基于 Hurst 参数的 DoS/DDoS 攻击实时检测技术研究

李金明[1,2],王汝传[1]
LI Jin-ming[1,2],WANG Ru-chuan[1]

1.南京邮电学院 通信工程系,南京 210003
2.北京锐安科技有限公司,北京 100037
1.Department of Communication Engineering,Nanjing University of Post and Telecommunications,Nanjing 210003,China
2.Beijing Ruian Science and Technology Co. Ltd,Beijing 100037,China

LI Jin-ming,WANG Ru-chuan.DoS/DDoS attack detection based on Hurst parameter.Computer Engineering and Applications,2007,43(6):20-23.

Abstract: Analyzed VTP method,the performance of real-time method based on VTP is evaluated,and the effectiveness of this method is proved.Using this method to calculate the Hurst parameter of traffic data set of Lincoln Lab of MIT,the law of change of Hurst parameter during DDoS attack is found,so a technology of detecting DDoS attack on real-time is proposed.
Key words: network security;DDoS;real-time detection;Hurst parameter

摘 要:在对 VTP 方法分析的基础上,对基于 VTP 的实时在线计算 Hurst 参数技术进行了性能分析,得出了其具有高效性的结论。并利用这种技术,对 MIT 的林肯实验室数据进行了分析,得出了 DDoS 攻击过程中,网络流量的自相似模型的 Hurst 参数变化规律,并由此总结出一种基于 Hurst 参数实时检测 DDoS 攻击发生的技术。
关键词:网络安全;分布式拒绝服务攻击;实时检测;Hurst 参数
文章编号:1002-8331(2007)06-0020-04 **文献标识码**:A **中图分类号**:TN914

1 引言

Dos/DDoS 攻击自从 20 世纪 80 年代第一次出现，就成了网络安全最令人头疼的问题。由于其利用网络协议或系统的一些缺陷,使得其易于实现,并且由于其伪造地址,使得确认攻击者的身份几乎是不可能的,因而也使得其成为网络黑客或是一些别有用心的人的终极网络攻击工具。2000 年 yahoo 等多家著名的网络公司的服务器遭受 DDoS 攻击,而在 2002 年,全世界的 14 个根域名服务器同时遭到 DDoS 攻击，其中的 8 个域名服务器被攻陷。在 2003 年 7 月份,以虚拟主机性价比享誉的商务中国(www.bizcn.com),一连遭受了几波极其猛烈的 DDoS 黑客攻击,在攻击过程中,商务中国托管在某地电信级 IDC 机房的服务器无一幸免。而到了 2004 年,开始出现一种趋势,那就是黑客们不再是盲目地想攻击哪家服务器就去攻击哪家服务器,而是带有了自己的经济利益。例如在 2004 年 10 月,先是一家网络游戏公司被黑客 DDoS 攻击,同时该站的李站长接到 Email,要求其在某个账户打入 10 万元钱。该公司的服务器托管在台湾的“中华电讯”,迫于无奈,最终关闭了服务器,并准备将服务器移至美国以躲避攻击。紧接着,腾讯公司的 QQ 服务器被 DDoS 攻击,对方同样提出了经济要求,来作为“修复费用”。黑客的这两次攻击,对网络游戏公司采用的是高速 DDoS 攻击,而对 QQ 服务器的攻击采用的低速的 DDoS 攻击。此后不久,QQ 服务器再次遭到了 DDoS 攻击,但这次的攻击更加狡猾,先是进行一段时间的高速攻击,然后再进行低速攻击,使得 DDoS 攻击的检测变得很困难。

正是因为 DDoS 攻击的日益猖獗，使得对 DDoS 攻击的防御，是目前网络安全的一项重要的也是极其困难的研究重点。为此,许多公司投入了大量的人力、财力。然而,如安全专家所言,就目前的情况来说,要想彻底杜绝 DDoS 攻击是不可能的,除非彻底的更改目前所使用的网络协议。就目前情况来说,这也不是很快就能实现的,因此,DDoS 攻击的防范技术研究也就显得尤为重要。

从目前的研究的情况来看,DDoS 攻击防御主要集中在这几个方面:DDoS 攻击源追踪技术、DDoS 攻击监测技术以及 DDoS 攻击包过滤技术。其中,DDoS 攻击监测技术是攻击源追踪技术与 DDoS 攻击包过滤技术的基础,只有快速准确地监测到 DDoS 攻击的发生，才能及时实施 DDoS 攻击源追踪与 DDoS 攻击包过滤。

