

IEEE 标准容限内以太网转发时延的测试与分析

梁永生^{1,2}, 张基宏¹, 张乃通²

(1. 深圳信息职业技术学院, 广东深圳 518029; 2. 哈尔滨工业大学通信技术研究所, 黑龙江哈尔滨 150001)

摘 要: 本文在介绍以太网时延和以太网线速的基础上, 对线速转发丢包问题进行分析. 以测试交换机和 Smartbits 测试卡作为两个对接交换设备的模型, 提出了时钟频偏是交换机转发时延的主要影响因素. 推导了以太网交换机内部转发时延和频偏时延的计算公式, 进行了以太网交换机一次、二次转发时延测试, 理论计算和实验研究的结果验证了提出理论的正确性.

关键词: 以太网时延; 线速; Smartbits 测试卡; 时钟频偏; 转发时延

中图分类号: TN915.06 **文献标识码:** A **文章编号:** 0372-2112(2008)01-0046-05

Measurement and Analysis of Forwarding Delay in Ethernet Architecture Within Tolerances of the IEEE Specifications

LIANG Yong-sheng^{1,2}, ZHANG Ji hong¹, ZHANG Nai tong²

(1. Shenzhen Institute of Information Technology, Shenzhen, Guangdong 518029, China;

2. Communication Research Center, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Based on ethernet delay and ethernet frame rate are briefly introduced, packet loss under the frame rate is analyzed in this paper. The model between a switch under tested (SUT) and a Smartbits card is presented and used for two interconnecting switches, this paper presents that clock frequency error (CFE) between a SUT and a Smartbits card is a leading factor of forwarding delay in an ethernet switch. The formulae of internal forwarding delay and forwarding delay caused by existing CFE are deduced. experimental measurement on primary Forwarding delay and secondary forwarding delay is implemented. Theoretical analysis and experimental study testify the accuracy of the above theory.

Key words: ethernet delay; frame rate; smartbits card; clock frequency error (CFE); forwarding delay

1 引言

以太网交换机或路由器是以太网架构中主要对接设备. 对接交换机或路由器之间的时延主要包括: 对接光纤/网线时延, 系统计算和处理时延和转发时延. 一般情况下, 前两者的时延通常在 μs 级, 因此转发时延的大小是决定网络时延的关键因素. 数据包的网络时延是交换网络性能的度量标准^[1-3].

以太网的线速表达式为:

$$S_{fr} = \frac{MXCF}{MFL + MIG + P + SFD} \quad (1)$$

其中: S_{fr} 为以太网线速 (Frame rate); $MXCF$ 为线路名义比特率 (Mac Xmit clock frequency); MFL 为最小帧长 (Minimum frame length); MIG 为最小帧间隔 (Minimum interframe gap); P 为前导码 (Preamble); SFD 为帧起始标志 (Start frame delimiter)

在标称时钟频率 125MHz 下, 由式 (1) 可以得到 GE

接口的 64 字节数据包的线速为:

$$\begin{aligned} S_{fr} &= \frac{MXCF}{MFL + MIG + P + SFD} = \frac{1000\,000\,000}{512 + 96 + 56 + 8} \\ &= 1488095 (\text{Packets/s, PPS}) \end{aligned} \quad (2)$$

根据文献[4]中时钟频偏 $\pm 100\text{PPM}$ 指标要求, 可以得到 GE 接口发送的最大帧速率在 1487946PPS ~ 1488244PPS 之间.

众多研究者对于转发时延的测量算法作了大量的工作^[5,6], Spirent 通信公司对 IEEE 标准容限内的性能测量进行研究^[7], 但是都未能从理论上对转发时延的机理及影响因素进行研究. 本文在介绍以太网时延和以太网线速的基础上, 以测试交换机和 Smartbits 测试卡作为两个对接交换设备的模型, 提出了时钟频偏是交换机转发时延的主要影响因素. 推导了以太网交换机内部转发时延和频偏时延的计算公式, 进行了以太网交换机一次、二次转发时延测试, 理论计算和实验研究的结果验证了提出理论的正确性.

2 线速转发丢包分析

最大吞吐量是指设备不丢包的最高转发速率,是设备正常工作时的速率.吞吐量与数据包长度的关系曲线如图 1 和图 2 所示.图 1 中不同的曲线显示理论和实际转发速率,表明当数据包长度达到一定值时,转发速率的实际值与理论值一致.图 2 以最大转发速率的百分比表示的吞吐量来规定图 1 的结果,表明当数据包长度达到一定值时,达到最大转发速率.图 1、图 2 的结论为以太网转发时延的测试与分析提供依据.

