플로우 전달 특성 기반의 P2P 헤비 트래픽 검출 알고리즘

(An Algorithm to Detect P2P Heavy Traffic based on Flow Transport Characteristics)

최 병 걸 † 이 시 영 †† 서 영 일 †††
(Byeong-Geol Choi) (Si-Young Lee) (Yeong-il Seo)
위 즈 빈 †††† 전 재 현 †††† 김 승 호 †††††
(Zhibin Yu) (Jae-Hyun Jun) (Sung-ho Kim)

요 약 최근 분산 컴퓨팅 환경이 확대되고 네트워크 기반의 응용프로그램이 다양하게 개발됨에 따라 네트워크 트래픽이 증가되고 있으며, 트래픽 종류도 P2P(Peer to Peer), 실시간 동영상등과 같이 다양해지고 있다. 네트워크 트래픽 중에서 P2P 트래픽이 지속적으로 증가되면서 많은 대역폭을 차지하고 있기 때문에 웹, 파일 전송 및 실시간 동영상등과 같은 다른 네트워크 응용프로그램의 서비스 품질을 보장하지 못하는 상황이 빈번하게 발생하고 있다. P2P 트래픽으로 인한 문제점을 해결하기 위해 기존에 포트 기반의 P2P 트래픽 검출 기법과 패킷들의 내용를 검사하는 DPI(Deep Packet Inspection) 방식의 검출 기법들이 제시되었으나 최근의 P2P 응용프로그램들이 고정된 포트를 사용하지 않으며, 패킷들의 내용을 암호화하여 전송함으로써 기존의 연구 방법을 P2P 트래픽 검출에 적용하기가 어려운 상황이다. 본 논문에서는 기존의 포트 기반의 P2P 트래픽 검출 기법과 DPI 기법의 문제점들을 해결할 수 있는 플로우(flow) 매개변수의 상관 관계를 이용한 플로우 전달 특성 기반의 P2P Heavy 트래픽 검출 알고리즘을 제시한다. 본 논문에서 제시하는 알고리즘은 P2P 트래픽 중에서 네트워크 대역폭을 가장 많이 차지하는 컨텐츠 다운로드 P2P 트래픽을 검출하는 것이다. P2P 트래픽은 컨텐츠를 가지고 있는 상대 노드(Peer)들을 검색하는 단계와 검색된 노드들 중에 하나 이상의 노드로부터 컨텐츠를 다운로드하는 단계로 이루어진다. 이러한 P2P 응용프로그램들의 특성을 P2P 플로우 패턴으로 정의하고 이를 기반으로 P2P Heavy 트래픽을 검출하는 알고리즘을 개발하였다.

키워드 : 플로우, 플로우 전달 특성, P2P 플로우 패턴, 소스 스캔 플로우, P2P Heavy 플로우

Abstract Nowadays, transmission bandwidth for network traffic is increasing and the type is varied such as peer-to-peer (P2P), real-time video, and so on, because distributed computing environment is spread and various network-based applications are developed. However, as P2P traffic occupies much volume among Internet backbone traffics, transmission bandwidth and quality of service(QoS) of other network applications such as web, ftp, and real-time video cannot be guaranteed. In previous research, the port-based technique which checks well-known port number and the Deep Packet Inspection(DPI) technique which checks the payload of packets were suggested for solving the

† 비 회 원 : 경북대학교 컴퓨터공학과
gulss73@gmail.com
†† 비 회 원 : 마하넷 연구소 연구소장
edward.lee.mahanet@gmail.net
††† 비 회 원 : KT 네트워크 연구소 수석연구원
syi@hana.ne.kr
†††† 비 회 원 : 경북대학교 전자전기컴퓨터학부
zbyu@mmlab.knu.ac.kr
jhjun@mmlab.knu.ac.kr
††††† 종신회원 : 경북대학교 컴퓨터학부 교수
shkim@knu.ac.kr

논문접수 : 2009년 7월 6일
심사완료 : 2010년 5월 11일

정보과학회논문지: 정보통신 제37권 제5호(2010.10)

problem of the P2P traffics, however there were difficulties to apply those methods to detection of P2P traffics because P2P applications are not used well-known port number and payload of packets may be encrypted. A proposed algorithm for identifying P2P heavy traffics based on flow transport parameters and behavioral characteristics can solve the problem of the port-based technique and the DPI technique. The focus of this paper is to identify P2P heavy traffic flows rather than all P2P traffics. P2P traffics are consist of two steps i)searching the opposite peer which have some contents ii) downloading the contents from one or more peers. We define P2P flow patterns on these P2P applications' features and then implement the system to classify P2P heavy traffics.

Key words : Flow, Flow transport characteristics, P2P flow pattern, Source Scan Flow, P2P heavy flow

1. 서 론

최근 인터넷 사용자가 전 세계적으로 급격히 증가하고, 네트워크 기반의 응용프로그램이 다양하게 개발되어 사용됨에 따라 네트워크 트래픽이 급격히 증가하고 있다. 이를 지원하기 위해 인터넷 서비스 사업자들(ISP, Internet Service Provider)은 네트워크를 계속해서 증설하고 있으며 최근 2.5Gbps~10Gbps 급의 백본 네트워크로 대용량화 되고 있고, 앞으로 몇 년 후에는 테라급의 네트워크가 구축될 것으로 예상된다. 이러한 증가된 네트워크 트래픽을 분석해보면 기존의 텍스트나 이미지 위주의 트래픽이 스트리밍 미디어 및 P2P 위주의 트래픽으로 빠르게 변화하고 있음을 알 수 있다. 특히 P2P 트래픽은 인터넷 백본 트래픽의 많은 대역폭을 차지하는 것으로 분석되어지고 있어서 P2P 트래픽을 실시간으로 검출하고 이를 제어함으로써 네트워크의 자원을 효율적으로 관리하는 것이 매우 중요한 문제로 부각되고 있다. 하지만, 현재의 네트워크 관리 시스템들은 트래픽 모니터링에 따른 통계값 제공 등의 단순 평면적인 정보만을 제공할 뿐, 네트워크 트래픽을 정밀 분석하고 그에 따른 대책을 마련하기에는 미약하다. 특히, 인터넷 백본 트래픽 중에 가장 많은 대역폭을 차지하는 P2P 트래픽을 실시간으로 검출하는 well-known port 기반의 P2P 트래픽 검출 기법 연구[1] 및 패킷들의 Payload를 보고 다양한 P2P 응용프로그램이 가지는 시그너쳐를 검사하는 DPI 검출 기법 연구[2-6]가 이루어졌으나, 이러한 P2P 검출 기법들이 가지는 문제점으로 인해 실제 네트워크에 적용하기에는 한계가 있다.

