# 웹 서비스 품질 보증을 위해 CBQ 모델을 사용한 클러스터 기반 웹 서버

김신형0 윤완오 정진하 최상방
인하대학교 전자공학과

# Cluster-Based Web Server using CBQ Model to Guarantee Quality of Web Service

Shin-Hyeong Kim0, Wan-oh Yoon, Jin-Ha Jung, Sang-Bang Choi
Dept. of Electronic Engineering, Inha University

**요 약**

인터넷의 급속한 성장과 함께 웹을 기반으로 하는 서비스들이 더욱 확산되며 중요해지고 있다. 하지만 제한된 네트웍 환경에서 웹 트래픽의 지수적인 증가로 인해 웹 서비스의 품질 보장 문제가 대두되고 있지만 현재의 인터넷은 최선의 서비스(Best-effort service)만을 제공하며, 품질 보장형의 서비스(QoS)는 제공하지 못하고 있는 것이 현실이다. 본 논문에서는 기존에 구축되어진 웹 클러스터 모델 중에서 서버의 부하 분산을 담당하고 있는 디스패처(Dispatcher)에 대기 정렬 큐잉(Class Based Queuing)의 패킷 전송 모델을 적용하고자 한다. 제안된 모델을 통하여 클러스터 기반 웹 서비스에서 원하는 클래스의 서비스 품질을 보장할 수 있다.

## 1. 서 론

인터넷 사용자의 급속한 증가와 더불어 인터넷에서 사용되고 있는 데이터들이 기존의 텍스트나 이미지 같은 작은 사이즈의 데이터에서 오디오, 비디오 같은 멀티미디어 데이터를 많이 이용하게 되었다. 그러나 기존의 네트워크 환경과 웹 서버에서 이러한 실시간을 요구하는 멀티미디어 데이터를 처리하면서 네트워크 병목 현상(bottle neck)과 시스템 부하 등의 문제점이 대두되었고, 그 해결책으로 많은 클러스터링 기술이 사용되어져 왔다. 적은 비용과 확장성이 용이한 클러스터 시스템을 사용하여 시스템 부하 문제를 해결할 수 있었지만 실시간성이 요구되는 스트리밍 데이터를 서비스 하는데 있어서 서비스 품질(QoS)을 보장할 수는 없다.

이런 이유로 인하여 웹 클러스터 분야에서 서비스 품질 문제가 대두 되었으며, 웹 서비스 품질 보증(QoWS)을 이루기 위하여 사용자들과 각각의 서비스 등급별 분류(Service classification), 성능 구분(Performance isolation), 사용자들의 요청 승인(Request admission)등의 방법이 사용되어져 왔다 [1]. 본 논문에서는 웹 서비스 품질 보증을 위한 방법의 일환으로 웹 서버에서 제공할 수 있는 몇 가지 서비스들을 각 클래스별로 나누고 우선순위를 할당하여 차등화 서비스(DiffServ) 모델을 이루고자 한다. 효율적인 패킷 전송을 위해 대기 정렬 큐잉(Class Based Queuing)을 웹 클러스터의 부하 분산 서버에 적용하여 멀티미디어 같이 실시간 서비스가 필요한 데이터에 대해 서비스 품질을 보장하는 차별화 된 웹 서비스 모델을 제안하고자 한다.

## 2. 웹 서비스의 품질 보장을 위한 방법

서버의 부하를 분산시키기 위한 웹 클러스터의 초기 모델은 최선의 서비스(Best-effort Service)로써 품질 보장형의 서비스는 제공해 주지 못하는 단점을 가지고 있었다. 웹 서비스 품질 보증이라는 단어는 네트워크에서 서비스 품질을 보장하기 위한 여러 가지 정책으로부터 파생되었다. 그러나 이것은 웹 시스템의 서버 쪽에 초점을 맞추고 있으며 네트워크에서의 트래픽 관리는 다루지 못하고 있다.

다음에 설명될 내용들은 서버의 부하 분산에 초점을 맞춘 기존의 웹 클러스터 모델에서 웹 서비스 품질 보장을 위해 제안되어진 몇 가지 방법들이다.

