 TCP 在面对拥塞时的快速回退策略,且通过配置该参数能够控制 QP 发送速率回退的快慢程度.为了限制速率变化范围,*CCTI* 定义了下界

CCTI_Min 和上界 *CCTI_Limit*,*CCTI* 初始化之后指向 *CCTI_Min*,*CCTI_Min* 的缺省值为 0,即不对报文发送速率做任何限制,而 *CCTI_Limit* 指向的表项是报文发送的最大间隔即速率的最小值.

表 1 QP 的 CCT 表相关参数配置

CCT 参数	定义	推荐值
<i>CCT Item</i>	报文间隔 <i>IRD</i>	/
<i>CCTI_Increase</i>	<i>CCTI</i> 每次增加的单位量	1
<i>CCTI_Limit</i>	<i>CCTI</i> 的上界	127
<i>CCTI_Min</i>	<i>CCTI</i> 的下界	0
<i>CCTI_Timer</i>	恢复速率的超时定时器	150

拥塞控制的另一个问题是在拥塞缓解之后如何通过提高发送速率来充分利用网络带宽资源,其中发送速率的增加是基于超时机制,当在 *CCTI_Timer* 时间内没有收到拥塞控制报文,那么相应 QP 流的 *CCTI* 值减少 1,相应的 *IRD* 值得到增加,提高了源发送速率,直到 *CCTI* 减少到 0 不限制报文注入速率.*CCTI_Timer* 参数的设置事实上决定了流量的波动快慢,过小的 *CCTI_Timer* 会造成流量的波动,反之如果偏大则使得拥塞控制流量快速下降之后难以及时恢复,CCT 管理配置算法如下.

算法 1. CCT 管理.

输入: CCT Configuration MAD, Packet with BECN
输出: Updated *IRD*

1. Initialize: Manager defines CC configuration file
2. SM parses it and send MAD to configure CC parameter in each QP: *CCT*, *CCTI_Increase*, *CCTI_Limit*, *CCTI_Min*, *CCTI_Timer*
3. Start *Timer*=0
4. if receive BECN && *CCTI*<*CCTI_Limit* then
5. *CCTI*=*CCTI*+*CCTI_Increase*
6. *IRD*=*CCT*[*CCTI*]
7. *Timer*=0
8. endif
9. if *Timer*=*CCTI_Timer* && *CCTI*>*CCTI_Min* then
10. *CCTI*=*CCTI*-1
11. *IRD*=*CCT*[*CCTI*]
12. *Timer*=0
13. endif
14. for every time unit
15. *Timer*=*Timer*+1

源端的限速是累积一定数量的报文(如 IB 规范中定义 4 个^[2])进行报文间隔计算,如果报文的间隔小于 *CCT*[*CCTI*],则会在判定后等待发送,从而保证源的发送速率不超过 CCT 表的定义.

算法 2. 源发送限速.

输入: IRD

输出: Controlled packet transmission

```

1. Initialize: arb_pkt_num=0
2. last_time=0
3. While(packet_to_send)
4. arb_pkt_num++
5. if (arb_pkt_num=aggregate)
6. if(CCT[CCTI]<(current_time-last_time))
7. next aggregate packets transmission starts
8. else
9. Wait for CCT[CCTI]-(current_time-last_time) to transmit next aggregate packets
10. last_time=current_time
11. arb_pkt_num=0
12. endif
13. endwhile

```

对于 FECN 标记,交换机主要有两个参数,一个是标记的阈值,另一个是标记报文的比率.根据 IB 规范的定义,IB 交换机的拥塞阈值从 0~15,0 表示没有拥塞标识,1 表示非常高的阈值,拥塞容易发生传播,而 15 代表非常低的阈值,代表即使互联拥塞程度不高也可能发生标记降速,降低拥塞传播的可能性.其中给出的参考是 15 表示 256 Bytes 的报文,报文之间的间隔为 176 Bytes,从而达到 60% 的带宽占用率.权重从 15~1 表示从平均 60% 的带宽占用以上的线性增加,阈值设置与交换机体系结构相关.