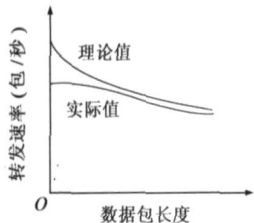


图 1 吞吐量与数据包长度的关系曲线

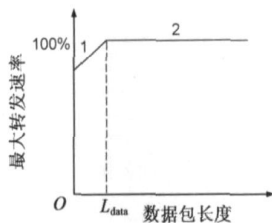


图 2 最大转发速率与数据包长度的关系曲线

转发时延研究、测试的前提是测试交换机的最大吞吐量无丢包.本文中假定:测试交换机的最大吞吐量至少达到 64 字节线速,以下的研究中以 64 字节数据包为依据.国际上通常采用 Smartbits 测试卡和相应软件 SMBApplication 来测试以太网交换机的转发时延.

在千兆以太网中,由于不可避免的时钟频偏,由式(2)可知,当测试交换机的时钟小于 Smartbits 测试卡的时钟一个 PPM 时,每一秒钟在交换芯片的内部聚留了 1.4881 个未转发出去的数据包.由于交换芯片有缓存而不会导致立即丢包.在缓存内会产生数据包的累积,累积速度为 1.4881 个数据包/秒,随着时间的增加,累积的数据包就会耗尽缓存的存储空间而导致丢包,丢包速度为 1.4881 个数据包/秒.

由表 1 可知:1 个数据包= 672 比特.

所以在标称 125MHz 频率下,在测试交换机开始丢包之前,每一秒钟产生的转发时延为:

$$1.4881 \times 672 \times \frac{1}{1 \times 10^9} \text{s} = 1\mu\text{s}$$

当测试交换机的时钟小于 Smartbits 测试卡的时钟一个 PPM 时,假定测试交换机的缓存为 4M 字节,则最少:

$$T_{\text{loss}} = \frac{S_B}{N_{ps} \times 672 \div 8} = \frac{4000000}{125} = 32000\text{s} \quad (3)$$

后开始丢包.

其中: T_{loss} 为交换机开始丢包前时延历时(s); S_B 为交换机缓存大小(Bytes); N_{ps} 为每秒钟累积的数据包的数量.

随着测试时间的增加,转发时延也在增加,转发时延的增加速度为每秒钟 $1\mu\text{s}$.数据包累积、时延增加和

测试时间的对应关系如图 3 所示.

在图 3 中,1 段表示:在未产生丢包之前,转发时延不断增加;2 段表示:交换机缓存耗尽,在产生丢包之后,转发时延不再增加.

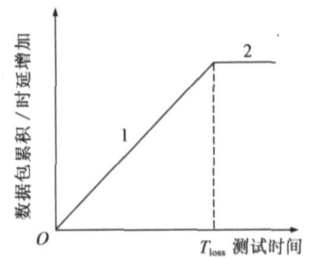


图 3 数据包累积、时延增加和测试时间对应关系图

3 转发时延测试

本文以测试交换机和 Smartbits 测试卡作为两个对接交换设备的模型,通常测试转发时延的组网是采用两个端口进行双向“灌流量”,本文选取一台带有四个 GE 口的以太网交换机 Cisco 4006 进行理论计算和实验研究.图 4 为 GE 口一次转发时延测试组网图,图 5 为 GE 口二次转发时延测试组网图^[8,9].测试交换机转发时延包括内部转发时延和频偏时延两部分.

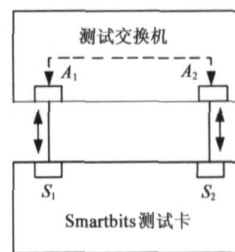


图 4 GE 口一次转发时延测试组网图

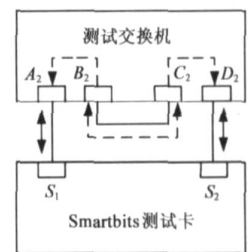


图 5 GE 口二次转发时延测试组网图

图 4、5 中画出了 GE 端口的双向流量,测试交换机的所有端口采用同一个晶振时钟作为发送时钟源,所以具有相同的时钟频偏指标.同样 Smartbits 测试卡的 S_1 和 S_2 端口时钟也具有相同的时钟频偏指标.由于测试交换机的两个流向具有相同的特性,在以下的分析中以图 4 中的“ $S_1 \rightarrow A_1 \rightarrow B_1 \rightarrow S_2$ ”和图 5 中的“ $S_1 \rightarrow A_2 \rightarrow B_2 \rightarrow C_2 \rightarrow D_2 \rightarrow S_2$ ”的单向流量为例.

