



(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2012년01월11일
(11) 등록번호 10-1103741
(24) 등록일자 2012년01월02일

(51) Int. Cl.

H04L 12/26 (2006.01)

(21) 출원번호 10-2010-0067430

(22) 출원일자 2010년07월13일

심사청구일자 2010년07월13일

(56) 선행기술조사문헌

'Internet Application Traffic Classification Using Fixed IP-port', Proc. of the Asia-Pacific Network Operations and Management Symposium (APNOMS) 2009, pp. 21-30. 2009.09.*

'응용 레벨 트래픽 분류를 위한 시그니처 생성 시스템 및 검증 네트워크의 개발', 2009 제31회 정보처리학회 춘계학술발표대회, 제16권 제1호, pp. 1288-1291. 2009.04.*

*는 심사관에 의하여 인용된 문헌

(73) 특허권자

시큐아이닷컴 주식회사

서울특별시 강남구 테헤란로83길 15, 1호 ~9층 (삼성동, 제이엠빌딩)

고려대학교 산학협력단

서울특별시 성북구 종암로 1 (안암동5가, 고려대역)

(72) 발명자

김명섭

충청남도 연기군 조치원읍 신흥리 푸르지오1차 107-401

오영석

서울특별시 노원구 상계3동 불암대림아파트 202동 203호

(뒷면에 계속)

(74) 대리인

특허법인다나

전체 청구항 수 : 총 18 항

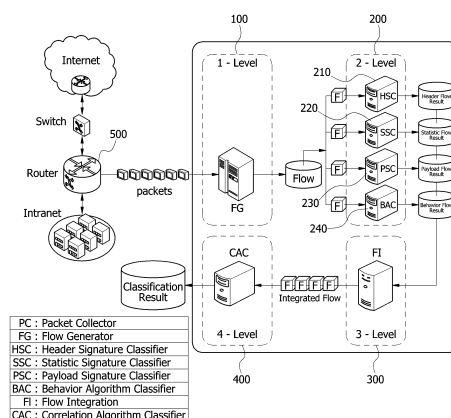
심사관 : 김대성

(54) 멀티 레벨 기반의 응용 트래픽 분석 시스템 및 방법

(57) 요약

본 발명에 따른 멀티 레벨 응용 트래픽 분석 시스템은 라우터로부터 패킷들을 수집하여 플로우 형태로 변환하는 플로우 생성부, 상기 플로우 생성부 후단에 위치하며 생성된 플로우를 입력 데이터로 하여 시그니처 매칭을 통하여 플로우의 응용을 결정하는 시그니처 기반 분류부, 상기 시그니처 분류부의 후단에 위치하며 응용이 결정된 플로우와 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 플로우 통합부, 및 통합된 플로우를 입력 데이터로 하여, 응용이 결정되지 아니한 미분류 플로우에 대하여 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 상관 관계 알고리즘 분류부를 포함하는 것을 특징으로 한다. 본 발명은 시그니처 기반 분석의 정확도를 활용하면서도 시그니처를 가지고 있지 않은 트래픽에 대해서도 플로우 간의 상관관계를 이용하여 추가적인 분석이 가능하게 함으로써 넓은 분석 범위(coverage)와 높은 분석률(Completeness) 및 정확도(Accuracy)를 갖는 트래픽의 응용 분류가 가능하다는 효과를 갖는다.

대표도 - 도1



(72) 발명자

박준상

전라남도 무안군 현경면 수양리 783

윤성호

경상북도 포항시 북구 용흥1동 현대타워 104동
1205호

박진완

충청남도 연기군 조치원읍 서창리 고려대학교세종
캠퍼스 과학기술대학 210-B

이상우

서울특별시 서초구 방배4동 878-24

이현신

충청남도 연기군 조치원읍 서창리 고려대학교세종
캠퍼스 과학기술대학 네트워크 관리연구실 210B

김경모

서울특별시 강남구 대치3동 949-11 포르쉐타워 5층

특허청구의 범위

청구항 1

라우터로부터 패킷들을 수집하여 플로우 형태로 변환하는 플로우 생성부;

상기 플로우 생성부 후단에 위치하며 생성된 플로우를 입력 데이터로 하여 시그니처 매칭을 통하여 플로우의 응용을 결정하는 시그니처 기반 분류부;

상기 시그니처 분류부의 후단에 위치하며 응용이 결정된 플로우와 시그니처를 가지고 있지 않아 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 플로우 통합부; 및

통합된 플로우를 입력 데이터로 하며, 응용이 결정되지 아니한 미분류 플로우에 대하여 선행적으로 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 상관 관계 알고리즘 분류부;를 포함하는 멀티 레벨 응용 트래픽 분석 시스템.

청구항 2

제1항에 있어서,

상기 상관 관계 알고리즘 분류부는 미분류 트래픽에 대하여 클라이언트가 통신한 서버가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 서버-클라이언트 기반 상관 관계 분석부; 설정 시간 그룹 내에 발생한 미분류 트래픽에 대하여 해당 시간 그룹 내에서 선행 분류에 의하여 응용이 확인된 다른 트래픽이 존재할 경우, 그 응용을 해당 응용으로 분류하는 발생 시간 기반 상관 관계 분석부; 미분류 트래픽에 대하여 양 통신 호스트가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 호스트-호스트 기반 상관 관계 분석부; 중의 적어도 하나 이상을 포함하는 멀티 레벨 응용 트래픽 분석 시스템.

청구항 3

제2항에 있어서,

상기 상관 관계 알고리즘 분류부는 상기 서버-클라이언트 기반 상관 관계 분석부, 상기 발생 시간 기반 상관 관계 분석부, 상기 호스트-호스트 기반 상관 관계 분석부를 순차적으로 적용하는 것인 멀티 레벨 응용 트래픽 분석 시스템.

청구항 4

제2항에 있어서,

상기 서버-클라이언트 기반 상관 관계 분석부는 IP 주소, 포트 번호, 통신 프로토콜의 3-튜플을 기준으로 통신한 서버를 확인하는 것인 멀티 레벨 응용 트래픽 분석 시스템.

청구항 5

제1항에 있어서,

상기 시그니처 기반 분류부는 특정 응용에서 사용하는 트래픽의 IP주소, 포트 번호 및 통신 프로토콜의 서버 3-튜플을 헤더 시그니처로 사용하는 헤더 시그니처 분류부; 플로우에 대하여 패킷 크기 및 방향을 기준으로 하는 시그니처를 이용하여 통계적으로 응용 프로그램을 구별하는 통계 시그니처 분류부; 플로우에 대하여 초기 M개(M은 자연수)의 패킷 페이로드에 존재하는 바이트의 시퀀스를 이용하여 시그니처 매칭을 수행하는 페이로드 시그니처 분류부 중의 적어도 하나 이상을 포함하는 것인 멀티 레벨 응용 트래픽 분석 시스템.