4.2 区别化 IB 带宽划分

相关研究^[7,25]证明在 IB 网络中若采用相同的 CCT 参数配置,各个 QP 之间能够达到很好的流级公平性,即各个 QP 应当处于相同的 CCTI 值.而通过对每个 QP 配置不同的 CCT 相关参数能够实现区别化的传输服务质量.

定理 1. 基于 ECN 的 IB 拥塞机制,对于任意 qp_i 和 qp_j ,通过设置 $CCTI_Increase_i = N \times CCTI_Increase_j$,发生拥塞时两条流稳定之后的发送窗口 $w_{i,k} : w_{j,k} = 1 : N$.

证明. 对于当前热点链路上的各条流 $i(i=1, 2, 3, \dots, n)$ 在时间周期 k 上的报文最小发送间隔限制为 $CCT[CCTI_{i,k}] = IRD_{i,k}$. 设 Δt 周期足够短使得可以认为带宽为某个固定值,第 k 个周期的时间 $t_k = k \times \Delta t$ ($k=0, 1, \dots, m-1$). 设平均报文长度为 l , qp_i 在时间 k 的发送窗口大小是 $w_{i,k} = l/CCT[CCTI_{i,k}]$.

设 α 为交换机采取 FECN 的标记报文比例,对于接受端的 CA 网卡,每 C 个报文采用 ACK 或者

发送 CN 实现 BECN 携带拥塞标记信息,则可得

$$\Delta CCTI_{i,k} = CCTI_Increase_i \times BECN_{i,k},$$

$$BECN_{i,k} = (\alpha \times w_{i,k}) / (C \times l),$$

$BECN_{i,k}$ 代表 QP 源收到的 BECN 报文数, $\Delta CCTI_{i,k}$ 是在时间周期 k 之后 CCTI 的变化量.下一周期的发送窗口 $w_{i,k+1} = l/CCT[CCTI_{i,k} + \Delta CCTI_{i,k}]$.

对于 $\forall qp_i, qp_j$, 如果 $w_{j,k} < N \times w_{i,k}$, 由于 $CCTI_Increase_i = N \times CCTI_Increase_j$,

$$\frac{\Delta CCTI_{i,k}}{\Delta CCTI_{j,k}} = \frac{CCTI_Increase_i \times BECN_{i,k}}{CCTI_Increase_j \times BECN_{j,k}} > 1 \quad (1)$$

满足 $l/CCT[a] - l/CCT[a+x] > l/CCT[b] - l/CCT[b+x]$ ($a < b$) 表示 CCT 表前端的表项 CCT[a] 导致的限速变化幅度大于后端的表项 CCT[b] 导致的限速变化,由于 $CCTI_{i,k} < CCTI_{j,k}$, 则 $l/CCT[CCTI_{i,k}] - l/CCT[CCTI_{i,k} + \Delta CCTI_{i,k}] > l/CCT[CCTI_{j,k}] - l/CCT[CCTI_{j,k} + \Delta CCTI_{j,k}]$, 由于式(1)中 $\Delta CCTI_{i,k} > \Delta CCTI_{j,k}$, 可得

$$\frac{l}{CCT[CCTI_{i,k}]} - \frac{l}{CCT[CCTI_{i,k} + \Delta CCTI_{i,k}]} > \frac{l}{CCT[CCTI_{j,k}]} - \frac{l}{CCT[CCTI_{j,k} + \Delta CCTI_{j,k}]} \Rightarrow \Delta w_{i,k} > \Delta w_{j,k},$$

对于 $w_{j,k} > N \times w_{i,k}$ 的情况,与式(1)同理 $\Delta CCTI_{i,k} / \Delta CCTI_{j,k} < 1$, 可证 $\Delta w_{i,k} < \Delta w_{j,k}$.

