高速以太网时延仿真分析

杨佳丽 窦 军
(西南交通大学信息科学与技术学院 成都 610031)

摘 要 随着以太网技术的不断发展,以太网已成为未来通信领域的核心技术之一。而影响用户业务体验服务质量(QoS)的参数通常包括时延、抖动、丢包率等。对于实时的语音和视频业务来说,业务数据的端到端时延则最为关键。主要在熟悉以太网标准的基础上,总结端到端时延的主要构成因素,对各个时延所占比重进行理论分析。最后对10G/100G以太网进行仿真模型设计,验证了理论分析的正确性。

关键词 10G/40G/100G以太网,端到端模型,服务质量,时延,队列调度算法

Delay Analysis and Simulation Design of High-speed Ethernet

YANG Jia-li DOU Jun
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, China)

Abstract With the rapid development of the Ethernet technologies, Ethernet has become one of the core technologies in future communication domain. And user experience may be affected by QoS parameters of services typically include delay, jitter, packet loss rate and so on. For real-time audio and video services, end-to-end delay of data is most critical. Based on the exist Ethernet standards, summarized the main factors of end-to-end delay, and theoretical analysis of different delayes on the proportion of totoal delay. Finally carried out a simulation model designed to 10G/40G/100G Ethernet, which proves the design is feasible and correct.

Keywords 10G/40G/100G Ethernet, End-to-end model, Quality of Service, Delay, Queue scheduling algorithm

1 引言

随着人们对网速要求的提高,包括在数据大集中趋势下数据中心对带宽增长的要求,以太网技术经历了诸多变革——从半双工的共享媒体10M/100M/1000M技术,发展到现在的10G/40G/100G局域网交换技术。

2009年11月中旬,Intel开始发售10GBase-T网卡,随着IEEE802.3az标准的成熟以及芯片技术的提高,10GBase-T的功耗会进一步降低。将来PC服务器甚至是普通电脑都会采用10Gbps以太网络。

所以在2006年10GBase-T标准出台后,IEEE根据网络发展的趋势,于当年6月就成立了HSSG(Higher Speed Study Group)研究小组[1]。并于2007年12月批准了PAR(Project Authorization Request)发展下一代网络40G/100G以太网标准[1]。按照计划下一代标准IEEE802.3ba将于2010年6月份出台。目前看来40G以太网与100G以太网,将来的应用主要集中在数据中,当服务器大范围采用10GBase-T的时候,核心交换的速度必须随之有一个大幅的提升。

笔者通过对现有10G端到端传输模型的分析与研究,建立了一个端到端以太网传输模型用于时延仿真。其转发节点只采用先进先出队列。通过对仿真结果的统计分析,初步验证了理论分析的正确性。

2 高速以太网技术及时延介绍

2.1 10G以太网内部协议层次

IEEE802.3ae的内部协议层次结构如图1所示[3]。图中IEEE802.3ae仅涉及数据链路层中的MAC子层(注:LLC由802.1定义,MAC控制子层只适用于半双工工作方式,10G以太网只支持全双工)和物理层。10GE的物理层的层次较复杂,可分为上、下两部分。上部分由调和子层(RS)和(可选择的)XGMII延伸单元(XGMII Extender)组成;下部分由物理编码子层(PCS)、物理介质接入子层(PMA)和物理介质相关子层(PMD)组成。对广域网,为了与SDH适配,还需要广域网接口子层(WIS)。

