Hybrid modeling applications for distributed information systems scaling tasks

Andrey A. Zenzinov

Junior research worker, Institute of mechanics, Moscow State University, +7(906)784-90-62,
andrey.zenzinov@gmail.com

Oleg A. Abankin

Student, Department of Mechanics and Mathematics, Moscow State University

1 Introduction

The development and distribution of information systems generally is accompanied by the increase in their architecture complexity and by the appearing of new requirements for software and resources. Modeling is an essential part of research on possible development directions of such systems.

Developers of productive information analysis systems (IAS) have to scale them sooner or later. Such need generally is caused by the increasing system complexity and the increasing users quantity which are geographically distributed. This leads to the research on scaling possibilities assessment and discovering optimal solutions in terms of ref source costs.

The Intelligent System for Thematical Analysis of Scientometric Data (russian abbreviation – ISTINA, for short in this paper – the System)[1] is a functional CRIS-system in Lomonosov Moscow State University (MSU). Within a pilot project, the System is being implemented in several scientific research institutes subordinated to the Federal Agency for Scientific Organizations (FASO) and distributed throughout Russia. In this regard there is a need of applying infrastructure development decisions aimed to satisfy new availability requirements caused by increasing number of geographically distributed users. The large distance between users and the System leads to increased data transmission delays in comparison with System usage in MSU.

In this paper, taking into account System usage in new organizations, we consider three following options:

- “cloud” mode, web-servers and database (DB) servers are located in a single data center (DC);
- web-servers are located in geographically distributed organizations while DB-servers are located in a single DC;
- web-servers and DB-servers are located in several geographically distributed DCs.

Testing and verification of developing architectural and software solutions on the productive System cause a number of difficulties containing the following:

- interaction between the System and the other functional information systems in organizations should be taken into account;
- productive System modification is difficult and accompanied by significant costs;
- distributed IAS consists of several components and it is necessary to consider their interaction and failure possibilities.

One of the common approaches described in similar studies is modeling of target systems. Such approach can be used to reduce costs and to assess the project solutions feasibility. The usage of hybrid modeling way as a conjunction of analytical, full-scale, virtual modeling and simulation, allows investigating the System behavior in various configurations, taking into account software and network infrastructure features.

2 Problem statement

The main purpose of this paper is to demonstrate the hybrid modeling applicability to complex distributed IAS simulation for solving scalability costs assessment tasks. By the scalability in this study we assume the possibility of architectural modifications conduction in order to provide desirable quality of service (QoS) in the context of involvement of new organizations which are using the System. These organizations can be located at a great distance from each other and their data can be logically distributed (e.g. different websites for different organizations). On the one hand the increase in users number leads to the increasing number of requests to the System. On the other hand the big distance between users and servers will cause a communication quality degradation.

Client organizations may establish the requirement of the System interaction with other information systems operating in them, including systems that process personal data. This may lead to the addition of new requirements for communication channels or requirement to place a part of the System servers in the same separate DC that hosts other servers for this organization. While the data related to scientific and pedagogical activity of different organizations personnel can be closely related to each other, there is a problem of DB-servers separation between organizations and choosing the data replication strategy. Latter can make a significant impact on development decisions but it is not defined for IAS “ISTINA” so far and this issue is outside the scope of this study.

A target problem of the System scalability assessment is investigation of dependence of objective (such as web page loading time) and subjective (QoS perception) user scores of the System web pages performance from network infrastructure parameters: network latency and packet loss rate.

Mean Opinion Score (MOS) [2] is the subjective user score, which values are logarithmically dependent on the network parameters [3]. This parameter will be thoroughly described in “QoS parameters” section.

The solution to the described problem may help to draw out some recommendations for the System architectural scaling.

Hybrid modeling can be applied in other development areas of the System such as following: decision-making on the project infrastructure development; systematic performance testing based on realistic scenarios with realistic infrastructure; continuous integration (automatic testing). Consideration of these approaches is out of this paper scope.

3 Description of the distributed IAS modeling system

In the context of this paper distributed information system model consists of three general components: nodes; network infrastructure; software installed on the nodes.

