高速网络流测量平台 WATCH1.0 处理器的结构设计

周明中 丁伟 高亚东

（东南大学计算机系 江苏省计算机网络技术重点实验室 江苏南京 210096）

摘 要 高速网络流测量平台 WATCH1.0 为基于流的网络行为研究提供了基本的试验条件。本文通过介绍 WATCH1.0 处理器的结构框架设计，分析处理器的性能并具体探讨可能出现的瓶颈及其对应的解决方案，并进一步明确了未来的研究方向。

关键词 高速网络 被动流测量 处理器 结构设计

High-speed network flow measurement platform WATCH1.0's disposer structure design

ZHOU Mingzhong DING Wei GAO Yadong

Abstract High-speed network flow measurement platform WATCH 1.0 provides the basic test condition for the flow based network behavior studies. This paper analyses performances of the disposer via introducing its structure design. And the bottlenecks of performance and theirs solutions are discussed deeply in this paper. At last, it is emphasized that the direction of future research direction.

Keywords High-speed network; passive flow measurement; disposer; structure design

1 引言

随着互联网技术的发展，对网络流量行为进行观测和分析的需求变得十分迫切。相应的网络测量工具层出不穷，如 Caida 的 CoralReef[1]，Berkley 的 libpcap[2]，UCSD 的 PMA[3]等等，但这些工具具有通用性较强但一般均只针对数据报文进行处理。而目前对网络行为的研究已从单纯分析数据报文特性转向报文和流特性分析并举，所以加强网络流量测量工具的流分析能力变得尤为重要。为此 CERNET 华东（北）网络中心组织开发了高速网络流测量平台 WATCH1.0。

2 WATCH1.0 处理器结构设计

WATCH1.0 是一个通过高速光纤以太网流量分光方式获取流量镜像并对其进行流分析的网络流测量平台。它基于被动测量技术，能在一段连续时间内，以低丢包率完成对双向全部 IP 报文头部的完全捕捉、组流、存储和简单测度计算，并以固定的格式存储，以便以后任意的分析程序的执行。由于 WATCH1.0 的处理对象是 IP 报文头部信息，在不引起歧义的情况下，本文所提到“报文”均代表 IP 报文头部信息（有特殊附加说明的除外）。

2.1 WATCH1.0 功能和总体结构简介

近年来高速网络流量的信息分类、分析和估计已经成为研究热点之一[6][7][8]，WATCH1.0 实现了三种不同粒度流量数据采集和存储：原始报文头部信息、报文流信息和流量统计测度信息。原始报文信息是用来分析网络性能和流量特性的最基本单位，但是由于对其进行存储、处理和传输所需要的资源是十分巨大的，所以只能有选择地进行存储；报文流信息是对原始报文信息的抽象和归纳，所保留的信息要小于原始报文信息，但是其所需资源也远小于原始报文的存储；流量统计测度信息则保持了固定时间间隔内报文各种考察指标的分布状况。WATCH1.0 所具有的采集和存储数据的功能可以基本满足目前基于

高速网络行为分析的需求。

WATCH1.0 的总体结构如图 1 所示，采集器接收从高速网络通过分光产生的流量镜像并截获其中的报文头部信息（每个报文头长 44 字节，以保证能截获完整的 IP 层和 TCP 层头部信息），并通过直连方式将其传输给处理器；处理器对报文头部信息进行相应的处理将其传送到存储器和数据库，并进行进一步处理和加以存储。

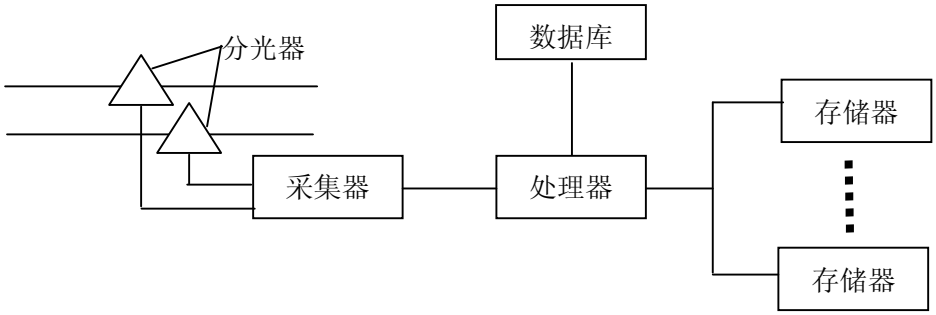

图 1　WATCH 1.0 总体结构图

在 WATCH1.0 中，不同的需求是通过作业表的形式提供给处理器的。当处理器接收到作业表后，将根据作业的时间要求依次启动存储器、处理器和采集器的相应模块，整个系统就开始按照作业的要求对报文进行分析、处理和存储，直到作业结束或存储资源耗尽。

