高速网络测量系统时钟同步的研究与分析[1]

林容容，丁伟，程光

（东南大学计算机科学与工程系 南京 210096）

摘要:在监测网络流量行为中，时钟同步是提供准确的网络测度测量值的前提保证。单向延迟等网络性能测度至少需要毫秒级的同步精度。然而，在高速、大规模、分布式的网络环境下，高速的网络流量影响了系统的时钟响应信号，使得原本是线性模型的相对时钟偏移模型受到影响。这一影响使得即便使用 NTP 或者 GPS，都无法很好地解决时钟同步问题。

本文以一个分布式的抽样测量监测系统 PERME 和 CERNET 地区网主干为依托，详细分析和讨论了高速 IP 网络中的时钟偏移问题。通过大量真实的实验数据对高速网络中的时钟同步问题进行了研究和分析。同时，对如何解决实验中出现的模型非线性的问题指出了自己的见解。

关键字：时钟同步；时钟偏移；高速网络；线性模型

中图分类号：TP393.1　　文献标识码：A　　文章编号：0401556

Clock Synchronization Experiment and Research in High-Speed Network

LIN Rong-rong, DING wei, CHENG guang

(Department of Computer Science & Engineering, Southeast University, Nanjing, Jiangsu Province 210096, China)

Abstract: While monitoring the network traffic behavior, clock synchronization is the guarantee to provide the accurate measurement of network metrics. However, in high-speed, large-scale, distributed network environment, the high-speed traffic hinders the clock single from duly response. As a result, the normal linear model of relative clock offset has been changed, which cause even NTP or GPS cannot resolve the clock synchronization problem.

Based on a distributed sample measure and monitor system PERME as well as the backbone of CERNET, clock offset problem in high-speed network has been discussed. Through large number of real experiments, clock synchronization problem has been researched and analyzed. Meanwhile, the paper has given its own opinion of that how to settle the non-linear model problem.

Key words: Clock Synchronization; Clock's Offset; High-Speed Network; Linear Model

1 引言

计算机时钟一般以振荡电路或石英钟为基础，每天的误差可达数秒，经过一段时间的累积就会出现较大的误差。随着分步式计算和网络技术的发展，不准确的计算机时钟对于网络结构以及其中应用程序的安全性会产生较大的影响，尤其是那些对时钟是否同步比较敏感的网络指令或应用程序。

在一个网络中，解决系统之间的时钟同步不可能完全依靠系统管理员手工使用 date 命令来调节各个系统的时钟，而是通过网络时间协议 NTP[1]。

[1] 收稿日期：2004-09-DD；修回日期：2004-MM-DD
基金项目：国家 973 项目(No.2003CB314803)

随着网络技术的不断发展，Internet 网络环境下的网络端至端 SLA 监测分析中，许多度量参数的准确性对时钟同步精确度的要求变得更高了，如单向延迟[4]的测量等。但由于不同主机的时钟振荡频率不一，即使采用 NTP 同步协议，在同步一段时间后都又会产生不同步。测试实验发现 5 分钟内两台主机之间的不同步大约为 30 毫秒，而在对单向延迟的测量中，需要毫秒级的同步精度。因而必须对时钟的偏移进行修正。

许多研究者对网络时钟同步问题展开了研究。分段线性最小法将测得的时钟偏移分为若干段，然后统计其变化情况，其同步结果很不理想。Paxson 提出了一种对往返两个方向上的单向延迟经处理后再使用线性规划方法测量时钟偏移以同步时钟的方法[5]。但这种方法在网络状况变化频繁的情况下性能很差。Moon 方法提出使用一种标准线性算法[6]逼近测量得到的延迟值，从而消除其中的延迟偏移。Li Zhang 方法基于 Moon 等人的研究提出类似凸包计算的修正算法[7]，并实现了在系统中存在时钟的瞬时校正情况下的同步工作。

无论以上何种方法，其基本出发点都是利用两端点的时戳报文所携带的时钟信息测量值进行统计计算从而实现时钟的同步。但在高速、大规模、分布式的网络环境下，高速的网络流量影响了时钟响应信号，使得时钟同步问题变得更加困难。为研究解决高速网络中的时钟同步问题，本文使用分布式抽样测量服务级别约定(SLA)监测系统 PERME 对高速 IP 网络中的时钟偏移展开实验研究。

