This article addresses the challenges of a digital cloud environment when it comes to global scalability with a large number of remote servers and an unsecure public cloud, such as Amazon. The goal of the study was to work out an approach for evaluating the reasonable system capacity under heavy workload. For that purpose, Zabbix monitoring solution is used and business metrics are applied in relation to existing system ones. A prediction data model is proposed to compute the future forecast of the user activity based on the collected historical statistics and to verify whether capacity adjustment is possible or not. The results of capacity planning are implemented at Genesys International Telecommunications Company. System analysis of the production environment indicates the possibility to downscale the capacity of certain virtual servers, which allowed savings to the annual operational costs of $3500 (50%) for each affected server.

1. Introduction

Fast growing internet information technologies (IT) based on globally distributed cloud computing (GCC) have allowed radical improvements in the efficiency of internet services as well as capacity and in the performance of virtual machines (VMs). Big international IT companies deploy thousands of VMs with production software in the cloud to provision multiple services for their customers worldwide.

 GCC implementation and operational support has both benefits and challenges, as the following:

- A cloud service is now an integration of multiple platforms and applications of the internet of things (IoT). 
- The well-known problems of big data, including storing client accounts in a database, protocoling user activity in the logs, transferring real time dataflow, and processing big data analytics.
- A fast growing and, therefore, unpredictable user workload.
- The implementation of mobile applications leads to human activity spikes, not only in peak daytime but also in the late evenings and early mornings.
- Continuous multiservice delivery and operational support in 24/7 mode.
- Initially developed in the USA, many International IT companies go global for provisioning GCC services in all regions and time zones worldwide.
- The virtualization of the digital environment and migration from private to public cloud services, such as software as a service (SaaS), platform as a service (PaaS), infrastructure as a service (IaaS), load balancer as a Service (LBaaS), database as a service (DBaaS), etc.
- Increasing cyber security problems and new sources of threats when moving from private to public clouds.
- State government standards, banking rules, and other local regional limitations.
- Existing monitoring solutions provide only built-in system metrics like memory, CPU, disk space, or network throughput, but not specific business items like HTTP requests per second, the number of actively logged-in users, server connections, calls count, and others.

These challenges are described in detail and proper solutions are defined in the sections below.

2. Related Works

A deep analytical overview of the advanced telecommunication technologies and the top 10 cyber threats in the networks, including the attack mechanisms, vulnerability and damage assessment, the object of the attacks, counter measures, modeling and prediction of cyber resistance, and the possibility of how to organize and coordinate the preventive actions, are provided in.

The article is an overview of the relevant research on the key threats in cloud cyber security, such as data breaches or loss, account or service traffic hijacking, insecure interfaces and applications of IoT, distributed denial of service (DDoS), malicious insiders, the abuse of cloud services, insufficient due diligence, and other shared technology vulnerabilities. In addition, the authors of the article show the dependency of cyber threats with the security controls and compliance models based on the corresponding standards. The performed theoretical research is implemented as a practical system to identify threats and protecting actions.

The article is the analytical study of various factors, which influence the cyber security on particular cloud platforms. The following ways of possible violation are pointed out:

- Confidentiality (C) of sensitive information
- Integrity (I) of the database (DB) information
- Availability (A) of cloud service Data (D)
Availability of a Virtual (V) server as a whole

For example, unauthorized access to a host would result in not only V violation but also in data manipulation (D violation) and discrediting of sensitive data as well (C violation). In another example, SQL injection to a DB, especially during data replication or backup (D violation), may cause an unacceptable data modification on the DB server side (I violation). In case of a very dangerous DDoS attack against a frontend server (both V and D violation), a cloud computing service will be entirely unavailable for internet users (A violation).

In terms of the monitoring of cloud computing resources, an analytical overview of the top five worldwide popular solutions is presented in [7]. International scientific conferences on facing the widespread problem of big data computing are addressed in [2].

Paper [13] proposes a recognition system in the IoT area to monitor real-time data, particularly on smartphones. Data fusion from multiple mobile sensors, including audio, Internet localization, etc., allows analysis of a person’s life patterns. This is helpful for taking care of either children or older adults.

Using predictive models in GCC monitoring systems is considered in [14][15], including the well-known problem of Java memory leak and auto-remediation of Java-based services running on a virtualized cloud environment based on the 4R options—restart, reboot, redirect, redeploy.

The novelty of this study is in the evaluation of VM capacity and cyber security when moving from a private corporative cloud (on premise) to a public one. The public cloud is much cheaper but less secure and, therefore, additional actions should be done to meet the cyber security policy.

