

통계적 분석을 이용한 HTTP 트래픽 모델링 및 분석[☆]

HTTP Traffic Modeling and Analysis with Statistical Process

전 의 수* 이 광 휘**
Uie-Soo Jun Kwang-Hui Lee

요 약

통신망을 효율적으로 설계하고 운영하기 위하여 통신망에 대한 구체적인 시뮬레이션이 필요하며 이에 관한 연구가 현재 활발히 이루어지고 있다. 본 논문에서는 시뮬레이션 시 요구되는 트래픽 모델을 수립하기 위하여 실제 수집된 트래픽 자료를 이용하여 HTTP 요구 수준에서 웹 트래픽을 모델링 한다. 그리고 또한 현재 논란이 되고 있는 인터넷 트래픽의 통계적인 특성이 포아송 과정을 따르는지 아니면 자기유사한(self-similar) 특성을 보이는지를 웹 트래픽의 특성 분석을 통해 알아보았다. 본 연구의 결과로 트래픽은 하나의 함수로 모델링 할 수 없었으며 특정 구간에서는 하나의 함수로, 그리고 다른 구간에서는 다른 함수로 모델링 될 수 있다는 사실을 알 수 있었으며, 그 모델링이 다르게 되는 영역은 HTTP 서버의 특성에 따라서 달라진다는 것을 알 수 있었다. 이러한 결과를 본 연구와 관련된 통신망 시뮬레이터(NetDAS)의 웹 트래픽 생성기의 설계 및 구현에 이용할 수 있도록 하였다.

Abstract

For efficient design and operation of a communication network, precise simulation of network characteristics is essential. This issue has been the focus of research by several groups. In this study, we first modeled the HTTP traffic which would be employed on simulation on the level of application using the real collected traffic data.

There are two different viewpoints on the characteristics of web traffic pattern, Poisson distribution and self-similar characteristics. In our study, the results show that web traffic characteristics do not depend on only one type of distribution, but the traffic can be modeled as composition of these depending on the size of response of Web server. This implicates that the web traffic can be modeled as the combination of two characteristics. We also found that the characteristics of Web traffic rely on the properties of web servers. This result was deployed as a traffic generator in implementing the network simulator (NetDAS).

· Keyword : HTTP, Traffic Engineering, Network Simulator,

1. 서 론

인터넷 사용의 급격한 증가와 다양한 서비스의 출현으로 인하여 데이터 트래픽의 양상이 점점 복잡해지고 있다. 따라서 기존의 경험적 지식만을 이용하여 통신망의 사용량이나 이용자 패턴을 조사하여 통신망을 설계하고 운영하는 것에는 한계가 있다. 이러한 관점에서 현재 트래픽 흐름

에 대한 통계적 분석과 더불어 시뮬레이션과 관련된 연구가 활발히 진행 중이다.

통신망 시뮬레이션 시 요구되는 목적을 위해서는 목적에 맞는 정확한 시스템의 구성과 트래픽 부하의 생성을 위한 모델이 가장 먼저 고려되어야 할 것이다. 따라서 본 논문은 통신망의 성능분석을 위한 도구인 NetDAS[1]에서 응용 트래픽 부하의 생성을 위한 모델에 관한 연구의 일부분으로써 트래픽 모델을 정의한다. NetDAS는 통신망의 망 수준 및 응용 수준의 트래픽 생성 및 시뮬레이션을 위하여 본 연구실에서 개발한 시스템이다. 인터넷 트래픽 부하를 생성하기 위해서는 먼저 각 응용별로 실제 자료 분석을 통해

* 정 회 원 : 창원대학교 전보전산원 연구원
fraser@cdcs.changwon.ac.kr(제 1저자)

** 정 회 원 : 창원대학교 컴퓨터공학과 교수
khlee@sarim.changwon.ac.kr(공동저자)

☆ 본 논문은 한국과학재단 산학협동연구(과제번호: 101-2001-000-00061-0)로 수행된 결과의 일부임.

가장 적합하게 모델링 되어지는 확률 분포와 파라미터를 추출함으로써 통계적인 방법으로 트래픽 부하를 생성할 수 있다. 인터넷 트래픽에는 여러 종류의 응용 서비스에 의해 발생하는 트래픽이 있지만 본 논문에서는 전체 인터넷 트래픽 중 HTTP 트래픽이 차지하는 비중이 높음으로써 웹(Web) 트래픽을 대상으로 모델링 하였다.

트래픽 모델링은 어떤 계층에서 하느냐에 따라 모델링 방법 및 분석 항목이 달라진다. 현재 웹 트래픽의 모델링은 패킷 및 응용 수준에서 주로 연구가 진행되고 있다[2-4]. 본 논문에서는 HTTP 요구 수준 단계의 웹 트래픽 모델링을 제시하고 액세스(access) 통신망의 성능 분석을 위해 필요한 통신망 트래픽의 특성을 파악하여 트래픽 부하의 생성을 위한 모델을 실제 수집된 자료를 분석하여 제시한다. 그리고 또한 현재 논란이 되고 있는 인터넷 트래픽의 통계적인 특성이 포아송 과정을 따르는지 아니면 자기유사한(self-similar) 특성을 보이는지를 본 연구에서는 웹 트래픽의 특성을 분석하여 살펴보았다. 본 연구의 결과는 어떠한 하나의 특성만을 가지는 것이 아니라 복합적인 특성을 가진다는 것을 알 수 있었다. 본 논문은 2장에서 관련한 기초 연구에 대하여 언급하고 3장에서는 수집한 데이터의 특성 분석을 보여준다. 4장과 5장에서는 이에 대한 원인 분석 및 결론을 맺도록 한다.

2. 관련 연구

2.1 HTTP(HyperText Transfer Protocol)

웹 트래픽을 모델링하기 위해서는 웹 서비스를 제공하기 위해 사용하는 HTTP의 이해가 필요하다. HTTP는 1990년대 중반 IETF(Internet Engineering Task Force)에서 처음 제안했으며 OSI(Open Systems Interconnection) 7계층 참조 모델 개념의 관점에서 응용계층 프로토콜으로써 단순 요청-응답 프로토콜로 설계되었다[5].

HTTP 응용들은 신뢰성 있는 데이터 전송을 위해 전송 계층 프로토콜인 TCP(Transmission Control Protocol)를 사용한다. 그렇기 때문에 클라이언트와 서버간의 데이터 전송을 위해 먼저 TCP 세션 설정이 요구된다. HTTP는 초기 버전 0.9에 이어 90년대 후반에 버전 1.1[6]이 발표되어 현재 널리 사용되고 있다. 먼저 HTTP의 동작 원리를 보면, 버전 1.0에서는 클라이언트가 서버로의 연결을 설정하고 요청을 보내면 이에 서버가 응답을 보낸다. 그리고 클라이언트가 응답을 받은 후 연결을 해지한다. 이러한 매커니즘에 의해 HTTP 프로토콜은 각 문서 당 TCP 연결을 수립하고 해지한다. 그러나 버전 1.1에서는 전송을 위한 요청 및 응답 마다 새로운 TCP를 수립하는 것이 아니라 하나의 TCP 연결을 통해 여러 요청 또는 응답을 보낼 수 있도록 수정하여 전송 속도의 향상과 TCP 연결 설정 및 유지에 따른 부하를 줄였다.

