第33卷 第15期 计 算 机 工 程 2007年8月
Vol.33 No.15 Computer Engineering August 2007
·网络与通信· 文章编号：1000—3428(2007)15—0121—03 文献标识码：A 中图分类号：TP393.08

基于抽样的IPv6高效地址扫描

李东宁，王振兴
(信息工程大学信息工程学院，郑州 450002)

摘 要：传统的地址扫描方法难以适用于IPv6巨大的地址空间，该文根据IPv6主机地址在地址空间中的分布状况，建立了随机地址扫描和抽样地址扫描2种效率分析模型。通过对2种模型的分析并结合IPv6网络特性，提出了基于抽样的高效地址扫描方法；通过在IPv6实验网中进行实验，证明了该方法是快速有效的。

关键词：网络安全；IPv6；地址扫描；抽样

Highly Efficient Address Scanning in IPv6 Based on Sampe

LI Dong-ning, WANG Zhen-xing
(School of Information Engieering, Information Engineering University, Zhengzhou 450002)

【Abstract】 As traditional address scanning approach is not available in huge IPv6 address space, this article establishs random address scanning and sample address scanning efficiency models based on IPv6 hosts addresses distribution in address space; By analyzing models and IPv6 network characteristics, a highly efficient address scanning approach based on sampling is presented; Through the test in an IPv6 experiment network, the approach is proved efficient.

【Key words】 network security; IPv6; address scanning; sample

对网络中存在的主机进行扫描，可以发现活动主机所使用的IP地址、开放的端口、使用的操作系统等信息。这些信息对于网络管理是必须的，因其是恶意攻击者进行网络入侵的前提，也是蠕虫传播过程中必不可少的部分。因此，关于网络扫描技术的研究对于网络安全具有重要的意义。

在IPv4网络中，IP地址长度只有32bit，通常采用遍历所有地址的方法进行扫描。当对一个含有255台主机的子网进行扫描时，假设每秒发出一个扫描数据包，那么完成对这个子网的扫描只需要大约5min。然而在IPv6网络中，IP地址长度扩大到了128bit，128bit的地址空间包含了2^{128}个可能的地址。即使是对1个含有2^{64}个地址的链路进行扫描，同样假设每秒发送1个数据包，也需要大约50亿年才能完成。因此，针对IPv6网络的地址扫描将会变得十分困难，同时也使得网络管理和蠕虫传播更加困难，所以，迫切需要研究针对IPv6大地址空间的高效扫描方法。

根据执行扫描任务的主机在网络中所处的位置，可以将IPv6地址扫描分为2种情况：(1)扫描主机与被扫描网络处于同一链路上，即对本地链路的扫描；(2)扫描主机与被扫描网络处于不同链路，即对远程网络的扫描。进行本地链路扫描时，扫描主机可以得到更多关于本链路的信息，而且可以使用更多的方法，所以相对较为容易，并且其某些方法具有一定的特殊性和局限性。而远程网络扫描则相对难度更大，其方法也具有通用性，所以，本文将着重讨论针对远程IPv6网络的地址扫描技术。

1 IPv6地址结构分析

IPv6地址分为单播地址、组播地址和任播地址3种类型。由于组播地址和任播地址都属于1个地址可以对应多台主机的情况，无法将1个组播地址或任播地址与1台特定的主机绑定，因此，地址扫描主要针对主机使用的单播地址。IPv6单播地址中又可分为全球单播地址、站点本地地址、链路本地地址等。站点本地地址和链路本地地址只能在站点内和本地链路内使用，不具有全球可路由的性质，下文讨论的是针对全球单播地址的扫描技术。

在全球单播地址结构中(图1)，前64bit是网络前缀，包含“001”、全球选路前缀、子网ID 3部分，Internet上的路由器通过它将数据包传送到主机所在的链路；后64bit是接口ID，用来区分同一链路中的不同主机[1]。在讨论远程网络地址扫描时，假设链路的网络前缀已从已知的主机地址中获得，即主机地址中的前64bit已知，扫描是针对地址中的后64bit，即对地址中接口ID的扫描。