3. Basic GCC Architecture Models and Cyber Security Issues

Big international IT companies have similar globally distributed infrastructure and provide cloud services in multiple geographic regions worldwide [16][17] (Figure 1). A region is usually named by a three letter abbreviation of the nearest international airport, for example SJC—San Jose California, USA; YYZ—Toronto’s main airport, Canada; AMS—Amsterdam, the Netherlands; SIN—Singapore; SYD—Sydney, Australia, etc.

![Figure 1. Typical globally distributed cloud computing (GCC) infrastructure and data flows.](image)

Every region is built of at least two data centers (DCs)—a primary unit and a standby one (Figure 2). Such redundancy is designed for the purpose of workload switchover in case of a local incident or an entire DC outage. Each unit supports a certain part of data (POD), including client account database (ADB) and a set of cloud service applications interconnected with the shared message system storage (MSS) for storing call prompts, DB queries, application logs, and other system events. Regular replication of ADB data between the DCs is configured to perform a switchover faster when initiated. For example, RingCentral International Telecommunications Company uses Oracle as a DB and Golden Gate as a third-party solution for optimized ADB synchronization based on triggers and data transaction logs [18].
To build a cloud IT environment, the three basic architecture models are practically used—private (on premises), public, and hybrid (Figure 3). On premises, the entire IT infrastructure is allocated in corporate DCs, giving full control over computing resources. Public cloud solutions allow IT companies to exclude the expenses for maintaining their own DCs. A hybrid cloud is an integrated solution with an attempt to inherit the advantages of both private and public models.

The key trend of GCC popularity among people and IT companies in the world is moving from private towards public clouds. Microsoft and Gartner analytics compiled in Table 1 shows that the growth of all public cloud services almost doubled over the last five years.[18][19][20]

Table 1. Gartner analytics of growth trend for main cloud services.

| World Market, $ billion | 2016   | 2017   | 2018   | 2019   | 2020   |
|-------------------------|--------|--------|--------|--------|--------|
| Software as a Service (SaaS) | 38.57 | 46.33  | 55.14  | 64.87  | 75.73  |
| Platform as a Service (PaaS) | 7.17  | 8.85   | 10.62  | 12.58  | 14.8   |
| Infrastructure as a Service (IaaS) | 25.29 | 34.6   | 45.56  | 57.9   | 71.55  |
| Load Balancer as a Service (LBaaS) | 90.26 | 104.52 | 118.52 | 133.57 | 151.1  |
The advantages of the public cloud are obvious:

- Significant decrease of IT expenses for maintaining DCs on premises
- Minimize the IT department to support end users
- Low level of IT personnel skills required
- Flexible scalability of cloud computing resources and capacity planning
- Support IT services related to backups, disaster recovery, etc.

In spite of the big benefits, the public cloud technologies have certain challenges, and cyber security is one of them. Personal sensitive information is stored on the external public network maintained by a vendor and is transferred via common internet channels. Tokenization of secret information, data encryption during transfer, and other approaches are good ways to reduce the risks. The other threats associated with the public cloud are shown in Table 2.

**Table 2.** The risks associated with the public cloud.

| Risk               | Risk Management                                      |
|--------------------|------------------------------------------------------|
| External fraud     | Review the company’s security policy and protect the data |
| Access issues      | Monitor health check to ensure the cloud service is stable |
| Internal fraud     | Ask external cloud provider to track and log the user's activities |
| Vendor failure     | Need alternative cloud vendor in case of a primary outage |
| Cloud outage       | Need backup network channel in case of a primary failure |

4. Public Cloud Computing Resources

This section describes the analysis of the specific GCC resources provided by Amazon Web Services (AWS) as an example, but the same approach can be applied to any other cloud provider, such as Microsoft Azure, IBM Cloud Services, Red Hat CloudForms, VMware Cloud Foundation, and other worldwide known leaders.

AWS is one of the leading public cloud vendors in the IT world, having a wide variety of VMs to implement all the main IT techniques as a service like SaaS, PaaS, IaaS, LBaaS, DBaaS, etc. VMs are available in AWS as a selection of various VM instance types and sizes, allowing flexible scaling of CPU, memory, storage, and network to the required workload of the applications (Table 3). Migration from one instance size to another is possible at any moment and is very easy—just select a proper type and restart the VM. If a misbalance of computing resources is needed, there are special VM types accelerated and optimized for computing, memory, or storage.