2.2 WATCH1.0 处理器内部结构

处理器是 WATCH1.0 的核心模块，其主要功能是根据需求的不同，完成数据报文的组流、简单的报文测度计算、指定 IP 的原始数据报文存储和对原始报文按不同的抽样方式进行抽样并存储等工作。主要由以下模块组成：控制模块，报文接收模块，报文测度信息提取和组流模块，发送模块。其中除报文测度信息提取和组流模块，其他各个模块均需要和采集器或/和存储器交互，其结构示意图如图 2 所示：

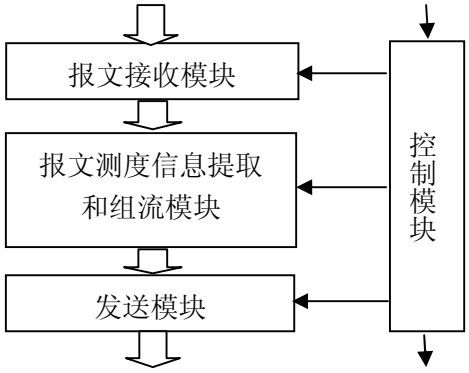

图 2 处理器结构示意图

（1）控制模块：主要功能是接收作业表信息，并根据作业表监视和控制整个处理器的运行。其中主要涉及与其他部分交互的信息：

- Ø 数据库：获得作业表信息；作业终止信息。
- Ø 采集器：采集开始信息；采集终止信息（作业终止/盘满终止）。
- Ø 存储器：作业表信息；作业终止信息。

（2）报文接收模块：其主要功能是接收从采集器发送过来的报文，将报文存储在一个暂存循环队列中，维护一个写指针，每次写入时判断写指针是否超越包测度信息提取和组流模块维护读指针，从而导致数据覆盖。

（3）报文测度信息提取和组流模块：该模块主要有两部分组成：报文测度信息提取并暂存

指定的原始报文以及对所有报文进行组流，是整个处理器的核心。

① 报文测度信息提取模块主要功能有三项：

- Ø 对每一个报文计算包测度信息；
- Ø 根据不同的抽样方式，对原始报文抽样，并暂存至一定量后发送到存储器；
- Ø 根据指定的源宿 IP 对要求，暂存和发送原始报文。

② 组流模块定义了两个 Hash 表空间。其主要功能是利用五元组定义流，通过 Hash 表和链表相结合存储流信息，在一个Hash表所存储的流到达一定阀值时，完成Hash表的切换，并由发送模块清空完成的 Hash。

（4）发送模块：发送模块在逻辑上是一个整体，用于将有关报文测度信息，原始报文和流信息发送到数据库和对应的存储器中。由于在具体实现时处理器也作为一个存储器使用，所以在物理上该模块由两个子模块构成：本地报文存储子模块和远程发送子模块。本地报文存储子模块是将处理器作为存储器使用。

3 处理器性能分析和优化

处理器是网络流测量平台的核心，其性能直接影响着系统的整体性能，是保证测量结果准确性的最关键部分。通过实验发现处理器的性能瓶颈主要集中在以下几个部分：（1）组流过程中可能导致的计算能力不足；（2）流信息维护的内存空间限制；（3）原始报文接收与测度信息、抽样报文信息和流信息发送的 I/O 冲突。而其中第 3 点在处理器实现初期表现得尤为明显。

