www.jusletter-it.eu

Philippe Gilliéron

Service Level Agreements (SLA) in cloud-based agreements

Best practices from a practitioner's standpoint

Dieser Beitrag ist nicht als wissenschaftlicher, sondern als wirtschaftsorientierter Beitrag gedacht, der die Erfahrungen des Autors widerspiegelt, nachdem er mehrere hundert Cloud-basierte Verträge und damit verbundene Service Level Agreements (SLAs) für zahlreiche Unternehmen ausgehandelt hat, von kleinen und mittleren Unternehmen (KMU/SME) bis hin zu multinationalen Unternehmen, in den meisten Fällen zugunsten von Kunden. Daher richtet sich dieser Beitrag vor allem an Praktiker, die sich in ihrer täglichen Arbeit mit SLAs auseinandersetzen müssen, d.h. die Leser werden keine Fussnoten oder wissenschaftlichen Referenzen finden, wie sie normalerweise in einem wissenschaftlichen Werk erwartet würden. (kg)

Kategorie: Beiträge
Region: Schweiz
Rechtsgebiete: IT-Recht; Vertragsrecht; Cloud Computing

Zitiervorschlag: Philippe Gilliéron, Service Level Agreements (SLA) in cloud-based agreements, in: Jusletter IT 23. Mai 2019

ISSN 1664-848X, jusletter-it.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

Table of contents

I. Preliminary remarks
II. Availability
  a. Definition
  b. Rate
  c. Credits
  d. Termination events
III. Response time
IV. Releases
V. Incident management
  a. Responsibility
  b. Channels of communication
  c. Support hours
  d. Priority levels
VI. Governance
VII. Conclusion

I. Preliminary remarks

[Rz 1] When my IT legal practice started around 2010, companies were strongly reluctant to adopt cloud-based solutions, which were still considered as to be avoided in favor of on-premise solutions. Truth is that, at that time, to engage with cloud vendors meant a loss of control over the companies' data that was not necessarily mitigated with a robust IS architecture from these cloud vendors. Over the years, however, things have changed: flexibility, scalability, easiness of SaaS to be rolled out globally in comparison with on-premise solutions coupled with robust IS standards and security controls now far outweigh these fears. From the exception, cloud-based solutions have become the rule. These services now play a key role in the digital transformation of companies, to a point that companies regularly make them today a pre-requisite in any RFP process and, conversely, that vendors give up in numerous instances their on-premise version of a solution in favor of a sole cloud-based one.

[Rz 2] The transfer of companies' data to a third party's infrastructure such as the hosting environment of a cloud vendor and the use of SaaS will obviously be coupled with the contractualization of additional safeguards in comparison with the mere licensing of an on-premise solution that would enable the company to keep control of its data. While data protection and security immediately pop into one's mind, rightfully so, SLAs will obviously also play a key role. To entrust a cloud vendor with some key data and processes for the companies indeed bear some risks in the absence of sufficient safeguards, as business continuity might potentially be endangered should the service be unavailable. The mitigation of such risks will be the objective of a well drafted SLA.

[Rz 3] The goal of this paper is to list the salient points of an SLA and to comment on the best practices on those points as of today.

II. Availability

a. Definition

[Rz 4] A company subscribing to a cloud-based solution will obviously expect the service to be running 24/7, 365/365. As unfortunate as it is, truth however is that bugs do exist and that, no matter how reliable a vendor is, some downtime of the service still is to be expected. As a result, it is important for the parties to an SLA to agree on the percentage of downtime of a service that is considered acceptable and reasonable for the parties, without triggering the potential allocation of service credits in favor of customers.

[Rz 5] To assess an acceptable availability percentage, several factors will come into play:

- First, will the availability be calculated on a weekly, monthly or quarterly basis? Choosing a certain periodicity may indeed bear some impact.

  To take an example, if one considers that a service has to be available 99% of the time on a daily basis (a rather uncommon periodicity to be true), downtime per day may amount at the most to 14.4 minutes. The same availability rate, however calculated on a monthly basis, leads a vendor to be entitled to have a downtime per month of 446.4 minutes, i.e. potentially more than 7 hours in a row in the absence of any safeguards...