这使得 qp_i 和 qp_j 有不同的降速快慢直到它们稳定在汇聚点 $w_{j,k} = N \times w_{i,k}$. 证毕.

定理 1 说明各个流通过呈比例的 CCTI_Increase 配置能够达到成比例的带宽划分.

如果链路上的总流量小于带宽 B , 那么对于流 i , 每经过 T_i 超时周期之后流 i 的 $CCTI_i$ 会减 1. S 为拥塞控制的阈值,而 s 表示超时增速开始时的缓冲区长度.超时的速率恢复过程如下式所示:

$$\sum_{i=1}^n \sum_{j=0}^{T/T_i-1} \frac{l \times T_i}{CCT[CCTI_i - j]} - B \times T = S - s \quad (2)$$

该速率增加过程使得两条流 i, j 的报文间隔比从 $CCTI[CCTI_i]/CCTI[CCTI_j]$ 变为 $CCTI[CCTI_i - (T/T_i - 1)]/CCTI[CCTI_j - (T/T_j - 1)]$, 当为 CCTI 较大流设定短的超时时间能够较好地维护原有的带宽比例关系.能够通过式(2)求出经过的超时增速时间周期.

除了 CCTI_Increase 外, CCT 表指针 CCTI_Min 和 CCTI_Limit 之间的值表示了报文间隔的变化范围,是实现区别化配置的重要参数. $CCT[CCTI_Min]$ 事实上代表报文之间的最小间隔,如 $CCT[0]$ 默认

值为 0, 说明当前没有任何限速, 其它情况下发送的最高速率为 $l/CCT[CCTI_Min]$, 而发送的最低速率控制在 $l/CCT[CCTI_Limit]$. $CCTI_Min$ 和 $CCTI_Limit$ 为流设定的阈值, 导致流的降速和升速受到限制. 通过设定该值能够限制当前 QP 吞吐量的变化范围, 从而达到区分化的带宽分配. 如果 CCT 表的表项配置相同, 可以得到以下结论.

对于 $\forall qp_i, qp_j$, 若其它参数相同则满足以下条件:

(1) 如果 $CCTI_i \leq CCTI_Min_j$ 则 $CCTI_j = CCTI_Min_j$.

(2) 如果 $CCTI_i \geq CCTI_Limit_j$ 则 $CCTI_j = CCTI_Limit_j$.

(3) 其它, 则 $CCTI_j = CCTI_i$.

$CCTI_Min$ 实现对每个网卡的限速阈值, 如 $CCTI_Min$ 设置为 1 能够使得流初始的发送间隔指向 CCT 表第 2 项, 这样即使不发生拥塞, 报文发送的间隔也被固定地限制在最小, 为 $CCT[1]$, 从而得到网卡限速的大小. 分区段的 CCT 配置事实上实现了静态的报文发送速率范围.

以上分析了为每个 QP 配置不同的拥塞控制参数(如 $CCTI_Increase$)能够达到区别化的端到端网络带宽资源划分, 从而实现传输性能隔离和基于应用层需求的传输机制等目标, 如根据各个租户的加权进行成比例的分配, 即某个租户无论建立多少流, 它所获得的带宽应当是对于当前链路的合理加权划分. 如下所示, 对于瓶颈链路上存在的租户 t , 其获得的带宽 B_t 为

$$Bw \times \frac{\omega_t}{\sum_t \omega_t} = B_t = \sum_j \frac{l}{CCT_j[CCTI_j]}.$$

当前该链路上租户 t 所有流的带宽求和应当满足其带宽的分配.

5 初步实验结果

为了对 IB 在数据中心网络中应用进行研究, 验证基于拥塞控制的区别化网络带宽分配, 我们搭建了真实的 IB 实验环境.

