基于 GPU 的分布式数据包采集与回溯分析系统设计研究

Design and Research of Distributed Packet Collection and Backtracking Analysis System Based on GPU

杨宏强

Hongqiang Yang

深圳市宏大联合实业有限公司　中国·广东 深圳 518000

Shenzhen Hongda United Industrial Co., Ltd., Shenzhen, Guangdong, 518000, China

摘　要：随着计算机网络技术的快速发展，越来越多的企业将其关键业务转移到网络上，使得数据采集成为一个非常重要的问题。作为从大数据中分析和提取网络的基础，数据包分析能够准确预测网络态势，并分析系统功能需求，根据需求来设计数据包采集系统，对于系统的采集效果、存储效果等进行分析与验证，确保系统的稳定性和可扩展性。

Abstract: With the rapid development of computer network technology, more and more enterprises are transferring their key businesses to the network, making data collection a very important issue. As the foundation for analyzing and extracting networks from big data, packet analysis can accurately predict network situations, analyze system functional requirements, design a packet collection system based on requirements, analyze and verify the collection and storage effects of the system, and ensure the stability and scalability of the system.

关键词：分布式数据包；数据采集；GPU；回溯分析系统

Keywords: distributed data packets; data collection; GPU; retrospective analysis system

DOI: 10.12346/csai.v2i1.9262

1 引言

随着不同业务、服务和 Web 应用程序的数量不断增长，网络流量也在不断增长。由于网络数据流量的大小和数量，网络管理员和维护人员在监控和管理网络方面面临巨大困难 [1]。为了解决网络流量管理中的许多问题，本系统在数据包采集和分析的基础上开发了网络数据包采集与分析系统，该系统功能包括数据包采集与分析、数据包存储管理、数据包回溯分析、基础数据管理、外部系统服务等，将众多功能集于一体，有效提升了分布式数据包采集系统的性能。

2 系统技术框架

图形处理器就是 GPU，是现代计算机中的一个重要设备，为科学计算提供了引擎。GPU 通常用于提供高密度和并行计算能力，并提高嵌入式系统和设备的计算性能。相较于传统 CPU 来说，在计算机图形和图像处理中具有更多的优势 [2]。GPU 可以总结为 N 个处理器同时工作，彼此之间内存共享。GPU 的组成包括多个单指令多数据处理器（SIMD，Single Instruction Multiple Data）和数千个并行线程。相较于 CPU 线程来说，GPU 的上下文切换、创建时间比较短等优势。

3 系统需求分析

随着通信技术与互联网的快速发展，连接到网络的系统数量和网络提供的服务类型不断增加，导致网络容量和数据流量爆炸性增长。另外，通过主动和被动网络性能的各种实际测量以及随后对网络流量的监测，通信介质通过接收大量数据来对于通信源的状态进行监测，并且其中含有大量的有价值信息 [3]。面对大量异构的网络资源，如连接故障信息、安全风险信息和攻击行为信息，如何恢复和分析网络上的大量数据流量源，以及如何有效地分析和管理这些资源，如何充分利用现有的网络数据源对于网络的状态进行评估成为网络研究人员的关注重点。

【作者简介】杨宏强（1987-），男，蒙古族，中国广东深圳人，本科，工程师，从事智能通信工程研究。

3.1 高速实时的数据包采集功能

分布式数据包的采集系统、分析系统的设计中，关键在于高速网络环境下手机数据包，然后对其进行深入分析。所以，分布式数据包的采集系统、分析系统必须实现高速、实时的数据包采集功能。论文基于 LIBCAP 的零拷贝技术，使得从网络收集数据包后的操作过程得到有效缩短，减少了数据包在内存中的复制和处理操作过程，从而提高了分布式数据包收集和反向系统分析的能力。

