광역 네트워크 트래픽의 장거리 상관관계와 1/f 노이즈

(Long-Range Dependence and 1/f Noise in a Wide Area Network Traffic)

이 창 용 †
(Chang-Yong Lee)

요 약 본 논문에서는 네트워크 트래픽의 수동적 측정치 분석을 통해 잘 알려진 장거리 상관관계가 광역 네트워크의 능동적 측정치에도 존재하는지 여부를 관련 분석법을 통하여 검정하고자 한다. 이를 위하여 PingER 프로젝트를 통하여 측정된 광역 네트워크 트래픽의 대표적인 능동적 측정치인 RTT(Round Trip Time)와 RTT의 변동성 시계열 데이터에 대하여 분석을 수행하였다. RTT 시계열 데이터는 장거리 상관관계 혹은 1/f 노이즈의 특성을 보였으며, RTT의 고차원 변화량으로 정의된 변동성은 로그정규분포를 따르며 변동성에 대한 장거리 상관관계는 고려하는 시간 간격이 짧은 경우 장거리 상관관계를 보이고, 시간 간격이 긴 경우에는 장거리 상관관계 혹은 1/f 노이즈를 따름을 밝혔다. 본 연구를 통해 볼 때 장거리 상관관계는 비단 패킷 도착의 시간 간격 등과 같은 수동적 측정뿐만 아니라 RTT와 같은 능동적 측정에서도 나타나는 특징이며, 특히 능동적 측정에는 수동적 측정에는 잘 나타나지 않는 1/f 노이즈 특성이 존재함을 밝혔다.

키워드 : 광역 네트워크, 장거리 상관관계, RTT, 변동성, 1/f 노이즈

Abstract In this paper, we examine a long-range dependence in an active measurement of a network traffic which has been a well known characteristic from analyses of a passive network traffic measurement. To this end, we utilize RTT(Round Trip Time), which is a typical active measurement measured by PingER project, and perform a relevant analysis to a time series of both RTT and its volatilities. The RTT time series exhibits a long-range dependence or a 1/f noise. The volatilities, defined as a higher-order variation, follow a log-normal distribution. Furthermore, volatilities show a long-range dependence in relatively short time intervals, and a long-range dependence and/or 1/f noise in long time intervals. From this study, we find that the long-range dependence is a characteristic of not only a passive traffic measurement but also an active measurement of network traffic such as RTT. From these findings, we can infer that the long-range dependence is a characteristic of network traffic independent of a type of measurements. In particular, an active measurement exhibits a 1/f noise which cannot be usually found in a passive measurement.

Key words : Wide Area Network, Long-range dependence, Round trip time, Volatility, 1/f noise

· 이 논문은 2008년 공주대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음

† 정 회 원 : 공주대학교 산업시스템공학과 교수
clee@kongju.ac.kr

논문접수 : 2009년 8월 26일
심사완료 : 2009년 12월 8일

정보과학회논문지: 정보통신 제37권 제1호(2010.2)

1. 서 론

학술 분야는 물론이고 사회생활 전반에 인터넷이 차지하는 비중이 갈수록 커짐에 따라 인터넷의 속도와 연결 상태 등 인터넷의 성능 문제가 지대한 관심사로 부각되기 시작하였다[1]. 인터넷은 네트워크 하드웨어의 발전에도 불구하고 사용자와 트래픽 양의 기하급수적인 증가로 인하여 정체(congestion), 패킷 손실(packet loss), 그리고 연결 지연(delay) 등의 문제가 발생하고 있다. 보다 빠르고 안정적인 인터넷 사용을 위해서는 라우터(router), 전송매체 등 하드웨어적 측면과 프로토콜(pro-

tocol) 및 응용 프로그램 등 소프트웨어 측면에서 인터넷의 성능을 향상시키는 노력도 중요하지만, 이에 더하여 인터넷과 같은 광역 네트워크 트래픽의 특성에 대한 체계적이고 심층적인 분석도 필요하다. 기존 연구 및 개발 동향을 통해 볼 때, 하드웨어와 소프트웨어 측면에서 인터넷의 성능 및 연결 상태를 향상시키고자 하는 연구는 다양한 각도로 진행되고 있지만, 인터넷 트래픽의 특성에 대한 연구는 상대적으로 부족한 현실이다.

전통적으로 네트워크 공학 분야에 속했던 인터넷에 관한 연구가 최근 들어 물리학적 개념 및 방법론을 접목하여 새로운 각도에서 연구가 활발하게 진행되고 있고, 일부 가시적인 성과를 내고 있다. 이것은 인터넷이 네트워크 하드웨어와 프로토콜들에 의하여 작동되는 공학적 시스템이지만, 대규모 자발적 조직체(massive self-organized system)로 물리학자들이 오랫동안 연구해온 시스템과 유사한 성질을 가지고 있기 때문이기도 하다. 따라서 인터넷은 이제 학제 간 연구로 자리매김 하고 있으며, 공학적 산물인 인터넷을 자연 과학 측면에서 수행한 연구들은 인터넷 트래픽을 이해하는데 큰 역할을 담당할 수 있을 것으로 예상된다.

