Paper Number 6217: On the Theory and Practice of Internet SLA's

Draft Version

2001.09.11

Abstract

The author has helped many leading Internet-related companies – both customers and providers – build SLA's. After an introduction to this rapidly evolving area, we develop a succinct but powerful "theory" of how SLA penalties should work, mostly understandable without calculus. We use this theory to discuss our experience with common SLA scenarios, as well as "game theory" for negotiating and implementing SLA's, and how this differs between customer, provider, and consultant. We close by characterizing new SLA contexts and the state and trajectory of the Internet SLA industry in 2001.

Introduction

As usual, a **Service Level Agreement** ("**SLA**") in this paper refers to a legally binding contract that specifies penalties for failure to meet specified behavioral, reliability, and/or performance goals. In general, an SLA is contracted between the **provider** and the **customer** of a service, although multi-party SLA's are increasingly common. However, within each term of an SLA, it is usually reasonable to think of the two main parties as provider and customer, as is done in this paper.

The field of SLA's for Internet-related services continues to be a hot topic in the year 2001. This has arisen out of several factors:

- Performance over the Internet is notoriously less reliable than the "five nines" (99.999% availability) by which mainframe services have come to be judged
- As the "e-conomy" has matured, a greater share of Internet service customers and providers have worked with previous mainframe-backed products, and thus have the higher mainframe reliability expectations.
- With the increasingly commercial use of the Internet, sufficiently many dollars are at risk that unreliability becomes tremendously uncomfortable and costly.
- The demand for performance guarantees counters the great reluctance of providers to guarantee what they cannot control explicitly. This friction makes Internet SLA's a topic of active discussion.
- The back-to-the-bottom-line mentality due to economic slowdown has focused competitive pressure on underwriting what matters to customers: end-user experience at the other side of the Internet cloud.
- At the same time, more distributed technologies (load-balanced server farms, CDN's, …) improve performance *probabilistically*, even if no web-page download is guaranteed individually. There is a need to measure and contract upon these improvements.

- New product offerings (such as *web services*) hope to become lucrative. This requires in some sense that they be "born adult" in order to have credibility with brick and mortar providers. Besides exposing new measurement needs, the new offerings raise fundamental obstacles to the very relevance of SLA's.
- A new level of sophistication in SLA's (along the lines of the probability-based insurance industry) is arising to meet the challenge.

Rigorous SLA's for the Internet began to reach critical mass around 2000, not just with Internet performance SLA's for web pages and transactions, but also for richer content such as streaming media. Compared to variable performance web pages, the latter are even more challenging candidates for SLA's, because not just the performance, but also the content is variable – for example, servers frequently "thin" streaming media by throwing out packets when server capacity is overwhelmed.

In 2001, offered Internet SLA's continue to evolve slowly but surely toward having more teeth (i.e. rigorous, enforceable terms.) This almost inevitable trend is being slowed by a general market dynamic of skepticism as well as a consolidation of Internet services into larger, more conservative companies. However, significant further evolution promises to continue.

We will discuss basic SLA questions in the light of experience gained with SLA reporting services first offered in 2001. We consolidate our recent experience in two ways. First, we develop a one-equation "theory" that has simple, yet far-reaching and general consequences for how SLA penalties need to work. Next, we use a survey of typical SLA constructs to demonstrate experiences we and our vendors have gained from offering and driving SLA's. We show how many of these elements are visible in the "SLA equation."

Although this paper is written both to providers and to customers, these parties tend to have characteristically different perspectives. We discuss both their characteristic differences and their compromises in these multi-party negotiations. If successful, such negotiations result in a formal agreement on problem prioritization, enhancing cooperation between provider and customer, and typically improving service. Thus, though the tug-of-war takes on predictable dimensions resolved necessarily under considerations such as costs, competition, and market clout, we emphasize the mutually desirable outcome of a conciliation.