There are several options for nodes modeling: a physical machine (real hardware and software); a virtual machine (real software, emulated hardware); a logical node in simulation (each node have its own representation in network without hard- and software modeling); aggregate nodes behavior modeling (nodes are logically indistinguishable). Like the nodes networks can be represented in different versions: physical network (real network hardware); virtual networks (NAT, network bridge, isolated network); simulated networks (simulated via NS-3 network simulator [4]). The entire network infrastructure can be composed of parts modeled in different ways.

The hybrid modeling approach allows a combination of these variants in different proportions depending on the modeling purposes and providing the necessary detail level.

User behavior simulation is carried out using the Apache JMeter application to load test functional behavior and measure web applications performance [5]. JMeter allows us to simulate scenarios consisting of different realistic HTTP-requests to the System from larger number of users.

The distributed system model deployment is carried out with special virtual deployment facilities developed at MSU. This software provides an ability to deploy a distributed system model in automatic mode with given nodes and networks specification. Such approach is convenient for dealing with frequent configuration changes. More details are described in [6] and [7], although since the publication of these articles the functionality of the developed tools has been significantly expanded.

The System model from an architecture perspective is represented by one or several web-servers and a DB-server. The System is driven by Django framework and due to features of its ORM (Object-Relational Mapping) technology implementation features, characterized by a large number of queries to DB, the main traffic passes between the web servers and the DB. Thus, in the current study network characteristics variation is modeled in this internal part of IAS network infrastructure.

4 QoS parameters

The Quality of Experience (QoE) parameter is commonly used to describe the parameters of the user's perception of the service. MOS – an integer value from 1 (user stops using the service) to 5 (full satisfaction with the service quality) – is a convenient metric to describe QoE. QoS parameter is also used for this purpose and desig-

nates measurable service parameters such as packet transmission delay and service response time. The MOS dependence on QoS is described in many studies [8-11]. MOS is a subjective parameter because it depends not only on QoS parameters but also on unmeasurable parameters too: user expectations; service type; usage purposes.

One of the most common scenario of the web service usage is web-surfing and search usage. Web-page loading time is the key characteristic for the quality assessment for this scenario and it can be classified as QoS parameter. User expectations formed before starting the use of service have a significant impact on user perceptions of the web-page performance. If the user expects 1 minute response but in fact response comes faster then the user takes this fact positively.

Jacob Nielsen, one of the most known user experience specialist, emphasizes three main response-time limits in his studies of websites user experience [12]: instantaneous response (0.1 s., feeling of direct manipulation, user doesn't notice waiting time); seamless load (1 s., user can sense a delay but feels comfortable, from 1 to 10 s. user wishes the service was faster); lost attention (10 s., user start thinking about other things). Such categorization assumes the highest user expectations on service performance. If he expects the time-consuming response his perception score will be different.

The attention loss parameter is determined by the type of using service as well as by the tasks to solve. For example, for information searching tasks response time is the important criterion while for analytic tasks such as report preparation it isn't. Web-page loading time cannot be a universal criterion for the QoE evaluation, since the web-page content in most tasks is more important for user perception but for report creation the essential part is the final result.

The dependence of user perception of service performance (in terms of MOS) on web-page loading time is depicted in Fig.1 taken from [11].

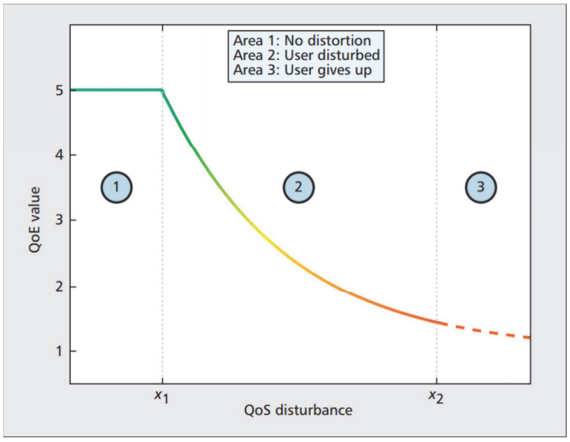

Fig. 1. Dependence of MOS on the page loading time

To determine the QoS numeric dependence on QoE the approach provided by recommendations [3] is used. QoS parameter is described by session time, and QoE – by

MOS. This document considers a scenario for searching information on the Internet, that consists of the following steps: a search page request; search query; search results displaying. For this scenario the following logarithmic dependence of MOS on the session time is valid:

$$MOS = \frac{4}{ln(Min/Max)} \cdot \big(ln(SessionTime) - ln(Min)\big) + 5,$$

where Min and Max – minimum and maximum session time obtained during experiments, respectively.

