УДК 004.75.05

A.V. GORBENKO

National aerospace university named after N. Ye. Zhukovskiy "KhAI", Ukraine

INSTABILITY ANALYSIS OF DELAYS CONTRIBUTING TO WEB SERVICE RESPONSE TIME

This paper reports our practical experience of benchmarking a complex System Biology Web Service, and investigates the instability of its behaviour and the delays induced by the communication medium. We present the results of our statistical data analysis and the distributions, describing the response time instability typical of Service-Oriented Architectures (SOAs) built over the Internet. We believe that performance uncertainty can be reduced by further optimisation of the internal structure and the right choice of components and technologies that suit each other and fit the system requirements better.

Keywords: *Web Service, Service-Oriented Architecture, Response Time, Instability.*

Introduction

Service-Oriented Systems are built as overlay networks over the Internet. Therefore, their dependable construction and composition are complicated by the fact that the Internet is a poor communication medium (has low quality and is not predictable). They can be vulnerable to internal faults from various sources and casual external problems such as communication failures, routing errors and network traffic congestions. Therefore, the performance of such system is characterised by high instability [1], i.e. it can vary over a wide range in a random and unpredictable manner.

While loosely-coupled Web Services (WSs) can be a desirable platform for building e-science applications [2], up to now there has not been sufficient experimental assessment of the dependability of such applications. Dependability is a major concern in service-oriented environments used for e-science projects, which are typically implemented through distributed global collaborations enabled by the Internet and involve formation of virtual organizations on an ad hoc basis. Ensuring dependability of services is particularly important in bioinformatics. Academic and non-commercial organisations deploy WSs to be used by scientists from the life science community without any prior service level agreements. Scientists have no other option but to use such services despite the fact that they may not always be reliable and/or available [3].

For complex bioinformatics workflows incorporating many different WSs some users may get a correct service, whereas others may perceive incorrect result of different types due to timing errors. These errors may occur in different system components depending on the relative position in the Internet of a particular user and particular WSs, and, also, on the instability points appearing during the execution. Thus, timing errors can become a major cause of inconsistent failures.

Recent works related to WS dependability, e.g. [4 – 6], have introduced several approaches to incorporating resilience techniques (including voting, backward and forward error recovery mechanisms and replication techniques) into WS architectures. There have been also some works on benchmarking and experimental measurements of dependability [7, 8]. However, this works do not refer to the uncertainty issue of WS.

In this work we use the general synthetic term *uncertainty* to refer to the unknown, unstable, unpredictable, changeable characteristics and behaviour of WS and SOA, exacerbated by running these services over the Internet. Understanding uncertainty arising in SOA is crucial for choosing right recovery techniques, setting timeouts, and adopting system architecture and its behaviour to such changing environment like the Internet and SOA.

In this paper, we present a set of new experiments we have recently conducted on a BASIS (Biology of Ageing E-Science Integration and Simulation System) Web Service [9] deployed at Newcastle University's Institute for Aging and Health as part of our research into dependability of WSs and SOA. This paper reports a continuation of our previous work aiming to measure the performance and dependability of e-science WSs from the end user's perspective [1, 10]. In the previous investigation we found evident performance instability existing in these SOAs and affecting dependability of both, the WSs and their clients. However, we were unable to capture the exact causes and shapes of performance instability.

The main difference between [10] and the work reported in this paper is that we have used an improved experimental technique and benchmarking software to measure two main delays contributing to the response time (RT) and its instability: the request processing time (RPT) (by a Web Service) and the network round trip time (RTT), i.e. RT = RPT + RTT.

Besides, we investigated how the performance of the BASIS WS and its instability changed during the 3 months since our previous large-scale experiment to check the hypothesis that once measured they stay true.

1. Method and Experimental Settings

The BASIS application is a typical, representative example of a number of SOA solutions found in e-science and grid. Being one of the twenty pilot projects funded under the UK e-science initiative in the development of the UK grid applications, BASIS aims at developing web-based services that help the biology-of-ageing research community in quantitative studies of the biology of ageing by integrating data and hypotheses from diverse biological sources. The BASIS WS integrates various components such as model design, simulators, databases, and exposes their functionalities as WSs to support simulation of biochemical reaction networks, metabolic networks, cell-signalling pathways and other kinds of systems studied in systems biology [11]. The experiments that we report here are a follow up to our previous work on dependability of WSs reported in [10]. In the new experiments we performed continuous remote WS testing (probing) to capture long-term performance trends and to unveil WS performance instabilities. This work was carried out in a way similar to the work reported in [10]. The BASIS WS, returning SMBL (Systems Biology Markup Language) simulation result of 500 Kb, has been invoked by the client software placed in five different locations (in Frankfurt, Moscow, Los Angeles and two in Simferopol) every 10 minutes during eighteen days starting from Apr, 11 2009 (more than 2500 times in total).

