A Data Filtering Method for Digital Resource Utilization Analysis Systems Supporting Dual-Stack and High-Speed Networks (Postprint)

Authors: Pan Zhuhong, Xiao Dehong

Date: 2017-10-11T00:00:00+00:00

Abstract

[Objective] To facilitate the rational construction, scientific management, and efficient utilization of digital resources in university libraries. [Application Background] The widespread adoption of IPv6 and 10-gigabit campus networks has posed challenges for network data acquisition. [Method] This paper proposes a network device port mirroring design approach that filters IPv4 and IPv6 network data prior to collection by the digital resource utilization analysis system. [Result] A digital resource utilization analysis system supporting IPv4/IPv6 dual-stack and 10-gigabit networks has been deployed in practice. [Conclusion] This enables the library digital resource utilization analysis system to adapt to dual-stack high-speed campus network environments.

Full Text

A Data Filtering Method for Digital Resource Usage Analysis Systems Supporting Dual-Stack and High-Speed Networks

Pan Zhuhong[1], Xiao Dehong[1,2]
[1](Information and Network Center, Xiamen University, Xiamen 361005, China)
[2](Library of Xiamen University, Xiamen 361005, China)

Corresponding Author: Pan Zhuhong, ORCID: 0000-0002-6837-5349, E-mail: zhpan@xmu.edu.cn

Abstract

[Objective] This study aims to promote the rational development, scientific management, and efficient utilization of digital resources in university libraries. **[Context]** The promotion of IPv6 and the popularization of 10 Gigabit campus

networks have created difficulties in network data acquisition. [**Methods**] We propose a port mirroring design method for network devices that filters IPv4 and IPv6 network data before collection by the digital resource usage analysis system. [**Results**] We deployed a practical digital resource usage analysis system supporting IPv4/IPv6 dual-stack and 10 Gigabit networks. [**Conclusions**] This enables library digital resource usage analysis systems to adapt to dual-stack, high-speed campus network environments.

Keywords: Digital resources; Information acquisition; IPv6; Port mirroring
Classification Number: TP393.1 G25

Introduction

University library digital resources encompass not only daily-used literature materials but also network academic resources such as shared information, software, and domestic and foreign databases within the library network platform. How to promote the rational construction, scientific management, and efficient utilization of university library digital resources is a key concern in current library work, with numerous studies [1] and practical explorations on library digital resource utilization evaluation systems already conducted.

Common investigation methods for digital resource utilization evaluation include user surveys [2] and access records provided by database vendors. Among these, data information transmitted via networks—such as digital resource login, retrieval, and download counts, as well as user group statistics—can be obtained from the campus network environment. It is understood that some university libraries have already deployed self-developed or commercial digital resource usage analysis systems. Limited by the high cost of dedicated hardware, common digital resource usage analysis systems typically implement packet header parsing in software on general-purpose processors. The network acquisition bottleneck for such systems is 1 Gbps [3], and they generally do not support IPv6. With the promotion of IPv6 in universities and the popularization of 10 Gigabit campus networks, digital resource usage analysis systems have become unable to handle campus network traffic.

This paper aims to design a data filtering method that supports IPv4/IPv6 dual-stack and 10 Gigabit network environments. This method completes data filtering before system data collection, outputting only library-related IPv4/IPv6 traffic to the data collection and analysis modules, thereby enabling digital resource usage analysis systems to function in high-speed, dual-stack campus networks.

2.1 Technical Background

(1) IPv6 Promotion

IPv6 is the next-generation IP protocol designed by IETF to replace the current IPv4 version, defined in RFC 1752 in 1994 [4]. The China Next Generation Inter-

net (CNGI) project, led by the National Development and Reform Commission, completed comprehensive IPv6 coverage for 100 universities in 2010 [5]. On June 6, 2012, IPv6 networks were officially launched globally [6]. IPv6 networks have been increasingly promoted and now carry some digital resource transmission. However, due to differences in protocol architecture, network management also varies significantly, and management analysis tools for IPv6 remain incomplete, making IPv6 one of the challenges for library digital resource usage analysis systems.

(2) Popularization of 10 Gigabit Campus Networks

After IEEE passed the 802.3ae 10 Gigabit Ethernet standard on July 18, 2002, 10 Gigabit Ethernet gradually became popular in university campus networks due to its bandwidth of up to 10 Gbps and various technical advantages. This also brought significant challenges to university library digital resource usage analysis systems.

