Enterprise 네트워크에서 Flow를 이용한 Host 분석 시스템*

*박진완, 박상훈, 김명섭
고려대학교 컴퓨터정보학과
e-mail:{*pakjw84, 2002270130, tmskim}@korea.ac.kr

Flow-based Host Analysis System on Enterprise Network

*Jin-Wan Park, Sang-Hoon Park, and Myung-Sup Kim
*Dept of Computer Information Science, Korea University

요 약

효율적인 네트워크 관리를 위해서는 해당 네트워크의 트래픽 분석 정보가 필요하다. 트래픽을 분석함에 있어 네트워크의 전체 트래픽에 대한 분석도 중요하지만, 총 호스트의 수, 호스트별 트래픽 현황, 호스트의 사용 중인 서비스 파악 등 호스트를 기반으로 한 트래픽 분석의 중요성이 날로 증가하고 있다. 본 논문에서는 호스트 트래픽 분석이 어떠한 정보를 제공해야 되는지 살펴보고, 해당 정보를 제공하는 Enterprise 네트워크 트래픽 분석 시스템을 설계하고 구현한 내용을 기술한다. 본 논문에서 설계 및 구현한 시스템은 기존의 Enterprise 네트워크에 적합한 Flow 기반의 실시간 트래픽 모니터링 시스템[1]을 확장시킨 형태로 호스트 트래픽 분석에 필요한 DB구성과 분석 정보 표현, 그리고 트래픽 추이 그래프 생성에 따른 유의 사항과 해결책을 제시한다. 본 논문에서 구축한 시스템은 학교 Campus 네트워크를 대상으로 구축되었다.

1. 서론

최근 네트워크 기술의 발전과 다양한 네트워크 어플리케이션의 등장으로 인해 네트워크 트래픽은 대용량화될 뿐만 아니라 복잡 다양해지고 있다. 이러한 상황 속에서 원활하고 안정적인 네트워크 서비스를 위해서는 네트워크 트래픽 분석 시스템이 요구된다.

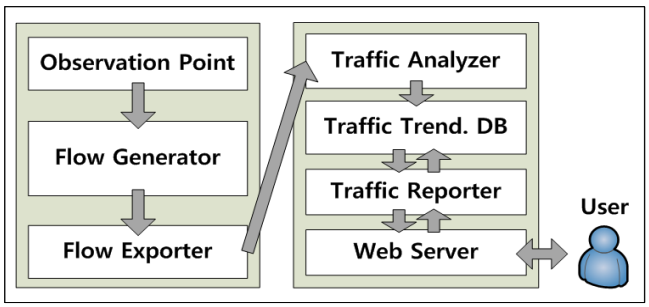
일반적으로 관리자는 트래픽 분석 시스템을 통해 특정 어플리케이션의 트래픽 정보, 비정상 트래픽의 발생 지점 등의 정보를 제공받기를 원한다. 하지만 현재 대부분의 네트워크 트래픽 분석 시스템들은 네트워크의 전체 트래픽 양과 같은 총체적인 트래픽 정보를 제공하는데 초점이 맞춰져 있어 이러한 정보들을 제공하지 못한다.

이러한 문제를 해결하기 위해서 본 논문에서는 특정 호스트의 Protocol/Port 별 트래픽 양과 같은 트래픽의 상세 정보를 제공하기 위해 네트워크의 각 호스트를 중점으로 한 트래픽 정보를 제공하는 시스템을 제시한다. 먼저 호스트 트래픽 분석 시스템이 일반적으로 요구되는 정보들을 파악하고 이 정보들을 제공하는 시스템을 설계 및 구현한 내용을 기술한다. 본 시스템은 기존의 Enterprise 네트워크에 적합한 Flow 기반 트래픽 모니터링 시스템(이하 KU-MON)[1]을 기반으로 시스템을 설계 및 구현하였다.

본 논문은 다음과 같은 순서로 기술한다. 2장은 관련연구로써 KU-MON 시스템에 대한 내용을 기술하고, 3장에서는 호스트 트래픽 분석 시스템의 요구사항을, 4장에서는 시스템을 설계한 내용을 기술한다. 5장에서는 설계를 바탕으로 시스템을 구현한 내용을 기술하고, 마지막으로 6장에서는 결론을 맺고 향후 보완점과 연구과제에 대하여 언급한다.

