Cloud Computing

Cloud computing is the use of IT infrastructure and services that are not kept on site on local computers but are leased as a service and accessed via a network, normally the internet. Rosenberg discusses in his book five principles that should be mentioned here. After that it should be considered if these are enough or if there is something missing that these principles don’t cover. In the end an application example, if your application grows at a steady state, what does this mean for your scaling needs and is going to the cloud still the best option should be discussed.

But first have a look at the five principles Rosenberg introduces for cloud computing. The five principles are:   
**1. Pooled Computing Resources**  
If a compony needs a server for its application it can choose one of the three basic operational deployment options. The first option is the self-hosting approach where the company has its own server and utilizes it for its application. Here the company has to spend a lot of money for the server and has also a lot of costs for the operational part. The second option is managed hosting. Here the company rents a server and can deploy its application on it. So, the difference between this approach and the self-hosting one is that here the company doesn’t has to spend a lot of money to buy and get the server to work. This approach seems better considering the financial part. But there even exist an even better and cheaper option, the cloud. The cloud is kind of a managed hosting, but the main difference is that at the cloud multiple applications can run on server while in the managed hosting option mostly only one application runs on a server. So, the cloud is utilizing the server more efficient, as will also be described in the second principle. Because of this the cloud I cheaper to use than the managed hosting option. Because of this it can be said that cloud computing utilizes pooled computing assets that are normally externally purchased and controlled. These pooled computing resources are contributing to a cloud if they are available to any subscriber. A subscriber can be everyone who is able to pay. In short it can be said that this principle says that cloud computing consists of pooled computing resources that are available to any subscribing user.  
This principle is essential for cloud computing as it describes that really everyone is able to use the cloud computing service. Not only a company can rent the service, but everyone can like you and me and it is even able to handle every application one can think of.  
**2. Virtualization of compute resources**  
Virtualization is being used for years, even before cloud computing was introduced. Virtualization is also an important part of cloud computing and efficiency at all. As the internet grows more and more servers are needed to handle all the traffic. Every server takes up physical space, has a huge power consumption and needs to be cooled properly. Getting a high utilization out of each server is vital to be cost effective. Here every server is portioned into multiple ones. Each of the portioned ones is separated from each other so they are not able to access data on other virtualized servers. The virtualization helps a lot utilize a server as much as possible and if the server can’t handle all the traffic anymore, the virtualized one gets shifted to another server which can handle all the traffic. This is quite good as so a company can react to the amount of traffic and workload they currently have to handle and therefore use as much computing power as needed.  
**3. Elastic scaling up or down according to need**As a cloud computing provider normally has a large number of servers a concept known as elasticity is being enabled. Elasticity is a synonym for dynamic scaling and refers to the ability to dynamically adapt the amount of resources being needed. Normally an application has base workload that is most of the time enough but then sometimes peaks are happening that also have to be considered such that the application doesn’t crash about the enormous amount of request it gets. Even the base level can increase over time, so that new servers have to be installed or rented in the self-managed or managed hosting option. In the cloud however the application can immediately utilize more servers or shut down servers as needed. For the self-hosting option, the company has to buy the server and also has enough computing power for the peaks and therefore buys a lot of hardware that is under normal circumstances not needed and therefore generates unnecessary costs most of the time. In the managed-hosting approach it takes longer to get more servers utilized such that the company can’t directly react to change in use. For the cloud however a company can say that under normal circumstances it needs that amount of computing power and as soon as the usage of its application grows over a certain point, that the cloud provider should provide more computing power such that all functionalities are still available for the users.  
**4. Automatic creation and deletion of new VM’s**  
Normally provision a new set of hardware resources takes at least weeks. For peak usage its way to slow so it is very useful that the cloud can adapt way faster, like in minutes. This ability, to automatically provision and deploy a new virtual instance and to deprovision it after it is not needed anymore is the fourth principle of cloud computing. Like described in the third principle the cloud can react to utilization of the virtual server. If it exceeds a threshold the cloud provisions and deploys an additional instance such that the application can execute all requests. As soon as the workload reduces again, the cloud deprovisions the added instance and only the base instances work again until another peak occurs. The company and user also has to pay for the additional service only for the amount of time it was really running, such that this way of handling peaks is very cost effective.  
**5. Resource usage billed only as used**  
Typically when going to the cloud the customer can allocate resources as needed and pays for them on an hourly basis. This is a big benefit as in this way the company really only has to pay for the resources it really utilizes. This is not only an advantage for big companies, but also for people who want to start a new one. They now don’t have to raise a lot of money in order to utilize hardware as they can now just rent the amount of hardware or resources they need. This is a huge difference in cost. The cloud has changed for them the play rules as now small business are more able to compete with bigger companies because of this.

All of these principles really show the advantages a cloud brings with it. In my opinion these principles are enough to describe cloud computing. There could be added one more which considers the focus of the company or user that wants to utilize a server. If someone wants to use a server he normally has to have knowledge in running a server and how to secure it. But if instead a cloud is being used, the company or user can focus on the application he wants to run and doesn’t has to think about how to secure that the server is running and secure and can therefore spent his time on more important tasks for him and simultaneously save money. As this is also one of the benefits of using the managed hosting option, I think that that’s why this is not a part of the cloud computing principles. Nothing else comes to my mind now when thinking about cloud computing that is not covered by the principles, so I would say that they are enough to describe cloud computing.

Now all the cloud computing principles were discussed, and the last part of this report should discuss the elastic scaling principle a bit further. For that an example when elastic scaling is necessary will be given. Elastic scaling is necessary when the workload/application requests increase in a short period of time. This can happen for example when a celebrity dies. Then a lot more people than usual go and search through the online news to find some information or use twitter to write about it. Then those sites have more traffic and need more resources to handle all of it. Even with elastic scaling it can happen that the application crashes, but it is not as likely as without using elastic scaling. Lower, a not as fast scaling, would probably be needed in times when the company and the users of the application grow. The growth can be seen over weeks or even months. Here the company could decide to build more server space or rent more in the cloud. Building new servers is always a question of does the company already have servers? Does it has enough capital to buy them and also does the company has enough space and infrastructure to build the new servers? If not all of these questions can be answered with yes, going to cloud is the best alternative. Even if they could be answered with yes, from economic side it can be smarter to head to a cloud provider instead of installing new server space. Growth is often steady and always forces the company to think about how to increase the server capability. Beside the financial and space aspects the company always has to consider if it will be able to handle and maintain all the servers. As the base usage grows, the peak also grows. Therefore, it is not only needed to add new servers that cover the growth. Servers are needed that also can cover the now even bigger peaks of usage. Considering all this I think that going to the cloud is still the best alternative as here the cloud provider cares about the servers. As this is the only job he does in his role, they are more likely to provide better more efficient servers and can handle problems with them faster.