3.1 内部转发时延分析

测试交换机内部转发时延可以通过计算图 5 中 GE 口的二次转发时延测试值和图 4 中 GE 口的一次转发时延测试值的差值得到.为了便于分析,我们定义:

D_{a1b1} — 图 4 中 A_1 到 B_1 的转发时延;

D_{a2b2} — 图 5 中 A_2 到 B_2 的转发时延;

D_{c2d2} — 图 5 中 C_2 到 D_2 的转发时延;

D_{swc} — 与 D_{a1b1} , D_{a2b2} 和 D_{c2d2} 都是交换机内部转发时延,由于测试交换机采用同一个交换芯片,所以内部转发时延基本相等;

D_{ckf} — 时钟频偏引起的频偏时延;

D_{lnk} — 测试交换机和 Smartbits 测试设备之间的光纤传输线路时延.

D_{snb} —测试系统总时延, Smartbits 测试设备计算和处理数据的总时延。

按照第2节的分析可知: 当测试交换机的时钟小于 Smartbits 测试卡的时钟时, 就会引起报文累积而产生时延—频偏时延。如果测试交换机的时钟大于或等于 Smartbits 测试卡的时钟, 则频偏时延等于 0。

在时延测试中得到的时延为:

$$D_{\text{total}} = D_{\text{snb}} + D_{\text{lnk}} + D_{\text{cfe}} + D_{\text{swc}} \quad (4)$$

由式(4)可以得到: 采用同一种型号的两台测试交换机, 由于频偏时延不同, 它们的转发时延测试结果不相同; 同样道理, 即使对于同一台测试交换机, 选用不同的 Smartbits 测试卡, 转发时延测试结果也不一样。

对于图 4:

$$D_{\text{total1}} = D_{\text{snb}} + D_{\text{lnk}} + D_{\text{cfe}} + D_{a1b1}$$

对于图 5:

$$D_{\text{total2}} = D_{\text{snb}} + D_{\text{lnk}} + D_{\text{cfe}} + D_{a2b2} + D_{c2d2}$$

忽略图 5 中的 B_2 到 C_2 端口的网线连接时延, 而且 $D_{\text{swc}} = D_{a1b1} = D_{a2b2} = D_{c2d2}$, 两个测试值相减就可以得到交换机内部转发时延为:

$$D_{\text{swc}} = D_{\text{total2}} - D_{\text{total1}} \quad (5)$$

由式(5)可以看出, 交换机内部转发时延排除了时钟频偏引起的频偏时延, 光纤传输线路/对接网线时延以及 Smartbits 测试设备的系统时延的影响。

可以通过测量多次转发时延值, 取相邻两次转发时延的差值的平均值作为交换机的内部转发时延。

3.2 频偏时延计算与分析

假定测试交换机的时钟小于 Smartbits 测试卡的时钟一个 PPM 时, 那么包缓存中聚留一个数据包需要的时间为:

$$T = 1 \div 1.4881 = 0.672\text{s}$$

这个值也就是不受时钟频偏影响的最小时延测试时间。根据上述的分析, 在规定的测试时间内, 不可能避免时钟频偏对测试结果的影响。如果时钟频偏不止一个 PPM, 最小时延测试时间将会更小, 具体的计算如下:

$$\Delta T = \frac{1}{1.4881 \times N_{\text{cfe}}} \text{s}$$

其中: N_{cfe} 是测试交换机和 Smartbits 测试卡之间的时钟频偏。

从上面的分析过程看, 在 ΔT 的时间后就会产生一个数据包聚留在包缓存中, 使后续的数据包时延增加 ΔT 。假定测试时间为 T_{test} , 则在测试时间段内, 有 $N = (T_{\text{test}} \div \Delta T)$ 个 ΔT 时间单位。

第 1 个 ΔT 内: $D_{\text{total}} = (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}})$

第 2 个 ΔT 内: $D_{\text{total}} = (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + D_p \times 1$

第 3 个 ΔT 内: $D_{\text{total}} = (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + D_p \times 2$

第 4 个 ΔT 内: $D_{\text{total}} = (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + D_p \times 3$

第 5 个 ΔT 内: $D_{\text{total}} = (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + D_p \times 4$

.....