2.2 자료 수집 방법

웹 트래픽 모델링을 위해 필요한 자료는 모델링 하고자 하는 수준과 목적에 따라 달라질 수 있다. 웹 트래픽을 모델링하기 위해 요구되는 기본적인 트래픽 자료들은 크게 세 가지 방법에서 얻을 수 있다. 이를 다음에서 간략하게 설명하도록 한다.

첫 번째로 웹서버 로그를 이용하는 기법이다 [7]. 특별히 관리자가 웹서버 구성을 바꾸지 않는 한 웹서버는 기본적으로 웹 로그를 남긴다. 이 방법은 자료 수집을 위한 별도의 작업이 필요 없다는 장점이 있다. 그러나 사용자가 여러 웹 브라우저를 띄워 동시에 다른 서버에 접속 시에는 사용자의 웹 액세스 패턴을 알기가 어렵다. 두 번째는 클라이언트 로그를 이용하는 것이다 [8]. 다양한 형태의 정보들을 클라이언트 측에 기록되게 할 수 있다. 그러나 이를 지원하기 위해서는 브라우저의 수정 등 부과적인 작업이 필

요하다. 마지막으로 패킷 트레이스(trace) 분석 기법을 들 수 있다[3]. 즉, 링크 상에 흐르는 패킷을 직접 수집하여 분석함으로써 사용자의 웹 액세스 패턴 및 다양한 정보들을 수집할 수 있다. 그러나 트래픽 모델링을 위해 필요한 기본적인 정보들을 구성하기 위하여 수집된 개개의 패킷들을 조합하고 분석하는데 그 어려움이 따른다. 따라서 본 논문에서는 서버 로그를 구할 수 있는 부분에서는 서버 로그 분석 기법을 사용하고 그렇지 못한 부분에서는 패킷 트레이스 분석 기법을 이용한다.

표 1은 서버 로그 자료의 예를 보여준다. 기본적으로 서버 로그에는 웹 사용자의 요청에 대한 웹서버의 응답 정보가 기록된다. 기록되는 항목으로는 자료 요청 사용자의 IP 주소, 응답 시간, 요청한 파일의 서버내의 위치, HTTP 버전, 상태 코드, 응답 크기 등이 있다.

일반적으로 패킷 트레이스 분석에서 링크상의 패킷을 수집하기 위해 패킷 수집 도구인 TCPdump[9]가 널리 사용된다. TCPdump는 TCP와 인터넷 게이트웨이(gateway)의 성능을 연구하고 향상시키기 위한 목적으로 Van Jacobson에 의해 고안되었으며, 로컬(local) 네트워크의 인터페이스 상에 흐르는 모든 패킷들을 대상으로 주어진 조건식에 맞게 필터링하여 헤더 정보를 보여 주는 도구이다.

표 2는 TCPdump를 이용하여 수집한 웹 패킷 자료의 예를 보여주는 것이다. 본 예는 3개의 패킷에 대한 TCP 헤더 정보를 보여주는 것이며,

기록되는 정보로는 패킷 수집 시간, 송,수신지 IP 주소 및 포트번호, TCP 플래그 정보, 시퀀스 번호, 패킷 크기, ACK 번호, 버퍼 크기 등이 있다.

2.3 통계적 분석법

일반적으로 인터넷 트래픽을 모델링 하기 위해서는 확률 분포와 통계적인 방법을 이용한다. 따라서 본 논문에서도 이러한 방법을 통해 어떤 샘플(sample) 데이터가 어떤 확률 분포로 적합하게 모델링 되는지를 찾는다. 이를 위해 사용하는 통계 방법으로 단순선형회귀분석법(simple linear regression analysis)[10]을 이용하였다.

통계적 분석을 통해 확률분포로 모델링된 트래픽은 NS-2[11], OPNET[12], COMNET III[13], NetDAS[1] 등의 시뮬레이션 도구에 직접적으로 이용되어질 수 있다. 이러한 통계적 분석에는 일반적으로 확률분포의 누적분포함수(cumulative distribution function)나 확률밀도함수(probability density function)를 이용한다. 본 논문에서 사용된 확률분포는 지수, 정규, 로그노말, 파레토 분포를 사용하여 웹 트래픽을 모델링한다. 파레토 분포의 경우, 그 특성이 잘 알려진 지수 분포나 정규 분포와는 달리 Heavy-Tailed 분포로써 자기유사한(self-similar) 특성[4,14,15]이 있으며 이 분포를 따라 생성되는 난수들의 평균과 분산이 특정 파라미터 값에 따라서는 무한한 값을 가진다. 이러한 특성은 분포의 꼬리 부분에 많은 확률 값이 존재함에 기인한다.

〈표 1〉 서버 로그 자료

203.246.6.96	--	[1/Nov/2002:18:07:39 +0900]	"GET / HTTP/1.1"	200	363
203.246.6.96	--	[1/Nov/2002:18:07:39 +0900]	"GET /html/menu.html HTTP/1.1"	200	1954
203.246.6.96	--	[1/Nov/2002:18:07:39 +0900]	"GET /html/body.html HTTP/1.1"	200	4267

〈표 2〉 패킷 트레이스 자료

18:18:52.792974	203.246.6.96.1730	>	203.246.5.100.80:	P 1:350(349)	ack 1	win 17520	(DF)
18:18:52.825107	203.246.5.100.80	>	203.246.6.96.1730:	P 1461:2921(1460)	ack 350	win 24820	(DF)
18:18:52.826234	203.246.5.100.80	>	203.246.6.96.1730:	P 1461:2921(450)	ack 350	win 24820	(DF)

2.4 기존 웹 트래픽 모델링

웹 트래픽은 혼잡 제어 등의 영향을 받는 TCP 패킷의 흐름을 기록하여 패킷 수준에서 모델링 될 수 있고, 웹 파일의 크기 분포 등을 이용하여 응용 수준에서 모델링 될 수 있다. 이를 각각에 대하여 다음에서 언급하도록 한다.

패킷 수준 모델링은 링크상의 흐르는 패킷을 TCPdump 또는 기타의 패킷 수집 방법을 이용하여 수집하고 그 흐름을 분석하여 트래픽의 특성을 모델링 하는 것이다. 그러나 패킷 수준에서 모델링 되는 항목인 패킷간의 시간 간격은 TCP의 흐름 제어 및 혼잡 제어 알고리즘에 영향을 받는다. 이렇게 수집된 트래픽 특성은 해당 트래픽 수집 시 링크 상에 흐르는 배경 트래픽에 따라 TCP 알고리즘의 영향을 받아 매 측정 때마다 그 특성이 달라질 수 있다. 그러므로 통신망 시뮬레이션 시 필요한 트래픽 부하를 모델링하기 위해서는 부적절하다[3].