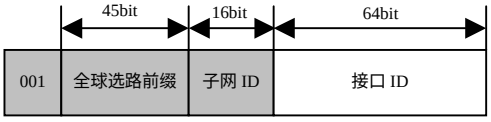

图1 全球单播地址结构

2 IPv6地址扫描效率分析模型

通过为地址扫描建立合理的数学模型，对影响其效率的各个因素进行分析，可以找出针对不同网络情况的最佳扫描策略。因为活动主机地址的分布状况在不同链路中是各不相同的，这也是影响扫描效率的关键，所以根据地址分布的特点，考虑地址均匀分布与非均匀分布2种情况，为地址扫描

基金项目：国家“863”计划基金资助项目(2003AA146010)
作者简介：李东宁(1983－)，男，硕士，主研方向：网络与信息安全，IPv6与下一代互联网；王振兴，博士、教授、博士生导师
收稿日期：2006-08-26 **E-mail：**ldnpla_wang@163.com

建立 2 种模型加以分析。

2.1 随机地址扫描模型

当链路中活动主机的地址均匀分布于地址空间中时，地址被扫描的先后顺序对整个扫描过程的效率没有影响，因此，可随机地选择被扫描的地址，传统 IPv4 中使用的地址扫描方法就属于这种情形。这一过程可以用随机地址扫描模型(random address scanning,RAS)来描述，估算扫描所用的时间。RAS 模型的符号说明如下：

H 为执行扫描任务的主机数量；

C 为扫描速率，执行扫描任务的主机每秒发送的数据包数量。C 与主机到被扫描网络的路径上的带宽有关，C 的值不能过大，否则会造成数据包的丢失，甚至可能造成网络的瘫痪；

V 为链路中活动主机的数量，这里假设链路中所有的活动主机均可以被扫描数据包检测到；

S 为扫描的地址空间，表示网络拥有的可能被使用的地址的个数；

$I(t)$ 为在时刻 t，已经被检测出的主机数量，$I(t)$是时间 t 的函数；

$S^{'}(t)$为在时刻 t，已经扫描过的地址空间，$S^{'}(t)$是时间 t 的函数。

在时刻 t，考虑 1 个时间间隔$\delta(\delta\to 0)$。在 δ 内，H 台主机总共可以发送 $H\cdot C\cdot\delta$ 个数据包，即 $S^{'}(t)=H\cdot C\cdot t$，而此时已经有 $I(t)$台主机被检测出来，则 $\frac{V-I(t)}{S-S^{'}(t)}$表示 1 个扫描数据包命中 1 台活动主机的概率。那么在间隔 δ 内新增的主机数量为

$$I(t+\delta)-I(t)=H\cdot C\cdot\delta\cdot\frac{V-I(t)}{S-S^{'}(t)}$$

$$\Rightarrow\frac{dI(t)}{dt}=H\cdot C\cdot\frac{V-I(t)}{S-S^{'}(t)}$$

$$\Rightarrow\frac{dI(t)}{dt}=H\cdot C\cdot\frac{V-I(t)}{S-H\cdot C\cdot t}$$

求解上面的微分方程，得到

$$I(t)=\frac{H\cdot C\cdot V}{S}\cdot t \qquad (1)$$

由式(1)可知，当主机地址均匀分布时，$I(t)$与 t 成线性关系。当进行扫描时，在 H，C，V，S 均已知的情况下，可以估算扫描所需要的时间。

2.2 RAS 模型的分析

由式(1)可知，在 t 相同时，增大 $I(t)$的方法有多种，可以增大 H、C、V，或是减小 S。但由于 H 和 C 受到可用资源和网络带宽的限制，不会比在 IPv4 网络中有很大的增长；V 值的上限是远程链路中所有活动主机的个数，也不可能无限制地增长。因此，提高扫描效率的可行方法是考虑如何减小地址空间 S，而 IPv6 网络所具有的一些特性和网络配置中人为因素的影响可以帮助减小 S。

