高速网络单点多信道测量与分析平台设计*

高亚东[1]，丁伟[1]，杨艳[1]

[1)] 东南大学计算机科学与工程学院，南京 210096

摘　要　设计和开发基于高速网络的测量和分析平台，将为网络行为学研究提供基础性的保障设施。本文给出了测量和分析平台的体系结构设计，具体探讨了平台几个宏观关键问题的解决方案，并指出了未来工作的重点。实验情况表明，平台中已经完成的系统能够适应基于高速网的测量要求。

关键词　高速网络，被动测量，结构

Preparation of Papers for the 8th Cross-Strait Information Technology Conference Proceedings

Author1[1], Author2[1], Author3[2]

[1)] School of Computer Science & Engineering, Southeast University, Nanjing 210096

(E-mail: xxxx1@xxx. edu.cn; xxxx2@xxx. edu.cn)

[2)] Department of Computer Science and Information Engineering, National Central University, Taoyuan 32001, Taiwan

(Email: yyy@yyy.edu.tw)

***Abstract*—These instructions give you basic guidelines for preparing papers for conference proceedings.**

***Keywords*—instruction, paper, conference**

1．引言

随着互联网迅速普及，网络带宽、用户数量、应用的种类和规模都在迅速增长，网络结构也日益复杂。在网络快速扩张的同时，各种各样的问题不断出现，影响了互联网的正常运行。这些问题中有的是因为网络基础设施滞后，而有的则是因为技术存在缺陷，可以通过技术改进加以解决。为了了解网络的运行状况，发现问题的原因所在，从而更好地为用户提供服务，有必要对网络行为进行系统性的测量和研究。网络测量分为主动测量和被动测量，其中被动测量技术通过在网络上设置测量点，获取网络流量镜像，其本身对网络运行没有任何干扰，不会产生额外流量，具有特殊的优势。但对于传输速率在 1Gbps 以上的高速主干信道，海量数据的捕获、存储，以及分析处理都是要解决的难题。

目前学术界开展的被动测量项目中比较著名的有美国应用网络研究国家实验室（NLANR）的 PMA（Passive Measurement and Analysis）和加州大学 Berkeley 分校与 IBM 共同开发的 SPAND（Shared Passive Network Performance Discovery）。CERNET 华东（北）地区网络中心承担的国家重点基础研究发展规划（973）课题“网络动态行为和传输控制理论”（2003CB314804）也是一个主要使用被动测量方法进行网络测量的重大课题。在该课题的支持下，已经实现了一个基于大规模高速网络的数据在线采集存储系统 Watcher1.1，采集点设置于江苏省教育和科研计算机网（JSERNET）的边界。JSCERNET 比较特殊，它具有唯一的边界，其边界路由器通过高速信道在 CERNET 华东（北）地区网络中心内实现和 CERNET 主干路由器连接，但该逻辑信道由多对物理光纤信道构成（目前是 3 对 1Gbps 光纤信道）。所以 Watcher1.1 是一个单点多信道采集存储系统，它采集的数据也分布在多台存储器上，需经过归并整理后才能成为研究数据。我们以 Watcher1.1 系统为基础，集成数据归并系统、数据分析系统、测量管理系统等其它系统，建立一个测量和分析平台，为 973 课题的

*

研究工作提供基础性的保障设施。本文将给出平台的体系结构设计，并就其中需要解决的关键问题进行阐述。

2. 平台体系结构

以数据流向作为设计出发点，构建网络测量和分析平台的体系结构，结构简图如图 1 所示。采集存储系统、数据归并系统和数据分析系统构成了平台的主骨架，将完成数据采集、归并整理和分析的全过程，而测量管理系统则是平台的辅助设施，是平台后期开发的重点。其中采集存储系统要在高速光纤信道实时采集数据，对设备性能有一定要求；采集的数据量非常庞大，数据归并系统需要采用特殊的策略来缩短数据归并整理的周期；数据分析系统完成各种基于 Trace 的分析工作，是平台的核心部分；测量管理系统设置管理信息库，使各项测量和分析工作处于该系统的监管之下，并方便研究成果的管理和检索。整个平台中，数据的采集存储、数据的归并策略、数据分析系统的设计，和测量管理系统需要支持的功能都是需要着力解决的关键问题。尤其是数据分析系统，它作为平台的核心部分，其设计是需解决问题中的重点。