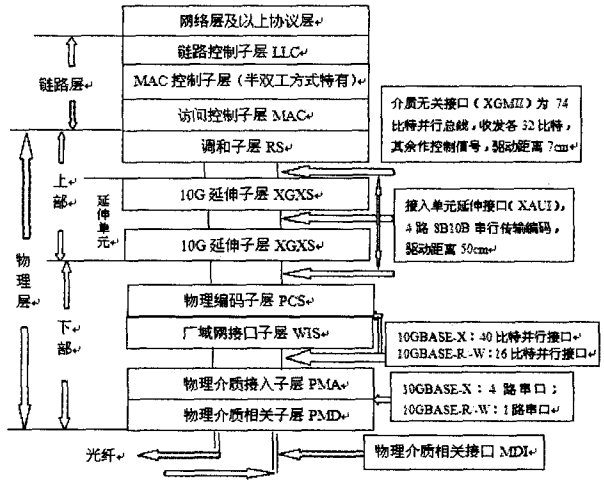

图1 10G以太网的内部协议层次示意图

图1中的MAC控制子层是半双工工作方式才需要的子层,即实现CSMA/CD功能的部分。调和子层隶属于物理

本文受国家自然科学基金(60773102)资助。

杨佳丽(1987—),女,硕士生,主要研究方向为网络与通信技术;窦 军(1963—),男,副教授,主要研究方向为网络体系结构。

层。两个可选的 AUI 延伸子层(XGXS)通过 XAUI 互联，XAUI 接口芯片为 16 根引线，允许驱动距离为 50cm。而 XGMII 接口的驱动距离仅 7cm，此接口允许调和子层与物理编码子层相距较远(相同或不同印刷板上)，故又称为长距离接口。

2.2 40G/100G 以太网传输方式

40GE 主要针对计算应用，而 100GE 主要针对核心和汇接应用。40GE 采用单个 MPO 连接器，4 根收，4 根发；100GE 采用 2 个 MPO 连接器，1 个收，1 个发[4]。发送系统将一个串行 40G 或 100GE 流分离成 4 个或 10 个并行通路。接收系统会将这 4 个或 10 个通路重新组合为单个 40G 或 100G 流。

40GE 的传输方式如图 2 所示。

图 2 40G 以太网的传输方式

100G 以太网的传输方式如图 3 所示。

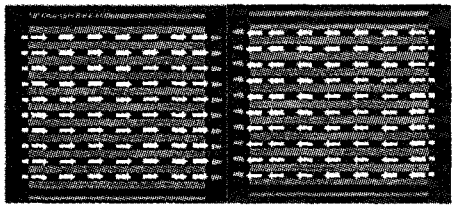

图 3 100G 以太网的传输方式

2.3 以太网时延分析

(1)共享式以太网

从发送端数据包刚好进入等待状态的时刻起到数据包最后一个字节刚好发送至接收端所经历的整个时间段称为端到端的通信时延。它主要包括 3 个时间区间：

1)等待时延 从数据包刚好进入等待状态到获得发送权所需要的时间。

2)发送时延 从数据包的第一个字节开始发送到最后一个字节发送完毕所经历的时间。用 L 表示帧的长度，V 表示以太网的通信速率，则发送时延(tsend)可以表示为 $tsend=L/V$。

3)传播时延 数据包在物理设备之间传输所经历的时间传输时延取决于传输数据的介质。传播时延＝信道长度/电磁波在信道上的传播速率。

由于所有的节点都接在同一冲突域中，不管一个帧从哪里来或到哪里去，所有的节点都能接收到这个帧。随着节点的增加，大量的冲突将导致网络性能急剧下降。而且集线器同时只能传输一个数据帧，这意味着集线器所有端口都要共享同一带宽。

(2)交换式以太网

随着交换式以太网的应用，影响网络时延的 CSMA/CD 碰撞策略得到优化。由于端口间的帧传输彼此屏蔽，从而减少了冲突[5]。但交换式以太网仍存在时延，由交换机引起的时延包括：

1)交换时延 交换时延是个定值，它与交换机的性能有关，可由交换机供应商提供。

2)帧转发时延 该时延可由交换机采用的转发模式以及帧的长度来确定。

3)缓冲时延 确定缓冲时延需要知道流量输入模式，如定期或不定期等。

3 仿真

3.1 仿真模型设计

本设计主要针对 10G/100G 以太网进行仿真，在模型中采用单向传输方式，网络域拓扑结构如图 4 所示，其中 src 节点为发送节点，dest 节点为接收节点，switch 节点为中间转发节点。

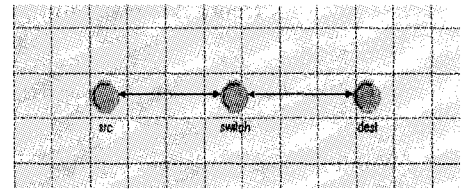

图 4 端到端网络域拓扑图

为了观察不同包格式在信道传输中对时延的影响，本方案中设置了最大包格式和最小包格式。根据 802.3as 的标准，最大包格式由 12 字节的帧间空载、7 字节帧前导、1 字节 SFD、6 字节目的地址、6 字节源地址、2 字节类型、1982 字节有效载荷、4 字节 CRC 校验构成，共 2020 字节。最小包格式不同于最大包格式在于其有效载荷为 494 字节，包长度为共 532 字节[5]。在数据包生成方式时，选择常数分布。