인터넷은 매우 복잡한 시스템임으로 인터넷 성능에 대한 연구를 위해 인터넷 전체를 연구 대상으로 삼는 것은 현실적으로 거의 불가능함으로 대부분의 경우 인터넷 트래픽 측정치를 분석하고 이를 통하여 트래픽 모델을 설정하는 방법을 취한다. 따라서 인터넷 트래픽 측정은 인터넷 성능 연구에 매우 중요한 요소인데, 트래픽 측정 방법에는 수동적 측정(passive measurement)과 능동적 측정(active measurement) 방법 등, 크게 두 가지 방법이 널리 사용된다. 수동적 측정 방법은 네트워크의 상태에는 영향을 미치지 않으면서 트래픽을 측정하는 방법으로 라우터나 서버에 전달되는 패킷, 혹은 어느 연결을 통과하는 패킷에 대한 정보를 수집하는 것이다. 따라서 수동적 측정에서 모니터링 사이트(monitoring site)는 수동적인 역할을 담당하며 단지 여러 원격 사이트(remote site)에서 전달되는 패킷의 크기, 시간 등에 대한 정보를 수집한다. 이에 반하여 능동적 측정은 네트워크 트래픽에 최소 간섭을 주면서 네트워크의 상태를 조사하는 방법으로 대부분의 경우 모니터링 사이트에서 일정한 크기의 패킷을 원격 사이트로 보내어 응답이 오는데 걸리는 시간인 RTT(Round Trip Time)를 측정하는 방법을 사용한다. 능동적인 측정 방법은 네트워크에 어느 정도의 간섭을 주는 대신 모니터링 사이트에서 패킷의 크기 및 측정 시간 간격 등을 조절할 수 있는 장점이 있다.

수동적 측정치를 사용한 기존의 연구는 주로 단기간(1분에서 1주일 정도), 단일 연결 상의 데이터 트래픽을 분석하는 연구가 주를 이루고 있는데, 데이터 트래픽의 특성 분석과 이에 대한 모델링은 최근 주요한 연구 영역으로 자리 잡고 있다. 특히 수동적 측정치를 분석하여 패킷 도착 시간 간격에 장거리 상관관계(long-range dependence)가 존재하며, 자기 유사성(self-similarity) 혹은 쪽거리(fractal) 특성이 있음을 밝힌 것은 매우 의미 있는 결과이다[2-4]. 이를 통하여 수동적 측정을 통한 데이터 트래픽은 기존 음성 트래픽(voice traffic, 전화망)의 경우 잘 적용된 포아송 모델(Poisson model)이 적절치 않다는 사실을 규명하였다. 수동적 트래픽의 이러한 특성을 설명하기 위하여 다양한 형태의 모델이 제안되었는데, 자기 유사성을 가지는 확률 과정(stochastic processes) 등에 대한 연구를 통하여 데이터 트래픽의 축척 없는 특성(scale-free characteristics)을 이해하고자 하는 연구가 진행되었다[5-7]. 최근 들어 데이터 트래픽 측정은 보다 큰 규모로 확장되었고, 시계열 형태를 사용한 분석과 더불어 송신과 수신 인터넷 주소에 기초한 분석이 이루어지고 있으며, 특히 네트워크 구조와 트래픽 흐름 간의 상호 관련성에 대한 연구도 시도되었다[8].

본 논문에서는 수동적 데이터 트래픽에서 잘 알려진 장거리 상관관계가 광역 네트워크의 능동적 데이터 트래픽에도 존재하는지 여부를 관련 분석법을 사용하여 검정하고, 그 차이점을 비교하고자 한다. 기존의 수동적 데이터 트래픽의 측정은 주로 단위 시간 당 특정 지점을 통과하는 패킷의 수를 측정하여 구성되는 시계열 데이터에 대하여 분석을 수행하는 반면, 능동적 측정은 두 사이트(모니터링-원격 사이트) 간에 RTT를 측정하여 시계열 데이터를 구성한다. 수동적 측정치에 대한 자기 유사성(self-similarity)과 장거리 상관관계는 여러 가지 방법론을 사용하여 입증된 결과이며, 이러한 특성이 능동적 데이터 트래픽에서도 존재하는지 여부를 밝히는 것이 이 논문의 내용 중 하나이다. 능동적 데이터 트래픽의 분석 결과를 통해 볼 때, 능동적 측정치인 RTT 시계열 자체는 자기 유사성 및 장거리 상관관계를 보이지 않고 $1/f$ 노이즈 현상을 보이며, RTT의 비선형 변환인 변동성은 시간 간격이 짧은 경우 자기 유사성 및 장거리 상관관계를 보이고 시간 간격이 긴 경우에는 장거리 상관관계 혹은 $1/f$ 노이즈 현상이 나타남을 알 수 있었다.

이러한 특성을 규명하기 위하여 본 연구에서는 SLAC (Stanford Linear Accelerator Center)의 IEPM(Internet End-to-end Performance Measurement)에서 PingER (Ping End-to-end Reporting) 프로젝트[9]를 통하여 전 세계 광역 네트워크에 대해 실시한 트래픽 모니터링 데이터를 사용하였다. PingER는 ICMP(Internet Control Management Protocol)에 기반하여 PING(Packet

·INternet Groper)과 유사한 프로그램을 사용하여 RTT (Round Trip Time)를 측정한다. 능동적 측정(active measurement) 중 하나인 RTT는 모니터링 사이트(monitoring site)에서 원격 사이트(remote site)로 ICMP 패킷을 보내어 응답이 오는데 걸리는 왕복 시간을 말한다. 특히 고정된 두 사이트(모니터링 사이트와 원격 사이트) 간에 같은 크기의 ICMP 패킷을 주기적으로 보내어 측정한 RTT 데이터는 시계열 데이터를 이루게 된다. RTT는 주로 지리적으로 고정된 사이트 간에 측정되기 때문에 두 사이트 간의 시간에 따른 네트워크 성능을 분석할 수 있을 뿐만 아니라, 서로 다른 시간에 측정한 RTT의 차이는 지리적 거리에 무관한 양이 되기 때문에 거리에 무관하게 데이터 트래픽을 연구할 수 있는 장점이 있다.