全文内容组织如下：第二章介绍非高速网络环境中时钟同步问题的基准实验及结果；第三章中介绍本实验所在的高速网络环境和具体实验体系；第四章详细论述实验结果及对该结果的分析研究；最后在第五章中给出结论，总结全文。

2 基准实验

在基准实验中，非高速网络环境中的相对时钟偏移值呈良好的线性模型，因而可以利用各种统计方法对其估计计算以进一步修正。

为使测量方法不仅可以有效地完成测量目的，而又不会对网络性能造成太大冲击，测量方法应满足简便的原则，并尽量使用已有的测量工具，使用得到广泛支持和充分实现的协议。由于 ICMP 协议在几乎各种主机和路由器上都得到支持，因此使用 ping 工具来测量是十分简便的方法。尽管 ping 的方法所测得的数据有一定的局限性，其性能和其他 TCP、UDP 或其他 IP 协议有一定的出入（一般，路由器给 ICMP 协议的优先性较低），但考虑 ping 工具及 ICMP 协议实现的普遍性，利用 ping 工具测量 Internet 网的性能，尤其在测量端到端性能的时候，是最普遍的做法。故本设计选用 ping 工具作为测量工具。

图 1 给出了两机器间时戳报文的测量体系图。

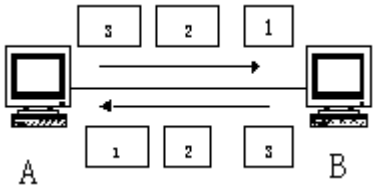

图 1 两机器时戳报文测量系统结构

从一台主机 A 发送带主机 A 发生报文时刻 $f_A(t_{i,1})$的时戳报文到另一主机 B，主机 B 一旦收到时戳报文立即将其接收时戳 f_B（$t_{i,2}$）记录在时戳报文中，同时转发报文至主机 A，并记录转发时戳 f_A（$t_{i,3}$），主机 A 收到该时戳报文产生应答报文到主机 A，主机 A 处记录接收该报文时戳 f_B（$t_{i,4}$），因此在主机 A 处可以得到四个时戳：f_A（$t_{i,1}$）、f_B（$t_{i,2}$）、f_A（$t_{i,3}$）、f_B（$t_{i,4}$）。其中，i 表示发送的是第几个时戳报文，假设总共时戳报文数为 n，$1 \le i \le n$。

图 2 为时戳报文的结构。其中，

SeqNum ——记录时戳报文的序号 i；

TimeSec ——记录时戳报文的时间（秒以上的部分）；

TimeUsec——记录时戳报文的时间（秒以下的部分）。

根据式（1）即可计算出各时刻的相对时钟偏移 $\Delta offset_i$ [2]。

$$\Delta offset_i = (t_{i,A} - t_{i,B}) = ((f(t_{i,4}) - f(t_{i,3})) - (f(t_{i,2}) - f(t_{i,1})))/2 \qquad \text{式（1）}$$

图 3 是该测量体系测得的点 A 和点 B 之间的时钟相对偏移 $t_{i,A} - t_{i,B}$ 曲线图，该组数据为每秒钟测量一次。

为验证测量器时间和两测量器相对时钟存在线形关系，评价模型的拟合程度，定义 R^2 测度[3]用于评价模型拟合测量数据的程度。R^2 测度值定义为：

$$R^2 = 1 - SSE / SST \qquad \text{式（2）}$$

SeqNum	$f_A(t_{i,1})$		$f_B(t_{i,2})$		$f_B(t_{i,3})$		$f_A(t_{i,4})$	
	TimeSec	TimeUsec	TimeSec	TimeUsec	TimeSec	TimeUsec	TimeSec	TimeUsec

