A Dynamic Priority Allocation Scheme of Messages for a Differentiated Web Services Satisfying Service Level Agreement

Dongjoon Kim, Sangkyu Lee, Sangyong Han, Ajith Abraham

Department of Computer Science & Engineering, Chung-Ang University, Seoul, Korea

djkim@tysystems.com, sklee@archi.cse.cau.ac.kr, hansy@cau.ac.kr, ajith.abraham@ieee.org

Abstract

Recently many enterprises are adopting web services as the standard between heterogeneous software in the XML message based distributed environment to carry out businesses from B2C to B2B. For effective application of web services, differentiated service quality must be guaranteed. However, majority of the current web services do not differentiate quality of messages and the current web servers do not reflect the quality factors of the service level agreement settled between the service provider and user. Our research analyzes the appropriate quality factor for the quality level where differentiated service is provided and suggests a method for assigning priorities to web service message processing processes based on these quality factors. The suggested method assigns the priority dynamically in order to satisfy the service level agreement as much as possible.

1. Introduction

Recently many enterprises carry out B2B business by adopting web services, which have settled as a standard in the XML message based distributed environment. In order to apply web services effectively, the service provider must be able to provide web services differentiated according to the various service levels. Furthermore, a Service Level Agreement (SLA) between the user and provider is necessary for specifying the various service levels [4]. The SLA is purported for defining the responsibility relation between the user and provider and to guarantee the quality of the service provided. Hence, the web service provider must be capable of guaranteeing the web service quality agreed by the SLA [13].

However, current web service technology standardization organizations have not yet established a standard for languages used in describing information on the web services quality evaluation factors and the evaluations themselves. At present only several vendors and universities are individually carrying out specifications and research [5, 6, 7, 8, 11]. Web service quality refers to the level of the functional service, such as performance, reliability, usability and security, and exists in various forms according to the categorizing standard.

Meanwhile, in IETF (Internet Engineering Task Format), DiffServ (Differentiated Service) is suggested for guaranteeing differentiated service quality at the network level [2]. But with DiffServ, the quality of end-to-end transfer via the Internet and the differentiated packet transfer function is implemented only in each separate section. Recently, to improve such disadvantages, some research initiatives on differentiating web services from web servers are underway [1, 3, 10, 12]. Since SOAP parsing time and business logic execution time is required in web services, the waiting time on web servers tends to be longer than the waiting time on the network, As so, the role of differentiated web services on the web server should be the more important. Current researches are purported for finding the way to schedule the processes according to the user's request in order to provide not only network level, but also application level services. At present, most of the web servers apply the FIFO and static priority scheduling methods. But such methods are not capable of dynamically assigning priorities to fit each particular situation and as a result a starvation of low priority processes occurs or the performance evaluation of quality information that had been provided in the past is not reflected.

In order to provide differentiated web services through the web server, our research analyzes various web services quality factors to define the appropriate web services quality factors applied in differentiated web services. Then this quality information is used to assign priorities dynamically to the processes that process messages according to the particular situation in order to suggest a scheduling method that satisfies the service level agreement as much as possible. Web services do not necessarily have to be operated on the web server, but, at present, the majority of web services interact through web servers. As so, the present study suggests a method for allocating priority ranks to request messages in order to provide differentiated web services in a web server environment.

2. Related Work

2.1 Web Services Quality

As web services continue to gain more importance, the service quality is greatly affected by the frequency of service use and reliability between the service provider and user. These factors are the key to the service provider's success in the business. Current web service technology standardization organizations are not putting sufficient effort in coming up with a language to describe information on web services quality evaluation factors or the evaluations themselves. At present only several vendors are individually carrying out specifications [6]. Web service quality refers to the level of the functional service, such as performance, reliability, usability and security, and exists in various forms according to the categorized standard.

Standardization of web services quality can be largely divided into two trends: standardization of performance and stability and standardization of web service and platform management [5]. The standard for performance and stability is not established by a standardization organization but by IBM's WSLA (Web Services Level Agreement) [6], Web service and platform management was first standardized by HP's WSMF and IBM's WS-Manageability, and afterwards these standards were integrated into OASIS's WSDM (Web Services Distributed Management).