2.2.1 利用接口 ID 由 MAC 地址生成减小 S

当使用无状态地址自动配置时，IPv6 主机地址的接口ID可以基于媒体访问控制(MAC)地址来生成。MAC地址长度是48bit，由 24bit的公司ID和 24bit的扩展ID组成。接口ID中的其余 16bit是固定不变的，因此，在对这类接口ID进行扫描时，S可由 2^{64}减为 2^{48}。通过进一步对IEEE为生产网卡的厂家分配的公司ID列表进行分析，可以发现 24bit的公司ID并没有全部被分配出去。截止 2006 年 3 月，已分配的公司ID个数为9032[2]，因此，$S=9032\cdot 2^{24}\approx 2^{37.2}$。

对于一部分公司和组织，其局域网中所使用的网卡很有可能出自少数几个生产厂家。通过对 2 个不同类型的大型局域网进行采样分析，1 个属于教育机构，具有 161 台主机；1 个属于社团，具有 227 台主机，它们都大约只具有 40 种不同厂家生产的网卡[3]。对于每 1 个公司ID，可以尝试只对MAC地址中的扩展ID部分进行扫描，S由 $2^{37.2}$减为 2^{24}。

2.2.2 利用 DNS 服务减小 S

由于IPv6 地址长度为 128bit，并不容易记忆，而且一些主机还可能会有多个地址，这就使得DNS在IPv6 中变得更加重要。当链路中的大部分主机都配置有 1 个域名时，可以将对主机IP地址的扫描转化为对主机域名的猜测[4]。

利用 DNS 的扫描可分为 2 个步骤：(1)由扫描主机生成一个域名字符串，向 DNS 服务器发送域名查询以判断此域名是否对应链路内的 1 台主机；(2)如果得到 DNS 服务器的回应，并且确定此域名对应链路内的一个主机地址，那么扫描主机再向得到的地址发送扫描数据包，以确定该地址对应的主机目前是否处于活动状态。其中，如何猜测域名字符串是这种方法能否成功的关键。

因为被扫描的主机处于同一链路，所以主机域名也可能存在着某些联系，在已知 1 台主机域名的情况下，可以尝试对其它域名进行猜测。例如，已知 1 台主机的域名是“xxx.yyyyy.zzz”，链路中的其它主机域名可能仅是与此域名在“yyyyy”部分有所不同，那么就可以在生成域名字符串时，只改变“yyyyy”部分的内容。这里假设猜测字符串的长度为 5，猜测字符由 26 个英文字母和 10 个阿拉伯数字组成，那么地址空间S的值可以认为是 36^5。

2.2.3 利用密集的编址规律减小 S

当网络中的主机地址由管理员手工配制，或是由DHCP服务器分配时，地址的接口ID可能被设置为人容易记忆的有规律的形式。例如，接口ID的生成不使用MAC地址映射，也不是随机生成，而是将链路内的主机进行顺序编号[5]。例如将一台主机地址设置为[Prefix]::1，而将另一台主机的地址设置为[Prefix]::2，如此顺序下去。因此，当远程链路中已知主机地址形如[Prefix]::xxxx:yyyy，或是[Prefix]::xxxx时，可以尝试只在“xxxx:yyyy”的范围里或“xxxx”的范围里顺序地进行扫描。这时，S将减为 2^{32}甚至是 2^{16}。

2.3 抽样地址扫描模型

通过上述讨论可知，因为 IPv6 网络提供了巨大的地址空间，所以活动主机地址的分布通常是非均匀的，此时如果仍然使用 RAS 模型来讨论，随机地选择被扫描地址进行扫描，则得不到理想的扫描效率。这里可以使用抽样地址扫描模型(sample address scanning, SAS)来描述扫描过程，引入统计学中的方法，采用整群抽样的思想，先将大地址空间划分为若干个大小相等的子空间区域，对每个区域进行抽样，利用得到的结果对各个区域内含有活动主机的概率进行预测分析、排序，得出拥有活动主机相对较多的区域，然后合理分配扫描资源优先对这些区域进行扫描，从而有效地提高扫描效率。