2. 관련 연구

최근 대용량 트래픽의 실시간 처리를 위해 네트워크 트래픽 분석 시스템은 Flow 기반 분석[2, 3]과 Cluster 형태의 분석 시스템 구조[4, 5, 6]를 이룬다. Flow 기반의 실시간 트래픽 분석 시스템인 KU-MON은 Enterprise 네트워크인 Campus 네트워크를 대상으로 구축되어 있으며, IETF IPFIX WG에서 제안한 Flow 구조[7]를 사용하여 구현되었다.



(그림 1) KU-MON의 전체 구조

* 이 논문은 2007년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2007-331-D00387)

(그림 1)은 KU-MON의 전체적인 구조이다. KU-MON 시스템은 다음과 같이 네트워크 트래픽 모니터링이 이루어진다. Flow Generator가 Observation Point에서 패킷을 캡쳐하여 1분 동안의 패킷 헤더 정보를 모은 후, Flow 정보로 변환 및 디스크에 파일로 저장한다. 저장된 Flow 파일은 파일 전송 모듈인 Flow Exporter에 의해 Traffic Analyzer로 전송된다. Traffic Analyzer는 Flow 정보를 분석 후 DB에 저장하는 역할을 하며, 마지막으로 Traffic Reporter가 DB에 저장된 분석 정보를 웹을 통해 사용자에게 제공한다. 사용자에게 제공하는 정보로는 Campus 네트워크의 분, 시, 일, 주, 월 단위의 전체 트래픽 양과 IP/Subnet별 상위 10개의 트래픽 양을 보여준다. 또한, 웹 기반의 사용자 인터페이스를 가짐으로써 사용자는 언제 어디서나 쉽게 네트워크를 모니터링 할 수 있다.

3. 호스트 트래픽 분석의 요구 사항

Enterprise 네트워크 내에서 Flow를 기반으로 한 호스트별 트래픽 정보가 트래픽의 다양한 분석에 기초적인 자료로 제공되기 위해서는 다음과 같은 사항들이 요구된다.

첫째, 각 호스트의 트래픽 양을 실시간에서부터 장기간 동안의 트래픽 양까지 나타낼 수 있어야 한다. 호스트별 트래픽 양을 파악하면 전체 네트워크에서 해당 호스트가 차지하는 비율을 파악할 수 있으며, 실시간의 분 단위 정보를 토대로 시, 일, 주, 월 단위의 각 호스트별 트래픽 누적량을 제공하면 해당 호스트의 네트워크 사용에 대한 패턴을 파악할 수 있다.

둘째, 트래픽 양의 변화를 그래프를 통해 확인할 수 있어야 한다. bps, pps, fpm 등의 트래픽 추이를 한 눈에 확인할 수 있는 방법이 그래프로 표현하는 방법이다.

셋째, 과거의 트래픽 분석 정보를 확인할 수 있어야 한다. 모니터링 시스템의 사용자가 실시간 트래픽 현황을 계속해서 지켜볼 수는 없다. 특정 시점의 트래픽 현황을 파악하기 위해서는 과거의 트래픽 정보를 저장해 놓아야 한다.

넷째, 호스트별 트래픽 분석 정보는 트래픽 양 뿐만 아니라 접속 IP, Port 및 Protocol 등 상세 정보를 제공할 수 있어야 한다.

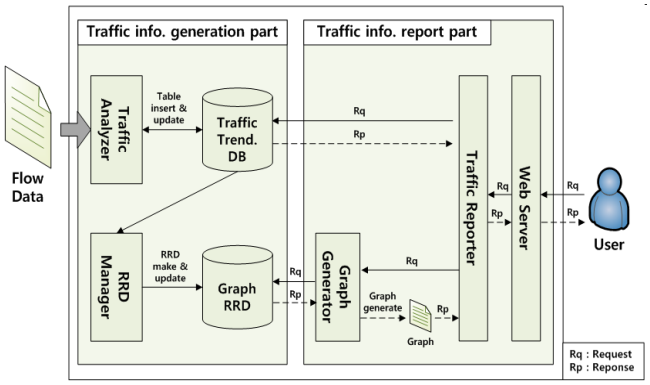
4. 시스템 설계

본 장에서는 KU-MON 시스템을 확장시켜 3장의 요구사항을 만족시키는 호스트 분석 시스템을 설계한 내용을 기술한다.