2.2 Service Level Agreement

An official agreement between the service provider and user is required to guarantee the defined level of the web service performance based on service quality factors. Such a service level agreement may be very comprehensive and at the same time very specific. The customer expects a certain level performance guaranteed by the service provider in accord to the individual service level agreement settled between each provider and user. The contract may also include the procedures to be followed when the provider has failed to provide the promised service level [4, 5].

IBM developed WSLA (Web Services Level Agreement)[6] for producing and monitoring service level agreements and standards for a web service environment. WSLA is a document of agreement defining the responsibilities of the web service provider and requester when using web services. It was designed considering the nature of the service level agreement environment. WSLA is composed of several elements as illustrated in Figure 1 [6].

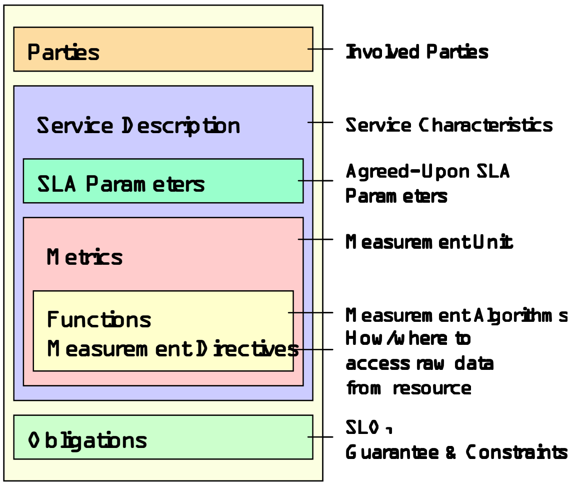

Figure 1. Structure of WSLA

2.3 Differentiated Service from Web Server

Network traffic is increasing tremendously along with the advance and popularization of the Internet. One of the biggest problematic matters in such a situation is the fact that all packets are processed equally at the highest level of performance for packet sending regardless of the service type. As a result, the service quality cannot be guaranteed and service providers also face limits in providing real time information telecommunication services. In order to solve this problem, DiffServ is suggested at IETF [2]. DiffServ classifies the packets to be transferred by the Internet into 8 to 64 service types according to the distinguishing method designated by the user, designates the processing function each packet exchanger must execute for each service type to allow differentiated Internet services.

But DiffServ does not guarantee the quality of end-to-end transfer via the Internet and the packet transferring function is implemented only for each separate section of the communication network. Due to the fact that the web server delay time is longer than the network delay time when the load is concentrated on the web server, many studies on the web server supporting differentiated services are underway [1, 3, 10, 12]. Among them, one study suggests the WebQoS structure. Upon receiving an HTTP request, this structure classifies services based on the classification policy and differentiates the services according to the class they belong to [1]. The structure is depicted in Figure 2.

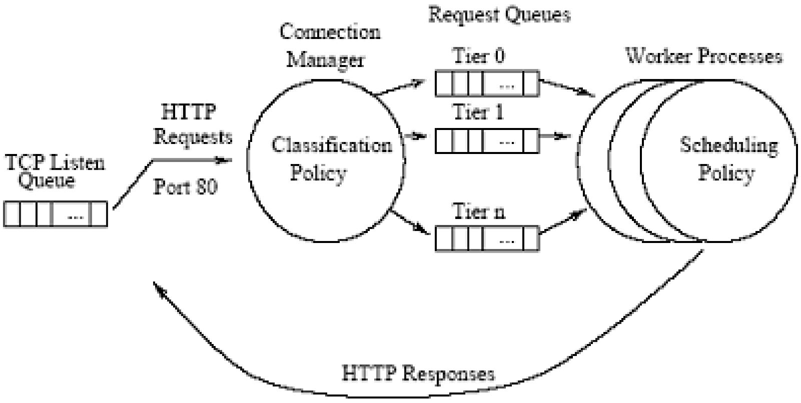

Figure 2. Structure of Web Server Supporting Differentiated Web Services [1]

Different from the current web server processing of HTTP requests using the FIFO method, WebQoS has a connection manager added to discriminate classes by the priorities as shown in Figure 2. The HTTP requests classified and stored in each class are executed according to the scheduling policy to support differentiated service quality. This paper suggests a scheduling policy improved using the WebQoS structure.