청구항 6

제5항에 있어서,

상기 헤더 시그니처 분류부는 플로우를 입력으로 하여 통신한 클라이언트 호스트의 개수(N_c), 사용시간(T_d) 및 체류시간(T_s)를 기준으로 고정 IP-포트를 자동으로 추출하는 헤더 시그니처 자동 추출 알고리즘에 의하여 추출

된 고정 IP-포트를 시그니처로 사용하는 것인 멀티 레벨 응용 트래픽 분석 시스템.

청구항 7

제5항에 있어서,

상기 통계 시그니처 분류부는 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 기준으로 하는 5-튜플이 동일한 플로우에 대하여 처음 N개(N은 자연수) 패킷의 방향과 페이로드 크기로 표시되는 N차원 벡터 대표값 및 그 벡터 대표값으로부터의 거리 임계값으로 시그니처 매칭을 수행하는 것인 멀티 레벨 응용 트래픽 분석 시스템.

청구항 8

제5항에 있어서,

상기 페이로드 시그니처 분류부는 HTTP 플로우에 대하여 파일 경로, URL 정보를 포함하는 스트링을 제외한 컨텐츠를 제거하는 HTTP 예비 필터링 단계, LCS(Longest Common Subsequence) 알고리즘에 의하여 후보 시그니처를 추출하는 단계, 및 후보 시그니처 중 최적의 시그니처를 선택하는 단계를 포함하여 선택된 시그니처를 사용하는 것인 멀티 레벨 응용 트래픽 분석 시스템.

청구항 9

- (a) 복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 수집하는 단계,
- (b) 상기 패킷들 중 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜의 5-튜플이 동일한 패킷을 그룹핑하여 플로우 데이터를 생성하고 응용이 확인된 플로우들에 대한 시그니처를 생성하는 단계,
- (c) 미분류 패킷을 수집하여 플로우를 생성하고 시그니처 매칭을 통하여 플로우의 응용을 결정하는 단계;
- (d) 이전 단계에서 응용이 결정된 플로우와 시그니처를 가지고 있지 않아 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 단계; 및
- (e) 이전 단계에서 통합된 플로우를 입력 데이터로 하며, 응용이 결정되지 아니한 미분류 플로우에 대하여 선행적으로 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 단계;를 포함하는 멀티 레벨 응용 트래픽 분석 방법.

청구항 10

제9항에 있어서,

단계(e)는

- (e-1) 미분류 트래픽에 대하여 클라이언트가 통신한 서버가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 단계;
- (e-2) 설정 시간 그룹 내에 발생한 미분류 트래픽에 대하여 해당 시간 그룹 내에서 선행 분류에 의하여 응용이 확인된 다른 트래픽이 존재할 경우, 그 응용을 해당 응용으로 분류하는 단계; 및
- (e-3) 미분류 트래픽에 대하여 양 통신 호스트가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 단계;를 적어도 하나 이상 포함하는 멀티 레벨 응용 트래픽 분석 방법.

청구항 11

제10항에 있어서,

단계(e)는 단계 (e-1), 단계 (e-2), 및 단계 (e-3)을 순차적으로 수행하는 멀티 레벨 응용 트래픽 분석 방법.

청구항 12

제10항에 있어서,

단계(e-1)은 IP 주소, 포트 번호, 통신 프로토콜의 3-튜플을 기준으로 통신한 서버를 확인하는 것인 멀티 레벨 응용 트래픽 분석 방법.

청구항 13

제9항에 있어서,

단계(b)는

특정 응용에서 사용하는 트래픽의 IP주소, 포트 번호 및 통신 프로토콜의 서버 3-튜플을 헤더 시그니처로 사용하거나;

플로우에 대하여 패킷 크기 및 방향을 기준으로 하는 시그니처를 사용하거나; 또는

플로우에 대하여 초기 M개(M은 자연수)의 패킷 페이로드에 존재하는 바이트의 시퀀스를 시그니처로 사용하는 것인 멀티 레벨 응용 트래픽 분석 방법.

청구항 14

제9항에 있어서,

단계(b)는 플로우를 입력으로 하여 통신한 클라이언트 호스트의 개수 임계값(N_c), 사용시간 임계값(T_d) 및 체류시간 임계값(T_s)를 기준으로 고정 IP-포트를 자동으로 추출하는 헤더 시그니처 자동 추출 알고리즘을 적용하는 것인 멀티 레벨 응용 트래픽 분석 방법.

청구항 15

제14항에 있어서,

응용을 알고 있는 서버의 3-tuple이 추출되면 3-튜플은 초기 상태로 되며, 통신한 클라이언트 호스트의 개수가 기준개수(N_c) 이상이면 후보 상태로 되고, 이러한 조건이 사용시간(T_d) 지속되면 시그니처로서 확정 상태로 되며, 각 상태에서 허용 체류시간이 경과되면 해당 플로우는 삭제되는 것인 멀티 레벨 응용 트래픽 분석 방법.

청구항 16

제9항에 있어서,

단계(b)는 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 기준으로 하는 5-튜플이 동일한 플로우에 대하여 처음 N개(N은 자연수) 패킷의 방향과 페이로드 크기로 표시되는 N차원 벡터 대표값을 시그니처로 포함하며,

단계(c)는 상기 벡터 대표값으로부터의 거리 임계값으로 시그니처 매칭을 수행하는 것인 멀티 레벨 응용 트래픽 분석 방법.

청구항 17

제16항에 있어서,

단계(c)는 플로우의 벡터 값이 알려진 응용의 벡터 대표값으로부터 설정된 시티 블록 거리 이내 일 때, 해당 응용으로 결정되는 것인 멀티 레벨 응용 트래픽 분석 방법.

청구항 18

제9항에 있어서,

단계(b)는 HTTP 플로우에 대하여 파일 경로, URL 정보를 포함하는 스트링을 제외한 콘텐츠를 제거하는 HTTP 예비 필터링 단계, LCS(Longest Common Subsequence) 알고리즘에 의하여 후보 시그니처를 추출하는 단계, 및 후보 시그니처 중 최적의 시그니처를 선택하는 단계를 포함하는 멀티 레벨 응용 트래픽 분석 방법.

명세서

기술분야

본 발명은 네트워크 트래픽을 효과적으로 분석하여 각 패킷을 응용 별로 분류하는 시스템 및 그 방법에 관한 것이다.