IB 交换机 IBS108Q 和 HCA 网卡是由国防科技大学课题组基于 Mellanox 芯片 SwitchX 和 ConnectX 自主设计研发的, 端口支持最高速率 QDR (40 Gbps), 设备均支持拥塞控制. 实验节点包括 6 台宝德 PowerLeader PR1760T 服务器, 服务器配置为 Intel Xeon E5540 双核 CPU、16 GB ECC RAM、2 个 8x PCIe 2.0 插槽、400 GB 硬盘, 所有服务器均运

行 RedHat Linux 6.1 x86 64 位操作系统, 其上采用开源 OFED(Open Fabrics Enterprise Distribution) 1.5.4 驱动和管理 HCA 设备.

OFED 的最新诊断包 infiniband-diags-1.6.0 提供了 `ibccquery` 和 `ibccconfig` 接口来读写拥塞控制参数. 我们主要利用 `ibccconfig` 来根据不同 QP 连接配置各自的 $CCTI_Increase$ 参数. IB 拥塞控制实验场景如图 2 所示, 我们设计实现了一个集中控制的流量行为发生器, 能够控制其它分布式的客户端建立 TCP 连接, 并根据来自集中控制器的命令向目的端发送报文, 类似 `iperf` 以尽可能大的速率发送报文. 由于 TCP 反应较慢, 在 IB 拥塞机制工作的情况下, TCP 拥塞控制事实上不发挥作用^[7]. 集中控制器通过一个配置文件, 设置客户端执行的命令, 如连接开始/持续的时间, 被控客户端/连接目的端的地址. 链路实际工作速率均设置为 10 Gbps.

图 2 IB 拥塞控制实验场景

在第 1 个实验中, 为了验证 QP 的公平性, 由 4 台服务器向服务器 C 发送报文, 最后一跳成为瓶颈, 该模式符合目前在数据中心网络中常见的 Incast^[8-9] 问题场景. QP 流每隔 1 s 建立连接, 持续 5 s 时间, 按照 100 ms 的粒度统计吞吐量. 如图 3(a), 如果拥塞控制关闭, 由于著名的停车位问题^[27-28], 从 3 s~5 s, 流量 $A \rightarrow C$ 加 $B \rightarrow C$ 的和等于 $D \rightarrow C$ 或者 $E \rightarrow C$ 的流量. 直到 $A \rightarrow C$ 传输结束才达到公平共享. 如图 3(b), 拥塞控制打开后虽然导致吞吐量的小幅抖动, 但是使得各条流能够很好地同步增速和减速以达到带宽的公平共享.

为了测试 IB 中 CCT 表配置对于 QP 连接限速的影响, 实验中将某条 QP 连接的 $CCTI$ 指向 1, 通过设置固定的不同 CCT 表项 IRD 值, 实现了不同速率的限制. 由于每次发送端聚合 4 个 1024 Bytes 的报文进行 IRD 比较, 如 $20.48 \mu s$ 的表项值, 带宽限制在 $(1024 \times 8) / (20.48 \times 4 \times 10^{-6}) = 100 \text{ MB}$, 同理 $IRD = 204.8 \mu s$ 将带宽限制在 10 MB, 而 $IRD = 2.048 \mu s$ 的带宽为 1 GB. 如图 4 所示, CCT 对于发送速率限制的效果比较明显.

图 3 相同配置下 QP 带宽的公平性

图 4 CCT 表的设置对于带宽限速的影响

第 2 个实验的目标是根据第 4.2 节的定理和分析,通过为各个 QP 配置不同的 $CCTI_Increase$ 参数实现租户级的公平带宽共享,即对于 S1 和 S2 之间的瓶颈链路,两个租户无论建立多少条流都应当达到公平的带宽占用.实验中,租户 1 有一个稳定的从服务器 $C \rightarrow A$ 的 QP 连接,而租户 2 的 QP 流顺序地加入和退出网络(租户 2 只有 $D \rightarrow B$ 的流存在于整个实验的时间区间).实验中 CCT 表项采用参考设计值 $CCT(i) = 14 \times i^2 / 106^2 (\mu s)$,可证该值满足第 4.2 节定理中 CCT 表前端的表项导致的限速变化幅度大于后端的表项.