第 N 个 ΔT 内: $D_{\text{total}} = (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + D_p \times (N - 1)$

其中: D_p 是一个数据包的时延, 即:

$$D_p = 672 \times \frac{1}{1 \times 10^9} \text{s} = 0.672 \mu\text{s}$$

我们测试的时延值为平均转发时延, 在 T_{test} 时间内的平均转发时延为:

$$\begin{aligned} \overline{D_{\text{total}}} &= (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + (D_p \times (0 + 1 + 2 + \dots \\ &\quad + (N - 1))) \div N \\ &= (D_{\text{snb}} + D_{\text{lnk}} + D_{\text{swc}}) + L_p \times (N - 1) / 2 \end{aligned} \quad (6)$$

对照式(4)和(6)可得:

$$D_{\text{cfe}} = D_p \times (N - 1) / 2$$

所以:

$$\begin{aligned} D_{\text{cfe}} &= 0.672 \times ((T_{\text{test}} \div \frac{1}{1.4881 \times N_{\text{cfe}}}) - 1) / 2 \\ &= 0.5 \times T_{\text{test}} \times N_{\text{cfe}} - 0.336 (\mu\text{s}) \end{aligned} \quad (7)$$

由式(7)可以得到: 频偏时延与测试时间成正比; 频偏时延与测试设备和 Smartbits 测试卡之间的时钟频偏成正比。

根据实际测试转发时延, 根据式(7), 可以得到时钟频偏的计算公式:

$$N_{\text{cfe}} = \frac{2 \times (D_{\text{cfe}} + 0.336)}{T_{\text{test}}} \quad (8)$$

4 实验研究

4.1 测试实例

在标时钟频率 125MHz 情况下, 对于千兆以太网口, 按照文献[9]的规范要求, 测试时间取为 60s。假定交换机的缓存为 4M 字节, 由式(3)和(7)可以计算出测试交换机和 Smartbits 测试卡之间时钟频偏, 频偏时延和丢包前时延历时之间的对应关系如表 1 所示。

表 1 时钟频偏, 频偏时延和丢包前时延历时的对应关系表

时钟频偏 N_{cfe} (PPM)	频偏时延 D_{cfe} (ms)	丢包前时延历时 T_{loss} (s)
0	0	∞
10	0.2997	3200
20	0.5997	1600
30	0.8997	1067
40	1.1997	800
50	1.4997	640
60	1.7997	533
70	2.0997	457
80	2.3997	400
90	2.6997	356
100	2.9997	320

由表 1 可知: 测试交换机和 Smartbits 测试卡之间的时钟频偏达到 33PPM 时, 频偏时延就已经达到 ms 级.

4.2 一次, 二次转发时延测试

实验选用图 4, 5 的测试组网图, 测试仪器卡采用 LAN3200A, 应用 SMBApplication 平台, 测试的时间取为 30 秒, 测试交换机为一台带有 4 个 GE 口的 Cisco 4006. 表 2, 3 所示为测试交换机不同大小数据包情况下的一次转发时延和二次转发时延的测试结果^[10].

表 2 数据包与一次转发时延测试结果对应表

数据包大小(Bytes)	线速(%)	一次转发时延(μ s)
64	100. 00	267. 8
128	100. 00	309. 1
256	100. 00	310. 2
512	100. 00	248. 5
1024	100. 00	273. 6
1280	100. 00	201. 3
1518	100. 00	173. 2

表 3 数据包与二次转发时延测试结果对应表

数据包大小(Bytes)	线速(%)	二次转发时延(μ s)
64	100. 00	270. 9
128	100. 00	313. 3
256	100. 00	315. 5
512	100. 00	256. 1
1024	100. 00	285. 5
1280	100. 00	217. 5
1518	100. 00	189. 7

由表 2, 3 可知, 最大吞吐量为 64 字节线速, 测试交换机的一次转发时延为 267. 8 μ s, 二次转发时延为 270. 9 μ s, 则测试交换机的内部转发时延为:

$$D_{swc}=270.9-267.8=3.1\mu s$$

通过多次转发测试得到的内部转发时延比较准确, 但是限于实验条件的限制, 本文不做三次转发时延测试. 测试交换机的内部转发时延包括交换芯片的转发处理、媒质访问控制地址查找和队列调度等时延.

