УДК 004.75.05

A.V. GORBENKO

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

DETERMINATION OF DISTRIBUTION LAWS OF DELAYS CONTRIBUTING TO WEB-SERVICES RESPONSE TIME

The paper reports results of finding distribution laws representing Web-Services response time and delays contributing. Theoretical investigations provided are based on real-live statistics. Results of hypotheses checking are reported. Response time simulation approaches are described. Experimental investigation and mathematical analysis of response time are reported. Our experiments have shown that delays arising in Service-Oriented Architecture have unstable characteristics make them really difficult to describe theoretically over a long period of time.

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

Introduction

Web Services are becoming a critical technology in building modern distributed information systems for e-business, e-science, e-medicine applications, etc. Concerning e-science, the use of Web Services is especially prominent in Bioinformatics and Systems Biology projects that focus on processing large datasets, and sharing and exchanging them across different organizations and institutes.

Different Web Services are orchestrated into workflows describing experiments that carry out in silico what used to be conducted in vivo in laboratories, but involve the use of computational resources such as data repositories and analysis/simulation programs available on the Internet [1]. Such in silico experiments may be long-lived due to the large volumes of data being analysed, whilst there may also be requirements on the timeliness of the workflow enactment.

As far as Service-Oriented Systems are mainly built as overlay networks over the Internet 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 [2], i.e. it can vary over a wide range in a random and unpredictable manner.

Inability of the WSs involved to guarantee a certain response time and performance and the instability of the communication medium can cause timing failures, when the response time or the timing of service delivery (i.e., the time during which information is delivered over the network to the service interface) differs from the time required to execute the system function. A timing failure may take the form of early or late response, depending on whether the service is de-livered too early or too late [3]. For complex bioinformatics workflows incorporating many different WSs some users may get a correct service, whereas others may perceive incorrect services 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 usually referred to as the Byzantine failures.

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. **The purpose of the paper** is to find a way to predict and represent the performance uncertainty in Service-Oriented Architecture by employing one of the theoretical distributions, used to describe such random variables like the WS response time. A motivation for this is the fact shown by many studies (e.g. [4, 5]) that the Exponential distribution does not represent well the accidental delays in the Internet and SOA. This work aims at estimating and predicting of evident performance instability existing in these Service-Oriented Systems and affecting dependability of both, the WSs and their clients.

© A.V. Gorbenko

1. Response Time Statistics

Our theoretical investigations reported in this paper are based on real-live statistics gathered during long-term benchmarking of BASIS (Biology of Ageing E-Science Integration and Simulation System) Web Service [6] deployed at Newcastle University's Institute for Aging and Health as part of our research into dependability of WSs and SOA.

BASIS WS 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 two main delays contributing to the WS response time (RT): network round trip time (RTT) and request processing time (RPT) by the Web Service.

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 one.

Performance trends of RPT, RTT and RT and its probability distribution series captured during eighteen days by Frankfurt client are shown at the fig. 1. Distribution series were built with the help of Matlab histfit (x) function.

It can be seen 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. Besides, 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 and more than five percents of RTT have value that is 80ms (1/5) less than the average one. All these factors makes doubt about real distribution of overall response time and different delays contributing to it.

2. Hypothesis Checking Technique

In this section we provide results of hypotheses checking about distribution law of WS response time (RT) and its component values RPT and RTT. In our work we use the Matlab numeric computing environment (www.mathworks.com) and its Statistics Toolbox (a collection of tools supporting a wide range of general statistical functions, from random number generation, to curve fitting).

The techniques of hypothesis checking consist of two basic procedures. First, values of distribution parameters are to be estimated by analyzing experimental sample.

Second, the null hypothesis that experimental data have a particular distribution with certain parameters should be checked.