**Table 3.** Amazon Web Services (AWS) instances of the M4 type with optimized storage.
Table 4. Prices for AWS instances of the M4 type running operating system (OS) Linux/Unix in the US East region.

| VM Instance | Virtual CPUs | Memory, GB | Storage Performance | Storage Volume, GB | Price per Hour for OS Linux |
|-------------|--------------|------------|---------------------|--------------------|-----------------------------|
| m4.large    | 2            | 8          | Optimized           | 1 x 4 SSD          | $0.1                        |
| m4.xlarge   | 4            | 16         | Optimized           | 1 x 8 SSD          | $0.2                        |
| m4.2xlarge  | 8            | 32         | Optimized           | 1 x 16 SSD         | $0.4                        |
| m4.4xlarge  | 16           | 64         | Optimized           | 1 x 32 SSD         | $0.8                        |
| m4.10xlarge | 40           | 160        | Optimized           | 2 x 40 SSD         | $2.0                        |
| m4.16xlarge | 64           | 256        | Optimized           | 2 x 80 SSD         | $3.2                        |

Tables 3 and 4 show that in case of migration from any AWS instance to the nearest one, the computing resources of both CPU and memory are doubled and the price doubles as well.

5. Monitoring System and Prediction Models

Although VM types in the public cloud can be scaled up and down very quickly, it should be done during planned maintenance to avoid service outage. For the purpose of capacity planning and performance evaluation, the monitoring system is needed. In this paper, Zabbix enterprise-class system is considered as one of the leading monitoring solutions in the world, having many built-in possibilities including:

- Open-source code, allowing customization and implementation of new data models (prediction models in our case).
- Enterprise-class product, supporting the server pool of up to 15,000 remote hosts with a total rate of 500 monitoring values per second.
- Free software to install from scratch and free for annual upgrades without affecting the customized code.
- Built-in templates with system metrics like CPU, RAM, disk I/O, etc., allowing monitoring of new hosts in “one click”.
- Enhanced reporting web user interface with graphs and analysis tools.
Zabbix supports a wide scope of built-in system metrics and triggers to monitor CPU, memory, disk space, etc., but business specific items should be implemented manually. Figures 4 and 5 show examples of such business items, measuring the incoming and outgoing user traffic at RingCentral and Genesys Companies accordingly.

User workload is usually cyclic but cannot be predicted using polynomial forecasting models due to Runge’s problem of oscillation for polynomial models of higher degrees (Figure 6). Instead, a linear model is applied for capacity evaluation where only maximum values in peak time are taken into account. As a result, the prediction looks more stable and precise.

---

**Figure 4.** Cyclic user workload with a stable prediction trend.

**Figure 5.** Cyclic user workload with a downgrading prediction trend.

**Figure 6.** Runge’s phenomenon of oscillation for polynomial models of 3 and higher degrees – the abscissa shows the degree of the model and the ordinate shows the error of the results in percentage, giving the same accuracy results and extra calculations are not required; red, blue and green curves is a polynomial models of corresponding degree.
The same approach can be applied to any monitoring tool other than Zabbix. The purpose of the prediction data model is to analyze the workload trend and verify whether VM downscale is possible or not. If the workload is stable (Figure 4) or slowing down (Figure 5), the downscale is reasonable, otherwise it is risky and possible that upscaling will be needed soon.

5. Implementation and Experimental Results

The proposed monitoring and forecasting models are implemented at Genesys International Telecommunications Company[16]. Production VMs of the customers with relatively low user workload are analyzed using Zabbix monitoring and prediction system. Some experimental results are presented in Figures 7 and 8.

Figure 7. Monitoring statistics for user workload.
Statistics in Figure 6 shows that the users (agents) are not as active as configured (20 logged on agents as maximum out of 75 configured that is less than 30%). System statistics in Figure 8 confirms less than 20% of CPU utilization even in peak time. Memory consumption is about 5.55 GB and is stable. In Table 3, all the instances of M4 type have at least 8 GB memory, which would be enough for running cloud services. Therefore, the VM instance could be scaled down without a risk of system overload.