광역 네트워크 트래픽 시계열 분석에 대한 연구는 다양한 각도에서 시도되었다. 트래픽 데이터의 특성을 설명하기 위해 제안된 중첩쪽거리(multifractal) 모델에 대하여 보다 정밀한 통계적 방법론을 적용하여 능동적 데이터 트래픽에서는 중첩쪽거리 현상이 존재하지 않거나 미약하게 존재함을 밝힌 논문[10]과 인터넷 백본(backbone)상의 트래픽을 웨이브릿(wavelet) 분석을 적용하여 작은 시간 영역(1-100 ms)에서 트래픽은 쪽거리 특성이 존재함을 밝힌 논문[11]은 음미할 만하다. 또한 인터넷 트래픽을 설명하기 위한 모델로 서비스 거부(DoS, Denial of Service)와 같은 예외적인 상황(anomaly)을 감지할 수 있는 장거리 상관관계 특성을 가지는 비정규(non-Gaussian) 확률 모델을 제안한 연구[12]가 있으며, 본 논문에서 사용하고자 하는 방법인 DFA(Detrended Fluctuation Analysis)를 사용하여 네트워크상에서 패킷을 전달할 때, 라우팅 전략의 유형에 따라 트래픽 시계열 데이터의 상관관계 정도가 다름을 밝힌 연구[13]도 있다.

과거 관련 연구와 비교해 볼 때 본 논문은 다음과 같은 새로운 특징을 가지고 있다. 첫째, 기존의 RTT 시계열 데이터에 대한 분석은 대부분 RTT 시계열 자체에 대한 분석인 반면 본 연구에서는 시계열 데이터의 비선형 변화량에 대한 연구를 수행하였다. 즉, 본 논문에서는 주식 시장(stock market)에서 유입되는 정보를 측정하는 양인 변동성(volatility) 개념을 적용하여 RTT에 대한 비선형 변화량에 대한 분석을 수행하였다. RTT 분석에서 변동성은 트래픽 데이터간의 고차원 상관관계를 조사할 수 있는 양으로 RTT 시계열 데이터와 함께 광역 네트워크 트래픽의 특성을 연구할 수 있는 중요한 양이다. 둘째, 본 논문에서는 기존 관련 연구에서 사용된 데이터 보다 비교적 장기간(약 3년 6개월) 동안 측정한 RTT를 사용하여 분석하였으며, 또한 분석 결과의 신뢰도를 더하기 위하여 시간에 대하여 지역적으로 존재할 수 있는 경향(trend)을 제거하면서 장거리 상관관계를 분석할 수 있는 DFA를 사용하였다.

2. 트래픽 시계열 데이터의 상관관계

PingER 프로젝트를 통해 측정된 데이터는 단일 모니터링 사이트에서 여러 원격 사이트에 대하여 동시에 RTT를 측정한 것임으로, 이를 사용하여 모니터링 사이트와 원격 사이트 각각에 대하여 RTT 시계열 데이터를 추출하였다. 분석을 위해 사용한 데이터는 pinger.slac.stanford.edu 모니터링 사이트에서 출발하여 IP 주소가 128.178.xxx.xxx와 133.11.xxx.xxx인 두 원격 사이트로 패킷을 보내어 측정한 RTT로 2004년 1월부터 2007년 7월까지 매시간 측정된 각각 30174개와 30225개의 데이터이다. 그림 1은 pinger.slac.stanford.edu에서 128.178.xxx.xxx로 매시간 패킷을 보내어 측정한 RTT 시계열 데이터를 나타낸 것이다. 그림 1을 통하여 볼 때 RTT는 가끔 갑작스러운 변동 현상(bursty phenomena)을 나타내며, 2006년 1월 경 부터 RTT 값이 그 이전에 비하여 증가하는 현상을 보이고 있음을 알 수 있다.

RTT 시계열 데이터에 대한 장거리 상관관계를 분석하기 위한 기법으로 R/S 통계량을 사용한 Hurst 지수[14]가 주로 사용된다. 그러나 Hurst 지수를 사용한 방법은 종종 모순되는 결과를 도출하는 경향이 있음이 밝혀졌음으로[15] 본 논문에서는 DFA(Detrended Fluctuation Analysis)[16]를 사용하고자 한다. DFA는 유동성이 많은 데이터(non-stationary data)에 존재하는 장거리 상관관계를 조사하는 방법으로 주어진 N개의 시계열 데이터를 $M=N/T$개의 부분 집합으로 나눈 후, T개의 데이터로 구성된 각 부분 집합에 대하여 선형 회귀 분석으로 추정한 데이터의 경향(trend)을 제거하여 순수한 상관관계를 살펴보는 방법이다. DFA의 장점은 크게 두 가지로 생각할 수 있다. 첫째, DFA를 사용함으로 지역적 혹은 전역적(global)으로 존재할 수 있는 상관관계와 무관한 경향을 제거할 수 있다. 둘째, DFA는 전역적인 장거리 상관관계뿐만 아니라 지역적으로 존재할 수 있는 상관관계 정도의 차이도 조사할 수 있다. 이것은 DFA를 통해 구하는 지수 값들이 부분 집합의 개수에 따라 달라질 수 있기 때문이며, 부분 집합에 속한 데이터의 개수는 지역적인 성질을 대변하기 때문이다.