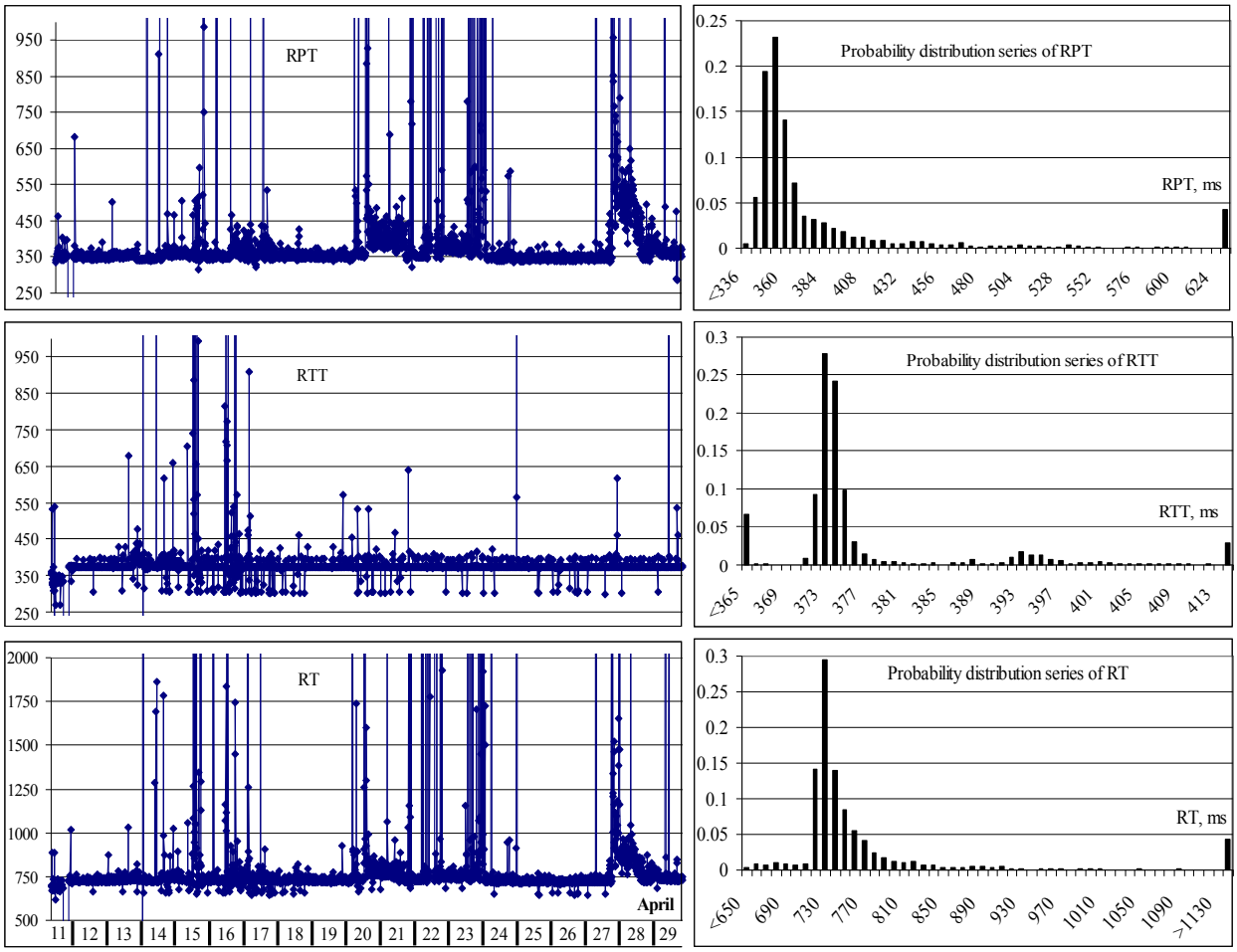

Fig. 1. Performance statistics and probability distribution series: RT, RTT and RPT

To check hypothesis itself we used the kstest function: [h, p] = kstest (x, cdf) performing a Kolmogorov-Smirnov test to compare the distribution of random variable x (i.e. response time statistic) to the hypothesized distribution defined by matrix cdf.

The null hypothesis for the Kolmogorov-Smirnov test is that x has a distribution defined by cdf. The alternative hypothesis is that x does not have that distribution. Result h is equal to "1" if we can reject the hypothesis, or "0" if we cannot reject that hypothesis. The function also returns the p-value which is the probability that x does not contradict the null hypothesis. We reject the hypothesis if the test is significant at the 5% level (if p-value less than 0.05).

Bellow we present an example of checking a hypothesis that the vector of ten samples *x* has the Exponential distribution $y = f(x \mid \mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$.

```
> x = [4;8;85;11;15;1;25;54;14;1]
> mu=expfit(x)
mu =   21.8000
> [h,p] = kstest(x, [x expcdf(x, mu)])
h =    0
p =    0.7574
```

As we can see, we cannot reject that hypothesis (h=0) and the p-value is good enough.

