Service availability (in the clouds)

Aleksander P. Czarnowski, AVET INS / EuroCloud Polska
**Table of Contents**

1. Introduction................................................................. 3
2. Availability concept for IaaS, PaaS, SaaS ........................................ 3
3. Availability and the type of cloud............................................. 3
4. Availability based on usage scenarios ........................................ 4
5. Availability for free of charge and commercial cloud services .......... 4
6. The availability matrix ........................................................ 5
7. The SLA problem in SME market ............................................ 6
8. Continuity ........................................................................... 7
9. Quality (of service) and performance ......................................... 8
10. Functionality ....................................................................... 8
11. Incident management .......................................................... 8
12. Monitoring ......................................................................... 9
13. Data access ....................................................................... 9
14. On-demand data access ........................................................ 10
15. On time data access and processing ........................................ 10
16. Security ........................................................................... 11
17. Cloud lock-in as availability threat ........................................... 12
18. Supply chains and availability ................................................ 12
19. What the industry have to say about the availability issue .............. 12
20. Summary .......................................................................... 14
1. Introduction

High availability is probably one of most important driving forces behind switching to the cloud, especially in case of SME sector. While overall prices for computing power have been decreasing constantly in last years, and the computing power of even most basic systems increased dramatically over last two decades, there are some areas that are still quite expensive. Energy, storage, backup, knowledge, are those resources that are not cheap. As the demand for computing power and/or storage increases due to business requirements, the energy consumption and heat generation increases consequently. In every organization lifecycle there is a breaking point where IT systems starts to play so important role that their availability and quality of services becomes a concern together with previous issues. Cloud computing promises – at least in theory – relatively cheap, high quality, high availability services to the masses. In this report I’m trying to show some issues that can be treated as possible obstacles to the above promise. In the same time I propose some solutions that may allow to turn theoretical promise into sound and fair business proposal.

2. Availability concept for IaaS, PaaS, SaaS

It is important to note that type of service offering is crucial for defining the term “availability” in aspects of any SLA agreement:

- **Infrastructure as a Service** – the availability applies only to the infrastructure availability, platform and software is being excluded from SLAs.
- **Platform as a Service** - the availability applies not only to the infrastructure availability on which the particular platform is being run but also concerns the platform availability as well. The software in case of services of applications being run on those platforms are still being excluded from the SLAs.
- **Software as a Service** - the availability applies to the software being run on particular platform exploiting particular infrastructure, so in reality all those 3 components form proper SLA even if only software is being mentioned.

3. Availability and the type of cloud

If we divide cloud types into following 3 categories:

- **Public clouds** – the infrastructure and software needed to deliver the service is available remotely through the internet, no part of the cloud (except web browser for example) is running on client side. Basically all customers shares the same infrastructure and software.
- **Private clouds** – the infrastructure and software needed to deliver the service is located within organization own networks and physically located there, no other organization can use this cloud for own purposes. There is no sharing of infrastructure, software or other resources with 3’rd parties. Every entity can use a bit different cloud.
• **Hybrid clouds** – this is the mix of public and private cloud.

It is becoming obvious that again the term “availability” can depend on the type of cloud service offering. For example for the end user in case of public cloud, availability of the cloud means not only the services provided by the cloud service provider but also he needs means to be able to access those services remotely. In most cases the cloud service provide does not provide and has no connection with telecommunication/internet provider. However some telco/internet provider already offer own cloud services to the end customer and they tend to target both consumers and SMEs.

Private cloud has very limited use in case of consumers however in medium size enterprises they can be used a bit wider. Since the cloud is running on end user infrastructure and is exploiting internal IT resources of the end user the availability from cloud provider perspective is again limited to the services. The problem of infrastructure availability is beyond his scope (unless he also delivered the infrastructure).