(3) Flow Mirroring Technology

Network data acquisition can be divided into dedicated hardware-based and general processor platform-based approaches. Dedicated hardware methods offer significant performance advantages in high-speed link environments, but their high costs lead most systems to adopt software-based packet header parsing on general-purpose processors. Limited by operating systems and hardware performance, current processing speeds can only reach 1 Gbps (1,488 Kpps) [3], making 10 Gigabit traffic links a bottleneck for deploying digital resource usage analysis systems.

Filtering and diverting data at the network device output stage before data capture is an effective means to solve system network acquisition bottlenecks. Common data output technologies include port mirroring and optical splitters. Splitters are physical layer tap devices that do not support data filtering. Port mirroring is a management function provided by network devices that copies packets from specified ports or VLANs to other ports [7]. Flow mirroring (ACL-based mirroring) is a type of port mirroring that only copies packets matching flow classification conditions to a specified destination [8] and is currently the only mirroring technology with data filtering capabilities.

Port mirroring technology consumes switch hardware and software resources. Even top-tier switches domestically and internationally are limited to 4 or even fewer mirroring groups. Flow mirroring technology is only supported in the latest software and hardware versions of some network devices, with significant functional limitations and stability that requires verification, generally supporting only one output [8]. While flow mirroring can support the data filtering needs of digital resource usage analysis systems, most campus network core devices do not support this technology. The limited flow mirroring resources on devices that do support it are often already used for user behavior analysis, attack protection, security detection, and public opinion control.

In summary, flow mirroring technology has significant limitations, high deployment costs, and supports only a small number of outputs, making it difficult to deploy on campus network core devices to provide data filtering for the system.

2.2 Composite Mirroring Technology

Port mirroring technology can be divided into three types: local mirroring, remote mirroring, and flow mirroring. [Figure 1: see original paper] shows the traffic flow patterns for various mirroring technologies.

Local mirroring performs additional copying of data during hardware forwarding to the mirroring destination port, as shown in Model A in [Figure 1: see original paper]. Remote mirroring technology sends mirrored packets to a remote mirroring reflector port, which then broadcasts and copies the packets within the remote mirroring VLAN, with data flow shown in Model B. Flow mirroring only copies packets matching flow classification conditions to a specified destination, as shown in Model C.

This paper proposes a technical concept: Based on the technical principles and characteristics of flow mirroring and multi-output remote mirroring, if flow-mirrored data streams could be directly sent to the mirror reflector port in the switch forwarding plane, and then broadcast by the reflector port to all ports configured as remote VLAN ports (as shown in Model D in [Figure 1: see original paper]), this would deeply integrate flow mirroring and remote mirroring to form a composite mirroring technology. This composite mirroring technology could achieve precisely controlled multi-path output at the hardware level, resolving the current contradiction where mirroring filtering and multi-output are difficult to coexist on most network devices, and providing data filtering support for digital resource usage analysis systems.

2.3 Overall Technical Approach

Information collection and analysis systems are generally divided into three functional modules: data collection and filtering, data storage and management, and data analysis and presentation [9]. Traditional systems have the data collection module capture all data before performing filtering.

The digital resource usage analysis system based on the proposed data filtering method for IPv4/IPv6 dual-stack and 10 Gigabit network environments advances the data filtering function to before data collection. The overall design consists of three functional modules: data filtering, data collection and storage, and data analysis and presentation, as shown in the overall model in [Figure 2: see original paper].

The data filtering module obtains a copy of network data containing both IPv4 and IPv6 packets from the core network. Through composite mirroring technology, it filters and diverts the data, outputting only information related to library digital resources to a data collection module that supports dual-stack

traffic. This solves the problem of existing digital resource usage analysis systems being unable to handle 10 Gigabit dual-stack campus network traffic.

The data collection and storage module captures and stores information related to digital resources. The data analysis and presentation module provides data display, query, and reporting functions, which are basically the same as traditional systems.

According to the overall model, the system is divided into three modules. The data collection and storage module and the data analysis and presentation module already have numerous research results and mature products. The research focus of this paper is to utilize the proposed composite mirroring technology-based data filtering method to complete data filtering before data collection.

3.1 Specific Technical Issues

Implementation of the data filtering module requires solving the following technical issues:

(1) Re-mirroring of Mirrored Traffic

As a technical concept, composite mirroring technology cannot be directly applied to production networks with high stability requirements. This module must be deployed on network hardware independent of the production network. Therefore, it faces the question of whether this hardware can identify and forward data sources obtained from core network devices.