P2P는 과거 FTP, HTTP, Telnet등의 클라이언트/서버 구조의 응용프로그램이 발생시키는 트래픽과는 달리 개인 사용자들간(peer to peer)에 컨텐츠를 공유하는 네트워크 구조를 띄고 있기 때문에 P2P에서 발생하는 네트워크 트래픽의 특성도 다르다고 할 수 있다. 또한 다양한 P2P 응용프로그램이 지속적으로 새롭게 개발되어지고 있고, 이러한 P2P 응용프로그램들이 고정된 포트를 사용하지 않으며, 패킷들의 내용을 암호화하여 전송함으로써 이에 대한 분석 및 연구가 이루어져야 한다. 특히 P2P 트래픽 중에 컨텐츠 다운로드 트래픽에 해당하는 P2P Heavy 트래픽을 검출할 수 있는 연구가 필요하다.

본 논문에서 제시하는 P2P Heavy 트래픽 검출 알고리즘은 플로우 매개 변수들의 상관 관계를 나타내는 플로우 전달 특성, 즉 P2P 응용프로그램은 사용자가 원하는 컨텐츠를 가지고 있는 Peer들을 검색하고 그 Peer들 중에 하나 이상의 Peer로부터 컨텐츠를 다운로드한다는 플로우 전달 특성을 기반으로 한다. 이러한 플로우 전달 특성 기반의 P2P Heavy 트래픽 검출 알고리즘은 기존에 연구된 P2P 검출 기법들이 가지는 문제점을 해결할 수 있으며, 새롭게 개발되는 P2P 응용프로그램이 발생시키는 트래픽의 검출도 가능하다.

본 논문의 구성은 다음과 같다. 제2장에서는 기존에 연구된 P2P 트래픽 검출 기법에 대해서 기술하고, 제3장에서는 P2P 트래픽을 분석하여 P2P 트래픽의 플로우 전달 특성 및 패턴을 도출한다. 제4장에서는 도출된 P2P 트래픽의 플로우 패턴을 이용한 P2P Heavy 트래픽 검출 알고리즘을 제시한다. 그리고 제5장에서는 본 논문에서 제시한 P2P 트래픽 검출 알고리즘의 검출 결과를 제시하고, 마지막으로 제6장에서 본 논문의 결론을 내도록 한다.

2. P2P 트래픽 분석 관련 연구

기존의 네트워크 제어 및 관리 장비들은 TCP/UDP 또는 IP의 패킷 정보를 기반으로 해당 장비들의 특정 단위 목표인 라우팅이나 QoS 보장 및 DDoS 방지 등을 달성하기 위해 노력해 왔다. 그러나 패킷 기반의 접근 방법은 상위 응용프로그램들의 통신 관계에 따른 정보들을 무시하고, 단순히 일시적인 정보 전달 단위인 각각의 분리된 패킷에 담겨 있는 정보들에만 의존함으로써, 적게는 처리 속도의 한계성과 크게는 적용성의 한계로 인해 패킷 라우팅을 위한 라우터, DDoS 공격을 방어하는 전용 시스템, 또는 트래픽 제어를 위한 DPI 시스템 등과 같은 독립적인 목표를 위한 단일 시스템의 형태로

제공되고 있다.

네트워크 트래픽을 분석하고 응용프로그램 별로 트래픽을 분류하는 기존 연구 중 대표적인 기법은 크게 Port 기반의 트래픽 분류[1], Payload 기반의 트래픽 분류[2-6], 호스트 동작 기반의 트래픽 분류[7-9], 그리고 마지막으로 Machine Learning 기반의 트래픽 분류 기법[10-21]으로 나뉜다. 이러한 트래픽 분류 기법 중에 실시간으로 트래픽을 분류할 수 있는 기법은 Port 기반과 Payload 기반의 트래픽 분류 기법이며, 나머지 두 개의 기법은 실시간성을 보장하지 못하여 단지 네트워크 트래픽 분석 용도로만 사용되고 있다.

Port 기반의 트래픽 분류 및 P2P Heavy 트래픽 검출 기법은 최근의 P2P 응용프로그램이 잘 알려진 포트를 사용하지 않고, 사용 Port를 사용자가 동적으로 설정하여 사용할 수 있도록 제공하고 있어서 P2P Heavy 트래픽 검출이 어렵다. 그리고 Payload 기반의 검출 기법은 현재 DPI 네트워크 장비에 도입되어 실제 네트워크에 적용되어 있으나 패킷의 Payload를 보고 P2P 응용프로그램별 시그너쳐를 검색하는 방법의 다음과 같은 문제점으로 인해 네트워크 적용의 한계점을 가진다.

첫째, 전송되는 모든 패킷의 Payload를 검사하기 때문에 프로세싱의 오버헤드가 너무 크다. 둘째, Payload가 암호화된 패킷일 경우 암호화를 풀 수 있는 방법이 없으므로 P2P 응용프로그램별 시그너쳐를 검출할 수 없다. 셋째, 패킷 Payload 검사로 인해 네트워크 응용프로그램 사용자의 프라이버시를 훼손할 수 있는 상황이므로 법적인 문제를 내포하고 있다. 넷째, 새로운 P2P 응용프로그램이 등장할 경우 P2P 응용프로그램의 시그너쳐를 알 수 없으므로 P2P Heavy 트래픽의 검출이 불가능하다. 즉, 알려지지 않은 P2P 응용프로그램에 대한 대응 자체가 불가능하다.

위와 같은 문제점을 해결하기 위해서 본 논문에서 제시하는 P2P Heavy 트래픽 검출 알고리즘은 호스트 동작 기반의 트래픽 분류 기법을 기반으로 P2P 응용프로그램이 생성하는 트래픽의 행위 패턴을 분석, 정의하여 적용하였다. 즉 3장에서 제시하는 P2P 플로우 패턴을 기반으로 포트 정보뿐만 아니라 다양한 플로우 매개 변수 정보를 이용하고, 패킷의 Payload를 조사하지 않음으로 인해 포트 기반 검출 기법과 Payload 기반의 검출 기법의 단점을 극복하였다.

