Performance evaluation for mobile access to composite Web services

Céline Boutrous-Saab
Laboratoire Lamsade
Université Paris Dauphine
75775 Paris cedex 16, France
Email: saab@dauphine.fr
Phone: (331) 44 05 41 83
Fax: (331) 44 05 40 91

Tarek Melliti
Laboratoire LRI
Université de Paris Sud
91405 Orsay cedex, France
Email: melliti@lri.fr
Phone: (331) 69 15 76 02
Fax: (331) 69 15 65 86

Lynda Mokdad
Laboratoire Lamsade
Université de Paris Dauphine
75775 Paris cedex 16, France
Email: mokdad@dauphine.fr
Phone: (331) 44 05 40 60
Fax: (331) 44 05 40 91

Abstract

In this paper, we propose a performance evaluation of mobile access to web services considering the gateway scenario. The model takes into account the gateway message processing strategies and the complexity of the web services. The evaluation of requests process is made by Markov chain. We show how the solution of this model is a product-form. Thus the performances measures can be deduced easily.

Keywords: Mobile Networks, Quality of Service, Mobile Web Services, Continuous Time Markov Chains.

1. Introduction

Web Services signal a new area of lightweight distributed application development. One of the design goals for Web Services is to allow companies and developers to share services with other companies in a simple way over the Internet. They are an easy way to create and consume services over the Internet. Web services are self contained, self-describing modular applications that can be published, located, and invoked across the Web.

However, certain type of applications requires developing more composite Web service in order to achieve more sophisticated application purposes. Then, with the success of mobile devices like cellulars and pda, Web service access becomes a necessity.

However, most existing mobile devices can not support XML web services. This is why an alternative mode of access should be provided. Thus, an intermediate module called gateway is usually used in order to play the role of interface between the mobile device and Web services. The drawbacks of such a module is that it only supports asynchronous communication and may have some effect on the performance and QoS. In this paper, we propose a performance evaluation study of mobile access to web services considering the gateway scenario. The model takes into account the gateway message processing strategies and the complexity of the web services. We use Markov chain in order to evaluate the performance of the requests process.

This paper is organised as follows : in Section 2, we present the Web service concept and introduce composite and mobile Web services. Section 3 de£nes the model after, we present some numerical computations in Section 4. Finally, last Section presents the main contribution of this work and gives future prospects.

2. Web Services

The huge success of Web Services is due to the fact that they show strong indications of cross-platform, cross-language compatibility [11]. In fact, Web services are built over XML and their framework is divided into three areas (communication protocol, service description, and service discovery) and speci£cations are being developed for each one:

- the Simple Object Access Protocol (SOAP) [10], which enables communication among Web Services,
- the Universal Description, Discovery and Integration (UDDI) [1], which is a registry of Web Services descriptions
- the Web Services Description Language (WSDL) [5], which provides a formal, computer-readable description of Web services.

Web services collaboration relies on an interaction model where the different components, service provider, directory service, and service consumer ensures the following roles:

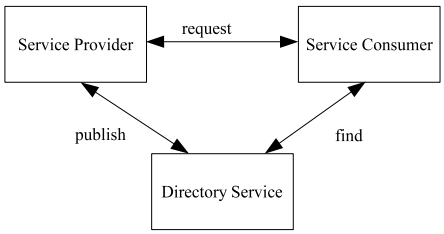

Figure 1. Interaction between the different components

- Service provider hosts an application which implements one or several web services. It should register the service into a directory server (UDDI).
- Service consumer locates the desired service by searching the directory server. Once located, it communicates directly with the service provider.
- Directory service is used by a provider to publish or update information concerning its services and by a consumer to search and locate services.

The relationship between these three architectural blocks is shown in the £gure 1.

A typical Web services scenario, is a business application (client) sends a request to a service at an URL. The service receives the request, processes it, and returns a response. An often-cited example of a Web service is that of a stock quote service, in which the request asks for the current price of a speci£ed stock and the response returns the stock price.