[0001]

배경 기술

- [0002] 최근 인터넷 사용자의 증가와 고속 네트워크의 보급으로 인해 네트워크 트래픽이 급증하였다. 이것은 단순히 WWW, FTP, e-mail 과 같은 정통적인 인터넷 서비스뿐만 아니라 멀티미디어 스트리밍, P2P(peer-to-peer) 파일 공유, 게임과 같은 다양한 멀티미디어 서비스의 대중화에 원인이 있다. 따라서 인터넷 트래픽이 급증함에 따라 효과적인 네트워크 관리를 위한 트래픽 모니터링 및 분석의 중요성이 커지고 있다.
- [0003] 네트워크 트래픽 분석은 네트워크 링크로부터 패킷을 수집하고 분석 기준에 맞게 각 패킷을 분류하는 일련의 과정이다. 이러한 분석 결과는 네트워크 관리와 제어의 관점에서 효과적으로 사용된다. 관리자는 네트워크 트래픽을 각각의 기준 별로 차단 및 제어를 할 수 있다.
- [0004] 트래픽 분석은 여러 기준에 의해 수행된다. 대표적인 예로는 응용레벨 프로토콜, 전송계층 프로토콜, 트래픽 타입, 응용 등이 있다. 많은 경우에 트래픽의 타입에 따라 트래픽을 분석하고 있다. 예를 들어, 통계적 특징과 행동유형이 유사한 msn, google talk 등과 같은 메신저들을 “CHAT” 이란 타입으로 통합 정의하여 분석한다. 그러나, 이러한 트래픽 타입의 기준을 실제 네트워크에 적용하였을 때 네트워크 관리 측면에서 많은 이점을 얻지 못한다. 이에 반해 실제 네트워크 관리에서 트래픽을 응용(Application Program) 별로 제어, 차단 하는 것은 많은 이점을 가진다. 예를 들어 응용 레벨 트래픽의 정확한 분류는 응용 별 종량제 과금, 응용기반 트래픽 제어, CRM(고객관계관리), SLA(Service Level Agreement) 지원, 응용 레벨 트래픽 보안 등에 직접적으로 활용이 가능하다.
- [0005] 종래에 트래픽의 응용별 분류를 위하여 시그니처 기반 분석 방법 등이 널리 제안되었고 어느 정도 양호한 분류 결과를 나타내고 있으나, 분석률과 정확도 측면에서 한계점은 여전히 존재한다. 따라서 종래에 제안된 시그니처 기반 분석 방법의 한계점을 극복하고 이를 보완하여 더욱 향상된 분석률과 정확도를 나타내는 응용 트래픽 분류 시스템 및 방법이 필요한 실정이다.

발명의 내용

해결하려는 과제

- [0006] 본 발명은 시그니처 기반 분석 방법에 의존한 트래픽 분류의 한계점을 보완하고 확충하여 더욱 향상된 분석률과 정확도를 나타내는 응용 트래픽 분류 시스템 및 방법을 제공하기 위한 것이다.

과제의 해결 수단

- [0007] 본 발명의 실시예에 따른 멀티 레벨 응용 트래픽 분석 시스템은 라우터로부터 패킷들을 수집하여 플로우 형태로 변환하는 플로우 생성부, 상기 플로우 생성부 후단에 위치하며 생성된 플로우를 입력 데이터로 하여 시그니처 매칭을 통하여 플로우의 응용을 결정하는 시그니처 기반 분류부, 상기 시그니처 분류부의 후단에 위치하며 응용이 결정된 플로우와 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 플로우 통합부, 및 통합된 플로우를 입력 데이터로 하며, 응용이 결정되지 아니한 미분류 플로우에 대하여 선행적으로 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 상관 관계 알고리즘 분류부를 포함하는 것을 특징으로 한다.
- [0008] 본 발명의 실시예에 따른 멀티 레벨 응용 트래픽 분석 방법은 (a) 복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 수집하는 단계, (b) 상기 패킷들 중 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜의 5-튜플이 동일한 패킷을 그룹핑하여 플로우 데이터를 생성하고 응용이 확인된 플로우들에 대한 시그니처를 생성하는 단계, (c) 미분류 패킷을 수집하여 플로우를 생성하고 시그니처 매칭을 통하여 플로우의 응용을 결정하는 단계; (d) 이전 단계에서 응용이 결정된 플로우와 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 단계; 및 (e) 이전 단계에서 통합된 플로우를 입력 데이터로 하며, 응용이 결정되지 아니한 미분류 플로우에 대하여 선행적으로 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 단계;를 포함하는 것을 특징으로 한다.

발명의 효과

[0009] 본 발명에 따른 멀티 레벨 응용 트래픽 분석 시스템 및 방법은 시그니처 기반 분석의 정확도를 충분히 활용하면서도 시그니처를 가지고 있지 않은 트래픽에 대해서도 플로우 간의 상관관계를 이용하여 추가적인 분석이 가능하게 함으로써 종래보다 넓은 분석 범위(coverage)와 높은 분석률(completeness) 및 정확도(accuracy)를 갖는 트래픽의 응용 분류가 가능하다는 효과를 갖는다. 또한 본 발명에서 제안된 시그니처 생성 방법은 계산처리의 부담을 경감하고 처리 속도를 높이는 장점을 갖는다.

도면의 간단한 설명

[0010] 도1은 본 발명의 일 실시예에 따른 멀티 레벨 응용 트래픽 분석 시스템의 구성도이다.

도2는 본 발명의 일 실시예에 따른 시그니처 추출 시스템의 구성도이다.

도3은 본 발명의 일 실시예에 따른 고정 IP-포트 상태 전이 다이어그램이다.

도4는 본 발명의 일 실시예에 따른 통계 시그니처 생성 방법의 다이어그램이다.

도5는 본 발명의 일 실시예에 따른 페이로드 시그니처 생성 방법의 다이어그램이다.

발명을 실시하기 위한 구체적인 내용

[0011] 이하에서는 본 발명에 따른 구체적인 실시예가 설명된다. 그러나 본 발명은 여러 가지 다양한 형태로 구현될 수 있으며 여기에서 설명하는 실시예에 한정되지 않는다. 본 발명에 첨부된 도면은 설명의 편의를 위한 것으로서 상대적인 척도는 변경될 수 있으며, 본 발명을 명확하게 설명하기 위해서 설명과 관계없는 부분은 생략되었다.

