Self-Similar 특성을 이용한 데이터 트래픽 특성에 관한 연구

이동철*, 김기문*, 김동일**
*한국해양대학교, **동의대학교

Self-Similarity Characteristic in Data traffic

Dong Chul Lee*, Ki Moon Kim*, Dong-il Kim**

*ETRI/PEC, **Dongui University
E-mail : dongspec.etri.re.kr

요 약

본 논문에서는 Self-similar 특성을 이용하여 트래픽의 특성을 분석하고, 최근 논문들에서 보고된 사례에 대한 동향과 실제 데이터 트래픽 특성에 대한 분석연구에 대하여 Self-similar 패턴의 분석 결과를 제시한다.

ABSTRACT

The classical queuing analysis has been tremendously useful in doing capacity planning and performance prediction. However, in many real-world cases, it has found that the predicted results form a queuing analysis differ substantially from the actual observed performance. Specially, in recent years, a number of studies have demonstrated that for some environments, the traffic pattern is self-similar rather than Poisson. In this paper, we study these self-similar traffic characteristics and the definition of self-similar stochastic processes. Then, we consider the examples of self-similar data traffic, which is reported from recent measurement studies. Finally, we wish you that it makes out about the characteristics of actual data traffic more easily.

I. 서 론

본 논문에서는 self-similarity의 개념과 self-similarity의 특성을 고려하여 self-similar 데이터 트래픽에 대한 연구 결과와 Poisson 트래픽에 비교되는 형태의 트래픽 성질관계, self-similar 트래픽 모델링과 기 파라미터 분석을 통하여 Self-Similarity Characteristic in Data traffic의 결과를 나타낸다.

II. Self-similarity 특성

self-similarity에 대한 예로 1-Mbps frame relay 라인과 4000 bit로 고정된 길이의 프레임을 모니터링한다고 가정하면 각 프레임이 전송되는 데 걸리는 시간이 4ms라고 한다. 또한, 다음과 같은 도착시간 즉, 각 프레임의 첫 bit가 도착하는 시간이 수신기에 기록되다고 가정한다.

<table>
<thead>
<tr>
<th>도착시간</th>
<th>0</th>
<th>8</th>
<th>24</th>
<th>32</th>
<th>72</th>
<th>80</th>
<th>96</th>
<th>104</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>216</td>
<td>224</td>
<td>240</td>
<td>248</td>
<td>288</td>
<td>296</td>
<td>312</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>648</td>
<td>656</td>
<td>672</td>
<td>680</td>
<td>720</td>
<td>728</td>
<td>744</td>
<td>752</td>
</tr>
<tr>
<td></td>
<td>864</td>
<td>872</td>
<td>888</td>
<td>896</td>
<td>936</td>
<td>944</td>
<td>960</td>
<td>968</td>
</tr>
</tbody>
</table>

여기서 어떠한 패턴이나 통계적인 특성을 식별하기는 어렵다. 그러나, 이 트래픽은 데이터 트래픽에서 예상했던 것처럼 비스트하게 보인다. 몇 개의 도착시간들은 서로 균등하게 있고 어떤 부분들은 약간의 차이가 있다. 이중에 5개 프레임 시간(20ms)이하의 집합을 모아서 aggregate(아greggate) 어떠한 프레임의 그룹이 되는 한 클러스터로 간주한다고 가정하고, 각 클러스터의 시작 시간을 기록 결과
는 아래와 같다.

$$0 \quad 72 \quad 216 \quad 288 \quad 648 \quad 720 \quad 864 \quad 936$$

따라서 집합 정도를 다 크게 하면, 10개의 프레임 시간(40ms)이하의 집합 모델 더 큰 클러스터가 정의되는데 이는 다음과 같은 도착시간 경로가 출력된다.

$$0 \quad 216 \quad 648 \quad 864$$

이 경우 결과는 216, 432, 216의 결과를 얻어지는는데, 패턴은 아래 그림에 의해 두 개의 클러스터가 반복되고, 둘째에 의해 나뉘는 두 개의 클러스터가 된다. 이전의 8개의 클러스터로 된 집합을 살펴보면, 이러한 패턴이 반복되는 것을 알 수 있다. 처음 4개의 도착 시간들은 '도착, 옷은 걸, 도착, 옷은 걸, 도착, 옷은 걸, 도착' 등의 패턴이 형성되고, 마지막 4개의 도착 시간들은 동일하게 반복된다. 원래 32개의 도착 데이터 집합으로 돌이켜서 살펴보면, 이러한 동일한 패턴이 8번 반복되는 것을 알 수 있다. 그러나 그 패턴은 원래의 데이터와 같은 모양을 가지는 서로 다른 패턴의 집합이 반복되는 것을 알 수 있다. 이것은 타임 시리즈가 resolution의 정도에 상관없이 동일한 패턴으로 나타나는 것이다. 이것이 self-similarity의 핵심이다.