3 Method for Assigning Priority using Web Services Quality Information

3.1 Drawing of the Applied Web Services Quality Factors

According to the research report published by the National Computerization Agency in Korea [5], web services quality factors are classified into aspects of performance and stability, middleware, possibility of management and interoperability etc. However, if too many quality factors are applied in determining the priority, calculations for determining the priority must be executed simultaneously with real time measurement of the factors and in result weakens accuracy and performance. Due to this reason, the present study adopts the quality factor of performance and stability, which is easy to monitor and most widely applied in service level agreements. The factor also reflects the performance level expected by the user. The formula for calculating the quality factor concerning performance and stability is defined below.

○ **Quality Factors Concerning Performance**

Service providers can provide good performing services to users by reflecting the quality factor concerning performance when assigning priority.

- Response Time = Total time taken for message processing / Number of messages processed
- Throughput = Total number of messages processed / Total processing time

○ **Quality Factors Concerning Stability**

The quality factor concerning stability determines the usability of web services. Web services should be differentiated in this aspect as well. Even when web services are available the user may consider them unusable if the performance is poor [9].

Furthermore, when a large number of service requests occur at the same moment of timeout, stability decreases. By considering this quality factor in allocating priority, service providers can provide differentiated services to their customers.

- Accessibility = Message processing rate for certain message
- Reliability = Number of messages processed / Number of messages requested

The processing rate in the accessibility formula is the same as the formula used for reliability (number of messages processed/number of messages requested). However, even when the server’s throughput exceeds the maximum throughput, the accessibility must still be considered in order to guarantee the processing rate for advanced users.

In order to guarantee the web services quality factors defined above, a service level agreement must be settled between the user and provider. Generally, these agreements are official contracts signed by the user and provider in order to guarantee measurable performance of applications, services and networks at a defined level. But since no standard exists for service level agreements, IBM’s WSLA 1.0 [6] is applied in this research. Such a service level agreement may be very comprehensive and at the same time very specific and includes the procedures to be followed when the provider has failed to provide the promised service level. However, it is regarded in this study that the agreement is made only considering the quality factors defined above. The present study’s main purpose is to realize quality factors, which are satisfactory to the maximum.

The <SLAParameter> element of WSLA 1.0 was used to express the services quality factors. The services quality factors defined above are expressed with WSLA 1.0’s <SLAParameter> as shown in Figure 3 below.

```
......
<SLAParameter name="AverageResponseTime"
                    type="float"
                    unit="seconds">
   <Metric>AverageResponseTime</Metric>
</SLAParameter>
<SLAParameter name="Throughput"
       type="integer"
       unit="transactions/hours">
   <Metric>Throughput</Metric>
</SLAParameter>
......
```

Figure 3. Example of WSLA <SLAParameter>

Based on the services quality factors defined in Figure 3, the user and provider must clearly define what is to be guaranteed. In the case of WSLA 1.0, two types of guarantees are provided using Service Level Objective and Action Guarantee. As observed in Figure 3, the Service Level Objective indicates the promise related to the SLA Parameter, in other words, the provider’s promise to maintain a certain state of service for a given period of time. Both parties, the provider and the user, can be responsible for this promise, but usually the provider is responsible for keeping the promise. The Action Guarantees are promises to execute a certain action and include notification of Service Level Objective violation and management operation calls. The study only applied the Service Level Objective in indicating the certain level the quality factors must satisfy. The indicated values were used for categorizing the messages and allocating priority. Figure 4 below shows an example of the Service Level Objective defined using WSLA 1.0.

```
<Obligations>
   <ServiceLevelObjective name="Ex_SLO">
      <Obliged>ServiceProvider</Obliged>
      <Validity>
<Start>2004-10-10T14:00:00.000-05:00</Start>
<End>2004-12-31T14:00:00.000-05:00</End>
      </Validity>
      <Expression>
         <Implies>
            <Expression>
               <Predicate xsi:type="Less">
<SLAParameter>AverageResponseTime</SLAParameter>
                  <Value>1000.0</Value>
               </Predicate>
            </Expression>
......
</Obligations>
```

Figure 4. Example of WSLA <ServiceLevelObjective>

3.2 Method for Assigning Priorities

The web server model supporting differentiated web services is illustrated in Figure 5. For this model, the service level agreement must be settled between the web service user and provider, and this agreement must contain the web service quality information measurements mentioned in Section 3.1. The web service message requested by the web service user is classified according to the classification policy and the execution queue is allowed to process messages up to the maximum number guaranteeing best processing performance. When requests exceed the limit, the remaining requests are put on standby at each buffer queue. The messages to be sent from the buffer queue to the execution queue are determined according to the priority allocation policy.