4.1 분석 시스템 전체 구조

(그림 2)는 분석 시스템의 전체적인 구조를 보여준다. Traffic Trend. DB, Graph RRD를 중심으로 트래픽 정보를 생성하는 부분과 트래픽 정보를 제공하는 부분으로 나누어진다.

트래픽 정보를 생성하는 부분은 매분 주기로 Traffic Analyzer, RRD Manager 2개의 모듈이 작동한다. Traffic

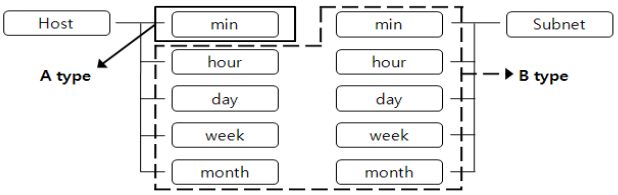


(그림 2) 분석 시스템의 전체 구조

Analyzer는 Traffic Trend. DB의 분, 시, 일, 주, 월 테이블에 해당 트래픽 정보를 분류하여 업데이트하며, RRD Manager는 각 호스트에 대한 RRD를 생성하고 업데이트하는 작업을 한다.

트래픽 정보를 제공하는 부분은 사용자의 요청에 의해 동작된다. 사용자가 웹을 통해 트래픽 정보를 요청을 하면 Traffic Reporter가 Traffic Trend. DB로부터는 트래픽 정보를 요청하고, Graph Generator에게는 그래프 생성 요청을 한다. 요청에 의해 응답이 완료되면 Traffic Reporter는 Web Server를 통해 사용자에게 트래픽 정보 및 추이 그래프를 제공한다.

4.2 Traffic Trend. DB



A type. Table attribute

Local_IP (4 byte)	Local_Port (2 byte)	Remote_IP (4 byte)	Remote_Port (2 byte)	Protocol (1 byte)	
InBytes (4 byte)	InPackets (4 byte)	InFlows (4 byte)	OutBytes (4 byte)	OutPackets (4 byte)	OutFlows (4 byte)

B type. Table attribute

Local_IP (4 byte)	InBytes (8 byte)	InPackets (8 byte)	InFlows (8 byte)	OutBytes (8 byte)	OutPackets (8 byte)	OutFlows (8 byte)

(그림 3) Traffic Trend. DB 테이블 스키마

Traffic Trend. DB는 호스트/서브넷을 기준으로 분, 시, 일, 주, 월 단위의 정보를 각각 하나의 테이블로 생성한다. (그림 3)에서 보는 바와 같이 총 10 종류의 테이블이 존재하며, 접두사 'Local'은 Enterprise 네트워크 내부를, 'Remote'는 외부를 나타낸다. 외부에서 내부로 향하는 트래픽은 'In'을, 내부에서 외부로 향하는 트래픽에는 'Out'이라는 접두사를 붙여 표현하였다.

분 단위 트래픽 정보는 짧은 기간의 정보로써 Local_Port, Remote_IP, Remote_Port, Protocol에 의해 분류된 상세한 트래픽 정보를 저장하기 위해 A type을 사용한다. 나머지 테이블은 외부의 수많은 IP/Port로 인해 분 단위와 같은 상세한 분류에 의한 정보 저장을 디스크의

제한으로 할 수가 없다. 따라서 분을 제외한 단위 시간 테이블은 Local_IP만을 분류 기준으로 하여 단위 시간동안의 트래픽 누적량을 표현하기 위해 B type을 사용한다.

일반적으로 호스트에서 발생할 수 있는 최대 Bandwidth가 100Mbps인 점을 고려한다면, 1분 동안의 Byte, Packet, Flow 양을 저장하기 위한 데이터의 크기는 4 byte이면 충분하다. 하지만, 시간 단위 이상에서는 4 byte의 데이터 크기로는 오버플로우의 발생으로 올바른 트래픽 양을 표현할 수 없기 때문에, 8 byte의 데이터 크기를 사용한다.

