MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
SUMY STATE UNIVERSITY
UKRAINIAN FEDERATION OF INFORMATICS

PROCEEDINGS

OF THE V INTERNATIONAL SCIENTIFIC CONFERENCE

ADVANCED INFORMATION SYSTEMS AND TECHNOLOGIES

AIST-2017

(Sumy, May 17–19, 2017)

SUMY
SUMY STATE UNIVERSITY
2017

Computing Resources Scaling Survey

Y. Kulakov, R. Rader
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", roman.rader@gmail.com

***Abstract* – The results of the survey about usage of scalable environment, peak workloads management and automatic scaling configuration among IT companies are presented and discussed in this paper. The hypothesis that most companies use automatic scaling based on static thresholds is checked. The insight into the most popular setups of manual and automatic scalable systems on the market is given.**

***Keywords* – cloud computing, auto scaling, computer cluster, survey.**

I. INTRODUCTION

In the modern world, the accelerating business demands on websites make the reliability and efficiency of managing resources crucial. It implies optimization of all the processes in mature companies. Business keeps increasing the dependence on the web applications; workloads and numbers of servers in the companies are growing. In these conditions, any optimization can influence the general system efficiency. In this paper, the optimization of the computing resources is considered. Obviously, the optimization should reduce costs without significant quality loss. In the conditions of unstable and unpredictable peak workloads, cloud computing can significantly reduce costs, making possible to scale up and release unused resources rapidly [1].

By the definition given by NIST, cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [2]. By this definition, one of the key features of cloud is the scalability, which is the ability to expand and add resources dynamically, to be scaled up to meet demand through replication and distribution of requests across a pool or farm of servers [3].

From the application perspective, although cloud allows to flexibly managing resources, the control and monitoring units should be present in the system.

Auto scaling or dynamic scaling implies the automatic control of the action of provisioning or de-provisioning web server virtual machine instances, a dynamic scaling algorithm based on relevant threshold or scaling indicator of the web application is developed [3].

Scaling indicators mentioned above and the approach to the control unit can vary. While the specific method usage depends on the application being developed, set of most popular approaches is described in [4] and covers most of systems (see section III).

In order to understand the current situation on the production systems serving websites and cloud services, this study has been conducted. Main questions to be answered in this study are: (1) how many companies are using scalable environment, (2) are they using its potential (automatic scaling) if they do, (3) how usually companies manage with unexpected peak loads, (4) how do they configure the auto scaling algorithm if applicable. The hypothesis to be check is: (1) most companies deploying in the scalable environment are using automatic scaling; (2) the resource control system is based on static thresholds; (3) thresholds are set empirically without a research.

II. METHODOLOGY

The survey was conducted online using Google Forms [5] service and distributed among software developers and system administrators community in Ukraine. Respondents were invited to answer the questions in chat rooms dedicated to DevOps and system administration topic.

Respondents were asked about the configuration of their clusters; essentially three main issues were to be uncovered: (1) ownership of computing resources; (2) resources behavior in edge cases; (3) the way of configuration of this behavior.

Questions were ordered from broad to narrow in order to find all the details of as many as possible setups.

A. Ownership of computing resources

Basically, to have an ability to optimize the cluster workload during off-peaks, a scalable environment is required. It can be any kind of cloud: the most affordable for any company is public cloud providers help companies to share resources with other companies. Most of the popular public cloud providers bill only the resources you use, so it's possible to save money from idle resources in off-peak hours by disposing of resources back to the pool.

Hence, to understand the share of companies using scalable environment it's important to know the way companies own their resources. In the survey respondents were asked whether they rent the computing resources or own them. The "rent" was divided into two options: (1) rent the physical machines and (2) rent the virtual resources from cloud providers.

B. Resources behavior in edge cases

Under the "edge case" the peak load condition in the system is meant here. Computer systems in the production environment should be configured to handle

different kinds of loads: peak and off-peak hours, increasing and decreasing of load during holidays, even stand the DoS attacks. To understand how companies manage with expected and unexpected peak workloads, they were asked to provide the information about their systems configuration, whether they provision resources manually or automatically, what approach is used.

C. The way of configuration of this behavior

As mentioned above, the approaches of managing the resources can vary. Manual control of the resources implies the presence of the monitoring system, which is used to make a decision about provisioning additional resources or releasing them by a human.

If the automatic resources control approach is used, the control system should be configured to scale the specific application on the specific environment.

Respondents were asked in what way their control unit was configured, whether some research was done to get the parameters for the system, parameters were set empirically by the system administrator, or their control unit doesn't require the configuration.

Also, respondents who use manual resource control approach were asked if they have any plans for a transition to the automatic system.

D. Confidence interval calculation

To be able to compare the values and show the statistically significant percentage we will calculate the confidence interval. Given sample size n and sample proportion $\hat{p}$, which is the percent of respondents in our case, the confidence interval for p - the population proportion of the survey responses for a single option of the question.

