인터넷 트래픽 분석과 억세스 네트워크 모델링

박형구[+], 장주욱[+], 김우현[*], 도진숙[*], 이경근[*], 이효진 [°]

[+]서강대학교 전자 공학과

[*]세종대학교 정보통신학과

[°] LG 텔레콤(주)

Internet Traffic Analysis and Access Network Modeling

Hyeong-Koo Park[+], Ju-wook Jang[+], Woo-Hyun Kim[*], Jin-Sook Do[*], Kyung-Geun Lee[*], Hyo-Jin Lee [°]

[+]Dept. of Electronic Engineering, Sogang Univ.

[*]Dept. of Information and Communications Engineering, Sejong Univ.

[°]LG Telecom

요약

최근 인터넷에서 트래픽의 유형과 크기가 WWW등 새로운 어플리케이션(Application)의 발전으로 급격히 변화하고 있기 때문에, 인터넷 트래픽의 특성에 대한 여러 가지 분석모델들이 공존하고 있다. 본 연구에서는 억세스(Access) 네트워크에 나타나는 트래픽의 통계적 특성을 정확히 예측하고, 사용하는 어플리케이션에 따라 트래픽의 통계적 특성이 달라짐을 기존 연구 및 실측 자료를 통해 보였다.

I. 서론

인터넷 트래픽의 통계적 특성을 정확히 예측, 분석하여 사용자와 공급자들의 요구를 정확하게 수용할 수 있는 억세스망 용량을 산출하는 것은 매우 중요하다. 억세스 망에서의 트래픽을 모델링하기 위해서는 큐잉 모델링, 패킷 수준의 모델링, 어플리케이션 수준의 모델링등의 방법이 있다.

억세스 망 트래픽을 예측하기 위해서는 다중화 이득의 관점에서 접근할 필요가 있다. 지금까지 발표된 연구들은 개개인의 접속 요구가 대수정규분포(Lognormal)를 가진다는 데 공감하고 있다, 그러나 중첩된(aggregate) 트래픽의 확률 분포에 대해서는 자기 유사성(Self-Similar) 분포를 가진다는 주장과 포아송(Poisson) 분포를 가진다는 주장으로 나뉘어져 있다[1,2,3,4,5,6].

이와 같이 상호 모순되는 주장들이 공존하는 이유는 WWW(World Wide Web)의 급격한 성장으로 말미암아, 인터넷의 트래픽 특성이 급격히 변하고 있기 때문이다. 즉, 1990년대 후반까지의 연구 결과에 의하면 인터넷 트래픽이 자기 유사성을 보인다는 주장이 주를 이루었지만, 최근에는 트래픽이 포아송 분포를 이룬다거나 포아송과 자기 유사성의 결합의 형태를 가진다는 연구들이 나타나고 있다.

본 연구에서는, 혼잡 제어의 영향을 받는 TCP 패킷의 흐름을 기록하는 패킷 수준에서의 트래픽 모델링과, 파일의 크기 분포, TCP 연결 수 분포, TCP 연결 유지 시간 분포, RTT의 분산 등을 이용하여 어플리케이션 수준에서 모델링하는 방법들을 사용하여, 억세스 망에서의 트래픽을 예측, 분석해 보았다. 그런데 트래픽 모델링에 있어서 패킷 수준의 모델링은 패킷 트레이스의 패턴이 트레이스가 발생할 시점의 트래픽 상황에 큰 영향을 받는 문제점이 있을 수 있다. 반면에 어떤 크기의 파일을 언제 전송하였는가 하는 어플리케이션 수준의 기록은 네트워크 상황의 영향을 거의 받지 않기 때문에, 어플리케이션 수준에서의 모델링도 사용함으로서 패킷 수준에서의 모델링의 단점을 보완하였다.

II. 인터넷 트래픽 모델링

2-1. 기존 연구의 비교 및 분석

억세스 망에서의 트래픽 패턴에 관한 기존 연구들은 대부분 트래픽이, 시계열의 확대나 이동에 상관없이 통계적 특성을 유지하는 자기유사성을 보인다라고 주장하고 있다. 만약 기존 연구들의 주장이 맞다면, 억세스 망에서 예상되는 트래픽의 높은 분산을 처리하기 위해서, 최번시(Peak Time)의 평균 대역폭 사용량보다 상당히 높은 억세스 망 용량을 설계해야만 할 것이다. 따라서 억세스 라우터들을 연결하는 백본 망의 용량도 증가하게 되어 상당히 많은 초기 투자가 필요하게 될 것이다.