3. P2P 트래픽 플로우 전달 특성 정의

본 장에서는 P2P Heavy 트래픽 검출 알고리즘을 도출하기 위해 P2P 응용프로그램 별로 발생되는 트래픽을 분석하여 P2P 플로우 전달 특성을 정의하도록 한다.

3.1 플로우 정의

네트워크 응용프로그램이 실행되는 세션에서 발생되는 트래픽은 하나 또는 그 이상의 플로우로 나뉘어질 수 있으며, 패킷 단위가 아닌 플로우 단위로 라우팅, QoS, 트래픽 분류 등을 수행할 경우 모든 패킷들을 검사할 필요가 없어지는 장점을 가지고 있다. 이러한 플로우는 패킷들의 주소 쌍(송신자 주소, 송신자 포트 번호, 수신자 주소, 수신자 포트), 호스트 쌍(송신자 네트워크 주소, 수신자 네트워크 주소), 프로토콜 등으로 명세되는 제한된 시간 내에 연속적으로 전달되는 IP 패킷들의 흐름이다. 이러한 플로우 모델은 Packet Train 모델로 처음 제안되었으나, 플로우의 정의는 사람마다 약간씩은 다를 수 있다. 본 논문에서는 일정 시간 내에 5-Tuple (Source Address, Destination Address, Source Port, Destination Port, Protocol) 값이 일치하는 연속적인 패킷들의 집합을 플로우로 정의한다. 플로우 정의를 수식으로 나타내면 식 (1)과 같다.

$$f = \{p_0, p_1, p_2, \cdots, p_i\} \; if \; T_{p_{i+1}} - T_{p_i} < T \quad (1)$$

식 (1)에서 p_i는 패킷, T_{p_i}는 패킷 p_i의 도착 시간, T는 패킷 p_i와 p_{i+1} 사이의 임계치 시간을 나타낸다. 플로우 f는 패킷 p_i 들의 연속된 집합이며, 패킷 p_i와 p_{i+1}의 패킷 도착 시간인 T_{p_i}, $T_{p_{i+1}}$의 차가 T 임계치보다 크면 p_{i+1} 패킷을 포함한 이후의 패킷들은 다른 플로우로 구별된다. 본 논문에서는 T의 값을 상용화된 플로우 라우터 장비에서 사용하는 임계치 값인 2초로 설정하여 네트워크 트래픽을 플로우 단위로 구분한다.

본 논문에서 제시하는 P2P Heavy 트래픽 검출 알고리즘은 패킷 단위의 프로세싱이 아닌 플로우 정의에 기반한 플로우 단위의 프로세싱으로 P2P 플로우 패턴을 도출한다.

3.2 플로우 매개 변수 정의

플로우 매개 변수와 전달 특성은 P2P Heavy 트래픽을 검출하기 위한 입력 데이터로 활용된다. 이러한 값들은 네트워크에서 검출되는 트래픽을 분석하고 분류하는데 사용되며, 응용프로그램 별로 각각 다른 매개 변수 값과 전달 특성을 가지고 있다. 플로우 매개 변수와 전달 특성은 크게 세가지로 정의된다.

첫째, 플로우 매개 변수(Flow Parameter)로 하나의 플로우가 생성되고 종료되기까지의 특성을 나타낸다. 플로우 매개 변수의 종류는 플로우의 5-Turple 정보, 플로우를 구성하는 패킷의 수를 나타내는 Packet Count, 패킷 크기들의 총합을 나타내는 Flow Size, 그리고 플로우의 지속시간을 나타내는 Flow Duration 이다. 플로우 매개 변수의 종류는 연구를 통해 지속적으로 도출되어야 하며, 이러한 플로우 매개 변수만를 이용하여 특정 응용프로그램에서 발생시키는 트래픽 특성을 표현할 수

도 있다. 즉, 플로우 매개 변수는 단일 플로우를 나타낼 수 있는 특성 값으로 정의한다.

둘째, 파생된 플로우 매개 변수(Derived Flow Parameter)로 플로우 매개 변수들의 조합된 결과로 나온 값이다. 파생된 플로우 매개 변수의 종류로는 Average Packet Size, Average Rate, Heavy 플로우 등이다. 예를 들면, Heavy 플로우는 일정 시간(Flow Duration)내에 Average Rate가 특정 임계치 값을 넘으면 그 플로우를 Heavy 플로우로 정의한다. Flow Duration과 Average Rate의 임계치 값 및 설정 기준에 대해서는 본 논문의 5장 P2P Heavy 트래픽 플로우 검출 결과에서 기술한다.

셋째, 플로우 패턴(Flow Pattern)으로 하나의 호스트에서 발생한 플로우들의 상관 관계로 정의된다. 예를 들면, 컨텐츠의 다운로드 서비스를 제공하는 P2P일 경우 컨텐츠를 검색하는 플로우들의 집합과 그 집합에 속한 플로우들의 목적지 Peer들 중에 하나 이상의 Peer들로부터 컨텐츠를 다운로드하는 플로우로 구성된다. 이러한 P2P 플로우들의 상관 관계를 P2P 플로우 패턴으로 정의한다. 플로우 패턴은 네트워크를 사용하는 각 응용프로그램의 동작 특성에 따라 다르게 정의될 것이며 본 논문에서는 P2P Heavy 트래픽 검출을 위해 P2P 플로우 패턴만을 정의하도록 한다.

3.3 P2P 플로우 패턴 정의

본 논문에서 제시하는 플로우 전달 특성 기반의 P2P Heavy 트래픽 검출 알고리즘 도출을 위해 다양한 P2P 응용프로그램에서 생성하는 트래픽을 분석하였으며, 그 결과 P2P 응용 프로그램은 항상 임의의 Peer들에게 연결 가능 유무, 컨텐츠 존재 유무, 컨텐츠 다운로드 가능 유무를 검색하는 단계와 실제적으로 컨텐츠를 다운로드하는 단계로 구분된다.

두 단계 중 첫번째 단계를 본 논문에서는 소스 스캔(Source Scan) 단계로 정의하고, 소스 스캔하는 트래픽을 소스 스캔 플로우로 정의한다. 그리고 소스 스캔 플로우들의 집합을 소스 스캔 그룹(Source Scan Group)으로 명명한다. 소스 스캔 단계에 이어서 소스 스캔한 목적지 Peer들 중에 하나 이상의 Peer로부터 컨텐츠를 다운로드 받는 단계가 수행되며, 이 단계에서 발생하는 플로우들 중에 그림 2의 조건에 부합하는 플로우를 P2P Heavy 플로우로 정의한다.