Web service technologies are currently limited to support simple services computations. This limitation is due essentially to the WSDL standard semantics [5]. In fact, WSDL speci£cation describes a Web service as a set of basic operation. It offers four operation models ; each model is de£ned by a formatted messages exchange schemas ; request-response and noti£cation-response, for two ways operation and solicitation, noti£cation for one way operation. For the two ways operation's model, the WSDL specify either a synchronous or an asynchronous execution of the data exchange (e.g RPC model for synchronous and document model for asynchronous).

Thus, certain type of applications require developing more composite Web service in order to achieve more sophisticated application purposes. This kind of services may often be encountered in two cases. First, when a web service is developed as an agent, it is composed of a set of accessible operations and a process model which schedules the invocation to a correct use of the service. Secondly, facing the capability limits of Web services, composite services may be obtained by aggregating (composing) existing Web services in order to create more sophisticated services (and this in a recursive way).

2.1. Web Service Composition

Composition requires the de£nition of collaboration activities and data exchange messages between involved Web services. Different mechanisms are used to describe the composition of Web services into more complex processes. Depending on the requirement, there are two manner of composition: orchestration or choreography.

In orchestration, a process takes control over the involved Web services, coordinates and manages their invocation. Thus, the involved Web services are unaware of the existing of other services and that they are involved into a composition process.

Choreography does not rely on a central process. Rather, each service is aware of the composition process and knows exactly what to do, how, and with whom to interact. It is more a collaborative work, where each service has his own role.

In both cases, the composition is totally transparent for the client. In fact, despite the manner a Web service is composed, it is still viewed as a simple Web service, i.e. it receives an invocation of a service, executes it and sends a reply. The only difference between a composite Web service and a simple one (not composed of other Web services) is that the composite one may invoke one or several simple or composite Web services in order to execute the service.

2.2. Mobile devices to access Web services

Today research and technological solution mixing web services and mobile devices go in two directions. The £rst consider the services as mobile entities moving in devices for a local access[6] [2]. The second consider mobile devices to access web services. The main objectif of mobile Web service to create Web service standards that will enable new business opportunities in the mobile space and to deliver integrated services across £xed (wired) and mobile networks[9, 8]. Mobile Web services use existing industry standard Extensible Markup Language (XML)-based Web services architecture to expose mobile network services to the broadest audience of developers.

In fact, mobiles are more and more used, not only to carry out basic tasks but also to offer Internet connection and the possibility of bene£ting from all available services. Thus, a customer does not limit himself any more to the use of cellular phone as a mean of communication but he would like also to use it, for example, to know the ¤uctuation of the market stock, etc. Actually most mobile network offers services that allow information to be pushed to mobile devices or to access some £xed services. Web services, in the context of mobile computing, is about the notion of devices

that can move in and out of service areas, and at the same time £nd and invoke Web services as needed. But, before it can be possible, certain architectural and techniques consideration must be addressed in order to handel mobility and performance aspects to access Web services. Two types of access can be considered.

direct access the £rst approach consider that mobile devices can directly access web services. Which means that such devices can support a SOAP client. This solution is very restrictive for universal access scenario and thus for several reasons. First of all, it supposes that all the devices (phones or PAD) can support Web services. Secondly, the access might have some performance costs related to slow data speeds while SOAP message are XML based and the XML data need a large bandwidth [7] [13] for SOAP latency studies. Finally, this solution requires additional tools to hide the complexity.

Indirect access the second approach, which is considered in this work, is a two phases access. It involves an intermediary entity called **gateway**[14]. A gateway plays the role of a SOAP/HTTP client by handling the request and the response. It returns results back to the mobile device in a supported format such as text-message, voice data or services Data. However, the gateway approche presents some drawbacks. Mainly, the con£guration can not handle a synchronous web service access while the user can switch area during the session. Synchronous and mobile devices are not terms that go well together. Thus, this con£guration requires asynchronous Web services [1].