A particularly exciting topic is the new generation of web services debuting this year. Web services include very rich user interactions, often more complex than a linear sequence of content elements. User experience for a given service can be interactive, and can have unclear boundaries to other services. Each of these cases underlines the need to answer such basic questions as "What constitutes a transaction?", or "What performance metrics really capture what users care about?" <<<CERTAIN CONTENT FOR THIS TOPIC CANNOT BE DISCLOSED AS OF 9/11/2001, BUT SHOULD BE SUBMITTED SHORTLY IN A REVISED PAPER.>>>

Finally, we situate the current state of Internet SLA's not just in relation to its history, but also in relation to emerging trends. By anticipating the direction of change, providers can better tailor (and time) offerings, and customers can better assess currently available value.

How SLA's exert force

> *I was once asked to speak at a large company where management typically signed SLA-free contracts with Internet service vendors. During the term of a contract, operational staff for the company would attempt to remedy questionable vendor performance by asking the vendor to agree to an SLA. Since the contract was already in place, the vendor was not motivated to expend additional resources or to lose income from SLA penalties, and so reacted to the SLA request at best with amusement.*

SLA's as a product feature

This story illustrates how little service levels are enforced during an SLA-free contract. Most customer leverage on the provider occurs during contract negotiation, after which vendor performance is less predictable, and is incented mainly by vocal complaints or much slower-acting concerns about contract renewals or legal pressure.

The tight influence of SLA's on product performance reveal SLA's as a clear product feature, usually commanding a price premium over the same promised offering without rigorous enforcement.

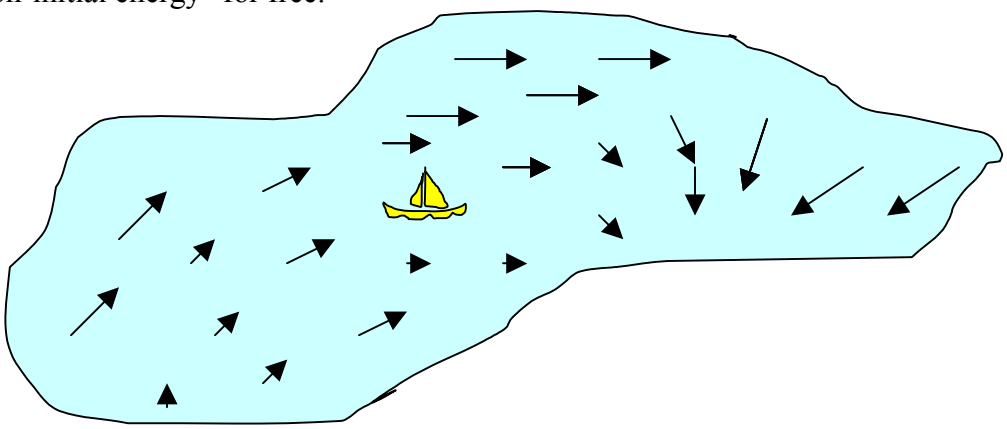
SLA's as a determinant of a business relationship

While providers of services might prefer to avoid SLA accountability, such evasion diminishes excellence (and thus long-term competitiveness.) Moreover, in the more cautious market of 2001 (driven largely by more experienced customers), it has become much more important for provider performance to be measurable, and for payment to vary with performance in cases of variability. Thus, it is in the interests both of providers and of customers to enter into the qualitatively different business relationships where an SLA provides ongoing traction on provider performance.

Within the context of these relationships, we next focus on how the motivational force of SLA penalties acts to affect performance.

Surprising ROI: a free energy field

In a sense, SLA penalties work like an energy field (analogous to the pattern of wind blowing over a lake) that stays active for the duration of the business contract. In this analogy, one can imagine that one gets to set wind direction and intensity at each point once and for all, and that one needs to pay no costs to keep these going. Thus, one obtains the non-initial energy "for free."



As we will see, the total energy available (the "*penalty budget*") does have to be paid for, as is established by negotiation. However, this paper explores how SLA penalties can exert fine-grained control over where to target the energy.