그러나 기존 연구들은 대부분 WWW의 사용이 본격적으로 사용되기 이전의(1990년대 중반) 트래픽을 측정하여 나온 결과라는 데에 주목할 필요가 있다. 최근에 발표된 연구 결과는 웹 트래픽의 경우 포아송 분포를 보인다고 보고하

면서, 그 근거로 짧은 시간 간격 내에서는 사용자들의 접근 패턴이 서로 상관관계가 낮다는 점을 들었다[1]. 또 한 연구에서는 대량의 데이터를 보내는 FTP와 같은 어플리케이션은 TCP의 혼잡 제어 프로토콜의 영향으로 인한 동기화에 의해 버스트(Burst)한 트래픽을 보여 분산이 높게 나타나지만, 소량의 데이터를 짧은 시간에 전송하는 웹 트래픽은 비교적 평탄한 트래픽 패턴을 보인다고 주장하고 있다[2]. 후자의 주장을 따를 경우에는 억세스 망 용량과 백본망 용량이 줄어들어 경제적인 네트워크 구성이 가능해 질 것이다.

억세스 망의 트래픽을 정확히 추정하는 것은 네트워크 구성 비용(공급자), 서비스 품질(사용자) 측면에서 매우 중요한 문제이므로 관련 연구들을 자세히 살펴보고, 현재 또는 미래에 예상되는 억세스망 트래픽 추정에 응용하고자 한다.

2-2. 자기 유사성과 포아송 분포 비교

인터넷 트래픽의 분포에 대해서 기존에 발표된 내용은 크게 두가지로 종합할 수 있다. 첫 번째는 인터넷 트래픽이 자기 유사특성을 보인다는 것이고, 두 번째는 포아송 분포를 이룬다는 주장이다.

인터넷 트래픽이 자기 유사특성을 보이는 경우와 포아송 분포를 이룰 때, 시간 t와 t+k에서의 트래픽량 간의 상관계수는 다음 식으로 모델링 할 수 있다.

자기 유사성 모델링: $C_S(t,t+k)=|k|^{-b}$ (1)

포아송 분포 모델링: $C_P(t,t+k)=a^{|k|}\ \ (0<a,b<1)$ (2)

식 (1), (2)를 살펴보면, 시간 k가 증가할수록 자기 유사특성 을 보이는 경우가 포아송 분포를 보이는 경우보다 시간 t, $t+k$에서의 트래픽간 상관관계가 더욱 강하게 나타난다는 것을 알 수 있다. 예를 들어 k=4, a=b=0.5라 할 때, 트래픽 간의 상관계수는, 자기 유사성인 경우에는 1/4 이고, 포아송 분포의 경우는 1/16 이 되어 자기 유사성을 보이는 경우의 상관계수가 포아송 분포인 경우의 상관계수 보다 높음을 알 수 있다. 또한 트래픽간에 시간상의 상관관계가 높은 자기 유사특성인 경우, 더욱 트래픽이 버스트해 진다는 것을 알 수 있다.

2-3. 트래픽이 자기 유사성을 보인다는 주장

Willinger와 Paxon은 자기 유사특성을 모델링 할 때 많이 사용하는 프랙탈 함수와 포아송 모델을 시뮬레이션 한 결과를 실제 인터넷 트래픽 자료(1995년도 트래픽 자료)와 비교함으로써 트래픽이 자기 유사특성을 가진다는 것을 주장하고 있다[3].

그 외에도, 최근에 측정한 인터넷 트래픽을 근거로 인터넷 트래픽의 자기 유사성이 병목 구간의 공유를 통해 인터넷 전체에 확산된다는 연구도 있다[4]. 이 연구는 트래픽간의 상관관계가 천천히 감소하는 이른바 장기간 의존성(Long-Range Dependence) 현상이 병목 구간을 통해 인터넷 트래픽에 전파된다는 것을 FTP 실험과 시뮬레이션을 통해 증명하고 있다.