仿真参数主要涉及链路速率(10Gbps、100Gbps)、hub 处理模块服务速率(目前常用 100Mbps、1000Mbps)及包长度(4256bit、16160bit)，针对相关参数设置下端到端模型进行仿真。

3.2 仿真结果及分析

当链路速率为 10Gbps，转发节点服务采用 100Mbps，采用以太网最小包格式下的仿真结果如图 5 所示。

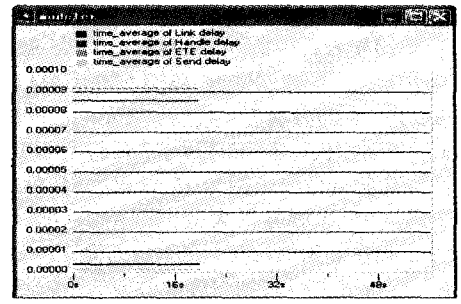

图 5 最小包格式下的仿真结果

图 5 可以从时间走向上分别比较各个时延，为了便于读取具体数据，可通过图 6 的网络报告获取各时延数值。

Statistic	Average	Maximum	Minimum
ETE delay	0.0000913	0.0000913	0.0000913
Handle delay	0.0000851	0.0000851	0.0000851
Link delay	0.00000307	0.00000307	0.00000307
Send delay	0.000000426	0.000000426	0.000000426

图 6 采用最小包格式产生的网络报告数据

在上述参数设置下，改变包长度，采用最大包时的仿真结果如图 7 所示。

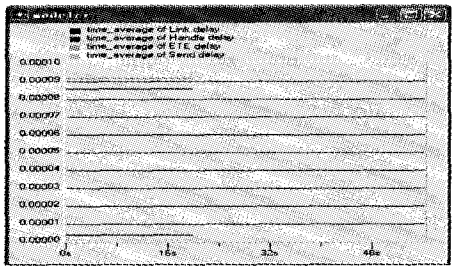

图 7　最大包格式下的仿真结果

图 8 为最大包格式下的网络报告数据。

Statistic	Average	Maximum	Minimum
ETE delay	0.000334	0.000334	0.000334
Handle delay	0.000323	0.000323	0.000323
Link delay	0.00000545	0.00000545	0.00000545
Send delay	0.00000162	0.00000162	0.00000162

图 8　最大包格式下的网络报告数据

比较不同包格式下产生的时延值可以看到，在链路速率为 10Gbps 情况下，采用以太网最小包格式，端到端时延及处理时延明显减小。产生这种结果的主要原因在于，对于不同长度的包格式，在信道中将采取不同的方式进行传输。超过一定长度的包将采用切割法传送。另外，在总的时延中，处理时延占有很大的比重，因此要减少总的时延，处理时延尤为重要。

当链路速率达到 100G 时，采用最小包格式，当中间转发节点的服务速率为 100M，仿真参数设置如表 1 所列。

表 1　链路速率为 100G 的参数设置

链路速率(data rate)	100Gbps
转发节点 hub 模块服务速率(service rate)	100Mbps
以太网包格式(packet format)	4256bit
数据包生成间隔时间	常数分布(constant)

表 1 参数设置下的仿真结果如图 9 所示。

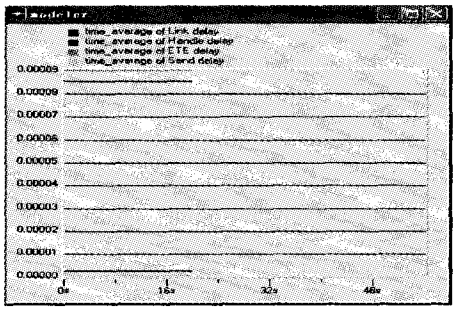

图 9　表 1 参数设置下的仿真结果

以上参数设置下网络数据报告如图 10 所示。

Statistic	Average	Maximum	Minimum
ETE delay	0.0000897	0.0000897	0.0000897
Handle delay	0.0000851	0.0000851	0.0000851
Link delay	0.00000231	0.00000231	0.00000231
Send delay	0.0000000426	0.0000000426	0.0000000426