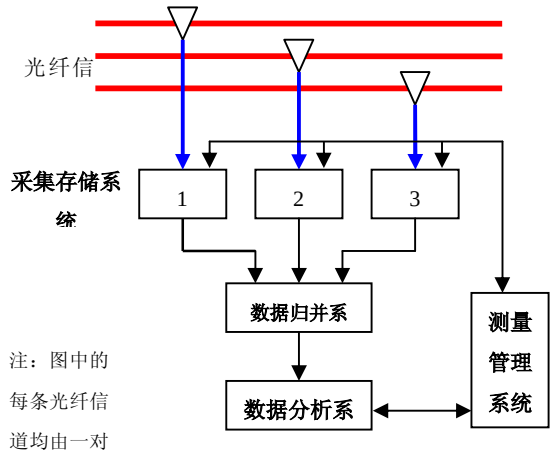

图 1 测量和分析平台构

3. 关键问题阐述

（1） 采集存储分级处理

根据实验，采集器将数据从千兆以太网卡发送到局域网比直接写入本机硬盘快得多，而且从高速光纤实时采集数据对机器的性能要求较高，因此基于性能的考量，将采集器和存储器分成两级，中间通过千兆铜缆线或者光纤直连。采集器以分光方式从光纤获取流量镜像，在采集器中根据采集前设定的配置信息，从报文头部截取一段，并打上此时的采集器系统时间作为报文的时间戳。采集器将若干个“时间戳＋报文头”拼成一个包，前面加上报文头的长度、数量等格式信息，并封装成以太帧，通过千兆以太网卡发送出去。存储器利用 pcap 程序库实现以太帧捕获，将以太帧负载的数据解析后获得报文头，写入文件。鉴于每台采集器只负责采集一对光纤，为了使时间戳误差尽可能减小，从而减小测量误差，采集器需定时而且较频繁地进行时钟同步。时钟同步的方式可以是采集器间互相同步，也可以是采集器和一台专用时钟服务器进行同步。

（2） 数据归并整理策略

如果将两台机器之间的一次通信定义为一个数据流，那么 JSERNET 边界路由器的负载均衡策略可能会使属于同一个流的不同报文选择不同的光纤信道进入 CERNET 主干。所以，对于采集的数据，同一个流的报文可能分布在多台存储器上。因此，我们在使用这些数据前，需要根据报文时间戳对报文进行归并整理，将分散的数据整理成一个规范的时间戳有序的报文流。因此，编写一个报文归并器，用于进行时间戳比较，将多份数据归并合一。为了提高数据整理的效率，我们利用东南大学计算机学院的高性能计算机集群进行报文归并。将各个存储器上数据按时间段划分切片，并按时间段进行归类。对于每一个时间段，数据归并系统在高性能计算机集群上开启一个归并器进程，将来源于不同存储器的切片归并到一起。若干进程并行运行，缩短了数据归并整理的时间。

（3） 数据分析系统设计

数据分析系统直接为用户提供服务，完成各种基于 Trace 的分析工作，是平台的核心部分，也是论文这一部分讨论的重点。由于数据分析需求各异，而且不断变化和增加，在系统中无法满足所有需求，所以该系统的定位是为用户提供一般性和基础性的数据分析服务。对于其它更高级的个性化数据分析工作，该系统起支持作用，为它们提供二次分析的源数据。

设计合理的体系结构是该系统的关键所在。系统在开发之时只能考虑到部分需求，所以系统结构应该具有良好的可扩展性，方便后期对系统功能进行扩展。我们考虑两种功能模块扩展方式，其一是编写新模块的源代码，在原系统的程序中增加相应语句，形成调用关系，并重新进行编译连接，生成新的系统执行代码。这种方式的好处是系统的整体性较好，缺点是耦