2.3. discussion

Each of the two scenarios (direct and indirect accesses) presents some advantages and drawbacks. The direct access seems to be more suitable for existing web services (it supports easily synchronous access, no modi£cation of the Web services implementation) but it requires more technological constraints which go against an universal access. The indirect scenario based on the gateways is more realistic but it decreases the Qos and the Web services performance. The performance decrease rate depends on the implementation of the gateways and the complexities of the web services (composite ones). In this paper, we propose a performance evaluation study of mobile access to web services considering the gateway scenario. The model takes into account the gateway message processing strategies and the complexity

[1]the Open Mobile Alliance (OMA) announced the public availability of new and "up-leveled" mobile speci£cations which provide guidelines for Web services implementations within the OMA architecture, and how to leverage SOA in the world of mobile devices

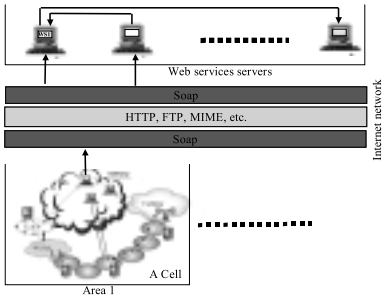

Figure 2. Architecture for mobile Web services

of the web services. We use Markov chain in order to evaluate the performance of the requests process.

2.4. Architecture for mobile device Web services access

In this paper, we consider an indirect access architecture using gateways. The wireless environment, in our architecture (see £gure 2), is composed of a set of cells. Each cell corresponds to a geographical area covered by the wireless network. In each cell, there is a set of Gateways aiming to handle users' requests in the corresponding area.

To access a Web service, the protocol is as following:

1. The User Side (US) can ask the gateways to £nd a Web service. The Gateway Side (GS) offers the search task as cell internal services and thus with the appropriate technology of the device (e.g Wap, sms, etc). The GS searchs in UDDI registries and returns the located Web services for choice.
2. The US can ask for information concerning a giving Web service. GS locates its WSDL and returns an abstraction (operation and needed information) in an appropriate format (e.g emulation of a SOAP client in WML page).
3. The US can choose the operations and send the request. GS creates a client for the corresponding service, formates the user information in SOAP message and invokes the service. Note that, in our architecture, we consider only asynchronous services. Web services developed to be used by mobile users must be asynchronous and must prevent a vocal response (number to call) or by SMS.

The £rst and the second steps are optional if the user already knows the Web services. So, the Web service invocation considered here is a request sent from the user in a context of speci£c cell to be processed by one of the associated gateways. In the frame of this architecture the performance[15]

of the Web services access depends on how to response to three questions:

1. how a given area Cell treat the user requests ?
2. how the Web service treat the client request?
3. is it an simple or a composed one?

In this study, we propose a performance evaluation model for this architecture in the context of a speci£c conceptual solution to each of the raised questions. First of all, we consider that each area or cell is composed of £xed number of gateways. Secondly, user requests, for a given cell, are queued in a shared buffer (a buffer for each cell). User requests are processed by the £rst free gateway according to FCFS (First Come First Served) discipline. A user request is moved to the target cell buffer if the user switches to another cell (before his request being processed). The gateway is free when it sends the SOAP request to the Web service. Finally, we suppose that all the web services server use the same strategy which mean that the client request are queued and processed according to FCFS discipline. In addition, we consider that a Web service is either a simple or a composite one within a £xed probability. So, a client request is processed when the most simple service (in the composition tree) processes the request.

3. The model

In this model, we present a mathematical model associated to a considered architecture. We consider a model in which users move along an arbitrary topology of K cells. Each cell has the same capacity of m channels. In our previous study [4], we have evaluate the performances of mobile networks without web services. In this paper, we consider that we have N web servers.

We suppose that the arrivals of requests in each cell i follow a Poisson process with rate λ_i, for $1 \leq i \leq K$. The distribution of the service time in each cell i is assumed to be an exponential distribution with rate μ_{c_i}.

The queuing discipline in the Web server queues is assumed to be FCFS. If a user from cell i did not change its position, its request is routed to the corresponding Web server queue j, with probability p_{i0j}, $1 \leq j \leq N$, else the request is routed to the cell k where the user is actually connected, with the probability p_{ik} for $1 \leq k \leq K$.