我们在集中控制器上定义了参数配置的时间点和值,初始化时所有流的 $CCTI_Increase = 1$,当租户 2 有两条流时, $CCTI_Increase = 2$,当流的数目增加为 3 时, $CCTI_Increase = 3$,随着流的离开再次修改该参数.如图 5(a)所示,从 C 到 A 的带宽基本稳定在 3.4 Gbps,没有受到租户 2 流加入和退出的影响.随着流 E 和 F 的加入,租户 1 应得的带宽在 2 和 3 条流之间公平的共享,该流降速收敛过程非常迅速的达到稳定.而随着流的离开,租户 2 当前的活跃流相对缓慢的增速,包括从 5 s~7 s 和 7 s~8 s 两个阶段,但是最终该过程还是能够达到租户内流的公平共享.

(b) 服务器D上收到的BECN数

图 5 基于 $CCTI_Increase$ 的流级区别化带宽划分

我们利用 OFED 中的 ibdump 工具在服务器 D 上抓取了 BECN 报文, $D \rightarrow B$ 的 QP 连接源收到的 BECN 数目如图 5(b)所示,除了在流加入和退出的快速突变, BECN 保持高度稳定.在 1 条、2 条和 3 条流的情况下 BECN 数目分别为 790, 480 和 320.所达到的 5 : 3 : 2 的比例几乎与带宽分配的目标 1 : 1/2 : 1/3 相同,从而验证了第 4.2 节中分析的 BECN 数目与发送窗口的正比关系.

传统的 IB 拥塞控制机制只能提供流级的公平性,因而在数据中心网络存在用户通过建立大量连接抢占带宽的情况. 在以太网领域已经提出静态预约和动态调整的多种算法^[13-15],但是目前尚没有如何在 IB 从端到端的传输角度保证带宽和提供隔离的机制. 我们利用 IB 拥塞控制算法的区别化参数配置实现成比例的带宽分配,能够避免限速器划分的带宽浪费问题,实验中我们以租户作为流聚合的单位,动态调整每条拥塞流的少量拥塞控制参数,已较小的开销实现租户级的带宽保证和性能隔离,从而达到本文提出的数据中心对于新型传输控制的要求.

6 下一步工作

基于 IB 实现的数据中心网络传输层控制以 QP 为核心,通过拥塞控制相关参数为用户提供了可配置的传输服务,同时利用 ECN 显式拥塞标记能够更及时的对网络拥塞做出反应. 我们下一步的主要工作首先是研究针对 IB 这种极低延时网络如何配置和改进拥塞控制算法,以适应数据中心存在的突发拥塞抖动,避免拥塞传播的同时减少带宽的浪费. 此外,IB 提供了 QoS 传输质量监控,我们需要研究在不能满足 QoS 需求的情况下,如何通过自适应路由等方式在多路径网络进行灵活的流量调度,从而配合拥塞控制机制缓解网络热点链路的拥塞.