테이블의 저장 기간은 디스크의 효율적인 사용을 위해 일정 기간을 유지해야한다. 분 단위 테이블의 개수가 DB의 총 데이터 양의 대부분을 차지한다. 분 단위 테이블의 유지기간이 길면 그 만큼 오래 전까지 분 단위 정보를 통한 상세한 분석이 가능하지만, 디스크의 용량을 많이 차지한다. 그 반대로 유지 기간이 짧으면 디스크의 용량은 적게 차지하지만, 상세 분석이 가능한 시간이 짧아 트래픽 분석에 어려움이 생길 수 있다. 따라서 상황에 맞는 적절한 유지 기간을 정해야 한다. 또한, DB로 Mysql 사용 시 대용량 데이터를 관리할 때 DB 접근 속도 저하 등 DB의 성능 저하가 발생하는 문제를 보였다. 그래서 호스트/서브넷 기준으로 두 개의 DB로 테이블을 나누어 저장하고 분 단위 테이블의 유지 기간을 하루(1440개)로 정하였다. 그리고 시 테이블은 하루(24개), 일 테이블은 일주일(7개), 주 테이블은 1년(51~52개), 월 테이블은 10년(120개)으로 유지 기간을 정하였다.

4.3 RRD Manager & Graph Generator

각 호스트의 bps, pps, fpm에 대한 추이 그래프를 생성하기 위해서 본 시스템에서는 RRDTool[8]을 사용한다. RRDTool은 MRTG와 더불어 트래픽 추이 그래프를 그리기에 용이한 Tool이다. 각 호스트의 bps, pps, fpm 추이 그래프를 그리기 위해 각 호스트마다 3개의 RRD가 필요하다. RRD 생성 및 업데이트와 그래프를 그리는 작업은 많은 시스템 자원을 사용한다. 따라서 이를 적절히 처리하는 것이 전체적인 분석 시스템의 시스템 부하를 줄이는 것이 된다.

먼저 각 호스트에 대한 RRD가 생성되어야 한다. Enterprise 네트워크 내의 모든 IP에 대한 RRD를 생성하면 분석에 용이하지만, 디스크의 용량 문제와 실제 사용하지 않는 IP에 대해서는 불필요하므로 바람직하지 않다. 따라서 실제 사용 중인 호스트 IP에 대해서만 RRD를 생성하는 작업이 필요하다. 포트 스캔 등의 이유로 사용하지 않는 IP에도 트래픽이 발생할 수 있으므로 단순한 트래픽 발생 유무로는 사용 중인 호스트 IP의 분류가 어렵다. 일반적으로 실제 사용 호스트라면 최소한 한 번은 일정 수준 이상의 Bandwidth를 차지할 것이라 감안하고, RRD Manager는 호스트 IP가 최초로 일정 수준 이상의 Bandwidth를 차지하는 시점에 RRD를 생성한다. 추후에 실제 사용 중인 IP를 찾는 지능적인 알고리즘을 RRD Manager에 적용시킨다면 좀 더 효율적인 RRD DB 관리가 이루어질 것이다.

RRD는 bps, pps, fpm에 대한 정보를 매분 업데이트해야한다. Enterprise 네트워크에 있는 많은 수의 호스트에 대한 RRD를 모두 업데이트하는 작업은 현재 분석 컴퓨터의 사양으로는 많은 시간이 걸린다. 그에 대한 해결책으로 bps, pps, fpm의 수치가 일정 수준 이상 올라간 호스트 IP에 대해서만 업데이트를 하는 방법이 있다. 그래프의 목적은 트래픽의 추이를 살펴보기 위한 것으로 상세한 트래픽 현황을 확인할 필요는 없다. 따라서 낮은 양의 트래픽은 그래프에서 표시를 하지 않아도 표시를 할 때와 크게 차이점이 없다고 판단된다. 본 시스템에서는 Download, Upload 트래픽 중 하나라도 100 Kbps, 10 pps, 50 fpm 이상인 호스트 IP에 대해서만 RRD 업데이트 작업을 행한다. 해당 RRD 업데이트 알고리즘을 적용해 본 결과 분당 평균 70개 정도의 호스트 IP가 업데이트 된다.