During each invocation we fixed four times the stamps that helped us to measure the main delays contributing to the WS response time (Fig. 1).

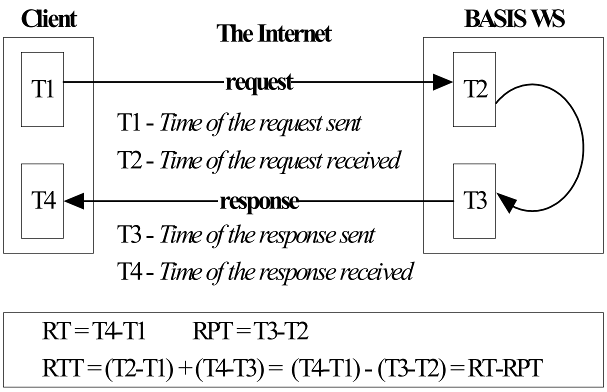

Fig. 1. Performance measurement

At the same time we traced the network route between the client and the BASIS SBML WS to understand how the Internet latency effects the WS invocation delay and to find out where possible the exact points of network instability.

After processing statistics for the all clients located in different places over the Internet we found the same uncertainty tendencies.

Thus, in the paper we report results obtaining only for the Frankfurt client.

2. Performance Trends Analysis

Performance trends of RPT, RTT and RT captured during eighteen days are shown at the Fig. 2.

It can be see that RTT and especially RPT have significant instability that contribute together to the instability of the total response time RT. Sometimes, delays were twenty times (and even more) longer than their average values (see Table 1).

In brackets we give estimation of the maximal and average values of RPT, RTT and RT and their standard deviations that were obtained after taking out of consideration ten the most extreme delays. A ratio between delay's standard deviation and its average value is used as the *uncertainty measure*.

As compared with our experiments of three month prescription we have observed a significant increase of the average response time (see Table 1) fixed by the Frankfurt client (889.7 ms instead of 502.15 ms).
In additional to this, an uncertainty of BASIS performance from the client-side perspectives has been increased in times (94.1% instead of 18.77%). The network route between the BASIS WS and the Frankfurt client has also changed significantly (18 intermediate routers instead of 11).

3. Probability Density Analysis

Probability density series corresponding to performance trends of RPT, RTT and RT (see Fig. 2) are shown at the Fig. 3.

In previous works [1, 10] we used ten bars in the histograms representing probability density series. In our current work we set the number of bars in the histogram equal to the square root of the number of elements in experimental data that is similar to Matlab histfit(x) function.

This allowed us to find out new interesting properties. In particular, we could see that about 5% of RPT, RTT and RT are significantly larger than their average values. It is also clear that the probability distribution series of RTT has two extreme points. This peculiarity of RTT causes an appearance of the observable left tail in the RT probability distribution series.

Table 1

Performance statistics: RPT, RTT, RT

	Minimal value (Min.), ms	Maximal value (Max.), ms	Average value (Avg.), ms	Standard Deviation (Std. Dev.), ms	Std. Dev. / Avg., %
RPT	287.0	241106.0 (8182.0)	657.7 (497.6)	4988.0 (773.5)	758.4 (155.4)
RTT	210.0	19445.0 (1479.0)	405.8 (378.2)	621.1 (49.2)	153.1 (13.0)
RT	616.0	241492.0 (11224.0)	1061.5 (889.7)	5031.0 (837.4)	474.1 (94.1)
Ping RTT	26.4	346.9 (50.4)	32.0 (31.9)	3.6 (0.9)	11.3 (2.8)

