Available online at www.sciencedirect.com
ScienceDirect
Computer Communications 31 (2008) 1993–2004
computer communications
www.elsevier.com/locate/comcom

Multiple-resource request scheduling for differentiated QoS at website gateway

Ying-Dar Lin, Ching-Ming Tien, Shih-Chiang Tsao *, Ruo-Hua Feng, Yuan-Cheng Lai

National Chiao Tung University, Department of Computer Science, Room 701, EIC Building, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan

Received 1 February 2007; received in revised form 29 December 2007; accepted 6 January 2008
Available online 26 January 2008

Abstract

Differentiated quality of service is a way for a website operator to provide different service levels to its clients. Traditional HTTP request scheduling schemes can achieve this, but they schedule requests to manage only one server resource, such as CPU or disk I/O. Actually, processing a request on the server will consume multiple resources. This paper presents a multiple-resource request scheduling algorithm, called mQoS, for differentiating the utilization of the server resource. The mQoS scheduler consists of several sub-schedulers and a main scheduler. Each sub-scheduler manages a server resource to differentiate its utilization among the classes. The main scheduler checks the availability of every server resource and triggers an appropriate sub-scheduler to balance the utilization of server resources. The implementation of the mQoS gateway is based on Squid and Linux. The evaluation compares the mQoS scheduling with no scheduling (nQoS) and single-resource request scheduling (sQoS). The mQoS scheduling reveals the accurate differentiation on every server resource. In addition, the total server throughput in the mQoS scheduling is improved by 21%, compared with the sQoS scheduling. The average user-perceived latency of the mQoS scheduling is also shorter than other scheduling.

Keywords: Multiple resources; Request scheduling; Service differentiation

1. Introduction

Web quality of service (QoS) is a way for a Web service provider to differentiate its service levels to users. Through *service differentiation*, a Web service provider can allow a specific group of users, e.g., paid users, to get better server throughput or user-perceived latency than other general users. There are many ways of enforcing Web QoS. The effort of some past researches was to modify the system kernel or the server daemon of a Web server, a caching proxy, or a cluster dispatcher for service differentiation. These QoS-enabled boxes intercept HTTP requests, perform request classification and *request scheduling* for dealing with the bottlenecked resource, such as bandwidth or processing power.

There are two issues in the above schemes. The first issue is where to deploy a QoS-enabled box. Many researches have been proposed in modifying the system kernel [1] or server daemon [2,3] of a Web server to have the capability of scheduling HTTP requests. However, this solution is hard to be deployed on a non-open operating system or server daemon. Some researches have been proposed in enforcing request scheduling on a dispatcher of a cluster server [4–6]. The QoS-enabled dispatcher schedules requests to the backend servers in a weighted round-robin fashion or according to the server loads. Some researches have proposed QoS-enabled content adaptation [7,8] or cache replacement algorithms [9] on caching proxies instead of request scheduling for service differentiation. The second issue is what resource for a request scheduling to manage. Common request scheduling schemes schedule requests by managing the bottlenecked resource, such as bandwidth or processing power. These request schedulers seem to perform the single-resource scheduling, which has a blind spot.

* Corresponding author. Tel.: +886 3 5731899.
E-mail address: weafon@cs.nctu.edu.tw (S.-C. Tsao).

0140-3664/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2008.01.010

Processing a request on a server needs to consume multiple resources, e.g., CPU, disk I/O, and bandwidth, rather than a single resource. In the single-resource scheduling, some resources may be wasted, when the managed resource is well utilized. A request scheduler should well utilize all resources by scheduling requests for managing all resource utilization. Some researches have discussed *multiple-resource request scheduling*, but many of them are applied on grid computing and multimedia applications [10–12], and few on HTTP request scheduling [13–15].

Considering the issues of QoS deployment and multiple-resource request scheduling, this paper presents a multiple-resource request scheduling algorithm called mQoS, which is deployed at a *website gateway* for controlling the requests toward a Web server. Today's gateways can perform firewall packet inspection, intrusion detection, virus scanning, and so on. A website operator can deploy a gateway for preventing attacks and providing value-added services. Hence, enforcing request scheduling at a website gateway is practical, and that can provide service differentiation without any modification on clients and the server.

There are three main functions in the mQoS gateway: *request profiling and server profiling*, *content-aware request classification*, and mQoS scheduling. The request profiling finds out the amounts of the server resources consumed by a request, whereas the server profiling measures the capacities of the server resources. The request classification mechanism inspects the headers or payloads of requests and puts requests into proper class queues. Specially, a service class has several queues, each of which stores specific resource-intensive requests. That is, when m service classes and n server resources exist, there are $m * n$ queues. The mQoS scheduling, derived from the *Deficit Round Robin* (*DRR*) scheduling [16], composed of one main scheduler and several sub-schedulers. One sub-scheduler, which has some deficit counters, manages one server resource. However, differing from the traditional DRR scheduling, the deficit counter of a class in a sub-scheduler can be decremented by any sub-scheduler because a request would consume multiple resources rather than a single resource. In addition, the main scheduler maintains the availability of the server resources in the resource availability counters. The main scheduler hence knows which resource is the most available and then triggers the corresponding sub-scheduler to service specific resource-intensive requests.

The mQoS gateway is implemented on Squid and Linux. The request and response modules of Squid are modified to be capable of classifying and scheduling requests. In the evaluation, the mQoS scheduling is compared with no scheduling (nQoS) and single-resource request scheduling (sQoS). The resource utilization, server throughput, and user-perceived latency of every scheduling algorithm are measured to demonstrate the effect of the mQoS scheduling. From the test results, the mQoS scheduling reveals its capabilities of differentiating server resource utilization, maximizing the total server throughput, and sharing resource.

The rest of this paper is organized as follows. Section 2 states the problems of resource management on a Web server. Section 3 introduces the architecture of the mQoS gateway and the designs of the request profiling and server profiling, content-aware request classification, and mQoS scheduling algorithm. Section 4 describes the implementation and evaluation of the mQoS gateway. Finally, Section 5 gives the conclusion and the future work of this research.

2. Problems of server resource management

The workload on a Web server will affect the utilization of the server resource. In a light-load situation, every HTTP request will get enough resources when being processed, but there could be unused resources on the server. Conversely, in a heavy-load situation, a request may be queued on the server and wait for being processed. If the server resources are inadequate for the requirements of the arrival requests, an HTTP request would experience long queuing and processing delay. For maximizing the utilization of the server resources and avoiding extra delay simultaneously, the resources on the server should be well managed.

