第34卷 第13期 计 算 机 工 程 2008年7月
Vol.34 No.13 Computer Engineering July 2008
·网络与通信· 文章编号：1000—3428(2008)13—0101—03 文献标识码：A 中图分类号：TP393

园区网边界流量采样及其可视化研究

夏晓忠[1,2]，肖宗水[1]，刘志刚[3]，仇一弘[1]
(1. 山东大学计算机科学与技术学院，济南 250061；2. 中国人民解放军72433部队，济南 250014；
3. 中国人民解放军武装警察部队山东省总队，济南 250014)

摘 要：基于NetFlow流技术通过提取园区网边界数据流的地址、端口、协议和流量等特征属性在三维空间中建立流的几何可视化模型，简化了网络流量的显示，设计了基于地址段*A*的可变坐标几何建模算法，优化了模型的渲染，为网络管理人员提供了一种通过流的3D可视化图形来感知网络运行性能的手段。实验表明该模型有效反映了网络流量的相关特征。

关键词：网络管理；流量；NetFlow流技术；可视化；建模

Study of Boundary Traffic Sampling and Visualization for Campus Networks

XIA Xiao-zhong[1,2], XIAO Zong-shui[1], LUI Zhi-gang[3], QIU Yi-hong[1]
(1. School of Computer Science and Technology, Shandong University, Jinan 250061;
2. Unit 72433 of the Chinese People's Liberation Army, Jinan 250014; 3. Chinese People's Armed Police Force in Shandong Province, Jinan 250014)

【Abstract】 Based on the NetFlow technology, this paper submits a visible flow geometric model in space of three dimensions by using some characters of campus network's boundary traffic. Those characters include IP, port, protocol and volume of flow. This model simplifies display of network's traffic. An algorithm of *A*-based alterable coordinates modeling is designed to optimize model rendering. The model and algorithm provides a method for network manager to apperceive network performance by 3D visible graph of flow. Experiment shows that the model effectively reflects relevant characteristics of traffic.

【Key words】 network management; traffic; NetFlow technology; visualization; modeling

园区网一般规模较大、结构复杂，网络边界通信量巨大、成分多样，如仅使用统计数据和报表，则往往会使网络管理员深陷在抽象的数字和复杂的相互关系之中，难以发现潜在的安全隐患和性能瓶颈。目前人们倾向于采用可视化技术来了解网络运行及被访问的整体情况，并对其中存在的问题进行分析。要进行可视化的前提和关键是网络的拓扑结构、流量等一些基础数据的提取，然而园区网的复杂性使得人们难以从中得到完整、准确的信息，而将获得的可用数据合理地组织、存储并加以利用就更加困难。本文针对以上需求设计了一个基于NetFlow流采样的园区网边界流量可视化原型。

1 相关知识

1.1 NetFlow

传统上通过SNMP协议从网络设备收集流量数据，得到的只是粗糙、简略的数据，只能让管理者发现问题，却无法进一步解决问题。网络探针等监听工具可以捕捉流过的所有数据包并加以翻译，找出数据包头中字段的相关信息，为进一步分析其内容还可以将所捕获到的数据包存储起来，分析其流量的内容和关联性，甚至可以再现网络流量，跟踪用户行为、确定故障责任，为管理员提供反映网络性能的详实信息。但监听工具通常专注在单一网络数据包的内容上，使网络管理者很难从所提供的信息来掌握整体网络的状态。此外，分析数据包非常耗费时间，而且数据包监听所储存并需要分析的数据量非常庞大，对资源和人员的消耗是惊人的，这种方式显然不适合大流量环境下的网络管理。

NetFlow协议是由Cisco公司开发的一套网络流量监测技术[1]，通过网络中的交换设备采集所有当前经过的流数据并将其存放到自身的缓存中，然后按一定的格式发送给指定的服务器，为网络流量计量、用户计费、网络规划、监控和数据挖掘提供依据。利用这种高性能设备的流缓存方式能很好地避免普通采集模式的丢包、网络带宽及运算资源过重的问题，保证了数据采集的完整性。

1.1.1 协议架构

NetFlow以流为数据统计的采集单位，流是一个特定来源和目的端的单向数据报文序列，也就是具有相同来源/目地地址、源/目地端口、传输协议、Tos字节和数据流入接口7个属性的报文整合成一个流。其协议的核心是对流缓存进行组织、管理，最终可提供遵循某种汇聚方法而得到流的统计数据。其处理方法为：在一定时间内，将流中的数据报文按照一定的聚类规则汇聚形成原始数据置于缓存中。超时时间到达或者缓存充满时，这些数据以UDP数据包按一定的格式发送到网络上指定的接收者。