参 考 文 献

- [1] Benson T, Akella A, Maltz D A. Network traffic characteristics of data centers in the wild//Proceedings of the IMC'10. Melbourne, 2010; 267-280
- [2] InfiniBand architecture specification. Release 1.2.1. Nov 2007
- [3] Oracle Corporation. An oracle white paper: Consolidating Oracle applications on Exalogic. March, 2012
- [4] Abts D, Marty M R, Wells P M, Klausler P, Liu H. Energy proportional datacenter networks//Proceedings of the ISCA'10. Saint-Malo, 2010; 338-347
- [5] Mellanox Corporation. Mellanox solution brief: Mellanox low latency, high bandwidth InfiniBand for Web 2.0 and Cloud deployments: A Microsoft Bing Maps Recipe, 2011
- [6] Ousterhout J et al. The case for RAMClouds: Scalable high-performance storage entirely in DRAM. SIGOPS Operating Systems Review, 2009, 43(4): 92-105
- [7] Gran E G, Eimot M, Reinemo S, Skeie T, Lysne O, Huse L P, Gilad S. First experiences with congestion control in InfiniBand hardware//Proceedings of the IPDPS'10. Atlanta, 2010; 1-12
- [8] Chen Y, Griffith R, Liu J, Katz R H, Joseph A D. Understanding TCP Incast throughput collapse in datacenter networks//Proceedings of the WREN'09. Barcelona, 2009; 73-82
- [9] Wu H, Feng Z, Guo C, Zhang Y. ICTCP: Incast congestion control for TCP in data center networks//Proceedings of the CoNEXT'10. Philadelphia, 2010; 13(1)-13(12)
- [10] Alizadeh M, Greenberg A, Maltz D A, Padhye J, Patel P, Prabhakar B, Sengupta S, Sridharan M. Data center TCP (DCTCP)//Proceedings of the SIGCOMM'10. New Delhi, 2010; 63-74
- [11] Jiang J, Jain R. Analysis of Backward Congestion Notification (BCN) for ethernet in datacenter applications//Proceedings of the INFOCOM'07. Anchorage, 2007; 2456-2460
- [12] Alizadeh M, Atikoglu B, Kabbani A, Lakshminantha A, Pan R, Prabhakar B, Seaman M. Data center transport mechanisms: Congestion control theory and IEEE standardization//Proceedings of the 46th Annual Allerton Conference. Urbana-Champaign, 2008; 1270-1277
- [13] Guo C, Lu G, Wang H J, Yang S, Kong C, Sun P, Wu W, Zhang Y. SecondNet: A data center network virtualization architecture with bandwidth guarantees//Proceedings of the CoNEXT'10. Philadelphia, 2010; 15(1)-15(12)
- [14] Shieh A, Kandula S, Greenberg A, Kim C. Sharing the data center network//Proceedings of the NSDI'11. Boston, 2011; 23-36
- [15] Lam T, Radhakrishnan S, Vahdat A, Varghese G. NetShare: Virtualizing data center networks across services. University of California, San Diego: Technical Report CS2010-0957, 2010
- [16] Kabbani A, Alizadeh M, Yasuda M, Pan R, Prabhakar B. AF-QCN: Approximate fairness with quantized congestion notification for multi-tenanted data centers//Proceedings of the HOTI'10. Mountain View, CA, 2010; 58-65
- [17] Fang S, Foh C H, Aung K M M. Differentiated congestion management of data traffic for data center Ethernet. IEEE Transactions on Network and Service Management, 2011, 8(4): 322-333
- [18] Wilson C, Ballani H, Karagiannis T, Rowstron A. Better never than late: Meeting deadlines in datacenter networks//Proceedings of the SIGCOMM'11. Toronto, 2011; 50-61
- [19] Vattikonda B C, Porter G, Vahdat A, Snoeren A C. Practical TDMA for datacenter Ethernet//Proceedings of the EuroSys'12. Bern, 2012; 225-238
- [20] Santos J R, Turner Y, Janakiraman G. End-to-end congestion control for InfiniBand//Proceedings of the INFOCOM'03. San Francisco, 2003; 1123-1133
- [21] Pfister G, Guzat M, Denzel W, Craddock D, Ni N, Rooney W, Engbersen T, Luijten R, Krishnamurthy R, Duato J. Solving hot spot contention using Infiniband architecture congestion control//Proceedings of the Invited paper in High Performance Interconnects for Distributed Computing. Research Triangle Park, 2005; 1-9
- [22] Yan S, Min G, Awan I. An enhanced congestion control mechanism in InfiniBand networks for High Performance Computing systems//Proceedings of the AINA'06. Vienna, 2006; 845-850