응용 수준 모델링은 패킷 수준 모델링의 단점을 극복하고자 하는 것으로 각 연결이 주고받는 데이터를 웹 파일 크기 등을 이용하여 언제, 얼마의 크기를 주고받았는지를 모델링 함으로써 웹 파일이 패킷 단위로 나뉘어 전송되는 패킷 수준에서와 같이 TCP 혼잡 제어 및 흐름 제어 알고리즘에 영향을 받지 않는다. 다시 말해서, 패킷들의 도착 시간 간격은 배경 트래픽에 영향을 받아 달라지더라도 응용 수준에서 언제 얼마의 크기가 전송되었는지는 변하지 않는다. [3]에서는 웹 응용의 성능분석을 위한 트래픽 부하 생성을 위해 응용 수준에서 경험적 트래픽 모델을 제안했으며, 다양한 관점에서 웹 트래픽의 특성을 모델링 했다. 그러나 응용 수준에서 사용자의 웹 액세스 패턴을 모델링 함으로써 모델링 해야 할 항목들이 많아지고 모델링된 트래픽 부하를 생성시키기 위한 트래픽 생성기 역시 복잡해진다는 단점이 있다.

3. 웹 트래픽 모델링

3.1 HTTP 요구 수준 모델링

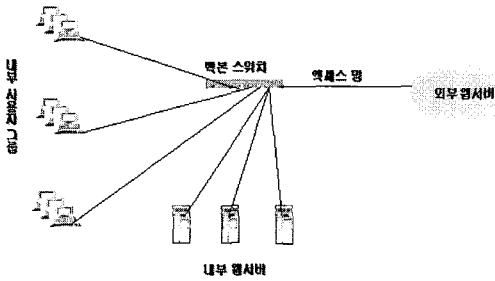
한 네트워크 내 총체적인 트래픽량은 웹 액세스를 통해 트래픽을 유발시키는 사용자의 수와 사용자당 웹을 액세스하는 시간에 따라 다르며, 일반화되어질 수 없는 요소이므로 대상이 되는 개별 네트워크의 실제 트래픽을 수집하여 모델링 되어야 한다. 따라서 액세스 네트워크 내의 총체적인 웹 트래픽을 모델링하기 위해서는 단일 사용자 관점이 아닌 네트워크 내에 존재하는 웹서버의 관점에서 트래픽을 모델링하며, 내부 사용자들이 외부 웹서버로의 요청에 의해 유발되는 웹 트래픽에 대해서도 단일 사용자 관점이 아닌 전체 사용자 관점에서 트래픽을 모델링 하는 것이 적절하다.

HTTP는 각 연결간의 데이터 전송을 위해 클라이언트의 요청이 서버로 보내지고 이에 상응하는 응답을 클라이언트에게 보내는 단순 요청-응답 메커니즘을 사용한다. 본 논문에서는 이러한 HTTP의 데이터 전송 메커니즘에 초점을 맞추어 HTTP 요구 수준에서 트래픽을 모델링 한다.

표 3은 HTTP 요구 수준에서 모델링될 항목들이다. 요청간의 시간 간격 항목은 모든 사용자 요청간의 시간 간격을 나타내며, 서버 로그의 경우 초 단위이고 패킷 트레이스인 경우 마이크로 초 단위이다. 요청 크기 항목은 클라이언트 측에서 서버로의 요청 정보를 보내기 위한 데이터의

〈표 3〉 HTTP 요구 수준에서의 모델링 항목

항목	단위	설명
요청간 시간 간격	seconds, micro seconds	모든 요청간의 시간 간격
요청 크기	bytes	HTTP 요구 길이
응답 크기	bytes	HTTP 응답 길이



〈그림 1〉 액세스 네트워크 토폴로지

크기 항목이고 응답 크기는 서버 측에서 클라이언트의 요청에 대한 응답으로 보내는 데이터의 크기 항목이다.

그림 1은 본 연구에서 웹 트래픽 모델링을 위해 실험의 대상이 된 액세스 네트워크의 구성도이며 그 구성은 다음과 같다. 먼저 본 네트워크는 3개의 서브(Sub) 네트워크로 나뉘어 있는 학내망이다. 네트워크 내 총 컴퓨터 수는 약 300대 정도이고 그 이용 인원수는 약 700명 정도로 추산되며, 약 10대의 주요한 웹서버가 현재 운영 중에 있다.

먼저 각 서버당 웹 트래픽을 내부 및 외부 사용자에게 의해 유발되는 트래픽으로 분리하였으며, 대상(target) 네트워크 내 모든 웹서버들의 서버 로그를 구하는데 따른 어려움으로 인해 분석 대상을 3대의 웹서버로 제한하였다. 그리고 내부 사용자들이 외부 웹서버들에 접속함으로써 유발되는 웹 트래픽은 웹서버들의 서버로그를 구할 수 없는 관계로 백본 스위치에서 직접 패킷을 수집하여 전체 내부 사용자 차원에서 모델링 하였다.

패킷 트레이스에서 모든 요청 및 응답 패킷을 기본 트래픽 자료로 사용하는 것과는 달리 서버

로그 자료에서는 성공 적으로 응답을 받은 것만을 대상으로 한다. 또한 서버 로그를 이용한 분석 시 요청 크기는 트래픽에 미치는 영향이 응답 크기에 비해 미비하고 서버 로그를 이용하여 모델링할 수 없기 때문에 웹서버를 대상으로 한 트래픽 모델링에서는 제외하고 패킷 트레이스를 이용한 부분에서만 그 특성을 모델링 한다.

본 논문에 사용된 자료는 표 4와 같이 3대의 서버 로그를 이용했다. 먼저 CE 서버는 학과 소개 및 학생들 간의 정보 공유를 위한 학과 홈페이지와 개인 사용자의 간단한 홈페이지로 구성되어 있으며, 하루 평균 접속자수는 약 120명 정도이다. 그리고 CSL 서버는 연구실용 서버이며, CE 서버와 유사한 목적으로 사용되고 있으며, 하루 평균 접속자수는 약 170명 정도이다. 마지막으로 HiBrain.Net 서버는 석,박사 채용정보 사이트이며, 현재 회원 가입자 수는 약 4만 명이며, 하루 평균 로그인 수는 약 5,000건 정도이고 하루 평균 접속자 수는 약 10,000건 정도이다.