4. **Availability based on usage scenarios**

Obviously the availability requirements means something different not only because of the (mis)conception of the end users and type of cloud they are using but is also based on type of the service they use. If it is social network service used for non-commercial use than the availability requirements and expectation are probably lower than in case of commercially used financial management system deployed in the cloud. So the end user usage scenarios and service functionality have direct impact on end user availability requirements and expectations.

5. **Availability for free of charge and commercial cloud services**

Obviously SLA levels offering can and will differ for free of charge and commercial cloud services. Despite the business model of particular service offering I believe that the SLA level should be clearly stated for all types of offerings and services independently of their business and funding model. It means that even in case of free of charge services end user should be informed about SLA level they can expect so SLA can be one of attributes used by them when choosing particular service or service provider. In free of charge services providing no SLA at all could also be possible, nevertheless the end user should be aware of it.

If there are different SLA levels possible for single service the end user should be given choice at:

- Starting using the cloud service for the very first time
- Any time using the cloud service offering if his needs has been changed

The pricing model for the SLA should be clear and well explained to the end user.
6. The availability matrix

In order to better understand issues presented further in this report and to cover all important issues in a systematic way I came up with cloud availability matrix. All issues and ideas discussed further will be focus on at least one part of this matrix.

![ Availability Matrix Diagram ]

Figure 1: The availability matrix

As already demonstrated the availability concept is quite complex one, therefore everybody can came up with own definition and very personal (and therefore quite unique) understanding of it. The same rules applies to consumers (private users) as well as business ones. The one important difference between consumer and a business user is that the later one should be able to define the requirements for his understanding of availability. In reality however it might not work this way, since large group within SME sector may not have the proper knowledge not only to define their availability needs in detail but also may not able even to monitor it.

<table>
<thead>
<tr>
<th>ID</th>
<th>AVAILABILITY COMPONENT</th>
<th>COMPONENT DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Continuity</td>
<td>Continuity ensures that the service is available for certain amount of time without any interruptions. Furthermore if there is an incident, continuity allows to restart the service and regain access to data and functionality of the service within particular time frame.</td>
</tr>
<tr>
<td>2</td>
<td>Quality (of service)</td>
<td>Quality describes attributes of service offering like access time,</td>
</tr>
</tbody>
</table>
number of supported users, amount of data processed in a way that is convenient for the user for example without any lags or timeouts.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>3</strong></td>
<td><strong>Functionality</strong></td>
<td>Functionality describes what end user can do and achieve in a particular time frame using selected services.</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td><strong>Incident management</strong></td>
<td>Incident management deals with incidents concerning services used by the end user. In case of any disruption of any above components the end user should be notified.</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td><strong>Monitoring</strong></td>
<td>In order to assure proper availability levels both the service owner and the service end user should be able to monitor it. It means that continuity, quality, functionality and security components should be measured and checked constantly.</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td><strong>Data access</strong></td>
<td>We can’t speak about availability if we forget the data either entered by the end user or data generated by the service based on end user data. Basically all the data end user enters and process including all results of this process should be available on-demand by the end user.</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td><strong>Security</strong></td>
<td>Obviously security is tightly connected with availability and its components like incident management, monitoring and data access. The end user expects that all those processes are in place and are working correctly. Furthermore as part of data access he expects his data to be secure i.e. no unauthorized access to his data is possible.</td>
</tr>
</tbody>
</table>

Table 1: Availability components list

7. The SLA problem in SME market

The Service Level Agreement and availability levels defined within it could turn out to be quite challenging to the SME users. There are several factors which play important role here:

1. Can SME end user define his real needs and requirements and check if SLA levels are correct when selecting cloud offer?
2. Can SME end user monitor SLA levels?
3. Can SME end user demand financial fees when SLA levels are incorrect assuming he is not able to properly define his availability needs and/or monitor the SLA levels himself?
4. How does service provider really estimates his SLAs? Most numbers like 99,999% of availability per year are imaginary numbers that marketing gurus came up with.