### 2.1 분류와 요청 승인

웹 서버에 접속하는 사용자들을 미리 정해진 기준에 의해 몇 개의 클래스로 분류하는 것이다. 이 아이디어는 클러스터가 미리 정해둔 임계 값 이상으로 부하가 걸리면 낮은 클래스의 사용자들의 요구에는 서비스를 거절하며, 시스템 부하가 매우 높을 때에는 높은 클래스의 사용자들에게도 서비스를 거절한다는 것이다. 클래스 분류와 더불어 요청 승인에 대한 조절은 웹 서비스의 성능 저하를 허락하지 않는다는 개념에서 시작되었다. 서버로 들어오는 요청은 일단 부하 분산 서버의 입력 큐에 쌓이게 되며 실제의 서버들로부터 거절 되어진다면 큐에서 지워질 것이다. Layer-7 웹 스위치는 요청 승인 기법을 실현하는 대표적인 것으로서 상대적으로 낮은 클래스에 속하는 사용자들에게는 서비스를 잠시 연기하는 Cherkasova가 제안한 알고리즘을 고려할 수 있다 [2].

### 2.2 성능 분류

성능 분류의 개념은 품질 보증 이론 중에서 가장 중요한 원리 중에 하나이다. 이 원리를 웹 클러스터에서 구현하는 가장 간단한 방법은 이미 정의되어 있는 서비스들을 많은

클래스로 구분 짓는 것이다. 서비스 차별화(Differentiated service) 정책은 웹 서버에서 제공하는 서비스 또는 웹 서버에 접속하는 사용자들에 근거하여 이루어지는데 근본적으로 이 정책은 서버들을 정적으로 구분 짓는 것을 의미하며, 정의 되어진 서비스 클래스를 각각 서버의 작은 집합들로 할당하는 것이다 [3].

### 2.3 높은 자원 활용

사용자들의 요구에 의해 어떤 서버 집합에 부하가 걸려있는 동안 다른 서버 집합이 충분히 사용되지 않고 있을 때에는 자원의 낭비가 생길 수도 있다. 그러므로 서버들을 정적으로 구분 짓는 방법은 변동이 심한 사용자들의 요구와 서버의 부하상태가 급격히 변하는 환경에서는 사용할 수가 없다. 한 단계 더 나아가 부하 상태와 실제적인 사용자들이 요구를 주기적으로 파악해서 웹 클러스터를 동적으로 구분 짓는 방법이 소개가 되었다. 즉, 동적 구분(Dynamic Partitioning) 방법과 요구 주도형의 서비스 차별화 정책 같은 기법을 사용함으로써 자원을 좀더 효율적으로 활용할 수 있게 되었다 [4].

그림 1은 기존의 웹 클러스터 모델로써 사용자들의 분류와 요청에 대한 승인, 성능에 의한 웹 서버의 분류를 보여주고 있다.

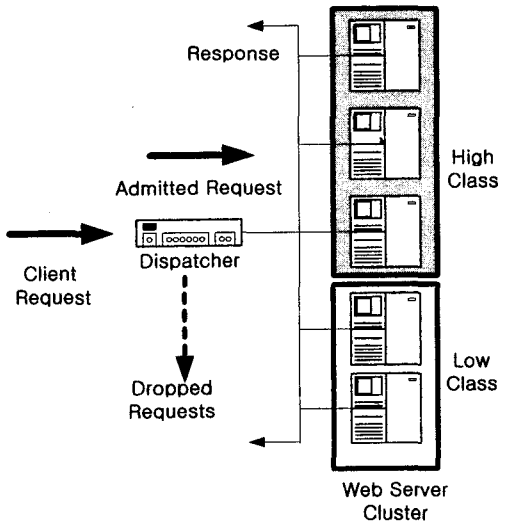


그림 1. QoS를 고려한 기존의 웹 클러스터 모델

## 3. 트래픽 관리

인터넷 환경에서 트래픽 폭주(Congestion)에 대한 해결책은, 현재 이루어지고 있는 최선의 서비스(Best-effort service) 보다는 품질 보증 이론에서 찾아야 한다. 기존의 일반적인 데이터를 위주로 서비스 하던 것과는 달리 시간 지연이 발생했을 때 큰 손실을 가져다 줄 수 있는 멀티미디어 서비스가 점점 더 많은 비중을 차지하면서 웹 서비스에서의 품질 보증은 반드시 고려해야 할 문제가 되었다.

기존의 웹 클러스터의 부하 분산 서버에서 우선순위를 배정하거나 서버들의 성능을 고려하여 부하를 분산시키는 방법은 실제로 많은 문제를 안고 있었다. 무엇보다도 품질 보증 개념을 부하 분산 서버에 추가하여 트래픽을 분류한다는 것은 새로운 기능이 추가된 것으로써 트래픽의 또 다른 지연문제를 야기 시킬 수 있으며, 그런 기능 자체가 병목 현상의 원인이 될 수 있다.