图 2　时戳报文的结构

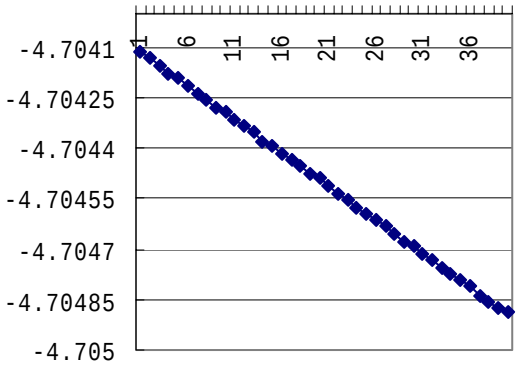

图 3 不同步时的相对时钟偏移 1

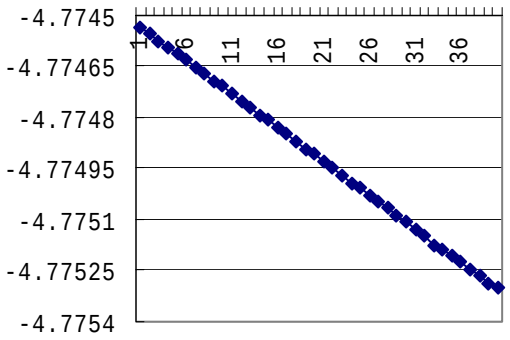

图 4 不同步时的相对时钟偏移 2

其　中　：　$SSE = \sum (y_i - \hat{y})^2$　，

$SST = \left(\sum y_i^2\right) - \left(\sum y_i\right)^2 / n$。当 R^2 在[0, 1]之间变化，并趋近于 1 时，表明模型具有良好的拟合效果。通过计算得到，R^2=1.000000，表明 y 与 x 线性相关。

在测量单向同步测量实验结束以后又做了相同的实验，其实验结果见图 4。计算此模型的 R^2 亦为 1.000000，即线性相关。

既然模型符合线性关系，因此将前后两次相邻的测量相对时钟偏移结果相减得到图 5 和图 6。从图中可以看出，相对时钟偏移在-0.0001 附近上下随机波动，这说明两相对时钟具有一定的随机波动性，但总的趋势来看是服从线性规律的。由于这种波动幅度在 0.02ms 上下波动,基本已经超出了计算机时钟频率精度，而且也超出单向延迟的测量值的精度范围，因此，相对时钟偏移的波动性无需讨论和研究。

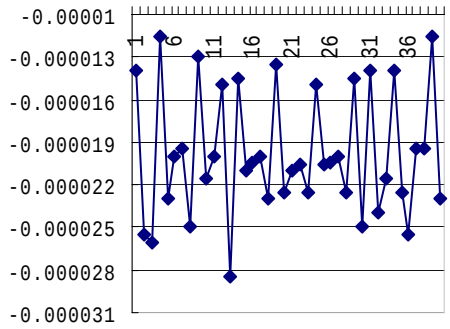

图 5 第一次相对时钟偏移的波动情况

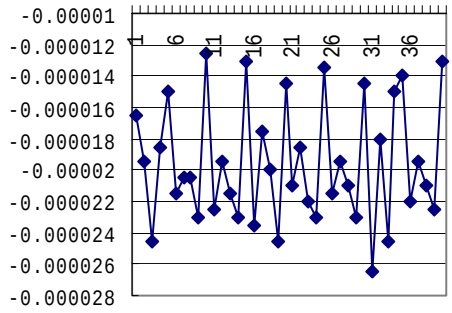

图 6 第二次相对时钟偏移的波动情况

以上实验是在非高速网络环境下进行的，测得的相对时钟偏移呈良好的线性模型分布状态。而必须指出的是，相对时钟偏移良好的线性模型的建立是以下列两个假设为前提的：

1. 用于测量的主机工作不繁忙，即：可忽略系统对时钟响应的延迟时间，将系统时间的不准确仅仅归结为时钟振荡频率的不同；
2. 时戳报文通过两台主机所需要时间相同，即假设通过 A、B 需要的时间和应答报文通过 B 回到 A 所需要的时间相同。