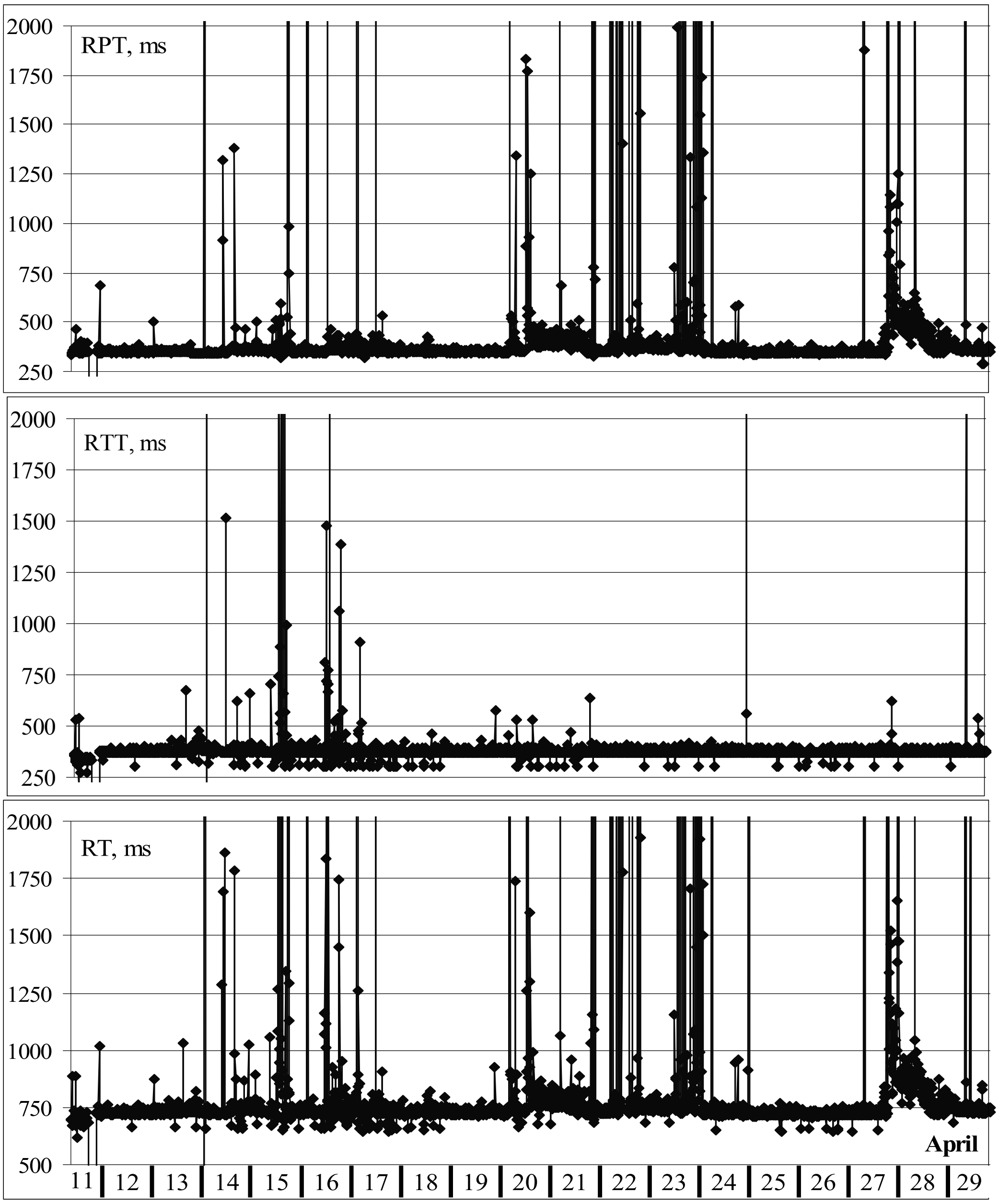

Fig. 2. Performance trends of RPT, RTT and RT

It also makes it difficult to find the theoretical distribution, representing RTT. Finally, it is clear that probability distribution series of neither RT, RPT nor RTT follow Exponential low. Tracing routes between the client and the service allows us to conclude that these fast responses were caused by shortening the network routes. This seems to be very unusual for RPT but should be typical for the Internet.