그림 1. Self-similar 시계열(time series).

(self-similarity)는 단지 최근에 들어서야 데이터 통신 트래픽 분석에 적합한 중요한 개념이다. 최근에 self-similarity가 여러곳에 산재해 있다는 것이 밝혀졌다(32), self-similarity는 차원(dimension)상의 서로 다른 확률분포나 서로 다른 스케일에서 보았을 때 동일하게 보이거나 동일하게 행동하는 자기 유사한 현상으로 나타난다.

(상기 예제에서 패턴은 근로법으로 보는 것이 좀 더 이해하기 쉽다. 그림 1.1)는 시간에 따른 프레임의 연속 도착을 그린 것이다. 각각의 수치 라인을은 수신자가 첫 bit에서 마지막 bit까지 전체의 프레임을 홀수하는데 걸리는 시간이 4ms에 빠르는 폭을 가진 한 개의 프레임을 나타낸다. 그림 1.2는 4개의 큰 클러스터로 모아진 데이터를 나타낸다. 이 그림에서 '도착, 옷은 걸, 도착, 옷은 걸, 도착'의 패턴이 데이터의 서로 다른 해상도에서 나타난다는 것을 알 수 있다.

앞의 예는 chaos, fractal에 관한 유명한 구조인 칸토르 집합(Cantor set)에서 유래된 것이다. self-similar한 특징들은 실제의 현상에서 불합하거나 유지되는 것은 아니다. 그렇지만 아주 큰 범위의 스케일을 통해 대부분의 현상들은 self-similarity를 드러낼 것이다.

III. Self-similar Data Traffic 특성


일반적으로 결정적이고 주기적인 신호는 시간에 따른 불변의 특징이 있다. 즉 그 신호는 시간상으로 여러가지 주기가 동일하다는 동일한 신호이다. 이에 비해, 정상확률과정에 대해서는 그 과정의 동형은 시간의 동일하다. 또한 평균은 시간에 독립적이고 자기교잡한 함수는 단지 시간의 차이에만 의존한다(34).

self-similar 확률과정은 기존의 논문들에서 여러가지 방법으로 정의되어 있다. 본 논문에서 는 단순 연속시간 확률과정에 대해 설명한 후, 데이터 트래픽과 관련된 이산시간 확률과정에 대해 설명하였다(35).

3.1 연속 시간 정의 분석 연구

Self-similar 확률과정의 일반적인 정의는 다음과 같이 연속시간 변수의 적합 스케일링에 기초한다. 어떤 연속 변수 $a > 0$에 대해, 확률과정 $X(a)\alpha$가 $X(t)$와 동일하게 동일한 특성을 가진다면, 확률과정 $X(t)$는 파라미터 $H(0.5 < H \leq 1)$를 가지고 통계적으로 self-similar하다. 이러한 관계는 다음의 3가지 조건으로 표현된다(36).

$$1. \quad E[X(t)] = E[X(at)]$$
$$2. \quad Var[X(t)] = \frac{Var[X(at)]}{a^{2H}}$$

**Mean**

<table>
<thead>
<tr>
<th>$\mu_X(t)$</th>
<th>$\mu_X(at)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(t)$</td>
<td>$X(at)$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\sigma_X^2(t)$</th>
<th>$\sigma_X^2(at)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(t)$</td>
<td>$X(at)$</td>
</tr>
</tbody>
</table>

**Variance**

<table>
<thead>
<tr>
<th>$\sigma_X^2(t)$</th>
<th>$\sigma_X^2(at)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(t)$</td>
<td>$X(at)$</td>
</tr>
</tbody>
</table>
3. \( R_c(t, s) = \frac{R_c(a, s)}{a^\gamma} \) Autocorrelation