- [23] Yan S, Min G, Awan I. Performance analysis of an efficient flow control mechanism with job threshold in InfiniBand networks//Proceedings of the AINA'08. Okinawa, 2008: 249-256
- [24] Lugones D, Franco D, Luque E. Adaptive multipath routing for congestion control in InfiniBand networks//Proceedings of the ICPP'09. Vienna, 2009: 222-227
- [25] Gran E G, Zahavi E, Reinemo S, Skeie T, Shainer G, Lysne O. On the relation between congestion control, switch arbitration and fairness//Proceedings of the CCGrid'11. Los

Angeles, 2011: 342-351

- [26] Escudero-Sahuquillo J, Gran E G, Garcia P J, Flich J, Skeie T, Lysne O, Quiles F J, Duato J. Combining congested-flow isolation and injection throttling in HPC interconnection networks//Proceedings of the ICPP'11. Taipei, China, 2011: 662-672
- [27] Galles M. Spider: A high-speed network interconnect. IEEE Micro, 1997, 17(1): 34-39
- [28] Dally J, Towles B. Principles and practices of interconnection networks. Morgan Kaufmann, 2004: 294-295

ZHANG Zi-Wen, born in 1984, Ph. D. candidate. His major research interests include datacenter networks and high performance interconnection.

LV Gao-Feng, born in 1980, Ph. D., lecturer. His major research interests include network security and software defined network.

SUN Zhi-Gang, born in 1973, Ph. D., professor. His major research interests include Next Generation Network and high performance interconnection.

WANG Cheng-Jun, born in 1983, Ph. D. candidate. His major research interests include congestion control algorithm and DTN.

LU Xi-Cheng, born in 1946, professor, Ph. D. supervisor, member of Chinese Academy of Engineering. His major research interests include computer network and distributed parallel processing.

Background

Datacenter has become a crucial infrastructure in cloud environment to provide resource and service. Based on the virtualization techniques, many different applications of multiple tenants host in one datacenter. Firstly, the traffic jitter and momentary burst nature has been widely observed which significantly impacts the utilization of the links. TCP protocol is not so reactive to this problem especially in low delay DCN since it only passively collects packet loss and RTT to infer network status. Secondly, applications have no isolation mechanism to avoid malicious user's access to the network bandwidth. Their transmission capacity cannot be limited and configured by the manager of datacenter to provide certain diverse capacity for different user. Last but not the least, datacenter transmission based on TCP is not aware of the network characteristic like multipath and traffic distribution which would result in the performance penalty. As the dominant interconnection technique in HPC, InfiniBand (IB) not only has significant performance advantages but also provides multiple new techniques to support stronger transmission control for datacenter.

This paper proposed a new-type transmission control model for DCN to systemically optimize the transmission in IB from three aspects. It realizes the differentiated application services by QoS manager which monitors the transmission quality and adopts tuning strategy when QoS of the QP (Queue Pair) is not satisfied. The strategy includes adjusting

VL mapping, congestion control, traffic reschedule over multiple path, etc. The multipath can be leveraged in IB due to the explicit different LID pairs for different paths. The network-aware capacity is provided by SA (Subnet Administration) database which stores the node and path QoS information.

The end-to-end transmission configuration and management is elaborated in this paper based on IB Congestion Control (CC). We investigate the CC algorithm and how its related parameters affect the QP's transmission capacity. The analysis and experiment proves flow level fairness can be achieved based on IB CC. And more importantly, different parameter configuration of each QP can guarantee the differentiated flow-level bandwidth allocation. The advanced bandwidth isolation strategy, such as tenant-level fairness, can be realized based on the parameter configuration.

This research is supported by Major State Basic Research Development Programs of China (No. 2009CB320503). The authors of this paper have been engaged in the work of datacenter network measurement and transmission control. This program targets to promote and optimize the application of IB in the field of datacenter. They have published some papers in this area. Since virtualization and bandwidth isolation has become an important issue in datacenter, this paper tries to resolve bandwidth allocation and configuration from the end-to-end aspect based on CC in IB.