Some researches have proposed admission control schemes to prevent new arrival requests from accessing a heavy loaded server [17–20]. With admission control, a server would drop new arrival requests when its resources cannot meet the requirements of the requests. However, admission control itself is not sufficient to support service differentiation because all arrival requests have the same probability to access server resources. The purpose of service differentiation is to allow different clients receive different treatments, such as server throughput and response time. For service differentiation, some researches have proposed request scheduling algorithms to manage the workload on a server [1,2,18,21,22]. The general schemes of the mentioned scheduling algorithms are to allocate different amounts of concurrent connections, request rate, or bandwidth among service classes.

A request entering a server requires several types of resources, e.g., CPU, disk I/O, and bandwidth, when being processed. The lack of any available resource would lead to a bottleneck. In other words, if there are n kinds of resources, there could be n kinds of bottlenecks on the server. Many of the mentioned request scheduling algorithms deal with the problems of single-resource scheduling. They manage a single resource for maximizing its utilization and differentiating its utilization simultaneously, but they cannot avoid the bottlenecks derived from the other resources. A resource can be managed well, while the other resources may be still non-fully utilized or inadequate for new arrival requests. Thus, a single-resource scheduling algorithm could lead to an inefficient or overloaded server. Actually, a request scheduling algorithm should consider the presence of multiple server resources. In the below, three request scheduling schemes, no scheduling, single-resource request scheduling, and multiple-resource request scheduling, are discussed. The assumption for the discussion is that

there are three resources, CPU, disk I/O, and bandwidth, on the server and a request will consume multiple resources. Besides, there are three service classes of clients issuing requests to the server, and the heavy-load situation is considered.

2.1. No scheduling (nQoS)

The nQoS scheduling is without any resource management scheme, such as admission control or request scheduling, enforced for the service differentiation. The requests originated from the three classes of clients contend for the server resources. The server works on a first-come-first-serve basis. The server workload of the nQoS scheduling is shown in Fig. 1a. The vertical axis stands for the resource utilization and c1, c2 and c3 stand for the class 1, class 2 and class 3, respectively. Due to the resource contention, every class of clients gets a third of each server resource. All server resource utilization is affected by the workload, but there is no any service differentiation. The pending requests would be queued on the server and wait for being processed, causing extra resource consumption, and prolonged user-perceived latency.

2.2. Single-resource request scheduling (sQoS)

In the sQoS scheduling, a request scheduler manages the utilization of one server resource. Fig. 1b shows the server workload of the sQoS scheduling. The CPU resource is managed for service differentiation, and the ratio of the resource allocated to the three classes of clients is 6:3:1. In this example, the sQoS scheduling indeed allocates the expected amount of the CPU resource to the three classes of clients, but it cannot take care the utilization of the other resources. The sQoS scheduling will stop scheduling any request to the server when the CPU resource is well utilized. However, the disk I/O and bandwidth resources are actually still affordable for the new arrival disk I/O- and bandwidth- intensive requests, respectively, causing the waste of these resources. Conversely, the sQoS scheduling will keep scheduling requests to the server when it finds the CPU resource is available. However, the disk I/O and bandwidth resources may be already fully utilized, causing an overloaded server and potentially prolonged user-perceived latency.

2.3. Multiple-resource request scheduling (mQoS)

In the mQoS scheduling, a request scheduler manages all server resources. The server workload of the mQoS scheduling is shown in Fig. 1c. The mQoS scheduling chooses the appropriate requests to well utilize all resources and at the same time allows the three classes of clients to use every resource proportionally. The mQoS scheduling eliminates the resource wasting or server overloading occurred in the sQoS scheduling, and the total server throughput can be improved. Due to scheduling the proper requests to the server, each resource utilization under the mQoS scheduling is better than that under the nQoS scheduling. The mQoS scheduling further avoids resource contention and enables service differentiation.

In the above discussion, the mQoS scheduling seems to be a better solution for server resource management. In this paper, a mQoS scheduling algorithm for service differentiation is presented. The mQoS scheduling algorithm has the capability of managing multiple server resources and it is deployed on a website gateway located in front of a Web server. The arrival requests are queued and wait for being scheduled on the mQoS gateway instead of the server. This has the advantage of avoiding extra resource consumption on the server. The server itself can concentrate on the request processing only.

3. The mQoS gateway architecture and scheduling algorithm

The purpose of the mQoS gateway is to avoid resource bottlenecks, provide differentiation of resource utilization, and maximize the server throughput. To do this, the mQoS gateway performs three tasks: request profiling and server

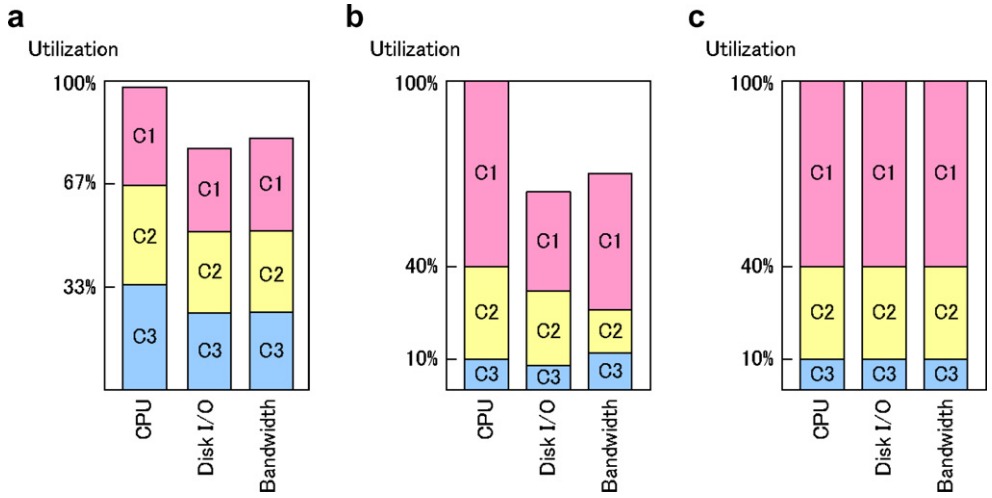

Fig. 1. Server resource utilization under different scheduling schemes. (a) nQoS scheduling. (b) sQoS scheduling. (c) mQoS scheduling.

profiling, request classification, and request scheduling. The request profiling and server profiling let the mQoS gateway know the resource consumption of a request and the capacity of each server resource. The request classification allows the mQoS gateway to classify requests into different service classes. The request scheduling determines the order and the time in which the mQoS gateway sends a request to the server.

The architecture of the mQoS gateway, as shown in Fig. 2, is composed of three components: server prober, request classifier, and request scheduler. The working flow of the gateway is described as follows. Before the on-line operation of the gateway, the server prober sends HTTP requests one by one to scan all Web pages on the server. The resource monitor program running on the server monitors the resource consumption for every request and reports this information to the server prober. The server prober records the URLs and resource consumption of the Web pages in the Web page table for the reference of the request classifier. The QoS policy table defines the service classes and their classification rules. Once the gateway starts to work, it incepts arrival requests. The request classifier classifies the incepted requests into different service classes according to the rules defined in the QoS policy table. Then the request classifier refers to the Web page table, tags the information of the resource consumption to each request, and puts the tagged requests into the corresponding queues. The request scheduler checks the availability of the server resources. If the available server resources are enough, the request scheduler fetches a request from a proper queue and sends it to the server. The detailed design of the server prober, request classifier, and request scheduler are described below.