대상 네트워크 내 사용자들의 외부 웹서버 액세스를 통해 유발되는 웹 트래픽을 모델링하기 위해서 그림 1의 백본 스위치에서 포트 미러링(port mirroring) 방법을 이용하여 링크 상에 흐르는 패킷을 직접 수집하였다. 패킷 수집을 위해 사용한 시스템은 펜티엄IV 1.2GHz의 CPU 2개 그리고 RAM 1GB가 장착되었으며, 운영체제는 레드햇 리눅스(LedHat Linux) 6.2를 사용했다. 그리고 패킷 수집을 위해 TCPdump 버전 3.7.1을 사용했으며, 패킷 수집 시 패킷 손실은 없었다. 표 5는 본 연구에 사용된 패킷 트레이스에 대한 설명이다. 패킷 단위로 수집된 데이터를 요

〈표 4〉 서버의 로그 자료 수집

웹서버	기간	요청횟수	총 응답크기(MB)
CE	01/09/18-01/11/11(59일)	350,368	2,193
CSL	02/06/17-02/09/01(78일)	570,040	20,980
HiBrain.Net	02/10/11-02/10/17(7일)	2,599,306	20,027

청 패킷과 응답 패킷으로 구분하고 단일 요청 및 응답으로 생성된 패킷들은 하나로 묶어서 크기를 산출하여 트래픽 분석에 이용하였다. 수집 기간

<표 5> 패킷 트레이스 자료

수집 위치	SUMMIT 48 Switch
기간	2002/10/29 23시 2002/10/30 22시 (23시간)
요청 횟수	591,709
총 응답 크기(MB)	7,517

은 패킷 트레이스의 크기를 고려하여 약 하루 동안의 데이터로 제한하였다.

3.2 요청간의 시간 간격 모델링

본 항목의 모델링은 각 내부 웹서버들에 들어오는 요청에 대한 시간 간격과 내부 사용자들에

의해 액세스된 모든 외부 웹서버들에 대한 요청간의 시간 간격을 대상으로 한다. 특히 전자의 경우, 내부 사용자와 외부 사용자 그룹으로 나눠 그 특성을 모델링 한다.

표 6과 표 7은 사용자 요청간의 시간 간격을 확률 분포로 모델링한 것이다. 내부 및 외부 사용자에 의한 3대의 내부 웹서버들과 내부 사용자에 의한 모든 외부 웹서버들을 대상으로 한 요청 시간 간격이 특정한 단일 분포로는 모델링되지 않았고 두 분포를 사용하여 전 구간을 모델링할 수 있었다. 그 결과, 하위 부분은 지수 분포로, 상위 부분은 파레토 분포로 모델링이 되었으며, 외부 웹서버들을 대상으로 한 항목을 제외한다면 나머지 모든 항목에서 파레토 분포를 따르는 부분은 총 분포 구간 중 상위 약 10%~28%로 작은 부분을 차지했다. R^2 은 실제 자료들이 얼마나 이론적인 분포에 적합한지를 나타내는 결정 계수로써 0과 1사이의 값을 가지며, 1에 가까울수록

<표 6> 내부 사용자의 요청간 시간 간격에 대한 분포 파라미터 추정치

웹서버	적용 비율	확률 분포	파라미터	R^2
HiBrain.Net	73%	지수	$\mu = 0.22$	0.83
	23%	파레토	$a = 0.99, k = 1.2$	0.96
CSL	74%	지수	$\mu = 0.22$	0.87
	26%	파레토	$a = 0.77, k = 3.34$	0.93
CE	72%	지수	$\mu = 0.99$	0.92
	28%	파레토	$a = 0.96, k = 2.37$	0.98
외부 웹서버	39%	지수	$\mu = 0.0035$	0.95
	61%	파레토	$a = 0.98, k = 0.01$	0.96

<표 7> 외부 사용자의 요청간 시간 간격에 대한 분포 파라미터 추정치

웹서버	적용 비율	확률 분포	파라미터	R^2
HiBrain.Net	90%	지수	$\mu = 0.21$	0.70
	10%	파레토	$a = 2.95, k = 0.47$	0.97
CSL	84%	지수	$\mu = 1.2$	0.94
	16%	파레토	$a = 1.27, k = 7.6$	0.92
CE	76%	지수	$\mu = 0.42$	0.91
	24%	파레토	$a = 1.0, k = 1.8$	0.96

적합하다고 할 수 있다. 이 수치가 HiBrain.Net의 외부 사용자에 대한 요청 간격 항목의 하위 지수 분포에 대한 부분을 제외한 모든 부분에서 0.8 이상으로 높게 나왔기 때문에 조사된 대부분 서버에서의 요청간의 시간 간격은 지수 분포와 파레토 분포의 혼합으로 잘 모델링 된다고 할 수 있다.

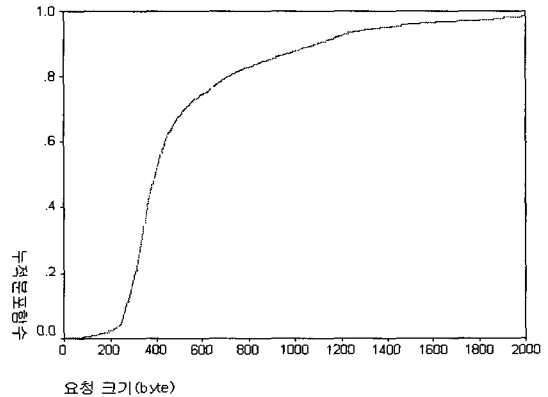
3.3 요청 크기 모델링

클라이언트가 웹서버로부터 데이터를 받기 위해 데이터 요청을 해당 웹서버로 보내게 된다. 이러한 요청을 위한 패킷 내의 정보로는 데이터 전송을 위해 사용되는 해당 프로토콜간의 통신을 위한 정보가 포함되며, 그 크기는 프로토콜별로 고정적이다. 또한 해당 클라이언트가 웹서버 내에 있는 웹 문서 등을 요청하는 경우 요청된 데이터의 웹서버 내 위치를 나타내는 정보가 있다. 그러나 클라이언트가 항상 웹서버로 요청 정보만을 보내는 것이 아니라 게시판에 글을 작성하거나 파일을 첨부하여 올릴 수도 있다. 이렇게 클라이언트 측에서 웹서버로의 업로드(upload)가 발생하면 패킷들의 크기가 자료에 따라 아주 커질 수 있다. 엄밀히 말하면, 이러한 클라이언트에 의해 웹서버로의 업로드는 데이터 등의 요청은 아니지만 본 연구에서는 요청 크기 항목에 포함하여 모델링 한다.

웹서버에 의해 자동적으로 기록되는 서버 로그에는 클라이언트 측에서 웹서버로의 자료 요청에 의해 유발되는 트래픽의 크기가 기록되지 않기 때문에 모든 외부 웹서버에 대한 사용자의 웹 트래픽을 모델링하기 위해 사용하는 패킷 트레이스 자료를 가지고 요청 크기를 모델링 한다. 수집한 패킷 트레이스 자료에 따르면 요청 크기가 전체 웹 트래픽에서 차지하는 비중이 5% 이하로 작은 비중을 차지하고 또한 그 특성이 특정 사용자 집단에 따라 독립적이지 않고 어느 정도 일반적인 특성이 있는 것으로 보이기 때문에 그림 1

의 네트워크의 성능 분석을 위한 시뮬레이션에서 개별 웹서버를 대상으로 한 웹 트래픽 생성 부분에서 사용하여도 된다.