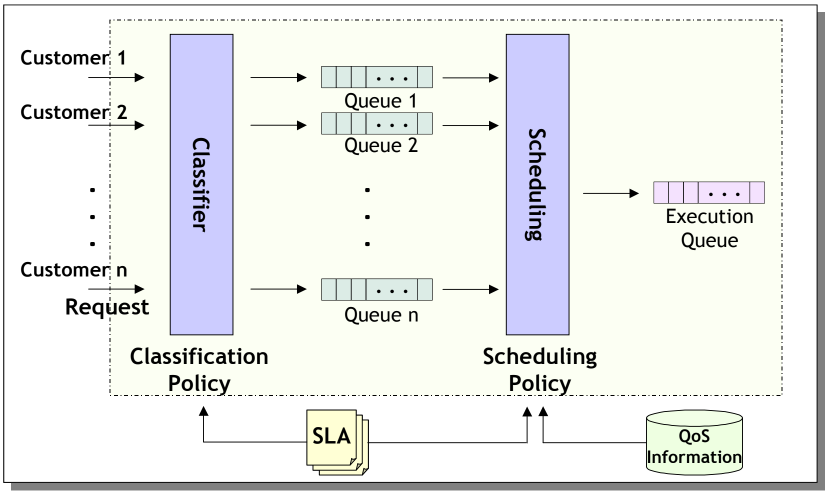

Figure 5. Model for Differentiated Web Server

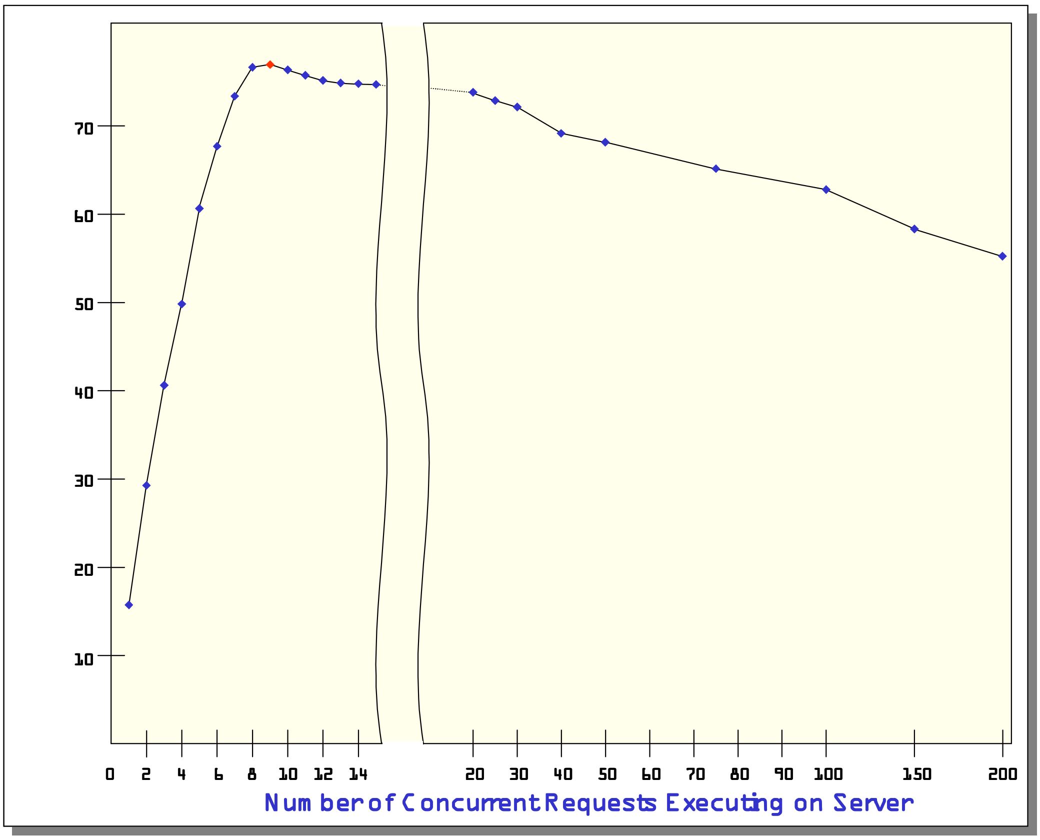

Figure 6. Measurement of Server's Throughput

As with this model, the optimum number of messages to be processed by the web server should be controllable. When messages requested to the web server increases, the message processing quantity reaches a saturation point and the response time is extended a tremendous amount. This is because performance is weakened due to collision between resources and conversion of contexts when passing the saturation point [4]. As so, providers should identify the saturation point before beginning web services or signing a service level agreement, and use this information in setting the appropriate number of messages for the execution queue.

Figure 6 shows the measurement of the server's throughput in the test environment. In this test, the best performance was observed when the execution queue length, the number of message requests executed simultaneously on the server, was set as 9. The execution queue length for the test in Chapter 4 was also set as 9.

Basically, two modules are required for constructing a differentiated web services supporting web server. The module for classifying each message and the module for assigning priorities to the classified messages. The message classification module classifies messages according to the service level agreement and discriminates the messages by the user IP address or the URL form. Meanwhile, the priority assignment module can be applied by using the accumulated web service quality information.

When the execution queue in the priority assigning module is saturated, messages standing by at each classified queue are assigned priorities and must be sent to the execution queue starting with the message with the highest priority. For this study, the priority here was determined by comparing the monitored value of the web services quality factor extracted in Section 3.1 and the quality factor value settled in the service level agreement in order to satisfy the service level agreement at a high level. Basically, the FIFO method is applied in the classified queues so that the first message that is received can be provided the first within the queue. When there are messages standing by in the classified queues, objects for comparison with the first arriving messages of each queue are selected and quality information values are calculated for these messages. The message with the smallest value is assigned the highest priority.

