웹 서비스 환경에서 평균 응답 시간의 제약조건을 만족하는 최대 객체 크기의 추정

이용진*
한국교원대학교 기술교육과 교수

Estimation of maximum object size satisfying mean response time constraint in web service environment

Yong-Jin Lee*
Professor, Department of Technology Education, Korea National University of Education

요 약 웹 서비스 환경에서 사용자가 원하는 서비스 품질을 만족하는 경제적인 방법의 하나는 객체의 크기를 조절하는 것이다. 이를 위해 본 연구에서는 서비스 품질로 평균 응답 시간이 임의의 임계값 이하로 주어질 때, 이 제약 조건을 만족하는 최대 객체의 크기를 구한다. 시스템이 정상 상태일 때, 라운드-로빈을 사용하는 결정 모델의 평균 응답 시간은 일반 분포를 따르는 큐잉 모델에서의 평균 응답 시간과 같아질 것으로 추론할 수 있다. 이를 기반으로 최대 객체 크기를 발견하기 위한 해석적 수식과 절차를 수립한다. 웹 트래픽은 서비스 분포로 Pareto 분포가 적합하므로 M/G(Pareto)/1 모델과 지수 분포를 사용한 M/G/1/PS를 적용하여 최대 객체의 크기를 구한다. 수치 계산을 통한 성능평가는 Pareto 분포의 형상 파라미터(shaping parameter)가 커짐에 따라 M/G(Pareto)/1 모델과 M/G/1/PS 모델의 최대 객체 크기가 같아짐을 보여준다. 본 연구의 결과는 경제적인 웹 서비스 제어를 위해 객체의 크기를 조절하는 환경에 사용될 수 있다.

주제어 : 최대 객체 크기, M/G/1/PS, Pareto 분포, 평균 응답 시간

Abstract One of the economical ways to satisfy the quality of service desired by the user in a web service environment is to adjust the size of the object. To this end, this study finds the maximum size of objects that satisfy this constraint when the mean response time is given below an arbitrary threshold for quality of service. It can be inferred that in the steady state of system, the mean response time in the deterministic model by using the round-robin will be the same as that of the queueing model following the general distribution. Based on this, analytical formulas and procedures for finding the maximum object size are obtained. As a service distribution of web traffic, the Pareto distribution is appropriate, so the maximum object size is computed by applying the M/G(Pareto)/1 model and the M/G/1/PS model using exponential distribution as computational experience. Performance evaluation through numerical calculation shows that as the shape parameter in the Pareto distribution increases, the M/G(Pareto)/1 model and M/G/1/PS model have the same maximum object size. The results of this study can be used to environments where objects can be sized for economical web service control.

Key Words : Maximum object size, M/G/1/PS, Pareto distribution, Mean response time

*교신저자 : 이용진(lyj@knue.ac.kr)
접수일 2023년 3월 16일 수정일 2023년 4월 20일 심사완료일 2023년 4월 23일

1. 서론

웹 객체의 크기는 종단 사용자가 요구하는 서비스 품질을 만족시키는 웹 서비스 시스템을 설계할 때 고려해야 할 사항 중의 하나이다. 대표적인 서비스 요구 품질로는 평균 대기 시간과 평균 응답 시간이 있다[1]. 웹 객체의 크기가 커질수록 종단 사용자의 대기 시간이 증가하는 반면에, 웹 객체의 크기가 작아질수록 웹 서비스를 지원하는 TCP (Transmission Control Protocol)의 슬로우 스타트 시간이 증가하기 때문에 네트워크의 성능이 저하된다. 또한 웹 서버에 동시에 접속하는 사용자(concurrent users)의 수가 늘어남에 따라 지연 시간도 늘어난다. 따라서 이러한 상황들을 고려하여 웹 객체의 크기를 결정해야 한다.

일반적으로 사용자의 요구가 웹 서버에 도착하는 통계적 과정은 푸아송 분포를 고려하고, 웹 서버에서의 서비스는 웹 객체의 크기에 따른 일반분포(General distribution)를 고려하는 M/G/1 모델을 사용한다[2].