그림 2는 요청 크기에 대한 누적분포함수 그



〈그림 2〉요청 크기에 대한 누적분포함수 그래프

래프를 나타낸 것이다. 이 그래프를 보면 약 1,200 바이트 이하의 요청 크기가 대부분을 차지하며, 그 이상에서 그래프가 긴 꼬리를 가지며 계속 이어지는 것을 볼 수 있다. 이러한 특성은 클라이언트 측에서 웹서버로의 업로드에 의해 발생하는 트래픽의 부분으로 판단된다. 그리고 또한 약 250바이트 에서 약 500바이트 사이에서 누적 확률 값이 급격히 상승하는 것을 보아 이 구간에 많은 요청 크기가 몰려 있음을 알 수 있다. 그러나 TCPdump를 이용한 패킷 트레이스 자료에서는 사용되는 프로토콜 헤더에 관한 정보 밖에 볼 수 없기 때문에 정확한 원인을 알 수 없으며, 원인 분석을 위해서는 추가적인 정보 수집이 필요하다.

표 8은 요청 크기를 확률 분포를 사용하여 모델링한 것이며, 그 결과 로그노말 분포로 가장 적합하게 모델링 되었다. 비록 적합도를 나타내는 결정계수의 수치가 0.76으로 낮게 산출되었지만 요청 크기가 전체 웹 트래픽에 미치는 영향이 다른 항목들에 비해 미비함으로 인해 시뮬레이션 시 트래픽 로드를 생성하기 위해 사용하여도 무

방하다고 생각한다.

〈표 8〉 요청 크기의 분포 파라미터 추정치

확률 분포	파라미터	R2
로그 노말	$\mu = 321, \sigma = 322$	0.76

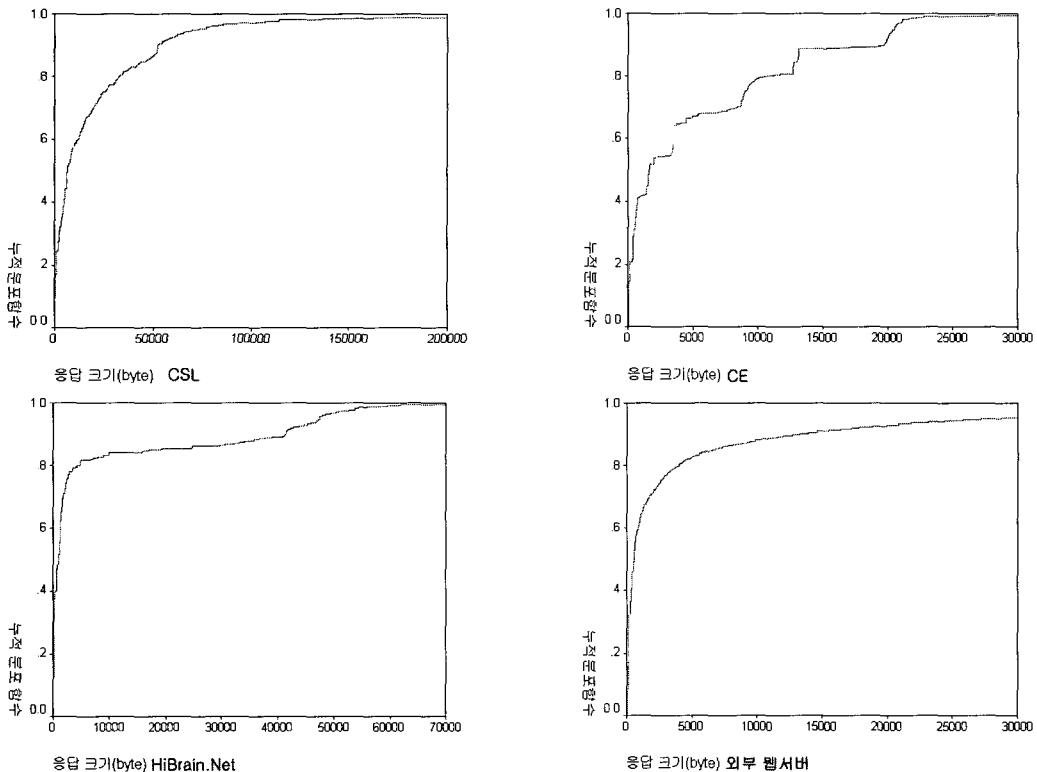
3.4 응답 크기 모델링

웹 트래픽의 응답 크기에는 HTML(HyperText Markup Language) 문서 및 멀티미디어 파일들과 같은 웹서버 내에서 크기가 정해져 있는 정적인 파일들과 CGI(Common Gateway Interface)로 구성된 게시판 액세스 등으로 인해 발생하는 동적인 자료의 전송에 의한 트래픽도 포함된다. 본 장에서는 조사된 대상 네트워크 내 웹 트래픽

에 영향을 미치는 웹서버들의 응답 크기를 먼저 분석해본다. 그리고 다른 네트워크상의 웹서버들에서도 같은 특성이 나타나는지를 확인해본다.

3.4.1 대상 네트워크 내 응답크기 모델링

응답 크기의 통계적 성질은 웹서버 내의 파일들의 특성을 따르며[8], 사용자 그룹에 따라 그 특성이 다르지 않을 것으로 가정하여 본 항목은 사용자들의 요청에 대한 응답 크기를 내부 및 외부 사용자의 요청에 의한 응답 크기로 나누지 않고 웹서버별로 응답 크기를 모델링 한다. 분석되는 웹서버들은 먼저 대상 네트워크 내에 존재하는 3대의 웹서버들과 내부 사용자들의 요청에 의한 외부 웹서버들의 총체적인 응답 크기를 외부 웹서버 항목으로 정의하여 총 4가지로 나뉜다.



〈그림 3〉 웹서버들의 응답 크기의 누적분포함수 그래프

그림 3은 각 웹서버 및 외부 웹서버들에 대한 응답 크기의 누적분포함수 그래프를 나타낸 것이다. 먼저 누적 확률값이 CSL의 경우 0.3, CE에서는 0.5까지 급격히 상승하고 그 이후에는 상승 폭이 조금씩 둔화되며, 뒷부분에서는 Heavy Tailed 분포의 특성과 같이 긴 꼬리를 형성하는 것을 볼 수 있다. 그러나 누적 확률값이 급격히 상승한다는 것은 그 부분에 응답 크기의 빈도가 높다는 것을 의미하며, 결국 이 구간에서는 분산값이 작다는 것을 알 수 있다. 이를 봄으로써 이 구간에서는 분산값이 큰 Heavy Tailed 분포로 모델링이 되지 않는다는 것을 알 수 있다. HiBrain.Net과 외부 웹서버들에 대한 그래프에서는 누적 확률 값이 급격히 상승하는 부분이 HiBrain.Net의 경우 약 0.8까지 지속되며, 외부 웹서버들의 경우 약 0.7까지로 CSL과 CE에서 보다는 높게 나왔다. 그러나 그 특성이 두 구간으로 나뉜다는 것을 역시 확인 할 수 있다. 결국, 응답 크기 항목을 확률 분포를 사용하여 모델링하기 위해서는 기존에 제시된 Heavy Tailed 분포인 파레토 분포 단독으로는 전 구간을 모델링할 수 없고 하위 부분에 맞는 추가적인 분포가 필요하다는 것을 알 수 있다.