2.3.1 通过抽样计算概率

对于被扫描的地址空间S，可将其划分为m个子空间区域 $S_k(1\le k\le m)$。通过抽样试验，对S_k中每个地址被活动主机使用的概率 P_k 进行预测。将对S_k发送n个扫描数据包看作是 1 次抽样，即样本容量为n，其中，扫描数据包探测的地址均匀地分布于S_k中，统计其返回结果，进而估算出 P_k 。设随机变量

$X_i(1 \le i \le n)$，定义如下：

$$X_i = \begin{cases} 1 & \text{第 } i \text{ 个数据包命中 1 台主机；} \\ 0 & \text{第 } i \text{ 个数据包命中 1 台主机。} \end{cases}$$

因为 X_i 服从(0-1)分布，所以

$$P\{X_i = x\} = P_k^x \cdot (1-P_k)^{1-x}, x = 0,1$$

$$E(X_i) = P_k, D(X_i) = P_k \cdot (1-P_k)$$

令 X 表示抽样中扫描到的主机数量，

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

因为通常情况下，$S_k >> n$，所以试验中每个扫描数据包是否命中主机可以近似看作是独立的过程。因此，有

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = n \cdot P_k \qquad P_k = \frac{E(X)}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X} \tag{2}$$

通过上述分析，可以看出 SAS 模型的适用条件是链路内活动主机数量相对较多，或是活动主机相对集中的分布在少数区域中的情形。因为不满足上述条件，很可能会出现抽样之后，利用式(2)得出的概率为 0 或十分接近 0 的情况，从而无法对概率进行排序，抽样也就没有意义了。当对各个区域进行一次抽样之后，无法得出理想的概率排序时，可以重复进行抽样，直到挑选出含有活动主机相对较多的区域。

2.3.2　相关参数的讨论

在地址空间相同的情况下，划分的粒度越细，选取的样本容量越大，得出的 P_k 值就越准确。但 RAS 模型并不要求得出 P_k 准确值，而只是对 P_k 的大小进行排序。另外，如果 m 选得过大，就会使得抽样算法的计算机实现变得复杂，占用过多的计算资源，实际操作中应结合 S 的大小、扫描资源和网络带宽确定 m 的取值。例如，当扫描主机开启 16 个线程来进行扫描时，可以取 m=16。同样如果 n 选得过大，花费在抽样上的时间就会太多，造成在含有少量或不含活动主机的区域中花费过多的时间和资源的情况，从而影响扫描的效率。所以，适当的选择 n 也是十分重要的，下面讨论关于 n 的计算公式。

根据统计学，对于来自(0-1)分布的总体 X，因样本容量 n 比较大，由中心极限定理，知

$$\frac{\sum_{i=1}^{n} X_i - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} = \frac{n \cdot \bar{X} - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}}$$

近似地服从 N(0,1)分布，于是有

$$P\left\{-z_{\alpha/2} < \frac{n \cdot \bar{X} - n \cdot p}{\sqrt{n \cdot p \cdot (1-p)}} < z_{\alpha/2}\right\} \approx 1-\alpha$$

$z_{\alpha/2}$ 表示置信水平为$(1-\alpha)$的双侧置信区间的置信上限。由此，可以得到在不重复抽样的条件下，允许误差 Δ 的公式

$$\Delta = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

从而得

$$n = \frac{N \cdot z_{\alpha/2}^2 \cdot \sigma^2}{(N-1) \cdot \Delta^2 + z_{\alpha/2}^2 \cdot \sigma^2} \tag{3}$$