Question options	x_1	x_2
Number of answers	n_1	n_2

where $n = n_1 + n_2$.

The dispersion of the sample is given by [6]:

$$D_s = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \quad (1)$$

Substituting n=2 and the values, the dispersion of the sample would be:

$$D_s = \frac{n \cdot n_i - n_1^2}{n} \quad (2)$$

The sample proportion $\hat{p}$ is:

$$\hat{p} = \frac{n_1}{n} \quad (3)$$

Substituting (3) into the (2):

$$D_s = \hat{p} \cdot (1 - \hat{p}) \quad (4)$$

The correction of the sample dispersion to calculate the population dispersion is not required since the population is greater than 30, $D_s \approx D_p$ [6] (see the III.A section).

Hence, the confidence interval is given by:

$$\hat{p} \pm t \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}} \quad (5)$$

For the confidence level of 95%, the value of $t = 1.06$.

III. Results

A. General numbers

Based on the data from State Statistics Service of Ukraine on 2015, there are 13617 companies registered in Ukraine by the "information and telecommunications technology" activity [7].

44 companies working in Ukraine responded to the survey. 41% of them are Ukrainian companies and 59% of them are official representatives of foreign companies in Ukraine.

B. Resources ownership

Among responded companies, 61.4% rent computing resources and 38% use their own hardware. Moreover, 96.3% of companies who rent resources, are renting them from cloud providers and only 3.7% are renting dedicated servers (see fig. 1).

TABLE I. OWNERSHIP OF RESOURCES

	Value and 95% confidence interval		
Ownership	***Own hardware***	***Rent computing resources***	***Total***
Responses in the survey	17 answers	27 answers	44 answers
Among IT companies	38.6%±14%	61.4%±14%	100%
Type of rented resources (among renting group)	***Cloud providers***	***Dedicated servers***	***Total***
Responses in the survey	26 answers	1 answer	27 answers
Among IT companies	96%±7%	4%±7%	100%

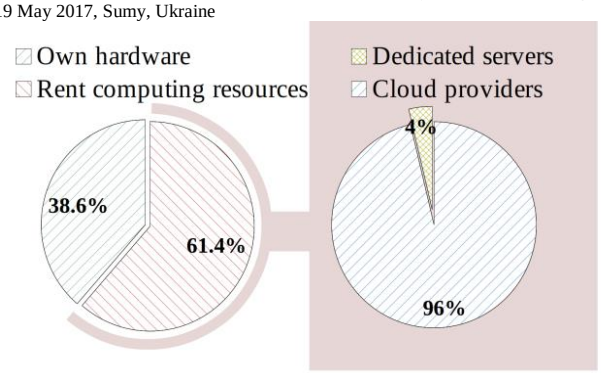

Figure 4. Ownership of resources

C. Behavior on edge cases

In this section, the results of the survey about the computer systems configuration are considered.

TABLE II. RESOURCES MANAGEMENT ON EDGE CASES

	Value and 95% confidence interval		
	Automatic	***Manual***	***Total***
Responses in the survey	18 answers	26 answers	44 answers
Among IT companies	40.9%±14%	59.1%±14%	100%

To get the insight in these two groups the approaches of resource management in each group were considered (see table III and table IV).

Respondents were asked how they act in conditions when their system experience higher workload than expected. There were two majorities of answers (see table II), (1) any kind of automatic resource management and (2) manual. Also, two respondents answered that their workload is highly predictable. For this survey purposes, these answers were assigned to the group of manual resource management.

TABLE III. TYPE OF MANUAL RESOURCES MANAGEMENT

	Value and 95% confidence interval			
	Manual management	***Queues and metrics thresholds as indicators***	***Regression model for workload prediction***	***Total***
Responses in the survey	17 answers	7 answers	2 answer	26 answers
Among IT companies	65.4%±18%	26.9%±17%	7.7%±10%	100%

TABLE I. TYPE OF AUTOMATIC RESOURCES MANAGEMENT

	Value and 95% confidence interval			
	Static thresholds	***Serverless architecture***	***Queueing theory***	***Total***
Responses in the survey	14 answers	3 answers	1 answer	18 answers
Among IT companies	77.8%±19%	16.7%±17%	5.6%±11%	100%

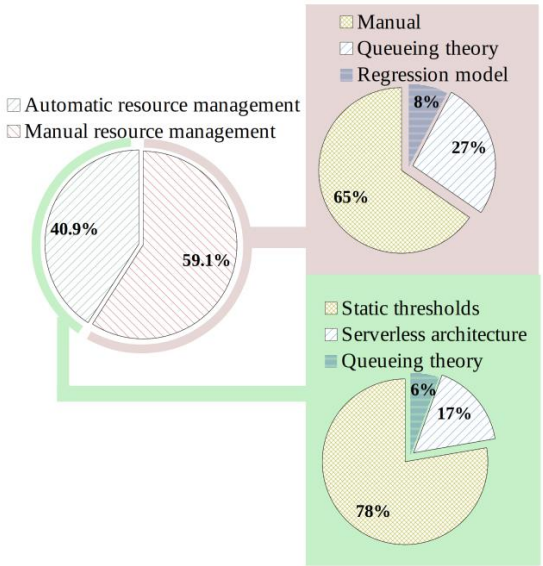

Figure 5. Resources management

D. Automatic resource management configuration

The automatic resource control unit may require a configuration to work in specific environment with concrete application. The ways of setting configuration parameters are considered in table V.

