Pricing Web Services for Optimizing Resource Allocation – An Implementation Scheme

Zhangxi Lin
The Rawls College of Business Administration
Texas Tech University
zlin@ba.ttu.edu

Huimin Zhao
School of Business Administration
University of Wisconsin – Milwaukee
hzhao@uwm.edu

Sathya Ramanathan
Institute of Research Information Management
Texas Tech University
sathya.ramanathan@ttu.edu

Abstract

Web services technology is becoming an important technological trend in web application development and integration. Based upon open standards such as SOAP, WSDL, UDDI, and BPEL4WS, web services allow web-based applications to communicate through standardized XML messaging and to form loosely-coupled distributed systems. While web services technology provides a new computing model, in which infrastructures and application systems are hosted by service providers and made available to service consumers via web services, computing resources such as network bandwidth, storage throughput, and CUP time need to be allocated such that the total benefits of both the service providers and the service consumers are optimized subject to the QoS requirements of service requests. This paper proposes an economics-based approach to resource allocation and pricing for web services and describes an implementation scheme.

Keywords: web services; resource allocation; resource pricing; incentive compatibility; QoS

1. Introduction

Firmly supported and actively advocated by influential computer companies such as IBM, Microsoft, and Sun, web services technology is emerging as an important trend in web application development and integration [14, 18, 21]. Based upon open standards such as SOAP, WSDL, UDDI, and BPEL4WS, web services allow web-based applications to communicate through standardized XML messaging and to form loosely-coupled distributed systems [8]. As important building blocks of electronic commerce and enterprise application integration architectures as well as on-demand computing (e.g., gird, wireless, and autonomic computing), web services are expected to greatly enhance the promise of distributed computing [14]. It is predicted that by 2007 web services demand will grow to $21 billion, and in the next decade a cumulative $184 billion will be spent on web services projects in the US [11].

Web services are autonomous programs that are available over the Internet and can interact with each other via a standardized XML messaging system. They are self-describing, meaning that the public interface of a web service, consisting of available public methods and their parameters and returns, should accompany the service. They are also discoverable, meaning that their interfaces and locations should be published in public or private registry systems for easy discovery by potential consumers. They can communicate with each other to form distributed systems. Contrast to the traditional EDI technology for application integration, which tightly couples the components via dedicated networks and proprietary messaging protocols, web services can be loosely coupled using open, text-based standards over the Internet, thus significantly reducing the costs of application integration.

Web services technology provides a new computing model, which greatly eases application integration within and across enterprises and is getting increasing adoption, especially in small to medium sized enterprises. These enterprises do not need to maintain expensive infrastructures and application systems necessary to conduct their businesses. They can adopt hosted solutions instead, where the infrastructures and application systems are maintained at service providers and are made simply available to them via web services. Such hosted solutions may significantly reduce the initial investment in system development and deployment as well as regular maintenance cost of many enterprises.

As hosted solutions are increasingly adopted, computing resources such as network bandwidth, storage throughput, and CPU time need to be allocated such that the total benefits of both the service providers and the service consumers are optimized subject to the *Quality-of-Service* (QoS) requirements of service requests. While purely technology-based approaches to resource allocation have serious limitations, an economics-based approach to resource management in an intranet environment has been proposed [10]. Because hosted solutions relying on web services technology closely resemble intranet environments, we adapt the approach proposed in [10] to the context of resource allocation for web services. Issues and technical features specific to web services are taken into consideration in the development of the new approach.

The paper is organized as follows. Section 2 briefly reviews the literature on QoS and network resource allocation, with a focus on traffic pricing. Section 3 proposes an economics-based approach to resource allocation and pricing for web services. Section 4 discusses technical issues in an implementation plan. Section 5 concludes the paper.

2. Literature Review

QoS refers to the ability of a network computing system to provide improved services to satisfy different kinds of requests. QoS has long been an active research area ever since computers and computer networks were brought into commercial use. By the mid 1980s, processor sharing had been a key topic of QoS research particularly because of the limited CPU speed (e.g., [2, 20]). With the advent of the Internet era and the constantly improving CPU capacities, QoS of computer networks became a more urgent research issue. In fact, the Internet has successfully survived from the serious problem of collapses that happened earlier due to congestion. This success is largely attributed to a group of flow-control algorithms such as *slow-start* and *congestion avoidance* proposed in the late 1980s and widely implemented today [1, 12, 20]. With the booming of the Internet in the last decade, research in QoS of bandwidth services has proliferated; many new ideas for improving QoS of bandwidth services have been proposed and quickly implemented. Some examples are *Random Early Detection (RED)* for router's queue management [6], *fair queuing (FQ)* [5] for bandwidth scheduling, *Resource Reservation Protocol (RSVP)* for real-time data flow routing [4], and *Differentiated Services* (DiffServ) [3].