图 10　表 1 参数设置下的网络数据报告

与 10GE 的仿真数据相比可知，虽然链路速率提高了一倍，但是端到端时延并没有得到很大幅度的减少。可见链路速率的提高并没有对时延产生很大的影响，主要因为系统产生了“瓶颈效应”，所以时延在很大程度上没有得到改善。

在表 1 参数设置中改变中间节点转发速率(1000Mbps)，仿真结果如图 11 所示。

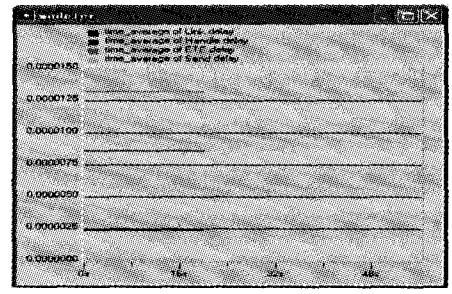

图 11　改变 hub 服务速率的仿真结果

当转发节点服务速率为 1000Mbps 时产生的网络数据报告如图 12 所示。

Statistic	Average	Maximum	Minimum
ETE delay	0.0000131	0.0000131	0.0000131
Handle delay	0.00000851	0.00000851	0.00000851
Link delay	0.00000231	0.00000231	0.00000231
Send delay	0.0000000426	0.0000000426	0.0000000426

图 12　改变 hub 服务速率下的网络报告数据

从图 10、图 11 的网络报告数据中可以看到，与 10GE 的仿真数据相比，端到端时延、处理时延、发送时延及传播时延明显有所减小。随着链路速率的提高，数据包从发送节点到目的节点的服务质量得到了显著的提高。虽然链路速率有所提升，但是由于转发节点服务速率的限制，时延值依然比较大。可见，在链路速率提升的同时，转发节点的转发速率及其采用的队列模型将在很大程度上影响时延。要改善时延，首先需要改进转发节点 hub 模块的队列模型。目前主要研究的队列模型有先进先出、严格优先级、加权时间片轮转等，本次仿真主要采用了先进先出队列，而对于其他队列算法，将在下一步的研究中加以实现。

理论分析的 10G 端到端模型中，端到端时延为 15.3064 us，处理时延为 7.5776us。本设计的实际端到端时延为 91.3us，与理论分析数值相比稍有误差。产生误差是由很多因素引起的，仿真场景在模型的设置和参数的设定上无法与真实场景保持一致，而且在链路参数的设定时，取消链路的纠错功能，设置链路的干扰模式为无干扰模式，同时仿真软件本身也存在误差，这些都会对仿真结果产生一定的影响。

结束语　本文以高速以太网为研究对象，结合以太网自身的特点，分析了高速以太网传输过程中的各个时延。然后，用 OPNET 网络仿真软件对以太网端到端传输进行了仿真，并通过网络报告显示数据的分析，验证了理论分析的正确性，同时也发现了本次设计的不足之处。首先，此次方案的设计比较简单，只能实现两个节点的数据包转发。其次，为了能更好模拟端到端模型的真实场景，应该设计一个更接近现实的方案，即中间的转发节点应使用更接近现实的交换机。在后续的工作中，笔者将会把此方案完善。最后，在 OPNET 中所进行的模拟是通过简化的网络模型来实现的，这样只能定性地说明问题，模型的完善还需要进一步的设计。

参 考 文 献

[1] 韦乐平. 城域电信级以太网的特征与新发展[J]. 电信科学, 2007(2)

[2] 刘韵洁, 张云勇, 张智江. 下一代网络服务质量技术[M]. 北京: 电子工业出版社, 2005

[3] 曾华燊. 现代网络通信技术[M]. 成都: 西南交通大学出版社, 2003

[4] Allan, Bragg, McGuire, et al. Ethernet as Carrier Transport Infrastructure[J]. IEEE Communications Magazine, February 2006

[5] 张奇智, 尹汝波. 交换式工业以太网的现状和研究[J]. 传感器世界, 2005, 2: 34-39

(上接第 304 页)

[3] orbit[EB/OL]. http://www.orbit-lab.org/

[4] Moment[EB/OL]. http://moment.cs.ucsb.edu/index.html

[5] Hull B, Bychkovsky V, Zhang Y, et al. CarTel: a distributed mobile sensor computing system[C]// Proceedings of the 4th international conference on Embedded networked sensor systems. Boulder, Colorado, USA, 2006: 125-138