### 3.1 대기 정렬 큐잉(Class-Based Queueing)

트래픽 관리를 위해 큐 관리 기법중의 하나인 대기 정렬 큐잉은 우선순위 큐잉 방법의 변형으로 서비스 클래스에 따라 각각의 큐를 정의하고 있으며, 특정 서비스 클래스가 상대적으로 낮은 우선순위로 인해 자원을 할당 받지 못하여 생기는 굶주림(Starvation) 현상을 막기 위해서 고안되었다. 대기 정렬 큐잉은 우선순위에 기초하여 트래픽의 형태에 따라 큐잉 서비스를 수행하고 특정 서비스의 클래스가 시스템 자원과 대역폭을 독점하는 것을 막음으로써 공평성을 제공하게 된다.

대기 정렬 큐잉 이론의 중요한 의미는 사용자와 서버 사이에 정해진 대역폭이 있을 때 우선순위가 높은 클래스에 의해 독점되는 것이 아니라 다른 응용 프로그램에 의해 사용되어질 수 있다는 것이다 [5]. 대기 정렬 큐잉은 각 큐의 우선순위에 따라 각기 다른 크기의 서비스 한계 값을 두고 이 값에 따라 스케줄링 라운드마다 각 큐를 서비스함으로써 특정 서비스 클래스가 영구히 서비스 받지 못하는 굶주림 문제를 막을 수 있으며 각 패킷들은 상당히 적은 양의 지연 값으로 서비스 되어진다. 즉, 우선순위 큐잉 방법과 비교해 낮은 우선순위를 갖는 트래픽에 대해 자원을 무조건 빼앗는 것이 아니라 약간의 일정한 자원을 할당하여 서비스를 수행하게 된다. 또한 대기 정렬 큐잉은 트래픽을 여러 가지 서비스 클래스로 분류하는 기본적인 방법으로, 각 서비스 클래스에 대해 링크 공유를 제공하여 보다 효율적으로 큐 자원을 관리하는 방법을 제공한다.

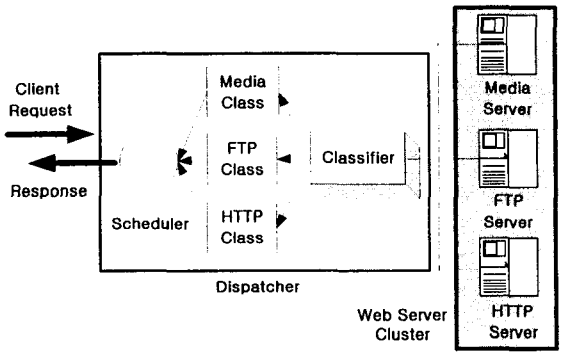


그림 2. 제안된 CBQ에 의한 클래별 웹 서비스 모델

### 3.2 CBQ 기반의 웹 클러스터

그림 2는 본 논문에서 제안하는 대기 정렬 큐잉에 의한 클래스별 웹 서비스 모델이다. 그림 1과 같은 기존의 웹 클러스터 모델에서는 디스패처에서 서버의 부하 상태를 조사하여 들어오는 사용자 요청에 대하여만 처리를 해주었다. 그러나 본 모델에서는 사용자들에 의해 요구된 데이터를

디스패처를 통해 출력이 될 때 각 클래스 별로 우선순위와 대역폭을 할당하여 원하는 클래스의 품질을 보장하도록 한다. 이 모델은 웹 서비스에서 주로 이루어지고 있는 3가지 서비스, 즉 멀티미디어 데이터를 위한 Media 클래스, TCP 기반의 FTP 데이터 전송을 위한 FTP 클래스, 일반적인 웹 데이터 전송을 위한 HTTP 클래스 구조를 가지는 모델이다.

각 과정을 살펴보면, 우선 사용자들의 요청에 의해서 Media 서버, FTP 서버, HTTP 서버에서 각각의 출력 데이터가 생성될 것이다. 이 때, 디스패처의 출력 모드에서는 각각의 서비스별로 큐가 만들어 질 것이며, 이미 정해진 우선순위, 대역폭 같은 변수 값에 의해 각 큐에서 데이터가 처리될 것이다. 본 논문에서는 HTTP 클래스에 50%, FTP 클래스에 30%, Media 클래스에 20%의 대역폭을 우선 할당한다. 비록 Media 클래스의 대역폭이 다른 서비스 클래스에 비해서 작지만, 스케줄러에 의해서 Media 클래스의 데이터가 끊김 현상 없이 출력되도록 모델을 구성한다. 지금까지 설명한 웹 클러스터의 모델과 대기 정렬 큐잉 방법에 의해서 본 논문에서 원하는 Media 클래스의 서비스 품질을 보장할 수 있다. 더 나아가, 실제로 구현을 할 때에는 현재 부족한 IP 주소 문제를 해결하기 위해 NAT(Network Address Translation)을 사용한 LVS(Linux Virtual Server)을 기본 모델로 잡고 있으며, 리눅스 가상 서버 중에서 부하 분산 서버로 사용되고 있는 디스패처(Dispatcher)에 CBQ 패킷 스케줄링 기법을 적용하고자 한다.