  Needless to say, while 14 minutes in a day might be acceptable (although not repeatedly which may lead to a termination event, an issue we shall discuss later on), 7 hours in a row may significantly impact business from an operational standpoint.

  As a result, depending upon the periodicity measurement agreed upon, a customer will (unlike the vendor) have an interest in ensuring that any downtime may not amount to more than X minutes per day for instance, so as to avoid jeopardizing operations without any remedial actions against the vendor[1].

  In practice, it is fairly common for the availability rate to be measured on a monthly basis and reported on a monthly or quarterly basis according to the author's experience.

- Second, parties will have to define whether maintenance windows will be taken into account or not in the availability rate to be agreed upon. In practice, maintenance windows are normally not taken into account into such measurement. To avoid significant maintenance windows that might, there again, have an impact on the availability and, consequently, over the business continuity, customers will have an interest in trying to agree with the vendor on a schedule that will set: (i) the periodicity of the maintenance windows (for instance no more than one per month), (ii) their duration (for instance no more than 3

[1] In theory, customers may try and argue that a 7 hours downtime in a single day amounts to a material breach of the agreement. Vendor will then try and rebut such argument in stating that its compliance with its SLA's obligations stands in contradiction with a customer view that a material breach might exist notwithstanding such compliance.

hours per maintenance windows) and (iii) the time when such maintenance windows will occur (for instance on Sunday between 2-5am in the time zone deemed the most relevant for the company if the company is a global player).

It is standard to consider that critical maintenance resulting for instance from a security incident will not be subject to the parameters agreed upon for the scheduled maintenance windows, and that these urgent handlings will not be considered as a downtime of the system counting towards the availability percentage agreed upon either. This emergency maintenance should however be immediately notified to customers and ideally addressed so as to minimize the potential impact upon customers' operations.

b. Rate

[Rz 6] Taking these variables into account, parties will then have to agree upon an availability rate, which will notably depend upon the criticality of the solution for the business[2]. While an availability as close to 100% as can be will be key if the solution is required to enable a proper performance of the supply chain of a company, and thus to its income, such rate may be lower if the solution aims at fixing bugs in code development or other ancillary services without any impact upon the business.

[Rz 7] Compliance with the availability rate will then be automatically checked and reported by the vendor in favor of customer on a periodicity to be defined, usually ranging from monthly to quarterly.

c. Credits

[Rz 8] Absent any sanction, the setting up of expected availability rates would hardly make sense. Credits will in most instances take the form of a certain percentage of the monthly or yearly service fee; the rate will obviously depend upon the leverage power of the parties, vendors aiming at a rate as low as can be, usually capped at a certain percentage no matter the number of service failures, while customers will try and obtain a rate high enough for vendors to have a strong incentive to comply, without any cap. The objective for customers in setting service credits indeed never is to cover potential damages resulting from a downtime, but rather to create an incentive for vendors to meet the customers' expectations. As a result, customers will always try and ensure that service credits do not function as liquidated damages, an attempt that vendors for obvious reasons will to the contrary try to pushback.

2 It goes without saying that the word «agree», as might be the case in several instances in this paper, might be an overstatement. While major customers may have sufficient leverage to negotiate an SLA, SMEs may find it hard to get any concession from a vendor. This is all the more true, including for major customers, in a cloud-based environment which, by definition, is a multi-tenant environment where customization is most of the time hard to implement.

[Rz 9] Ultimately, an availability matrix may look as follows, the rate of 99.9% availability of the service on a monthly basis reflecting a fairly standard customer expectation for critical services[3]:

Availability rate	Credits
Below 99.9%	10% of annual fee
Below 98.5%	20% of annual fee
Below 96%	30% of annual fee

[Rz 10] In more complex scenarios, parties may agree to set minimum and expected thresholds, in which case exceeding the expected thresholds might entitle vendors to get a bonus or, alternatively, set off credits it may have suffered as a result of another SLA breach.