利用NetFlow协议执行流量采集的系统一般由3个部分组成：

(1)产生NetFlow流量数据的路由器或者交换设备；

(2)用来采集流数据的设备，称为流收集器；

(3)利用流数据进行各种数据分析和处理的设备，称为流分析器。

作者简介：夏晓忠(1975－)，男，硕士研究生，主研方向：计算机网络管理；肖宗水，副教授；刘志刚，工程师；仇一弘，硕士研究生
收稿日期：2007-08-17 **E-mail**：xiaxiaozhong@sdu.edu.cn

1.1.2　协议数据格式

NetFlow 协议目前包括多个版本，版本之间差异主要表现在对流采用的汇聚方法不同。以网络监控或规划为目的而部署 NetFlow 要求获得流的较多细节，因此常采用版本 5。该版本采集到的流数据可以支持不同维度的统计分析，数据流信息格式及说明参见文献[1]。

1.2　流量可视化

种类繁多的信息源产生的大量数据，远远超出了人脑分析解释这些数据的能力。因此，美国计算机成像专业委员会提出了解决方法——可视化。可视化把数据转换成图形，给予人们深刻且意想不到的洞察力，在很多领域使科研人员的研究方式发生了根本变化。

可视化的主要过程是建模和渲染。建模是把数据映射成物体的几何图元。渲染是把几何图元描绘成图形或图像。将可视化过程应用于园区网边界网络流量的主要技术就是合理选取可以有效反映网络特性的流量特征，对其进行多维标量建模，然后在屏幕上渲染出来，以加深网络管理员对流量数据含义的理解、指引检索过程、加快检索速度，使其在杂乱无章的海量数据中发现其中隐藏的问题。

2　流量可视化技术

2.1　地址空间的基本可视化建模

原型从数据流中抽取 4 个特征数值，分别是源 IP 地址 *Is*、目的 IP 地址 *Id*、源端口数值 *Ps* 和目的端口数值 *Pd* 作为基本参数，在三维空间中对一个数据流建立模型。其中对 IP 地址在二维笛卡儿坐标平面中按一定规则进行分配，表示为(*Xip, Yip*)，端口数值均匀分布在三维笛卡儿坐标系的 *Z* 轴的正方向上，表示为 *Zp*，则一个数据流的源端点表示为(*Xips, Yips, Zps*)、目的端点表示为(*Xipd, Yipd, Zpd*)，然后用一条直线连接 2 点。这样就可以在三维空间坐标系第一象限上对一个流进行基本的可视化建模。

利用这 4 个数值作为笛卡儿空间的参数，就可以在三维空间里绘制每个流，使每个流成为一条线段。对于网络中一对主机间的一次会话由于 2 个方向数据包的源地址、目的地址、源端口和目标端口互异，NetFlow 将其归并为 2 个流，分别记录在数据库中，在空间坐标系中，由于 2 组参数值相同，2 个流将重合为同一线段。

2.2　协议和流量的可视化建模

流可视化模型如图 1 所示。

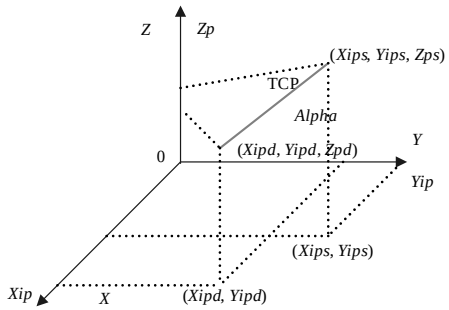

图 1　流的可视化模型

协议和流量是流的重要特征，不仅是信息统计的重要来源，而且通过两者不同的组合，及流的时序和地址分布可有效发现各种蠕虫的攻击行为。原型对协议的建模是通过线段的颜色定义的，对每种协议结合各色彩的心理视觉敏感度[2-3]的高低分配特定的*RGB*值，使重要和感兴趣的协议在模型中较为突出；通过线段的*Alpha*通道值对流的量进行表征，可较为直观地感知流量的大小。

2.3　地址段平移的均匀分布策略

在对山东大学南校区学生宿舍路由交换机连续 24 h 进行 1:1 比率的 NetFlow 流采样，得到一个 8 039 451 条流的记录集，其中交换机所转发的 IP 地址数量统计及分布情况见表 1。

表 1　转发地址统计

分类	IP 地址个数
源 IP 地址	925 766
目的 IP 地址	1 246 307
网内 IP 地址	1 290
网外 IP 地址	1 498 268

根据表 1 可知，交换机转发IP地址与地址空间的比率为(网内IP地址+网外IP地址)/2^{32}≈0.000 35，不足地址空间的万分之四，说明了园区网边界转发地址的有限性。