P2P 플로우 패턴은 그림 1의 소스 스캔 그룹과 그림 2의 Heavy 플로우로 나타낼 수 있다. 플로우는 5-Turple 정보를 이용하여 하나의 플로우를 하나의 라인으로 표현할 수 있으며, 그림 1은 하나의 호스트에서 N개의 상대 Peer 호스트를 대상으로 연결 가능 유무, 컨텐츠 존재 유무, 컨텐츠 다운로드 가능 유무 등을 검색할 때 발생하는 플로우들을 나타낸 것이다.

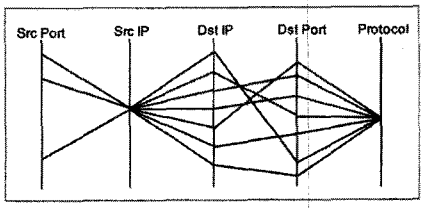

그림 1 소스 스캔 그룹

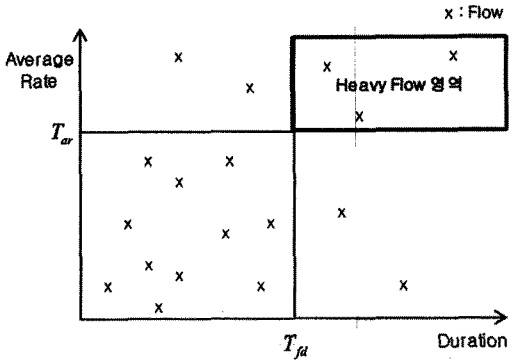

그림 2 Heavy Flow 조건

그림 2는 네트워크 트래픽 플로우 중에 Heavy 플로우에 속하는 조건을 나타낸 것이며, Flow Duration과 Average Rate의 임계치 값 설정에 대한 자세한 내용은 본 논문의 5장 P2P Heavy 트래픽 플로우 검출 결과에서 기술한다. P2P 플로우 패턴 FP_{p2p}를 수식으로 표현하면 식 (2)와 같다.

$$FP_{p2p} = S_{ssg} \; and \; f_{heavy}$$

$$S_{ssg} = \{f_0, f_1, f_2, \cdots, f_{n-1}\} \; if \; PC(f_i) < T_{pc}$$
$$f_{heavy} = f_i \; if \; AR(f_i) > T_{ar}, FD(f_i) > T_{fd} \qquad (2)$$

식 (2)에서 S_{ssg}는 P2P 소스 스캔 플로우들의 집합을 나타내며, f_{heavy}는 Heavy 플로우를 나타낸다. $PC(f_i)$는 f_i의 Packet Count, T_{pc}는 소스 스캔 플로우 조건을 검사하는 Packet Count의 임계치 값을 나타낸다. 그리고, $AR(f_i)$은 f_i의 Average Rate, $FD(f_i)$는 f_i의 Flow Duration, T_{ar}, T_{fd}는 각각 Heavy 플로우 조건을 검사하는 Average Rate, Flow Duration의 임계치 값을 나타낸다.

그림 1에서 보는 바와 같이 P2P 소스 스캔은 하나의 호스트에서 다수의 목적지 Peer들과의 많은 플로우를 생성한 후에 그 중에 하나 이상의 목적지 Peer로부터 P2P Heavy 트래픽 플로우가 발생한다.

$$P2P\, f_{heavy} = f_{heavy} \; if \begin{cases} S_{ip}(f_{heavy}) == D_{ip}(f_i), \\ f_i \in S_{ssg} \;\; (0 \le i < n) \end{cases} \qquad (3)$$

즉, 식 (3)에서 보는 바와 같이 플로우 f_{heavy}의 소스 주소 $S_{ip}(f_{heavy})$는 소스 스캔 플로우들 중에 임의의 한 플로우의 목적지 주소인 $D_{ip}(f_i)$와 반드시 일치한다. 다음 장에서 설명할 P2P 트래픽 플로우 검출 알고리즘은 위와 같은 P2P 플로우 패턴 특성을 기반으로 하여 작성되었다.

4. P2P 트래픽 플로우 검출 알고리즘

본 장에서는 P2P 트래픽 플로우 전달 특성 정의에서 명시한 P2P 플로우 패턴을 기반으로 P2P Heavy 트래픽 검출 알고리즘을 기술한다. 플로우 전달 특성 기반의 P2P Heavy 트래픽 검출 알고리즘은 컨텐츠를 검색하는 플로우들의 집합인 소스 스캔 그룹을 검출하는 단계, P2P Heavy 플로우의 후보가 되는 Heavy 플로우를 검출하는 단계, 그리고 마지막으로 Heavy 플로우 중에 소스 스캔 그룹에 속하는 P2P Heavy 플로우를 검출하는 단계로 구분된다. 그림 3은 본 논문에서 제시하는 알고리즘의 동작 순서를 다이어그램으로 표현한 것이다.

그림 3에서 나타내는 P2P Heavy 트래픽 검출 과정은 P2P 소스 스캔 플로우들의 집합인 소스 스캔 그룹의 검출율 결과에 의해 P2P Heavy 트래픽의 검출율이 결정된다.

소스 스캔 플로우를 검출하는 과정을 pseudo 코드로 나타내며 그림 4와 같다.

그림 4와 같이 소스 스캔 후보 플로우 중에 잘 알려진 포트를 사용하는 플로우와 패킷 개수가 T_{pc} 임계치 보다 큰 값을 가지는 플로우는 소스 스캔 플로우에서 제외시키고, 그 외의 플로우들을 소스 스캔 그룹 S_{ssg}에 포함시킨다.