웹 객체는 홈페이지와 같은 정적 객체와 홈페이지에 포함된 이미지 등의 동적 객체로 분류된다. 정적 객체의 크기는 무거운 꼬리 (heavy tailed) 특성을 갖는데, 실제 웹 트래픽 분석 결과 형상 파라미터(shape parameter)가 1.16에서 1.5 사이의 값을 갖는 Pareto 분포를 하는 것으로 알려져 있다[3-8].

M/G/1 모델을 사용하면 평균 대기 시간과 시스템 내에서의 평균 객체 크기를 계산할 수 있다. 하지만 사용자가 요구하는 평균 응답 시간을 만족하는 최대 객체의 크기는 직접 계산할 수 없다. M/G/1 모델에서 사용하는 서비스의 확률분포(G)로는 지수 분포, 초기하분포 및 와이블(weibull) 분포 등이 있다[9,10].

웹 서비스 정책으로는 주로 FIFO가 사용되지만 프로세서 공유 정책도 많이 사용된다. 이 방식은 라운드-로빈 스케줄링과 유사하며 M/G/1/PS로 표기한다[11,12]. 이 밖에 Bouded Pareto 분포를 사용하는 M/BP/1 모델도 웹 서비스를 표현하는 데 사용된다[13,14].

사용자가 원하는 평균 응답 시간을 만족하는 최대 객체의 크기를 구하려면 먼저 결정 모델(deterministic model)을 수립하여 평균 응답 시간을 계산하고, 이것을 M/G/1 모델의 평균 응답 시간과 같은 것으로 추론하여 객체의 크기를 구한다.

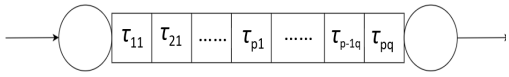
기존 연구에서는 초기하분포와 와이블 분포에 대해 평균 대기 시간을 만족하는 최대 객체의 크기를 구한 바 있으나[15], 본 연구에서는 지수 분포와 Pareto 분포에서 평균 응답 시간을 만족하는 최대 객체 크기를 구하고자 한다.

2. 최대 객체 크기 추정 모델

이 장에서는 사용자의 평균 응답 시간을 만족시키는 최대 객체 크기를 추정하는 데 필요한 수식 및 절차에 관해 설명한다.

먼저, 웹 서버가 하나의 링크를 통해 외부와 연결되고 p명의 사용자가 동시에 접속한다고 하자. 또한 웹 객체에 포함된 패킷의 수를 q라고 하자.

[Fig. 1]에서와 같이 i번째 사용자의 j번째 패킷을 서비스하는 시간을 τ_{ij} (i=1,2,..,p: j=1,2,..,q)라 하고, 패킷 기반의 라운드-로빈 스케줄링을 적용하면 결정 모델에서의 평균 응답 시간(E(T_D))은 식 (1)과 같다. ρ는 시스템 부하(load), τ는 패킷 서비스 시간, 그리고 E(x)는 서비스 시간의 기댓값이다[10]. 단, 모든 (i, j)에 대해 τ_{ij} = τ이다.



[Fig. 1] Mean waiting time in the deterministic model

$$E(T_D) = \frac{(p-1)(2q-1)\tau}{2} + E(x) \qquad (1)$$

M/G/1 모델에서 평균 응답 시간(E(T))은 식 (2)로 주어진다[1,2]. E(x^2)은 서비스 분포의 2차 모멘트이다.

$$E(T) = \frac{\rho E(x^2)}{2(1-\rho)E(x)} + E(x) \qquad (2)$$

시스템이 정상 상태(steady state)에 있는 경우, E(T_D) = E(T)로 간주할 수 있다. 따라서 식 (1)과 식 (2)로부터 p에 대해 정리하면 식 (3)을 얻는다.

$$p = 1 + \frac{\rho E(x^2)}{(1-\rho)(2q-1)E(x)\tau} \qquad (3)$$

Little의 공식에 의해 시스템에 있는 평균 사용자의 수(E(N))는 λE(T)이므로 식 (4)와 같다.

$$E(N) = \rho + \frac{\lambda^2 E(x^2)}{2(1-\rho)} \quad (4)$$

시스템이 정상 상태일 때, p와 E(N)이 같아질 것으로 추론할 수 있다. 따라서 식 (3)과 식 (4)로부터 q에 대해 정리하면 식 (5)를 얻는다.