3.1. Server prober

The mQoS gateway is deployed in front of any type of Web servers. The gateway has to know the server resource consumption of a request and the capacity of each server resource. For this, the server prober is used for request profiling and server profiling. The request profiling is the process of measuring the resource consumption of a request, whereas the server profiling is the process of measuring the maximum capacity of each server resource.

For measuring the resource consumption of a request, the server prober sends HTTP requests one by one to scan all Web pages on the server. Starting from the homepage, the server prober recursively parses every Web page and finds the URLs of the embedded objects and hyperlinks until the website is traversed. During the traversing, the monitor program running on the server monitors the amounts of server resources consumed for each request and reports this information to the server prober. As an example, a query page consumes 15 U of CPU, 5 U of disk I/O and 8 U of bandwidth per second. To increase the validity of the measurement, the probed results are verified before being used. That is, when the prober sends multiple requests to the server concurrently, the amount of the resource consumption is multiplied as the number of concurrent requests being processed on the server. Notice that this information is not directly used by the request scheduling algorithm because the actual percentage of the resource consumption is not known yet.

In order to calculate the percentage of the resource consumption of a request, the server prober has to measure the maximum capacity of each server resource. Thus, the server prober sends huge amount of specific resource-intensive requests at the same time to the server and checks the resource utilization. The maximum capacity can be measured when the resource is fully utilized. After all resource capacities are measured, the actual capacities of the server resources and the percentages of the resource consumption of a request are derived. The maximum capacity of a server resource can be derived from multiplying the number of the concurrent requests on the server by the resource consump-

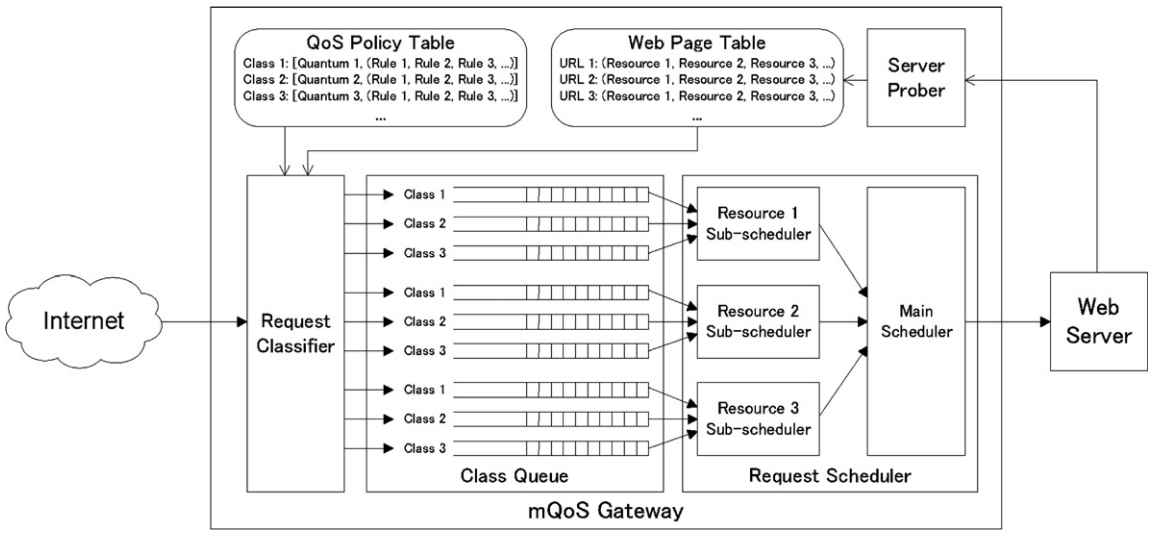

Fig. 2. Architecture of the mQoS gateway.

tion of a request. As an example of measuring the CPU capacity, if there is 100 requests being processed by a fully loaded server and the CPU resource consumption of each request is 15 U, then the maximum CPU capacity is 1500 U. The percentage of the CPU resource consumption of a request can be also derived from dividing its CPU resource consumption by the CPU capacity. In the above example of a query page, its percentage of the CPU resource consumption is 1% (derived from 15/1500). The server prober finally records the URLs and resource consumption information in the Web page table for the use of the request classifier and request scheduler.

Notably, although the server may generate different pages for a request due to the parameters attached with the request. However, since the server usually uses the same program to serve a request, the amount of resources spent on generating these pages is similar [20]. Therefore, even though the server may have infinite pages, it has finite requests and scanning all requests in a server to get the resources spent on each request should be practicable. However, this work does not consider the influences of caching and prefetching techniques on the resource consumption of requests. The issue about the influences deserves to an individual study.

3.2. Content-aware request classifier

The request classifier is used to identify the class and the resource tendency for each request. The classification is based on the predefined rules in the QoS policy table. The header and payload of a request will be inspected by the request classifier to check whether it matches a rule. If yes, the request will be classified into this corresponding class; otherwise, it will be compared with the other rules until classified. Once a request is classified, its URL will be inspected to match the URLs in the Web page table. The purpose is to find out the expected resource consumption and judge the tendency of the resource consumption. For example, a request consuming 9% of CPU, 5% of disk I/O and 7% of bandwidth is regarded as a CPU-intensive request. After a request is matched with the QoS policy table and Web page table, the request classifier tags the information of the resource consumption to this request and put it into an appropriate queue. Every service class has several queues, each of which stores specific resource-intensive requests. If there are m service classes and n server resources, there are totally $m * n$ queues. The requests wait in the queues for being scheduled by the request scheduler.

3.3. Multiple-resource request scheduler

The request scheduler schedules the requests in the class queues to manage the server resources in order to provide service differentiation. The key idea of the mQoS scheduling is derived from the deficit round robin (DRR) scheduling for packet scheduling. A traditional DRR scheduler serves the head-of-line (HOL) packet of every non-empty queue which the value of the deficit counter is greater than the packet size. If it is lower, then later the deficit counter is incremented by a given value called quantum. A deficit counter is decremented by the size of packets being served. However, some considerations should be noticed on scheduling requests using the concept of the DRR scheduling. The traditional DRR schedules packets to manage the bandwidth of a link, whereas the presented mQoS scheduler schedules requests to manage the multiple resources of a server. The utilization of the server resources has to be balanced. None of the resources should be overused or underused; otherwise a resource bottleneck would happen or a server resource would be wasted.