그림 4에서 생성된 각각의 S_{ssg}들은 적게는 몇 개의 소스 스캔 플로우만을 가질 수도 있고, 많은 경우는 수

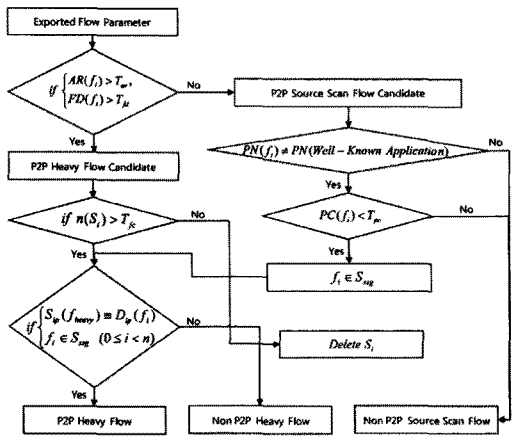

그림 3 P2P Heavy 플로우 검출 동작 순서도

```
Algorithm Filter_SSF (Source Scan Flow)
    input:
        structure of non-heavy flow( f_i ) parameter
        port number table of well-known applications
    output:
        true or false (true if p2p source scan flow, otherwise false)
    define variable:
        boolean, result ← false

    if( PN(f_i) ≠ PN(Well-KnownApplication) )
    {
        if( PC(f_i) < T_pc )
        {
            f_i ∈ S_ssg ;
            result ← true;
        }
    }
    return result;
```

그림 4 소소 스캔 플로우 검출 pseudo 코드

```
Algorithm Filter_SSG (Source Scan Group)
    input:
        an array of S_i (S_i ∈ {S_0, S_1, S_2, ···, S_n-1})
        threshold value of source scan flow number
    output:
        a set of source scan group
    define variables:
        an integer, flowcount ← 0

    while( eachS_i )
    {
        Flowcount ← n(S_i) ;
        if( Flowcount < T_fc )
        {
            delete S_i;
        }
    }
    return (array of source scan groups)
```

그림 5 소스 스캔 그룹 필터링 pseudo 코드

천, 수만 개의 소스 스캔 플로우를 가진다. 그러나 P2P 응용프로그램이 발생시키는 트래픽을 분석한 결과 P2P 소스 스캔 플로우는 컨텐츠를 검색하는 Peer들의 개수만큼 생성되고, 그 개수는 적어도 수십 개 이상이다. 따라서, 몇 개의 소스 스캔 플로우를 가지는 S_{ssg}는 P2P 응용프로그램이 생성시키는 소스 스캔 그룹이 아닐 가능성이 높다. 따라서, 그림 5에서 제시한 소스 스캔 그룹 필터링에서 T_{ssg} 임계치 이상의 개수를 가지는 S_{ssg}만을 P2P 소스 스캔 그룹으로 판단한다.

$$P2P\ S_{ssg} = S_i\ if\ n(S_i) > T_{fc} \quad (4)$$

본 본문에서 제시하는 알고리즘은 위의 수식에 부합하는 S_{ssg}들을 이용하여 P2P Heavy 트래픽을 검출하며, 검출 알고리즘의 Pseudo 코드는 그림 6과 같다.

그림 6에서 보는 바와 같이 P2P Heavy 트래픽 플로

```
Algorithm Detect_P2P_Heavy_Flow
    input:
        structure of heavy flow(f_heavy) parameter
    output:
        true or false (true if P2P heavy flow, otherwise false)
    define variable:
        boolean, result ← false

    // Phase 1
    filter_SSF( );
    filter_SSG( );
    build_hashtables(); // build hash table for each S_i

    // Phase 2
    while( each hash table of hash (D_ip(f_j)) (0 ≤ j < n) )
    {
        if( hash _table ( hash (S_ip(f_heavy))) ≠ null )
        {
            result ← true;
        }
    }
    return result;
```

그림 6 P2P Heavy 플로우 검출 pseudo 코드

우를 검출하는 Phase 2의 성능 향상을 위해 소스 스캔 그룹 집합들에 속한 모든 플로우들의 목적지 주소 값을 Hash Table로 구성하고 현재 발생한 Heavy 트래픽 플로우 소스 주소의 Hash Function 값이 Hash Table 내에 존재하면 그 플로우를 P2P Heavy 플로우 판단한다. 본 장에서 제시한 알고리즘의 검출율은 다음 장에서 기술한다.

5. P2P Heavy 트래픽 플로우 검출 결과

본 장에서는 4장에서 제시한 알고리즘을 검증하기 위한Heavy 플로우 검출 조건 분석, P2P Heavy 플로우 검출 환경, 그리고 검출 결과에 대해서 기술한다.

5.1 Heavy 플로우 검출 조건 분석

본 논문에서 제시하는 플로우 전달 특성 기반의 P2P Heavy 플로우 검출 알고리즘은 P2P 응용프로그램에서 생성되는 모든 트래픽을 검출하는 것이 아니라, P2P 트래픽 중에 네트워크 대역폭의 많은 부분을 차지하는 P2P Heavy 트래픽만을 검출하는 것이다. 따라서 Heavy 플로우를 검출하는 조건인 Flow Duration과 Average Rate의 임계치 값이 타당해야 한다.

Heavy 플로우 검출 조건의 타당성을 검증하기 위해서 실제 네트워크 트래픽을 이용하여 Flow Duration과 Average Rate의 임계치 값을 변경했을 경우 그 조건에 속하는 Flow들의 총 볼륨의 합이 전체 트래픽 볼륨 중에 어느 정도 차지하는지를 분석하였다.

실망 데이터는 특정 ISP 사업자의 가입자 망 1Gbps 링크에서 캡처한 데이터이며 22시 29분에서 22시 59분 사이의 30분간 트래픽이고, 볼륨은 약 25Gbytes이다. 이러한 실망 데이터를 플로우 단위로 Flow Duration과