$$q = \frac{1}{2}\left[1 + \frac{2\rho E(x^2)}{\{\lambda^2 E(x^2) - 2(1-\rho)^2\}E(x)\tau}\right] \quad (5)$$

이제 평균 객체 크기(mean object size)를 Ψ, 최대 세그먼트 크기(maximum segment size)를 S로 놓으면 $q \geq \Psi/S$이다. 따라서 q를 식 (5)에 대입하여 정리하면 식 (6)을 얻는다.

$$\Psi \leq \frac{S}{2}\left[1 + \frac{2\rho E(x^2)}{\{\lambda^2 E(x^2) - 2(1-\rho)^2\}E(x)\tau}\right] \quad (6)$$

이제, 사용자가 원하는 평균 응답 시간을 T_r로 놓으면 식 (2)에서 E(T) 〈 T_r이다. 식 (2)를 E(x^2)에 대해 정리하면 식 (7)을 얻는다.

$$E(x^2) \leq \frac{2(1-\rho)(T_r - E(x))}{\lambda} \quad (7)$$

식 (7)을 식 (6)에 대입하면, 사용자가 원하는 평균 응답 시간을 만족하는 최대 객체의 크기는 식 (8)로 주어진다.

$$\hat{\Psi} \leq \frac{S}{2}\left[1 + \frac{2(T_r - E(x))}{\{\lambda(T_r - E(x)) - (1-\rho)\}\tau}\right] \quad (8)$$

3. 웹 서비스 모델에 따른 최대 객체 크기의 추정

2장에서 제시한 식 (8)을 이용하여 최대 객체 크기를 구하려면 특정한 웹 서비스 모델을 가정해야 한다. 이 장에서는 M/G/1/PS 모델과 G가 Pareto 분포인 M/G/1 모델을 고려한다.

3.1 M/G/1/PS 모델

이 모델은 데이터 네트워크에서 사용자의 서비스 모형으로 사용하는 것으로 [Fig. 1]에 묘사된 패킷 기반의 라운드-로빈 스케줄링과 유사하다. 특히, M/G/1/PS 모델을 사용했을 때 정상 상태에서의 사용자의 수가 p인 확률은 M/M/1 모델에서 FIFO로 서비스할 때 사용자의 수가 p일 확률과 같다[11]. 따라서 M/M/1 모델을 사용하여 최대 객체 크기를 구할 수 있다.

지수 분포(M)의 확률 밀도 함수는 식 (9)와 같고, 기댓값은 식 (10)과 같다.

$$f(x) = \begin{cases} \mu e^{-\mu x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (9)$$

$$E(x) = \frac{1}{\mu} \quad (10)$$

식 (10)을 식 (8)에 대입하여 정리하면 사용자의 평균 응답 시간의 제약 조건을 만족하는 최대 객체 크기를 얻는다.

$$\hat{\Psi}_{M/G/1/PS} \leq \frac{S}{2}\left[1 + \frac{2(\mu T_r - 1)}{\{\lambda(\mu T_r - 1) - (\mu - \lambda)\}\tau}\right] \quad (11)$$

3.2 M/G(Pareto)/1 모델

웹 트래픽 특성의 연구를 통해 웹 브라우저에 의한 트래픽 패턴이 자기-유사적(self-similar)으로 밝혀졌다[3,4]. 특히 웹 브라우저를 ON/ OFF 소스로 모델링한 데이터들은 형상 파라미터(shape parameter)가 1.16에서 1.5 사이의 값을 갖는 Pareto 분포를 따른다[4].

따라서 이 절에서는 서비스 분포(G)가 Pareto 분포를 따르는 M/G/1/FIFO 모델에서 최대 객체 크기를 추정한다.