Correlation of this statement with the subjective users score depends on the users expectations. According to [3] for 60 s. waiting time an achieved correlation was 0.95 and for less than 6 s. waiting time – 0.72.

Thus, the obtained formula evaluates QoE on QoS dependence with a high precision for 60 s. waiting time and with a medium precision for less than 60 s. waiting time.

5 Experiments

In this section a scenario of information search and browsing the System web-pages is considered and it consists of the following steps: signing in; searching worker in the System by his name; browsing personal web-page of worker found; browsing a journal web-page; signing out. According to statistics from the Google Analytics these web-pages are the most visitable and due to the nature of this scenario we can use the MOS evaluation approach described in the previous section.

The System model is implemented using the described modeling system. The web-server and DB-server nodes are represented by virtual machines, network infrastructure is simulated by NS-3 and Netem (Traffic Control) [13]. In the first case the System network infrastructure is completely modeled by NS-3, in the last one – virtual networks are used and network characteristics are set by Netem.

The session time for this experiment will be considered equal to the total execution time of all scenario steps. Users expectations were calculated as a session time with no latency and no packet loss. The expectation of worker web-page loading time is about 35 s. with the given parameters.

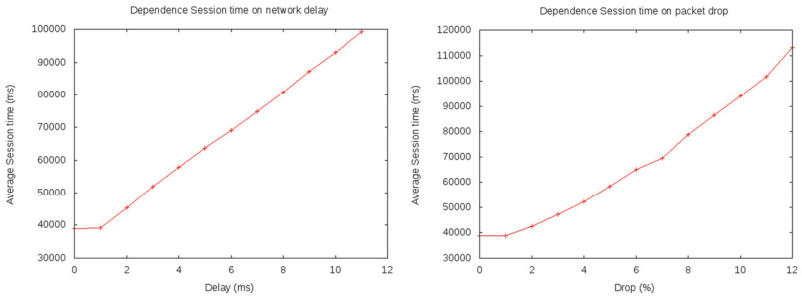

Fig. 2. Dependence of session time on the delay time and packet loss, NS-3 networks.

The measurements were performed for variation of latency from 0 to 12 ms. and for the packet loss variation from 0 to 10%. Experiment results showed that the session time is linearly depended on latency for each of the network simulation approach. In the case of NS-3 the simulated networks session time varied from 35 s. (0 ms. delay, 0% packet loss) to 100 s. (12 ms., 0%). These results are illustrated in Fig 2. In the case of using Netem – from 40 s. (0 ms., 0%) to 120 s. (12 ms., 0%). The linear nature of this dependence can be explained by the fact that during traffic processing a consequently processing of operations is performed. Packet loss variation doesn't affects the session time linearly. With given 0 ms. delay session time varied from 35 s. to 95 s.(10% packet loss) in the case of NS-3 driven networks (see Fig. 2) and from 40 s. to 240 s for Netem tuned networks. This non-linear nature can be explained by the features of TCP-protocol realization.

MOS values were calculated with the provided formula and for the delay variation the results are following (see Fig. 3): 5 (0 ms. delay, 0% packet loss), 4.4 (2 ms., 0%); 3.4 (4 ms., 0%), 2.5 (6 ms., 0%), 1.9 (8 ms., 0%), 1 (11 ms., 0%). MOS values for packet loss variation: 5 (0 ms., 0%), 4.4 (0 ms., 1%), 3.6 (0 ms., 2%), 3.1 (0 ms., 3%), 2.6 (0 ms., 4%), 2.1 (0 ms., 5%), 1.7 (0 ms., 6%), 1.4 (0 ms., 7%), 1 (0 ms., 8%).