병목 구간을 흐르고 있는 배경 트래픽과, 그 구간을 지나가는 TCP 트래픽간의 관계는 식 (3)으로 나타낼 수 있다.

$$TCP(t) = C - B(t)$$
$$B(t)=\sum_{i=1}^{N} B_i(t) \qquad (3)$$

여기서 C는 병목 구간의 총 대역폭을 의미하고 $B(t)$는 많은 수의 Connection들 ($B_i(t)$, i=1,2,…,N)에 의해서 발생된 트래픽들이 이루는 배경 트래픽을 의미한다. 이때 병목 구간을 공유하는 TCP 연결 TCP(t)는 TCP의 혼잡제어 프로토콜의 특성상 가능한 많은 네트워크 용량을 차지하려 하게 된다. 따라서 FTP 어플리케이션을 위해 맺은, TCP 연결의 트래픽은 병목 구간의 총 용량에서 배경 트래픽이 차지하고 있는 용량을 뺀 만큼을 모두 차지하려 하게 된다.

실제 트래픽 측정 결과와 NS2(Network Simulator)를 이용한 시뮬레이션 결과로부터, 식 (3)처럼 TCP 트래픽이 배경 트래픽의 특성을 잘 따라가고 있다는 것을 알 수 있다[4]. 따라서 어떤 병목 구간을 지나는 여러 트래픽들이 모인 배경 트래픽 $B(t)$가 자기 유사특성을 보인다면, 그곳을 지나가는 TCP 연결에 의한 트래픽도 자기 유사성을 보이게 될 것이다. 더 나아가서, 병목 구간을 지나면서 자기 유사특성을 가지게 된 TCP 트래픽이 네트워크의 다른 지역을 지나면서, 마찬가지 방식으로 다른 트래픽들에게 자기 유사성을 옮기게 될 것이다. 그 결과 전체 인터넷 트래픽들이 자기 유사성을 가지게 된다는 것이다.

2-4. 트래픽이 포아송 분포라는 주장

Morris와 Lin의 연구는 하바드 대학과 루슨트 연구소에서 측정한 트래픽(1998년도)중 웹 트래픽만을 고려한 연구이다. 인터넷이 급속한 성장하면서 웹이 차지하고 있는 비중이 상당히 커지고 있고, 웹이 여러 가지 인터넷 서비스들을 통합해 가고 있는 추세이므로 웹 트래픽을 분석 한 것이 큰 의미를 가지게 된다.

인터넷 트래픽이 포아송 분포를 이룬다는 주장의 근거이론은 식 (4)와 같다.

$$Var(X)=\sum_{i}^{N}\sum_{j}^{N} Cov(Xi,Xj) \qquad (4)$$

식 (4)에서 각 사용자간의 공분산(Covariance) 값은 사용자간의 상관관계가 상당히 낮다면 0의 값에 수렴하게 될 것이고, 사용자간의 상관관계가 높다면 1이란 값에 수렴하게 될 것이다. 따라서 식 (4)에 의해서 사용자간의 상관관계가 낮을 때는 X의 분산이 $O(N)$의 형태로 증가하는 포아송 분포를 이루고, 사용자간의 상관관계가 높을 때는 X의 분산이 $O(N^2)$의 형태로 증가하는 자기 유사특성을 보이게 된다.

또한 Morris와 Lin은 하바드 대학에서 측정한 웹 트래픽에서 대역폭과 분산과의 관계를 분석해 내었다[1]. 분석 결과를 보면 트래픽의 분산이 $O(N^2)$으로 증가하는 것이 아니라, $O(N)$의 형태로 선형 증가함을 알 수 있다. 따라서 TCP 연결 유지 시간이 짧고 전송 파일 크기가 비교적 작은 웹 트래픽은 자기 유사 특성을 보이지 않고, 포아송 분포를 이룬다는 것을 증명하고 있다.

2-5. 어플리케이션의 구성에 따라 다르다는 주장

인터넷 트래픽이 자기 유사성 또는 포아송하다는 주장과

는 달리, 어플리케이션에 따라 달라질 수 있다는 연구 결과들이 최근 발표되고 있다.[2, 5]

You와 Chandra의 연구는 트래픽을 이루는 어플리케이션의 구성비에 따라서 트래픽의 특성을 달리 봐야 한다는 것이다[5]. 연구 결과 웹 트래픽은 time-stationary한 성질을 보였고, 반면에 FTP 트래픽이 강한 non-stationarity를 나타낸 다는 것을 밝혔다.