合太紧密，必须要熟悉系统的人员甚至是系统开发人员才能进行功能扩展，适用于系统开发阶段，而且这种方式无助于体现系统具有可扩展性。第二种扩展方式是新增的功能模块自身就是一个可独立运行的程序，扩展时将模块的可执行文件放入某一指定位置，在系统注册文件中添加该模块的信息，系统运行时根据用户的配置通过系统调用自动加载相应功能模块。这种方式下系统承担了管理的功能，新模块在系统中登记注册，即可实现系统功能扩展。其优点是系统开发完成后，不需要修改系统代码就可以进行功能扩展，适用于系统正式运作之后。这也是本系统体现可扩展性的方式。

系统的性能也是设计系统体系结构时需要着重考虑的因素。增加系统不同模块的并行度，可以很大程度提升系统的性能，而且有利于同时满足用户的多种数据处理需求。这主要通过多线程的方式实现。对于用户

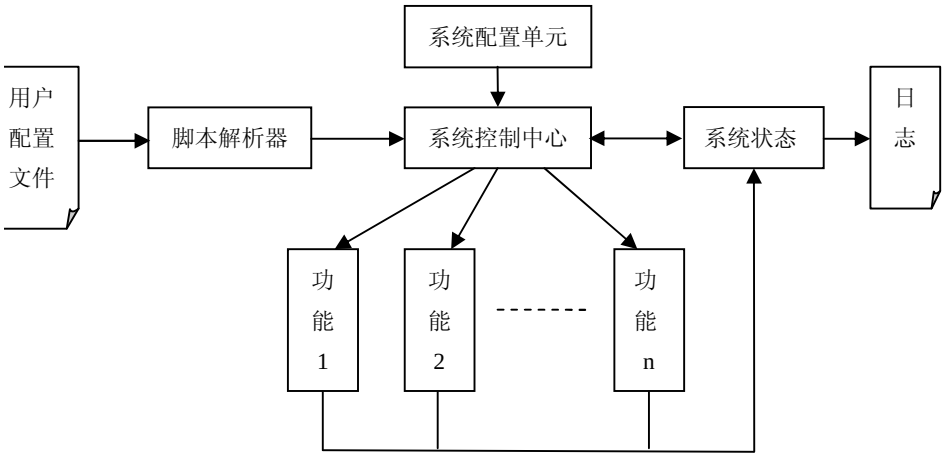

图 2（a） 控制框图

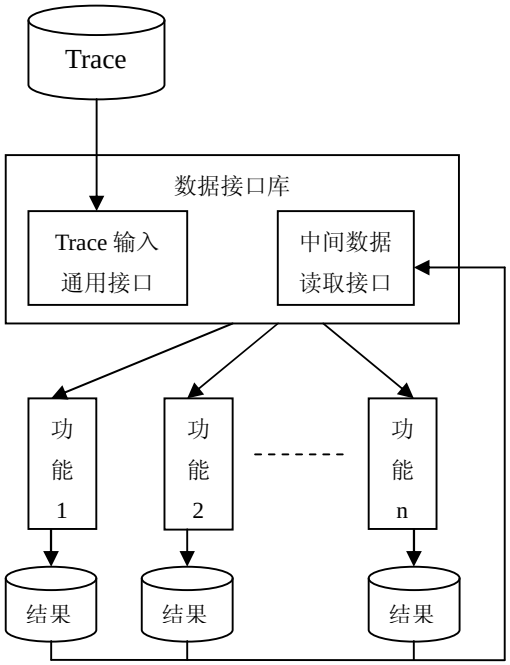

图 2（b） 数据框图

的配置信息，和相关的数据格式，可利用某种语言进行描述，并在系统中设计脚本解析器进行解析。综合上述几点考虑，从控制流和数据流两个角度设计系统的框架，图 2 是相关框图