The response time is a factor the user can directly feel during quality factor evaluation, and it is often one of the most important factors taken into account in service level agreements. In order to apply this response time, the value can be calculated using (1).

$$V_{RT} = RT_{SLA} - QT - RT_{mean} \qquad (1)$$

In (1), RT_{SLA} is the value of the average response time agreed on by the user and provider in the service level agreement, QT is the time the message stood by in the queue and RT_{mean} is the measured average response time. Basically, the message with the lowest V_{RT} value, in other words, having the least time left till the time limit, is assigned with the highest priority. In this aspect, this method is similar to the EDF method.

The EDF (Earliest Deadline First) method allocates priority to messages with closest deadlines. Among the by standing messages, the one with the closest deadline is selected and sent to the execution queue. The priority of each class changes as time passes. This method's strong points are the fact that the deadline can be estimated and the respond time is satisfied to a considerable degree compared to other methods.

This paper suggests other quality factors as well as the response time. The satisfaction rate of each quality factor is calculated in order to improve the overall satisfaction rate including the response time satisfaction rate and the satisfaction rates of other factors. The satisfaction rate (V_T) for amount of message processed is calculated by using (2).

$$V_T = T / T_{SLA} \qquad (2)$$

In (2), T_{SLA} is the throughput agreed on in the service level agreement and T indicates the measured throughput. Furthermore, the accessibility and reliability satisfaction rates (V_A and V_R respectively) can be calculated as.

$$V_A = A / A_{SLA} \qquad (3)$$

$$V_R = R / R_{SLA} \qquad (4)$$

In (3) and (4), A_{SLA} and R_{SLA} are the accessibility and reliability values agreed between the user and provider in the service level agreement and A and R indicate the accumulated accessibility and reliability respectively.

The priorities are assigned using the values calculated by formulas (1) to (4). The values to be applied are calculated as below.

$$V_P = V_{RT} \times [(V_T + V_R)/2] \qquad (5)$$

Equation (5) is used for the first messages arriving and standing by in each standby queue to calculate and compare the

V_P value and thereby assign the highest priority to the message with the smallest value. This is based on the response time. Compared to the throughput and reliability values with the service level agreement values, the service level agreement satisfaction rate rises as time passes.