Since some (if not many) users will based their selection of cloud service offering based on price and availability it seems that fair information about the real service availability options must be included
in pre-contractual information. However this does not solve the problem that many end users will not be able to define their real needs in terms of availability. Since I’ve already demonstrated that the availability is quite a complex thing, providing purely technical information to the end user will not be a proper solution. While technical specification should be always available, many end users will not be able to use it due to lack of knowledge.

Obviously the problems described above are not new, and since the market does not like empty space there is already a number of cloud service comparison sites trying to aid even less tech savvy end users to make a proper choice. It is important however to note that those sites should not be associated with any of the cloud service delivery companies. Another issues to remember about is since the availability needs can and will differ sometimes simple comparison will not be enough. While alone this is not necessary a problem, if end users will based their decision on incomplete comparison without realizing it, their choices can be wrong based on misleading information.

1. Comparison sites should inform in clear way what is their affiliation with other could service providers
2. Comparison sites should inform in clear way user if the comparison report is not taking all important availability attributes into account
3. SLA based on magic or imaginary numbers should be prohibited – cloud service provider should document and explain his SLA levels and their source
4. Pre-contractual information should contain availability specification and availability options
5. Pre-contractual information should contain clear information on SLA fees
6. Pre-contractual information should contain information about SLA monitoring facilities delivered / offered by the cloud service provider

Another crucial problem is how the SLA is being defined. Main concerns are:

- Are the service windows included or excluded from the SLA and what is their impact on availability
- How the SLA defines the interruption of the service
- How the SLA deals with following situation: the service has been unavailable every day for 5 seconds every hour within one month or the service has been unavailable for whole 24 hours within a year. The first situation makes the service complete unusable while in theory it is available for more than 23 hours per day.

8. Continuity

One of the key factors for SME market to switch into the cloud is foreseeable preservation of continuity of their business critical IT needs. In theory building complex and effective Business Continuity Management solution including Disaster Recovery Plan is not only time consuming but also requires expert knowledge and resources. SMEs might not be able due to financial reasons to meet any of the above requirements. Therefore switching into environment where continuity is
being preserved by 3rd party in a lot cheaper way is very tempting opportunity. However in order to provide any possible availability of the service in a long term the cloud service provider really needs to have in place complex and effective Business Continuity Management solution including Disaster Recovery Plan based around sound incident management process. If the cloud service provider fails to do that he puts in risk not only his own business but also the business and data of his customers/users. In turn continuity is critical component of availability.

9. Quality (of service) and performance

Quality of service is important element of availability of any service – if the end user must wait long for the service response it is not in operational state any longer, even thou from technical perspective the service is working. This is important aspect in terms of contract since SLA agreements should also define quality of service level plus expected performance as well and such information should be available to the customer in pre-contractual phase already. Putting real numbers into the signed contract is already too late.

10. Functionality

Just like in case of continuity, the preservation of minimal functionality is critical from SMEs business perspective. If crucial functionality is missing due to removal or fundamental change done by cloud service provider during the contract the business of the customer might be in danger. However this is very tricky part due to very dynamic evolution of the cloud technology and services provided by the cloud service providers. In turn nobody wants to limit the development and innovation of cloud computing. However it is important to ensure that the end user will have a time widow when functionality he desires is available and he is aware of this fact before entering the contract.

11. Incident management

If the cloud service provider does not have sound incident management process in place he cannot guarantee any business reasonable SLA and in turn can no provide any availability assurance.

The incident management process is also important from end user perspective, especially in case of commercial services. If there is availability problem due to hardware or software failure or due to popular DDoS (Distributed Denial of Service) attack the end user should be informed not only about the problem but also about expected downtime due to this incident. After the incident has been handled and service is operation again, he should be informed again that he can restart using the service. However recovering from the incident can be a time consuming process, especially in case of DDoS attack for example, so restoration of full quality of service might not be possible at once while the service in principle becomes available again.
12. Monitoring

Monitoring in cloud services is not necessary trivial thing considering the amount of resources and assets that could be and should be monitor for availability, performance and quality of service. If the end user is not given any monitoring tools by his cloud service provider he might not be able to monitor the cloud service himself due to several reasons:

- Lack of knowledge
- Lack of resources
- Lack of access to critical cloud service assets that should be monitor in order to be able to monitor the SLA

There are however some services that tries in one or another way monitor cloud service provider. They start with monitoring simple hosting services (which are obviously not the cloud) and end with quite complex monitor tools.

