2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

인터넷 액세스망의 효율적 구성을 위한 어플리케이션별 특성 분석

*김우현, *도진숙, *이경근,
**박형구, **장주욱, ***양원석
*세종대학교 정보통신공학과
**서강대학교 전자공학과
*** LG 텔레콤(주)

Internet Traffic Analysis for Efficient Access Network Design

*Woo-Hyun Kim, *Jin-Sook Do, *Kyung-Geun Lee,
**Hyeong-Koo Park, **Ju-wook Jang, ***Won-Suk Yang
*Dept. of Information and Communications Engineering, Sejong Univ.
**Dept. of Electronic Engineering, Sogang Univ.
***LG Telecom

요약

인터넷 액세스망을 효과적으로 구축하기 위하여 급격히 확산되는 인터넷 트래픽을 수용할 기반 구축의 방법이 최근 주요 과제로 부상하고 있다. 기존의 엑세스망 설계는 전체 인터넷 트래픽 특성에 대한 가정으로부터 모델링을 하였으나 본질적으로는 특정 어플리케이션에 중점을 둔 결과를 보였다. 일반적으로 요즘에는 어플리케이션의 종류도 다양해졌고 트래픽의 양도 비약적으로 증가하였다. 그래서 이러한 설계방식은 다양한 어플리케이션를 제공하는 인터넷에서는 효율적인 엑세스망 설계를 저해하는 요인이 될 수가 있다. 본 논문에서는 계속적으로 증가하는 어플리케이션을 중심으로 어플리케이션에 따라 달라지는 트래픽의 통계적 특성을 조사하여 분류하고 설계에 필요한 요소들을 도출하고자 한다.

I. 서론

요즘 인터넷 사용자와 트래픽 양은 거의 기하급수적으로 증가하고 있다. 더불어 예전에는 파일전송(FTP), 메일(E-mail) 등 극소수의 어플리케이션만이 인터넷에서 사용되어져 왔으나 월드와이드웹(World Wide Web : 이하 웹)의 등장으로 어플리케이션의 수가 급속하게 증가하고 있음을 부인할 수 없다. 또한 이러한 어플리케이션은 사용법과 제어방법, 전송방법, 트래픽 특성등 여러 가지 면에서 저마다 다른 형태를 나타내는 것이 일반적이다.

기존의 논문에서 트래픽 패턴은 자기 유사성(Self- similarity) 또는 장기간 의존성(long-range-dependence)을 보고하거나 추측하고 있다[1][2]. 트래픽 특성이 자기 유사성에 가깝다면 억세스 망 트래픽에서 예상되는 높은 분산을 처리하기 위해서는 평균 사용 대역폭보다 높은 억세스망 용량을 설정해야 된다. 또한 억세스 라우터들을 연결하는 백본망의 용량도 따라서 증가하게 되어 상당히 많은 초기 투자가 필요하게 된다. 최근에 발표된 연구 결과에 의하면 인터넷 트래픽의 상당 부분을 차지하는 웹 트래픽의 경우 예상과 달리 포아송 분포에 가깝다는 연구가 있었는데 그 근거로 짧은 시간 간격 내에서는 사용자들의 접근 패턴이 서로 상관 관계가 낮다는 이론이 제시되었다[3]. 또 다른 연구에서는 대량의 데이터를 보내는 FTP와 같은 어플리케이션은 TCP의 혼잡 제어 프로토콜의 영향으로 인한 동기화에 의해 버스트한 트래픽을 보여 분산이 크게 나타나지만 소량의 데이터를 짧은 시간에 전송하는 웹 트래픽은 비교적 평탄한 트래픽 패턴을 보인다고 주장하였다[4]. 만약 후자의 주장을 따를 경우, 억세스 망 용량과 백본 망 용량이 줄어들어 경제적인 네트워크 구성이 가능해 질 것이다.

본 논문에서는 인터넷 트래픽을 이루는 대표적인 어플리케이션들이 일반적으로 자기 유사성을 나타내는지 포아송 특성을 나타내는지 캠퍼스 네트워크에서 측정된 여러 관측자료를 비교하여 어떠한 이론이 타당한지를 검증하고자 한다. 다음절에서는 어플리케이션별 하향(Downlink)과 상향(Uplink)비율을 분석하고 어플리케이션별 평균과 표준 편차의 비율을 조사한다. 또한 전체 최번시와 어플리케이션별 최번시 상대비율을 조사하고 활성화율(Activity Factor)을 살펴 본다. 그리고 결론에서 어플리케이션별 특성을 종합분석하고 향후 연구 방향을 제시하고자 한다.