An exercise in communication

A young, retired dot com-er once advised me: "*one great way to make money is to do what other people aren't willing to do.*" An obvious example that comes to mind is the often laborious SLA planning discussion within a provider company that must reconcile conflicting points of view and conflicting interests. In fact, SLA planning is quite unusual in surfacing these differences, in that a legal contract is quite a bit more rigorous than product plans and deadlines, or even the typical level at which inter-corporate negotiations occur.

Common differences are outlined below in the section "game theory within providers."

The natural trajectory of such a discussion is to settle on a lowest common denominator, which generally includes no substantive SLA. Alternately, unachievable goals may come down from upper management, leading to failure and smaller guarantees at the next iteration.

If there is sufficient interest or customer pressure (or managerial pressure) to offer more rigorous guarantees, the high perceived cost of these guarantees tends to promote very efficient prioritization and collaboration. That subset of providers willing to lock themselves in to rigorous guarantees is then highly likely to also optimize their process – hopefully by averting penalties rather than by reacting to them.

Thus, the mere existence of a rigorous SLA is a strong indicator of a provider's maturity and effectiveness.

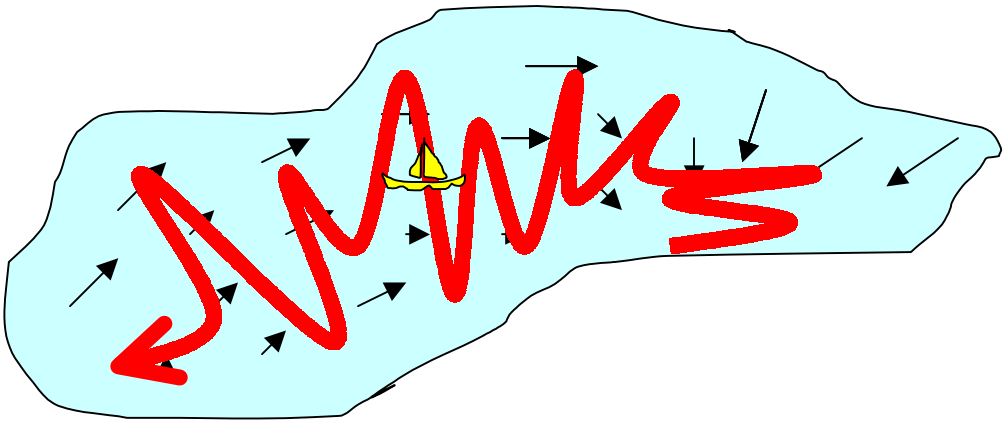
The large effect of nagging little penalties

Part of the surprising effectiveness of even small SLA penalties derives from their *consistent* action. Similarly, in the absence of strong water currents and intentional sailing in other directions, a boat blown by even a mild wind eventually moves in the direction blown.

In order to take advantage of this small but steady force, we recommend that SLA penalties should be assessed at relatively small time periods – preferably daily. In the case of a quarterly penalty period, for example, initial motivation is diluted by the impression that there is still plenty of time to fix problems, and later motivation may be diluted by a notion that it is too late to compensate for prior poor performance.

Circumventability

The wind/sailboat analogy for SLA's is problematic in that the speediest sailing is partially upwind, as illustrated by the red path below:



Similarly, poorly thought through SLA's may have unintended fulfillment strategies highly detrimental to the other contracting party. The following are all real-life examples:

- A host makes a web site so unattractive that the user load diminishes to what the server can handle.
- Given specific download performance targets, a content provider + host strips down content without bothering to optimize delivery. This results in a sparser user experience than necessary.
- In an SLA that allows scheduled downtime, a provider decides to "schedule" downtime whenever the site has crashed.
- Given a 100% SLA penalty on the host based on specific performance targets, a customer wishing to avoid payment makes the content so gigantic that it cannot load in time.
- In a case where the provider (or customer) houses the SLA reporting, they face a conflict of interest between honest reporting and avoidance (respectively exaggeration) of penalties.