在非高速网络环境中，这种假设是符合实际情况的。

3 高速实验环境

基准实验的结果代表了普通实验环境下的时钟偏移情况，本文在此基础上重点研究高速实验环

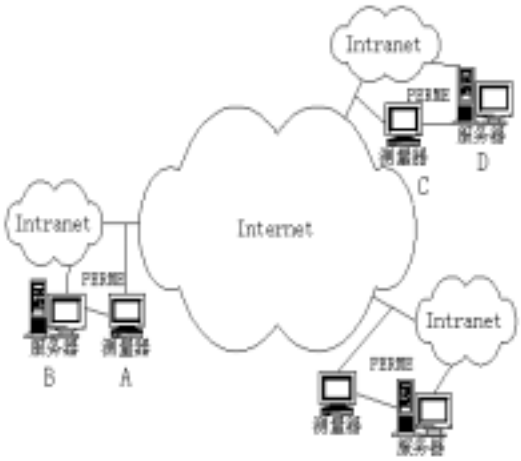

图 7 PERME 系统实验环境图分布图

境下的时钟偏移，首先介绍一下本文的高速实验平台。

实验进行的环境是面向高速网络流量的 PERME 系统。该系统是一个分布式抽样测量服务级别约定(SLA)监测系统，它面向高速 IP 网络（CERNET 地区网主干），采用被动的抽样测量方式，向用户提交涉及单个端点或两个端点的性能测量参数的 SLA 报告。同时对传输层协议和周知端口进行报文总数和字节总数的统计，为网络行为学的研究提供原始数据。

如图 7 所示，PERME 系统采用以测量域为单位的自治结构，并以联邦方式支持合作和协同。每个测量域内有一个测量点，可以独立完成单点的测量功能。PERME 系统测量和提供各种反映网络服务质量的性能参数。性能参数量化了终端用户对网络服务质量的直观感受。

对许多度量参数来说，时间的准确同步是保证所得各种测度值准确的前提，如：单向延时的测定就必须依赖于高精度的时间同步。

实验展开的具体环境如图 7 中所示，测量器分别是图中的 A 机器和 C 机器，服务器为 B 机器和 D 机器。在此实验系统中，机器 A、C 分别是两个测量域中的测量点，独立完成单个端点性能参数的测量功能。

在此测量系统中，测量器主机直接面向高速网络，采集各种数据报文，服务器主机根据采集来的数据报文统计各测度结果。数据报文的时戳是由测量器主机的时钟决定的，因此测量器主机间的时钟同步成为首要问题。

4 实验结果分析与讨论

为讨论在高速、大规模、分布式的网络环境下，

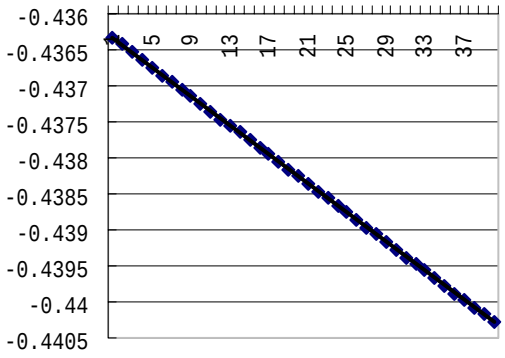

图 8 使用实验内核的时钟偏移情况

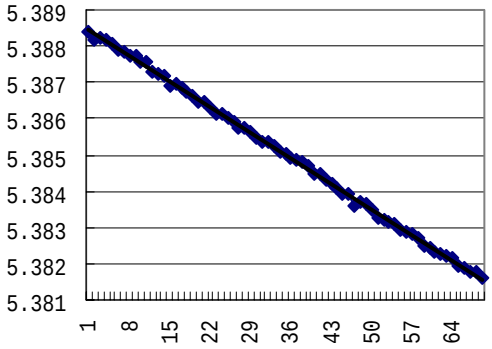

图 9 使用千兆网卡的时钟偏移情况

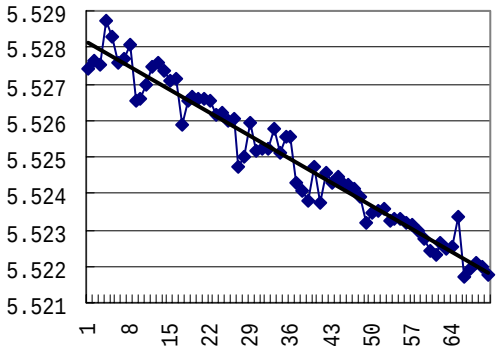

图 10 千兆网卡+实验内核

高速的网络流量如何影响时钟信号，如何解决时钟同步问题，我们在第三节所介绍的高速网络平台上对图 7 中的 A、B、C、D 四台机器两两进行测量。在此环境中，第二章中所论述的相对时钟偏移的线性模型仍是合理的，但由于网络流量和主机繁忙程度的增加，这种线性模型将可能受到严重的影响。