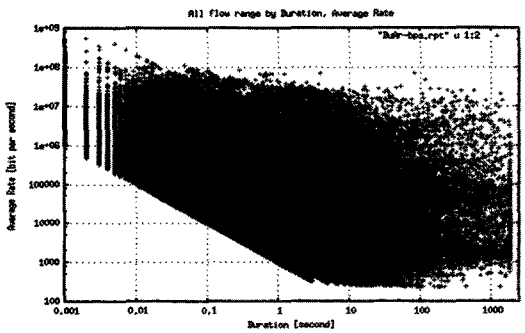

그림 7 Flow Duration과 Average Rate 그래프

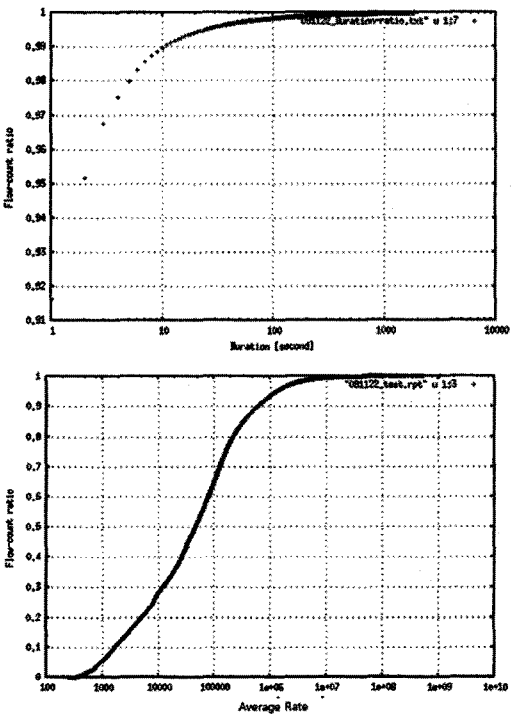

그림 8 Duration, Average Rate별 Flow 수 그래프

Average Rate의 상관 관계 그래프로 나타내면 그림 7과 같다.

그림 7에 보는 바와 같이 Flow Duration이 짧고 Average Rate가 낮은 플로우가 전체 플로우에서 많은 수를 차지하고 있다. 이에 대한 정확한 통계 정보를 그래프로 나타내면 그림 8과 같다.

그림 8에서 보는 바와 같이 Flow Duration이 1초 미만인 플로우의 수가 약 90% 이상을 차지하고, Average Rate가 1Mbps 미만인 플로우가 약 94% 정도 차지하고 있지만, Flow Duration과 Average Rate별로 트래픽의 볼륨을 그래프로 나타내면 Flow Duration이 1초 미만

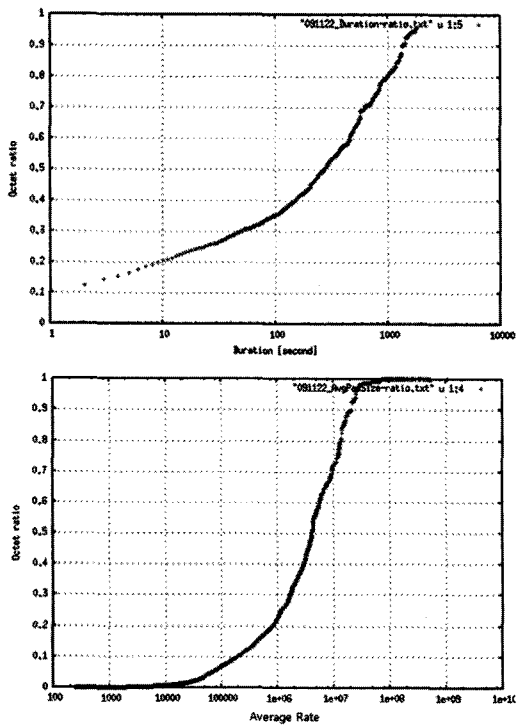

그림 9 Duration, Average Rate별 트래픽 볼륨 그래프

의 플로우는 전체 트래픽 볼륨에서 약 10%를 차지하고, Average Rate가 1Mbps 미만인 플로우는 약 20%만을 차지하고 있음을 알 수 있다. 그림 9는 Flow Duration과 Average Rate별 트래픽 볼륨 상관 관계 그래프를 나타낸 것이다.

결론적으로 Flow Duration이 일정 시간 이상이고, Average Rate가 일정 속도 이상인 플로우들이 네트워크 대역폭에 큰 영향을 준다는 것을 알 수 있다. 이런 분석 결과를 기반으로 Flow Duration과 Average Rate 임계치를 특정 범위내에서 변경하면서 전체 트래픽 볼륨 중에 몇 %를 차지하는지 분석하였으며, 그 결과는 표 1과 같다.

표 1에서 보는 바와 같이 Heavy 플로우를 검출하기 위한 Flow Duration과 Average Rate 의 임계치 값이 하나의 값으로는 정해지는 것이 아니라 인터넷 서비스 사업자의 네트워크 망 정책에 의해 결정되어야 한다.

본 논문의 실험에서는 표 1의 임계치 값 중에 Flow Duration은 5초, Average Rate는 400Kbps 값을 이용하여 실험하였으며, 이 때에 전체 트래픽 중에 차지하는 볼륨은 약 73.11%이다.

5.2 P2P Heavy 플로우 검출 환경

본 논문에서 제시하는 플로우 전달 특성 기반의 P2P Heavy 플로우 검출 알고리즘을 검증하기 위해 테스트

표 1 Duration과 Average Rate별 플로우 수 및 트래픽 볼륨

Duration (second)	Average Rate (Kbps)	Percentage per Flow Count(%)	Percentage per Traffic Volume(%)
3	100	0.64%	81.73%
4	100	0.47%	79.95%
5	100	0.36%	78.86%
6	100	0.30%	77.91%
7	100	0.25%	77.03%
3	200	0.41%	79.14%
4	200	0.30%	77.45%
5	200	0.24%	76.44%
6	200	0.20%	75.56%
7	200	0.17%	74.73%
3	300	0.31%	77.26%
4	300	0.22%	75.63%
5	300	0.18%	74.69%
6	300	0.15%	73.86%
7	300	0.13%	73.08%
3	400	0.24%	75.54%
4	400	0.18%	74.00%
5	400	0.15%	73.11%
6	400	0.13%	72.31%
7	400	0.11%	71.55%
3	500	0.21%	74.25%
4	500	0.16%	72.77%
5	500	0.13%	71.91%
6	500	0.11%	71.13%
7	500	0.10%	70.39%

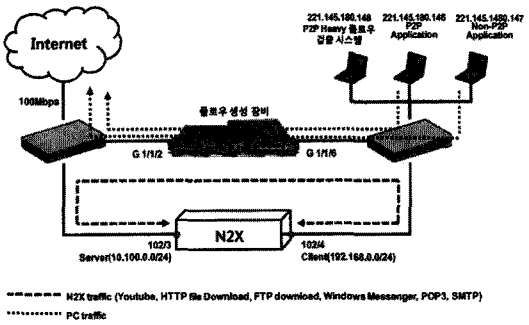

그림 10 검증을 위한 테스트 베드 환경

베드를 그림 9와 같이 구성하여 실험하였다. 5.1절에서 언급한 1Gbps 링크에서 캡쳐한 트래픽 데이터를 비교 검증하는 것이 올바르나 이를 위해서는 Payload 기반의 트래픽 분석 방법이 사용되어야 하고, 또한 Payload 기반의 트래픽 분석에 필요한 P2P 응용프로그램별 시그너쳐(Signature)가 확보되어야 한다. 하지만 P2P 응용프로별 시그너쳐의 확보가 불가능하여 그림 10과 같은 테스트 베드 환경에서 검증하였다.