3. Goodness-of-Fit Analysis

In our experimental work we have checked six hypotheses that experimental data conform Exponential, Gamma, Beta, Normal, Weibull or Poisson distributions. These checks were performed for the request processing time (RPT), round trip time (RTT) and response time (RT) as a whole. Our main finding is that none of the distributions fits to describe the whole performance statistics, gathered during 18 days. Moreover, the more experimental data we used the worse approximation were provided by all distributions! It means that in the general case an uncertainty existing in Service-Oriented Architecture can not be predicted and described by analytic formula.

Our further work focused on finding the distribution law that fits the experimental data within limited time intervals. We have chosen two short time intervals with the most stable (from 0:24:28 of Apr, 12 until 1:17:50 of Apr, 14) and the least stable (from 8:31:20 of Apr, 23 until 22:51:36 of Apr, 23) response time.

The first time interval includes 293 request samples. Results of hypothesis checking for RPT, RTT and RT are given in Tables 2, 3 and 4 respectively. The p-value, returned by the kstest function, was used to estimate the goodness-of-fit of the hypothesis. As it can be seen, Beta, Weibull and especially Gamma (1) distributions fit the experimental data better than others. Besides, RPT is approximated by these distributions better than RT and RTT.

$$y = f(x \mid a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}, \qquad (1)$$

Typically, the Gamma probability density function (PDF) is useful in reliability models of lifetimes. This distribution is more flexible than the Exponential one, which is a special case of the Gamma function (when a=1). It is remarkable, that the Exponential distribution in our case describes experimental data worst of all. However, close approximation even by using the Gamma function can be achieved only within the limited sample interval (25 samples in our case). Moreover, RTT (and sometimes RT) can hardly be approximated even under such limited sample length.

For the second time interval all six hypotheses failed because of the low confidence of the p-value (less than confidence interval of 5%). Thus, we can state that the deviation of experimental data significantly affects goodness of fit. However, we also should mention that the Gamma distribution also gave better approximation than other five distributions.

4. Response Time Simulation

In many theoretical and experimental studies of the performance and dependability of distributed queuing systems it is necessary to simulate response time. It can be easily done if we know a distribution law describing this random variable. However, we have to remember that in practice (in accordance with our current study and [4]) theoretical distributions can approximate the response time in service-oriented systems well only within a limited time frame. Nevertheless, two simulation approaches are possible. Firstly, RT can be directly simulated by using a particular distribution function (i.e. Gamma) with the certain parameters. Secondly, we can take into account the fact that RT = RPT + RTT, where RPT and RTT are independent variables. In this case we deal with so called "composition" (2) of two distribution laws $f_1(\mathrm{RPT})$ and $f_2(\mathrm{RTT})$. In this section we are trying to answer the question what simulation approach is more accurate.

$$\begin{aligned} g(\mathrm{RT}) &= g(\mathrm{RPT}, \mathrm{RTT}) = f_1(\mathrm{RPT}) f_2(\mathrm{RTT}) = \\ &= \int_{-\infty}^{+\infty} f_1(\mathrm{RPT}) f_2(\mathrm{RT} - \mathrm{RTT}) d\mathrm{RPT} = \\ &= \int_{-\infty}^{+\infty} f_1(\mathrm{RT} - \mathrm{RTT}) f_2(\mathrm{RTT}) d\mathrm{RTT} \end{aligned} \qquad (2)$$

As observed in the previous section, the Exponential distribution does not fit the stochastic processes happening in the Internet and Service-Oriented Systems.

Within the limited time interval the Gamma distribution gives the best approximation of RPT, RTT and RT as a whole. Thus, $f_1(\mathrm{RPT})$ and $f_2(\mathrm{RTT})$ can be Gamma functions with individual parameters.