Joo와 Ribeiro의 연구는 어플리케이션 별로 전송하는 파일 크기의 확률 분포가 다르기 때문에, 트래픽 특성도 다르게 나타난다는 것을 보이고 있다[2]. 즉, 웹 트래픽의 경우 작은 TCP 전송이 여러 번 행해지고 각 TCP 연결은 slow start중에 있을 가능성이 높아 가용 대역폭을 보다 잘 차지할 수 있다는 것이다. 반면에 주로 대용량의 파일을 전송하는 FTP 트래픽의 경우는 혼잡 회피(congestion avoidance) 상태에 있을 가능성이 높아 대역폭을 최대한 이용하기까지 시간이 걸리게 된다. 결국 대용량의 FTP 트래픽은 트래픽에 큰 사이클을 일으켜서 대역폭 활용률이 떨어지는 자기 유사특성을 보이게 되는 반면에, 웹 트래픽은 잔잔한 트래픽의 변화로 여유 대역폭을 효과적으로 차지하여 대역폭 활용률이 높아지고 분산이 비교적 작은 특징을 보인다는 것이다.

III.트래픽의 측정과 분석

3-1. 사용자수와 대역폭과의 관계 분석

그림 1은 서강대 라우터에서 2000년 4월 19일 오후 5시부터 4월 24일 오전 7시까지 총 111시간 동안 측정한 트래픽 자료를 가지고 사용자수와 대역폭간의 관계를 분석해 본 것이다. 서강대 트래픽의 경우 사용자수와 대역폭간의 상관계수는 0.83이란 높은 수치가 나와서, [1]의 연구 결과처럼 두 요소 사이에 높은 상관관계가 있음을 확인할 수 있었다. 실험 결과로부터 사용자수와 대역폭간에는 높은 상관관계가 있으므로, 사용자수의 예측으로서 억세스 망의 용량을 결정할 수 있다는 실질적인 의미를 찾을 수 있다.

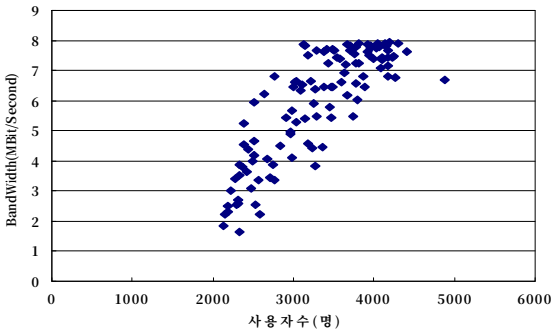

그림 1. 사용자수와 대역폭간의 관계(서강대)

3-2. RTT의 측정을 통한 트래픽 분석

패킷 트레이스를 통한 트래픽 분석을 보완하기 위해서 RTT(Round Trip Time)의 측정을 통한 트래픽 분석을 해 보았다. RTT를 측정하기 위해서 서강대 망의 Subnet에 있는 Linux 환경에서 Ping 프로그램을 사용하였으며, 또한 Ping프로그램이 사용하는 ICMP 메시지를 이용하여, 시퀀스 넘버와 RTT 값을 측정하였다.

만약 트래픽이 자기 유사성을 보여서 트래픽에 큰 사이클이 나타난다면, 네트워크 상황을 반영하는 RTT 값에도 그런 현상이 반영될 것이다. 반면에 트래픽이 포아송 분포를 보인다면, RTT 값의 분산은 RTT값에 $O(N)$의 형태로 비례하게 될 것이다.

다음은 Yahoo(www.yahoo.com)와 버클리 대학 연구소(www.lbl.gov)에 초당 1회 RTT 값을 측정하여 1분(60개의 표본) 동안의 평균 RTT값과 RTT 값의 분산, 표준편차를 구해 본 것이다.(2000년 8월 28일 ~ 8월 29일)