800×200 그래프를 RRDTool로 생성하는 데 걸리는 시간은 약 0.2초이다. 따라서 각 호스트 당 3개씩 모든 호스트에 대해 그래프를 매분 그리는 것은 비효율적이다. 본 시스템에서는 호스트 트래픽 추이 그래프를 매분 그리지 않고, 웹 페이지에서 사용자의 요청이 발생하는 시점에 그래프를 그리는 방식을 택한다. 사용자 요청 발생 시, Traffic Reporter는 그래프를 생성해 주는 Graph Sever에게 그래프 생성 요청을 해서 해당 그래프를 생성한다. Traffic Reporter와 Graph Generator 사이의 통신은 UDP를 사용한다.

5. 시스템 구현

5장에서는 4장의 설계를 통하여 Flow 기반 호스트별 네트워크 분석 시스템을 구현한 내용을 기술한다.

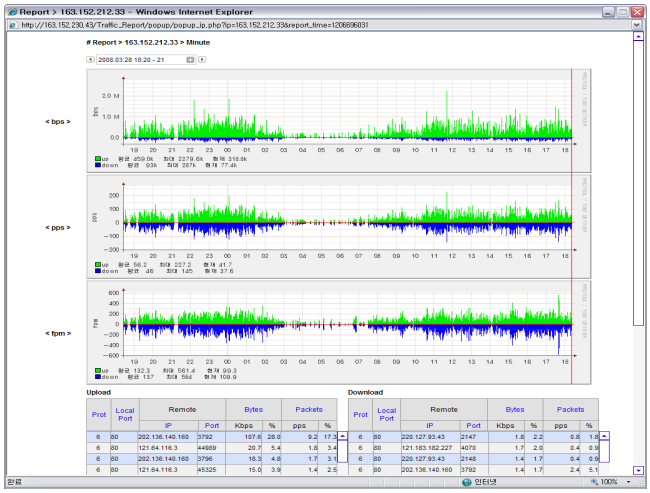
<표 1> KU-MON 시스템 개발 환경

OS	Linux Fedora 6
Web Server	Apache 2.2.3
Language	C, PHP5
DB	MySql 5.0.22, RRD
Tool	RRDTool, Cron

본 시스템의 개발 환경은 <표 1>과 같다. Linux 환경에서 Apache 웹 서버를 사용하였고, DB는 MySql와 RRD를 사용하였다. 웹을 만드는 Traffic Reporter는 PHP를 이용하여 구현하였고, 나머지 모듈들은 C를 이용하여 구현하였다. Flow 정보를 만드는 모듈과 분석하는 모듈, RRD Manager는 매 분 주기로 실행하기 위해 Linux에서 제공되는 Cron 데몬을 사용하였으며, 마지막으로 시간추이 그래프를 생성하기 위해 RRDTool을 이용하였다.

기본적으로 분, 시, 일, 주, 월 단위의 전체 네트워크에 대한 트래픽 양과 bandwidth 추이 그래프를 제공하고, 각 단위 시간의 모든 호스트/서브넷을 파악할 수 있다. IP에 연결된 링크를 통해 해당 호스트의 상세 트래픽 정보를

Local Port, Remote IP/Port, Protocol로 분류하여 제공하며, bps, pps, fpm에 대한 추이 그래프를 추가로 제공하여 상세한 정보를 얻을 수 있게 구성되어 있다. 호스트의 상세한 트래픽 정보를 제공하는 웹 페이지의 UI는 (그림 4)에서 보여주고 있다.



(그림 4) 호스트 정보 페이지의 UI

구축된 시스템은 현재 Campus 네트워크의 트래픽 모니터링을 위해 운용되고 있으며 아래의 URL을 통하여 웹으로 확인할 수 있다.

http://nmlab.korea.ac.kr/sys/traffic_report/

5.1 Campus 네트워크 트래픽의 특징

본 절에서는 본 시스템을 Campus 네트워크에 적용하여 관찰된 트래픽 특징에 대하여 설명한다.