The mQoS scheduler consists of a main scheduler and several sub-schedulers, as shown in Fig. 3. A sub-scheduler

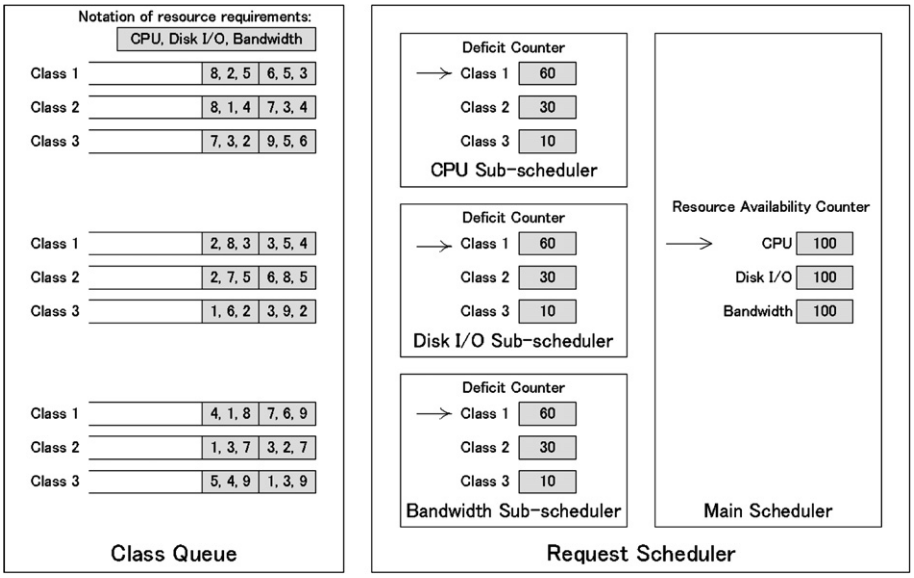

Fig. 3. mQoS scheduler.

services the class queues of a server resource for differentiating the resource utilization among the classes, and the main scheduler triggers an appropriate sub-scheduler according to the availability of the server resources. In a sub-scheduler, there are several deficit counters (DCs), each of which is associated with a class to record the unused quantum. However, differing from the traditional DRR scheduling, the DC of a sub-scheduler can be decremented by any other sub-schedulers because a request would consume multiple resources rather than a single resource. Each sub-scheduler has a round-robin pointer that indicates which class queue to be serviced. When the round-robin pointer moves back to the first class queue, every DC of this sub-scheduler is incremented by the predefined quantum.

In the main scheduler, resource availability counters (RACs) are used to record the availability of the server resources. Each RAC contains the percentage of the availability of a server resource. By checking the RACs, the main scheduler knows which resource is the most available and then triggers the corresponding sub-scheduler to service a specific resource-intensive request. Therefore, the main scheduler can maximize the resource utilization and balance the utilization among the resources.

3.4. Multiple-resource request scheduling algorithm

The mQoS scheduling algorithm works as follows. Initially, the value of each RAC is set to 100, which means each type of server resource is 100% available. Each round-robin pointer in these sub-schedulers moves to the first class queue. In the traditional DRR scheduling, a DC is incremented only when the round-robin pointer moves to its corresponding queue. However, here all DCs of a sub-scheduler are incremented at the same time by the predefined quantum because the DC of a sub-scheduler could be decremented by another sub-scheduler. The main scheduler checks the values of the RACs to find out which resource is the most available. A sub-scheduler will be triggered for scheduling the corresponding resource-intensive requests to effectively utilize the most available resource. The main scheduler randomly triggers a sub-scheduler, when there is no resource more available than the others.

The triggered sub-scheduler inspects the resource consumption information of the HOL request of the queue which the round-robin pointer locates. If no request waits in this queue, the sub-scheduler moves the round-robin pointer to the next queue and the remaining deficit will be carried over to the next service cycle in the DC. The resource requirements of this request are then compared with the values of the RACs. If any resource is not enough, the sub-scheduler will move the round-robin pointer to the next queue without scheduling this request. If all resource requirements are satisfied, the sub-scheduler will check the values of the DCs of the same class from all the sub-schedulers to see whether this class has enough values in the DCs. If no, the sub-scheduler will move the round-robin pointer to the next queue without scheduling the request. If yes, the sub-scheduler fetches the request from the queue, decrements the amounts of the resource requirements from the DCs and RACs, and sends this request to the server. When the response from the server comes back, the RACs will be incremented by the amounts of the resource requirements of the corresponding request to reflect the releasing of the consumed resources. The main scheduler continues to trigger a sub-scheduler. A sub-scheduler continues to serve the requests from a queue until the queue becomes empty, or the resource requirements cannot be satisfied. Since the scheduler has to be aware

```
/* definitions of variables
r: number of resources
c: number of classes
DC: deficit counter
RAC: resource availability counter
RRP: round-robin pointer
RR: resource requirement
Q: quantum */
-----------------------------------------------------
Initialization module:
For (i = 0; i < r; i = i + 1)
  RAC_i = 100;
  move_pointer(RRP_i, 0);
  For (j = 0; j < c; j = j + 1)
    DC_ij = 0;
-----------------------------------------------------
Request enqueuing module: on arrival of request p
j = get_class(p);
RR = get_resource_requirements(p);
k = get_resource_tendency(RR);
enqueue(Queue_kj, p);
```

Fig. 4. Pseudo code of the mQoS scheduling algorithm.

of the responses, the mQoS scheduler is not proper to work with direct routing.

The pseudo code of the mQoS scheduling algorithm is shown in Figs. 4 and 5. Some details are ignored in the pseudo code. The enqueuing module performs the request classification to put a request into an appropriate queue. The dequeuing module executes the mQoS scheduling algorithm to schedule the requests in the class queues. The response processing module checks the finish of a response and increments the RACs.

Fig. 6 exhibits an example of the mQoS scheduling. In this example, the requests are classed into three service classes: class 1, class 2, and class 3. The ratio of the service weights of classes is set to 6:3:1, hence the quantum assigned to each class is 60, 30 and 10, respectively. The server resources to be managed are CPU, disk I/O, and bandwidth. Because there are three service classes and three server resources, totally nine class queues exist. The initial stage is shown in Fig. 3. The main schedules randomly triggers the CPU sub-scheduler. The CPU sub-scheduler inspects the HOL request of the class-1 queue and knows the resource requirements of this request are (CPU: 6, disk I/O: 5, bandwidth: 3). The CPU sub-scheduler compares the amounts of the resource requirements to the values of the RACs (CPU: 100, disk I/O: 100, bandwidth: 100) and concludes the server resources are enough. Then it compares the resource requirements to the values of the DCs of the CPU, disk I/O, and bandwidth sub-schedulers for class 1 (CPU: 60, disk I/O: 60, bandwidth: 60) and concludes the values in the DCs are enough. The CPU sub-scheduler now sends the request to the server and decrements the DCs and RACs. The results of the decrements on the DCs and RACs are shown in Fig. 6a. Now the main scheduler triggers the bandwidth sub-scheduler because the bandwidth resource is the most available. The bandwidth sub-scheduler sends the HOL request of the class-1 queue to the server. The result after this request scheduling is shown in Fig. 6b. Now the disk I/O resource becomes the most available, hence the main scheduler triggers the disk I/O sub-scheduler to send a request. Suppose the server has finished responding the first request after the request sent by the disk I/O sub-scheduler. The final values of the RACs and DCs are shown in Fig. 6c.