However, it is not easy to apply (5) directly. When the V_P values are negatives, the priority can be reversed if the priority is settled based on the minimum value. As so, the V_P values must be corrected as shown below. Here, the $TimeOut$ value is used for making the V_{RT} values into positives. The sum of V_{RT} and $TimeOut$ adopts the value of $\ln$ to be compared with other V_T and V_R values.

$$V_P = \ln(V_{RT} + TimeOut) \times [(V_T + V_R)/2] \quad (6)$$

Using (6), the service level agreements of all requesters can be satisfied within the server's processing range. But if the throughput exceeds the fixed limit, in other words, if the number of requests is larger than the server is capable of processing, someone must give up the service. Here, the service requesters who have contracted the accessibility value in the service level agreement are regarded as advanced users and the number of requests per second when the server performs the best is set as the threshold value. Then, (7) is applied to the messages of users who have contracted the accessibility value and (8) to messages of ordinary users. When the number of requests exceeds the threshold value, the reliability value is applied to guarantee the advanced user the maximum performance and the value of throughput is applied to guarantee the ordinary user the minimum performance.

$$V_P = \ln(V_{RT} + TimeOut) \times V_A \quad (7)$$
$$V_P = \ln(V_{RT} + TimeOut) \times V_T \quad (8)$$

In this study, (6), (7) and (8) are suggested for assigning priorities. But if factors other than the response time are omitted, other equations excluding the particular factors can be applied instead.

4. Experiment and Evaluation

In this Section, we present some experiment results to demonstrate the suggested priority assigning method for differentiated web services and the performance is compared with the EDF method.

4.1 Experiment Scenario

The network situation is excluded and an application level simulation is executed in the LAN environment. The processing time for each single web service is set at approximately 0.06 seconds, and the reliability, response time and throughput were measured.

The web service messages were classified into three types. Part of WSLA used in the experiment is depicted in Figure 7. As observed in Figure 7, For messages placed in Class 1, it is supposed that the service level agreement defines the response time as two seconds or less, throughput as 30 or more and reliability as 95% or higher. In this experiment, the service quality is guaranteed in the Class 1 > Class 2 > Class 3 order. For Class 2, the response time achieved is 2.5 seconds, throughput of 25 or more and a reliability of 90% or higher. For messages sorted into Class 3, it is supposed that the service level agreement defines the response time as 3.5 seconds, throughput as 20 and reliability as 80%. The same type of WSLA suggested in Figure 7 can be created.

4.2 Results and Analysis

As shown in Figures 8, 9 and 10, the maximum performance value capable of satisfying all of the service level agreements in the test environment is when the number of request per second is 120. Figure 8 shows the response time deviation between the EDF method and the suggested method. Both methods satisfy the service level agreement when the number of request per second is 120. As evident, with reference to reliability and throughput, the EDF method does not satisfy the service level agreement sufficiently when the number of requests per second is 120. This means that the number of requests per second must stay under 120 to have the EDF method satisfy all of the service level agreements. On the other hand, the suggested method is capable of satisfying all of the service level agreements even when the number of requests per second is 120.

```
. . . . .
<Obligations>
. . . . .
  <Expression>
    <Predicate xsi:type="Less">
      <SLAParameter>
        AverageResponseTime</SLAParameter>
      <Value>2000</Value>        <!-- 2sec -->
    </Predicate>
  </Expression>
  <Expression>
    <Predicate xsi:type="Greater">
      <SLAParameter>
                   Throughput</SLAParameter>
      <Value>30</Value>
    </Predicate>
  </Expression>
  <Expression>
    <Predicate xsi:type="Less">
      <SLAParameter>Availability_CurrentDown
                        Time</SLAParameter>
      <Value>0.01</Value>
    </Predicate>
  </Expression>
  <Expression>
    <Predicate xsi:type="Greater">
      <SLAParameter>
      Accessibility_Transaction</SLAParameter>
      <Value>0.95</Value>         <!-- 95% -->
    </Predicate>
  </Expression>
  <Expression>
    <Predicate xsi:type="Greater">
      <SLAParameter>Reliability_Transaction
                          Rate</SLAParameter>
      <Value>0.95</Value>        <!-- 95% -->
    </Predicate>
  </Expression>
 <EvaluationEvent>NewValue</EvaluationEvent>
  </ServiceLevelObjective>
</Obligations>
. . . . .
```