在对麻省理工大学林肯实验室的数据进行分析的基础上,

基金项目:国家自然科学基金(the National Natural Science Foundation of China under Grant No.60573141,70271050);国家高技术研究发展计划(863)(the National High-Tech Research and Development Plan of China under Grant No.2005AA775050);江苏省自然科学基金(the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2005146);江苏省高技术研究计划(BG2004004,BG2005037);江苏省计算机信息处理技术重点实验室基金(kjs050001);江苏省高校自然科学研究计划(05KJB520092)。
作者简介:李金明(1972-),男,博士研究生,主要研究方向为计算机软件、计算机网络、信息安全等;王汝传(1943-),男,教授,博士生导师,主要研究方向是计算机软件、计算机网络和网格、信息安全、移动代理和虚拟现实技术等。

提出了基于 Hurst 参数的快速 DDoS 攻击检测技术。该技术是基于这样一种思想:正常的网络流量模型是符合自相似模型的,而 DDoS 攻击所产生的流量,将改变正常的网络流量的自相似特性,因而可以利用这一点来检测 DDoS 攻击的发生。

2 网络流量的自相似性

1994 年,Leland[1]等人对 Bellcore 的局域网测试与分析的结果显示,实际网络流量模型具有统计自相似性。这完全不同于以往的通信领域的传统的业务量模型——基于泊松(连续时间)或贝努利(离散时间)过程,这些模型是短时相关的。1995 年,Beran 等通过对大量的不同类别的可变比特率视频流数据的统计发现,它们也同样表现出一种长相关特性[2,4]。另外,对 WAN[3]、FASTPAC[5]等网络的测量,同样发现这些网络业务量表现出长相关的特性。而 A.Veres 等人通过模拟产生单个 TCP 对话流量的自相似现象[6],得出 TCP 流量控制机制也是一个能产生自相似现象的确定性因素。网络流量的这种特性,不仅使得传统的模型无法准确对其描述、分析,同时,也对实际的网络性能有着重要的影响。

那么网络流量的自相似特性是怎样产生的呢?通过对自相似现象成因进行分析后可知,在网络终端用户及多种业务的共同作用下,网络业务流量表现出的突发性是造成自相似性的主要原因[6,7],例如,网络用户的个体行为的突发性与随意性,文件的重尾分布等原因。而 A.Veres 等人通过模拟产生单个 TCP 对话流量的自相似现象[8],得出 TCP 拥塞控制机制也是一个能产生自相似现象的确定性因素。

目前有着多种对自相似过程定义,且它们并不是完全等价的。这里采用了文献[2,9]的描述:考察一个广义平稳过程 $X=(X_i,i=0,1,2,3,\cdots)$,X_i 表示第 i 个单位时间内到达的网络流量单元的数目(如到达的数据包的个数,或者到达的字节数),记 $\mu=E[X_i]$,$\sigma^2=E[(X_i-\mu)^2]$,自相关函数 $r(k)=E[(X_i-\mu)(X_{i+k}-\mu)]/\sigma^2$。令 $X_k^{(m)}=(X_{km-m+1}+\cdots+X_{km})/m$,$k=1,2,3\cdots$,称为 $\{X_i\}$ 的 m 阶的聚合过程,对每个 m,$X^{[m]}$都表示一个广义平稳随机过程,$r^{(m)}$为其对应的自相关函数。

定义 1 如果对所有的 m,m 阶聚合过程 $X^{[m]}$都具有与原过程 X 同样的相关函数结构,即:

$$r^{(m)}(k)=r(k)\sim k^{-\beta} \quad 0<\beta<1 \tag{1}$$

当 $k\to\infty$ 成立,则称 X 为精确二价自相似过程,并称 $H=1-\beta/2$ 为其自相似参数。也就是说至少 $X^{[m]}$与 X 直到二阶统计特性是不可分的。

定义 2 如果:

$$r^{(m)}(k)\to r(k)\sim k^{-\beta} \tag{2}$$

当 $k\to\infty$,$k=0,1,2,\cdots$则称 X 为渐近二价自相似过程,且具有自相似系数 $H=1-\beta/2$。参数 H 是表述自相似特性的唯一参数,其值越大,过程的自相似程度越高,取值范围是(1/2,1)。

对自相似过程参数的求解,前人已经提出了许多的方法,其中有 R/S 法[1]、Whittle 估计法[10]、基于小波分析的 EM 法[11],基于周期图的半参数估计法以及方差-时间图(Variance-Time,VTP)分析方法[12]。