The distribution of the service time in each Web server i is assumed to be an exponential distribution with rate μ_{s_i}. When a request takes place in the Web server queue i, $1 \leq i \leq N$, it is served as follows: at this end of the service in Web server queue i, if the current request needs to be served by another Web server, the request is routed to the server queue j with probability q_{ij} otherwise it is routed outside with probability q_{i0}.

The considered model can be described by a continuous time Markov chain denoted $X(t)$. To describe this chain, we de£ne the state x by $(x_{c_1}, x_{c_2}, ..., x_{c_K}, x_{s_1}, x_{s_2}, ..., x_{s_N})$ where

x_{c_i} is the number of requests in the cell i, and x_{s_j} is the number of requests in Web server s_j, $\forall 1 \leq c_i \leq K$ and $\forall 1 \leq s_j \leq N$.

Thus, $X(t)$ constitutes Markov chain in a continuous time. We denote by $\Pi(x)$ its stationary probability distribution.

In order to understand the behavior of $X(t)$, we give its state evolution of $x(t)$.

$$
\begin{array}{lll}
x & (1). \rightarrow & (\ldots, x_{c_i}+1, \ldots), \\
 & & with\ rate\ \ \lambda_i \\
 & (2). \rightarrow & (\ldots, x_{c_i}-1, \ldots, x_{c_j}+1), \ldots) \\
 & & with\ rate\ \ x_{c_i}\ \mu_{c_i} p_{ij} \\
 & & if\ \ x_{c_i} < m\ \ or\ \ m\mu_{c_i} p_{ij}\ \ else \\
 & (3). \rightarrow & (\ldots, x_{c_i}-1, \ldots, x_{s_j}+1), \ldots) \\
 & & with\ rate\ \ x_{c_i} \mu_{c_i} p_{i0j} \\
 & & if\ \ x_{c_i} < m\ \ or\ \ m\mu_{c_i} p_{i0j}\ \ else \\
 & (4). \rightarrow & (\ldots, x_{s_i}-1, \ldots) \\
 & & with\ rate\ \ \mu_{s_i} q_{i0}\ \ if\ \ x_{s_i} > 0 \\
 & (5). \rightarrow & (\ldots, x_{s_i}-1, \ldots, x_{s_j}+1), \ldots) \\
 & & with\ rate\ \ \mu_{s_i} q_{ij}\ \ if\ \ x_{s_i} > 0
\end{array}
$$

In the previous evolution equations, the different lines mean:

- (1) means that there is an arrival of a request in cell i, $1 \leq i \leq K$ with rate λ_i.
- (2) means that there is a service of request in cell i, with rate $x_{c_i}\ \mu_{c_i}$ if $x_{c_i} < m$ else with rate $m\mu_{c_i}$. but the user have already moves from cell i to cell j, thus the request is moved to the corresponding cell j with probability p_{ij} to be served by this cell.
- (3) is the same as (2) but the user is still in the same cell, so the request is routed to the corresponding Web server s_j to be served with the probability p_{ij}.
- (4) and (5) mean that the request is served by the Web server s_i with rate μ_{s_i}. In (4), the request is routed outside with probability q_{i0} and in (5), the request needs to be served by another Web server s_j, so it is routed with probability q_{ij}

The considered model ful£l the following assumptions:

- External arrival is a Poisson process with rate λ_i

- In the cell queue i, the distribution of the service times is exponentially distributed with rate μ_{c_i} and the discipline of the service is FCFS (First Come First Served).
- type of cell queues are (M/M/m)
- type of Web server queues are M/M/1

The considered model is Jackson network [3], thus, it has a product-form solution as follows:

$$\pi(x_{c_1}, x_{c_2}, ..., x_{c_K}, x_{s_1}, x_{s_2}, ..., x_{s_N} = \Pi_{i=1}^{K} \pi_{c_i}(x_{c_i}) \Pi_{i=1}^{N} \pi_{s_i}(x_{s_i})$$