Figure 7. WSLA of Class 1

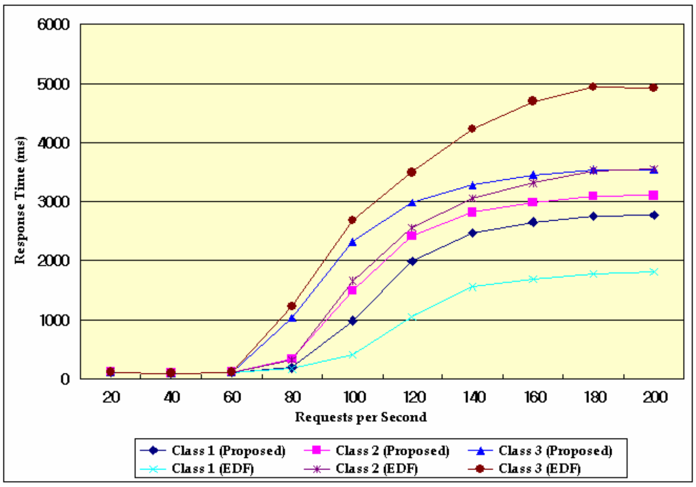

Figure 8. Response time of EDF method and proposed method

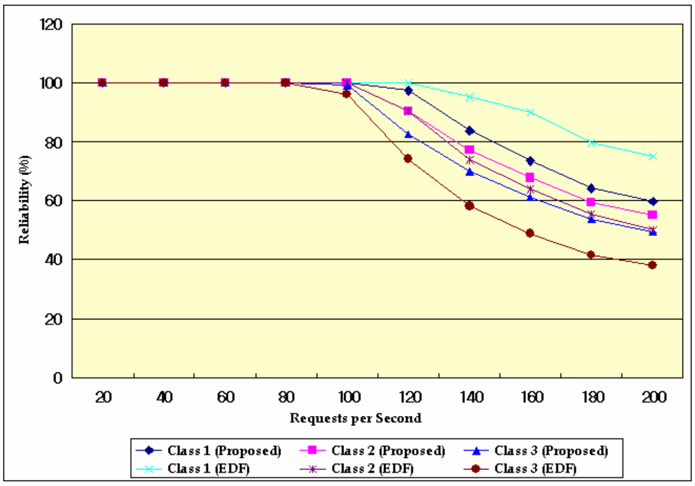

Figure 9. Reliability of EDF method and proposed method

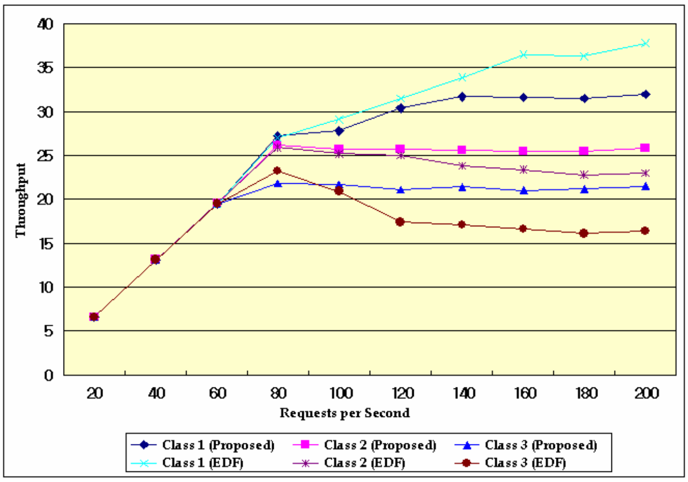

Figure 10. Throughput of EDF method and proposed method

The number of requests per second exceeding 120 can be regarded as exceeding the server's maximum throughput. Hence, all of the service level agreements cannot be satisfied and the requests from ordinary users can only be ignored in order to answer the advanced users' requests. Figure 8 suggests the reliability depending on whether accessibility is considered or not. As observed here, the advanced user can be guaranteed the best performance by using accessibility.