4. Implementation and evaluation

4.1. Implementation

The implementation of the mQoS gateway is based on the Squid package and Linux operating system. The Squid

```
Request dequeuing module:
While(TRUE)
  m = get_most_available_resource(RAC);
  If (new_round_of_scheduling == TRUE) then
    For (j = 0; j < c; j = j + 1)
      /* increment deficit counters by quanta */
      DC_mj = DC_mj + Q_j;
  RR = get_resource_requirements(p);
  For (i = 0; i < r; i = i + 1)
    /* check resource availability and deficits */
    If ((RAC_i < RR_i) or (DC_ij < RR_i)) then
      If (j == c - 1) then
        /* move round-robin pointer to the first class */
        move_pointer(RRP_m, 0);
      Else
        /* move round-robin pointer to the next class */
        move_pointer(RRP_m, j + 1);
    Else
      For (i = 0; i < r; i = i + 1)
        /* decrement deficit counters */
        DC_ij = DC_ij - RR_i;
        /* decrement resource availability counters */
        RAC_i = RAC_i - RR_i;
      send_request(p);
-------------------------------------------------------
Response processing module: on arrival of response q
For (i = 0; i < r; i = i + 1)
  RR = get_resource_requirements(q);
  /* increment resource availability counters */
  RAC_i = RAC_i + RR_i;
send_response(q);
```

Fig. 5. Pseudo code of the mQoS scheduling algorithm.

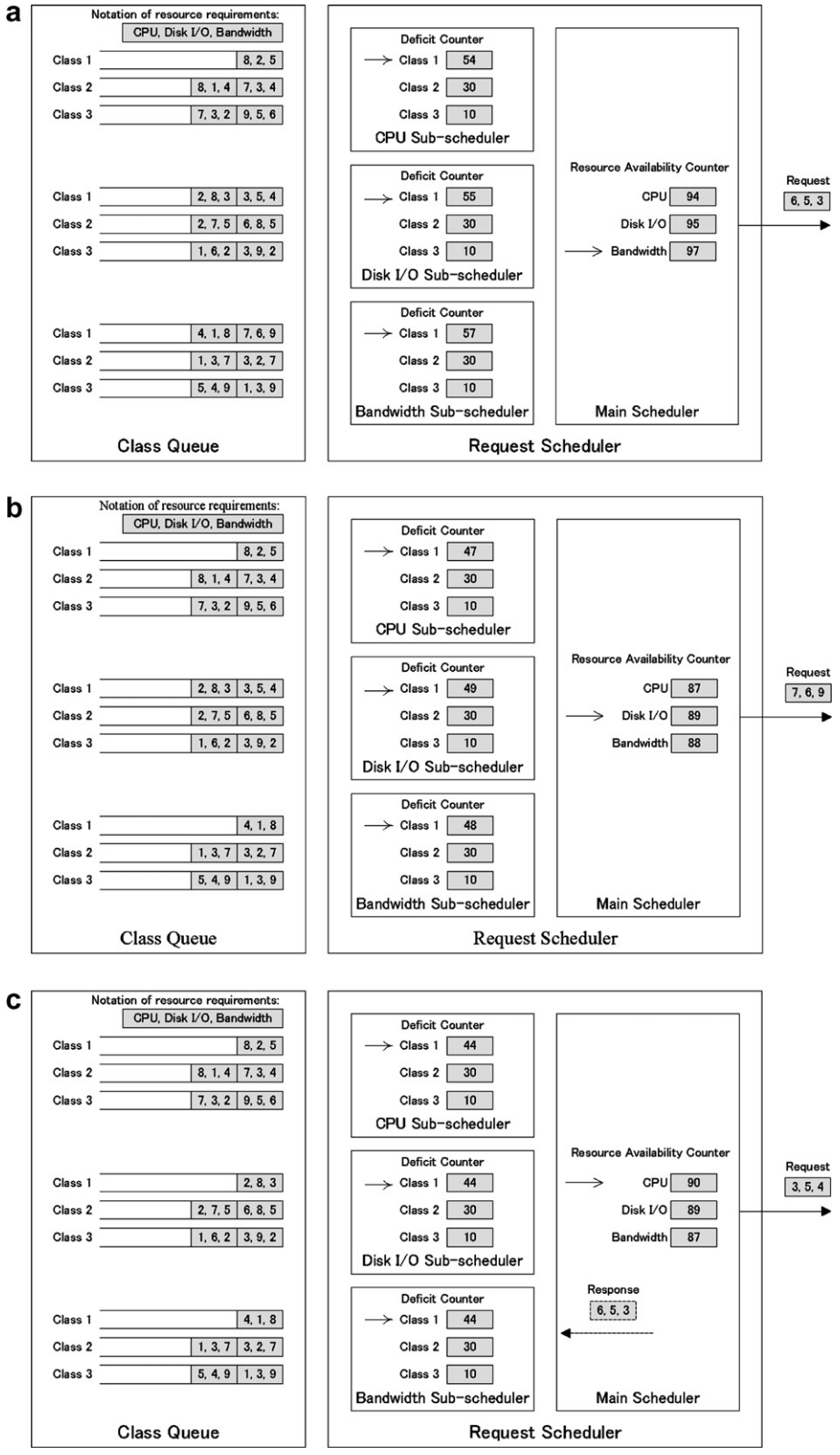

Fig. 6. An example of the mQoS scheduling. (a) The CPU sub-scheduler sends a request. (b) The bandwidth sub-scheduler sends a request. (c) The disk I/O sub-scheduler sends a request and then a response returns.

package is modified to be capable of request classification and request scheduling. Squid is of a single-process event-driven architecture, which uses the select( ) system call to simultaneously wait for events on all connections being handled. When select( ) delivers one or more events, the main loop of Squid invokes handlers for each ready con-

nection. The performance and scalability of the mQoS gateway is good because it does not need to fork a child process for each request. The server prober and resource monitor program are implemented as the server daemons running on the gateway and server, respectively. When a request enters the gateway, the iptable utility rewrites the destination IP address and port number of this incoming packet to redirect it to Squid. Such a redirection mechanism makes the mQoS gateway works transparently to clients and the server. The Squid gateway performs request classification and scheduling and sends the request to the server. The Squid gateway then receives the response from the server without caching the response and sends it to the client.