其中，σ^2 表示样本的方差，$\sigma^2 = p_k \cdot (1-p_k)$；$N$ 表示总体容量，这里 $N = S_k$。由于在抽样前，P_k 的值无法确定，这里可以用其最大值 $p_k = V/S_k$ 来代替，则有 $\sigma^2 = V/S_k \cdot (1-V/S_k)$。$z_{\alpha/2}$ 的值可以在确定置信水平后查表获得。因为利用抽样的结果只是为了得到不同 P_k 之间的顺序，并不要求得到十分准确的 P_k，所以允许误差 Δ 可以取最大值，设 $\Delta = p_k = V/S_k$。因为通常情况下，有 $V/S_k \to 0$，所以 $(1-V/S_k) \to 1$。将 σ^2 和 Δ 代入式(3)，并化简得

$$n \approx \frac{z_{\alpha/2}^2 \cdot S_k}{V} \tag{4}$$

至此，已得到了关于 n 的计算公式，在模型应用中可以根据实际情况改变相关变量的值，从而得到不同的 n。

3　2 种模型的实验分析

在校园 IPv6 实验网中，通过对主机地址不同的配置，分别模拟 RAS 模型和 SAS 模型的适用环境，在网络带宽、扫描资源等其它条件相同的情况下，分别使用 RAS 模型和 SAS 模型给出的扫描方法进行实验，对结果进行比较分析。

3.1　模拟 RAS 模型适用环境的实验

RAS模型适用于活动主机地址均匀分布的情况，实验中将主机地址的接口ID配置成由MAC地址生成，由于实验网内的机器大都出自同一厂家，因此MAC地址的公司ID是相同的，只对MAC地址中的扩展ID进行扫描，并且近似地认为扩展ID是均匀分布的。取S=2^{24}，m =16，根据式(4)，取置信水平为 0.95，可得n=40 000，其余参数见图 2。

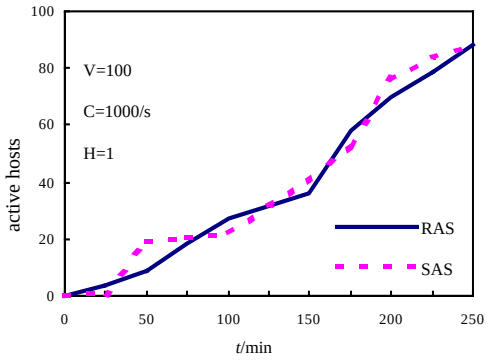

图 2　模拟 RAS 模型适用环境的实验结果

由图 2 可知，当活动主机地址均匀分布时，RAS 模型和 SAS 模型的效率基本上是相同的，由于扩展 ID 并不是真正的均匀分布，因此 RAS 模型的结果并不是一条直线。

3.2　模拟 SAS 模型适用环境的实验

SAS模型适用于活动主机地址非均匀地分布在地址空间中的情况。实验中将网内所有主机的地址限定在“2006::80:0000~2006::80:9999”范围内，使用密集编址方法。仍然取S=2^{24}，m =16，n=40 000，其余参数如图 3 所示。

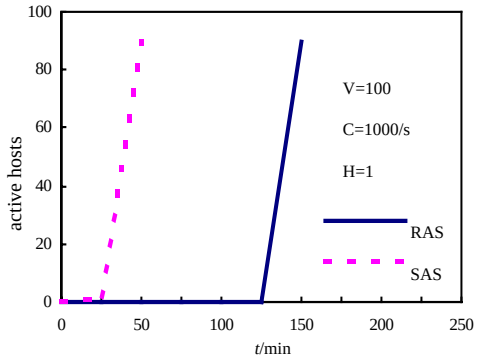

图 3　模拟 SAS 模型适用环境的实验结果

由图 3 可知，在对主机地址非均匀地分布的网络进行扫描时，SAS 模型比 RAS 模型的效率提高了很多，对于实验中的这种网络配置，扫描时间节省了约 80%。

（下转第 126 页）