Similar calculations are verified for some more VMs and the results are put into pivot Table 5. CPU utilization is computed given multiple cores. Price savings are converted from “$ per hour” to “$ per year”, taking into account the same VM type for primary and backup servers and multiple regions for some customers, using the following formulas:

\[
\text{Price per year} = \text{Price per hour} \times 24 \times 365;
\]

\[
\text{Price saving per year} = \text{Price saving per hour} \times 24 \times 365 \times 2 \text{ backups} \times 2 \text{ regions}.
\]

| AWS Region/VM | # of Configured Agents | Max # of Logged on Agents | Max/avg CPU Usage per Core, % | Max Memory Usage, % | AWS Instance Type | Recommended AWS Type | Price Savings, $ per Year |
|---------------|------------------------|---------------------------|--------------------------------|---------------------|-------------------|---------------------|--------------------------|
| euw1/vmp-400  | 75                     | 20                        | 87/25                          | 40                  | m3.large          | m3.medium           | $3504                    |
The estimated price saving is $3504 per year for each affected VM. If evaluating the whole GCC infrastructure of a low loaded customer, the other VMs most probably could also be reduced in size and reach the total price saving up to 50%, that is about $13K per year (Figure 9).

Some instances could probably be downscaled twice depending on the actual monitoring statistics. Reducing the size of VMs is based on objective, historical statistics and safety. If the workload increased at some point, the triggers of the monitoring system will detect this and VM could be upsized back at any moment.

References

1. Mescheryakov, S.; Shchemelinin, D. Deployment and Monitoring of Internet Multiservice System in the Cloud; Monograph. Scientific and Practical Edition; LAP Lambert Academic Publishing: Saarbrücken, Germany, 2018; 101p, ISBN 978-613-9-96321-8. Available online: https://www.lap-publishing.com/catalog/details/store/fr/book/978-613-9-96321-8/deployment-and-monitoring-of-internet-multiservice-system-in-the-cloud?search=mescheryakov (accessed on 12 March 2020).
2. Rudenko, A.O.; Mescheryakov, S.V.; Shchemelinin, D.A. ASE International Conferences on Big Data Science and Computing. St. Petersb. State Polytech. Univ. J. 2015, 1, 110–119, doi:10.5862/JCSTCS.212.10.
3. Efimov, V.; Mescheryakov, S.; Shchemelinin, D. Adaptive Control of Cloud Computing Resources in the Internet Telecommunication Multiservice System. In Proceedings of the 6th International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT), St. Petersburg, Russia, 6–8 October 2014; pp. 387–393, ISBN 978-1-4799-5291-5, doi:10.1109/ICUMT.2014.7002117. Available online: https://ieeexplore.ieee.org/document/7002117?arnumber=7002117 (accessed on 12 March 2020).
4. Efimov, V.V.; Mescheryakov, S.V.; Shchemelinin, D.A.; Yakovlev, K.A. Integration and Continuous Service Delivery in Globally Distributed Computing System. Hum. Sci. Univ. J. 2017, 30, 13–20.
5. Efimov, V.; Mescheryakov, S.; Shchemelinin, D. Integration Data Model for Continuous Service Delivery in Cloud Computing System. In Communications in Computer and Information Science; Springer International Publishing: Basel, Switzerland, 2017; Volume 700, pp. 87–97, ISBN 978-3-319-66836-9, doi:10.1007/978-3-319-66836-9_8. Available online: https://www.springer.com/gp/book/9783319668352 (accessed on 12 March 2020).
6. Ardulov, Y.; Mescheryakov, S.; Shchemelinin, D. Dynamic Load Balancing and Continuous Service Delivery in a Big Cloud Infrastructure. In Proceedings of the 42nd International IT Conference by Computer Measurement Group (CMG), La Jolla, CA, USA, 7–10 November 2016. Available online: https://www.cmg.org/publications/conference-proceedings/conference-proceedings2016/ (accessed on 11 November 2016).
7. Mescheryakov, S.V.; Shchemelinin, D.A. Analytical Overview of Zabbix International Conference 2013. St. Petersb. State Polytech. Univ. J. 2014, 1, 91–98.
8. Buinevich, M.; Izrailov, K.; Vladyko, A. Metric of vulnerability at the base of the life cycle of software representations. In Proceedings of the 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea, 11–14 February 2018; ISBN 979-11-88428-01-4, doi:10.3931/ICACT.2018.8323940.
9. Buinevich, M.; Vladyko, A. Forecasting Issues of Wireless Communication Networks' Cyber Resilience for an Intelligent Transportation System: An Overview of Cyber Attacks. Information 2019, 10, 1–22.
10. Buinevich, M.; Izrailov, K.; Stoilyarova, E.; Vladyko, A. Combine method of forecasting VANET cybersecurity for application of high priority way. In
Keywords

internet of things; cyber security; public cloud; digital environment; virtual machine; digital transformation; monitoring system; prediction data model

Retrieved from https://encyclopedia.pub/1364