3 Hurst 参数的 VTP 分析法

由定义 1、定义 2,可以得出相等价的式(3):

$$Var(X^{(m)})\sim \alpha m^{-\beta} \quad 当\ m\to\infty \tag{3}$$

式中 α 是一正常数,β 的含义与定义 1、定义 2 中相同。

对式(3)两边取对数得出:

$$\log(Var(X^{(m)}))\sim -\beta\ \log(m)+c \quad 当\ m\to\infty \tag{4}$$

由式(4)可以看出,当序列 X 是自相似过程时,如果 $m\to\infty$,则可以通过式(4)求出参数 β,再通过 $H=1-\beta/2$,就可以解出 Hurst 参数。具体的求解过程如下所示:

假设有一个自相似过程的 $X=\{X_i,i=0,1,2,3,\cdots,N\}$,将 X 中的每 m 个叠加,形成新的时间序列 $X^{(m)}=\{X_1^{(m)},X_2^{(m)},X_3^{(m)},\cdots\}$,其中:$X_k^{(m)}=(X_{km-m+1}+\cdots+X_{km})/m$,$k=1,2,3,\cdots$。从中也可看出,原始的时间序列 X 其实也就是 $m=1$ 的序列 $X^{(1)}$。例如:

$X=\{4,7,5,0,2,3,4,4,7,6,9,1\}$,

则有:

$X^{(2)}=\{(4+7)/2,(5+0)/2,(2+3)/2,(4+4)/2,(7+6)/2,(9+1)/2\}$
$=\{5.5,2.5,2.5,4,6.5,5\}$

$X^{(3)}=\{(4+7+5)/3,(0+2+3)/3,(4+4+7)/3,(6+9+1)/3\}$
$=\{5.3,1.7,5,5.3\}$

对每个时间序列 $X^{(m)}$,计算它的方差 $Var(X^{(m)})$:

$$Var(X^{(m)})=\frac{1}{N/m}\sum_{k=1}^{N/m}(X_k^{(m)})^2-(\frac{1}{N/m}\sum_{k=1}^{N/m}X_k^{(m)})^2$$

根据计算出的 $Var(X^{(m)})$与 m 值,可以作出$(\log(m),\log(Var(X^{(m)})))$图,则可以根据曲线的斜率 t 得出 β 值,$\beta=-t$。而 $H=1-\beta/2$,这样就可以估算出自相似时间序列 X 的 H 参数值。

在时间-方差分析法的基础上,文献[13]提出了一种在线实时求解 Hurst 参数的方法,正是在此基础上,得出了通过 Hurst 参数的在线实时求解,来在线实时检测 DDoS 攻击的发生。

实时求解真实网络流量的 Hurst 参数方法如图 1 所示,在图中,以 l 代表一次计算 Hurst 参数的时间序列长度,每计算一次 Hurst 参数以后,都将这个时间序列的前面长度为 a 的数据丢弃,而在后面重新添加新采样的长度为 a 的时间序列,再次计算 Hurst。这样,每一次计算 Hurst 参数,仅需要重新采样长度为 a 的数据即可,而不需要每次都要重新采样 l 时间长度的数据,不仅解决了计算 Hurst 参数所需的一定长度的时间序列问题,又解决了 VTP 方法中每次计算 Hurst 参数不能准确代表网络流量实时变化的问题。

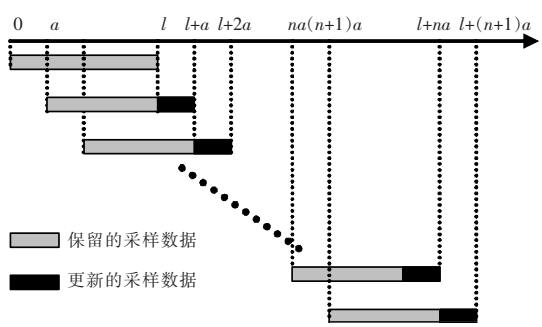

图 1 实时计算 Hurst 参数示意图

4 实验数据分析

此次分析所采用的数据是麻省理工大学林肯实验室的入侵检测系统数据库的数据。在这次实验中,整个网络模型被划分为 3 个域:inside(代表美国空军研究所的内部网络),outside