그림 10의 테스트 베드 환경은 Agilent 사의 네트워크 트래픽 생성 및 성능 검증용 장비인 N2X 장비, 플로우 생성 장비, P2P 트래픽 생성 호스트, P2P 이외의 트래픽 생성 호스트, 그리고 P2P Heavy 플로우 검출 시스템으로 구성된다. Agilent사의 N2X 장비에서는 170

개의 호스트에서 Youtube, HTTP File Download, FTP Download, Windows Messager, POP3, SMTP등의 트래픽을 생성하도록 설정하였으며, N2X 트래픽 생성기 이외의 다른 두 개의 호스트에서 P2P와 P2P가 아닌 트래픽을 생성하였다. 그리고 플로우 생성 장비에서는 플로우를 생성하여 P2P Heavy 플로우 검출 시스템에 전송하도록 하였다.

5.3 P2P Heavy 플로우 검출 결과

본 장에서는 4장에서 제시한 알고리즘을 실제 트래픽에 적용하였을 경우 P2P Heavy 트래픽 검출 결과에 대해서 기술한다. 본 논문에서 제시한 알고리즘의 검증은 테스트 환경의 제약성으로 인해 실제 네트워크 망(가입자 망, 인터넷 백본 망)에서 발생하는 트래픽을 대상으로 하지 못하고, 테스트 베드에서 검증하였다. 검증 대상 P2P는 웹하드형 서비스를 제외한 국내의 푸르나, 싼타 25, 파일구리, 큐파일, 디비고, 프레앙, 오렌지파일, 토마토팡, 송사리, 엔피, 빅파일, 메가파일, 가제트, 파일팜, 겜플, 소리바다, 그리고 몽키 3 응용프로그램과 해외의 uTorrent를 검증하였다.

그 결과는 표 2에서 명시된 것처럼 토마토팡과 엔피를 제외한 모든 P2P 응용프로그램의 P2P Heavy 트래픽 플로우를 정확하게 검출하였다.

표 2에서 보는 바와 같이 토마토팡과 엔피 P2P도 모든 P2P Heavy 플로우를 검출하지 못한 것이 아니라 일부 플로우를 검출하지 못하였다. 그 이유는 토마토팡과 엔피 P2P인 경우 상대 Peer들에게서 컨텐츠를 다운로드하는 P2P Heavy 플로우에 해당하는 소스 스캔 플로우가 존재하지 않는 경우가 발생하였기 때문이다. 즉, 그림 6의 알고리즘 중에 Phase 2의 조건에 부합하지 않아서 검출하지 못하였다. 그리고, 겜플, 소리바다, 몽키 3 P2P인 경우는 음악 파일 다운로드 서비스로 Heavy 플로우가 검출되지 않았다. 따라서 겜플, 소리바다, 몽키 3 P2P는 네트워크 대역폭에 큰 영향을 주지 않는다는 결론을 내릴 수 있다.

표 2 P2P Heavy 트래픽 플로우 검출 결과

테스트 프로그램	Heavy 플로우 수	TP 플로우 수	TN 플로우 수	FP 플로우 수	FN 플로우 수
uTorrent	1320	80	1240	0	0
프루나	961	1	960	0	0
싼타25	946	17	929	0	0
파일구리	961	1	960	0	0
큐파일	903	14	889	0	0
디비고	967	2	965	0	0
프레앙	964	1	963	0	0
오렌지파일	961	1	960	0	0
토마토팡	1032	6	1025	0	1
송사리	960	1	959	0	0
엔피	986	20	959	0	7
빅파일	924	6	918	0	0
메가파일	950	3	947	0	0
가제트	878	37	841	0	0
파일팜	964	4	960	0	0
겜플, 소리바다, 몽키3	Heavy Flow 검출이 안됨				

TP : True Positive, TN : True Negative
FP : False Positive, FN : False Negative

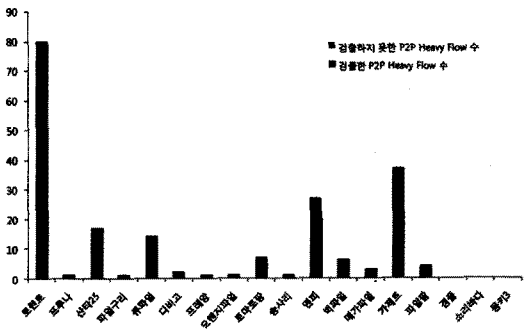

그림 11 P2P Heavy 플로우 검출 그래프

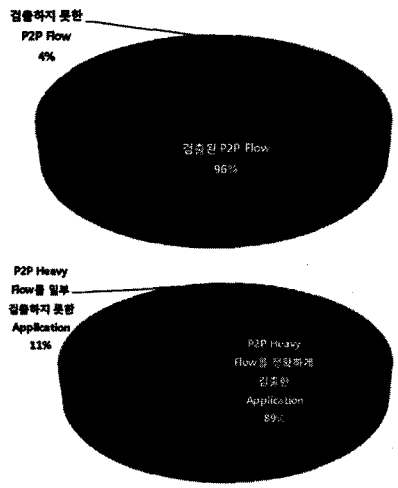

그림 12 P2P Heavy 플로우 검출율 그래프

표 2의 결과 중에 True Positive 플로우 개수와 False Positive 플로우 개수를 기준으로 막대 그래프로 나타내면 그림 11과 같다.

그림 12의 상단 그래프는 P2P Heavy 플로우의 총 개수 중에 검출된 P2P Heavy 플로우 개수를 기준으로 검출율을 나타낸 것이고, 하단 그래프는 분석한 P2P 응용프로그램의 총 개수 중에 검출된 P2P 응용프로그램의 개수를 기준으로 검출율을 나타낸 것이다. 그림 12에서 P2P Heavy 플로우를 하나라도 검출하지 못하면 미검출된 P2P 응용프로그램에 속하도록 하였다. 그림 12에서 보는 바와 같이 플로우 개수 기준으로 96%의 검출율을 보였으며, P2P 응용프로그램 개수 기준으로는 89%의 검출율을 보였다.