표 9는 웹서버들에 대한 응답 크기의 특성을 확률 분포를 사용하여 모델링한 결과를 나타낸 것이다. CSL과 CE 그리고 외부 웹서버들의 응답 크기는 누적분포함수 그래프에서 살펴본 것과 같이 하위 부분은 지수 분포로 그리고 상위

부분은 파레토 분포로 잘 모델링 되었으며, 지수 분포로 모델링된 하위 부분의 적용 비율이 누적분포함수에서와 같이 누적 확률값이 급격히 상승하는 부분까지 잘 적용되는 것을 볼 수 있다. 그러나 HiBrain.Net의 경우 본 연구에서 조사된 다른 서버들 및 [3,8]과는 달리 파레토 분포로는 모델링이 되지 않았고 전 구간에서 지수 분포로 더욱더 적절히 모델링 되었다.

그림 3의 HiBrain.Net 서버의 누적분포함수 그래프를 다시 살펴보면, 누적 확률값이 0.8까지 급격한 상승을 보인다. 이러한 특성은 다른 웹서버에서도 나타나지만 HiBrain.Net의 경우, 그 수치가 상대적으로 다른 웹서버에 비해 높다. 그리고 누적 확률값이 0.8까지 급격히 상승하지만, 그 이후 급격히 둔화되어 응답 크기의 약 40KB까지 거의 누적 확률값이 증가하지 않는다. 이러한 특성은 이 구간에서 응답 크기의 빈도가 상대적으로 아주 작다는 것을 나타낸다. 그리고 약 40KB~60KB 까지 적은 범위에서 가파른 상승세를 보이며, 누적 확률값이 거의 1.0에 가까워진다. 다시 말해서 누적 확률값의 급격한 증가세가 둔화되는 부분에서 누적분포 값이 1.0까지의 분포 그래프가 파레토 분포의 특성을 갖는 다른 웹서버들과는 현저히 다른 것을 알 수 있다. 따라서 이런 특성으로 인하여 상위 부분이 파레토 분포로 모델링이 되지 않고, 오히려 전 구간에 걸쳐 지수 분포로 모델링 된다는 것을 알 수 있다.

〈표 9〉 응답 크기의 분포 파라미터 추정치

웹서버	적용 비율	확률 분포	파라미터	R ²
HiBrain.Net	100%	지수	$\mu = 15,621$	0.98
CSL	30%	지수	$\mu = 676$	0.85
	70%	파레토	$a = 0.94, k = 2,117$	0.94
CE	53%	지수	$\mu = 526$	0.95
	47%	파레토	$a = 1.53, k = 2,903$	0.82
외부 웹서버	70%	지수	$\mu = 406$	0.97
	30%	파레토	$a = 1.13, k = 1,800$	0.99

〈표 10〉 외부 웹서버들의 서버 로그 자료

웹서버	기간	요청횟수	총 응답크기(MB)
Sarim	02/11/17-02/11/26(10일)	4,603,933	64,831
NeoSight	02/10/04-02/10/14(11일)	1,567,087	11,099

〈표 11〉 외부 웹서버들의 응답 크기의 분포 파라미터 추정치

웹서버	적용 비율	확률 분포	파라미터	R ²
Sarim	43%	지수	$\mu = 428$	0.96
	57%	파레토	$a = 0.91, k = 1,295$	0.98
NeoSight	44%	지수	$\mu = 323$	0.91
	56%	파레토	$a = 0.90, k = 1,000$	0.86

3.4.2 다른 통신망 상의 웹서버들의 응답크기 모델링

본 절에서는 대상 통신망 내 웹 트래픽에 영향을 미치는 웹서버들의 응답 크기의 특성이 다른 통신망 상에 위치한 웹서버들에서도 나타나는지를 확인하기 위해 외부에 위치한 두 웹서버를 분석해 본다.

표 10은 외부에 위치한 두 대의 웹서버들의 기본 트래픽 자료를 나타낸다. 먼저 Sarim 서버에는 대학의 메인 홈페이지가 운영되고 학교 및 학내 행사의 홍보 등 다양한 용도로 사용되며, 하루 평균 접속자 수는 약 9,500명 정도로 추산된다. 그리고 NeoSight 서버에는 회사 홍보용 웹사이트가 운영되고 회사 홍보를 위한 자료로 구성되어 있으며, 하루 평균 접속자 수는 약 3,200명 정도이다.

표 11은 응답 크기의 통계적 특성을 확률 분포를 사용하여 분석한 것이다. 그 결과 표 9의 내부 웹서버들의 응답 크기의 통계적 특성과 같이 하위 부분에서는 지수 분포로, 상위 부분에서는 파레토 분포로 적합하게 모델링 되었다. 결국, 웹서버들의 응답 크기는 기존에 제시된 파레토 분포만으로 모델링 하는 것보다 하위 부분에서는 지수 분포로, 상위 부분에서는 파레토 분포로 더

욱더 적절하게 모델링 된다는 것을 알았으며, 내부 웹서버인 HiBrain.Net의 경우 지수 분포로 모델링 되는 부분이 100%를 차지한다고 볼 수 있다.

4. 웹 트래픽 특성 분석

4.1 액세스 망에서의 웹 트래픽 특성 분석

기존 연구를 보면 LAN 및 WAN 환경에서 웹 트래픽의 특성이 자기유사하다는 주장 [2,8,16,17]도 있고, 이와 상반되게 포아송 과정을 따른다는 주장[18-20]도 있다. 어떠한 트래픽이 자기유사한 특성을 가진다는 것은 트래픽의 분산이 크다는 것이다. 이러한 트래픽을 적절히 처리하기 위해서는 최번시(busy hour)의 트래픽량보다 훨씬 높은 트래픽을 수용할 수 있는 대역폭 및 통신망 장비가 요구된다. 반면에 트래픽이 포아송 과정을 따른다면 최번시의 트래픽량보다 조금 높은 수준의 트래픽을 처리할 수 있는 수준에서 통신망을 설계할 수 있다. [8]에서는 웹 트래픽의 자기 유사한 특성을 확인하고 이러한 특성이 클라이언트의 요청에 의한 서버로부터의 응답 크기들의 특성을 따른다는 것을 보였다. 그리고 또한 [21]에서는 다중화된 IP 트래픽들은

자기유사성을 띄고, 이 같은 자기유사성은 트래픽 패턴에 영향을 미치는 전송 계층, 혼잡제어 알고리즘, 통신 환경과 같은 여러 인자와 무관함을 주장하였다. 이러한 주장에 따르면, 본 논문에서 사용된 내부 3대의 웹서버 중 2대의 웹서버와 외부 웹서버들에 대한 응답 크기 항목이 표 9와 같이 상위 부분에서 자기유사한 특성이 있는 파레토 분포로 모델링 됨으로써 웹 트래픽의 특성이 자기유사한 특성이 있음을 말해주지만, 반대로 HiBrain.Net 서버의 경우 응답 크기가 지수 분포를 따르기 때문에 또한 포아송 과정을 따른다고 볼 수 있다. 결국 본 연구 결과, 액세스 망에서의 웹 트래픽의 특성은 대상이 되는 통신망에 영향을 미치는 웹서버들의 응답 크기 특성에 따라 달라진다고 할 수 있다.