Recently, grid computing is emerging as a new direction of Internet computing with the combined power of individual computers and the Internet [15]. In grip computing, QoS requirement is a multi-facet issue, which involves CPU, bandwidth, and data resource services [6]. As an enabling technology for grid computing but specialized in the prevailing web paradigm, web services technology is subject to a similar QoS problem: how to optimize the allocation of limited network computing resources to cope with ever-increasing demands.

While computer scientists and engineers have achieved progresses in improving the QoS over the Internet, the problem of traffic congestion stays largely unresolved. Internet performance has not been significantly improved in the last several years. In 2000, Keynote System's *Keynote Business 40 Internet Performance Index* showed that the best response time for a web site access was about 1.5 seconds and the worst average was about 15 seconds [17]. In the week of August 11-15, 2003, the best response time for a web site access was about 2.08 seconds and the worst average was about 14.73 seconds [13], showing no improvement after three years.

Since early 1990s, QoS research has started to investigate economics-based network resource allocation mechanisms that support usage-based pricing underpinned by the principle of incentive compatibility [9, 17]. The main idea is to use market mechanisms to suppress low-value data traffic and reduce traffic, which will result in a better network traffic condition for those data flows with high values. A widely accepted method is dynamic pricing, e.g., the GSW model, which is a general equilibrium model with a resource-price structure that is incentive compatible for network resource allocation. An optimal traffic pricing formula in a quadratic form has been derived in the GSW model for the first-in-first-out (FIFO) M/G/1 queueing system and has been well tested under various scenarios by a series of experiments. Lin et al. [17] extended the GSW model from FIFO queueing to round-robin (RR) queueing and conducted extensive experiments based on a network testbed. In addition, they presented an analytical expression for traffic pricing complying with the pipeline-effect of packet-switching networks, which was critical for the Internet and was not previously investigated in the GSW model.

While researchers have theoretically demonstrated that the economics-based approach to network resource allocation has the potential of solving Internet congestion problems, attempts in implementing a working traffic pricing system started just a few years ago. Gupta, Stahl and Whinston [10] conceived a detailed architecture for computing resource pricing in intranets, named *intranet Resource Management Unit* (iRMU). Li et al [15] tested network traffic pricing using human subjects and obtained similar results as were obtained from simulations. However, it still requires more effort and an appropriate technical environment to convert the theory into a real system. The emergence of web services has provided an ideal platform for implementing and validating network resource allocation and pricing mechanisms.

3. Implementing Resource Pricing for Web Services

We have developed an approach to resource pricing for web services, in light of some ideas of iRMU. We adapt the ideas on resource pricing proposed in previous research to the web services context and focus on implementation of a resource pricing system for web services. The approach is described in the following sub-sections.

3.1 *The Economic Objective of a Pricing System*

An important issue for a web services pricing system is the economic objective of the system. Two kinds of pricing models have been typically considered, that is, profit maximization and welfare maximization. If a profit maximization model is adopted, an organization charges its clients prices for its services with the sole objective of maximizing its own benefits without considering clients' utilities. Alternatively, when an organization adopts a welfare maximization model, it sets prices aiming at the best total of its own benefits and clients' utilities. In the web services context, the latter is preferred, as it is better suited to the following, arguably major, scenarios:

1. Web services are provided within an intranet operated by an organization. In this scenario, the organization should seek service welfare maximization. This is also the case of iRMU [10].
2. Web services are hosted by a service provider that is independent of service consumers. Naturally, the provider is to maximize the benefit from the service provision, in which the investment on the web services is counted as a portion of business costs. As the main benefits of web services provision is not from the resource utilization by consumers but from the e-business processes supported by the web services, the provider takes into account the QoS of the web services in order to operate the business processes more smoothly. In this scenario, the provider and a consumer form a virtual partnership or a virtual organization, and the provider should maximize the benefits from the operation of the business.

Therefore, the objective of web services pricing is to optimize the overall benefit of the business operation based on web services by charging the usage of prioritized computing resources in accordance with dynamically updated prices.