(代表美国空军研究所的外网部分),dmz(连接 inside 和 outside 的子网)。其中 outside 域包括有 Linux、Solaris、SunOS 及 MacOS 等多台主机,以及 SNMP Monitor、外网网关、外网 Web 服务器和思科 2514 路由器。dmz 域包括有多台主机与嗅探器以及防火墙、路由器等。在 inside 域中包括近 40 台的主机以及防火墙等。

在这次实验数据采集时间,被划分为 5 个阶段:(1)黑客探测主机 mill.eyrie.af.mil,这台主机是公用 DNS 服务器;(2)入侵主机 mill.eyrie.af.mil;(3)通过 FTP 上传 DDoS 攻击工具 Mstream 以及攻击脚本,并且侵入更多的主机,上传 Mstream 攻击工具的被控制端;(4)通过 Telenet 登录 mill.eyrie.af.mil,并初始化 Mstream 攻击工具的被控制端;(5)通过 Telnet 登录到 mill,并且 Telnet 本地的 6723 端口,连接到 Mstream 的攻击工具控制端,发动了一段时间的 DDoS 攻击。整个过程的实验数据如图 2 所示。

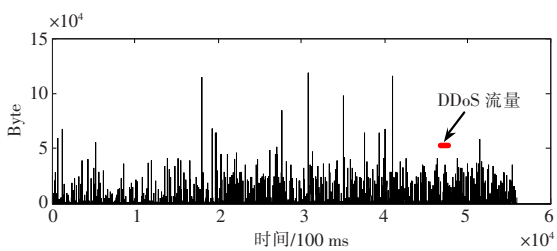

图 2　麻省理工大学林肯实验室数据示意图(100 ms)

将图 2 中的 DDoS 流量放大如图 3 所示。

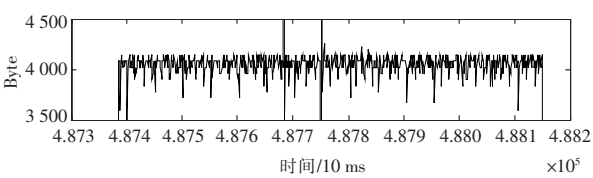

图 3　DDoS 攻击流量放大示意图(10 ms)

黑客在发动 DDoS 攻击前的网络流量,都是黑客在做 DDoS 攻击准备的网络行为所产生网络流量,相对于 DDoS 攻击的流量来说,这都是正常的网络流量。在对发生 DDoS 攻击前的数据流量分析可知,此时的网络流量模型的 Hurst 参数在正常的范围内小幅度波动,如图 4 所示。

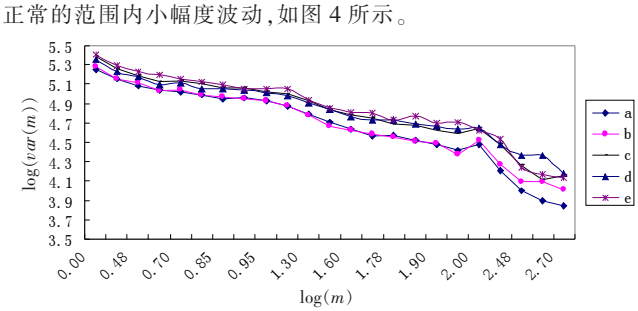

图 4　正常网络流量的 log(*var*(*m*))–log(*m*)图

图 4 中,a 系列的数据曲线是图 1 中所示长度 l(此处 l=10 000)是根据 VTP 分析方法计算出 log10(*var*(*m*))与 log(*m*)的对应值绘制的,聚合度 m 取值为 1、2、3、…、9、10、20、30、…、100、200、…、500。b 系列是 a 系列的数据丢弃图 1 中所示的长度为 a(此处=1 000)的数据,并在后面添加长度为 a 的新采集数据所构成的新的序列,计算出的 log10(*var*(*m*))与 log(*m*)的对应值绘制的。同理,c、d、e 序列的曲线都是依次处理所得。从图 4 中可以看出,这些曲线的斜率变化趋势基本是一致的,也就是说 β 值基本是差不多的,只是在小幅度内波动。由图 4 所得的各个序列的 β 值如表 1 所示。

表 1　正常流量各序列所对应的 Hurst 参数值

序列	a	b	c	d	e
H 参数值	0.796	0.785	0.807	0.823	0.815

也就是说,在正常网络行为下所产生的网络流量,其 Hurst 参数值是基本稳定的。并且其值也是在正常的自相似模型范围内。

再来看看当发生 DDoS 攻击时的网络流量对 β 值的影响。图 5 是从新添加数据中含有 DDoS 攻击流量开始的第一个序列计算出的一些 log(*var*(*m*))–log(*m*)图。