本实验对测量器 A 和测量器 C 的相对时钟偏移情况展开。与第二节的基准实验相比，不同的是本实验环境需要启动实验内核及千兆网卡，系统将在此平台上采集大量的数据报文。

首先，考虑实验内核是否对相对时钟偏移模型存在影响。如图 8 所示的实验结果是在不启动千兆网卡的情况下使用实验内核测得的测量器 A 和测量器 C 之间的相对时钟偏移的测试结果。该曲线的

$R^2 = 0.99995959$，说明该曲线具有良好的线性特征。该结果证明，使用实验内核时的线性模型是很明显的。为全面检验实验内核的稳定性，在启动千兆网卡后再将其停用，此时测得的相对时钟偏移的曲线的 R^2 测度值为 0.99997399，仍然是良好的线性关系模型。

这说明：当前的实验内核对相对时钟偏移的线性模型并不存在明显的影响。

从另一方面考虑，PERME 系统的实验环境的突出之处就在于它的高速性和大规模。在启动千兆网卡后，也就意味着网络将承受大流量的冲击。

实验证明，启动千兆网卡后，相对时钟偏移仍然是符合线性模型的。如图 9 所示，该曲线具有良好的线性关系数据，经拟和计算得到其 R^2 测度值为 0.99911394，。即：启动千兆网卡但使用普通内核的情况下，相对时钟偏移的线性关系也很明显，而且这种线性关系与不启动千兆网卡时的模型情况可以说是无区别的。如此看来启动千兆网卡对相对时钟偏移的线性模型也不存在明显的影响。

最后，图 10 的曲线是在启动千兆网卡同时使用实验内核的情况下测量得到的相对时钟偏移值。虽然该曲线的总体趋势仍基本上是呈直线状的，但其线性关系远不如之前的实验结果那么良好。计算得到其 R^2 测度值为 0.95028743，这也就证实了此时的相对时钟偏移线性模型受到了一定程度的干扰。

研究发现，影响相对时钟偏移模型的线性关系程度的主要原因不是千兆网卡的启动也不是实验内核的使用，其主要原因是在于网络的繁忙程度。图 9 中的结果由于是在普通内核下测得的，故 PERME 系统并未采集任何数据。而当 PERME 系统开始面向大规模高速网络进行数据采集时，也就是图 10 所示的情况下，相对时钟偏移的线性模型明显受到了影响。

进一步的实验结果显示：当处于高速网络环境的 PERME 系统的采集率不断提高时，相对时钟偏移模型的线性关系程度也就愈加不明显。图 11 的数据是在采集率为 5／256 时测得的。而当采集率为 10／256 时，如图 12 所示，相对时钟偏移模型的线性程度变得更加的差。

经分析，造成这种实验现象的原因可能是由于网络高速，测量器主机十分繁忙，因此系统难以及时响应时钟同步信号，而响应时钟滞后的延迟使得该实验结果与基准实验结果的不一致。

从理论上说，单向延迟由三个部分组成：光速的传播延迟；发送一个数据单元花费的时间以及网络内部的排队延迟。光速的传播延迟是由物理介质上的传输速度决定的。由于速度在通常情况下总是一定的，而本实验中的测量器之间的物理链路长度