The original Squid is a caching proxy used to cache the responses from a Web server. It is deployed between clients and servers to incept requests and responses. When a client issues a request, Squid reads the request, parses the request, and checks whether the response of this request is already in the cache. If yes, Squid fetches the cached data from the cache and sends it to the client. Otherwise, Squid prepares to forward the request and sends the request to the server. When the server returns a response, Squid reads the response, parses the response, and stores or replaces the response data in the cache. Squid then prepares to forward the response and sends the response to the client.

In the mQoS gateway, the request and response processing modules of Squid are modified to be capable of request classification and request scheduling. The cache module of checking in the request direction and the module of cache storing or replacing in the response direction are bypassed. Instead, the request classification is performed before Squid prepares to forward a request. Afterward, the request scheduling is performed before Squid sends a request to the server. When Squid finishes reading and parsing a response, the request scheduler updates the resource availability counters and then prepares forwarding the response to the client.

4.2. Evaluation

The effect of server resource management is discussed theoretically in Section 2. Here the implementations of the nQoS, sQoS, and mQoS scheduling are practically evaluated on server resource utilization, server throughput, and user-perceived latency. The evaluation environment consists of a traffic generator, a gateway, and a Web server. The gateway and server platforms are Pentium III 700 MHz systems with 256 MBytes main memory and 100 Mbps Ethernet network adaptors. Spirent's Avalanche software and SmartBits platform are used as the traffic generator. Avalanche emulates a large number of clients to issue HTTP requests to the server and gathers the statistics. The gateway performs the traditional DRR scheduling to manage the CPU resource of the server for the sQoS scheduling, or the mQoS scheduling algorithm to manage the CPU, disk I/O, and bandwidth resources. In the nQoS scheduling, the gateway only forwards requests and responses between the traffic generator and the server without any processing. The Web server is based on Apache and PHP. There are three kinds of pages in the server, and different pages will lead to different consumptions of the multiple resources when being accessed. The accesses to the pages of CGI scripts are CPU-intensive. The accesses to the pages of photos are disk I/O-intensive. The accesses to the pages of streaming media are bandwidth-intensive. In the evaluation, three service classes are defined in the QoS policy table, and the ratio of their quanta is set to 6:3:1. The workload contains three kinds of resource intensive requests, but the traffic generator issues more CPU-intensive requests than the other types of requests in order to test the capabilities of the mQoS scheduling.

4.2.1. Differentiation on the resource utilization

Different request scheduling schemes result in different utilization of the server resources, shown in Fig. 7. From observing Fig. 7a, in the nQoS scheduling, every class gets a third of every server resource due to the resource contention. Although three resources are well utilized, there is no differentiation on the resource utilization among three classes. From observing Fig. 7b, in the sQoS scheduling, the gateway schedules requests to well utilize the CPU resource of the server and simultaneously to differentiate the resource utilization to the ratio of 6:3:1. However, the gateway stops sending requests to the server when the CPU resource of the server is well utilized, causing the waste of the disk I/O and bandwidth resources of the server. In Fig. 7c, the mQoS scheduler sends appropriate requests to the server to well utilize the three server resources. Furthermore, the differentiation of the resource utilization is evidently observed from that every server resource is utilized by the three classes according to the defined ratio of 6:3:1.

4.2.2. Differentiation on the server throughput

The amount of the utilization of every server resource will affect the server throughput, as presented in Fig. 8. In the nQoS and mQoS scheduling, the maximum total throughput is close to 300 requests per second which is limited by the server capacity. However, in the sQoS scheduling, due to the waste of the disk I/O and bandwidth of the server, the total throughput is only 260 requests per second. The mQoS scheduling improves the total throughput by 21% from the sQoS scheduling. Another finding is that there is no differentiation on the server throughput among the three classes in the nQoS scheduling. However, the sQoS and mQoS scheduling reveal the differentiation on the server throughput because they schedule requests for different classes. The ratio of the server throughput of the three classes is close to 6:3:1.

In Fig. 8, the server throughput of the nQoS scheduling is close to that of the mQoS scheduling. This is because the maximum server throughput is limited by the server capacity. The workload in the nQoS and mQoS scheduling make

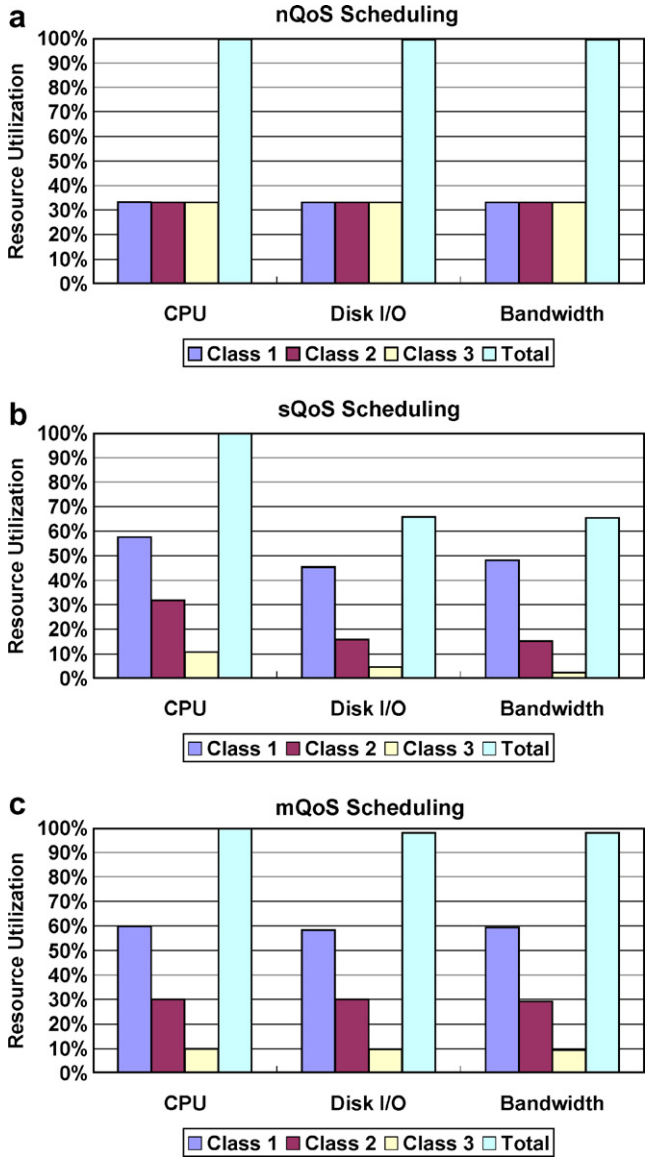

Fig. 7. Server resource utilization of the nQoS, sQoS, and mQoS scheduling. (a) Resource utilization in the nQoS scheduling. (b) Resource utilization in the sQoS scheduling. (c) Resource utilization in the mQoS scheduling.

the server resource well utilized. In nQoS scheduling, the server faces uncontrolled heavy request arrival rate, whereas in the mQoS scheduling, the server faces the scheduled request arrival rate which can well utilize the server. Due to the uncontrolled request arrival rate, the nQoS scheduling has the longer user-perceived latency than the mQoS scheduling.