주어진 시계열 데이터 $x_1, x_2, \cdots, x_N$에 대하여 M개의 DFA 함수 $F_k(T)$는 다음과 같이 정의 된다.

$$F_k(T) = \sqrt{\frac{1}{T}\sum_{i=kT+1}^{kT+T}(x_i - z_{k,T})^2} \qquad (1)$$

여기서 $k=0,1,2,\cdots,M-1$이며, $z_{k,T}$는 데이터의 개수

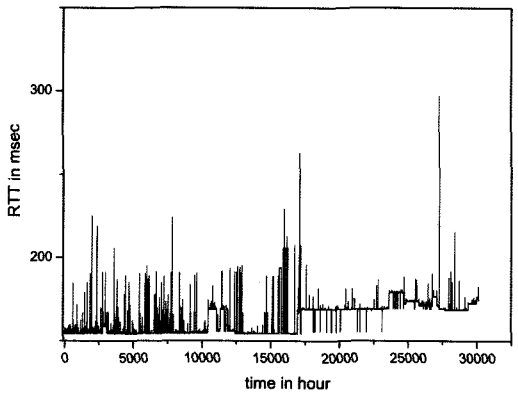

그림 1 원격 사이트의 IP가 128.178.xxx.xxx 인 사이트에 ICMP 패킷을 보내어 매시간 측정한 RTT 시계열 데이터. RTT는 milli-second 단위로 측정되었으며, 가로 축은 매시간 측정된 RTT 순서를 나타낸다.

가 T인 k번째 부분 집합에 속한 데이터들을 선형 회귀 분석으로 추정한 지역적 경향을 나타낸다. 주어진 T에 대하여 모든 k에 대한 $F_k(T)$의 평균을 $<F(T)>$라 하면 부분 집합의 섭동(fluctuation)을 구할 수 있다. 일반적으로 $<F(T)>$는 데이터 개수 T가 클수록 증가하며 만약

$$<F(T)> \propto T^{\alpha} \text{ 혹은 } \log<F_k(T)> \propto \alpha \log T \quad (2)$$

의 관계가 존재하면 고려하는 시계열 데이터 사이에 축척(scaling) 관계가 존재하며, 지수 α값에 따라 장거리 상관관계 특성이 결정된다. 만약 $\alpha=0.5$이면 시계열 데이터는 소위 백색 노이즈(white noise)라 불리며 서로 독립이며, 따라서 장거리 상관관계가 존재하지 않는다. 그러나 $0.5<\alpha<1.0$이면 시계열 데이터 사이에는 멱급수(power-law)로 표현되는 장거리 상관관계가 존재하며, $0<\alpha<0.5$이면 장거리 반상관관계(anti-correlation)가 존재한다. 특히 $\alpha=1.0$인 경우에는 $1/f$ 노이즈(noise)라 불리며, $\alpha=1.5$인 경우는 백색 노이즈를 적분한 것으로 브라운 노이즈(brown noise)라 부른다. 또한 $\alpha \geq 1.0$이면 상관관계가 존재하나 멱급수적으로 표현될 수 있는 것은 아니다[17].

그림 2는 위의 두 RTT 시계열 데이터에 대하여 DFA 분석을 실행한 결과이다. 그림 2를 통하여 알 수 있듯이 원격 사이트의 IP 주소가 128.178.xxx.xxx인 경우에는 $\alpha \approx 1.01$로 $1/f$ 노이즈 현상[18]을 나타내고 있다. 플리커 노이즈(flicker noise)라고도 불리는 $1/f$ 노이즈는 물리, 생물, 전기전자, 심지어 경제계에서 널리 나타나는 현상이다. 특히 DFA에서 지수 α는 시계열 데이터의 울퉁불퉁한 정도(roughness)를 나타내는 것으로 간주할 수 있는데, 이런 측면에서 보면 $1/f$ 노이즈를 보이는 RTT는 완전히 예측 불가능한 백색 노이즈와 요철이 없고 평탄한 브라운 노이즈의 중간 형태를 띠고 있다고 할 수 있다. $1/f$ 노이즈에 대한 많은 연구가 진행되고 있으나 아직까지 $1/f$ 노이즈를 생성하는 간단한 수학적 모델은 존재하지 않는 실정이며, 대부분의 경우 특별한 경우에 한정된 모델이 존재한다. 따라서 $1/f$ 노이즈는 아직 매우 중요한 연구 주제로 남아있다.

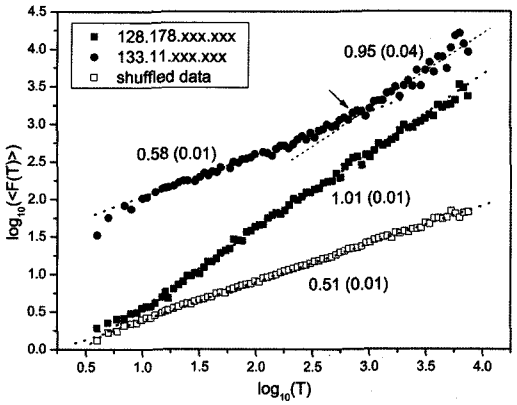

그림 2 두 시계열 데이터에 대하여 DFA를 수행한 결과. ■는 원격 사이트의 IP 주소가 128.178.xxx.xxx인 경우이며, ●는 원격 사이트의 IP 주소가 133.11.xxx.xxx인 경우이고 화살표는 교차점을 나타낸다. 또한 □는 시계열 데이터를 무작위로 섞어서 생성한 시계열 데이터에 대한 DFA 수행 결과이다. 그림에서 점선들은 기울기 α를 추정한 것으로 숫자들은 추정한 기울기 α를 나타내며 추정치에 대한 오차는 괄호 안에 표현하였다.