3.2 高效的数据包解析功能

系统的数据包分析和拦截功能非常中，可以对数据包内的有效信息进行分析，并对于网络流量实现实时监控；同时，系统还能够添加自己的协议，并调用自己的协议进行数据包分析。

3.3 数据包回溯分析功能

该系统旨在将特定时间的网络流量数据回访，并显示各个网段、IP、应用程序等的详细信息。同时，该系统还具有显示数据和信息的功能，以多维方式向用户显示类似的结果。

3.4 高效的数据存储功能

该系统的目标应是有效地存储数据包，根据需要实现分组，创建数据存储索引，消除重复和数据噪声，并根据规则对数据进行分类。同时拥有分布式存储和海量数据的恢复等功能。

3.5 灵活的系统扩展与配置功能

随着网络的发展，系统需要具备灵活的可扩展性才能满足不断发展的功能需求。系统需要增加一个可扩展的接口，能够灵活地响应外部应用程序的插入、对其他系统或设备的访问等 。同时，数据包采集系统必须坚持易于安装和实现的原则，以创造友好的用户体验。从系统管理员的角度来看，他们必须管理用户和用户组的权限；设置不同用户的基础数据处理权限；限制外部用户访问系统。

3.6 非功能性需求分析

分布式数据包采集与回溯分析系统需要拥有几方面特征：①高性能：系统单设备采集不小于 100Mps，存储能力总体不小于 20TB，单个采集设备的存储能力不小于 1TB。所以，系统需要为用户提供高精准服务，排除不良因素的影响，降低不可预知的故障发生率，确保系统的高性能。②高可靠性：系统的可靠性是稳定运行的关键，也是检验系统是否完善的参考依据。所以，分布式数据包采集和回溯分析系统需要拥有一套完整可行的设计方案，实现系统的高可靠性。③高可扩展性：系统设计需要根据一定规模的网络环境设计方案，但是随着网络发展，系统的性能和存储能力总会出现不足，所以系统需要设计一套完整可行的方案进行系统扩展。并预留相应接口，方便插件、其他系统的接入。

4 数据包采集与回溯分析系统设计

4.1 系统总体功能框架设计

分布式数据包分析采集系统的组成如图 1 所示，包括网络信息包分析采集模块、基础数据管理模块、数据包存储模块、网络数据包采集模块、GPU 分析处理模块、系统视图模块、外部服务模块等。

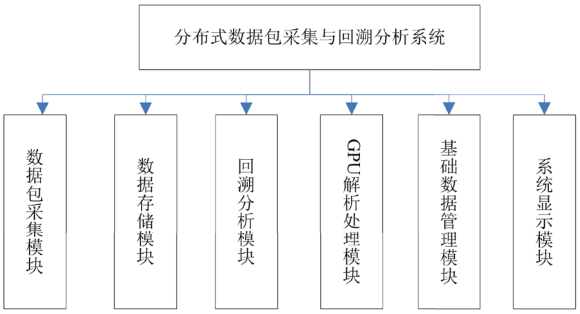

图 1 数据包采集与回溯分析系统功能组成图

4.2 数据包采集模块

系统中的采集分析模块是非常重要的模块之一，主要负责数据包收集和分析，从而实现网络数据包的收集功能。该模块可以实时采集网络数据包，并通过 PCAP 格式读取 HDFS 文件系统中分布的历史数据包。模块基于 LipcapLibpcap（Packet Capture Library）数据包捕获函数库，实现网络数据包的收集工作。如果数据包收集器收集的数据包不需要临时分析，则数据包可以保存为 HDFS 文件系统中的 PCAP 文件。

4.3 GPU 解析处理模块

GPU 并行处理模块主要负责的是数据报协议分析与数据包字段分析工作。该模块主要是利用 GPU、MapReduce 等并行技术，可以加快数据包在分析过程中分析速度以及处理速度，有效提高系统性能。在数据包协议分析过程中，系统预先设计 TCP 和 UDP 协议分析方法 。同时，为了能够提高系统的可扩展性，分布式数据包收集和跟踪系统为用户提供了 LUA 脚本接口，允许用户扩展协议解析部分，提高系统的可伸缩性。

4.4 数据存储模块

该模块通过关系数据库（Postgrec1）与 HDFS 文件系统结合，实现数据包、数据包字段、数据分析结果以及数据包中的 PCAP 文件的管理功能。HDFS 用于存储大容量文件，关系数据库可以储存数据包中的解释字段、各种处理结果与分析结果等。该系统将分析模块分析的包中的每个字段存储在关系数据库中，使用分钟、小时和天数作为时间尺度，轻松访问和分析数据包中的字段。为了根据需要加快数据包的存储和操作，该系统使用 PG-STROM 技术和并行 GPU 技术来实现灵活的海量数据存储和高效恢复。系统收集的原始数据包将会采用 PCAP 文件的形式存储在 HDFS 文件系统中，