II. 인터넷 트래픽에 대한 이론

인터넷 트래픽을 측정한 데이터 간격을 0.1초, 1초, 10초, 100초로 늘려 가면서 트래픽을 비교하면 측정된 트래픽의 패턴이 측정 간격을 작게할 경우 포아송 모델과 큰 차이를 보이지 않으나 시간 간격을 점점 증가시키면 트래픽이 평탄해 지는 경향을 보이는데 이는 인터넷 트래픽에 대한 기존 가정과 많은 차이를 나타낸다.

한편 자기 유사성(Self-similarity)을 보이는 프랙탈 함수를 사용

하여 비교하였더니 실제 여러 논문에서 가정하는 인터넷 트래픽과 비슷한 높은 분산폭을 가짐을 알 수 있었다[1]. 최근에 측정한 인터넷 트래픽을 근거로 인터넷 트래픽의 자기 유사특성이 병목 구간의 공유를 통해 인터넷에 확산된다는 연구도 있는데 이 연구는 트래픽 간의 상관관계가 상당히 천천히 감소하는 이른바 장기간 의존성 현상이 병목 구간을 통해 인터넷 트래픽에 전파된다는 이론이다[2].

인터넷의 주요 어플리케이션인 웹 트래픽 경우는 포아송특성을 갖는다는 연구가 있다. 사용자간의 상관관계가 높을 때에는 트래픽이 자기 유사성의 특성을 가진다. 즉 사용자간의 상관관계가 높다면 사용자들이 동시에 많은 양의 트래픽을 발생시키거나 동시에 트래픽을 보내지 않거나 하기 때문에 트래픽이 버스트하게 나타나는 특성을 보인다.

한편 사용자간의 상관관계가 낮을 때에는, 트래픽이 포아송 분포를 보인다. 왜냐하면 사용자간의 상관관계가 낮다면, 사용자들이 발생시키는 트래픽이 골고루 분산되어서, 트래픽의 발생이 평균 대역폭에 평탄하게 수렴하게 될 수 있기 때문이다.

그리고 사용자와 대역폭간에 얼마나 상관관계가 있는가를 나타내기 위해서 상관계수(Correlation coefficient)를 구한 결과 하바드 대학의 경우에는 사용자와 대역폭간의 상관계수는 0.88, 루슨트 연구소의 경우는 사용자와 대역폭간의 상관계수가 0.84로서 사용대역폭과 사용자수와는 상당히 높은 상관관계가 있음을 알 수 있다. 이 두 가지 사실을 종합하면 웹의 경우는 포아송 분포의 특성을 가진다는 주장이다[3].

Ⅲ. 어플리케이션별 트래픽 특성

3-1. 실험 환경

본 데이터는 세종대와 서강대에서 2000년 4월 18일부터 23일까지 측정한 자료를 기반으로 분석한 자료이다. 1,500여대 호스트 규모의 세종대와 3,600여대 호스트 규모의 서강대 규모는 일반인을 기준으로 한 어플리케이션 분포를 파악하는데 유용하다고 볼 수 있다. 본 연구에서는 인터넷의 수많은 어플리케이션 중 현재 가장 많이 사용되어지고 있으며 기존의 연구에서 다른 특성을 보일 것이라 예측되는 웹, FTP, E-mail에 대해 조사를 하였다. 각 어플리케이션의 특성을 파악하기 위해 어플리케이션별 하향과 상향의 트래픽 비율, 표준 편차와 평균의 비율, 전체 최번시와 서비스별 최번시 상대비율, 활성화율(Activity Factor)을 분석하였다.

3-2. 하향(Downlink) vs 상향(Uplink) 비율

인터넷 트래픽은 알려진 바와 같이 비대칭의 경향이 있는데 이러한 특성에 따라 망 설계시 하향과 상향의 비율을 고려한다면 효율적인 망 설계를 할 수 있을 것이다. 그래서 이러한 하향과 상향의 비율이 어플리케이션에 따라 어떻게 나타나는지를 분석하고 한다. 아래의 식과 같이 하향 대 상향 데이터 비율을 r_1 이라 정의한다.

$$r_1 = \frac{\text{총 하향 데이터 양}}{\text{총 상향 데이터 양}} \quad (1)$$

FTP와 웹에 대하여 측정한 r_1의 분포는 그림 1과 그림 2와 같다.