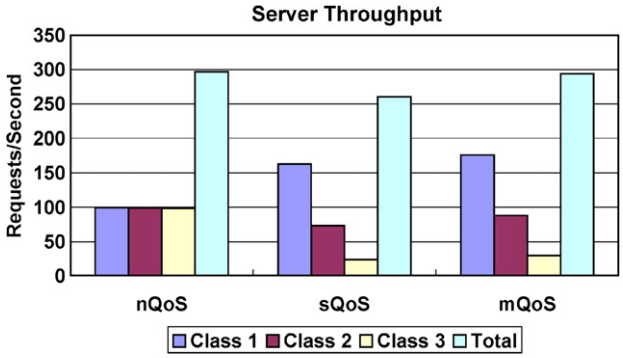

Fig. 8. Server throughputs of the nQoS, sQoS, and mQoS scheduling.

The throughput improvement in the mQoS scheduling results from the fact that the gateway sends appropriate requests to the server to effectively utilize the three server resources. Fig. 9 compares the types of outstanding requests between the sQoS and mQoS scheduling. In the sQoS scheduling, the gateway does not try to balance the utilization of the server resources. However in the mQoS scheduling, the main scheduler takes effect to balance the utilization on every resource. Also the three sub-schedulers differentiate the utilization of every resource among the three classes with a ratio close to 6:3:1.

4.2.3. Differentiation on the user-perceived latency

User-perceived latency is the time between issuing a request and receiving a response back at the client. Fig. 10 shows the user-perceived latency of the nQoS, sQoS, and mQoS scheduling. For the nQoS scheduling, there is no differentiation on the user-perceived latency among the three classes. Because the heavy workload leads to requests queued on the server, the average latency is longer than mQoS scheduling. For the sQoS scheduling, although the user-perceived latency is differentiated, the average latency is longer. For the mQoS scheduling,

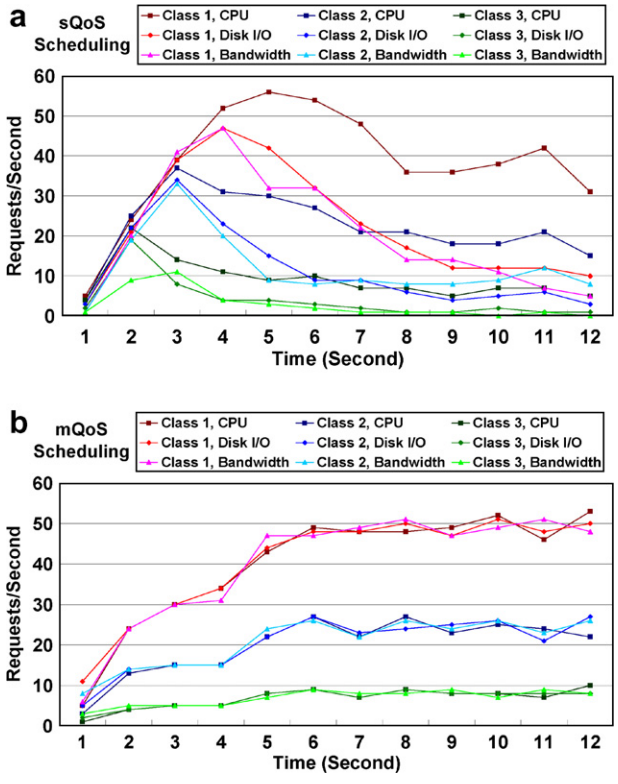

Fig. 9. Types of requests sent to the server by the sQoS and mQoS scheduling. (a) Types of requests sent by the sQoS scheduling. (b) Types of requests sent by the mQoS scheduling.

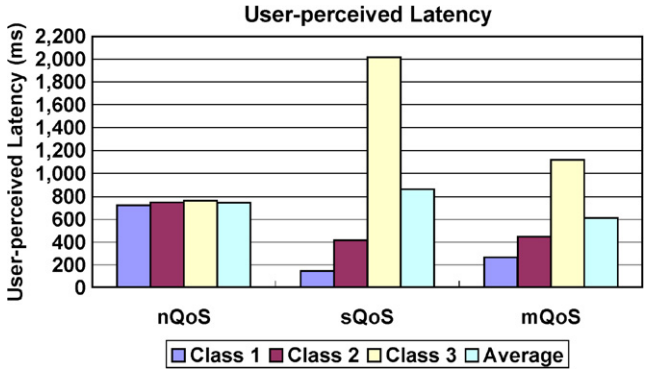

Fig. 10. User-perceived latency of the nQoS, sQoS, and mQoS scheduling.

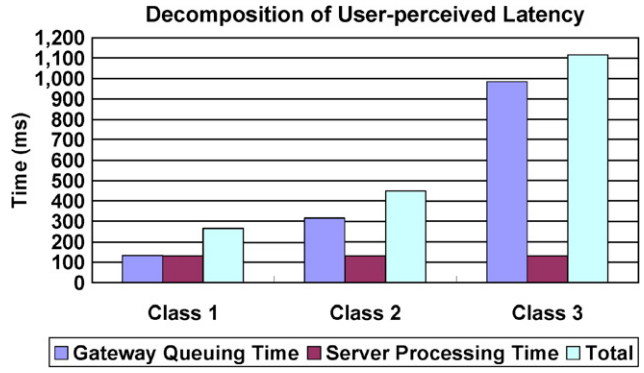

Fig. 11. Decomposition of the user-perceived latency in the mQoS scheduling.

besides the mQoS gateway differentiates the resource utilization, the user-perceived latency is also differentiated but the ratio is not exactly 6:3:1. Furthermore, the average user-perceived latency of the mQoS scheduling is shorter than those of the nQoS and sQoS scheduling.

4.2.4. Decomposition of the user-perceived latency

The user-perceived latency in the mQoS scheduling mainly consists of the gateway queuing time and server processing time. The gateway queuing time is the time between accepting a request from the client and scheduling this request to the server at the gateway. The server processing time is the time between accepting a request from the gateway and sending the response to the client at the server. Fig. 11 shows the decomposition of the user-perceived latency in the mQoS scheduling. The server processing time is almost the same among the three classes, whereas the queuing time of every class is different. Different queuing times lead to the differentiation on the user-perceived latency.