In each of these cases, the sailing "upwind" is more dramatic, but counterproductive to the cooperative goal of the SLA. Generally, effective SLA design includes a Machiavellian thought experiment: "*How could the other party take advantage of this*?"

As a legal note, in the United States, intentional setting of traps is often grounds for their dismissal from contracts. However, evasion of unexpectedly disadvantageous terms to a party is much harder if they cannot prove they were intentionally misled. Thus, there is no substitute for due diligence in considering ways to cheat the system. In our sailboat analogy, one would want to confine the path of the boat to narrow channels, avoiding the zigzagging that allows upwind sailing.

For the rest of this paper, we will assume that SLA's are designed with enough foresight that they are not easily cheatable. However, this still leaves open important strategies for providers to *circumvent* SLA penalties:

- improving performance to a level where penalties are extremely unlikely
- offering only tiny penalties, and not on the most uncertain outcomes

The chaos of exaggerated penalties

When potential SLA penalties become large relative to expected income, the revenue model truly approaches that of an insurance underwriter. Given that only partial movement in this direction is uncomfortable to most Internet providers, one can see how such a situation tends to undermine a provider's good will.

Such large payouts may occur for example when a web host is asked to indemnify for lost business, or more generally when a small provider services a much more powerful customer. This should be a particular concern in the increasingly litigious Internet-related business climate of 2001.

This is one important example where circumvention becomes the only escape. We have seen providers attempt to interpret the high-risk SLA language in such a way that the measurements became irrelevant, but achievable.

One should understand that even reasonably solid SLA's may often be stretched at least partially. Thus it is important to promote the good faith of both parties in finding a mutually acceptable compromise.

Building a theory of SLA's

Before we survey common offerings or make specific recommendations, we take a step back to build a unifying framework for how SLA's operate. This extra work rewards us with a simple framework for viewing the business processes and legalese documents that often become highly contorted.

The penalty budget: an SLA's main leverage

The worst-case penalty is a "**volume knob**" that determines the entire motivational budget of the SLA. We define the **penalty factor** as the fraction of service cost that could potentially be lost as an SLA penalty. In fact, the penalty factor may be fuzzy, because SLA penalties are often paid in other units than dollars (such as free bandwidth or additional service usage.)

A penalty factor of zero is just an SLA-free situation. In general, providers hesitate to have their entire revenues up for grabs, and so the penalty factor is typically strictly less than one.

On the other hand, from the customer's point of view, absolutely abysmal service should be penalized even more strictly than by a full refund, since in that case not only has the customer's received nothing of value and not only has their business credibility been damaged, but they have had to invest resources to obtain this poor service. Thus, customers might prefer penalty factors *above* one. These usually take the form of indemnity for lost revenue. Powerful customers certainly put the provider on the defensive when the latter propose penalty factors less than one.

We will have more to say later on how this volume knob is negotiated, but for now, note that it allows a quantitative re-negotiation without the overhead of having to re-implement the entire SLA.

In our wind analogy, the penalty budget correspond to the sum of all the energy blowing over the entire lake. The wind analogy raises both downsides and upsides of SLA's. One disadvantage is that the penalty budget represents a *fixed maximum* that must be apportioned wisely among all possible scenarios.

On the plus side, unlike weather uncertainty, an SLA codifies penalties once and for all. Thus, even if the total energy behind the SLA had to be negotiated, there is no expense to keep this energy in force.

Three main ingredients: utility, probability, and cost

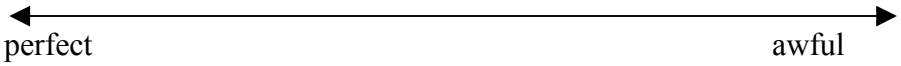
If we assume for the moment that SLA's are engineered carefully enough to limit circumvention, we can describe a theoretically optimal solution, first in terms of a single performance measurement "x", and later more generally.