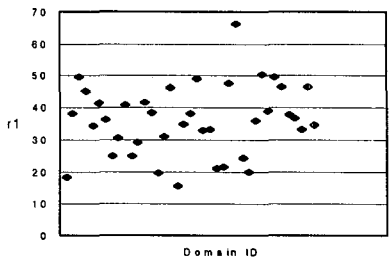

그림 1 FTP의 하향 대 상향 비율(r_1) 분포

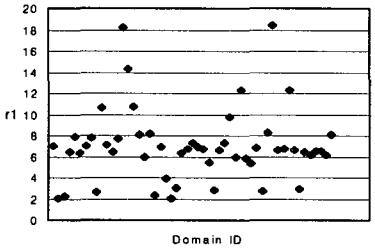

그림 2 웹의 하향 대 상향 비율(r_1) 분포

그림 1과 그림 2에서 보듯이 FTP의 r_1의 값이 웹과 크게 차이를 보이는 이유는 웹은 보통 작은 크기의 데이터로 페이지가 구성되어 Connect와 Disconnect 관계가 자주 발생됨으로써 일방향성 하향 트래픽이 크게 나타나는 FTP에 비해 데이터를 교환하는 동작이 더 빈번하게 발생하기 때문이다. 그림 3에서 웹의 경우는 서비스별 특성에 의해 r_1이 다르게 나타나는 것을 볼 수 있다. r_1의 값이 인터넷 폰, 문자 채팅과 같이 1 대 1에 가까운 성향의 서비스가 있고 증권이나 쇼핑 등과 같이 상대적으로 하향 데이터 중 이미지가 많은 서비스들은 좀 더 r_1 값이 높게 나타난다.

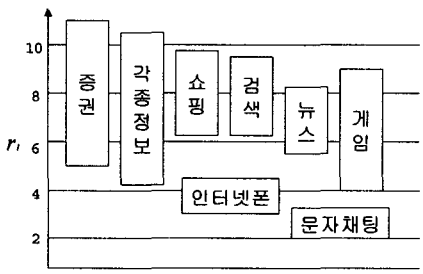

그림 3 r_1에 따른 웹 서비스 분류

3-3. 표준 편차와 평균의 비율

일반적으로 표준편차(σ)와 평균(m)의 비율이 높다는 것은 버스트한 특성이 있다는 것이며 이러한 비율을 이용하여 망 설계시 평균 트래픽 보다 어느 정도 큰 용량을 설정해야 하는지를 파악할 수 있을 것이다. 본 연구에서는 표준 편차와 평균의 비율을 r_2로 정의한다.

$$r_2 = \frac{\sigma}{m} \quad (2)$$

그림 4는 일별로 어플리케이션별 표준 편차와 평균의 비율을 측정한 결과이다.

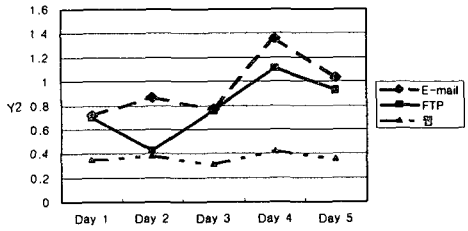

그림 4 표준편차와 평균의 비율(r_2)

그림 4에서 보듯 일일 측정값 변동이 약간 있으나 일반적으로 웹의 경우는 E-mail이나 FTP와 비교하여 r_2의 값이 낮음을 알 수 있다. 즉 웹은 메일이나 FTP보다 더 포아송 특성이 강하게 나타나는 어플리케이션이며 메일이나 FTP의 경우는 웹에 비해 버스트한 특성을 보이고 있음을 알 수 있다

3-4 전체 최번시와 어플리케이션별 최번시 상대비율

네트워크의 용량을 결정하는데 사용되는 것이 최번시(Busiest hour)의 트래픽이다. 아무리 평소에 트래픽이 없어도 특정 시간대 나 몇 분간 이라도 네트워크가 체증(Congestion)으로 동작하지 않는다면 좋은 설계라 할 수 없다. 전체 최번시에서의 어플리케이션의 데이터 양(B_1)과 어플리케이션별 최번시에서의 데이터 양(B_2)의 상대비율을 조사하면 어플리케이션별 최번시와 전체 최번시에서의 어플리케이션간의 관계를 알 수 있다. 즉 전체 최번시의 어플리케이션 비중을 알면 그 어플리케이션의 최번시 데이터 양을 알 수 있다. 이러한 상대 비율을 통해서도 버스트한 경향이 있는지를 알 수도 있다. 즉 상대비율이 작다는 것은 버스트한 경향이 크다는 말과 같으며 상대비율이 크다면 그 반대일 것이다. 상대비율(r_3)은 식 (3)과 같이 정의한다.