4.2 웹 트래픽 특성의 원인 분석

액세스 망에서의 웹 트래픽 특성을 좌우하는 응답 크기의 특성이 [8]에서는 서버 내 파일 크기의 특성과 일치한다는 것을 보였으며, 궁극적으로 웹 트래픽의 특성이 서버 내 파일들의 크기의 특성에서 기인한다고 주장하였다. 본 연구에서도 이러한 성질이 있는지를 분석하여 정확한 원인을 알아보았다.

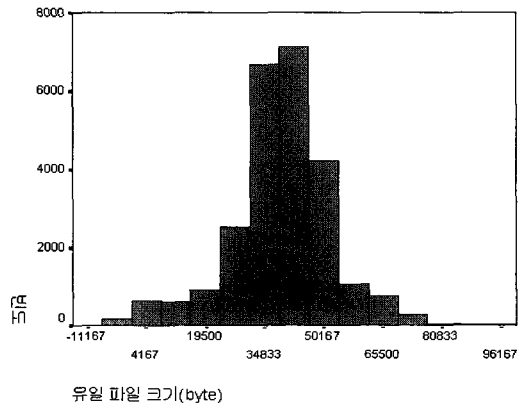
웹서버 내 파일 중 웹 서비스에 의해 제공되는 파일들의 크기를 직접 조사하는 것은 쉽지 않은 관계로 본 연구에서도 [8]에서처럼 응답 크기를 이용하며, 사용자에게 의해 한번 이상 액세스된 모든 파일들을 대상으로 하였다.

표 12는 서버 내의 파일 크기들을 확률 분포를 사용하여 가장 적합한 분포와 파라미터를 산출한 것이다. 표 12를 보면 각 서버 내 파일 크기들의 분포가 CSL과 CE 서버의 경우, 응답 크기에서와 같이 파레토 분포로 잘 모델링이 되었다. 하지만 HiBrain.Net의 경우 응답 크기에서와 같이 지수 분포로 모델링 되지 않고 형태가 완전히 다른 정규 분포로 모델링 되었다. 그림 4에서

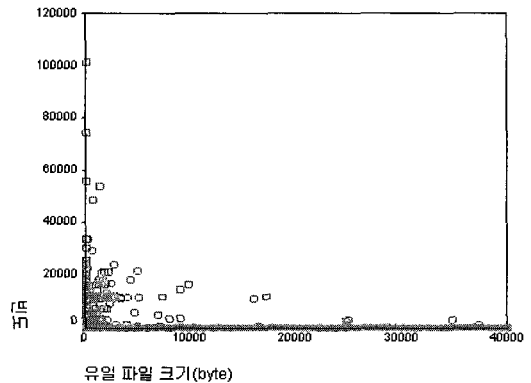
도 분포의 모양이 정규 분포와 유사한 것을 확인할 수 있다. 이렇게 HiBrain.Net의 경우처럼 응답 크기와 서버 내 파일 크기의 특성이 다르다는

〈표 12〉 서버 내 파일 크기들의 분포 파라미터 추정치

웹서버	확률 분포	파라미터	R^2
CSL	파레토	$\alpha = 1.09$	0.98
CE	파레토	$\alpha = 0.81$	0.96
HiBrain.Net	정규	$\mu = 39,872, \sigma = 10,486$	0.99



〈그림 4〉 HiBrain.Net 서버 내 파일 크기들의 빈도 그래프



〈그림 5〉 HiBrain.Net 서버 내 파일들에 대한 사용자 액세스 패턴

것은 사용자의 파일에 대한 요청이 특정 파일들

에 편중적으로 이루어진다는 것을 알 수 있다. 그림 5는 HiBrain.Net 서버 내 파일들이 사용자에 의해 액세스된 빈도를 파일 크기 순으로 나타낸 것이다. 그래프를 통해 사용자의 웹 파일 액세스 패턴이 전체 파일들에 골고루 액세스 되지 않고 소수의 작은 크기의 파일들에 편중되는 것을 확인 할 수 있다.

이러한 분석을 통해 CSL 서버와 CE 서버에서는 응답 크기의 특성과 서버 내 파일 크기의 특성이 일치함으로써 사용자의 웹 파일 액세스 패턴에 의해 응답 크기의 특성이 영향을 받지 않지만, HiBrain.Net 서버의 경우 직접적으로 그 영향을 받는 것으로 보아 서버 내 파일 크기와 응답 크기의 특성이 서버에 따라서는 다르게 나타날 수 있다는 것을 알았다.

HiBrain.Net 서버 내 파일 크기의 성질이 CSL과 CE 서버와는 달리 자기유사한 특성이 없는 정규 분포와 같이 모델링이 되는 이유를 살펴보면, 3대의 웹서버에서 주로 게시판용을 사용하여 사이트를 구성하였지만 CSL과 CE 서버에서는 게시판에 사용자들이 첨부하여 업로드할 수 있는 파일의 최대 크기를 제한하지 않는 반면, HiBrain.Net의 경우 최대 크기를 5MB로 제한하였다. 그리고 또한 HiBrain.Net 서버의 경우 멀티미디어 관련 파일 등과 같이 큰 파일들이 거의 존재하지 않음으로써 파일 크기들의 범위가 작고 이로 인해 분산 또한 작다. 결국 이러한 원인에 의해 서버 내 파일 크기의 분포가 자기유사한 특성이 없는 정규 분포로 모델링 되고, 이러한 특성이 사용자의 웹 파일 액세스 패턴에 영향을 받아 응답 크기 항목이 지수 분포의 특성을 가지게 됨을 알 수 있다.

5. 결론

본 논문에서는 통신망 성능 분석을 위한 시뮬레이션 시 필요한 트래픽 생성기의 설계를 위해 실제 트래픽 자료를 수집, 분석하여 HTTP 요구

수준에서 통계적 방법을 통해 확률 분포로 모델링 하였다. 본 연구에서 주요한 결론의 한 가지는 기존에 단일 확률 분포만을 사용하여 모델링 하던 것을 본 연구에서 두 분포의 혼합으로 웹 트래픽의 특성을 모델링하고 이러한 특성이 나타나는 원인을 분석해 보았다.

모델링 되는 항목 중의 하나인 응답 크기는 전체적인 웹 트래픽 특성에 미치는 영향이 큰 만큼 정확한 분석을 통해 그 특성을 파악하는 것이 중요하다. 하지만 지금까지는 파레토 분포만을 사용하여 그 특성을 모델링 하였지만, 본 연구에서는 파레토 분포를 따르지 않는 하위 부분은 지수 분포로 적합하게 모델링이 된다는 것을 분석을 통해 알았다. 그러므로 응답 크기는 지수 분포와 파레토 분포의 혼합으로 적합하게 트래픽 부하를 생성할 수 있다.