在系统中，设计和实现数据接口库，用于 Trace 和数据结果的读取。在开发时，主要设计四种基本类型的算法，分别是流处理、报文抽样、报文测度计算和流测度计算。报文抽样主要有随机抽样、系统抽样和分层抽样三种；报文测度计算和流测度计算，针对某些指标，进行统计，获取指标值；流处理的算法比较复杂，而且因流的定义而异。根据 Ryu B 等人在参考文献[1]中给出的定义，数据流是符合特定的流规范（specification）和超时（timeout）约束的一系列数据包的集合，不同的流规范和超时约束导致不同的流定义。目前 973 课题所涉及的流主要采用（源 IP、宿 IP、源端口、宿端口、协议类型）五元组和超时进行定义，而其它方式定义的流与其相比有很多相似之处。对于采用五元组和超时进行定义的流，其组流算法，我们已在参考文献[6]中给出模型。其它方式的流定义，一般都比采用五元组和超时约束的流定义简单，所以它们的组流模型可以通过简化我们已有的模型建立。

（4） 测量管理系统功能

测量管理系统是测量平台中的辅助设施，它独立于平台主骨架之外，其目标是实现对原始数据、研究内容、实验代码、实验数据、实验结果，以及各种实验文档的有效管理，提供便捷的查询方式，并对用户行为进行有效监督。其中重点是对实验代码、实验数

据、实验结果和实验文档等的管理，因为它们是研究人员的科研成果，高效的管理有助于减少重复劳动，提高科研效率。该系统是平台构建的后期工作重点，目前尚未进行具体设计，这里仅讨论它的功能需求。它所支持的功能主要体现在四个方面：a.具有可靠的安全管理和用户管理策略，并能够对用户行为进行监督和审计；b.系统以原始的报文数据 Trace 为管理核心，对于和 Trace 相关的研究工作，系统能够对其管理；c.系统具有友好的人机接口，方便用户发布与检索研究文档和数据；d.系统能根据用户需求，生成适合用户提交数据的表单，以及直观的示意图。

鉴于平台比较复杂庞大，上面更多地从宏观上讨论了这四个问题。几个系统在具体实现时必然存在很多技术和细节上的困难，这有待我们进一步解决。

4．总结和展望

网络测量和分析平台中采集存储系统和数据归并整理系统已经投入使用。实际运行情况表明，采集存储系统能够满足高速信道的采集要求，丢包率控制在很低的数值；数据归并整理系统在高性能计算机集群上能够以较高效率进行数据归并整理，性能优良。目前，数据分析系统正在开发之中，而测量管理系统则在构思设计之中。未来，我们将继续优化已经完成的系统，对于未完成的系统，则按计划进行设计开发。其中重点是做好有关系统的衔接，例如数据分析系统和测量管理系统如果能做到无缝衔接，将有助于提高平台的整体性，而且给研究工作带来极大的方便。此外，数据分析系统的功能也将不断扩展，有关功能模块的算法也可以继续改进。

参考文献

[1] Ryu B, Cheney D, Braun H.W. Internet Flow Characterization: Adaptive Timeout Strategy and Statistical Modeling[J]. In Workshop on Passive and Active Measurement(PAM), Apr, 2001.

[2] Jürgen Quittek, Tanja Zseby, Georg Carle, Sebastian Zander, Traffic flow measurements within IP networks: Requirements, Technologies, and Standardization, Applications and the Internet (SAINT) Workshops, 2002. Proceedings.2002.

[3] Jun Li, Minho Sung, Jun Xu, Li Li, Large-scale IP traceback in high-speed Internet: practical techniques and theoretical foundation, Security and Privacy, 2004. Proceedings.2004 IEEE Symposium on 9-12 May 2004 Page(s):115-129.

[4] Baek-Young Choi, Jaesung Park, Zhi-Li Zhang, Adaptive random sampling for traffic load measurement, Communications, 2003. ICC '03. IEEE International Conference on Volume 3, 11-15 May 2003 Page(s):1552-1556 vol.3.

[5] 张宏莉，方滨兴，胡铭曾等，Internet 测量和分析综述，2003 Journal of Software, Vol.14, No.1, P110-116.

[6] 高亚东，周明中，丁伟，高速网络中的数据流信息提取模型，计算机时代，Vol.22,No.12,p31-33，2004 年 12 月。

[7] WATCH1.1 设计文档，东南大学计算机学院江苏省网络技术重点实验室，2005 年 10 月。