Pareto 분포의 확률 밀도 함수는 식 (12)와 같다. α는 형상 파라미터이고, x_m은 최빈수이다.

$$f(x) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} \qquad x \geq x_m \quad (12)$$

기댓값은 식 (13)과 같다.

$$E(x) = \frac{\alpha x_m}{\alpha - 1} \qquad \alpha > 1 \quad (13)$$

식 (13)을 식 (8)에 대입하여 정리하면 사용자의 평균 응답 시간의 제약 조건을 만족하는 최대 객체 크기를 얻는다.

$$\hat{\Psi}_{M/G(Pareto)/1} \leq \frac{S}{2}\left[1+\frac{2\{T_r(\alpha-1)-\alpha x_m\}\alpha x_m/(\alpha-1)}{\{\rho(T_r(\alpha-1)-\alpha x_m)-(1-\rho)\alpha x_m\}\tau}\right] \quad (14)$$

4. 성능평가

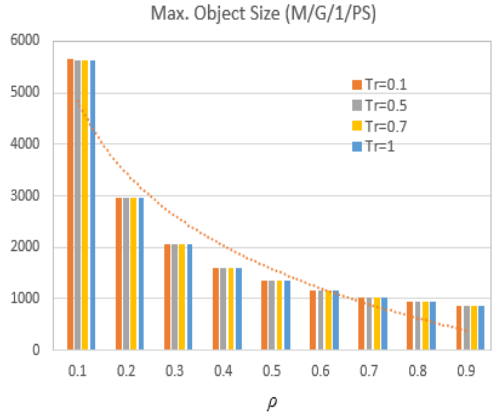
이 장에서는 3장에서 유도한 최대 객체 크기에 관한 수식을 이용하여 얻은 수치 결과를 설명한다.

4.1 M/G/1/PS 모델

〈Table 1〉은 식 (11)을 사용하여 M/G/1/PS 모델에 나타낸다. 이때 사용된 성능 파라미터는 S(maximum segment size)=536B, C(link capacity)=100Mbps, τ(time slice)= $S\times 8/C$, E(x)=τ이다.

〈Table 1〉 $\hat{\Psi}_{M/G/1/PS}$ satisfying Tr in M/G/1/PS model

ρ	T_r = 0.1	T_r = 0.5	T_r =0.7	T_r = 1
0.1	5649	5632	5631	5630
0.2	2953	2949	2949	2948
0.3	2056	2055	2055	2055
0.4	1609	1608	1608	1608
0.5	1340	1340	1340	1340
0.6	1162	1161	1161	1161
0.7	1034	1034	1034	1034
0.8	938	938	938	938
0.9	864	864	864	864
평균	1953	1953	1953	1956



[Fig. 2] Maximum object size satisfying Tr in M/G/1/PS model

[Fig. 2]는 〈Table 1〉을 그림으로 표시한 것이다. 이 예제에서는 T_r 값의 미세한 변화 때문에 최대 객체 크기가 거의 변하지 않았지만, S, C, τ의 크기가 변함에 따라 최대 객체 크기 역시 변하게 된다.

4.2 M/G(Pareto)/1 모델

이제 M/G(Pareto)/1 모델에서 주어진 평균 응답시간 (T_r=1.0)을 만족하는 최대 객체 크기를 계산해 보자. 이때 사용된 성능 파라미터는 α = 1.1, 1.2, 1.3, 1.4, 1.5, 5000, S(maximum segment size)=536B, C(link capacity)=100Mbps, τ(time slice)=$S\times 8/C$, E(x)=τ이다.

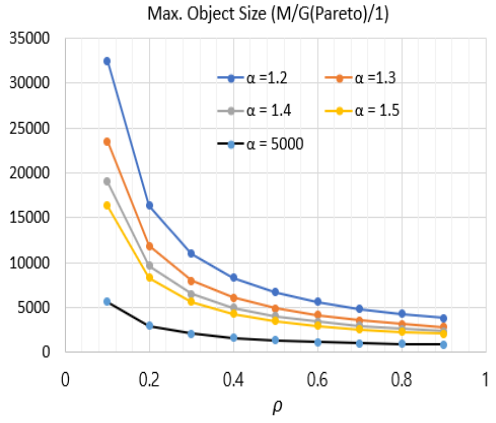
〈Table 2〉는 식 (14)를 사용하여 형상 파라미터(α)가 1.16에서 1.5 사이의 값을 갖는 Pareto 분포를 감안하여 α = 1.2에서 α = 1.5까지 변화시켰을 때 최대 객체 크기를 계산한 것이다.