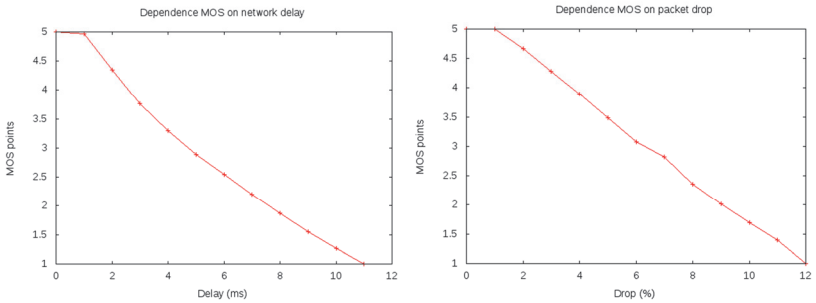

Fig. 3. Dependence of MOS on the delay time and packet loss, NS-3 networks.

Taking into account the obtained results we can formulate recommendations for the required network characteristics of the network section between the web-server and DB-server. Given that the satisfactory MOS value is 3, it is recommended to provide the connection quality with the network delay below 4 ms. and the packet loss rate below 3%. The System web-servers thus should be placed in the same DC with DB-servers if possible. Placing a web-server at a significant distance from the DB-servers will lead to severe delays.

6 Conclusions

This paper describes the distributed IAS hybrid modeling approach and demonstrates its possible applications for various scalability assessment tasks and for other tasks arising during the complex systems development.

Current work also provides an example of proposing recommended requirements for the network characteristics of network infrastructure of the System in the context of its distribution for new client organizations. The dependence of session time on network parameters such as latency and packet loss rate was obtained and analyzed.

The results provided in this paper can be useful for IAS "ISTINA" development processes. In particular, the obtained nature of the session time dependence on the network delay indicates that significant distances between the web-servers and DB-servers (e.g. if they are placed in different remote DCs) will cause a significant user characteristics decrease.

References

1. The Intelligent System for Thematical Analysis of Scientometric Data (ISTINA) / V.A. Sadovnichiy, S.A. Afonin, A.V. Bakhtin et al. – Moscow University Publishing, 2014. – P. 262. (in Russian)
2. Mean opinion score. // URL https://en.wikipedia.org/wiki/Mean_opinion_score (request date 30.06.2017)
3. ITU-T G.1030 : Estimating end-to-end performance in IP networks for data applications // URL https://www.itu.int/rec/T-REC-G.1030/en (request date 30.06.2017)
4. What is NS-3? // URL: https://www.nsnam.org/overview/what-is-ns-3/ (request date: 30.06.2017)
5. Apache JMeter // URL http://jmeter.apache.org/ (request date 30.06.2017)
6. Vasenin V.A., Roganov V.A., Zenzinov A.A. An environment for the study of the information sercurity facilities in grid and cloud systems. // Program engineering. – 2014. – No 3. – P. 21–33 (in Russian)
7. Zenzinov A. Automated deployment of virtualization-based research models of distributed computer systems // Proceedings of the 7th Spring/Summer Young Researchers' Colloquium on Software Engineering (SYRCoSE 2013). – National Research Technical University Kazan, Russia: Kazan, 2013. – P. 128–132.
8. Daisuke Yamauchi and Yoshihiro Ito A METHOD OF EVALUATING EFFECT OF QOS DEGRADATION ON MULTIDIMENSIONAL QOE OF WEB SERVICE WITH ISO – BASED USABILITY // International Journal of Computer Networks & Communications(IJCNC). – Vol.7, No.1, January 2015
9. ITU-T Nazrul Islam Vijaya John David Elepe On Factors Affecting Web Browsing-QoE Over Time // Master Thesis Electrical Engineering Thesis No: MEE100038. – January 2014
10. S. Canale, F. Delli Priscoli, S. Monaco, L. Palagi, V. Suraci A reinforcement learning approach for QoS/QoE model identification // Control Conference (CCC), 2015 34th Chinese. – July 2015
11. Markus Fiedler, Phuoc Tran-Gia A Generic Quantitative Relationship between Quality of Experience and Quality of Service // IEEE Network (Volume: 24, Issue: 2, March-April 2010). – March 2010
12. Jakob Nielsen. Website Response Times. 2010 // URL: https://www.nngroup.com/ articles/website-response-times/ (request date: 30.06.2017)
13. networking:netem // Linux Foundation Wiki // URL: https://wiki.linuxfoundation.org/ networking/netem (request date: 30.06.2017)