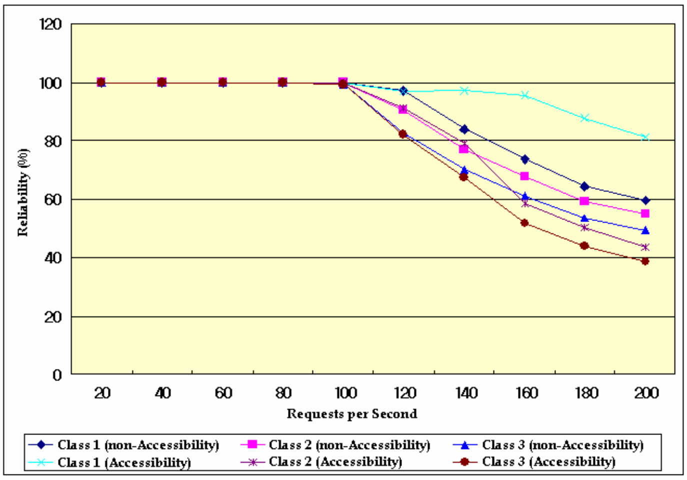

Figure 11. Reliability considering accessibility

5. Conclusions and Future Work

Recently many enterprises are adopting web services, which has settled as the standard between heterogeneous software in the XML message based distributed environment, to carry out businesses from B2C to B2B. For successive business with web services, providers should be capable of offering differentiated services.

However, differentiated web services on the network level do not guarantee end-to-end transfer quality whereas the network environment is improving. Other than that, since web services send messages using XML, XML parsing is inevitable and the majority of current web services use web servers for communication.

For such reasons, this paper suggests a priority assignment method for providing differentiated web services on the web server, not at the network level but at the application level. For implementing this method, the web services quality factors most used in web services were analyzed to extract the factors required for assigning priorities. Using these quality factors and IBM's WSLA, the method for assigning priorities to messages in the web server is produced.

Since this study did not take network speed into account, a method that is compatible with differentiated services at the network level such as DiffServ should be implemented.

6. References

[1] Bhatti, N., Friedrich, R. (1999). Web Server Support for Tiered Services. Network, IEEE. Vol. 13. Issue 5.

[2] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., Weiss, W. (1998). An architecture for differentiated services. IETF RFC 2475.

[3] Chen, X., Mohapatra, P. (2002). Performance Evaluation of Service Differentiating Internet Servers. Computers, IEEE Transactions on. Vol. 51. Issue 11.

[4] Dan, A., Ludwig, H., Pacifici, G. (2003). Web services differentiation with service level agreements. White Paper, IBM.

[5] Heo, J.H. (2003). A Study on Technical Trends and Deployment Strategies of Web Service Quality Management. Research Report, National Computerization Agency, Korea.

[6] Keller, A., Ludwig, H. (2003). The WSLA Framework: Specifying and Monitoring of Service Level Agreements for Web Services. IBM.

[7] Mani, A., Naqarajan, A. (2002). Understanding quality of service for Web services. IBM.

[8] Menasce, D.A. (2002). QoS Issues in Web Services. Internet Computing, IEEE. Vol. 6. Issue 6.

[9] Park, S.K. (2003). Web Service Availability Measurement Model based on Service Degradation Factor. Ph.D Thesis, Chonnam National University, Korea.

[10] Vasiliou, N., Lutfiyya, H. (2000). Providing a Differentiated Quality of Service in a World Wide Web Server. Proceeding of the Performance and Architecture of Web Servers Workshop, Santa Clara, California, USA.

[11] Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., Sheng, Q.Z. (2003). Quality Driven Web Services Composition. Proceedings of the 12th international conference on World Wide Web, ACM.

[12] 12. Zhou, X., Cai, Y., Godavari, K.G., Chow, C.E. (2004). An Adaptive Process Allocation Strategy for Proportional Responsiveness Differentiation on Web Servers. Proceedings of IEEE International Conference on Web Services.

[13] Dongjoon Kim, Sangkyu Lee, Sangyong Han and Ajith Abraham, Improving Web Services Performance Using Priority Allocation Method, IEEE International Conference on Next Generation Web Services Practices, Seoul, Korea, IEEE Computer Society Press, ISBN 0-7695-2452-4, pp. 201-206, 2005.