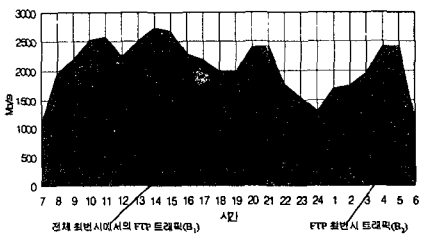

그림 5 전체 최번시에서와 최번시에서의 트래픽 정의

- A 영역 : 시간대별 전체 트래픽의 변화
- B 영역 : 시간대별 FTP 트래픽의 변화

$$r_3 = \frac{B_1}{B_2} \qquad (3)$$

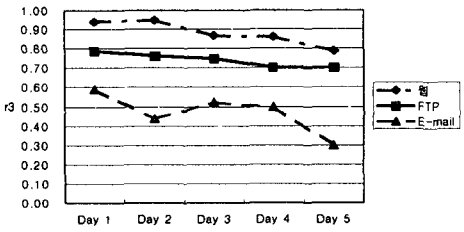

그림 6 전체 최번시와 어플리케이션별 최번시 상대비율(r_3)

각 어플리케이션에 대한 r_3을 나타낸 그림 6에 따르면 웹의 경우는 전체 최번시에서의 웹 데이터 양과 웹 최번시의 데이터 양이 비슷하다는 결과를 나타낸다. 즉 엑세스망 설계시 전체 최번시에 웹이 차지하는 데이터 양보다 약간 크게 설정을 해도 웹 트래픽을 수용할 수 있다는 것을 말한다. 그러나 E-mail의 경우는 두 배 가깝게 차이가 나며 변화의 폭이 큼을 알 수 있다.

3-5 활성화율(Activity Factor)

트래픽의 활성화율이 용량 설계 주요 요소로 될 것으로 보는데 과거의 연구에 따르면 활성화율이 낮다면 트래픽은 포아송 분포에 가깝게 될 것이고 활성화율이 높다면 자기 유사성한 특성을 보일 것으로 생각되는데 두 어플리케이션의 활성화율을 조사하여 어떠한 분포를 이루는지를 알아보고자 한다. 먼저 본 절에서 사용될 용어에 대해 살펴보고 어플리케이션별 활성화율을 계산하여 트래픽 특성을 파악하고자 한다.

r_4를 정의하기 위한 측정값들은 다음과 같이 정의하며 그 한 예가 그림 7에 나타나 있다.

- Active on : 한 Pageview를 위해 서버로부터 데이터를 수신하는 시간
- Active off : 한 Pageview를 위해 Connect 후 서버로부터 데이터를 받지 않는 시간이다.
- Inactive off : 한 Pageview를 위해 서버로부터 데이터를 모두 수신받고 사용자가 그 Page를 보는 시간
- Download Time : 한 Pageview를 위해 서버로부터 데이터를 모두 전송 받는데 까지 걸리는 시간
- Time Per Pageview : 한 Pageview에 대해 데이터를 전송 받고 보는 시간.
- Activity Factor :실제 데이터를 받은 시간에 한 Page를 보기 위해 소요되는 모든 시간으로 나눈 값으로 식 (4)와 같다.

$$r_4 = \frac{T_{active\ on}}{T_{active\ on} + T_{active\ off} + T_{inactive\ off}} \qquad (4)$$

이에 따라 활성화율 r_4를 계산하면 표 1과 같다.

그러나 웹의 경우는 페이지의 크기와 밀접한 관련이 있다. 즉 웹 페이지의 크기가 클수록, DB 검색 시간이 클수록 활성화율이 증가하나 조사한 서비스에서는 0.03에서 0.10의 값- 뉴스 0.03, 증권 0.05, 검색 0.06, 쇼핑 0.10 -을 얻었다.