Sometimes the monitoring information can be gather from 3rd parties sites (like Netcraft for example – see Figure 2) or industry reports. Some of those products are free of charge but in such case they are usually not very detailed.

13. Data access

In terms of availability of cloud service it is not just about the availability of the service itself but also about secure data access to:

- Data entered by the end user
- Data that is the result of computing operation based on end user data

If either the access to data is not available or access is possible but only in insecure manner this is should be clearly treated as violation of any commercial SLA agreement. In case of services that do not process privacy, financial or health care data secure access might not be a requirement and in fact might not even be available from the beginning.

Data access is also important during contract termination and data retention period. Access to data should be granted under fair conditions for a reasonable time frame.
14. **On-demand data access**

Obviously the cloud should provide on-demand access to customer data stored / processed in the cloud. Despite normal business cases for data access in cloud service, there are also for example fiscal requirements defining data retention period and being able to access data upon 3rd party request. A great example could be a fiscal control case. Lack of access to fiscal data in such case can result in charges against control subject.

15. **On time data access and processing**

Another interesting case is when end user is basing his decision based on data provided by the cloud service. If the data access is granted too late the end user can make wrong decision based on not accurate information leading to potential lose. Another case is when the end user is not able to close transaction on time. Such situation can also lead to potential (financial) losses. The important
question here is: how the cloud service provider can repair potential damage and how contractual penalties can or should be enforced in such cases.

From technical point of view it is important to note, that not every case described above can be easily monitored and therefore the proof of potential loss or damage can be impossible. However there is still a possibility for proof of negligent service provider behavior so both situations should be considered under assumption that the end user may still not be able to monitor any of above cases due to several factors:

- Limitation of monitoring functionality
- Limitation of end user knowledge
- Limitation of end user resources
- Other

### 16. Security

Last but not least security has direct impact on all availability components and final, real SLA levels. In fact if we take a look at well-known CIA triangle we will find out that the availability is one of key components of it:

- Confidentiality
- Integrity
- Availability

We should consider availability an integral part of any cloud that is being marketed and advertised as secure. However this does not define the exact SLA level unfortunately but that simplistic approach perfectly fits SME end users needs and understanding of availability issue. The one solution that comes to mind is to propose a baseline SLA for a public cloud service offering to be considered “secure”. Such SLA should contain minimally:

- Maximum downtime
- Recovery Point Objective (RPO)
- Recovery Time Objective (RTO)
- Minimal and maximum capacity of the service
- Minimal and maximum quality of the service

If we follow this conception, in order to be compliant with above requirements, service providers should monitor Recovery Time Actual (RTA) and eliminate possible gap based on following equation: 

\[ RTA - RTO = \text{gap} \]
17. **Cloud lock-in as availability threat**

One crucial thing not covered by the cloud availability matrix presented at the beginning of this report is a threat of cloud lock-in. This is an important threat to any type of availability since in case of lock-in customer cannot move his data to different service provider. Secondly in dynamically changing world of cloud computing services and service providers end user might not even be aware of possible lock-in problem in case when his service provider (or one of providers inside link of providers) is going out-of-business. In such case the data might become unavailable for an indefinite period of time and the risk of losing all data entered into the cloud cannot be underestimated.

18. **Supply chains and availability**

Depending on the type of the service it could be deliver by one supplier or a chain of suppliers. From end user perspectives this could be unnoticeable until something goes wrong i.e. one of chain suppliers has technical problems or is going out of business for example. Obviously the end user has some form of a contract with front supplier. Therefore – at least in theory – the front supplier should have appropriate contracts with the rest of suppliers within his chain. Such contracts should define fees, SLA levels etc.