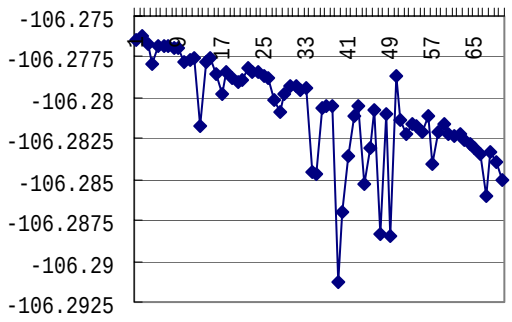

图 11 采集率 5／256

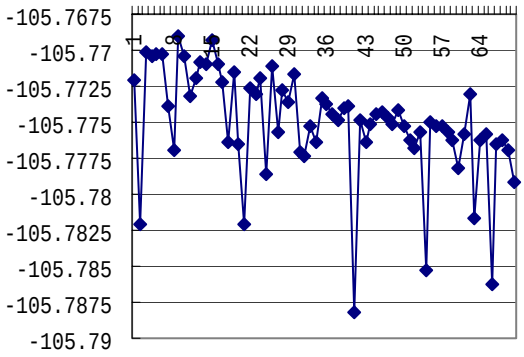

图 12 采集率 10／256

也是固定的，故该部分的延迟是一定的。而后二者则与当前的网络运行状况有关。其中，设 Size 是数据传输中的分组的大小，Bandwidth 则是网络分组带宽，则发送一个数据单元花费的时间 Transmit 可由下式得到：

Transmit = Size / Bandwidth　　式（3）

而分组交换机在将分组转发出去之前通常还需要将其存储一段时间，这就是排队延迟。

在高速大规模网络环境中，网络流量大，测量器系统的采集工作繁忙，CPU 的计算能力有限，在高负荷运作下 CPU 无法及时处理时戳报文。这使得测量器主机系统对时钟响应的延迟时间有所增加，即第二节中的假设 1 不再成立。两台测量器主机的繁忙程度不一样，Transmit 时间及排队延迟无法预测也不再等同，使得假设 2 中对时戳报文通过两台主机所需要时间相同的假设也不再成立。此时，在图 1 所示的测量体系下，设：时戳报文通过 AB 需要的时间和通过 BA 所需要的时间分别为 $delay_{i,AB}$ 和 $delay_{i,BA}$。此时：$delay_{i,AB} \neq delay_{i,BA}$。

于是有：

$$f_B(t_{i,2}) = f_A(t_{i,1}) + delay_{i,AB} + \Delta offset_i$$

$$f_B(t_{i,3}) = f_A(t_{i,4}) + delay_{i,BA} - \Delta offset_i$$

两式相减得到：

$$\Delta offset_i = (t_{i,A} - t_{i,B}) = ((f_A(t_{i,4}) - f_A(t_{i,3})) - (f_B(t_{i,2}) - f_B(t_{i,1})) - (delay_{i,AB} - delay_{i,BA}))/2 \quad \text{式（4）}$$

式（4）中的 $delay_{i,AB}$ 和 $delay_{i,BA}$ 受该时刻的网络负载情况影响，对相对时钟偏移的测量来说，是一种不确定因素。同时，网络负载严重使得时戳 $f_A(t_{i,1})$、$f_B(t_{i,2})$、$f_A(t_{i,3})$ 和 $f_B(t_{i,4})$ 中记录的时间中包含了时钟信号的延迟。正是这些因素造成了对相对时钟偏移模型的非线性测量结果，使得在高速网络中的时钟同步工作变得更加困难。从图 11 与图 12 的结果比对看来：网络的负载越重，测量的结果越紊乱，试图从相对时钟偏移的估计计算中同步测量器的时钟也就越难实现。

时钟的同步要依靠 CPU 对时钟信号的响应，而繁忙的负载工作使得时钟响应信号产生延迟，从而最终导致相对时钟偏移的非线性模型。为解决这一问题，可以提高测量器端的系统性能，增加其系统处理能力，使得时钟信号能够及时得到响应。但随着网络的飞速发展，仅靠优化端系统硬件的方式来解决问题是远远不够的。另一种方法是人为地调整时钟响应的优先级，使得时钟同步得以保证。然而做这样的调整的同时，测量器本身的实时测量能力则有可能受到影响。比较可行的一种方法是：首先，在测量器开始采集数据之前测量和统计两主机的相对时钟偏移模型。此时系统虽然面向高速网络，但采集工作尚未开始，主机的工作负载较轻，因而测量得到的相对时钟偏移将完全符合基准实验的线性模型。然后，在测量器完成数据采集工作之后再次测量和统计两主机的相对时钟偏移模型，此时的相对时钟偏移也呈线性模型。根据这前后两次的测量结果即可估算出数据采集工作中的相对时钟偏移值，以修正采集过程中得到的数据的相对时钟偏移量，完成网络中时钟的同步功能。