For the computation of $\pi_{c_i}(c_i)$, it is the solution of a classical M/M/m queue which is given in the following :

$$\pi_{c_i}(x_{c_i}) = \begin{cases} \pi_{c_i}(0) \frac{(m*\rho_{c_i})^{x_{c_i}}}{x_{c_i}!} & if \;\; 0 \leq x_{c_i} \leq m \\ \pi_{c_i}(0) \frac{\rho_{c_i}^{x_{c_i}} m^m}{m!} & if \;\; x_{c_i} \geq m \end{cases} \quad (1)$$

with $\rho_{c_i} = \frac{\lambda}{m\mu_{c_i}}$ and

$$\pi_{c_i}(0) = [\sum_{k=0}^{m-1} \frac{(m\rho_{c_i})^k}{k!} + \frac{(m\rho_{c_i})^m}{m!(1-\rho_{c_i})}]^{-1}$$

For the computation of $\pi_{s_i}(s_i)$, it is the solution of a classical M/M/1 queue which is given in the following:

$$\pi_{x_{s_i}} = (1-\rho_{s_i})\rho_{s_i}^{x_{s_i}} \text{ and } \rho_{s_i} = \frac{\lambda}{\mu_{s_i}}$$

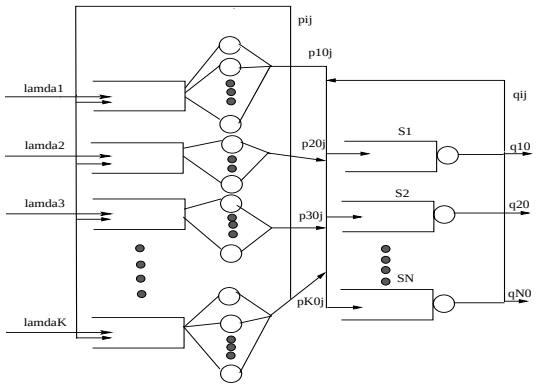

Figure 3. Model

The computation of the stationary probability distribution Π of the considered model is quite simple as explain before because it is independant of the values of K and N.

4. Numerical computation

In this Section, we give some numerical results of performance measures. All our computations were done on a Pentium-PC 1.6Ghz, 256MB, with Scilab 2.7.2 [12] under Windows 2000. We summarize all model parameter values in the table 1.

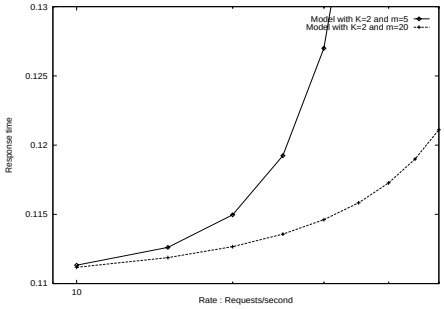

Figure 4. Response time for K=2

We plot the response time by varying the request arrival rate. Thus, in £gure 4, we have considered two cases. In one case, we have considered two cells and each cell is composed of 5 channels and in the other case, each cell is composed of 20 channels. We can notice that, in this £gure, we have obtained the expected results, in fact, the response time with $m = 5$ is upper than with $m = 20$. In £gure 5, we have also considered two cases. In one case, we have considered three cells and each cell is composed of 5 channels and in the other case, each cell is composed of 20 channels. We make the same remark that in the case with two cells.

Thus, we can see in both £gures 4 and 5 that the response time increases when λ_i increases and of course the response time decreases when the channel number in each cell increases.

We did not vary all the parameters because the objective of the paper is to show how to model the considered architecture by a Markov chain and how the resolution is very simple (product-form network). Thus, we can deduce easily several performance indices from the solution of the model.

