Scalability Issues in Cloud Computing and Solution Approaches

Daniel Bretschneider
ic19b035@technikum-wien.at
University of Applied Sciences Technikum Wien
Vienna, Austria

Ferhat Dövme
ic15b046@technikum-wien.at
University of Applied Sciences Technikum Wien
Vienna, Austria

Behnam Ezazi
xyz@technikum-wien.at
University of Applied Sciences Technikum Wien
Vienna, Austria

julius Kosa
xyz@technikum-wien.at
University of Applied Sciences Technikum Wien
Vienna, Austria

ABSTRACT

Scalability is the fundamental attribute of every network, system or infrastructure to increase or reduce its performance, resources and functionalities in order to meet the demands of a growing number of users and devices. High scalability results in an optimization of the overall system efficiency and cost-savings, while poor scalability eventuates in poor system performance necessitating the replication of system components, for example.

Cloud computing is a big shift from the traditional way businesses think about IT resources, bringing several benefits that encourage more and more organization to outsource their services and data into the cloud. Another big issue is the evolving sector of IoT with upcoming billions of devices inter-connected via the Internet and will only be possible due the paradigm of cloud computing.

When scaling a system or network, very different types of problems can occur.This paper further contains several approaches on solving scalability related issues in cloud systems in order to improve system performance.

KEYWORDS

Scalability, Cloud, Cloud Computing, Issues, Problem localization, horizontal scaling, vertical scaling, IoT

1 INTRODUCTION

This section contains an introduction on cloud computing in general and clarifies why scalability – among other concepts – plays such an important role. Also provide examples of scalable cloud-based systems (IaaS, PaaS, etc.)
Cloud computing is sharing software and hardware resources, location independent, via the internet. Examples of cloud-based systems are Server Scalability available by Infrastructure as a Service (IaaS), Scaling of the Network with the need to scale by consolidated data centers that host several VMs per physical machine (often achieved by overprovisioning resources), Scaling of the Platform by Platform as a Service (PaaS) offer ready to use execution environments and convenient services for applications.

1.1 Related Work

There are many papers related to scalability, cloud computing and IoT, but there are only a few papers that combine all three topics. Because of the growing devices connected to clouds and the emerging IoT sector the paper is an attempt to provide solutions to current problems.

2 SCALABILITY - CHANGE!

Scalability is the ability to provide sufficient performance despite increasing demands. Grow or shrink, scaling is a change of size and does not always mean increasing. Adjusting to changing requirements is very important. Also declare what does not concern to scalability (like replacing or something). Scalability of a system can be measured along at least three different dimensions [Neuman, 1994]. First, a system can be scalable with respect to its size, meaning that we can easily add more users and resources to the system. Second, a geographically scalable system is one in which the users and resources may lie far apart. Third, a system can be administratively scalable, meaning that it can still be easy to manage even if it spans many independent administrative organizations. Unfortunately, a system that is scalable in one or more of these dimensions often exhibits some loss of performance as the system scales up.[3]

2.1 Vertical Scalability

What happens when a system is being scaled vertically? The system is built up on different depending layers. User interface layer, application layer and database layer are the typical layers for the three-tier architecture. Each layer can be placed either on the client machine or the server machine (cloud). Depending on how the different layers are established we distinguish the different kind of possible vertical scalability.

2.2 Horizontal Scalability

Horizontal scalability means to allocate the vertical Scalability to different physical machines. Difficulties are the requirements of certain services, which are not provided by every physical machine and the requirements to latency.

2.3 Comparison: Which is the better one?

Discuss the both scaling methods based upon several different factors. Those factors are: Pros/cons, when should what method be used, how easy are they achieve, what problems cloud possibly occur etc.

Introduce the concept of diagonally scaling.

2.4 Diagonally Scalability

Briefly explain how vertical and horizontal scalability can be brought together to benefit from the advantages of both methods.

2.5 Diagonally Scalability

Briefly explain how vertical and horizontal scalability can be brought together to benefit from the advantages of both methods.

3 MODIFICATIONS

Explain different issues concerning scalability in cloud systems and define which items / components actually interfere with scalability. Where should the overall focus should be.

3.1 Lack of Standardization

Today there are many different standards...