Table 1

RPT Goodness-of-fit approximation

Number of requests	Approximation goodness-of-fit (p-value)					
	Exp.	***Gam.***	Norm.	***Beta***	*Weib.*	Poiss.
293 (all)	7.8E-100	1.1E-06	9.5E-63	9.3E-25	2.3E-11	4.9E-66
First half	1.1E-99	0.0468	1.2E-62	0.0222	0.00023	1.1E-65
Second half	1.3E-47	*0.2554*	5.1E-30	***0.2907***	*0.0729*	1.6E-31
First 50	6.9E-18	***0.2456***	2.3E-11	*0.2149*	*0.0830*	7.5E-12
First 25	2.3E-09	***0.9773***	5.1E-06	*0.9670*	*0.5638*	2.9E-06
Second 25	2.5E-09	***0.2034***	5.2E-06	*0.1781*	*0.0508*	3.1E-06

Table 2

RTT Goodness-of-fit approximation

Number of requests	Distribution's goodness-of-fit (p-value)					
	Exp.	Gam.	Norm.	Beta	Weib.	Poiss.
293 (all)	2.1E-94	5.1E-30	4.4E-59	7.0E-39	5.0E-38	7.5E-85
First half	6.5E-52	2.6E-17	9.1E-33	1.1E-16	2.6E-19	1.0E-45
Second half	5.0E-44	2.5E-11	1.8E-27	4.6E-16	4.6E-13	8.1E-40
First 50	8.1E-18	1.9E-04	2.1E-11	2.9E-04	2.0E-07	2.1E-15
First 25	2.7E-09	0.004	4.2E-06	0.0043	0.0133	4.6E-08
Second 25	1.6E-09	6.0E-04	4.0E-06	5.4E-04	3.5E-04	4.8E-08

Table 3

RT Goodness-of-fit approximation

Number of requests	Distribution's goodness-of-fit (p-value)					
	Exp.	***Gam.***	Norm.	***Beta***	*Weib.*	Poiss.
293 (all)	1.6E-96	1.8E-14	4.4E-60	4.4E-29	1.0E-19	4.0E-67
First half	2.6E-52	0.0054	9.4E-33	0.0048	1.1E-06	2.6E-35
Second half	1.0E-45	9.8E-08	1.9E-28	5.2E-15	9.1E-09	2.2E-32
First 50	6.1E-18	***0.1159***	2.1E-11	*0.1083*	*0.1150*	6.1E-12
First 25	2.4E-09	*0.8776*	4.2E-06	***0.8909***	*0.7175*	2.7E-06
Second 25	1.9E-09	***0.0843***	4.5E-06	*0.0799*	0.0288	2.8E-06

To simulate g (RT) directly with the help of the Gamma distribution we should fit its parameters beforehand in a way similar to that described in section 4.2. Matlab function gamfit (x) can be used here. Another Matlab function gamrnd (a,b) that generates vector of gamma random numbers with parameters a and b can be used to simulate RT. An accuracy of simulated RT as compared to actual data obtained experimentally can be evaluated by use of the kstest2(x, y) function. This function performs a two-sample Kolmogorov-Smirnov test to compare the distributions of values in the two data vectors x and y. The null hypothesis for this test is that x and y have the same continuous distribution. The whole sequence of Matlab commands implementing the first simulation approach is as it shown bellow.

```
> RTpar = gamfit(RT)
> y = gamrnd(RTpar(1),RTpar(2),25,1)
> [h,p] = kstest2(y,RT)
```

The second simulation approach composing RPT and RTT can be easily implemented in the Matlab environment as well:

```
> RPTpar = gamfit(RPT)
> RTTpar = gamfit(RTT)
> x = gamrnd(RPTpar(1),RPTpar(2),25,1)
   + gamrnd(RTTpar(1),RTTpar(2),25,1)
> [h, p] = kstest2(x,RT)
```

Here, RT, RPT and RTT are vectors of the first 25 samples of the response time, the request processing time and the round trip time gathered experimentally starting from 0:24:28 of Apr, 12.

Average p-values corresponding to the first and the second simulation approaches are 0.69 and 0.57. They were estimated after performing thirty rounds of random generation. This shows that both simulation approaches can be used, however the first one provides better approximation to the experimental data.

Conclusion

Our main finding is that none of the distributions fits to describe the long-termed performance statistics. The more experimental data we used the worse approximation were provided by all distributions. It means that, in the general case, an uncertainty existing in SOA can not be predicted and described by analytic formula. According to section 3, goodness of fit was significant only within short time intervals which include no more than 20-30 samples.