Utility

Many researchers in economics and organizational psychology quantify costs and benefits of outcomes using *utility* metrics. Besides capturing a valuation of dollar values, these may include such other factors as goodwill, perceived convenience, likely effect on future benefits.

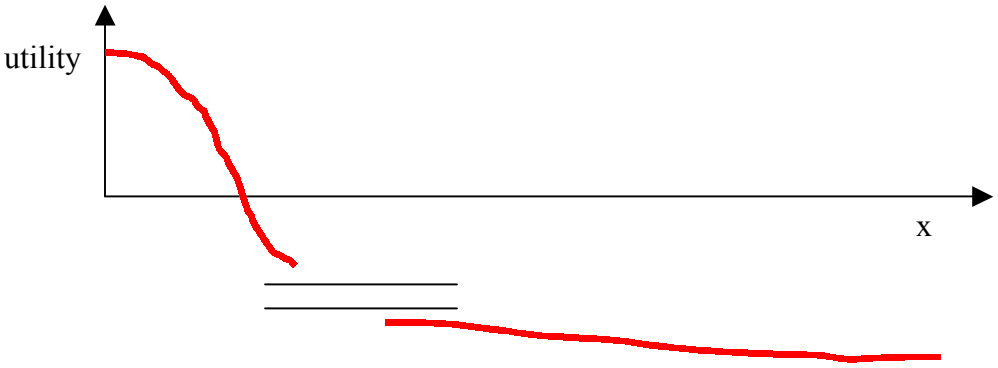
The concept of utility captures non-linear valuations of money. For example, most people would find ten million dollars to be less than "ten times" as valuable to them as one million dollars, since the lifestyles made possible do not seem that qualitatively different. Conversely, many people are willing to buy lottery tickets (with negative expected earnings), because to belief in the possibility of winning outweighs the negative dollar valuation.

Assume a single variable ("x") that ranges from perfect to as bad as imaginable:



For example, the variable x might be "unavailability", ranging from 0% to 100%. In general, we will use the term "zero" to refer to the perfect value of x.

One may graph the expected utility of a service as it depends on the input variable x:



Several features of this graph are worth noting:

- The utility corresponding to optimal performance (the y-value for x=0) is often confused with the actual utility the service will have. Instead, this should serve as a starting point, from which price should be negotiated down.
- Small enough values of x are effectively undetectable, and so need not be penalized. We have seen SLA's attempt to guarantee perfection, and then respond to the unachievability of this goal by removing penalties. The utility curve shows that this is unnecessary.
- As x increases, eventually we reach a point of neutrality, where performance is so bad that any value of the service is equalized by performance problems.
- As x increases further, utility becomes negative, and may become several times as negative as the optimal value for x=0. The two black lines show that the utility curve is interrupted.

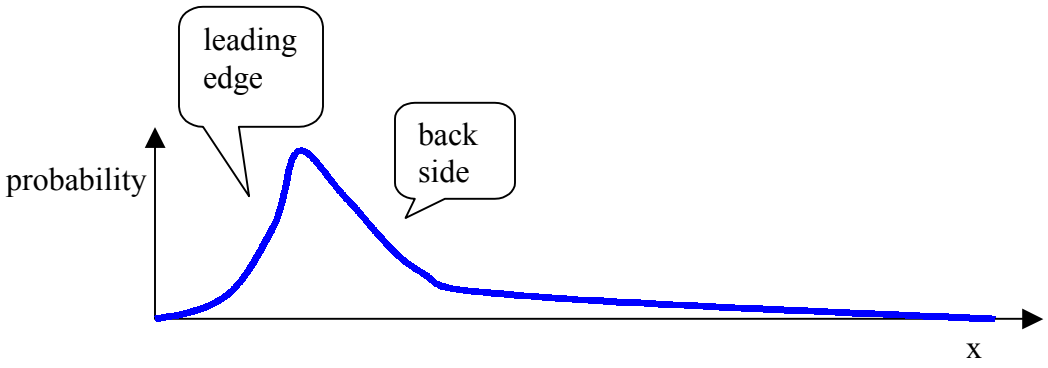
Whose utility function is this?