以太网 IP 地址的点分十进制表示形式为 *A.B.C.D*，其中 *A, B, C, D* ∈ [0~255]，对边界交换机转发地址结构进行统计，发现其中地址 *A* 段部分各数值出现几处较大的转发量，而其余部分十分平坦且几近为零，反映出 IP 地址的转发具有局部性。这种现象的出现是由于全球 IP 地址分配的有序规划使得国家、地域、组织获取相对集中的地址资源，以及人们在使用网络中表现出规律性行为所导致的综合表现。

从网络边界流量转发地址有限性和局部性的角度出发，原型提出一种基于地址段 *A* 的可变坐标几何建模算法(algorithm of alterable coordinates modeling based-*A*)，以便将转发地址均匀分布在三维空间中。原型将地址分为 2 部分，即 *A.B* 和 *C.D*，映射到二维几何平面上，分别放置在 *X* 轴和 *Y* 轴。由于 *C, D* 2 段各数值在边界转发过程中出现的频率几乎相等，因此将其所表示的 65 536 个数值均匀分布在 *Y* 轴上；*B* 段各数值在样本统计中具有一定的波动性，但幅度较小，亦适合采用均匀分布方式；*A* 段各数值分布不均，波动幅度很大，采用可变坐标的方式处理。基于地址段 *A* 的可变坐标几何建模算法如图 2 所示。

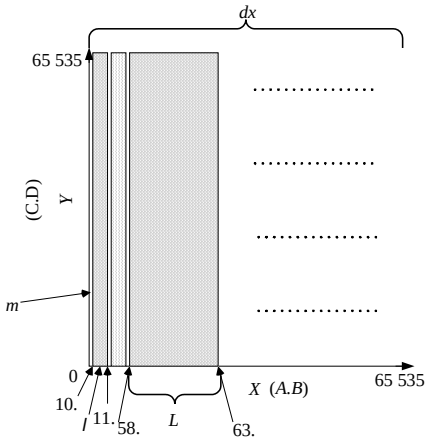

图 2　地址段 *A* 可变坐标图

设*X*轴显示长度*dx*，阈值*k*≥0，在一次模型渲染期中，所需渲染的边界转发地址*A*段各数值a_i，*i*∈[0, 255]，其全集*A*= $a \cup \bar{a}$。N_i为渲染期a_i出现次数的统计量，若$N_i \geq k$，则$a_i \in a$，否则$a_i \in \bar{a}$。*N*表示集合*a*中元素的个数，$\bar{N}$表示集合$\bar{a}$中元素的个数。l_i为a_i渲染宽度，L_{ij}为地址段$a_i a_{i+1} \cdots a_j$的渲染宽度。其中$j > i$, $\bigcup_{t=i}^{j} a_t \in a$。若$a_i \in \bar{a}$，则$l_i$=*m*, (*m* ≥ 1)，否则

$l_i=\frac{dx-m\cdot N'}{N}$，$L_{ij}=l_i\cdot(j\text{-}i+1)$。由此确定地址$a_i$在$X$轴的渲染位置和渲染宽度，然后将$a_i$地址段中所有$B$段地址均匀分配在$l_i$上。$C.D$均匀分配在$Y$轴上。

k 刻画的是地址转发频率对该地址在建模过程中不同处理的界限，对于转发频率小于 k 的地址，将粗略渲染在宽度为 m 的范围内。k 和 m 的取值应根据实际网络边界流量进行选取，对于中等校园网分别设为 100 和 1 较为适合。

该可视化建模较现有的其他网络流量可视化模型，如 VisFlowConnect[4], NVisionIP[5]等，能更充分利用图形图像的视觉元素，集中提供更多的流量信息。

3 整体结构设计

原型的体系结构由 4 部分组成：流发生器(支持流的路由器或交换机)，流收集器，流分析器和流可视化生成器。原型的体系结构及数据流向如图 3 所示。

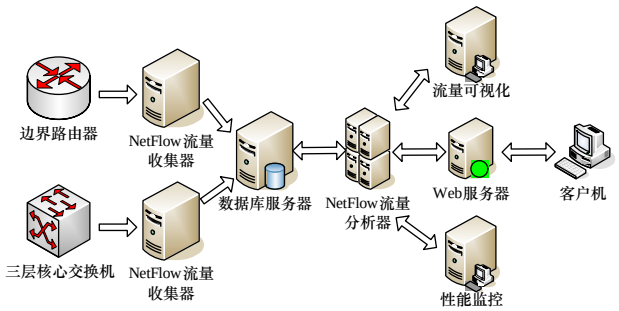

图 3　体系结构及数据流向

流发生器是由支持 NetFlow 的路由器或交换机担任，用于依不同应用向流收集器发送相应版本的 NetFlow 流，版本 5 可满足目前一般应用需要；流采集器和分析器运行在高性能的主机上，完成流的采集、存储和分析任务；流可视化生成器可作为独立客户端程序与数据库和分析器一起组成完整的可视化应用，也可在 Web 服务器端形成 B/S 模式的流量可视化解决方案，由具有浏览器的客户机随时访问。对于原型所获得的流数据也可为其他网络流量计费、性能监控及评估应用提供基础性的数据，从而组成通用、开放的网络信息平台。