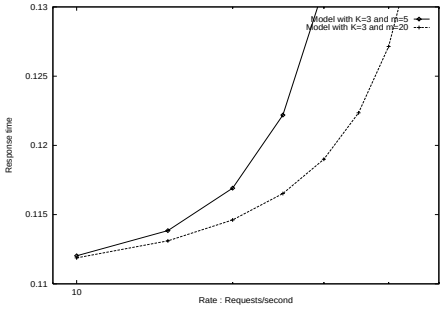

Figure 5. Response time for K=3

Cell number in the network	$K = 2$ or $K = 3$
Channel number in each cell	$m = 5$ or $m = 20$
Web server number	$N = 2$
Request arrival rate	varied from λ_i=10 requests/second to λ_i=50 requests/second requsets/second
Routing probabilities from cell 1	$p_{101} = 0.45$, $p_{102} = 0.45$, $p_{12} = 0.05$, $p_{13} = 0.05$
Routing probabilities from cell 2	$p_{201} = 0.45$, $p_{202} = 0.45$, $p_{22} = 0.05$, $p_{23} = 0.05$
Routing probabilities from cell 3	$p_{301} = 0.45$, $p_{302} = 0.45$, $p_{32} = 0.05$, $p_{33} = 0.05$
Routing probabilities from Web server 1	$q_{12} = 0.05$, $q_{10} = 0.95$
Routing probabilities from Web server 2	$q_{21} = 0.05$, $q_{20} = 0.95$

Table 1. Model parameter values

5. Conclusion

In this work, we have proposed a model based on Markov chains in order to evaluate an architecture based on mobile access to composite web services considering the gateway scenario The model takes into account the gateway message processing strategies and the complexity of composite web services. The advantage of this model is that it is a product-form queuing network. The considered model ful£l the Jackson assumptions. Thus, it has a product form solution and so, it can be solved easily and several performance measures can be deduced.

In future work, we will consider the requests with different priority levels in order to take into account their Quality of Service (QoS). We will also consider in the model different strategies for service disciplines.

References

[1] T. Bellwood, L. Clment, and C. von Riegen. Universal description, discovery and integration. Technical report, OASIS UDDI Speci£cation Technical Committee, mar 2002. http://www.oasis-open.org/cover/uddi.html.

[2] Boualem Benatallah, Quan Z. Sheng, and Zakaria Maamar. On composite web services provisioning in an environment of £xed and mobile computing resources, March 25 2003.

[3] G. Block, S. Greiner, and K.S. De Meer, H. Trivedi. *Queueing Networks and Markov chains*. John Wiley and sons, 1998.

[4] H. Catel-Taleb and L. Mokdad. Performance measure bounds in mobile networks by state space reduction. *The 13th IEEE / ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS-05)*, 2005.

[5] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana. Web services description language (WSDL) 1.1. Technical report, World Wide Web Consortium, mar 2001. http://www.w3.org/TR/wsdl.

[6] Frank P. Coyle. Mobile computing, web services and the semantic, August 22 2002.

[7] M. Parashar D. Davis. Latency performance of soap implementations. In *In Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid2002)*, pages 377 – 382, 2002.

[8] David Dahlem, Jens H. Jahnke, Luay Kawasme, and Yury Bychkov. Towards context oriented web services for smart personal object technologies (COWSPOTS) david dahlem, yury bychkov, luay kawasme, jens H. jahnke, September 17 2003.

[9] Kamal Elbashir. Transparent caching of web services for mobile devices, June 10 2004.

[10] M. Gudgin, M. Hadley, N. Mendelsohn, J. Moreau, and H. Nielsen. Simple object access protocol (soap) 1.1. Technical report, World Wide Web Consortium, may 2000. http://www.w3.org/TR/SOAP/.

[11] S. Haddad, T. Melliti, P Moreaux, and S. Rampacek. Modelling web services interoperability. In *In Proceedings of the 6th Int. Conf. on Entreprise Information Systems (ICEI04)*, pages 14–17, 2004.

[12] INRIA. Scilab home page: `http://www.scilab.org`, 2002.

[13] Arun Iyengar, Heiko Ludwig, Isabelle Rouvellou, and Richard King. Performance and service level considerations for, November 03 2003.

[14] Niels Christian Juul and Niels Jrgensen. WAP may stumble over the gateway (security in WAP-based mobile commerce), August 20 2001.

[15] H. Ritter, J. Schiller, M. Tian, T. Naumowicz, and T. Voigt. Performance considerations for mobile web services, May 22 2003.