忽略 Smartbits 测试卡自环时延, 则一次转发时延与内部转发时延的差值即为测试交换机和 Smartbits 卡之间的频偏时延, 即:

$$D_{dc}=267.8-3.1=264.7\mu s$$

时钟频偏引起的频偏时延是无法避免的, 因为在网络上不同设备对接都会存在“时钟频偏”和“时延不确定”现象, 严重情况下可能会达到 ms 级.

由式(8), 我们可以计算出测试交换机和 Smartbits 测试卡之间的时钟频偏, 即:

$$N_{dc}=\frac{2\times(264.7+0.336)}{30}=17.67PPM$$

由以上分析可知, 测试交换机和 Smartbits 测试卡之间的时钟频偏引起的频偏时延远大于内部转发时延, 这也验证了时钟频偏是交换机转发时延的主要影响因素的正确性.

5 结束语

本文以测试交换机和 Smartbits 测试卡作为两个对接交换设备的模型, 网络时延主要包括测试交换机和 Smartbits 测试卡之间的频偏时延、对接光纤/网线时延以及系统计算和处理时延. 在忽略不计后两者的前提下, 提出了时钟频偏是交换机转发时延的主要影响因素. 推导了测试交换机的内部转发时延和频偏时延的计算公式, 进行了一次、二次转发时延测试, 理论计算和实验研究的结果验证了提出理论的正确性, 为实时性要求极高的应用领域的网络性能测量和分析提供依据.

参考文献:

[1] Jamli N Ayoub, Barbara, M At Ramadna. Average network delay in packet switched networks[A]. Proceedings of the 6th IEEE Intemational Conference on Electronics, Circuit and Systems[C]. IEEE Press, 1999, 1: 389- 392.

[2] 蒋婕. Internet 网络延时分析[J]. 广西通信技术, 2002, (2): 19- 22.

Jiang Jie. Analysis of internet network delay[J]. Guangxi Communication Technology, 2002, (2): 19- 22. (in Chinese)

[3] Sue B Moon, Paul Skelly, Don Towsley. Estimation and removal of clock skew from network delay measurement[A]. Proceedings of the 18th Annual Joint Conference of the IEEE Computer and Communications Societies[C]. IEEE Press, 1999, 1: 227- 234.

[4] Lan Man Standards Committee of the IEEE Computer Society. IEEE Standard 802. 3 Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications[M]. The Institute of Electrical and Electronics Engineers, Inc. 1998: 36- 79.

[5] J H Choi, C Yoo. Analytical derivation of one way delay[J]. Electronics Letters, 2003, 39(25): 1871- 1872.

[6] Spirent Communications, Inc. Performance Measurement within the Tolerances of the IEEE Specifications[R]. USA: Spirent Communications White Paper, 2003: 1- 7.

[7] 王雷, 蔡安妮, 孙景鳌. 延时估计和实时数据的同步控制[J]. 通信学报, 1999, 20(2): 46- 52.

Wang Lei, Cai Anni, Sun Jingao. Delay estimation and synchronization of realtime data[J]. Journal of China Institute of Communications, 1999, 20(2): 46- 52. (in Chinese)

[8] Network Working Group. Request for Comments: 2285[S]. European Network Laboratories Category, 1998: 1- 25.

[9] Network Working Group. Request for Comments: 2544[S].

Harvard University, 1999: 1- 31.

- [10] 梁永生. 网络中心战中动态组合关键问题的研究[D]. 哈尔滨工业大学博士后研究报告, 2005: 39- 52.

Liang Yongsheng. Research on Key issues of Dynamic Clustering in Network Centric Warfare Architecture[D]. Harbin Institute of Technology Postdoctoral Dissertation, 2005: 39- 52. (in Chinese)

作者简介:



梁永生 男, 1971 年生于黑龙江, 1999 年于哈尔滨工业大学获工学博士学位, 现为深圳信息职业技术学院教授. 研究方向为计算机网络与数据通信、信号处理与模式识别.

E-mail: liangys@szit.com.cn



张基宏 男, 1964 年生于江苏, 1992 年于东南大学获工学博士学位, 现为深圳信息职业技术学院院长, 教授. 研究方向为数据通信、信号处理和多媒体信息处理.

E-mail: zhangjh@szit.com.cn



张乃通 男, 1934 年生于江苏, 哈尔滨工业大学教授, 中国工程院院士.

E-mail: ntzhang@hit.edu.cn.