[Rz 11] The way service credits will be allocated will also have to be addressed in an SLA agreement. Such allocation may take place in different ways. If the service fee has to be paid on a monthly basis, parties may agree to automatically credit the service on the next invoice; if the payment takes place on an annual basis, parties may agree to aggregate the credits on a yearly basis to then be credited on the next yearly invoice.

d. Termination events

[Rz 12] While service credits may be helpful in creating an incentive for vendors to comply, the repeated allocation of service credits may not be considered as a satisfactory remedy if the SLAs are breached month after month. In that case, customers may be willing to have more coercive remedies at disposal, such as the entitlement to terminate the contract[4]. These so-called «termination events» may come into play when the availability rate appears to be much lower than anticipated to a point that customers cannot trust the service anymore, or if the SLAs are breached repeatedly over the months, such periodicity to be agreed upon by the parties.

[Rz 13] Once termination events are taken into account, the availability matrix may look as follows[5]:

Availability rate	Credits	Termination events
Below 99.9%	10% of monthly fee	5x in a row or 6x in a year
Below 98.5%	20% of monthly fee	4x in a row or 5x in a year
Below 96%	30% of monthly fee	2x in a row or 3x in a year
Below 95%		Termination event as such

3 This is obviously only an example, an aggressive one favoring customer. Service credits of that magnitude will be rare and require a significant negotiation power from customers that will seldom be met. In vendors' perspective, credits will rather be calculated on a monthly basis, at a lower percentage.

4 Although such repeated breaches might be considered as a material breach under the underlying service agreement, customers will likely prefer to be entitled to immediately terminate the agreement rather than have to accept a cure period usually linked to the notion of material breach.

5 Please take note that this only is an example and is not meant to reflect an industry standard.

III. Response time

[Rz 14] The average response time is another criterion to measure the service performance. Schematically, it consists of calculating the time between a click and the display of the relevant page. To implement an SLA on the response time is a way for customers to ensure that vendors will maintain state-of-the-art adequate infrastructure to ensure that an increase in the total number of active users will not cause latency or increase in response times.

[Rz 15] Vendors are fairly reluctant to commit on average response times, especially when these are bound to service credits. Truth is that the response time to display a webpage will depend upon several variables that vendors do not necessarily have under their control; as an example, slowdown of the network, power outage or even the datacenter location chosen by a customer on the other side of the world are all factors that will have an impact upon such response time. To mitigate such risks, vendors may be expected to maintain or engage network acceleration service or solution to ensure that there will not be any latency or increase in response times due to the geographic location.

[Rz 16] In any case, although it may thus be hard to set an SLA on such a metric, best practice now consists for customers to try and obtain an average response time below 1 second, upon a periodicity that will there again have to be agreed upon by the parties but is typically calculated on an monthly basis. Ideally, such an SLA should be coupled with a service credit, that will normally be lower than the one set for the availability rate. If the average response time proves to be particularly slow, customer may try and obtain that an average response time in excess of X seconds over a given month might be considered as downtime and, thus as an unavailability of the service. A rather aggressive average response time matrix favoring customer may look as follows:

Response time	Service Credits
1–2 seconds	5% monthly fee
2–3 seconds	10% monthly fee
3–4 seconds	15% monthly fee
4–5 seconds	20% monthly fee
Above 5 seconds	Considered as unavailability

[Rz 17] While one should never say «never» in terms of contract negotiation, the author has never faced any termination event upon an average response time SLA. In practice, as stated, an excessive average response time may be considered as a downtime and, potentially, as a material breach.

IV. Releases

[Rz 18] Software are not static but dynamic products. Over time, to keep up with technology (and potentially generate new revenues), vendors will improve their products, which may take the form of an update, upgrade or a new release version («Releases»).

[Rz 19] When software is installed on-premise, it will be key for multinational companies to control the release cycle of their vendors' products. The reason for that is that the rolling out of

a Release in several markets having each several sites will take time, often measured into months rather than weeks. As a result, if a vendor releases 5 Releases a year for instance, multinational companies will find it hard to cope with the rhythm and to be always running on the latest version.