## 4. 모의 실험 및 결과

모의실험에서는 대기 정렬 큐잉을 적용하지 않은 기존 모델과 대기 정렬 큐잉을 적용한 두 모델을 비교하여 각 클래스의 큐에서 처리되는 처리량을 비교하여 원하는 클래스의 서비스 보장이 이루어지고 있는지 확인 하고자 한다.

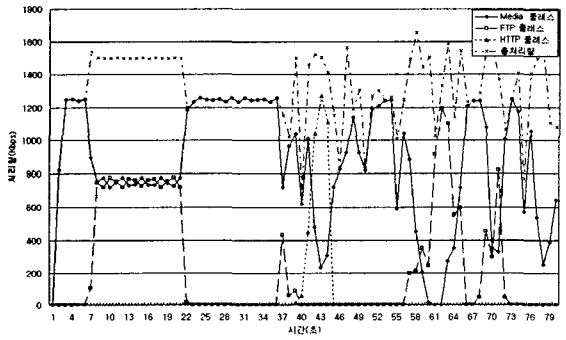

그림 3. 기존 웹 클러스터의 클래스별 처리량

그림 3은 NAT 방식으로 구성된 웹 클러스터 모델에서의 각 서비스별 처리량(Bps)을 보여주고 있다. 여기서 FTP 클래스나 HTTP 클래스에서 사용자로부터 요구가 있어 데이터의 출력이 갑자기 많아질 경우에는 Media 클래스의 서비스가 보장되지 못함을 볼 수 있다.

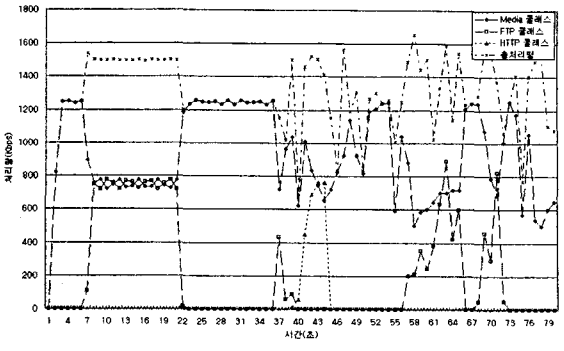

그림 4. CBQ를 적용한 클러스터의 클래스별 처리량

그림 4는 이 논문에서 제안하는 모델로써 NAT 방식의 클러스터 모델에 대기 정렬 큐잉를 적용하여 각 서비스 클래스별로 우선순위를 할당한 후의 모의실험 결과이다. 이번 결과에서는 다른 서비스 클래스의 데이터가 사용자 요청에 의해 출력이 갑자기 많아질 경우에도 Media 클래스의 서비스는 일정한 값 이상의 처리량을 보여줌으로써 웹 서비스의 품질 보장이 이루어지고 있음을 알 수 있다.

기존 웹 클러스터의 부하 분산 서버에서는 주기적으로 서버들의 부하 상태를 점검하기 때문에 실제로 처리속도의 지연이 발생하였고, 정확한 서버의 부하 상태를 알아내기가 힘들었다. 또한 데이터의 무분별한 출력으로 인해 손실이 컸던 것이 사실이다. 그러나 본 논문을 통해 기존의 웹 콜러스터 모델에 고가의 하드웨어의 추가 없이도 사용자들이 원하는 서비스 클래스의 품질을 보장해 줄 수 있을 것이다.

## 참 고 논 문


[1] V. Cardellini, E. Casalicchio, M. Colajanni, and M. Mambelli, "Web switch support for differentiated services," *ACM Performance Evaluation Review*, Vol. 29, No. 2, pp. 14-19, Sept. 2001.

[2] L. Cherkasova and P. Phaal, "Session based admission control: A mechanism for improving performance of commercial web site," In *Proc. Int'l workshop on Quality of Service*, London, June 1999

[3] Resonate Inc., Central Dispatch, http://www.resonate.com.

[4] H. Zhu, H. Tang, and T. Yang, "Demand-driven service differentiation in cluster-based network servers," In *Proc. IEEE Infocom 2001*, Anchorage, Alaska, Apr. 2001.

[5] Alistair Croll and Eric Packman, Managing Bandwidth, Prentice Hall, 2000.