원격 사이트의 IP 주소가 133.11.xxx.xxx의 경우에는 T값에 따라 기울기 α가 다르며 따라서 교차점(cross-over)이 존재한다. 교차점 아래에는 $\alpha \approx 0.58$로 RTT 시계열 데이터들은 거의 독립적이거나 미약한 장거리 상관관계가 존재한다고 할 수 있으나, 교차점 위쪽에서는 $\alpha \approx 0.95$임으로 장거리 상관관계를 나타내거나 혹은 거의 $1/f$ 노이즈 현상이 나타남을 볼 수 있다. 또한 위에서 발견한 장거리 상관관계와 $1/f$ 노이즈가 의사(擬似)적으로 형성된 것이 아님을 입증하기 위하여 RTT 시계열 데이터를 무작위로 섞어서 생성한 새로운 시계열 데이터에 대하여 DFA를 수행하였다. 무작위로 섞은 시계열 데이터 사이에는 상관관계가 존재하지 않아야 하는데, 그림 2의 결과를 통해 볼 때 기울기가 $\alpha \approx 0.51$임으로 무작위로 섞은 시계열 데이터 사이에는 장거리 상관관계가 존재하지 않고 서로 독립적임을 알 수 있다. 따라서 RTT 시계열 데이터에 존재하는 $1/f$ 노이즈와

장거리 상관관계는 의사(擬似)적으로 형성된 것이 아님을 알 수 있다.

3. 변동성(volatility)을 통한 트래픽의 상관관계

광역 네트워크에서 데이터 트래픽과 주식 시장(financial market)의 주가 변동은 많은 컴퓨터 혹은 증권 투자자(traders)들이 복잡한 상호작용을 하며 외부의 영향(패킷 혹은 주식 시장 정보)에 대해 반응하는 복잡한 시스템이라는 공통점을 가지고 있다. 주식 시장의 주가 변동에 관한 연구에서는 소위 변동성(volatility)[19]라 불리는 양을 통해 주식 시장의 섭동과 이에 따른 상관관계가 중요한 연구 주제이다. 변동성은 주식 시장에 유입되는 정보를 정량화하기 위하여 도입된 정량적인 측도(measure)로 주식 시장에 유입되는 정보의 양이 크면 주식 거래량 역시 증가하는 특징을 가지고 있다. 휘발성이라고도 불리는 변동성은 일반적으로 일정 기간 동안 발생하는 주가 변동에 대한 절대값으로 정의된다. 광역 네트워크의 데이터 트래픽에 대한 변동성 역시 주식 시장의 변동성과 유사하게 정의할 수 있으며, 이 변동성은 절대값으로 정의되기 때문에 트래픽의 고차원 상관관계를 조사할 수 있는 양이 된다.

RTT 시계열 데이터 $x_1, x_2, \cdots, x_N$의 시간 간격 τ에 대한 변동성 $v_\tau(i)$를 아래 식 (3)과 같이 정의하면, $N-\tau$개로 구성된 변동성을 유도할 수 있으며 변동성 역시 시계열 데이터를 이룬다. 즉, 각 $v_\tau(i)$는 τ개의 RTT 변화의 절대값에 대한 평균으로 아래와 같이

$$v_\tau(i) = \frac{1}{\tau}\sum_{k=1}^{\tau}|x_{i-\tau/2+k} - x_{i-\tau/2+(k-1)}| \qquad (3)$$

주어지며, 여기서 $i=\tau/2+1, \tau/2+2, \cdots, N-\tau/2$이다. 이때 변동성은 광역 네트워크에서 트래픽의 고차원 변동을 나타내며 트래픽에 대한 섭동을 추정할 수 있는 양이 된다.

위의 두 원격 사이트에 대한 RTT 데이터를 사용하여 변동성 $v_\tau(i)$를 구하여 확률 분포를 추정하였으며, 원격 사이트가 128.178.xxx.xxx인 경우 그 결과를 그림 3에 나타내었다. 그림 3은 서로 다른 τ에 대하여 로그정규화된 변동성(log-normalized volatility)에 대한 확률 분포를 나타낸 것으로 로그정규화된 변동성은 표준정규분포를 따름을 알 수 있다. 서로 다른 τ에 대하여 v_τ를 $v_\tau \to (\ln v_\tau - \mu_\tau)/\sigma_\tau$로 로그정규화하면 v_τ의 확률분포 $P(v_\tau)$는 $P(v_\tau) \to \sigma_\tau P(v_\tau)$로 치환되고, 만약 로그정규화된 확률변수 $(\ln v_\tau - \mu_\tau)/\sigma_\tau$가 표준정규분포를 따르면 v_τ는 로그정규분포(log-normal distribution)[20]를 따른다. 따라서 변동성 v_τ에 대한 확률 분포 $P(v_\tau)$는 τ에 무관하게 식 (4)와 같은 로그정규분포를 따름을 알 수 있다.