[0012] 이하에서, 어떤 부분이 어떤 구성요소를 "포함"한다고 할 때, 이는 특별히 반대되는 기재가 없는 한 다른 구성요소를 제외하는 것이 아니라 다른 구성요소를 더 포함할 수 있는 것을 의미한다. 또한, 명세서에 기재된 "--부", "--기", "--모듈" 등의 용어는 적어도 하나의 기능이나 동작을 처리하는 단위를 의미하며, 이는 하드웨어나 소프트웨어 또는 하드웨어 및 소프트웨어의 결합으로 구현될 수 있다.

[0013] 도1은 본 발명에 따른 멀티 레벨 응용 트래픽 분석 시스템의 전체적인 구성도이다. 이 시스템은 라우터(500)로부터 패킷들을 수집하여 플로우 형태로 변환하는 플로우 생성부(FG, 100), 상기 플로우 생성부(100) 후단에 위치하며 생성된 플로우를 입력 데이터로 하여 시그니처 매칭을 통하여 플로우의 응용을 결정하는 시그니처 기반 분류부(200), 상기 시그니처 분류부(200)의 후단에 위치하며 응용이 결정된 플로우와 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 플로우 통합부(FI, 300), 및 통합된 플로우를 입력 데이터로 하여, 응용이 결정되지 아니한 미분류 플로우에 대하여 선행적으로 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 상관 관계 알고리즘 분류부(CAC, 400)를 포함한다.

[0014] 상기 시그니처 기반 분류부는 특정 응용에서 사용하는 트래픽의 IP주소, 포트 번호 및 통신 프로토콜의 서버 3-튜플을 헤더 시그니처로 사용하는 헤더 시그니처 분류부(HSC, 210), 플로우에 대하여 패킷 크기 및 방향을 기준으로 하는 시그니처를 이용하여 통계적으로 응용 프로그램을 구별하는 통계 시그니처 분류부(SSC, 220), 플로우에 대하여 초기 M개(여기서 M은 자연수)의 패킷 페이로드에 존재하는 바이트의 시퀀스를 이용하여 시그니처 매칭을 수행하는 페이로드 시그니처 분류부(PSC, 230)를 포함한다.

[0015] 상기 통계 시그니처 분류부(220)를 좀더 상세히 설명하면 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 기준으로 하는 5-튜플이 동일한 플로우에 대하여 처음 N개(N은 자연수) 패킷의 방향과 페이로드 크기로 표시되는 N차원 벡터 대표값 및 그 벡터 대표값으로부터의 거리 임계값으로 시그니처 매칭을 수행한다. 플로우의 벡터 값이 알려진 응용의 벡터 대표값으로부터 설정된 시티 블록 거리 이내 일 때, 해당 응용으로 결정된다. 시티 블록거리(city-block distance)란 두 벡터간의 거리를 나타내는 유클리디안 거리에 대체되는 개념으로서 각 차원 요소값 차이의 절대값의 합으로 정의된다. 이는 많은 양의 데이터를 처리해야 하는 트래픽 분석에서 처리 속도가 빠른 장점을 제공한다.

[0016] 그 외에 추가적으로 Skype 등 특정 응용프로그램에 대하여 동작(Behavior) 형태에 기반한 알고리즘을 통하여 플로우를 분석하는 동작 알고리즘 분류부(BAC, 240)를 시그니처 기반 분류부(200)와 통합하여 또는 시리얼적으로 부가하는 것도 가능하나, 이하에서는 동작 알고리즘 분류부(BAC)가 사용되지 아니하는 실시예를 기준으로 설명한다.

[0017] 상기 시스템의 동작을 전체적으로 설명하면, 첫 번째 단계(1-level)에서는 라우터(500)에서 수집된 모든 패킷들

은 FG(100)에 의해 플로우(Flow) 형태로 변환된다. 생성된 모든 플로우들은 두 번째 단계(2-level)에서 각 시그니처 분류부(210, 220, 230)의 입력 데이터로 사용된다. 각 분류부는 매 분마다 분류되지 않은 플로우를 동일하게 전달받는다. 패킷이 들어오면 해당 패킷의 플로우가 생성되어 있는지 확인하고 업데이트를 한다. 이미 확인된 플로우라면 다음 패킷을 검사하고 응용이 결정되지 않은 플로우이면 세 가지 시그니처 기반 분류부(210, 220, 230)의 시그니처 리스트와 입력된 플로우를 비교함으로써 분석한다. 세 가지 시그니처 리스트 중 하나라도 매칭이 되면 그 플로우는 해당 응용의 플로우로 결정되고, 세 가지 리스트 모두 매칭이 되지 않으면 어떤 응용의 트래픽인지 결정을 할 수 없으므로 다음 패킷을 검사한다. 분류가 확인된 플로우나 분류가 확인되지 아니한 플로우들은 세 번째 단계(3-level)에서 플로우 통합부(FI, 300)에 의해 통합된다. 마지막 네 번째 단계(4-level)에서는 상관 관계 알고리즘 분류부(CAC, 400)가 앞선 단계(3-level)에서 통합된 플로우를 입력으로 하여 두 번째 단계(2-level)에서 분석되지 않은 플로우들에 대하여 상관관계 알고리즘을 이용하여 추가적인 분류를 수행한다.

[0018] 상기 상관 관계 알고리즘 분류부(400)는 미분류 트래픽에 대하여 클라이언트가 통신한 서버가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 서버-클라이언트(Serve-Client) 기반 상관 관계 분석부; 설정 시간 그룹 내에 발생한 미분류 트래픽에 대하여 해당 시간 그룹 내에서 선행 분류에 의하여 응용이 확인된 다른 트래픽이 존재할 경우, 그 응용을 해당 응용으로 분류하는 발생 시간 기반(Time-based) 상관 관계 분석부; 미분류 트래픽에 대하여 양 통신 호스트가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 호스트-호스트(Host-Host) 기반 상관 관계 분석부; 중의 적어도 하나 이상을 포함한다.

[0019] 이하에서는 본 발명의 멀티 레벨 기반 분석 시스템의 상관 관계(Correlation) 기반 분석 알고리즘을 설명한다.

[0020] 트래픽 상관관계 기반 분석은 인터넷 트래픽의 3-tuple 레벨 주소 체계(IP address, port number, transport protocol)와 트래픽 발생 시점, 발생 형태 등의 특성을 바탕으로 트래픽 플로우들 사이에 연관성을 가중치로 표현하고 가중치의 임계값을 적용하여 트래픽을 응용 프로그램 별로 구분하는 방법이다. 이 방법은 우선 시그니처 기반 분석 방법으로 트래픽을 정확하게 분석하고 플로우 사이의 다양한 연관성을 찾아 이를 바탕으로 분석물을 향상 시키는 것이다.