3.2 *The Architecture of a Resource Pricing System for Web Services*

The operation of iRMU is essentially an internal economy, with the departments of the organization running an intranet as the agents involved in the e-business. However, iRMU is still at a conceptual stage and needs more details in the system architecture. In addition, to implement such a system in a web services environment, the specific characteristics of web services technology have to be considered. We build upon the conceptual components of iRMU and propose an architecture for a web services pricing system, with more specific considerations in accordance with the web services technology. We adopt a layered model that is typically used in network modeling as well as in web services modeling. Using a layered model allows us to decompose a complicated system into modules that can be dealt with independently. The modules of the pricing system are related to the layers in the existing web services stack.

Layered model of web services pricing system

Figure 1 shows the modules of the web services pricing system and their relationships with the layers of the web services stack proposed by IBM. The modules include:

?? Service Access Utilities (BPEL4WS/UDDI) - assists clients in optimizing service priority selections and handling their service payments.

?? Service Admission (BPEL4WS/UDDI) - receives and grants web services requests.

?? Service Billing (BPEL4WS/UDDI) – bills web services accesses.

?? Resource Pricing (UDDI/WSDL) – periodically updates resource usage prices.

?? Traffic Measuring (SOAP/Network) – measures traffic and load status of web services

?? Resource Scheduling (SOAP/Network) – assigns a proper priority to each service request and collects service usage information for billing.

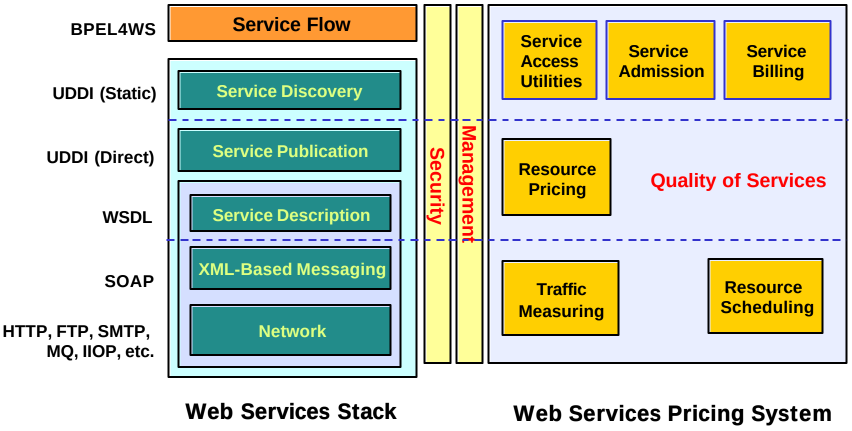

Figure 1: A Web Services Pricing System with Reference to the Web Services Stack

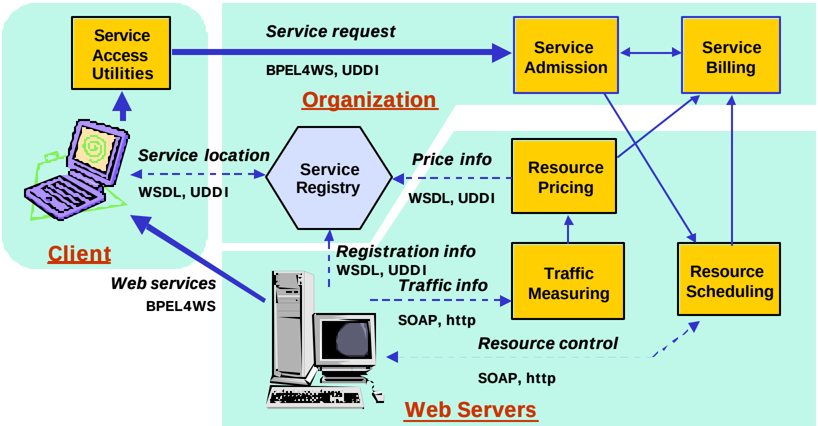

Figure 2: Interaction among the Modules of a Web Services Pricing System

The workflow of the pricing system is shown in Figure 2 and briefly described as follows.

?? Web services, along with pricing information, are registered in UDDI registries.

?? The Traffic Measuring module measures the estimated traffic load (bandwidth, CPU, database access, etc.).

?? The Resource Pricing module periodically calculates the prices for different services and updates the price information in the UDDI service registry.

?? Potential clients locate the web services and obtain the prices from the service registries.

?? A client sends a request for web services with QoS requirements via the Service Access Utilities module.

?? The Service Admission module receives the service request and contacts the Service Billing module to charge the client.