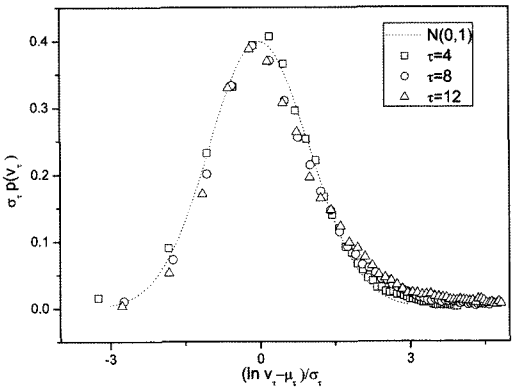

그림 3 원격 사이트 IP 주소가 128.178.xxx.xxx의 경우 서로 다른 $\tau=4, 8, 12$에 대한 변동성을 로그정규화한 확률변수의 확률분포. 여기서 μ_τ와 σ_τ 값은 최대우도추정(maximum likelihood estimation) 방법을 사용하여 추정한 값을 사용하였고, 점선은 표준정규분포를 나타낸다.

$$P(v_\tau) = \frac{1}{\sqrt{2\pi}\,\sigma_\tau v_\tau} e^{-(\ln v_\tau - \mu_\tau)^2/2\sigma_\tau^2} \qquad (4)$$

여기서 μ_τ와 σ_τ는 각각 확률 변수 $\ln v_\tau$에 대한 평균과 표준편차를 나타낸다. 원격 사이트의 IP 주소가 133.11.xxx.xxx인 경우에도 변동성의 확률 분포는 로그정규분포를 따름을 알 수 있었다. 변동성에 대한 로그정규분포는 비단 광역 네트워크의 RTT에서만 나타나는 현상이 아니라 주식 시장과 지역 네트워크 트래픽에서도 발견되는 현상이다[4,21,22]. 따라서 변동성에 대한 로그정규분포는 매우 광범위한 현상으로 이에 대한 심층적 분석과 로그정규분포를 따르는 확률 과정(stochastic process)에 대한 연구가 필요하다.

변동성 시계열 데이터에 대한 장거리 상관관계를 살펴보기 위하여 식 (1)과 (2)를 사용하여 DFA를 적용하였으며, 원격 사이트의 IP 주소가 128.178.xxx.xxx 인 경우의 결과는 그림 4와 같다. 그림 4를 통해 볼 때, 모든 τ에 대하여 $\log\langle F_k(T)\rangle$와 $\log T$ 간에 T값에 따라 기울기 α가 달라지는 교차점(crossover)이 존재한다. 교차점은 모든 τ에 대하여 거의 동일하며 약 $T \approx 500$에서 일어남을 알 수 있다. $T \le 500$에서 지수 α는 $0.68 \le \alpha \le 0.72$임으로 장거리 상관관계가 존재하며 상관관계 정도를 나타내는 지수 α는 τ가 클수록 증가하는 추세를 보이고 있다. 또한 $T \ge 500$에서 변동성은 $1/f$ 노이즈 성질을 가지고 있는데, 이것은 $T \ge 500$의 영역에서 변동성은 완전한 무작위(random) 상태와 매우 평탄한(smooth) 변화 사이에 존재함을 의미한다.

원격 사이트의 IP 주소가 133.11.xxx.xxx 인 경우의

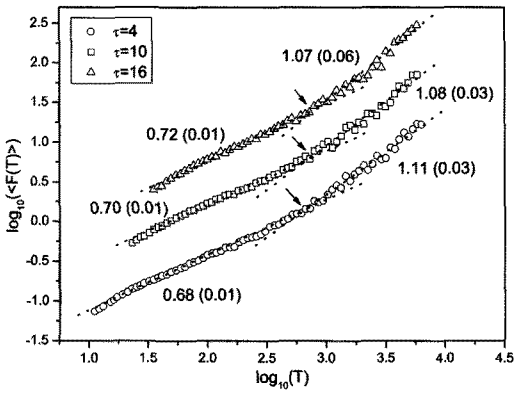

그림 4 원격 사이트 IP 주소가 128.178.xxx.xxx 인 변동성 시계열 데이터에 대하여 서로 다른 τ에 따른 DFA 분석 결과. 그림에서 점선들은 기울기 α를 추정한 것으로, 숫자들은 추정한 기울기 α를 나타내며 추정치에 대한 오차는 괄호 안에 표현하였다.

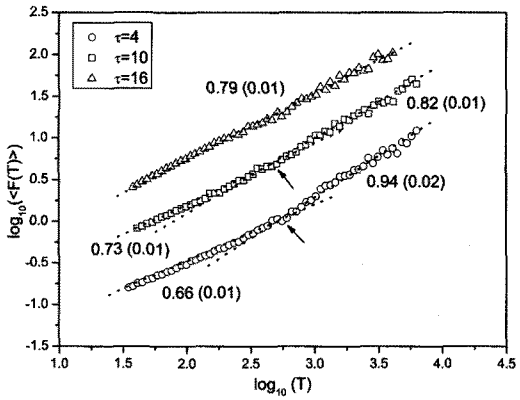

그림 5 원격 사이트 IP 주소가 133.11.xxx.xxx 인 변동성 시계열 데이터에 대하여 서로 다른 τ에 따른 DFA 분석 결과. 그림에서 점선들은 기울기 α를 추정한 것으로, 숫자들은 추정한 기울기 α를 나타내며 추정치에 대한 오차는 괄호 안에 표현하였다.