[0021] 서버-클라이언트(Server-Client) 기반 상관관계 알고리즘은 하나의 서버 3-tuple(IP address, port number, transport protocol)을 동시에 하나의 응용만이 사용한다는 특성을 바탕으로 한다. 만약 선행 시그니처 기반 분석 방법으로 분석된 트래픽 중 서버를 포함하는 분석 결과를 가지고 있다면, 해당 서버의 3-tuple을 해당 응용이 사용한 것으로 간주하고 분석 대상이 되는 네트워크에 해당 서버와 통신하는 트래픽을 해당 응용으로 분석한다.

[0022] 발생시간(Time-based) 기반 상관관계 알고리즘은 한 호스트에 한해 일정 기간 안에 발생하는 트래픽은 같은 응용일 가능성이 높다는 특성을 바탕으로 한다. 선행 방법으로 분석되지 않는 트래픽들을 호스트와 짧은 시간 간격으로 그룹핑한 후, 해당 그룹 중 분석된 트래픽이 있을 경우 그 그룹에 속한 모든 트래픽을 해당 응용으로 분석하는 방법이다.

[0023] 호스트-호스트(Host-Host) 기반 상관관계 알고리즘은 두 호스트 간에 일어나는 통신은 하나의 응용에 의해 발생될 가능성이 높다는 특성을 바탕으로 한다. 즉, 호스트 간의 통신에서 일 부분만 선행 방법으로 분석이 되었다면 분석되지 않은 트래픽을 이미 분석된 트래픽의 응용으로 분석하는 것이다. 데이터를 주고 받는 통신일 경우 제어를 위한 통신과 실제 데이터를 전송하기 위한 통신이 서로 다른 포트에서 이루어 진다. 따라서 이러한 형태를 보이는 호스트들 사이의 트래픽을 호스트-호스트 상관관계를 이용하여 효과적으로 분석할 수 있다.

[0024] 상기 상관 관계 알고리즘 분류부(400)는 상기 서버-클라이언트 기반 상관 관계 분석부, 상기 발생 시간 기반 상관 관계 분석부, 상기 호스트-호스트 기반 상관 관계 분석부를 우선순위를 두어 순차적으로 적용하는 것이 바람직하다. 이는 정확도가 가장 큰 Server-Client기반 상관관계 방법론의 우선순위를 높게 부여하여, 선행 분류 단계에서 분류되지 못한 미분류된 트래픽 만을 분석 대상으로 하는 추가적인 분류의 정확도를 극대화 하기 위한 것이다. 따라서 위와 같은 플로우 상관관계를 이용하면 시그니처를 가지고 있지 않은 트래픽에 대해서도 응용 분류가 가능하므로 종래보다 넓은 분석 범위(coverage)와 높은 분석률(Completeness) 및 정확도(Accuracy)를 갖는 트래픽의 응용 분류가 가능하다.

[0025] 이하에서는 도2를 이용하여 본 발명에 사용된 시그니처 추출 시스템에 대해 설명한다. 패킷 수집기(PC, 610)는 특정 네트워크에서 인터넷으로 향하는 모든 트래픽을 수집한다. 플로우 생성기(FG, 620)는 수집한 패킷을 이용하여 플로우(630)를 생성한다. 플로우는 통상 소스 IP(source IP), 목적지 IP(destination IP), 소스 포트

(source port), 목적지 포트(destination port) 및 전송 계층 프로토콜(protocol)의 동일한 5-튜플을 갖는 패킷의 집합으로 정의될 수 있다. 트래픽 측정 서버(TMS, 650)는 네트워크 내의 트래픽 측정 에이전트(TMA, 640)가 설치된 여러 단말 호스트들로부터 TMA 로그를 수집하고 정답지 생성기(GTG: Ground Truth Generator, 660)에게 전달한다. GTG는 앞서 FG가 생성한 플로우와 TMS로부터 전달받은 TMA 로그 정보를 비교하여 정답지 트래픽 데이터(Ground Truth Traffic Data, 670)를 생성한다. 이 정답지(Ground Truth) 정보를 바탕으로 헤더 시그니처 생성기(HSG, 681), 통계 시그니처 생성기(SSG, 682), 페이로드 시그니처 생성기(PSG, 683)는 각각의 시그니처를 생성한다. 위의 각 시그니처 생성기는 헤더 시그니처(691), 통계 시그니처(692), 페이로드 시그니처(693)를 생성하고, 이들은 앞서 설명된 헤더 시그니처 분류부(HSC, 210), 통계 시그니처 분류부(SSC, 220), 페이로드 시그니처 분류부(PSC, 230)의 시그니처 리스트로 제공된다.

[0026] 상기 헤더 시그니처 생성기(681)는 플로우를 입력으로 하여 통신한 클라이언트 호스트의 개수(Nc: Number of clients), 사용시간(Td: Duration) 및 체류시간(Ts: Time allowed to stay)을 기준으로 고정 IP-포트를 자동으로 추출하는 헤더 시그니처 자동 추출 알고리즘에 의하여 고정 IP-포트를 시그니처로 추출한다.

[0027] Header 시그니처는 특정 응용을 사용할 때에만 나타나는 트래픽의 서버 3-tuple(IP address, port number, transport protocol)을 의미한다. 즉, 많은 인터넷 응용들은 실제 서비스를 제공하기 전에 사용자 인증 과정을 거치게 되는데 이러한 과정을 특정 로그인 서버가 담당한다. 따라서 로그인 서버의 3-tuple이 해당 응용의 Header 시그니처 중 하나가 될 수 있다. 여기서 트래픽에 대하여 고정 IP-포트의 구분만으로 거의 완벽하게 TCP 또는 UDP 등으로 구분되므로 전송 프로토콜은 헤더 시그니처에서 생략 가능할 수도 있다.

[0028] 도3은 헤더 시그니처 자동 추출 알고리즘의 상태 전이 다이어그램이다. 최초 모든 플로우의 3-tuple은 해제(Discard) 상태에 속한다. 만약 발생시킨 응용을 알고 있는 서버의 3-tuple이 추출 시스템에 입력 데이터로 들어오면 해당 3-tuple은 초기(New) 상태로 이동한다. 이후, 통신한 클라이언트 호스트의 개수가 Nc 이상이면 후보(Candidate) 상태로 되고, 이러한 조건이 사용시간(Td) 동안 지속되면 확정(Fixed) 상태로 된다. 또한 각 상태에서 허용 체류시간(Ts)이 초과되면 해당 플로우는 메모리로부터 삭제되어, 오버헤드를 경감할 수 있게 한다. 여기서 각 상태에 대하여 Ts의 구체적 값은 서로 달리 설정될 수 있다. 확정(Fixed) 상태에 속한 3-tuple들이 고정 IP-port가 되고 충돌(conflict) 여부를 검사하여 Header 시그니처로 결정된다.