5. Conclusion and future work

Resource management on a Web server allows a website operator to control the utilization of the server resources and provide differentiated quality of service. Traditional single-resource request scheduling cannot manage multiple server resources well, that leads to resource wasting or overloading. This research presents a multiple-resource request scheduling algorithm, mQoS, deployed at the website gateway to provide service differentiation. The mQoS gateway consists of a server prober, a request classifier, and a request scheduler. The server prober profiles the resource consumption of every Web page and the capacity of every server resource. The content-aware request classifier determines the resource tendency and the service class of requests, and puts them into different class queues. The mQoS scheduler consists of several sub-schedulers and a main scheduler. Each sub-scheduler manages a server resource and differentiates the resource utilization among the classes. The main scheduler checks the availability of the server resources and triggers an appropriate sub-scheduler to balance the utilization among the resources. The mQoS scheduling algorithm is work-conservative to the server to keep the server resources well utilized. However, it is non-work-conservative to the class queues because the scheduler remains idle when there are no enough resources for servicing a request.

The mQoS gateway is implemented on Squid and Linux. The mQoS scheduling is compared with no scheduling (nQoS) and single-resource request scheduling (sQoS). The nQoS scheduling owns no differentiation, and the sQoS scheduling owns the differentiation only on the utilization of one server resource. However, the mQoS scheduling holds the differentiation on the utilization of every server resource. Because all server resources are well utilized in the mQoS scheduling, the total server throughput is improved by 21%, compared with the sQoS scheduling. Moreover, the user-perceived latency is also differentiated among the classes in the mQoS scheduling due to the differentiation of the gateway queuing delay. The evaluation reveals that the mQoS scheduling has the capabilities of differentiating the server resource utilization, maximizing the server throughput, and sharing resource.

In the future, we will consider the influences of caching or prefetching techniques on estimating the resources consumption of requests. Besides, we will revise the presented mQoS scheduling algorithm to support a cluster of servers. The more complex multiple-resource, multiple-server request scheduling algorithm can be implemented on a server load balancer. The issues of service differentiation, resource utilization, and server load balancing should be completely considered in the design of the new algorithm.

Acknowledgement

This work was supported in part by the Taiwan National Science Council's Program of Excellence in Research, and in part by grants from Cisco and Intel.

References

[1] J. Almeida, M. Dabu, A. Manikutty, P. Cao, Providing differentiated levels of service in Web content hosting, in: Proceedings of the 1st Workshop Internet Server Performance, June 1998.

[2] L. Eggert, J. Heidemann, Application-level differentiated services for Web servers, World Wide Web Journal 2 (3) (1999) 133–142.
[3] R. Pandey, J.F. Barnes, R. Olsson, Supporting quality of service in HTTP servers, in: Proceedings of the 7th Annul ACM Symposium on Principles of Distributed Computing, June 1998, pp. 247–256.
[4] V. Cardellini, E. Casalicchio, M. Colajanni, M. Mambelli, Web switch support for differentiated services, ACM Performance Evaluation Review 29 (2) (2001) 14–19.
[5] H. Zhu, H. Tang, T. Yang, Demand-driven service differentiation in cluster-based network servers, in: Proceedings of the 20th Conference of the IEEE Communications Society, vol. 2, April 2001, pp. 679–688.
[6] K. Shen, H. Tang, T. Yang, A flexible QoS framework for cluster-based network services, in: Proceedings of the 2002 USENIX Annual Technical Conference, December 2002.
[7] S. Chandra, C.S. Ellis, A. Vahdat, Application-level differentiated multimedia web services using quality aware transcoding, IEEE Journal on Selected Areas in Communications 18 (12) (2000).
[8] C.C. Hung, L.Y. Hong, Adaptive proxy-based content transformation framework for the world-wide Web, in: Proceedings of the 4th International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region, vol.2, May 2000, pp. 747–750.
[9] Y. Lu, T. Abdelzaher, C. Lu, G. Tao, An adaptive control framework for QoS guarantees and its application to differentiated caching services, in: Proceedings of the 10th International Workshop on Quality of Service, May 2002.
[10] W. Leinberger, G. Karypis, V. Kumar, Job scheduling in the presence of multiple resource requirements, in: Proceedings of the 7th International Conference on High Performance Networking and Computing, April 1999.
[11] W. Leinberger, G. Karypis, V. Kumar, Load balancing across near-homogeneous multi-resource servers, in: Proceedings of the 9th Heterogeneous Computing Workshop, May 2000, pp. 60–71.
[12] C. Lee, J. Lehoczky, D. Siewiorek, R. Rajkumar, J. Hansen, A scalable solution to the multi-resource QoS problem, in: Proceedings of the 20th IEEE Real-Time Systems Symposium, December 1999.
[13] M.E. Crovella, R. Frangioso, M. Harchol-Balter, Connection scheduling in Web servers, in: Proceedings of the 1999 USENIX Symposium on Internet Technologies and System, October 1999.
[14] M. Aron, P. Druschel, W. Zwaenepoel, Cluster reserves: a mechanism for resource management in cluster-based network servers, in: Proceedings of the SIGMETRICS Conference on Measurement and Modeling of Computer Systems, June 2000, pp. 90–101.
[15] E. Casalicchio, M. Colajanni, A client-aware dispatching algorithm for Web clusters providing multiple services, in: Proceedings of the 10th International World Wide Web Conference, May 2001, pp. 535–544.
[16] M. Shreedhar, G. Varghese, Efficient fair queuing using deficit round-robin, IEEE/ACM Transaction on Networking 4 (3) (1996) 375–385.
[17] X. Chen, P. Mohapatra, H. Chen, An admission control scheme for predictable server response time for Web accesses, in: Proceedings of the 10th World Wide Web Conference, May 2001, pp. 545–554.
[18] K. Li, S. Jamin, A measure-based admission control Web server, in: Proceedings of the 9th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, March 2002, pp. 651–659.
[19] L. Cherkasova, P. Phaal, Session-based admission control: a mechanism for peak load management of commercial Web sites, IEEE Transactions on Computers 51 (6) (2002) 669–685.
[20] S. Elnikety, J. Treacy, E. Nahum, W. Zwaenepol, A method for transparent admission control and request scheduling in e-commerce Web sites, in: Proceedings of the 13th International World Wide Web Conference, May 2004, pp. 276–286.
[21] N. Bhatti, R. Friedrich, Web server support for tiered services, IEEE Network 13 (5) (1999) 64–71.
[22] V. Kanodia, E.W. Knightyly, Ensuring latency targets in multiclass Web servers, IEEE Transaction on Parallel and Distributed Systems 14 (1) (2003) 84–93.