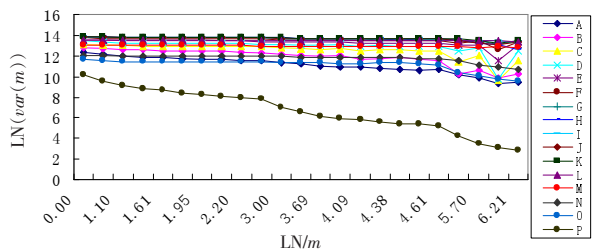

图 5　含有 DDoS 攻击流量的序列的 variance–time 图

为了更清楚地了解 DDoS 攻击流量对 Hurst 参数的影响,在此次的数据分析中,开始的 10 个序列,每次更新的数据长度 a 为总序列长度的 1%左右,而不是如前面的图 4 中所示序列的更新长度约为 10%。图 5 中,各个序列中 DDoS 流量占整个序列的流量的比例情况:A 序列 0%,B 序列约为 1%,C 序列约为 2%,D 序列约为 3%,E 序列约为 4%,F 序列约为 7%,G 序列大约为 6%,H 序列约为 7%,I 序列约为 8%,J 序列约为 9%,K 序列约为 10%,L 序列约为 75%,M 序列约为 97%,N 序列约为 98%,O 序列约为 99%,P 序列是完全由 DDoS 攻击流量构成。可以发现 P 序列的曲线斜率变化趋势明显不同于其他的曲线。把 P 序列去掉,再把图 5 上面部分的曲线进行放大,如图 6 所示,来看看当 DDoS 攻击流量占有比例逐渐加大时的各序列的曲线斜率变化情况。

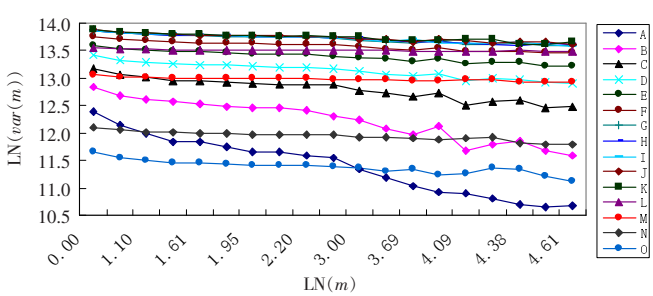

图 6　图 5 部分序列的放大图

从图 6 中,可以看出,从 A 序列(A 序列为正常的网络流量序列)的曲线开始,随着 DDoS 攻击流量的逐步增大,variance–time 图的曲线渐趋平缓,亦即 β 值越来越小,H 值越来越大。而且还可以发现,在开始的 1%、2%的 DDoS 攻击流量对正常流量的 H 参数的影响是很大的。然后,随着 DDoS 攻击流量逐步增大,其对 H 参数的影响逐渐变小,如表 2 所示。当

DDoS 攻击流量占到 10%时,*H* 参数已经达到了 0.976, 如表 2 的 K 序列所示 *H* 值。也就是说 DDoS 的攻击流量,在此时实际上增加了网络流量突发性。实际上,这也可以很直观地认识到,当 DDoS 攻击流量加入正常网络流量时,相对于正常的网络流量, 这是一批突发的网络流量, 因而加大了网络流量的突发特性。

表 2 含有 DDoS 流量各序列的 Hurst 参数值

序列	A	B	C	D	E	F	G	H
H 值	0.812	0.871	0.928	0.949	0.954	0.965	0.969	0.974
序列	I	J	K	L	M	N	O	P
H 值	0.973	0.975	0.976	0.994	0.990	0.972	0.960	0.443

再从表 2 的 L、M、N、O 这几个序列的 *H* 值变化情况来看,在 DDoS 攻击流量占有绝大多数时,当 DDoS 攻击流量进一步加大时,*H* 值在逐渐变小, 也就是说此时,DDoS 流量的比例逐渐增大, 在削弱网络流量的突发性。然而, 也可以看出来,当 DDoS 比例已经很大了时,如 O 序列,DDoS 流量已经占有整个序列的 99%,*H* 值仍然有 0.960,依然表现出极大的突发特性。这是因为,此时那部分的正常的网络流量,相对于 DDoS 攻击流量来说,是一种负方向的突发,因而,此时整个序列依然表现出很大的突发特性。当网络流量完全是 DDoS 攻击流量时,其表现出了另外一种特性,已经不再具有突发性,其 *H* 值也从 O 序列的 0.960 突降到 0.443。