3.2 Volume

Cloud scalability has to deal with various volumes of users, resources and data involved in service provision. Due the evolving IoT area, billions of devices will be inter-connected by the year 2020[2].

3.3 Lack of Ensuring autonomous scalability service management.[3]

Whatever.

4 IMPROVING SCALABILITY IN CLOUD SYSTEMS

Now that the reader knows that several scalability related problems exist in cloud environments, this section will inform him about concepts and methods of solving these problems.

4.1 Establish an international standard to support scalability in between clouds.

blah. bluh. blih.

4.2 IoT-Centric Cloud approach

Cloud Computing is not only sharing the resources but also maximizing the resources location independent. Virtualization of physical devices in cloud based IoT to share the devices and bring IoT functionalities into the cloud. Distribution over heterogeneous platforms, spanning multiple management domains. The ecosystem consists of local clouds and a global cloud for real time big data and analytics. A local Cloud can be created on-demand and provides service to users in that geographical area. It can involve a large number of nodes (sensors, actuators, smartphones, etc.) The global Cloud is the "backbone-infrastructure" and increases business opportunities for service providers. It increases and provides a more dynamic resource management and orchestration techniques, dynamically offloading from clients/hosts to cloud. It provides a reliable real-time communication from objects to applications and all of that executed across borders.[1]

4.3 Service Scalability assuring Process

The approach is to establish an autonomously managing process:

(1) Quality metrics for measuring services are defined (e.g.: throughput, - efficiency of handling service invocations within a given time)
(2) Acquire raw data items from monitored services
(3) Compute scalability metrics. In case the metrics reveal an acceptable scalability level, the control goes back to step 2 and repeat steps 2 and 3. If the metrics show a need to take actions steps 4,5 and 6 are performed.
(4) Devise a remedy plan for enhancing suffered scalability based on the current states of monitored service. (scalability assuring schemes)
(5) Run the selected scalability assuring schemes according to the plan. Many, if not all, of the schemes should be able to run without human administrators.
(6) Analyze the result of applying the remedy plan and learn from the whole process of enhancing scalability. Making the whole scalability framework more intelligent.

Many Scalability assuring schemes are hardware-based solutions. But there are also software-oriented schemes. Such as Service Replication and Service Migration.[3]

4.4 Benefits of Cloud Scalability

Regarding performance, cost-efficiency etc.

5 CONCLUSION

As the section title says this will be the conclusion. Here will summarize our "findings" and explain how cloud systems should be scaled in order to prevent different types of problems to even come up.

6 CITATIONS AND BIBLIOGRAPHIES

The use of BibTeX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the `\end{document}` command:

```
\bibliographystyle{ACM-Reference-Format}
\bibliography{bibfile}
```

where "`bibfile`" is the name, without the "`.bib`" suffix, of the BibTeX file.

Citations and references are numbered by default. A small number of ACM publications have citations and references formatted in the "author year" style; for these exceptions, please include this command in the **preamble** (before "`\begin{document}`") of your LaTeX source:

```
\citestyle{acmauthoryear}
```

Some examples. A paginated journal article [**?**], an enumerated journal article [**?**], a reference to an entire issue [**?**], a monograph (whole book) [**?**], a monograph/whole book in a series (see 2a in spec. document) [**?**], a divisible-book such as an anthology or compilation [**?**] followed by the same example, however we only

output the series if the volume number is given [**?**] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [**?**], a chapter in a divisible book in a series [**?**], a multi-volume work as book [**?**], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [**?**], a proceedings article with all possible elements [**?**], an example of an enumerated proceedings article [**?**], an informally published work [**?**], a doctoral dissertation [**?**], a master's thesis: [**?**], an online document / world wide web resource [**? ? ?**], a video game (Case 1) [**?**] and (Case 2) [**?**] and [**?**] and (Case 3) a patent [**?**], work accepted for publication [**?**], 'YYYYb'-test for prolific author [**?**] and [**?**]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [**?**]. Boris / Barbara Beeton: multi-volume works as books [**?**] and [**?**]. A couple of citations with DOIs: [**? ?**]. Online citations: [**? ? ?**]. Artifacts: [**?**] and [**?**].

7 APPENDICES

Here comes the appandix.