Hurst 또는 self-similarity 파라미터 \( H \)는 self-similarity의 핵심적이다. 다시 말하면, \( H \)는 분
계적 인 환상의 지속성(persistence)에 대한 측도이
고 확률과정의 각각의 종속에 대한 측도이다.
\( H=0.5 \)의 값은 self-similarity의 부재를 나타내고,
\( H=1 \)에 가까운수록, 지속성의 정도 또는 장기간
의 종속의 정도는 더욱 커진다. 즉, \( H>0.5 \)에 대하
여 파괴와 미래의 연속에 대한 상관성을 없어지
고, \( H<0.5 \)에 대하여 지속성의 두드러진 특성을
가진다.

3.2 이산 시간 정의 분석 연구

정상 시계열(stationary time series) \( x \)에 대해,
\( m \)-aggregated 시계열 \( x^{(m)}(t) \) \( m=1, 2, \ldots \),을 연립
\( m \)-크기의 흐름을 결정 없이(nonoverlapping)
원래의 시계열을 핵심적으로써 정의한다. 이것은
\( m \)-기가 표현된다.

\[
x^{(m)}_k = \frac{1}{m} \sum_{i=1}^{m} (x_{k+i-m})
\]  

예를 들어, \( x^{(3)}(3) \)는 다음과 같이 정의된다.

\[
x^{(3)}(3) = \frac{x_{1} + x_{2} + x_{3}}{3}
\]  

aggregated 시계열을 완료하는 한 방법은 타임
스케일을 압축하는 기법이다. 즉, \( x^{(m)} \)는 \( m \)-크기의 시계열의 최대 해상도이고, \( x^{(3)} \)는 비율 3으로 축
소한 것이다. 만약 이에 확률과정의 통계(mean,
variance, correlation, etc)가 동일한 양축 사분을
간직하고 있다면, self-similar 과정으로 나를 수
있다.

\( x^{(m)} \)의 예르고드적 과정에 대해, 시간평균은 조화
평균과 동일하고, 시간평균의 분산은 조화평균과
동일하다. 만약 시간평균의 분산이 \( m \)이 매우 커
짐에 따라 zero로 수렴하게 되면, 이것은 self-
similar 과정이 아니다. 만약 확률과정 \( x \)가 모든
\( m=1, 2, \ldots \)에 대해서 다음과 같다면 파라미터 \( \beta (0<
\beta <1) \)에 대하여 정확하게(exactly) self-similar하
다고 한다.

\[
Var(x^{(m)}_k) = \frac{Var(x)}{m^\beta} \quad \text{Variance}
\]

\[
R_{c^{(m)}}(k) = R_c(k) \quad \text{Autocorrelation}
\]

따라서, 이러한 self-similarity의 정의에 의해서,
aggregated 과정의 자기상관은 원래의 확률과정
과 동일한 형태를 가진다. 이것은 변이성 또는 버
스트한 정도가 서로 다른 타임스케일에서 동일
하게 나타난다는 것을 의미한다.

3.3 장기간 의존성 분석 연구

self-similar 과정의 중요한 특징들 중 하나는 장기
간 의존성으로 나타난다. 일반적으로, 단기간의
존재성은 자기공분산이 최소한 지수적으로 급격
히 감소한다는 조건을 만족시킨다.

\[
C(k) \sim a^{-\beta}, \quad |k| \rightarrow \infty, \quad 0 < \beta < 1
\]

이와 같이, 장기간 의존성은 hyperbolical하
게 감소하는 자기공분산을 가진다.

\[
C(k) \sim |k|^{-\beta}, \quad |k| \rightarrow \infty, \quad 0 < \beta < 1
\]

여기서 \( \beta \)는 양에서 정의된 파라미터이고
\( H=1-(\beta /2) \)이기 때문에 Hurst 파라미터와 연관이
있다. 이 경우 \( \sum C(k) = \infty \)이 될 수밖에
있다. 그에 \( \sum C(k) = \infty \)이 될 수 있다.