TABLE II. CONFIGURATION APPROACH

	Value and 95% confidence interval				
	Research has been done	***Parameters set empirically***	***Load testing***	***Nonparametric algorithm***	***Total***
Responses in the survey	8 answers	5 answers	1 answer	4 answers	18 answers
Among IT companies	44.4%±22%	27.8%±20%	5.6%±10%	2.2%±7%	100%

E. Trends of future

Respondents were asked whether they are in process of migration to auto-scaling solutions now or not. Among those who are not using auto scaling currently, 39% (±14% among all IT companies) of respondents answered they are in process of development or adoption the auto-scaling control unit for their systems.

IV. DISCUSSION

Main questions this survey was intended to answer are (1) how many companies are using scalable environment, (2) are they using its potential (automatic scaling) if they do, (3) how usually companies manage with unexpected peak loads, (4) how do they configure the auto scaling algorithm if applicable. Also, the hypothesis was put forward to check that (1) most companies deploying in the scalable environment are using automatic scaling; (2) the resource control system is based on static thresholds; (3) thresholds are set empirically without a research.

A. How many companies are using scalable environment?

Considering the results that 61.4% are renting resources and 96% of them are renting them from cloud providers, the answer is 59% of companies have a scalable environment (see table I). This result is consistent with study by RightScale in 2017 [8], considering the confidence interval and focus of this study on the Ukrainian companies.

B. Are companies using the potential of auto-scaling and how they manage with peak loads?

Among respondents, 40.9% are using auto-scaling (see table II). Considering that 59% have a scalable environment (see IV.A), we can state that majority of companies (70%) are using the potential of the auto-scaling environment, which confirms the (1) of the hypothesis. Moreover, most of them (77.8%) are using the method of static thresholds for their resource control systems (see table III), which confirms (2) of the hypothesis.

The amount of respondents is the limitation to get statistically significant results of using less popular auto-scaling techniques than usage of the static thresholds approach.

C. How usually companies manage with unexpected peak loads?

The study showed that less than half of companies set the parameters of their auto-scaling system empirically (see table V). The majority (44.4%) made some kind of research to get them. This result rejects the (3) of the hypothesis. Also, since customers are ready to make a research to configure the auto-scaling system, for researchers that mean that customers need the balanced solution that both easy to setup and has flexible, easily understood configuration parameters to make the auto-scaling system adaptable for a specific application.

CONCLUSIONS

The purpose of the study is described, main questions to answer by the study are put and the hypothesis is articulated. The methodology of conducting the survey is given. The results with calculated margins of errors of the conducted survey are presented in this paper. Results summarize and interpretation of them are described, the limitations of given results are described. The hypothesis is confirmed by two items and rejected by one, interpretation is provided.

REFERENCES:

[1] T. Dillon, C. Wu, and E. Chang, "Cloud computing: issues and challenges." 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), April 2010.

[2] P. Mell, and T. Grance, "The NIST definition of cloud computing." National Institute of Standards and Technology, Special Publication 800-145, 2009.

[3] T. C. Chieu, A. Mohindra, A. A. Karve, and A. Segal, "Dynamic scaling of web applications in a virtualized cloud computing environment", 2009 IEEE International Conference on e-Business Engineering, pp. 281-286, October 2009.

[4] T. Lorido-Botrán, J. Miguel-Alonso, and J. A. Lozano, "Auto-scaling techniques for elastic applications in cloud environments." Department of Computer Architecture and Technology, University of Basque Country, Tech. Rep. EHU-KAT-IK-09-12, 2012.

[5] R. Rader "Google Forms: Computing Cluster Scaling Survey 2017", docs.google.com/forms/d/e/1FAIpQLSfYzDhmRU3_NebtCep77DlZmVjpDFlCkUB2k1uHhjYIcAuhBg/viewform, accessed April 15, 2017.

[6] V. E. Gmurman "Probability theory and mathematical statistics." Moscow: Vysshaya Shkola, 1972 (in Russian).

[7] State Statistics Service of Ukraine "Number of business entities by economic activities in 2015" ukrstat.gov.ua/operativ/operativ2014/fin/osp/ksg/ksg_u/ksg_u_15.htm, 2015, accessed April 15, 2017 (in Ukrainian).

[8] K. Weins, "Cloud Computing Trends: 2017 State of the Cloud Survey", www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2017-state-cloud-survey, February 2017, accessed April 17, 2017.