从以上序列的分析过程中,可以看出,当正常序列中含有 DDoS 攻击流量时,随着 DDoS 攻击流量所占比例的逐步加大,Hurst 参数值是逐步增大的,并且是在初期变化明显,而后变化逐渐趋小。当 DDoS 攻击流量所占流量比例进一步加大到一定程度时,Hurst 参数值开始跟 DDoS 参数成反比例变化,但 Hurst 参数值依然很大,表现依然是网络流量的突发特性。但当网络流量完全是 DDoS 攻击的流量时,Hurst 参数会有一个突降的过程,而此时的网络流量特性也就不再具有突发性了。

如果按照图 4 所示的网路流量模型,每次更新的数据长度占到总的数据序列的长度的 10%, 则如表 2 中 A 序列直接到 K 序列,Hurst 参数变化有 0.164 之多, 比正常流量的 Hurst 参数小幅度变化要大得多。

5 结论与下一步的工作

从第 4 章的实验数据分析中可以看出,采用第 3 章所说的 VTP 分析法及在线实时计算 Hurst 参数技术,可以从计算结果中发现 DDoS 攻击流量对 Hurst 参数的影响: 在 DDoS 攻击刚刚开始时,Hurst 参数有增大的趋势,且变化较大,不同于正常网络流量的 Hurst 参数小幅度变化;当序列中的正常流量从有到无时,Hurst 参数有一个从很大到很小的突变, 如表 2 中,从 O 序列到 P 序列的 Hurst 参数突变。因而,从 Hurst 参数的变化上,可以判断出 DDoS 攻击的发生。

当然,要将这种方法实际应用,还有许多的工作要做。首先,如果黑客意识到这种检测技术,采用间歇式的 DDoS 攻击方法,将会增大检测难度;其次,如何确定最合适的序列长度以及更新序列长度也还有许多的工作要做。

(收稿日期:2006 年 12 月)

参考文献:

[1] Leland W E,Taqqu M S,Willinger W,et al.On the self-similar nature of Ethernet traffic(extended version)[J].IEEE/ACM Trans on Networking,1994,2(1):1-15.

[2] Beran J,Sherman R,Traqqu M S,et al.Long range dependence in variable bit rate video traffic[J].IEEE Trans on Communication,1995,43(2/3/4):1566-1579.

[3] Paxson V,Floyd S.Wide area traffic:the failure of poisson modeling[C]//Proc ACM Sigcomm'94,1994:257-268.

[4] Garrett M W,Willinger W.Analysis,modeling and generation of self-similar VBR video traffic [C]//Proc ACM Sigcomm'94,1994:269-280.

[5] Addie R.Fractal traffic:measurements,modeling and performance evaluation[C]//Proc of INFOCOM'95,Boston,MA,1995:977-984.

[6] Crovella M E,Bestavros A.Self-similarity in World Wide Web traffic-evidence and possible cause[C]//Proceedings of ACM Sigmetrics'96,1996:160-169.

[7] Willinger W,Taqqu M S,Sherman R,et al.Self-similarity through high-variability:statistical analysis of Ethernet LAN traffic at the source level[J].IEEE/ACM Transactions on Networking,1997,5(1):71-86.

[8] Veres A,Boda M.The chaotic nature of TCP congestion control[C]//Proceedings of the IEEE Infocom'2000,2000.

[9] Tsybakov B,Georganas N D.On self-similar traffic in ATM queues:definitions,overflow probability bound,and cell delay distribution[J].IEEE/ACM Trans on Networking,1997,5(3):397-408.

[10] Garrett M.Contribution toward real-time service on packet switched networks[D].New York:Columbia University,1993.

[11] Wornell G W,Oppenheim A V.Estimation of fractal signals from noisy measurements using wavelets[J].IEEE Trans on Signal Processing,1992,40(3):611-623.

[12] Zhang H F,Shu Y T,Yang O.Estimation of Hurst parameter by variance-time plots Communications[C]//IEEE Pacific Rim Conference on Computers and Signal Processing'10 Years PACRIM 1987-1997-Networking the Pacific Rim',1997(2):883-886.

[13] Hagiwara T,Doi H,Tode H,et al.High-speed calculation method of the Hurst parameter based on real traffic[C]//LCN 2000:Proceedings 25th Annual IEEE Conference on Local Computer Networks,2000:662-669.