장기간 의존성은 self-
similar 과정들의 저장적 현상을 반영한다.
즉, 모든 타임스케일들의 균질 및 버스트한 특
성의 존재를 나타낸다.
3.4 Heavy-tailed Distributions

앞의 aggregated 시계열, 장기 의존성에 관한 앞의 흐름은 동등한 정의이다. heavy-tailed 분포는 대로 다음과 흐름적인 특징으로 동등한 분포를 가진 평균 세부적인 특징을 갖고 있지만, 본질적으로 이러한 분포로 self-similar 확률과정을 정의하는 것이 가능하다. heavy-tailed 분포 집계 전문가의 예로는 사설레이션 모델들을 다루기 쉽게 한다는 것이다.

heavy-tailed 분포는 폐겨의 도착(interarrival) 시간 및 비스트 길이와 같은 트래픽 과정을 설명하는 확률적 확률을 나타내기 위해 사용되며 잘 수 있다. 랜덤변수 X의 분포가 다음과 같다면 heavy-tailed하다고 한다.

\[ 1 - F(x) = \Pr[X \leq x] \approx \frac{1}{x^a} \quad \text{as } x \to \infty, \quad 0 < a < 1 \]  

(7)

일반적으로 heavy-tailed 분포를 가진 랜덤변수는 높은 분산(성어)이 무한대의 분산)을 갖는다. 가장 단순한 heavy-tailed 분포는 파라미터 k와 \( a(k, a > 0) \)를 가진 Pareto 분포이다. 일반적 최 분포함수는 다음과 같다.

\[ F(x) = \frac{k}{k + x}, \quad F(x) = 1 - \left( \frac{k}{x} \right)^a \quad \text{for } x \geq 0 \]  

(8)

따라서 기대값은 (10)식과 같이 나타난다.

\[ \mathbb{E}[X] = \frac{a}{a-1} k \quad \text{for } a > 1 \]  

(9)

파라미터 k는 랜덤변수가 취할 수 있는 최소의 값을 제한한다. 파라미터 a는 랜덤변수의 기대값 및 분산을 결정한다. 만약 \( a \leq 2 \)이면, 분포함수는 무한 분산을 가지고, \( a \leq 1 \)이면, 무한한 기대값과 분산을 가지게 된다. 그림 3은 log-linear 스케일의 Pareto 및 치수밀도 함수를 비교한 것이다.

이 그림에서 지수밀도 함수는 거의 직선으로 나타나고, Pareto 분포의 tail은 지수함수보다 매우 더 서서히 감소한다. 그러므로 'heavy tail' 분포 함수라 한다.

![그림 2. 장기적 의존성과 단기적 의존성의 비교](image)

![그림 3. Pareto 및 지수 확률 밀도 함수](image)

IV. Self-similar 데이터 트래픽 특성 연구

최근 몇 년 동안 여러 연구에서 실제의 네트워크 환경에서 데이터의 트래픽 패턴이 self-similar 과정에 의해 아주 잘 모델링된다는 것이 증명되었다. 이 장에서는 몇몇 논문에서 발표되었던 두 가지의 self-similar한 데이터 트래픽의 예를 소개한다.

4.1 Ethernet Traffic 특성 연구

Ethernet traffic에서는 기존의 Poisson 트래픽 가정을 사용하는 straightforward 유형분석이 모든 네트워크 트래픽을 모델링하는데 적합하지 않음이 밝혀졌고, Ethernet traffic에 대한 새로운 모델링과 분석의 접근법이 대두되었다.

그림 4는 이러한 Ethernet 트래픽 특성과 Poisson 모델링의 문제점을 단적으로 보여준다. 그림 4는 실제의 Ethernet 트래픽이 여러 가지 해상도의 단계에 따라 나뉘는 것이다. 여기에서 외도로운 것은 모든 그림들의 분포적인 관점에서 볼 때, Ethernet 트래픽은 큰 스케일(hours, minutes)이나 작은 스케일(seconds, milliseconds)에서도 그 비스트링이 잘 나타나있다. 즉, 모든 타임스케일에서도 비스트링을 나타낸다.

이와 대조적으로, 그림 5는 Ethernet 그룹과 동일한 방식으로 생성되었지만, 합성된 트래픽 데이터를 사용해서 그린 것이고, 이를 데이터는 실제의 데이터와 거의 비슷한 평균패킷 사이즈와 도착률을 구해 Poisson 모델을 사용해서 생성된 경우이다. 