[Rz 20] To mitigate that risk, multinational companies will typically have their vendors make commitments of the following kinds:

- Have the vendor share its product roadmap for the next 12 months;
- Ensure that there will be a minimum of Releases per year to demonstrate continuous improvement, but no more than a certain number which would make it difficult for customers to roll them out as explained above (for instance at least 1 major release per year, and a maximum of 3 minor releases);
- Ensure that these Releases will only be implemented during the agreed scheduled maintenance windows;
- Ensure that full support will not always be provided for the latest Release, but also for the two prior ones (and potentially have a residual support for even prior ones if deemed appropriate);
- Ensure that vendors will inform customers of any Release a certain time prior to its implementation (for instance one month).

[Rz 21] Such risks do not exist in a cloud-based environment, where Releases will automatically be implemented at a global level by nature. This may be one of the explanations for multinational companies to favor cloud-based services rather than on-premise solutions for global solutions.

[Rz 22] Although thus far limited, risks resulting from Releases in a cloud-based environment are not totally excluded. Such notably is the case of rebundling, which consists to remove certain functionalities from a product to have them implemented in a different product. Such rebundling may obviously be detrimental to customers that may have subscribed to a service for certain functionalities which, over time, may be removed through Releases and require them to subscribe to a different service to keep that feature. To avoid such an unfortunate outcome, customers will try and ensure that, no matter the Release, vendors commit to ensure that existing features and functionalities as of the date of the subscription will not be removed from the service, respectively that the service subscribed to will not suffer from any material downgrade during the subscription period (which may include potential renewal periods). Customer may furthermore try and ensure that they will be entitled to any successor product at no additional fee, an entitlement that however usually proves hard to obtain for obvious reasons.

V. Incident management

[Rz 23] Probably one of the most important points to be addressed in an SLA relates to incident management procedures. Although these procedures will obviously vary from one vendor to another depending upon several factors such as criticality of the service or its geographical scope, the following items will be contractualized:

a. Responsibility

[Rz 24] The first point will be to allocate the support responsibilities between the customer and the vendor. This is traditionally done in accordance with L1 (level 1) to L3 (level 3) support classification: L1 support includes interacting with the end users, understanding their problems and then creating a «ticket» against it. The L1 support team is in the front line to interact with end users and may be able to solve the issue when the request is minor and does not require technical knowledge. When the issue is too complex to be addressed by the L1 support team, it will reroute the ticket and escalate it to the L2 support team, that will possess some technical knowledge. Ultimately, when the L2 support team proves unable to solve the problem, it will then reroute the ticket and escalate it to the L3 support team. L3 is the last line of support and usually comprises of a developer team which addresses the technical issues that require code/development related fixes.

[Rz 25] While each level of support will be provided by the vendor in most SMEs, it is fairly common within multinational companies to have their IT department address L1 and L2 support after having been trained by the vendor, and to only have the L3 support issues addressed by the vendor.

[Rz 26] Truth is that the L1 to L3 classification will make less sense in a cloud-based environment than for solutions that are installed on-premise. As a cloud-based environment is in most cases a multi-tenant environment that is not controlled by customers, vendors will manage the end-to-end support with limited customers' involvement. This categorization may however still play a role notably in hybrid or private cloud environments.

b. Channels of communication

[Rz 27] Parties will have to agree upon the channels of communication and ways to report incidents. While the use of a ticketing system is fairly common, emails or even calls might be required by customers depending upon the issue at stake. While the use of a ticket for a minor issue which does not disrupt operations might be good enough, such will certainly not be the case for a major defect that would affect all sites on a global basis and impact operations at global level; in the latter case, customers will want to be able to get a proper phone line.

[Rz 28] The triage of tickets might also become an issue when the vendor provides both the infrastructure (IaaS) and the service running on that infrastructure (SaaS). In that case, it is not uncommon for customers to have to face two support teams, one related to the infrastructure and one related to the application. Considering the fact that the end user may have difficulty in assessing whether the issue he is confronted with relates to the infrastructure or to the application, customers should try and ensure that the management of the ticket will be addressed by the vendor internally. Should the root cause analysis carried out demonstrate that the issue which the customer thought to be related to the application actually relate to the infrastructure, the vendor's application support team should reroute the ticket to the infrastructure support team on its own and inform customer of such rerouting. As this transfer will have an impact upon the resolution, such rerouting should be taken into account and addressed by the parties in the SLA.

c. Support hours

[Rz 29] Vendors will usually provide different packages ranging from a standard service offering support during their business hours five days a week, up to a platinum service offered 24/24 seven days a week. Anything in between may also be available or negotiated depending upon the parties, such as 24/24 five days a week, or even 8/24 seven days a week.