- End user SLA must be compatible with SLAs within the whole chain in order to protect end user rights
- In case of switching from one supplier to a different one within supply chain the end user should be informed and be given ability to quit contract if the new SLA is not exactly the same with the one had before – this applies both to free of charge and commercial services

The last case brings up question if country of data processing of supplier should be treated as a part of SLA. I believe this is not the case since SLA should deal with availability, quality of service, capacity and security related issues. The exact location maybe the part (due to certain sector or local regulations) of the contract but should not be part of SLA in order to make easy to define and monitor.

19. **What the industry have to say about the availability issue**

This report would not be complete without at least one point of view of the cloud service provider.

“High availability is a key area of interest for organizations embracing cloud computing. Therefore there are a lot of areas that need to be addressed to ensure a high-availability and security on the whole cloud solutions, such as network vulnerability, multisite redundancy and storage failure. These aspects are generally represented by implementing the same or better solutions as they are in traditional IT environments. Here we can find a lot of solutions which can be implemented by cloud providers on different levels. Traditional public cloud availability delivers “Triple Nines,” or 99.9%. Taking into consideration recent clients’ requirements from the service provider perspective we see
that Enterprise Market requires SLA at 99.99% uptime (“4-Nines”) and even more i.e. in IaaS services availability: 5-Nines (99.999%)

To meet these requirements cloud providers must pay attention also on another of the most important area - the cloud’s infrastructure, including hardware such servers, routers, storage devices, network, power supplies, cooling and fire protection systems, and other components that support operations. It is essential that the cloud infrastructure is highly resilient, by implementing highly secure and highly quality devices and reasonable maintenance program. The cloud infrastructure must be built and delivered with availability at its core. Redundancy and solutions with concurrent maintenance is a key thing in a “high availability cloud computing world”

All of these areas must be evaluated by official measurable standards to give customers clear view of SLA performance. In other words: cloud providers and enterprise data centers need improved tools, processes and procedures aimed at proactively monitoring SLAs.

Thinking about availability we have to meet with requirements regarding safeguards such as physical security include the adequate control and monitoring of physical access using bio-metric access control measures and closed circuit television (CCTV) monitoring.

Cloud and data centers providers selling their services should compare all above aspects with local law and regulations. We cannot forget that all UE countries have different rules and both users and owners of cloud solutions have to respect them.” – Piotr Mańka, Manager DC Design and Maintenance at LinxTelecom

Here is another opinion from the same company but focusing on the offering for enterprise – it is important I believe to see the difference between SME offer and enterprise offerings:

“We do not target SME and definitely not consumers. We target the high end of the market so our IaaS proposition is way too expensive for SME, SoHo and Consumers.

In our risk analysis we identified high availability as a top priority for the high-end customers; therefore we implemented several costly controls to reduce the impact and probability of outages, which enables us to provide our LinxCloud Service under an SLA that guarantees 99,9% availability:

- our platforms are based on the validated Secure Multi-Tenant (SMT) design by Cisco, NetApp and VMware
- we have implemented the High Availability (HA) variant, which means we run the server blades in a 7:1 redundant configuration (7 active, 1 hot stand-by) with seamless transition of all VM states if a blade fails and we have duplicated networking and storage
- of course the IaaS platforms are fully monitored 24*7 from our Tallinn NOC
- our platforms are hosted in ISO27001 certified datacenters designed to comply with TIA-942 Tier3 requirements. This means high security, dual power feeds and systems and dual cooling systems resulting in a DC availability of 99,982%
- we have off-site back-up platforms (in Warsaw this in LIM -1, in St Pete this is in Borovaya) connected by our own dual diversely routed fibers
• *if customers are willing to pay for it, we can also use VMotion to move their entire environment to another physical location (at the moment only St Pete, but Moscow and Tallinn later this year) in case of an outage*