6. 결론 및 향후 연구

본 논문에서는 플로우 매개 변수 및 그들의 상관 관계를 나타내는 플로우 전달 특성과 플로우 패턴을 기반으로 P2P Heavy 트래픽 플로우를 검출하는 알고리즘

을 제시하였다. 이 알고리즘은 컨텐츠를 공유하는 P2P 응용 프로그램일 경우 항상 소스 스캔하는 플로우들의 집합과 그 중에 하나 이상의 Peer로부터 Heavy 트래픽 플로우가 발생한다는 플로우 패턴을 기반으로 도출하였으며, 검증 결과 국내 점유율이 높은 대부분의 P2P 응용프로그램에서 발생하는 P2P Heavy 트래픽을 검출하였음을 보여준다.

향후에 본 논문에서 제시하는 알고리즘의 개선을 위해서 실제 가입자 네트워크 망과 인터넷 백본 망에서 발생하는 트래픽으로 검증이 필요하며, 또한 본 논문에서 제시하는 알고리즘으로 검출되지 않은 토마토팡과 엔피 P2P와 같은 P2P 트래픽에 대한 추가적인 분석 및 알고리즘 개선이 필요할 것이다.

참 고 문 헌

[1] CoralReef. http://www.caida.org/tools/measurement/coralreef/

[2] T. Choi, C. Kim, S. Yoon, J. Park, B. Lee, H. Kim, and H. Chung, "Content-aware internet application traffic measurement and analysis," *IEEE/IFIP NOMS*, April 2004.

[3] P. Haffner, S. Sen, O.Spatscheck, and D.Wang, "Automataed construction of applicatin signatures," *ACM SIGCOMM MineNet Workshop*, August 2005.

[4] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Multilevel traffic classification in the dark," *ACM SIGCOMM*, August 2005.

[5] A. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," *PAM*, April 2005.

[6] S. Sen, O, Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," *WWW*, May 2004.

[7] M. Iliofotou, P. Pappu, M. Faloutsos, M. Mitzenmacher, S. Singh, and G. Varghese, "Network monitoring using traffic dispersion graphs," *ACM IMC*, October 2007.

[8] T. Karagiannis, A. Broido, M. Faloutsos, and kc claffy, "Transport layer identification of p2p traffic," *ACM IMC*, October 2004.

[9] T. Karagiannis, K. Papagiannaki, and M. Faloutsos, "Multilevel traffic classification in the dark," *ACM IMC*, August 2005.

[10] A. McGregor, M. Hall, P. Lorier, and J. Brunskill, "Flow clustering using machine learning techniques," *PAM*, April 2004.

[11] A. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," *ACM SIGMETRICS*, June 2005.

[12] M. Roughan, S. Sen, O. Spatscheck, and N. Duffield, "Class-of-service mapping for qos: a statistical signature-based approach to ip traffic classification," *ACM IMC*, October 2004.

[13] L. Bernaille, R.Teixeira, and K. Salamatian, "Early application identification," *ACM CoNEXT*, December 2006.

[14] J. Erman, M. Arlitt, and A. Mahanti, "Traffic Classification Using Clustering Algorithm," *ACM SIGCOMN MineNet Workshop*, September 2006.

[15] J. Erman, M. Arlitt, and A. Mahanti, "Traffic Classification Using Clustering Algorithm," *ACM SIGCOMN MineNet Workshop*, September 2006.

[16] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," *IEEE Transactions on Neural Networks*, vol.18, no.1, pp.223-239, January 2007.

[17] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," *ACM SIGCOMM CCR*, vol.37, no.1, pp.7-16, January 2007.

[18] S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," *IEEE LCN*, November 2005.

[19] J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/Realtime Traffic Classification Using Semi-Supervised Learning," *IFIP Performance*, October 2007.

[20] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," *ACM SIGCOMM CCR*, vol.36, no.5, pp.7-15, October 2006.

[21] Z. Li, R. Yuan, and X. Guan, "Accurate Classification of the Internet Traffic Based on the SVM Method," *ICC*, June 2007

[22] Cisco, White Pagers, "NetFlow Service and Application," http://www.cisco.com/warp/public/cc/pd/iosw/ioft/neflct/tech/napps_wp.htm

최 병 걸

1997년 2월 경북대학교 컴퓨터 공학과 졸업. 1999년 2월 경북대학교 컴퓨터공학과 석사. 2001년~2002년 ㈜블루코드 주임. 2003년~2008년 ㈜모비루스 팀장. 2009년~현재 경북대학교 컴퓨터공학과 박사과정. 관심분야는 Traffic Classification. P2P, 멀티미디어

이 시 영

1995년 2월 경북대학교 컴퓨터공학과 졸업. 1997년 2월 경북대학교 컴퓨터공학과 석사. 2003년~2008년 ㈜테크바일 연구소장. 2009년~현재 ㈜크레블 부사장. 관심분야는 Cloud computing, Network virtualization, Flow, Traffic Classification

서 영 일

1994년 2월 경북대학교 전자공학과 졸업
1996년 2월 경북대학교 전자공학과 석사
1996년~현재 KT 재직, KT 네트워크 연구소 부장. KT Premium 백본 구축, ITU-T IPTV FG의 Editor 및 IETF L2VPN design team등 표준 활동 수행
관심분야는 Network, VPN, P2P, CDN

위 즈 빈

2005년 7월 Harbin 공업대학 Thermal Energy and Engineering 졸업. 2009년 3월~현재 경북대학교 전자전기컴퓨터학부 석사과정. 관심분야는 멀티미디어, 동기식 이더넷, 기계학습

전 재 현

2009년 2월 대구가톨릭대학교 전자공학과 졸업. 2009년 3월~현재 경북대학교 전자전기컴퓨터학부 석사과정. 관심분야는 멀티미디어, 다시점 동영상, 동기식 이더넷 등

김 승 호

1981년 2월 경북대학교 전자공학과 졸업
1983년 2월 한국과학기술원 전산학과 석사. 1994년 2월 한국과학기술원 전산학과 박사. 1985년~현재 경북대학교 컴퓨터학부 정교수. 관심분야는 알고리즘, 멀티미디어, 다시점 동영상, 감시 시스템, 동기식 이더넷