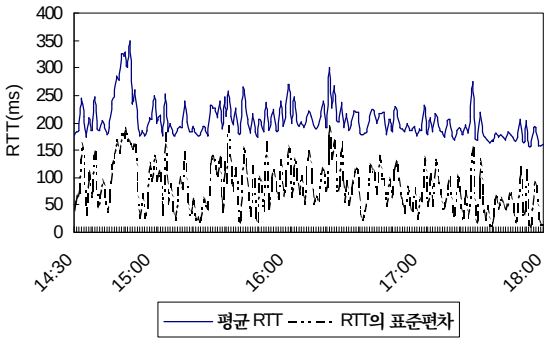

그림 2. Yahoo의 RTT분석

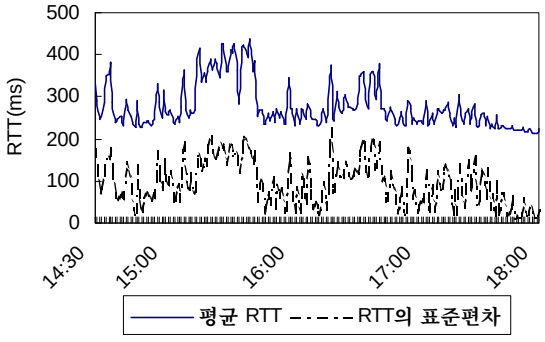

그림 3. 버클리대학 연구소의 RTT분석

그림 2와 3의 실험 결과에서 알 수 있듯이 RTT의 표준편차가 RTT의 변화를 잘 따라가고 있는 것을 알 수 있다.

즉 RTT의 값이 증가하면 RTT의 표준 편차도 증가하고, RTT값이 감소하면 RTT의 표준 편차도 감소하는 현상을 뚜렷이 볼 수 있다

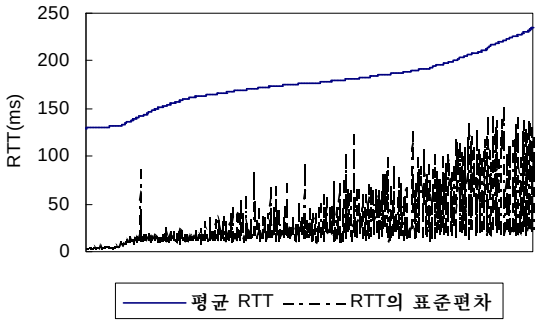

그림 4. RTT의 증가에 따른 표준 편차와 평균값의 경향

하루 동안의 측정 결과를 RTT의 오름차순으로 정렬해 나타낸 그림 4를 보면, RTT값이 증가함에 따라 표준편차의 진폭이 점점 더 커지는 경향을 확실히 알 수 있다.

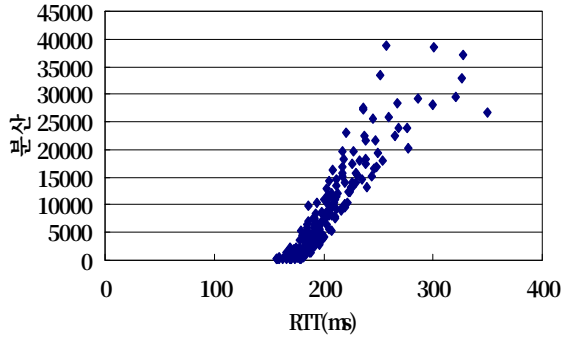

그림 5. Yahoo의 RTT와 RTT의 분산과의 관계

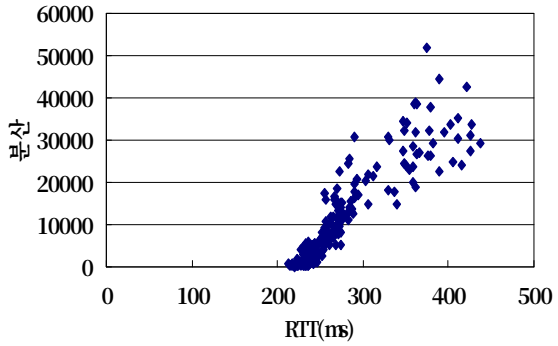

그림 6. 버클리 대학 연구소의 RTT와 RTT의 분산과의 관계

그림 5와 6은 Yahoo와 버클리 대학 연구소에 대한 RTT와 RTT의 분산과의 상관 관계를 알아본 것이다. 실험 결과에서 두 사이트 모두 공통적으로, RTT값의 분산은 RTT값에 $O(N)$의 형태로 선형증가 하고 있는 것을 알 수 있다. 얼마나 두 요소사이의 상관관계가 높은 지를 알아보기 위해, 상관 계수를 구해 보았더니 Yahoo의 경우 0.916, 버클리 대학 연구소의 경우 0.907로서 상당히 높은 상관관계가 있음을 알 수 있었다. 또한 RTT의 분산이 $O(N^2)$의 형태로 증가하지 않고 $O(N)$의 형태로 증가하는 것으로 보아, 트래픽이 자기 유사특성을 보이는 것이 아니라, 포아송 분포에 더 가깝다는 것을 알 수 있다.