Network devices identify and forward traffic by matching destination MAC addresses of data frames. As a network device management function, packet mirroring replication occurs before the network device identifies and forwards the packet. This replication mechanism can be analogized to a Hub [5]. Even if network devices do not forward certain mirrored packets due to transmission flags such as VLAN tags, they will first send the packets to the mirroring destination port, ensuring complete mirroring output of all packets. Experiments have also proven this mechanism. Therefore, re-mirroring output data packets through independent network devices is feasible.

(2) Specific Implementation of Composite Mirroring Technology

Switch devices are closed functional hardware that do not provide interfaces for modifying underlying hardware forwarding. To verify the feasibility of composite mirroring technology in actual environments, we use flow mirroring' s flow definition ACL to filter network traffic, sending only qualified traffic to a specific port, and then perform remote mirroring on that port again to achieve controllable multi-path output. The difference between this experimental scheme and the composite mirroring technology design is that an additional physical interface forwards the traffic that would have been directly sent to the remote mirroring reflector port in the composite mirroring design.

Experiments proved that directly performing remote mirroring on the flow mirroring destination port did not produce the expected output. The reason is that the mirroring destination port is marked as a non-forwarding port, making it impossible to directly perform remote mirroring on the flow mirroring destination port. To trigger the mirroring function again, the flow mirroring destination port is connected to other ports on the same device through external physical lines, sending the filtered traffic back to the network output system, and then performing remote mirroring on this loopback port to output to multiple ports, achieving multi-output.

(3) Support for IPv4/IPv6 Dual-Stack Traffic

Mirroring filtering is achieved by copying only data streams matching flow classification conditions to the destination port. Conventional flow classification supports only IPv4 or IPv6 protocols, but not both simultaneously. Supporting dual-stack traffic filtering requires implementing flow classification that can match both IPv6 and IPv4 conditions.

This system has implemented simultaneous matching of IPv4/IPv6 flow classification. For hardware systems that cannot support simultaneous matching of both protocols, dual-stack traffic filtering can also be achieved by performing composite mirroring and local mirroring plus composite mirroring on the source data separately and outputting to the same destination port.

3.2 Deployment Scheme

Both physical layer splitter transmission and data link layer port mirroring output technologies can be used to output from the production network. Given that splitter transmission is inflexible to deploy in campus network environments with multi-link load balancing and adds additional network failure points, this system adopts the local mirroring approach.

[Figure 3: see original paper] shows the specific deployment model of the data filtering module, primarily implemented based on a mid-range data center three-layer 10 Gigabit switch that supports flow mirroring. As shown, campus network equipment sends a copy of mirrored data to the switch where the data filtering module resides. This switch filters IPv4/IPv6 traffic from the mirrored data and outputs it, then loops it back to the same switch, and performs remote mirroring on the loopback port to output to multiple ports for data collection hardware. The data filtering module can provide data support for multiple data collection systems.

The IPv4/IPv6 flow filtering rules of the data filtering module are shown in . The database resource group in the IPv6 flow rules currently has no members.

The data collection and analysis module is primarily implemented based on a mid-range network security device, mainly providing functions such as digital resource access analysis, excessive file download alerts, and data security monitoring.

4.1 Data Filtering Method Experiment Process

Four three-layer switches and two PC terminals were designed to implement the data filtering method proposed in this paper. Switches A/B/C simulate the production network. Switch D is a mid-range three-layer 10 Gigabit switch. Terminals A/B use Wireshark to capture network packets.

The experimental topology is shown in [Figure 4: see original paper]. Switch B sends data from port G0/1 to port G0/20 through local mirroring. Port G0/20 connects to Switch D' s T1/0/4 port. Flow mirroring is configured on port T1/0/4 to output to port T1/0/16. An optical jumper connects T1/0/16 back to T1/0/15, and remote mirroring is configured on T1/0/15 to output to terminals A/B for packet capture.

Flow control information configuration is as follows:

```
acl number 3000
rule 0 permit ip destination 10.0.5.0 0.0.0.255
rule 5 permit ip destination 10.0.8.0 0.0.0.255
acl ipv6 number 3100
rule 10 permit ipv6 destination 2001:DA8:E800:40::/64
rule 15 permit ipv6 destination 2001:DA8:E800:43::/64
```

Flow classification configuration is as follows:

```
traffic classifier mirror-2 operator or
if-match acl 3000
if-match acl ipv6 3100
```

Switch A sends 5 ping packets to 4 IPv4 addresses (including 10.0.5.2) and 4 IPv6 addresses (including 2001:DA8:E800:40::2) for verification. According to the control rules, terminals A/B should simultaneously capture packets sent to 10.0.5.2, 10.0.8.2, 2001:DA8:E800:40::2, and 2001:DA8:E800:43::2. Other packets should be filtered by flow rules.