웹 트래픽의 특성을 파악하기 위해 모델링 되는 항목 중의 하나인 응답 크기의 특성을 웹서버 별로 파악해 봄으로써 그 특성이 개별 웹서버에 대해 독립적이라는 것을 확인할 수 있었으며, 그 원인을 분석해 보았다. 또한 이러한 응답 크기의 특성이 액세스 망의 웹 트래픽 특성을 반영한다는 기존 연구를 통해 그 특성이 자기유사한 특성만이 있는 것이 아니라 웹서버에 따라 달라질 수 있다는 것을 알았다. 그러므로 조사된 환경에 따라 웹 트래픽의 특성이 자기 유사한 성질이 있을 수 있고 또한 포아송 과정을 따를 수 있다는 것을 분석을 통해 알았다. 또한 기존 연구를 보면 응답 크기의 특성이 서버내 파일 크기의 특성을 따른다고 하지만 분석 결과 그 특성이 일부 웹서버에서는 다르다는 것을 알았다. 그러므로 응답 크기의 특성은 서버 내 파일 크기의 특성을 그대로 반영하는 것이 아니라 사용자의 웹 문서 요청의 편중화 현상에 영향을 받아 그 특성이 달라질 수 있다는 것을 분석을 통해 알았다.

향후 연구로는 본 연구에서와 같이 웹 트래픽이 지수 분포와 파레토 분포의 혼합으로 생성되지만, 두 분포의 비율 및 파라미터 값은 웹서버

에 따라 다음으로 인하여 이러한 파라미터를 빠르고 정확하게 파악 할 수 있는 자동화 도구의 개발이 필요하며, 또한 통신망 상의 전체적인 트래픽 부하를 생성하기 위해 웹 트래픽 이외의 다른 응용의 트래픽들의 특성 분석이 요구된다.

참고문헌

- [1] 최재원 외 7명, "네트워크 분석을 위한 시뮬레이션 시스템 설계", 2002년도 한국통신학회 하계종합학술발표회, Vol. 25, 무주, Jul. 2002.
- [2] A. Veres et al., "On the Propagation of Long-Range Dependence in the Internet", *Computer Communication Review*, Vol. 30, No. 4, Oct. 2000.
- [3] Bruce A. Mah, "An Empirical Model of HTTP Network Traffic", in *Proceedings of IEEE INFOCOM, Japan*, pp. 592-600, Apr. 1997.
- [4] Peter B. Danz and Sugih jamin, "tcplib: a Library of TCP Internetwork Traffic Characteristics", Technical Report CS-SYS-91-01, Computer Science Department, University of Southern California, Los Angeles, 1991.
- [5] T. Berners Lee et al., "Hypertext Transfer Protocol - HTTP/1.0", IETF RFC-1945, May 1996.
- [6] R. Fielding et al., "Hypertext Transfer Protocol - HTTP/1.1", IETF RFC-2068, Jan. 1997.
- [7] Martin F. Arlitt and Carey L. Williamson, "Web Server Workload Characterization: The Search for Invariants", in *Proceedings of the ACM SIGMETRICS Conference on Measurement & Modeling of Computer Systems*, pp. 126 - 137, May 1996.
- [8] Mark E. Crovella and Azer Bestavros, "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes", *IEEE/ACM Transactions on Networking*, Vol. 5, No. 6, pp. 835-840, Dec. 1997.
- [9] Lawrence Berkeley National Laboratory, Berkley Packet Filtering Tool, <http://www.tcpdump.org/>
- [10] 이외숙 외 3명, "통계학 입문", 서울:경문사, 2000.
- [11] Lawrence Berkeley National Laboratory, "Network Simulator-2", <http://www.isi.edu/nsnam/ns/>
- [12] OPNET Technologies Inc., "OPNET", <http://www.mil3.com/>
- [13] CACI Inc., "COMNET III", <http://www.caci.com/>
- [14] Vern Paxson and Sally Floyd, "Wide- Area Traffic: the Failure of Poisson Modeling", in *Proceedings of ACM SIGCOMM, London*, pp. 257 - 268, Sep. 1994.
- [15] 김철, 김영탁, "IP-over-ATM 체계의 Internet Traffic에 대한 Self-Similarity 특성 분석", 한국통신학회, 하계학술발표회 논문집, Vol. 19, 관동대학교, pp. 865-868, Jul. 1999.
- [16] Xusheng Tian et al., "A Unified Framework for Understanding Network Traffic Using Independent Wavelet Models", in *Proceedings of IEEE INFOCOM, New York*, Jun. 2002.
- [17] S. Uhlig and O. Bonaventure, "Understanding the Long-Term Self-similarity of Internet Traffic", QOFIS2001, Portugal, pp. 286 - 298, Sep. 2001.
- [18] R. Morris and D. Lin, "Variance of Aggregated Web Traffic", in *Proceedings*

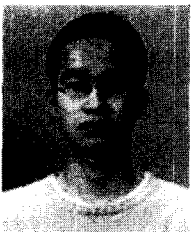
of IEEE INFOCOM, Israel, pp. 360 - 366, Mar. 2000.

- [19] 박형구 외 5명, “인터넷 트래픽 분석과 역세스 네트워크 모델링”, 한국통신학회 추계 학술발표회 논문집, Vol. 22, 연세대학교, pp. 523-526, Nov. 2000.
- [20] Jin Cao et al., “Internet Traffic Tends To

Poisson and Independent as the Load Increases”, Bell Labs. Technical Report, Murray Hill, 2001.

- [21] 홍승범 외 3명, “다중화된 IP 트래픽의 자기유사성에 관한 연구”, 한국통신학회 추계 학술발표회 논문집, Vol. 22, 연세대학교, pp. 519-522, Nov. 2000.

● 저자 소개 ●



전 의 수

2001년 경남대학교 컴퓨터공학과 졸업 (공학사)
2003년 창원대학교 대학원 컴퓨터공학과 졸업 (공학석사)
2003년~현재 창원대학교 정보전산원
관심분야 : 트래픽 모델링, 망관리시스템, QoS 관리, 멀티캐스트 프로토콜
E-mail : fraser@cdcs.changwon.ac.kr



이 광 휘

1983년 고려대학교 전자공학과 졸업 (공학사)
1985년 고려대학교 대학원 전자공학과 졸업 (공학석사)
1989년 고려대학교 대학원 전자공학과 졸업 (공학박사)
1988년~현재 창원대학교 컴퓨터공학과 교수
1991년~1992년 영국 Walse대학(Swansea) 및 Newbridge Networks사 연구원
1994년~1995년 영국 런던대학(UCL) 연구원
1997년~1999년 영국 Reading대학 연구원
2000년~현재 영국 Nortel Networks사 연구원
관심분야 : 망관리시스템, 멀티캐스트 프로토콜, 이동통신망, QoS 관리, 분산시스템
E-mail : khlee@sarim.changwon.ac.kr