?? After the Service Billing module has verified user's payment account, the Service Admission module issues the Resource Scheduling module to assign the requested services for the client.

?? The Service Billing module charges the service fee to the user's account according to the price received from the Resource Pricing module and the actual usage of resources reported by the Resource Scheduling module.

4. Implementation Considerations

We are in the process of designing a prototype of the pricing system to validate the utility of the proposed approach. Some critical considerations in the implementation plan are described in the following sub-sections.

4.1 *Implementing the System Modules as Web Services*

The modules in the pricing system, including Traffic Measuring, Resource Pricing, Resource Scheduling, Service Admission, and Service Billing, are all to be developed as web services. The interaction between the modules is to be implemented using standard XML messaging (i.e., SOAP). As this is the first effort in implementing and validating an approach to resource pricing for web services, it is expected that many components of the approach and the corresponding modules in the prototype will experience extensive modification. Web services can significantly reduce the costs of developing and integrating the modules. They can be developed independently and easily assembled. Modifying one module incurs little disruption on others.

4.2 *Resource Pricing for the Services Involving both Web Server and Database Server*

In a two-tier web service architecture with a shared CPU between the web server and database server, the performance of the web server and database server is directly correlated to the efficiency of CPU utilization. The database load is determined based on the average time per query or the current number of active process and worker threads. The throughput of the web server could be measured by issuing a certain number of requests and counting the number of replies within a particular time. In a three-tier web service architecture, a database server runs in another CPU behind the web server. Two CPUs will be priced separately. However, we can adopt a single pricing service that is associated to the web server to price both web service and database service. In this way, the database services are treated as extensions of web server's services, and the communications between the web server and affiliated database server are necessary.

4.3 *Resource Utilization and Service Request Status Measuring*

In network traffic pricing, bandwidth demand can be well measured according to the status of the buffer for the queueing service. Clients are also able to estimate the throughput time in accordance with the data flow size and the available bandwidth. In web services resource pricing, the situation is more complicated. CPU and the throughput of storage devices are to be priced in addition to the bandwidth usage. These are less measurable. For example, when a client requests a CPU service for a job, the consumption of CPU time is not precisely predictable, unless the same job was done before. A potential solution is as follows:

1. At the service side, the server estimates the status of the load based on the jobs waiting for services or being served in a round-robin fashion. Service prices are determined based on this estimation.
2. At the client side, it is the client's responsibility to estimate the service amount for each job and make submission decisions.
3. Clients are charged according to the actual resource utilizations posterior to the services.

Based on the above protocol, previous bandwidth pricing schemes are applicable to other resource types that need to be considered in a web services resource pricing system.

4.4 *Publication of Pricing Information*

Pricing information is to be published in UDDI registries, along with information about the service provider and links to the WSDL documents about the provided services. When potential service consumers search for the web services in the registries, the pricing information is also retrieved. The Resource Pricing module periodically updates the price information in the registries to keep it up-to-date. An appropriate updating frequency needs to be determined via experiments.

4.5 *Resource Pricing System Deployment*

From an implementation perspective, we need to consider where these subsystems are deployed, how these subsystems interact with existing web services applications, and how these subsystems are coordinated network-wide. We can identify three subsystems (summarized in Table 1) for the web services pric ing system, that is, *Resource Scheduling and Pricing System*, *Billing and Admission System*, and *End User Assistant System*, which are used by three types of agents, that is, organizations, end users, and web servers. The Resource Scheduling and Pricing System is not directly accessible to end users and is deployed at lower layer for each web server. Normally, an organization only needs a single Billing and Admission System that controls and supervises the Resource Scheduling and Pricing System. The End User Assistant System currently contains one module, which consists of a set of utilities to facilitate the interaction between service requests of end users and the web service pricing system. Other possible utilities, such as user account management, are not relevant to web services access and not mentioned here. Since a user may access web services provided by different organizations, the protocol for the interactions between the End User Assistant System and the Billing and Admission System should be standardized.

Table 1: Subsystems and Their Configurations

Subsystem	Deployment	Modules	Notes
Resource Scheduling and Pricing System	Deployed for each web server	Traffic Measuring	
		Resource Pricing	
		Resource Scheduling	
Billing and Admission System	Deployed for each organization running web services	Service Admission	
		Service Billing	Accessing and interacting with user's online payment accounts
End User Assistant System	Downloadable from any service providers in this web services pricing enabled system; Installed on each client computer	Service Access Utilities	Users may set up their payment accounts by the services of other third parties.