Based on our experimental investigation and mathematical analysis reported in the paper we can state that RPT has higher instability than RTT, however, in spite of this RPT can be better represented using a particular

theoretical distribution. At the same time the probability distribution series of RTT has unique characteristics making it really difficult to describe them theoretically. Among the existing theoretical distributions the Gamma, Beta and Weibul capture our experimental response time statistics better than others.

The Matlab numeric computing environment provides powerful toolboxes and functions for statistical analysis of the experimental data in the types of the experiments we have been conducting. However, improving the prediction of WS performance needs more sophisticated procedures for experimental data processing (e.g. using dynamic time slots, rejecting some extreme samples, etc.) beforehand.

Our work supports the claim that dealing with the uncertainty inherent in the very nature of SOA and WSs, is one of the main challenges in building dependable SOA. Uncertainty has two consequences. First, it is difficult to assess the dependability and performance of services, and hence it is difficult to choose between them and gain confidence in their dependability. Secondly, it is difficult to execute fault tolerance mechanisms in a (close to) optimal manner, since too much data is missing to make good decisions and exploit all features of the dependability mechanisms.

Uncertainty of the Internet and service performance instability are such that on-line optimization of redundancy can make a substantial difference in perceived dependability, but currently there are no good tools available for the company to carry out such optimisation in a rigorous manner. We believe that uncertainty can be resolved by two means: uncertainty removal through advances in data collection and uncertainty tolerance through smart algorithms that improve decisions despite lack of data (e.g., by extrapolation, better mathematical models, etc.).

References

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

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

3. Basic Concepts and Taxonomy of Dependable and Secure Computing [Text] / A. Avizienis [et al.] // IEEE Trans. Dependable and Secure Computing. – 2004. – №1 (1). – P. 11-33.

4. Reinecke P. Experimental Analysis of the Correlation of HTTP GET invocations [Text] / P. Reinecke, A. van Moorsel, K. Wolter // Proc. 3rd European Performance Engineering Workshop (EPEW'2006). – Budapest (Hungary), 2006. – P. 226-237.

5. Maurer S. M. Restart strategies and Internet congestion [Text] / S. M. Maurer, B. A. Huberma // Journ. of Ec. Dyn. and Control. – 2004. – № 25. – P. 641-654.

6. Benchmarking Dependability of a System Biology Application [Text] / Y. Chen [et al.] // Proc. 14th IEEE Int. Conference on Engineering of Complex Computer Systems (ICECCS'2009). – Potsdam, 2009. – P. 146-153.

Поступила в редакцию 12.01.2010

Рецензент: д-р техн. наук, проф., зав. кафедры компьютерных систем и сетей В.С. Харченко, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина.

ВИЗНАЧЕННЯ ЗАКОНІВ РОЗПОДІЛУ ЗАТРИМОК, СКЛАДАЮЧИХ ЧАС ВІДКЛИКУ WEB-СЛУЖБ

А.В. Горбенко

У статті представлено результати пошуку законів розподілу, що описують час відклику Web-служб та його складові. Теоретичні дослідження базовані на реальній статистиці. Представлені результати перевірених гіпотез. Описані підходи до моделювання симуляції часу відклику. Обгрунтовано експериментальне дослідження і проведений математичний аналіз часу відклику. Визначено, що затримки, які виникають у Сервіс-Орієнтованих Системах мають дуже невизначені статистичні характеристики, що значно ускладнюють їхнє представлення впродовж тривалого часу за допомогою теоретичних законів розподілу випадкової величини.

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

ОПРЕДЕЛЕНИЕ ЗАКОНОВ РАСПРЕДЕЛЕНИЯ ЗАДЕРЖЕК, СОСТАВЛЯЮЩИХ ВРЕМЯ ОТКЛИКА WEB-СЛУЖБ

А.В. Горбенко

В статье представлены результаты поиска законов распределения времени отклика и других временных характеристик Web-служб. Теоретические исследования основаны на реальной статистике. Представлены результаты проверенных гипотез. Описаны подходы к моделированию симуляции времени отклика. Обосновано экспериментальное исследование и проведен математический анализ времени отклика. Установлено, что задержки, возникающие в Сервис-Ориентированных Системах имеют высокую степень неопределенности статистических характеристик, что существенно затрудняет их описание с помощью теоретических законов распределения случайной величины.

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

Горбенко Анатолий Викторович – канд. техн. наук, доц., доц. кафедры, Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина, e-mail: A.Gorbenko@csac.khai.edu.