M/G(Pareto)/1 모델에서 α가 매우 커지는 경우를 가정하여 α = 5000인 경우도 계산하였는데, 이때의 최대 객체 크기는 $\alpha \geq 5000$일 때의 최대 객체 크기와 같아져서 특정한 값에 수렴한다.

〈Table 2〉 $\hat{\Psi}_{M/G(Pareto)/1}$ satisfying T_r = 1.0 for varying α in M/G(Pareto)/1 model

	$\hat{\Psi}_{M/G(Pareto)/1}$				
ρ \ α	1.2	1.3	1.4	1.5	5000
0.1	32503	23534	19053	16367	5631
0.2	16365	11890	9654	8312	2949
0.3	10994	8014	6524	5630	2055
0.4	8311	6076	4959	4289	1608
0.5	6702	4914	4021	3484	1340
0.6	5629	4140	3395	2948	1162
0.7	4863	3586	2948	2565	1034
0.8	4288	3171	2613	2278	938
0.9	3841	2849	2352	2055	864
평균	10388	7575	6169	5325	1953

[Fig. 3]은 〈Table 2〉를 그림으로 표시한 것이다. 동일한 ρ 값일 때 α 값이 커지면 최대 객체의 크기는 작아지고, 동일한 α 값일 때 ρ 값이 커지면 최대 객체 크기는 큰 변화가 없다.



[Fig. 3] Maximum object size satisfying Tr =1.0 for varying α in M/G(Pareto)/1 model

4.3 M/G/1/PS 모델과 M/G(Pareto)/1 모델의 비교

〈Table 3〉은 평균 응답 시간(T_r)의 변화에 따른 M/G/1/PS 모델과 M/G(Pareto)/1 모델의 최대 객체 크기를 비교한 것이다.

〈Table 3〉에서 보듯이 모든 T_r에 대해 M/G/1/PS 모델의 최대 객체 크기와 α = 5000일 때, M/G(Pareto)/1 모델의 최대 객체 크기는 같아진다.

〈Table 3〉 Comparison of $\hat{\Psi}_{M/G(Pareto)/1}$ and $\hat{\Psi}_{M/G/1/PS}$ for varying Tr

	$\hat{\Psi}_{M/G/1/PS}$	$\hat{\Psi}_{M/G(Pareto)/1}$		$\hat{\Psi}_{M/G/1/PS}$	$\hat{\Psi}_{M/G(Pareto)/1}$	
	T_r = 0.1			T_r = 0.5		
ρ \ α	-	1.5	5000	-	1.5	5000
0.1	5649	16537	5650	5632	16385	5633
0.2	2953	8350	2953	2949	8316	2949
0.3	2056	5644	2057	2055	5631	2055
0.4	1609	4296	1609	1608	4290	1608
0.5	1340	3488	1341	1340	3485	1340
0.6	1162	2950	1162	1161	2948	1162
0.7	1034	2566	1034	1034	2565	1034
0.8	938	2279	938	938	2278	938
0.9	864	2055	864	864	2055	864
평균	1953	5352	1956	1953	5328	1954
	$\hat{\Psi}_{M/G/1/PS}$	$\hat{\Psi}_{M/G(Pareto)/1}$		$\hat{\Psi}_{M/G/1/PS}$	$\hat{\Psi}_{M/G(Pareto)/1}$	
	T_r = 0.7			T_r = 1.0		
ρ \ α	-	1.5	5000	-	1.5	5000
0.1	5631	16375	5632	5630	16367	5631
0.2	2949	8314	2949	2948	8312	2949
0.3	2055	5630	2055	2055	5630	2055
0.4	1608	4289	1608	1608	4289	1608
0.5	1340	3485	1340	1340	3484	1340
0.6	1161	2948	1162	1161	2948	1162
0.7	1034	2565	1034	1034	2565	1034
0.8	938	2278	938	938	2278	938
0.9	864	2055	864	864	2055	864
평균	1953	5327	1954	1953	5325	1953

이제 다음의 차이율을 정의한다.

$$diff = \frac{\hat{\Psi}_{M/G(Pareto)/1} - \hat{\Psi}_{M/G/1/PS}}{\hat{\Psi}_{M/G(Pareto)/1}} \times 100\ (\%) \quad (15)$$

〈Table 3〉의 각 값에 식 (15)를 적용하면 전체 평균 $diff$는 25%가 되어 두 모델의 최대 객체 크기의 차이가 근사함을 유추할 수 있다. 물론, 최대 객체 크기는 성능 파라미터값에 의해 영향을 받으므로 이 결과의 일반화에는 주의가 필요하다.