DFA 결과는 그림 5에 나타내었다. 그림 5에서 볼 수 있듯이 비교적 작은 τ의 경우 교차점이 존재하며, 교차점의 위치는 128.178.xxx.xxx 의 경우와 마찬가지로 약 $T \approx 500$ 에서 일어남을 알 수 있다. 또한 교차점을 기준으로 장거리 상관관계 정도가 달라지는데, $T \leq 500$에서는 장거리 상관관계가 존재하며, $T \geq 500$ 에서 변동성은 128.178.xxx.xxx 인 경우와 달리 장거리 상관관계가 존재하거나 거의 $1/f$ 노이즈 성질($\alpha \approx 0.94$)을 가지고 있다. 비교적 큰 τ의 경우에는 교차점이 존재하지 않으며 모든 T값에 대하여 장거리 상관관계가 존재한다.

변동성 시계열 데이터가 로그정규분포를 따르며 장거리 상관관계 혹은 $1/f$ 노이즈 특성이 존재한다는 결과는 매우 중요한 의미를 내포하고 있다. 일반적으로 로그정규분포를 따르는 확률변수는 승법과정(multiplicative process)을 통해 생성될 수 있으나, 승법과정을 통해 생성된 값들이 장거리 상관관계 혹은 $1/f$ 노이즈 특성을 가지고 있는 모델에 대한 연구가 아직 부족하며 이러한 메커니즘이 규명되지 않은 실정이다. 다만 변동성은 유동적(non-stationary)이고 시간에 의존하기 때문에 전역적으로는 안정적이고 유동적이지 않으나 지역적으로는 고정되지 않는 변동을 가지는 데이터를 설명할 수 있는 모델인 ARCH[23] 혹은 GARCH[24] 등을 사용한 모델을 통하여 설명될 수 있는 가능성이 있다.

4. 요약 및 결론

본 논문에서는 대표적인 능동적 측정치인 RTT를 사용하여 광역 네트워크 트래픽의 특성을 시계열 차원에서 분석하였다. PingER 프로젝트를 통하여 측정된 RTT 데이터를 DFA를 적용하여 분석한 결과 RTT 시계열 데이터는 장거리 상관관계를 나타내거나 혹은 $1/f$ 노이즈의 특성을 보였다. 또한 광역 네트워크 트래픽에 내재한 고차원 상관관계를 조사하기 위하여 주가 변동에서 널리 사용되는 고차원 변화량인 변동성을 네트워크 트래픽에서도 유사하게 정의하여 변동성의 특성을 분석하였다. 고차원 변화량으로 정의된 변동성은 로그정규분포를 따르며 또한 변동성에 대한 장거리 상관관계는 고려하는 시간 간격에 따라 다른 특성을 가지고 있음을 밝혔다. 변동성은 상대적으로 짧은 시간 간격에서는 장거리 상관관계가 존재하였고, 장거리 상관관계 정도를 나타내는 지수 α는 τ가 클수록 증가하는 추세를 보였으며, 상대적으로 긴 시간 간격에서는 장거리 상관관계 혹은 $1/f$ 노이즈를 현상을 나타내고 있음을 알 수 있었다. 따라서 장거리 상관관계는 비단 지역 네트워크 트래픽뿐만 아니라 광역 네트워크 트래픽에서도 나타나는 특징이며, 특히 광역 네트워크에서는 $1/f$ 노이즈를 현상이 존재함을 알 수 있었다.

RTT와 변동성 시계열 데이터가 장거리 상관관계와 $1/f$ 노이즈 현상을 보이는 것은 광역 네트워크를 구성하고 있는 컴퓨터와 라우터들 간에 비선형적 상호 작용이 존재함을 의미한다. 이러한 특성과 변동성의 확률 분포가 로그정규분포라는 사실을 적용하여 광역 네트워크의 RTT를 이해할 수 있는 모델에 대한 연구는 향후 광역 네트워크의 트래픽을 이해하는데 매우 중요한 역할을 담당할 것으로 예상된다.

대부분의 경우 능동적 측정은 지리적으로 멀리 떨어져 있는 두 사이트(모니터링-원격 사이트) 간의 패킷을

전송하는데 걸리는 시간인 RTT를 측정하기 때문에 능동적 측정에 대한 분석 결과는 광역 네트워크 내부의 트래픽 상태를 조사할 수 있는 '탐침'(probe) 역할을 할 수 있다. 이것은 인터넷을 통한 데이터 통신의 안정성과 효율성을 가름할 수 있는 양적인 지표로 사용될 수 있다는 것을 의미한다. 특히 동일한 모니터링-원격 사이트 간에 시간에 따라 측정된 RTT는 트래픽의 정체 정도에 관한 정보를 담고 있기 때문에 RTT에 대한 분석은 컴퓨터 네트워크의 대기 시간(latency) 정도를 가늠하여 정체 현상을 개선하는 연구에 활용될 수 있으며, 네트워크의 유효성(throughput) 향상 및 패킷 손실(packet loss)을 감소시키는 네트워크 설계에도 활용될 수 있다.