并且用于后续的系统分析和快速访问其他数据分析平台。

4.5 回溯分析模块

一方面，可追溯模块根据时间、周期、IP 地址和源端口、应用协议、TCP 会话、UDP 会话、IP 地址与目的端口等对网络数据包进行分析，提取一段时间内的网络流量信息。另一方面，可以根据需要从原始数据文件中读取包，并通过内部接口为其他模块提供服务，利用回溯分析模块和系统显示模块向用户呈现出结果。

5 系统部署与测试

5.1 系统性能测试

为了测试系统性能是否达到系统需求，因此对于系统性能开展了功能测试。通过使用大小不同的数据，测试了 Postgrecl 数据库、单 CPU 计算、GPU 计算（一个加装 NVIDIA K20C GPU 卡节点）的节点的计算性能。测试数据由源 IP 地址、目的端口号、源端口号、目标 IP 地址、协议和数据包大小等一组数据包属性构成。三组测试数据分别为 2000 万、13000 万和 50000 万。计算任务是测试 IP 地址搜索的有效性，结果见表 1。

表 1 分类统计排序性能对比表

记录数（万）	2000	13000	50000
PostgreSQL 计算时（s）	65.20	243.35	743.62
CPU 计算时长（s）	142.62	287.28	875.25
GPU 计算时长（s）	7.46	15.92	21.60

从上述结果看出，添加 GPU 计算节点有效提升了分类的性能；同时，处理时间与数据量有关，不同的数据记录会根据自身的复杂性而导致处理时间可能存在显著差异。

5.2 系统性能测试

经过系统测试发现，分布式数据包收集和追溯分析系统可以满足设计需求分析。从测试结果表 2 中可以看出，在对模块进行功能测试时，已经涵盖了需求分析过程的主要功能点。

表 2 系统性能测试结果表

编号	模块名称	测试要求	测试结果
01	登录模块	1. 输入用户名或密码错误时是否可以给出错误信息提示	测试成功
		2. 是否可以登录系统的主界面	
		3. 用户身份是否对应操作权限，如无操作权限，是否已隐藏没有权限操作的功能	
02	采集模块	1. 能否采集到接口的数据包	测试成功
		2. 采集到的数据包能否及时高效地传递到数据包分析模块	
03	实时分析模块	1. 采集到的数据包是否能够解析	测试成功
		2. 分析界面能否实时显示 UDP 实时流速、总流速、TCP 实时流速等信息	
04	回溯分析模块	1. 历史流量趋势是否可以解析，是否可以绘制流量趋势图	测试通过
		2. 能够显示各节点流量占比及其饼状图	
05	基础数据管理模块	1. 能否对自定义协议、应用系统、节点信息等进行修改、增加、删除等操作	测试成功
		2. 能否增加用户或修改用户	
06	数据存储模块	1. 能否存储经过解析后的数据包各字段	测试成功
		2. 存储模块能否及时响应其他模块的数据请求	
		3. 是否可以根据 PCAP 格式存储原始数据包	

6 结语

综上所述，分布式数据包的采集与回溯分析系统设计过程中，GPU 并行分析模块、回溯分析模块、数据包采集模块和数据存储模块是重要构成部分。同时，该系统使用 GPU 等相关技术进行并行分析和数据包存储，提高了采集和回溯分析系统的性能。根据系统的测试结果，可以看出该系统符合设计要求。但是还需要对于回溯分析进一步完善，通过利用数据挖掘算法，实现在数据中有效信息的挖掘。

参考文献

[1] 谭超,段俊明,辛亮,等.基于STM32与异步FIFO的乒乓模式高速数据采集系统设计[J].电工材料,2023(1):55-59+63.

[2] 张祖刚,贾琨.基于MQTT协议的数据中心动环系统数据采集器设计[J].智能建筑电气技术,2023,17(1):61-64.

[3] 马云龙,王黎明,张法业,等.用户侧多能源数据采集系统设计与应用[J].电子设计工程,2023,31(2):53-58.