IV. 결론

인터넷 사용자수와 인터넷 서비스 품질에 대한 요구가 증가하고 있는 상황에서, 억세스 망에서의 트래픽의 특성을 정확히 추정하는 것이 중요한 일이 되고 있다. 특히 스트리밍에 기반한 실시간 멀티미디어 서비스에 대한 사용자들의 요구가 높아지는 상황에서 네트워크의 통계적 특성을 파악하는 것은 반드시 필요한 일이다.

관련된 연구들을 분석해보면, 대개의 경우 사용자들의 요구는 Lognormal(또는 Pareto)한 특성을 가진다는 데에는 공감을 하고 있다. 그러나 중첩된(aggregated) 트래픽에 대하서는 자기 유사특성을 보인다는 주장과 포아송 분포를 이룬다는 주장이 맞서 있다.

먼저 억세스 망에서의 트래픽에 있어서 기존 연구들이 대부분 자기 유사성을 주장한데 비해서, 최근에는 임의의 소규모 전송이 많은 웹 트래픽의 확산으로 포아송 분포의 경향을 보인다는 연구가 최근 발표되고 있으며, 본 연구에서도 서강대 인터넷 트래픽 분석과 국외 특정 사이트에 대한 RTT분석을 통해 확인할 수 있었다.

억세스 라인에서의 다중화 이득의 관점에서 보았을 때, 사용자들의 접속의 크기가 Lognormal한 분포를 가질 경우, 매우 큰 접속 요구가 순간적으로 나타나는 자기 유사 특성을 보이고, 이런 트래픽을 적절히 수용하기 위해서는 높은 억세스 망 용량을 필요로 하게 된다. 반면에 접속의 크기가 지수 함수적인(Exponential) 분포를 가질 경우, 발생하는 트래픽이 분산이 그다지 크지 않은 포아송 분포를 보여서, 비교적 적은 억세스 망 용량으로도 사용자들의 요구를 만족시킬 수 있게 된다.

이런 상황은 인터넷에서의 사용 패턴이 WWW의 급속한 성장으로 인해 급격한 변화를 가져왔기 때문이라고 본다. 또한 스트리밍에 기반한 멀티미디어 서비스에 대한 요구가 급증하고 있는 상황에서, 트래픽을 사용 어플리케이션에 따라서 다른 특성으로 이해해야 한다고 본다. 즉, 웹 트래픽의 경우 중첩되었을 때 트래픽의 분산이 비교적 작은 포아송 분포를 보일 것이고, FTP나 멀티미디어 서비스의 경우에는 트래픽이 버스트한 특성을 보이는 자기 유사특성이 나타날 것으로 본다.

WWW이 전체 트래픽의 상당 부분을 차지하는 현재 상황에서 억세스 망의 용량은 포아송 분포쪽으로 설계의 기준을 삼을 수 있다고 본다. 다만, 버스트한 트래픽 특성을 가지는 멀티미디어 서비스 등의 요소로 인해 자기 유사특성에 대한 여유 분을 다소 부가하는 것이 바람직하다고 본다.

V. 참고문헌

[1]R.Morris and D.Lin. "Variance of Aggregated Web Traffic", Infocom, 2000.

[2]Y.M.Joo, V.Ribeiro, A.Feldmann, A.Gilbert and W.Willinger, "On the impact of variability on the buffer dynamics in IP networks,"

[3]W.Willinger and V.Paxon "Where Mathematics meets the Internet," ASM, 1998.

[4]A.Veres, Zs.Kenesi, S.Molnar and G.Vattay. "On the Propagation of Long-Range Dependence in the Internet," SIGCOMM, 2000.

[5]C.You and K.Chandra. "Time Series Models for Internet Data Traffic," 24th Conference on Local Computer Networks, 1999.

[6]M.Crovella and A.Bestavros. "Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes," IEEE/ACM TRANSACTIONS ON NETWORKING, 1997.

[7]MathWorld: 확률 및 통계분포 http://mathworld.wolfram.com