첫째, 트래픽 양이 높은 상위 10개의 호스트가 전체 Bandwidth의 70% 이상의 비율을 차지하는 경우가 대부분이다. 이 호스트들의 트래픽 특징은 Bandwidth를 많이 차지할 뿐만 아니라 많은 양의 Flow 수를 가진다.

둘째, 일반적으로 서버가 많지 않은 Enterprise 네트워크에서는 Upload의 양보다는 Download의 양이 많을 것이라 예상된다. 하지만 장기간의 트래픽 정보를 보여주는 월단위 트래픽 정보를 볼 때, Upload, Download의 전체 트래픽에 대한 비율이 예상보다는 큰 차이를 나타내지 않았다.

셋째, 학교의 특성상 방학 중에는 학기 중보다 네트워크를 사용하는 인원이 상대적으로 적다. 따라서 학기 중과 방학 중의 트래픽 양의 차이가 많아야 하지만, 실제 방학 중과 학기 중 트래픽 양은 많은 차이를 보이지 않는다. 이는 공용 호스트가 아닌 개인 호스트에서 발생하는 트래픽 양이 전체 트래픽 양의 대부분을 차지하는 한다는 사실을 뒷받침한다.

넷째, 실제 사용하는 호스트 수 이상의 많은 IP가 측정되는 경우가 가끔 발생한다. 이러한 이유는 포트 스캔 등에 의해 발생한 트래픽으로 실제 네트워크를 사용하는 데에는 불필요한 트래픽으로 간주된다.

6. 결론

본 논문에서는 현재 대용량 Enterprise 네트워크의 트래픽을 분석하는 방법에 있어 호스트 트래픽 분석을 제시하고 있다. 다양하고 복잡한 트래픽의 발생으로 네트워크의 전체적인 분석뿐만 아니라 호스트를 기반으로 하는 상세한 분석에 대한 요구가 앞으로 계속 증가할 것이다. 따라서 본 논문은 호스트 트래픽 분석이 요구하는 사항들을 정리하였으며, 시스템 설계 시 여러 가지 문제 사항들을 살펴보고 해결책을 제시함으로써 향후 호스트 트래픽 분석 시스템을 구축하려는 연구에 도움을 줄 것으로 기대한다. 또한 호스트 트래픽 분석 정보를 통해 어플리케이션 분류 등 다양한 연구를 위한 기초 정보로 제공되기를 기대한다.

향후 보완점 및 연구 과제로는 시스템의 성능 및 효율성의 개선 방안으로 Enterprise 네트워크에서 실제 사용중인 IP를 구별하는 알고리즘을 적용하는 방법을 들 수 있다. 그리고 Campus 네트워크의 특징들로 보아 P2P를 사용하는 호스트가 많은 트래픽 양을 발생시키는 것으로 보인다. 따라서 현재 호스트에서 사용하는 P2P 어플리케이션에 대한 트래픽을 분석하는 연구를 계획하고 있다.

참고문헌

[1] 박상훈, 박진완, 김명섭, "Flow 기반 실시간 트래픽 수집 및 분석 시스템", 정보처리학회 추계학술대회, 목포대학교, 전주, Nov. 9-10, 2007, pp. 1061.

[2] Cisco Systems, White Papers, "Introduction to Cisco IOS NetFlow," May. 2007.

[3] Myung-Sup Kim, Hun-Jeong Kang, Seong-Cheol Hong, Seung-Hwa Chung, James W. Hong, "A Flow-based Method for Abnormal Network Traffic Detection", Accepted to appear in the Proc. of NOMS 2004, Seoul, Korea, April 2004.

[4] N. Brownlee, C. Mills and G. Ruth, "Traffic Flow Measurement: Architecture", IETF RFC 2722, October 1999.

[5] N. Brownlee, "Traffic Flow Measurement: Experiences with NeTraMet", IETF RFC2123, March 1997.

[6] Se-Hee Han, Myung-Sup Kim, Hong-Taek Ju and James W. Hong, "The Architecture of NG-MON: A Passive Network Monitoring System", LNCS 2506, DSOM 2002, October 2002, Montreal Canada, pp. 16-27.

[7] IETF Working Group IPFIX (IP Flow Information Export), http://www.ietf.org/html.charters/ipfix-charter.html.

[8] RRDTool, http://www.rrdtool.org.