The actual results are shown in . The results match expectations, proving that dual-stack composite mirroring technology is feasible and the data filtering method is accurate and effective.

4.2 System Operation Status

Xiamen University Library has completed experimental deployment of a digital resource usage analysis system supporting IPv4/IPv6 dual-stack and 10 Gigabit networks. [Figure 5: see original paper] and [Figure 6: see original paper] show network traffic before and after filtering by the data filtering module. Comparison reveals that the filtered data volume is approximately 1-2% of the pre-filtered volume, a traffic level that common network information collection systems can handle.

shows the web access log volume collected by the digital resource usage analysis system and two comparison systems on the same day. The comparison systems are two currently highest-configuration domestic security management devices. Comparison system 1 collects data from the campus network IPv4/IPv6 dual-stack export mirroring traffic (i.e., pre-filtered network data). Comparison system 2 is a campus network boundary security device that does not support IPv6, collecting data from the campus network IPv4 export traffic.

Under high-traffic data loads, comparison system 1 is inferior to our system in terms of logging functionality and accuracy, while comparison system 2 produces data results close to our system. This proves that the data filtering module in our system can stably filter irrelevant traffic and completely output data information of interest to the library, and the data collection and storage module is sufficient to handle the filtered traffic.

The main functions of the digital resource usage analysis system include digital resource access analysis, excessive file download alerts, and data security monitoring. [Figure 7: see original paper] shows a weekly report of site requests by common library databases; [Figure 8: see original paper] shows a weekly report of downloads by common digital resource file formats; [Figure 9: see original paper] shows the virus detection alert function in the security protection features.

4.3 System Defects

The method proposed in this paper is primarily based on professional network equipment and requires personnel with professional network knowledge for configuration and maintenance, resulting in poor usability. The implementation of composite mirroring technology uses a port loopback approach, which has certain deployment complexity and stability risks. The system currently only completes basic recording and statistical analysis functions for digital resource network transmission, providing relatively raw data. Future work should further integrate with digital resource utilization evaluation systems to provide more accurate and intuitive data support for digital resource assessment.

References

[1] Pan Bubu. Study on Utilization Assessment of Digital Resource [J]. Research on Library Science, 2012(13): 86-89.

[2] Zhang Jing. An Empirical Study of Library Users' Behavior in Utilizing Digital Resources[J]. Journal of Guangdong University of Technology: Social Sciences Edition, 2012, 12(4): 74-78.

[3] Fusco F, Deri L. High Speed Network Traffic Analysis with Commodity Multicore Systems[C]. In: Proceedings of the 10th Annual Conference on Internet Measurement. 2010.

[4] Bradner S. The Recommendation for the IP Next Generation Protocol [J/OL]. [2013-03-02]. https://www.rfc-editor.org/rfc/pdfrfc/rfc1752.txt.pdf.

[5] Ma Yan. Member of CNGI-CERENT2 Actively Deployed IPv6 [J]. World Telecommunications, 2012(6): 61-62.

[6] Next Generation Internet Protocol IPv6 Officially Launched [J]. Silicon Valley, 2012(12): 14.

[7] Configuring Traffic Mirroring [EB/OL]. [2015-03-16]. http://www.cisco.com/c/en/us/td/docs/routers/asr90 1/interfaces/configuration/guide/hc51xasr9kbook/hc51span.html.

[8] H3C. Network Management and Monitoring Configuration Guide [EB/OL].[2015-03-16]. http://download.h3c.com.cn/download.do?id=1034399.

[9] Wang Jilong, Wu Jianping. An Internet Performance Monitoring Model Design and Implementation [J]. Journal of Computer Research and Development, 2000, 37(4): 443-452.

Author Contribution Statement

Pan Zhuhong, Xiao Dehong: Proposed research ideas, designed research plan, drafted and revised the final version of the paper; Pan Zhuhong: Collected, cleaned and analyzed data, conducted experiments.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data [1-3] can be found in the journal' s online version http://www.infotech.ac.cn; Supporting data [4] is self-archived by the authors, E-mail: zhpan@xmu.edu.cn.

Note: Figure translations are in progress. See original paper for figures.