![그림 3. Pareto 및 지수 확률 밀도 함수](image)
상도가 점점 낮아짐에 따라 smooth out해진다.
이는 Poisson 모델에서 정의된 것처럼 정상 및
예로고 디 확률과정으로부터 기인한 것이다.

그림 4c는 실제 Ethernet 트래픽의 Hurst 파라미터를
추정한 값인 H=0.9를 사용한 self-
similar 트래픽 모델로 생성된 것이다. 이 그림
은 실제의 Ethernet 트래픽의 그것과 일반적으로
동일한 특성을 보여준다. 즉, 모든 타임 스펙트럼에
서 Ethernet 트래픽의 베타함을 아주 잘 보여주
고 있다.

4.2 World Wide Web 트래픽 연구

50만개 이상의 Web 도ку먼트들에 대한 요청이
수반된 Web 트래픽 연구의 예로서, 웹 브라우저에
의해 생성된 트래픽 패턴이 self-similar 하다는 것
이 증명되었다. 각각의 브라우저는 ON/OFF 소
스로 모델링하였고, 그 데이터가 Pareto 분포하
며 수 잘 들어있고, 다양한 측정을 통해 그것의
1.16~1.5범위의 α 값을 가지 Pareto 분포라는 것
이 발견되었다[17].

V. Self-similar 트래픽의 성능관계

이러한 데이터 트래픽의 self-similar한 특성이
실제의 성능에 어떠한 영향을 미치는가에 대한
문제가 제기되었고, 많은 연구가 행해졌다.
그중 가장 중요한 문제는 self-similarity가 성능에
박해한 영향을 미친다는 것이다.

특히, Ethernet에 대한 중요한 발견은 Ethernet
상의 부하가 점점 더 높아질수록, 계산된 Hurst
파라미터 H가 더욱 높아지거나, 이에 상응하는
self-similarity의 정도가 더 높아졌다는 것이다.
이러한 결론은 성능의 정점에서 가장 관련 있는
높은 부하정도에서 정확하기 때문에 극히 중요한
것이다. 또한 Ethernet 분석에 대한 중요한 결론
는 성능을 예측하기 위한 전통적인 Poisson모델들이
적당하다는 것이었다[17]. 예를 들어, 데이터 트래
픽에 관련된 일반적인 가설은 많은 수의 독립적
인 트래픽 스트림들을 베타함을 대상하는 것이
Poisson성으로 귀착된다는 것이다. 기존의 이러
한 가정과 결과적인 유휴분석이 초창기의 ATM
switch 제조업자들이 매우 적은 비파(10~100cells)
冗가진 1세대 스위치를 생산하도록 했고, 이러한 스위치들이 혼잡에 대처되지 실패한 트래픽 수용했을 때, 기대한 범위를 초과 넘어선 cell 손실들이 발생하게 되었고, 그 스위치들을 다시 설계하도록 하는 결과를 초래했다. 이를 예로 알 수 있듯이, 입력이 self-similar하다면, 증가되는 자연과 증가되는 비례 사이즈의 요구조건은 self-similar 스트림의 어떠한 멀티플렉싱으로 나타날 것이라는 결론이 나온다[9]. 이것은 ATM, frame relay와 100BASE-T와 같은 스위치들와 WAN 라우터들, Ethernet과 같은 공유전송 링과 같은 통합적으로 멀티플렉싱에도 적용된다.

VI. 결론

본 논문에서는 self-similar의 정의와 self-similarity를 나타내는 특성과 데이터를 고찰 보았고, 데이터 트래픽에서의 실제 예를 통해 기존의 Poisson 뮤임의 분산이 self-similar 특성을 가진 트래픽 분석에서는 적합하지 않음을 설명했다. 그러나 동일한 뮤임의 분산이 이제 부정적으로 써지는 의미는 아니다. 즉 self-similarity가 모든 데이터 트래픽에서 적용될 수 있는 것은 아니다.

여기서 고찰한 Self-similar 트래픽의 특성을 토대로 하여, 추후 연구에 있어서는 실제 트래픽을 측정한 결과 적절 Hurst 파라미터를 추정해 보고, 그 결과를 Network Simulator를 사용해 시뮬레이션한 결과를 Self-similar 트래픽 모델의 적용성 및 타당성에 대해서 검증해 볼 것이다.

참고문헌

[7] Corvella, M., Béstavros, A. Self-similarity in World-Wide Web Traffic: Evidence and