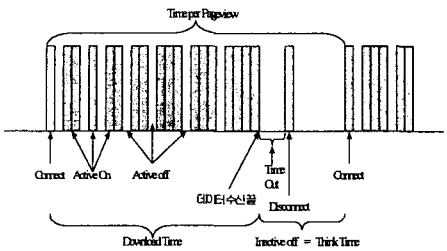

그림 7 활성화율 관련 타이밍 다이어그램

표 1 웹과 FTP의 활성화율(r_4)

어플리케이션	r_4
웹	0.05
FTP	0.97

결론적으로 볼 때 웹은 한 페이지를 보는 시간(Time per Pageview)에 비해 데이터를 받는 시간은 아주 짧음을 알 수 있다.

따라서 이러한 관점에서 볼 때 웹은 자기 유사성이 아닌 포아송 특성을 따르고 있음을 알 수 있다. FTP의 경우는 거의 실시간 적으로 데이터를 전송하므로 Inactive off가 존재하지 않으며 네트워크 체증이나 프로세싱 지연에 의한 약간의 Active off가 있을 뿐이고 대부분은 Active on 상태임을 알 수 있다. 즉 활성화율이 거의 1에 가까우면 이는 곧 자기 유사성의 특성으로 이어짐을 말해준다.

Ⅳ. 결론 및 향후 연구 방향

본 연구에서는 인터넷을 대표하는 웹, FTP, 메일 세 가지의 어플리케이션에 대해 여러 가지 특성을 분석하였다.

전체적으로 볼 때 어플리케이션별로 다른 특성을 나타내는데 상하향 비율인 r_1의 경우 FTP에 비해 웹은 작은 값을 가짐을 알 수 있다. 웹의 경우는 좀 더 다양한 서비스가 존재하는데 서비스별로 특성에 의해 r_1이 다르게 나타나는 것을 볼 수 있다. 웹의 경우 r_1을 통해 서비스의 성격을 FTP와 비슷한 성향의 서비스, 일 대 일 성이 강한 서비스, 페이지의 크기나 DB에 의해 좌우되는 서비스로 구분하여 각 어플리케이션의 이용도에 따라 용량 추정을 하는 것이 효율적인 방법이라 하겠다.

표준편차 대 평균의 비율은 r_2를 통해 FTP나 메일이 웹에 비해 버스트한 특성이 강하게 나타나는데 이는 체증의 빈도가 FTP나 메일이 웹보다 적은 사용자 특성을 반영한 것이라 할 수 있다.

최번시 상대비율인 r_3을 통해 두 가지 사실을 얻을 수 있는데 그 중 하나가 전체 최번시에서의 어플리케이션 트래픽을 측정하여 어플리케이션 최번시의 트래픽 양을 유도할 수 있다는 것이며 이를 통해 트래픽의 변동폭을 또한 알 수 있다는 것이다.

정의된 활성화율인 r_4를 통해 사용자가 소비하는 시간에 대한 트래픽 양과 Connection 관계를 알 수 있다. 웹에 대해 r_4도 r_1과 유사한 특성을 보임을 알 수 있다.

결론적으로, 점차적으로 인터넷 트래픽의 많은 부분을 차지해 가고 있는 웹은 포아송 특성이 메일이나 FTP에 비해 상대적으로 강하며 그에 반해 메일이나 FTP의 경우는 자기 유사성이 강함을 알 수 있다. 인터넷 트래픽이 증가하는 상황에서 망 설계는 각 어플리케이션의 구성에 따라 달라지며 전체적으로 '버스트' 하다거나, '포아송' 하다고 가정하는 것은 효율적인 설계를 저해할 수 있는 요인이 된다고 하겠다.

효율적인 인터넷 액세스망 구축을 위해서는 어플리케이션별 통계적 특성을 고려해야 하며 본 연구에서 논의하지 않은 다른 특성-트래픽 패턴의 비교, 유사한 어플리케이션 분류, *Interarrival Time* 등-에 대해 지속적인 연구가 필요하다.

V. 참고문헌

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

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

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

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

[5] M.Nabe, M.Murata and H.Miyahara, "Analysis and modeling of World Wide Web traffic for capacity dimensioning of Internet access lines", Performance Evaluation 34, 1998.

[6] M.Molina, Paolo Castelli and G.Foddis, "Web Traffic Modeling Exploiting TCP Connections" Temporal Clustering through HTML-REDUCE", IEEE, 2000.

[7] P.Barford and M.Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation", SIGMETRICS, 1998.

[8] B.Mah, "An Empirical Model of HTTP Network Traffic", Infocom, 1997.

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

[10] G.Babic, B.Vandalore and R.Jain "Anaysis and Modeling of Traffic in Modern Data Communication Networks Ohio State University Department of Computer and Information Science," Ohio State University Technical Report, 1998.