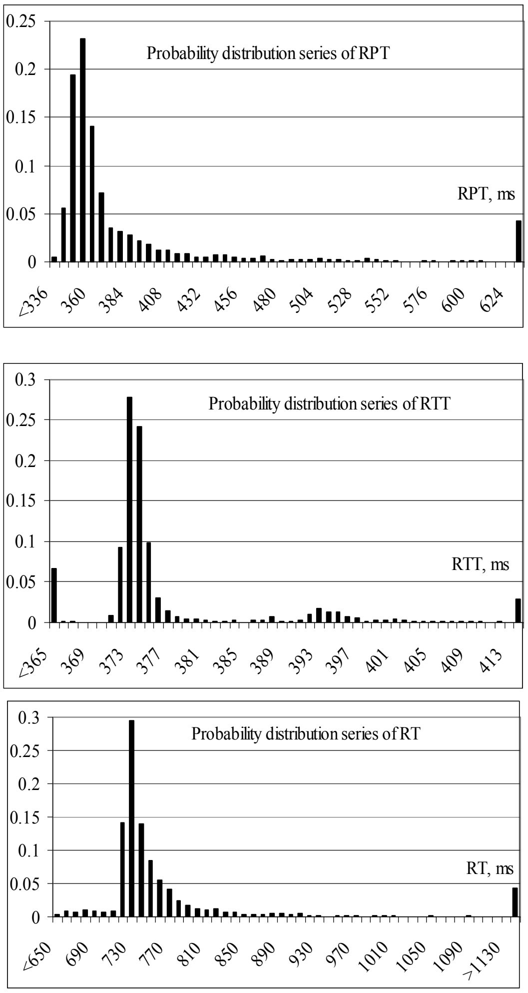

Fig. 3. Probability distribution series of RPT, RTT and RT

As the reliability concern we should mention that BASIS WS was unavailable for four hours (starting from 19:00 Apr, 11) because of the network rerouting. Besides, two times the BASIS WS reported an exception instead of returning the normal results.

Besides, more than five percents of RTT have value that is 80ms (1/5) less than the average one.

Conclusion

In the experiments performed the major uncertainty came from the BASIS WS itself, whereas in the experiments conducted three month before this (during the Christmas week [10]) the Internet was the main cause of the uncertainty. The reason for this might be in the low internal load of the BASIS server and just a few clients using BASIS SW during that week. At the same time it is worth noting that the overall Internet activity typically grows during this time as social networks (e.g. Facebook) and forums experience a sharp growth during the holidays [12].

An important fact we found is that RPT has higher instability than RTT. The probability distribution series of RTT has unique characteristics making it really difficult to describe them theoretically. We also should notice that performance and other dependability characteristics of WSs could become out of date very quickly.

Thus, the BASIS response time has changed significantly after three month in spite of the fact that there were no essential changes in its architecture apart from changes of the usage profile and the Internet routes. Once measured the non-functional characteristics of WSs cannot be assumed to be true forever.

The BASIS WS is a typical example of a number of SOA solutions found in e-science and grid. It has a rather complex structure which integrates a number of components, such as a SBML modeller and simulator, database, grid computing engine, computing cluster, etc., typically used for many *in silico* studies in systems biology.

We believe that performance uncertainty, which is partially due to the systems themselves, can be reduced by further optimisation of the internal structure and the right choice of components and technologies that suit each other and fit the system requirements better. To this end, developing a general dependable architecture (or set of architectural patterns) for system biology applications would be a significant step toward more dependable and predictable e-science services.

Finally, our concrete suggestion for bio-scientists using BASIS is to set up a time out that is 1.2 times longer than the average response time estimated for 20-25 last requests. When the time out is exceeded, a recovery action based on a simple retry can be effective most of the time in dealing with transient congestions happening in the Internet and/or the BASIS WS.

A more sophisticated technique that would predict the response time more precisely and set up the time out should assess the average response time and the ratio between standard deviation of the response time and its average value.

To be more dependable, clients should also distinguish between different exceptions and handle them in different ways depending on the exception source [13].

References

1. Gorbenko A. The Threat of Uncertainty in Service-Oriented Architecture / A. Gorbenko // Proc. 1st Int. Workshop on Software Engineering for Resilient Systems (SERENE'2008). – Newcastle upon Tyne (UK), 2008. – P. 49–54.

2. Oinn T. Taverna: a tool for the composition and enactment of bioinformatics workflows / T. Oinn // Bioinformatics. – 2004. – №20(17). – P. 3045-3054.

3. Stevens R. Exploring Williams-Beuren Syndrome Using myGrid [Text] / R. Stevens // Bioinformatics. – 2004. – №20(1). – P. 303-310.

4. Salatge, N. Fault Tolerance Connectors for Unre-

liable Web Services / N. Salatge, J.-C. Fabre // Proc. Int. Conf. On Dependable Systems and Networks (DSN'2007). – Edinburgh (UK), 2007. – P. 51–60.

5. Zheng. Z. A QoS-Aware Fault Tolerant Middleware for Dependable Service Composition / Z. Zheng, M. Lyu // Proc. Int. Conf. On Dependable Systems and Networks (DSN'2009). – Lisbon (Portugal), 2009. – P. 239– 248.