[6] Broustis I, Eriksson J, Krishnamurthy S V, et al. A Blueprint for a Manageable and Affordable Wireless Testbed: Design, Pitfalls and Lessons Learned[C]// Testbeds and Research Infrastructure for the Development of Networks and Communities. TridentCom, 2007: 1-6

[7] Chereddi C, Kyasanur P, Vaidya N H. Design and implementation of a multi-channel multi-interface network[C]// Proceedings of the 2nd international workshop on Multi-hop ad hoc networks. Florence, Italy, 2006: 23-30

[8] Patra R, Nedevschi S, Surana S, et al. Wildnet: Design and implementation of high performance wifi based long distance networks[C]// 4th USENIX Symposium on Networked Systems Design and Implementation. Cambridge, MA, USA, 2007: 87-100

[9] Dhekne A, Uchat N, Raman B. Implementation and Evaluation of a TDMA MAC for WiFi-based Rural Mesh Networks[C]// 3rd ACM Workshopo on Networked Systems for Developing Regions(NSDR). Montana USA, 2009

[10] Kandhalu A, Rowe A, Rajkumar R R, et al. Real-Time Video Surveillance over IEEE 802.11 Mesh Networks[C]// IEEE Real-Time and Embedded Technology and Applications Symposium(RTAS). San Francisco USA, 2009: 205-214

[11] Guo F, Chiueh T. Software TDMA for VoIP applications over IEEE802.11 wireless LAN[C]// IEEE International Conference on Computer Communications (INFOCOM). Alaska USA, 2007: 2366-2370

[12] Verkaik P, Agarwal Y, Gupta R, et al. SoftSpeak: Making VoIP Play Well in Existing 802.11 Deployments[C]// USENIX Symposum on Networked Systems Design and Implementation (NSDI). Boston, USA, 2009

[13] Costa R, Portugal P, Vasques F, et al. A TDMA-based mechanism for real-time communication in IEEE 802.11e networks [C]// Emerging Technologies and Factory Automation (ETFA). Bilbao, Spains, 2010: 1-9

[14] Scheible G, Dacfey D, Endresen J, et al. Unplugged but connected-Design and Implementation of a Truly Wireless Real-Time Sensor/Actuator Interface[J]. Industrial Electronics Magazine, IEEE, 2007, 1(2): 25-34

[15] Korber H J, Wattar H, Scholl G. Modular Wireless Real-Time Sensor/Actuator Network for Factory Automation Applications [J]. IEEE Transactions on Industrial Informatics, 2007, 3(2): 111-119

[16] Rao A, Stoica I. An overlay MAC layer for 802.11 networks[C]// The International Conference onMobile Systems, Applications, and Services (MobiSys). Washington, USA, 2005

[17] Djukic P, Mohapatra P. Soft-TDMAC: A Software TDMA-Based MAC over Commodity 802.11 Hardware[C]// IEEE International Conference on Computer Communications (INFOCOM). Rio de Janeiro, Brazil, 2009: 1836-1844

[18] Sharma A, Belding E M. FreeMAC: framework for multi-channel mac development on 802.11 hardware[C]// Proceedings of the ACM workshop on Programmable routers for extensible services of tomorrow. Seattle, WA, USA, 2008: 69-74

[19] Neufeld M, Fifield J, Doerr C, et al. SoftMAC: A Flexible Wireless Research Platform[C]// Workshop on Hot Topoics in Networks(HotNets). Maryland, USA, 2005

[20] MadWiFi[OL]. http://madwifi-project.org/

[21] Leffler S. TDMA for Long Distance Wireless Networks [OL]. people.freebsd.org/~sam

[22] Neira-Ayuso P, Gasca R M, Lefevre L. Communicating between the kernel and user-space in Linux using Netlink sockets[J]. Software: Practice and Experience, 2010, 40(9): 797-810

[23] iproute2[EB/OL]. http://www.linuxfoundation.org/collaborate/workgroups/networking/iproute2

[24] OLSRd. OLSRD: Ad-hoc Wireless Mesh Routing Daemon[EB/OL]. http://www.olsr.org

[25] Chintalapudi K K, Venkatraman L. On the Design of MAC Protocols for Low-Latency Hard Real-Time Discrete Control Applications over 802.15.4 Hardware[C]// Proceedings of the 7th international conference on Information processing in sensor networks. 2008: 356-367