在初步设计时，组流模块是使用对 IP 报文源宿 IP 地址进行高效 Hash 运算来定位和匹配对应的流，并使用链表解决流信息冲突。由于 Hash 函数效果和 IP 地址分布的不均匀性及不确定性等原因，导致了相当部分链表长度过长，从而导致处理器在匹配流时耗费大量 CPU 时间，系统的整体性能下降，以至于在流量突发情况下不能正常工作。对此，我们提出了两种解决方案：采用二叉树替代链表，改进 Hash 函数使其所得值分布更均匀。采用二叉树替代链表可以将匹配的时间复杂度从 O(n)降低到 O(log(n))，在最坏情况下也能保证时间复杂度为 O(n)；对于 Hash 函数值分布均匀性的改进，我们选取了折叠偏移算法[9]，从理论的角度证明其对 IP 对 hash 的均匀性，并通过多次试验对结论进行了验证。

维护流信息所需要的内存空间和网络当前的运行状况有很大的关系，当网络中存在大量长而慢的流时，流信息所占用的空间将大大超出正常值，这不仅影响处理器的内存空间，而且对其处理能力也是很大的考验，所以必须保证处理器可以在这种极端状况下工作正常。我们提出的解决方案是将组流任务分解，让存储器承担其中一部分。具体的方案如下：

（1） 在处理器中维护两个相同大小的 Hash 表 A 和 B，当工作的 Hash 表 A 中维护的流数等于某个阀值时，完成表之间切换：Hash 表 B 进入工作状态，表 A 中的流被发送到指定的存储器；如此循环往复；

（2） 当前工作的存储器根据 Hash 表号和对应的流 ID 号完成进一步的组流工作，并将结束的流存储在本地指定文件中。

这样就分两步完成了整个组流，将组流任务所需的计算量分解为处理器和存储器共同完成，从而有效地减轻了处理器的压力。

由于处理器既需要接收来自采集器的原始数据报文又需要向存储器和数据库传输不同的流信息，报文信息和报文测度信息，所以解决可能产生的 I/O 冲突显得十分重要。其重点是保证接收不丢报文的情况下将待发送信息尽快发送给后端设备。WATCH1.0 采用不同的线程接收和传输数据，首先保证接收进程有足够的 I/O 接收数据不丢包；其次将待发的数据封装成足够大的包使用 Socket 直接和后端设备通讯，这样就保证了处理器在 CERNET 江苏省网边界极限情况下（双向流量 1.4Gbps，200Kpps）可以正常工作。

4 未来研究方向

在江苏省网边界的实际测试表明，WATCH1.0 可以在目前流量极限下完成设计目标，但是由于网络的发展，流量还在不断增加，要保证系统在更高流量情况下还能正常工作，必须提高处理器的处理能力。主要体现在以下几个方向：

（1） 提高组流过程的性能：主要寻找具有更大空间且能保证 IP 对的 hash 值分布均匀的 Hash 函数和更有效的冲突解决方案；

（2） 对报文测度信息提取的扩展，以满足对 IP 报文头部信息更完整提取的需求。

（3） 更有效的 I/O 冲突处理机制。

参考文献

[1] CoralReef Status. http://www.caida.org/tools/measurement/coralreef/status.xml. 2004.9.

[2] Time Carstens. Programming with pcap. http://www.tcpdump.org/pcap.htm 2004.9

[3] PMA Site index: configuration and status. http://pma.nlanr.net/.2004.9

[4] WATCH1.0 设计文档。东南大学计算机系江苏省网络技术重点实验室，2004 年 7 月。

[5] K.C.Claffy. Internet traffic characterization. Dissertation for the degree Doctor of Philosophy. University of California, San Diego.1994

[6]A.Kumar, Mh.Sung, et al. Data Algorithms for Efficient and Accurate Estimation of Flow Size Distribution, ACIM SIGMETRICs, June 13, 2004.

[7] N.Duffiels,C.Lund, and M. Thorup. Estimation flow distribution from sampled flow statistics. In Proc. ACM SIGCOMM, Aug.2003.

[8]N.Hohn and D.Veitch. Inverting sampled traffic. In Proc. ACM SIGCOMM Internet Measurement Conference, Oct. 2003.

[9] 程光，丁伟，龚俭。面向 IP 流的哈希算法研究。 新一代互联网体系结构理论研究第三次研讨会，湖南，长沙。2004 年 8 月

[10] R.Jain and S.A.Routhier. Packet trains- measurements and a new model for computer network traffic. *IEEE JSAC*, 4:986-995, 1986.