[0029] 상기 통계 시그니처 생성부(682)는 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 기준으로 하는 5-튜플이 동일한 다수의 플로우에 대하여 처음 N개(N은 자연수) 패킷의 방향과 페이로드 크기로 표시되는 N차원 벡터를 통계적으로 처리하여 그 대표값을 시그니처로 생성한다. 방향은 양수와 음수로 표현되며, TCP의 경우 양수는 클라이언트에서 서버로 향하는 패킷, 음수는 서버에서 클라이언트로 향하는 패킷을 의미한다. UDP는 서버/클라이언트의 구분이 명확하지 않기 때문에, 양수/음수의 의미는 단지 방향이 서로 반대라는 것만 표현할 수 있다. 따라서, UDP의 경우에는 첫 패킷을 양수로 표현하고 뒤에 이어지는 패킷은 첫 패킷을 기준으로 방향이 같으면 양수, 다르면 음수로 표현한다.

[0030] 도4는 통계 시그니처 생성 과정을 보여준다. 먼저 네트워크 GTG(660)으로부터 전달된 GT(Ground Truth) 트래픽을 입력으로 하여 수집된 트래픽을 플로우 단위의 그룹핑을 하기 위해 전처리(preprocessing) 작업을 거친다. 전처리가 끝난 플로우들은 페이로드의 크기와 방향을 이용하여 그룹핑(Grouping)을 하고 그룹핑 된 그룹들은 최적화(Optimization) 작업을 거치고 각 그룹들 사이의 충돌을 해결함으로써 통계 시그니처를 생성한다. 이 때 그룹 최적화 및 충돌처리를 수행하여 해당 응용으로 분류될 수 있는 벡터 대표값으로부터의 거리 임계값을 결정하여 시그니처 매칭시 사용한다.

[0031] 도5를 참조하면 상기 상기 페이로드 시그니처 생성부(683)는 HTTP 플로우에 대하여 파일 경로, URL 정보를 포함하는 스트링을 제외한 콘텐츠를 제거하는 HTTP 예비 필터링 단계, LCS(Longest Common Subsequence) 알고리즘에 의하여 후보 시그니처를 추출하는 단계, 및 후보 시그니처 중 최적의 시그니처를 선택하는 단계를 수행하여 시그니처를 선택한다.

[0032] HTTP 예비 필터링(Pre-filtering) 단계는 HTTP플로우에 대한 시그니처의 정확성을 높이고, 추출 시스템의 부하를 줄이기 위한 단계이다. 시그니처 추출 결과 HTTP를 사용하는 응용프로그램의 시그니처는 요청하는 파일에 대한 Path정보, 서버의 URL 정보에서 추출된다. 이 단계에서는 파일 경로, URL 정보를 포함하는 스트링을 제외한 콘텐츠를 제거하는 역할을 수행한다.

[0033] 후보 시그니처를 추출(Extract Candidate Signature) 단계는 LCS알고리즘을 기반으로 가능한 모든 후보 시그니처를 추출하는 단계이다. 기본적인 LCS에 의해서 추출되는 시그니처는 하나의 결과로 나타나지만, 실제 LCS에

의해서 추출 가능한 결과는 다양한 경우의 수로 나타난다. 따라서 추출 가능한 시그니처에 대한 모든 경우의 수를 고려해야 정확한 응용프로그램 시그니처의 추출이 가능하다.

[0034] 최적의 시그니처 선택(Best Signature Selection) 단계는 후보 시그니처에서 최적의 시그니처를 찾는 과정이다. 이 단계는 두 가지 구성된다. 첫 번째 단계는 후보 시그니처로부터 최적의 시그니처를 추출하는 과정으로서 길이가 가장 긴 substring이 존재하는 경우와 페이로드 시작 부분을 포함하는 경우를 선택하여 최적의 시그니처를 선택한다. 두 번째 단계는 첫 번째 단계에서 선택된 시그니처에서 시그니처로서 변별력을 갖지 못하는 substring을 제거하는 과정이다. 이때 적용되는 규칙은 Substring의 최소 길이는 2바이트로 제한하고, Common substring 제거하는 규칙을 적용한다. 시그니처가 결정되면 기존의 유효한 시그니처 목록과 충돌 관계를 검사하여 최종적으로 시그니처 기반 분석 시스템의 입력데이터로 제공된다.

[0035] 시그니처 생성과 트래픽 분류를 전체적으로 통합한 멀티 레벨 응용 트래픽 분석 방법은 (a) 복수의 클라이언트 호스트와 복수의 서버 호스트 사이의 패킷들을 수집하는 단계, (b) 상기 패킷들 중 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜의 5-튜플이 동일한 패킷을 그룹핑하여 플로우 데이터를 생성하고 응용이 확인된 플로우들에 대한 시그니처를 생성하는 단계, (c) 미분류 패킷을 수집하여 플로우를 생성하고 시그니처 매칭을 통하여 플로우의 응용을 결정하는 단계; (d) 이전 단계에서 응용이 결정된 플로우와 응용이 결정되지 아니한 플로우를 시계열적으로 통합하는 단계; (e) 이전 단계에서 통합된 플로우를 입력 데이터로 하며, 응용이 결정되지 아니한 미분류 플로우에 대하여 선행적으로 응용이 결정된 플로우와의 상관 관계를 이용하여 해당 응용을 추정하는 단계;를 포함하여 구성된다.

[0036] 상기 단계(e)는 상관 관계 알고리즘을 이용하는 분류단계로서 (e-1) 미분류 트래픽에 대하여 클라이언트가 통신한 서버가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 단계; (e-2) 설정 시간 그룹 내에 발생한 미분류 트래픽에 대하여 해당 시간 그룹 내에서 선행 분류에 의하여 응용이 확인된 다른 트래픽이 존재할 경우, 그 응용을 해당 응용으로 분류하는 단계; 및 (e-3) 미분류 트래픽에 대하여 양 통신 호스트가 동일한 선행 분류된 트래픽이 존재하는 경우, 그 응용을 해당 응용으로 분류하는 단계;를 적어도 하나 이상 포함한다. 더욱 바람직하게는 단계 (e-1), 단계 (e-2), 및 단계 (e-3)을 순차적으로 수행함으로써 보다 우선 순위가 높은 상관 관계를 먼저 수행한다.