광역 네트워크는 중앙 통제를 받지 않는 열린 시스템(open system)으로 광역 네트워크에서 데이터 트래픽에 대한 연구는 아직까지 초보적인 단계로 체계적인 연구가 부족한 실정이다. 이런 점을 감안할 때 데이터 트래픽의 특성에 관한 본 연구 결과는 향후 차세대 인터넷 개발의 중요한 기초 연구 자료로 사용될 수 있다. 특히 이 분야의 연구는 광역 네트워크 구조 측면과 연관되어 효율적 네트워크 구조에 대한 연구로 이어질 수 있으며, 나아가 효율적인 네트워크 구축에도 본 연구 방법론은 적용될 수 있다. 또한 본 연구를 통해 밝힌 장거리 상관관계와 1/f 노이즈 현상은 카오스(chaos) 및 비선형 동력학에서 자주 나타나는 현상으로 타 학문과 연계 측면에서도 학술적 가치가 있음으로, 향후 이러한 측면에서 연구가 더욱 진행되어야 할 것으로 생각된다.

참 고 문 헌

[1] Paxson, V., "Growth trends in wide-area TCP connections," *IEEE Network*, vol.8, no.4, pp.8-17, 1994.

[2] Leland, W., Taqqu, M., Willinger, W., and Wilson, D., "The Self-similar nature of Ethernet traffic," *IEEE/ACM Trans. Net.*, vol.2, no.1, pp.1-15, 1994.

[3] Paxson, V. and Floyd, S., "Wide-Area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Net.*, vol.3, no.3, pp.226-244, 1995.

[4] Lee, C., "Higher-Order Correlations in Data Network Traffic," *J. Korean Phys. Soc.*, vol.45, no.6, pp.1664-1670, 2004.

[5] Willinger, W., Taqqu, M., Sherman, R., and Wilson, D., "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Trans. Net.*, vol.5, no.1, pp.71-86, 1997.

[6] Takayasu, M., Takayasub, H., and Fukudac, K., "Dynamic phase transition observed in the Internet traffic flow," *Physica A*, vol.277, pp.248-255, 2000.

[7] Nakayama, A., "Critical behaviors and self-similarity in models of computer network traffic," *Physica A*, vol.293, pp.285-296, 2001.

[8] Claffy, K., "Internet measurement and data analysis: topology, workload, performance, and routing statistics," *NAE '99 workshop*, 1999; available at http://www.caida.org/outreach/papers/1999/Nae

[9] http://www-iepm.slac.stanford.edu/pinger/

[10] Veitch, D., Hohn, N., Abry, P., "Multifractality in TCP/IP traffic: the case against," *Computer Networks*, vol.48, pp.293-313, 2004.

[11] Ribeiro, V., Zhang, Z., Moon, S., Diot, C., "Small-time scaling behavior of Internet backbone traffic," *Computer Networks*, vol.48, pp.315-334, 2005.

[12] Scherrer, A., Larrieu, N., Owezarski, P., Borgnat, P., Abry, P., "Non-Gaussian and Long Memory Statistical Characterizations for Internet Traffic with Anomalies," *IEEE Transactions of dependable and secure computing*, vol.4, pp.56-70, 2007.

[13] Xiao-Yan, Z., Zong-Hua, L., Ming, T., "Detrended Fluctuation Analysis of Traffic Data," *Chinese Phys. Lett.*, vol.24 pp.2142-2145, 2007.

[14] Beran, J., *Statistics for Long-memory Processes*, Chapman & Hall/CRC, 1994.

[15] Karagiannis, T., Molle, M., and Faloutsos, M., "Long-Range Dependence: Ten Years of Internet Traffic Modeling," *IEEE Internet Computing*, vol.8, no.5, pp.57-64, 2004.

[16] Peng, C. et al., "Mosaic organization of DNA nucleotides," *Phys. Rev. E.*, vol.40, no.2, pp.1685-1689, 1994.

[17] Peng, C. et al., "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," *Chaos*, vol.5, no.1, pp.82-87, 1995.

[18] Bak, P., Tang, C., and Wiesenfeld, K., "Self-organized criticality: An explanation of the 1/f noise," *Phys. Rev. Lett.*, vol.59, no.4, pp.381-384, 1987.

[19] Müller, U. et al., "Volatilities of different time resolutions - Analyzing the dynamics of market components," *J. Empirical Finance*, vol.4, pp.213-239, 1997.

[20] Hogg, R. and Craig, A., *Introduction to mathematical statistics*, 4th Ed., Macmillan Publishing Co., New York, 1978.

[21] Liu, Y., Gopikrishnan. P., Cizeau, P., Meyer, M., Peng, C-K., and Stanley, H. E., "Statistical properties of the volatility of price fluctuations," *Phys. Rev. E.*, vol.60, no.2, pp.1390-1400, 1999.

[22] Lee. C., "Characteristics of the volatility in the Korea composite stock price index," *Physica A*, vol.388, pp.3837-3850, 2009.

[23] Engle, R., "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, vol.50, no.4, pp.987-1007, 1982.

[24] Bollerslev, T., "Generalized autoregressive conditional heteroskedasticity," *J. Econometrics*, vol.31, no.3, pp.307-327, 1986.

이 창 용

1983년 서울대학교 계산통계학과 졸업(이학사). 1995년 미국 텍사스 주립대학교(Univ. of Texas at Austin) 물리학과 졸업(이학박사). 1996년~1998년 한국전자통신연구원 선임연구원. 1998년~2007년 공주대학교 산업정보학과 교수 2007년~현재 공주대학교 산업시스템공학과 교수. 관심분야는 진화 연산 알고리즘 및 최적화 문제, 복잡계 네트워크(complex networks), 생물정보학(bioinformatics) 등