5. 결론

본 연구에서는 다중 사용자가 동시에 웹 서버에 접속할 때 평균 응답 시간의 서비스 품질을 만족하는 최대 객체 크기를 구하였다. 수치 계산을 위해 서비스 분포로 M/G/1/PS 모델과 M/G(Pareto)/1 모델을 사용하여 최대 객체 크기를 계산한 결과, Pareto 분포의 형상 파라미터값이 커짐에 따라 두 모델의 최대 객체 크기가 같아짐을 확인할 수 있었다. 본 연구의 결과는 웹 서비스 품질을 만족하는 객체의 크기를 조절하는 데 사용가능하며 앞으로의 연구에서는 더욱 정확한 수리적 모델의 개발이 기대된다.

REFERENCES

[1] M. Hachol-Balter, Performance Modeling and Design of Computer Systems, pp.13-15, Cambridge Press, 2013.

[2] S. Ross, Introduction to Probability and Statistics for Engineers and Scientists, pp.538-540, Academic Press, 2012.

[3] M. E. Crovella and A. Bestavros, "Self-similarity in

World Wide Web Traffic: Evidence and Possible Causes", IEEE/ACM Transactions on Networking, Vol.5, No.6, pp.835-846, 1997.

[4] W. Stallings, High-speed Networks, pp.182-198, Prentice Hall, 2002.

[5] V. Paxson and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling," IEEE/ACM Transactions on Networking, Vol.3, No.3, pp.226-244, 1995.

[6] V. Paxson, "End-to-End Routing Behavior in the Internet," ACM SIGCOMM Communication Review, Vol.36, No.5, pp.41-56, 1996.

[7] J. Reeds and M. Jorgensen, "The Double Pareto-Lognormal Distribution- A New Parametric Model for Size Distributions," Communications in Statistics- Theory and Methods, Vol.33, No.8, pp.1733-1753, 2004.

[8] Y. M. Tripathi, C. Petropoulos, and M. Jha, "Estimation of the Shape Parameter of a Pareto Distribution," Communications in Statistics- Theory and Methods, Vol.47, No.18, pp.4459-4468, 2018.

[9] R. Khayari, R. Sadre and B. R. Haverkort, "Fitting World-wide Web Request Traces with the E-M-algorithm", Performance Evaluation, Vol.52, No.2, pp.175-191, 2003.

[10] A. Riska, V. Diev and E. Smirni, "Efficient Fitting of Long-tailed Data Sets into Hyper-exponential Distributions," in Proceedings of the IEEE Global Telecommunication Conference, Vol.3, pp.3513-2517, 2002.

[11] H. Lee, Queuing Systems for Engineers, Hong-Reung, 2008.

[12] Y. Lee, "Mean Object Size Comparison of M/G/1/PS and TDM System," ICIC Express Letters, Vol.12, No.5, pp.417-423, 2018.

[13] Y. Lee, "On the Comparison of Mean Object Size in M/G/1/PS Model and M/BP/1 Model for Web Service," International Journal of Internet, Broadcasting and Communication, Vol.14, No.3, pp.1-7, 2022.

[14] Y. -J. Lee, "Mean Object Size considering Average Waiting Latency in M/BP/1 System", International Journal of Computer Networks and Communications, Vol.12, No.5, pp.73-80, 2020.

[15] Y. Lee, "Maximum Web Object Size satisfying M/G/1 Queueing Delay constraint in Multiple User Access Environment", The Journal of Korean Institute of Information Technology, Vol.12, No.6, pp.119-124, 2014.

이 용 진(Yong-Jin Lee) [종신회원]

■ 2005년 9월 ~ 현재 : 한국교원대학교 기술교육과 교수

〈관심분야〉

인터넷 기술, 모바일 컴퓨팅, 성능평가