[Rz 30] In most instances, to insist on an extensive support and have to pay the resulting high costs will not make sense; a standard support will suffice for the vast majority of the services. In certain cases, customers may be able to negotiate a hypercare period during the first weeks or even up to two to three months after the implementation of a service within the company.

[Rz 31] Depending upon the criticality of the issue and their bargaining power, customers may be able to obtain such extensive support for P1 (i.e. critical one) only. In that case, vendors will implement what is referred to in the industry as the «*follow-the-sun*» regime. In accordance with that mechanism, each geographical zone will work on the resolution of the issue during its business hours, before passing such resolution along to its counterpart located on the next zone, thus from Europe to the East Coast for instance, then on the West Coast, etc. Support centers may also be defined with relevant contact details per zone.

d. Priority levels

[Rz 32] Not all issues will be addressed with the same degree of urgency. As a result, parties will agree upon a matrix defining the expectations and commitments depending upon the criticality of the incident. An incident matrix can take numerous forms and may for instance look as follows:

Priority	**Incident severity**	**Definition**	**Response time**	**Resolution time**	**Example**
1	Emergency	Service not available (all users and functions)	20 min.	2 hours	Virus, email server failure, server crash, network failure
2	Critical – site impact	Significant degradation of the service (large number of users or business critical functions affected)	1 hour	4 hours	Internet outage, finance dpt software not working, site-wide printer outage
3	Major – Department impact	Limited service degradation (business process can continue – limited number of functions or users affected)	2 hours	8 hours	Application fault, shared file unavailable
4	Normal – User impact	Small service degradation (business process can continue – one user affected)	4 hours	12 hours	Single virus, users machine crashed, internet outage for a user
5	Low – Nuisance issues	Issues which do not lead to service degradation	8 hours/2nd business day	16 hours	Unwanted popup popping up, slow computer

Source: https://cnsit.com/wp-content/uploads/2014/05/CNS-SLA-Matrix.png

[Rz 33] As end users will be the ones raising tickets, it will normally be up to them to assess the priority level of the ticket they raise. The classification made by the end user may obviously be revisited by the vendor, in a duly documented way, usually subject to agreement with customers. In the absence of any such agreement to take place quickly, escalation path may have to be followed urgently, especially if the discussion is around a P1 or P2 issue.

[Rz 34] Response time will normally require a manual response to the raised issue, and not a mere automatic acknowledgment of receipt. Goal of such response indeed is for customers to ensure that someone within the vendor is aware of the issue and has started to work on it.

[Rz 35] While customers will want to get some commitment as to the time to resolve the raised issue, vendors will obviously be reluctant to make any such commitment. Truth is that it always is fairly hard for a vendor to engage on a resolution time without even knowing what the issue and its origin are; prior to having carried out a root cause analysis, vendors thus usually consider that they cannot make any such commitment. As an alternative, parties may also agree that the resolution time will consist of providing the customer with a temporary workaround, and that they will agree in good faith upon a timing for the definitive resolution of the issue after having carried out a proper analysis and diagnosis. Another option consists of setting in between

the response and resolution times a restoration time, that will in any case usually consist of the implementation of a workaround as well until the issue is definitively fixed.

[Rz 36] In addition to the above, customers in particular may want to be regularly updated as to the progress made towards the resolution of the issues, notably if such issue is a P1 or a P2. In this case, the matrix may further provide:

Priority	**Incident severity**	**Definition**	**Response time**	**Resolution time**	**Update**
1	Emergency	Service not available (all users and functions)	20 min.	2 hours	hourly
2	Critical – site impact	Significant degradation of the service (large number of users or business critical functions affected)	1 hour	4 hours	2 hours
3	Major – Department impact	Limited service degradation (business process can continue limited number of functions or users affected)	2 hours	8 hours	4 hours
4	Normal – User impact	Small service degradation (business process can continue – one user affected)	4 hours	12 hours	N/A
5	Low – Nuisance issues	Issues which do not lead to service degradation	8 hours/2nd business day	16 hours	N/A

[Rz 37] It only is once the issue has been definitively fixed that the ticked can be closed by the vendor.