[0037] 클라이언트가 통신한 서버를 확인하는 단계는 IP 주소, 포트 번호, 통신 프로토콜의 3-튜플을 기준으로 하는 것이 통상적이다.

[0038] 상기 단계(b)는 플로우들에 대한 시그니처를 생성하는 단계로서 특정 응용에서 사용하는 트래픽의 IP주소, 포트 번호 및 통신 프로토콜의 서버 3-튜플을 헤더 시그니처로 사용하거나; 플로우에 대하여 패킷 크기 및 방향을 기준으로 하는 통계 시그니처를 사용하거나; 플로우에 대하여 초기 M개(M은 자연수)의 패킷 페이로드에 존재하는 바이트의 시퀀스를 페이로드 시그니처로 사용할 수 있음은 앞서 설명한 바와 동일하다.

[0039] 헤더 시그니처 생성은 플로우를 입력으로 하여 통신한 클라이언트 호스트의 개수 임계값(N_c), 사용시간 임계값(T_d) 및 체류시간 임계값(T_s)를 기준으로 고정 IP-포트를 자동으로 추출하는 헤더 시그니처 자동 추출 알고리즘을 적용할 수 있다. 이때, 응용을 알고 있는 서버의 3-tuple이 추출되면 3-튜플은 초기 상태로 되며, 통신한 클라이언트 호스트의 개수가 기준개수(N_c) 이상이면 후보 상태로 되고, 이러한 조건이 사용시간(T_d) 지속되면 시그니처로서 확정 상태로 되며, 각 상태에서 허용 체류시간(T_s)이 경과되면 해당 플로는 삭제되는 방법이 사용된다. 물론 각 상태에 대하여 T_s 의 구체적 값은 서로 달리 설정될 수 있다.

[0040] 통계 시그니처 생성은 소스 IP, 목적지 IP, 소스 포트, 목적지 포트, 프로토콜을 기준으로 하는 5-튜플이 동일한 플로우에 대하여 처음 N개(N은 자연수) 패킷의 방향과 페이로드 크기로 표시되는 N차원 벡터 대표값을 시그니처로 사용하며, 상기 벡터 대표값으로부터의 거리 임계값이 결정된다. 이때, 플로우의 벡터 값이 알려진 응용의 벡터 대표값으로부터 설정된 시티 블록 거리 이내 일 때, 해당 응용으로 결정되는 방법이 사용된다.

[0041] 페이로드 시그니처 생성은 HTTP 플로우에 대하여 파일 경로, URL 정보를 포함하는 스트링을 제외한 콘텐츠를 제거하는 HTTP 예비 필터링 단계, LCS(Longest Common Subsequence) 알고리즘에 의하여 후보 시그니처를 추출하는 단계, 및 후보 시그니처 중 최적의 시그니처를 선택하는 단계를 포함한다.

[0042] 위에서 개시된 발명은 기본적인 사상을 훼손하지 않는 범위 내에서 다양한 변형예가 가능하다. 따라서, 앞에서 개시된 실시예들은 모두 예시적으로 해석되어야 하며, 한정적으로 해석되지 않는다. 따라서 본 발명의 보호범위는 상술한 일 실시예가 아니라 첨부된 청구항에 따라 정해져야 한다. 첨부된 청구항의 균등물로의 치환은 첨부된 청구항의 보호범위에 속하는 것이다.

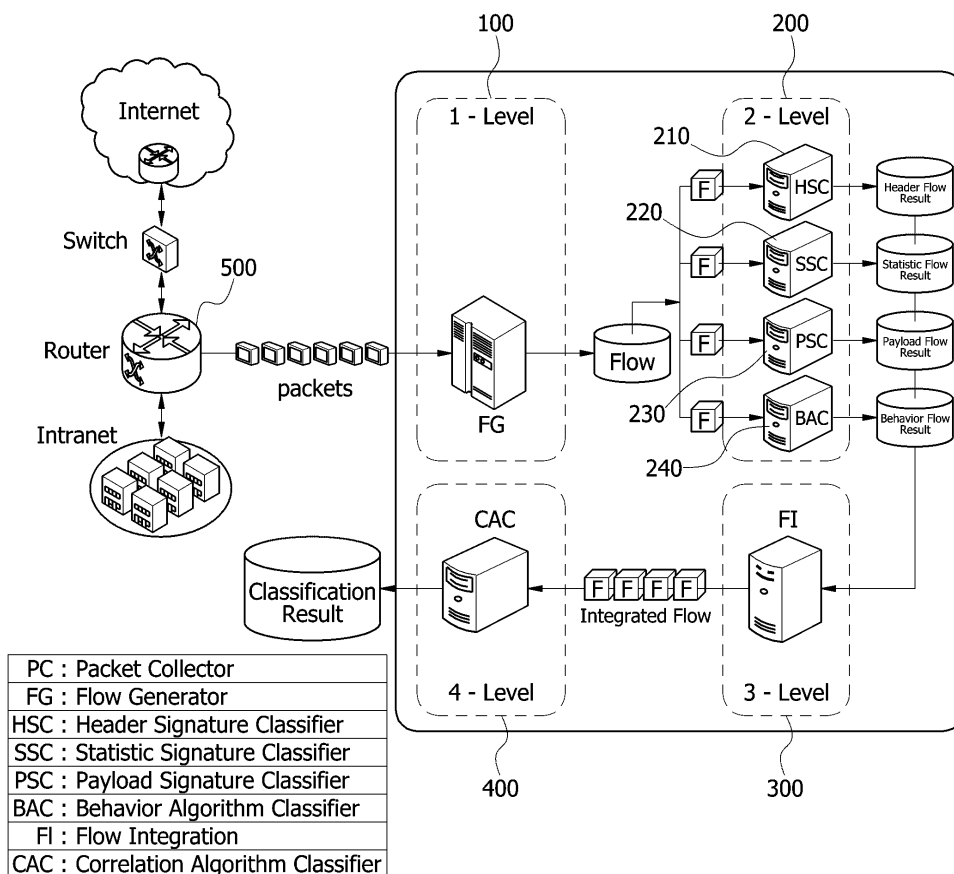
부호의 설명

[0043]