5. Conclusion and Further Work

We have described an economics-based approach to resource pricing for web services in this paper. Such an approach can potentially improve the utilization of resources and increase the benefits of service providers and consumers. We have outlined the architecture of a pricing system based on the proposed approach and discussed some technical issues that need to be considered in an implementation. This research is currently still in progress. We are implementing the proposed approach on the Microsoft .NET platform and validating its utility via simulation experiments. Upon successful implementation and validation of the approach, we plan to extend it to other settings, such as network traffic pricing that has been theoretically studied for over a decade and has never been actually implemented. Such an empirical research will shed some light on the potential commercial value of the theories.

References

[1] M. Allman, V. Paxson, and W. Stevens, "TCP Congestion Control," Internet RFC2581, 1999, http://www.ietf.cnri.reston.va.us/rfc/rfc2581.txt.

[2] P. Barta, F. Németh, R. Szabó, and J. Bíró, "Call Admission Control in Generalized Processor Sharing Schedulers with Tight Deterministic Delay Bounds," *Computer Communications*, Vol. 26, No. 2, 2003, pp. 65-78.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," Internet RFC2475, 1998, http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc2475.html.

[4] R. Braden, Ed., L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification," Internet RFC2205, September 1997, http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc2205.html.

[5] Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queuing Algorithm," *Internetworking: Research and Experience*, Vol. 1, 1990, pp.3-26.

[6] S. Floyd and V. Jacobson, "Random Early Detection (RED) Gateways for Congestion Avoidance," *IEEE/ACM Transactions on Networking*, Vol. 1, 1993, pp. 397-413.

[7] L. Gommans, C. de Laat, B. van Oudenaarde, and A. Taal, "Authorization of a QoS path based on generic AAA," *Future Generation Computer Systems*, Vol. 19, No. 6, 2003, pp. 1009-1016.

[8] K. Gottschalk, S. Graham, H. Kreger, and J. Snell, "Introduction to Web Services Architecture," *IBM Systems Journal*, Vol. 41, No. 2, 2002, pp. 170-177.

[9] A. Gupta, D. O. Stahl, and A. B. Whinston, "A Stochastic Equilibrium Model of Internet Pricing," *Journal of Economic Dynamics and Control*, Vol. 21, 1997, pp. 697-722.

[10] A. Gupta, D. O. Stahl, and A. B. Whinston, "Managing Computing Resources in Intranets: an Electronic Commerce Perspective," *Decision Support Systems*, Vol. 24, 1998, pp. 55–69.

[11] IDC, "U.S. Web Services Market Analysis," 2002, http://www.idc.com/getdoc.jhtml?containerId=28493.

[12] V. Jacobson and M. J. Karels, "Congestion Avoidance and Control," *SIGCOMM'88*, 1988.

[13] Keynote, Keynote Business 40 Internet Performance Index, August 2003, http://www.keynote.com/solutions/performance_indices/business_index/business_40.html.

[14] H. Kreger, "Fulfilling the Web Services Promise," *Communications of the ACM*, Vol. 46, No. 6, 2003, pp. 29-34.

[15] D. Li, Z. Lin, D. O. Stahl, and A. B. Whinston, "Bridging Agent-based Simulations and Direct Experiments - An Experimental System for Internet Traffic Pricing," *AMCIS'01*, Boston, August 3-5, 2001.

[16] B. Liljeqvist and L. Bengtsson, "The Wire Speed Grid Project," 2002, http://www.ce.chalmers.se/staff/labe/Wire_Speed_Grid_Project.htm.

[17] Z. Lin, P. S. Ow, D. O. Stahl, and A. B. Whinston, "Exploring Traffic Pricing for the Virtual Private Network," *Information Technology and Management*, No.3, October 2002, pp. 301-327.

[18] G. Miller, "The Web Services Debase: .Net vs. J2EE," *Communications of the ACM*, Vol. 46, No. 6, 2003, pp. 64-67.

[19] J. Nagle, "Congestion Control in TCP/IP Internetworks," Internet RFC 896 (1984), http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc896.html.

[20] T. M. O'Donovan, "Direct Solutions of M/G/1 Processor-Sharing Models," *Operations Research*, Vol. 22 No. 6, Nov/Dec 1974, pp. 1232-1235

[21] J. Williams, "The Web Services Debase: J2EE vs. .Net," *Communications of the ACM*, Vol. 46, No. 6, 2003, pp. 59-63.