5 结论

时钟同步是提供准确的网络测度测量值的前提保证。在监测网络流量行为中，单向延迟等网络性能测度至少需要毫秒级的同步精度。由于一般的基准测量实验中，相对时钟偏移呈良好的线性模型，因而许多研究基于该模型对网络时钟同步问题提出各种解决方案。然而，在高速、大规模、分布式的网络环境下，高速的网络流量影响了系统的时钟响应信号，使得原本的相对时钟偏移模型受到影响和破坏。时钟响应信号的延迟使得即便使用 NTP 或者 GPS，都无法很好地解决网络时钟同步问题。

本文在分布式抽样测量服务级别约定(SLA)监测系统 PERME 测试环境中对高速 IP 网络中的时钟偏移现象进行了测量和分析。测量发现，网络的负载越重，端系统越繁忙，则时钟的偏移模型越不明显。本文中这些实验数据可以为进一步研究高速网络中的时钟同步问题提供原始参考资料。对于模型中的非线性，本文提出通过测量器主机测量前及测量后的相对时钟偏移的统计，进而估计计算测量中得到的数据的准确时戳值的解决方法。

参考文献

[1] Mills, D., “Network Time Protocol (Version 3) Specification, Implementation and Analysis”, RFC 1305, March 1992.

[2] 程光，大规模高速 IP 网络流量抽样测量及行为分析研究，东南大学博士学位论文，2003，pp：80-83.

[3] 邓勃编著，分析测试数据的统计处理方法，清华大学出版社，1995.

[4] J. Mahdavi, M. Mathis, “Framework for IP Performance Metrics”, RFC2330, May 1998.

[5] Vern Paxson, “On calibrating measurements of packet transit times”, in Proceedings of SIGMETRICS’98, pages: 11 - 21 , June 1998.

[6] Moon, S.B., Skelley, P. and Towsley, D., Estimation and Removal of Clock Skew from Network Delay Measurements. In Proceedings of the IEEE INFOCOM Conference on Computer Communications, page 227-234, March 1999.

[7] Li Zhang, Zhen Liu and Cathy Honghui Xia, Clock Synchronization Algorithms for Network Measurements. in: Proceedings of INFOCOM 2002, p. 6. 2002.

作者简历：

林容容，女，1981 年，硕士生，研究方向：网络测量；Email: rrlin@njent.edu.cn

丁伟，女，1963 年，教授，博士生导师，研究方向：网络行为学，网络安全，网络测量；Email: wding@njent.edu.cn

程光，男，1973 年，讲师，研究方向：网络行为学，网络安全，网络测量。Email: gcheng@njent.edu.cn

传真：83614842　　电话：83794000　　邮编：210096

英文图注：

Fig.1 Architecture of Timestamp Message Measurement System between Two Endpoints
Fig.2 Format of Timestamp Message
Fig.3 Non Synchronous Relative Clock Offset 1
Fig.4 Non Synchronous Relative Clock Offset 2
Fig.5 Fluctuate of Relative Clock Offset for the First time
Fig.6 Fluctuate of Relative Clock Offset for the Second time
Fig.7 Experimental Entironment of PERME System
Fig.8 Relative Clock Offset with Experimental Kernel
Fig.9 Relative Clock Offset with Kilomega Net Card
Fig.10 Relative Clock Offset with both Experimental Kernel and Kilomega Net Card
Fig.11 In the Collection ratio of 5 to 256
Fig.12 In the Collection ratio of 10 to 256

中文参考文献的英文译文：

[2] 程光，大规模高速 IP 网络流量抽样测量及行为分析研究，东南大学博士学位论文，2003，pp：80-83.
[3] 邓勃编著，分析测试数据的统计处理方法，清华大学出版社，1995.

[2] GuangCheng, Research on Traffic Sampling Measurement and Behavior Analysis in Large-Scale High Speed IP Networks，Doctor Thesis of Southeast University, 2003, pp：80-83.
[3] BoDeng，Statistical Method of Analyzing the Testing Data, TsingHua University Press, 1995.