4 具体实现

4.1 NetFlow 数据库设计

原型运行在 Linux 平台上，使用 MySQL 数据库。由于大型园区网络边界设备所转发的流的数量往往是巨大的，每分钟可达数千条，甚至上万条。为对流进行高效处理，在数据库的设计方面采用以下措施：

(1)规划表字段，精简字段数量；有效设置索引，优化查询。设置了源/目的地址、源/目的端口、协议、数据包、字节数、插入时间戳和更新时间戳字段，索引字段为前 5 个。

(2)对边界流的进、出流量分别存储，进一步减小数据表大小。

(3)提高边界流采样比率可以显著减少存储开销。这需要平衡存储空间和观测精度之间的关系，一般可设置在 1:500 之内。

(4)定时建数据表，设定每小时创建一个由月-日-时为名的数据表。定时对过期数据表进行处理，维持数据库规模。

4.2 可视化渲染算法的主要步骤

流的渲染首先要确定合理的渲染范围，转发量较大时渲染范围过大往往会使异常的或是感兴趣的流湮没在正常流的渲染过程中，原型选取每分钟对流模型进行一次渲染。具体步骤如下：

(1)读取近 1 min 的流信息；

(2)初始化三维坐标轴，设定各坐标轴显示长度、投影视角；

(3)取流记录，对流运行基于地址段 A 的可变坐标几何建模算法确定流的源节点和目的节点的平面坐标位置；

(4)根据流的源端口和目标端口确定其三维空间模型；

(5)对协议和流量进行建模；

(6)在屏幕上渲染该流；

(7)全部渲染完毕退出，否则转(3)。

5 结束语

园区网边界流量采样及可视化系统对山东大学学生宿舍路由交换机发送的 NetFlow 流进行采样，以监测该网络的运行情况，其 C/S 模式下的客户端采用 Delphi 开发，流的渲染如图 4 所示。

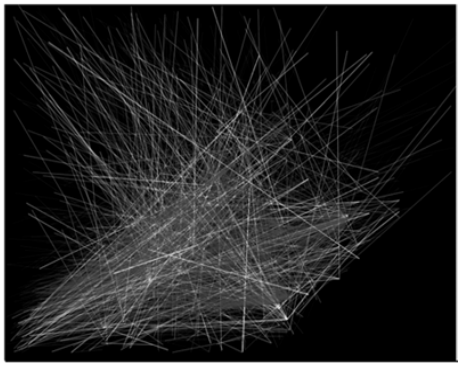

图 4　流的渲染效果

通过流的可视化渲染后，网络管理员可以直观地发现网络性能瓶颈和蠕虫早期特征[6]，但这需要管理员具有丰富的管理经验和网络知识。因此，对流可视化图像的智能判别和自动预警将是下一步重点研究的方向。

参考文献

[1] Cisco Inc. NetFlow[EB/OL]. [2007-07-17]. http://www.cisco.com/en/US/tech/tk812/tsd_technology_support_protocol_home.html.

[2] Gonzalez R C, Woods R E. Digital Image Processing[M]. 北京: 电子工业出版社, 2002.

[3] Berin B, Kay P. Basic Color Terms: Their University and Evolution[M]. Berkeley, CA, USA: University of California Press, 1991.

[4] Yin Xiaoxin, Yurcik W, Treaster M, et al. Visflowconnect: Netflow Visualizations of Link Relationships for Security Situational Awareness[C]//Proc. of 2004 ACM Workshop on Visualization and Data Mining for Computer Security. Washington D. C., USA: [s. n.], 2004: 26-34.

[5] Lakkaraju K, Yurcik W, Lee A, et al. Nvisionip: Netflow Visualizations of System State for Security Situational Awareness[C]//Proc. of 2004 ACM Workshop on Visualization and Data Mining for Computer Security. Washington D. C., USA: [s. n.], 2004: 65-72.

[6] Kim H. Kang I. Real-time Visualization of Network Attacks on High-speed Links[J]. IEEE Network, 2004, 18(5): 30-39.