[Rz 38] Similarly to the situations we have described for the availability rate and average response time, having such a matrix in place without any consequence in the absence of compliance hardly makes sense. Although customers may be willing to try and receive credits for any failure from vendor to comply with the response and resolution times agreed upon, vendors can hardly be expected to comply 100% with the timing agreed upon, so that some targets will be set, such as:

Priority	**Incident severity**	**Definition**	**Response time**	**Resolution time**	**Update**	**Target**
1	Emergency	Service not available (all users and functions)	20 min.	2 hours	hourly	100%
2	Critical – site impact	Significant degradation of the service (large number of users or business critical functions affected)	1 hour	4 hours	2 hours	95%
3	Major – Department impact	Limited service degradation (business process can continue limited number of functions or users affected)	2 hours	8 hours	4 hours	90%
4	Normal – User impact	Small service degradation (business process can continue – one user affected)	4 hours	12 hours	N/A	80%
5	Low – Nuisance issues	Issues which do not lead to service degradation	8 hours/2nd business day	16 hours	N/A	70%

[Rz 39] The periodicity of measurement of such target may depend, but will regularly be measured on a yearly basis. In addition to these metrics, parties may agree on a maximum number of P1 to P3 issues that are considered acceptable within a year, linked to a continuous improvement plan that will be one of the objectives of each SLA review meeting.

[Rz 40] Should the vendor fail to achieve those targets, customer will be entitled to receive a credit that may be credited in the next invoice. Ultimately, the matrix may thus look as follows:

Priority	**Incident severity**	**Definition**	**Response time**	**Resolution time**	**Update**	**Target**	**Credits**
1	Emergency	Service not available (all users and functions)	20 min.	2 hours	hourly	100%	30% monthly fee
2	Critical – site impact	Significant degradation of the service (large number of users or business critical functions affected)	1 hour	4 hours	2 hours	95%	20% monthly fee
3	Major – Department impact	Limited service degradation (business process can continue limited number of functions or users affected)	2 hours	8 hours	4 hours	90%	10% monthly fee
4	Normal – User impact	Small service degradation (business process can continue – one user affected)	4 hours	12 hours	N/A	80%	5% monthly fee
5	Low – Nuisance issues	Issues which do not lead to service degradation	8 hours/2nd business day	16 hours	N/A	70%	Improvement plan

VI. Governance

[Rz 41] SLAs are not meant to be static, but rather dynamic. To ensure continuous improvement in complex IT projects, one can recommend customers to have SLA review meetings at agreed

time intervals, for instance on a quarterly basis (or more as might be considered appropriate), to measure several key performance indicators (KPIs), such as:

- Service levels and service availability achieved;
- Average response time;
- Number of incidents per priority;
- Response and resolution rates of incidents.

[Rz 42] Should the vendor's performance prove unsatisfactory, vendor should commit to engage on an improvement plan to be agreed upon between the parties. Absent any agreement, one can recommend the parties to agree on an escalation process to avoid facing a bottleneck that might ultimately be detrimental to the overall project or even relation between the parties.

VII. Conclusion

[Rz 43] At a time when IT resources are increasingly externalized, it is no surprise to see the significance of cloud-based services increase accordingly. Business-critical functions are now depending upon third parties, thus potentially putting business continuity at risk. To mitigate those risks, it is therefore key to ensure that the service provided by such third party is supported by strong SLAs. The goal of this paper was to share my experience as a practitioner and draw the readers' attention upon some salient points to pay close attention to when reviewing SLAs. If the reader comes to the conclusion that this is the case, this paper will then have reached its objective.

Philippe Gilliéron, Professor at Lausanne Law School, Attorney at Law at TIMES Attorneys (Lausanne/Zurich, Switzerland).