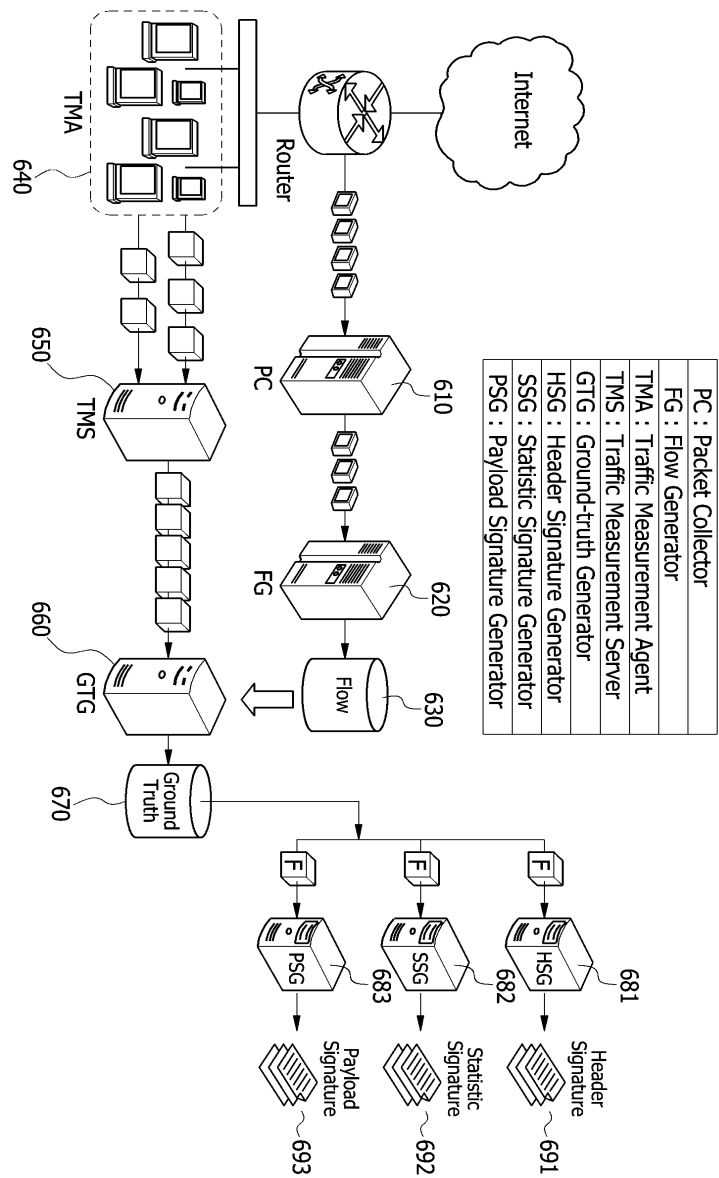
100: 플로우 생성부	200: 시그니처 기반 분류부
300: 플로우 통합부	400: 상관 관계 알고리즘 분류부
500: 라우터	210: 헤더 시그니처 분류부(HSC)
220: 통계 시그니처 분류부(SSC)	230: 페이로드 시그니처 분류부(PSC)
610: 패킷 수집기	620: 플로우 생성기
630: 플로우	640: 트래픽 측정 에이전트(TMA)
650: 트래픽 측정 서버(TMS)	660: 정답지 생성기(GTG)
670: 정답지	681: 헤드 시그니처 생성기(HSG)
682: 통계 시그니처 생성기(SSG)	683: 페이로드 시그니처 생성기(PSG)
691: 헤더 시그니처	692: 통계 시그니처
693: 페이로드 시그니처	

도면

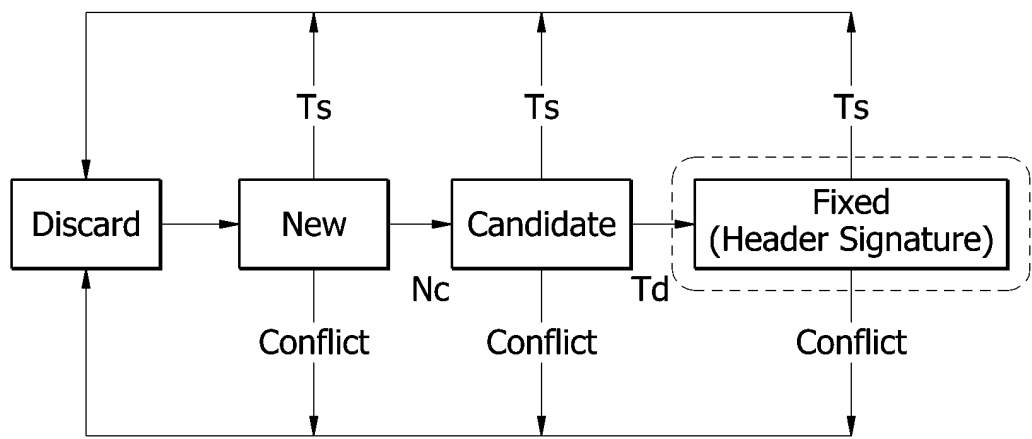
도면1



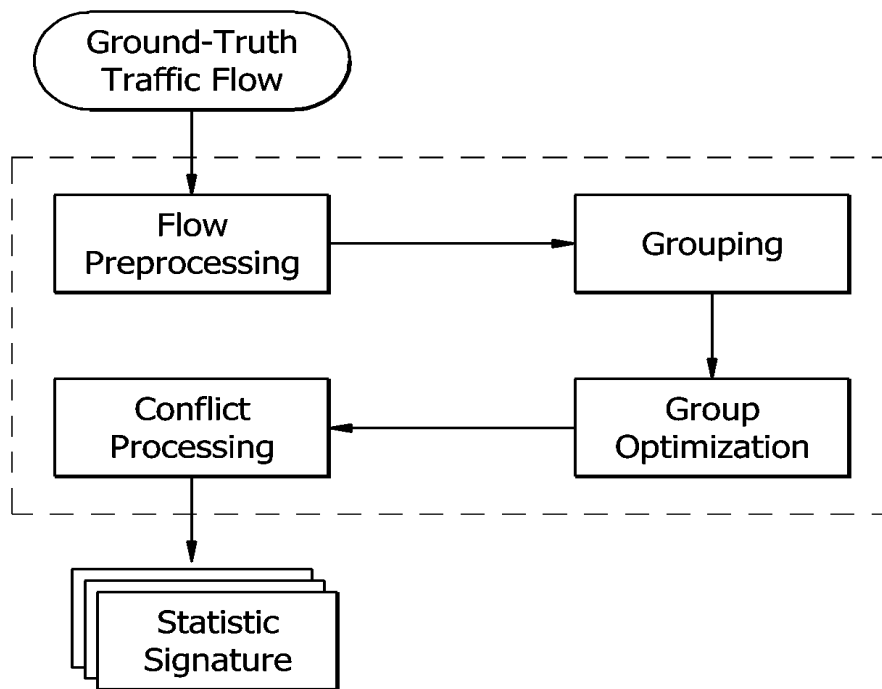
도면2



도면3



도면4



도면5