“—Niko Bel, Product Manager at Linxtelecom

20. **Summary**

Due to complexity of the “availability” term and different meanings of this concept and large spectrum of needs and requirements of cloud services for SME and consumers there is no magic silver bullet solution. In long term I believe that further actions should directed towards:

1. Assurance of fair and clear description of SLA levels and their pricing and distinction between offerings within single provider.
2. Education for SME sector so they can better define their availability needs and requirements and select most appropriate solution
3. Modification of regulation that would result in end user protection when their data is not available due to cloud service offering failure making them inaccessible on-demand as expected
4. Assurance of minimal service functionality for particular time frame

In any case we should always remember about dynamic changes in cloud computing and therefore we shall not limit its development nor innovation, however the minimal availability requirements for business oriented services should be in one way or another enforce in order to protect SME end users.
Main report questions:

1. What are cloud users' legitimate expectations regarding the availability of the cloud service? Would it be possible to define minimum criteria that should always be satisfied in terms of different components of availability? In particular:

   - **Continuity and Incident management**: Should there be a right to minimum service to ensure that critical business IT needs are always fulfilled? In case of incident, what measures shall the cloud service provider take and which information should be given to the user when and after an incident occurs? (e.g., back-up-policy, reason of the incident, period of time during which the service is down, period of time until which the quality of the service will be fully restored, payment of fees for restoration depending on the cause of the incident...?) If the cloud provider is required to have effective business continuity processes in place, how should these processes be designed in order to be effective?

   - **Quality (of service) and performance**: How would you define "quality of the service", by contrast with mere "access" to service? What information should the user receive about the quality of service when choosing a cloud solution?

   - **Functionality**: Taking into account the need not to hamper the development of innovation of cloud computing, what measures shall be taken by the cloud service provider to maintain appropriate functionality of the service and how should the consumer be made aware of this?

   - **Data access and on-demand data access**: What could be fair conditions or a reasonable timeframe to ensure ordinary access to which kind of data? Should there be special conditions or timeframe to ensure access to data in specific cases, such as tax control?

   - **Security**: Security means both ensuring that all processes work correctly and preventing unauthorized access to or disruption of the service. When informing the user about availability, should a simple statement confirming that security is provided be sufficient or should the cloud service provider need to demonstrate that it complies with a series of pre-defined security requirements? If so, how should these security requirements be defined?

2. The Service Level Agreement and availability levels defined within it can be challenging for SME users. With this respect, how could SMEs (as an end user) be helped to: (i) define their real needs and requirements and check if SLA levels are appropriate when selecting a cloud offer? ; and (ii) monitor SLA levels?

3. How should the availability of the particular service be measured / assessed and presented, in a comparable manner, to enable the cloud user to have a clear understanding of what he will get?
4. Can you identify any cloud specific reasons (other than data protection) why a cloud service provider would not be responsible for the availability of its service and liable in case of disruption of availability towards its users because the cloud service provider relies on its subcontractors?

- Should there be special contractual liability rules between cloud sub-contractors and the main CSP?
- Would a joint "network liability" of the provider and all its subcontractors towards the user make sense?
- Should the user have a direct claim against the sub-contractor in case of availability related damages imputable to the sub-contractor? Would this be workable in practice in a cloud service environment?
- Should the subcontractor have rights towards the user? If so, how should they be granted? Should the subcontractor be allowed to access user's data? Should there be a link between the subcontractor's liability and user's behavior? A duty of the user to cooperate with subcontractors?

5. Availability was mentioned as a key element to be part of the pre-contractual information: what level of details on availability should be part of the pre-contractual information?

6. Horizontal distinctions to take into account when responding to the above questions: when responding to the above questions, please bear in mind the distinction between consumers and SMEs, between free and paid services, and whether the service supplied is an SaaS, PaaS, or IaaS.