6. Looker, N. Increasing Web Service Dependability Through Consensus Voting / N. Looker, M. Munro, J. Xu // Proc. 2nd Int. Workshop on Quality Assurance and Testing of Web-Based Applications at COMPSAC'2005. – Edinburgh (UK), 2005. – P. 66–69.

7. Vieira, M. Assessing Robustness of Web-services Infrastructures [Text] / M. Vieira, N. Laranjeiro, H. Madeira // Proc. Int. Conf. On Dependable Systems and Networks (DSN'2007). – Edinburgh (UK), 2007. – P. 382–401.

8. Gonczy, L. Dependability Evaluation of Web Service-Based Processes [Text] / L. Gonczy, S. Chiaradonna, F. di Giandomenico // Proc. European Performance Engineering Workshop (EPEW'2006). – Bertinoro (Italy), 2006. – P. 166–180.

9. Towards an E-Biology of Ageing: Integrating Theory and Data [Text] / B. L. Thomas [et al.] // Journal of Nature Reviews Molecular Cell Biology. – 2003. – №4.– P. 243–249.

10. Benchmarking Dependability of a System Biology Application [Text] / Y. Chen, A. Romanovsky, A. Gorbenko, V. Kharchenko, S. Mamutov, O. Tarasyuk // Proc. 14th IEEE Int. Conference on Engineering of Complex Computer Systems (ICECCS'2009). – Potsdam (Germany), 2009. – P. 146-153.

11. Welcome to BASIS [Electronic resource] / Institute for Ageing and Health, Newcastle University. Access mode: http://www.basis.ncl.ac.uk/.

12. Goad, R. Social Xmas: Facebook's busiest day ever, YouTube overtakes Hotmail, Social networks = 10% of UK Internet traffic [Electronic resource] / R. Goad. Access mode: weblogs.hitwise.com/robin-goad/2008/12/facebook_youtube_christmas_social_networking.html

13. Experimenting with exception propagation mechanisms in service-oriented architecture [Text] / A. Gorbenko [et al.] // Proc. 4th Int. Workshop on Exception Handling. – Atlanta (Georgia), 2008. – P. 1–7.

Поступила в редакцию 12.01.2010

Рецензент: д-р техн. наук, проф., В.С. Харченко, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков.

АНАЛІЗ НЕСТАБІЛЬНОСТІ ЗАТРИМОК, ЩО СКЛАДАЮТЬ ЧАС ВІДКЛИКУ WEB-СЛУЖБ

А.В. Горбенко

У статті представлено результати випробувань й практичного дослідження нестабільності часових характеристик Web-служби для проведення системних досліджень в галузі біо-інформатики. Представлені результати статистичної обробки експериментальних даних і побудови рядів розподілення випадкових затримок, які є складовими часу відклику Web-служби та характеризують нестабільність типової сервіс-орієнтованої архітектури, що розгорнута в мережі Інтернет. Ми впевнені, що нестабільність продуктивності може бути зменшена за рахунок оптимізації внутрішньої структури та правильного вибору компонент та технологій, що пасують один одному, а також більш пасують вимогам системи.

Ключові слова: Web-служба, сервіс-орієнтована архітектура, час відклику, нестабільність.

АНАЛИЗ НЕСТАБИЛЬНОСТИ ЗАДЕРЖЕК, СОСТАВЛЯЮЩИХ ВРЕМЯ ОТКЛИКА WEB-СЛУЖБ

А.В. Горбенко

В статье представлен отчет о результатах тестирования и практического исследования нестабильности временных характеристик Web-службы для проведения системных исследований в области био-информатики. Представлены результаты статистической обработки экспериментальных данных и построения рядов распределения случайных задержек, составляющих время отклика Web-службы и характеризующих нестабильность типовой сервис-ориентированной архитектуры, развернутой в Интернет. Мы уверены, что нестабильность производительности может быть уменьшена за счет оптимизации внутренней структуры и правильного выбора компонент и технологий, которые подходят друг другу, а также лучше подходят требованиям системы.

Ключевые слова: Web-служба, сервис-ориентированная архитектура, время отклика, нестабильность.

Горбенко Анатолий Викторович – к.т.н., доцент, доцент кафедры Компьютерных систем и сетей Национального аэрокосмического университета им. Н.Е. Жуковского «ХАИ», Харьков, Украина, e-mail: A.Gorbenko@csac.khai.edu.