Suppose a web host believes that full utility is achieved if they internal pings of their server have never failed. In fact, perhaps the pings only occur once an hour, and perhaps the Internet connectivity is so poor that users experience substantial outages. In this case, there are different candidate utility functions between provider and customer. In using the theory that follows, both parties must consider utility from their point of view, in order to decide whether a given SLA is appropriate.

In fact, it may be the case that metrics proposed by the provider account for only a few percent of the utility experienced by the customer. In this case, a corresponding penalty budget for these metrics is less relevant, and thus need not be bigger than a few percent of revenues. However, an informed customer would want to know how to cover all the rest of utility in an SLA.

Probability

Besides considering utility, one must consider the *probability* that the service will actually perform at a particular level:



While the probability curve will very with the situation, and even over time, certain common features are worth noting:

- In general, x will never be perfect. So while perfect performance is a wonderful intention, it has less use as an SLA objective.

- The probability curves for performance tend to have the shape shown: asymmetry, with a heavy right tail. Generally, the median (50^{th} percentile) sits near the middle of the main hump, but the expected value (the mean) is way off to the right, and depends in a very unstable way on outliers.
- The "leading edge" – the slope where probability increases – tends to be relatively steep. It represents the area where providers can move to as they improve their service.
- The "back side" – the slope where probability decreases as performance degrades further – tends to be more drawn out. It reaches all the way to the worst possible performance.
- While the points far off to the right are unlikely, they are important, since they are disasters with very low utility!

The discussion of probability raises some crucial questions. First, one is in no position to negotiate a reasonable SLA without some understanding of the probability curve. Not knowing it, one could budget all one's energy in a part of the curve so unlikely to occur that the SLA will not incent better performance. Thus, understanding of historical performance (and reason to believe that this benchmark predicts future performance) is part of the due diligence for SLA design.

Whose probability?

Given that probability is an attempt to predict the future, one must consider that a providers usually have better information, so that their probability curves tend to differ from that of customers. However, the nature of the difference is quite ambiguous. On the one hand, a provider wants to instill confidence, by shifting the perceived curve to the left (the "marketing impulse".) On the other hand, as we see below, this encourages stricter SLA's. Thus, the provider's communication of the probability curve requires *management of expectations*.

If SLA penalties are weak, the marketing tends to predominate, resulting in a disconnect between advertisement and promise. This is the typical situation, and is an excellent measure of market immaturity and the distrust one should confer upon advertisement. Thus, one should not compare SLA's to marketing statements, but only to other SLA's.

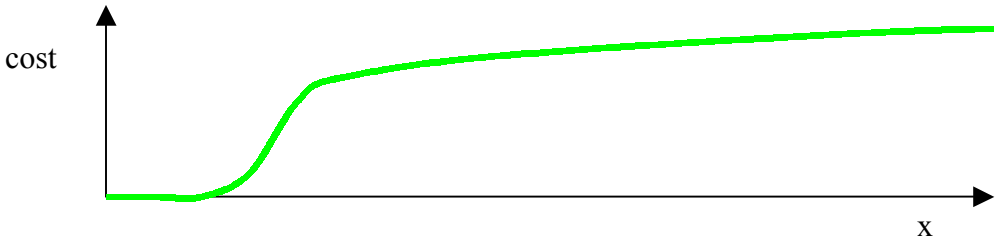
On the other hand, a customer's skepticism, while lowering perceived value, shifts the probability curve to the right. Once an SLA has been signed, this is great for the provider, who then has an easier job meeting expectations.

Probability changes over time

Much of the Internet industry is still to immature to guarantee end-to-end performance. New technologies have had drastic effects on performance over the last two years, and the changing commercial landscape argues for periodic revalidation of SLA assumptions. We recommend currently that probability curves (and hence SLA penalty functions) be redrawn once or even twice a year, until performance stabilizes.

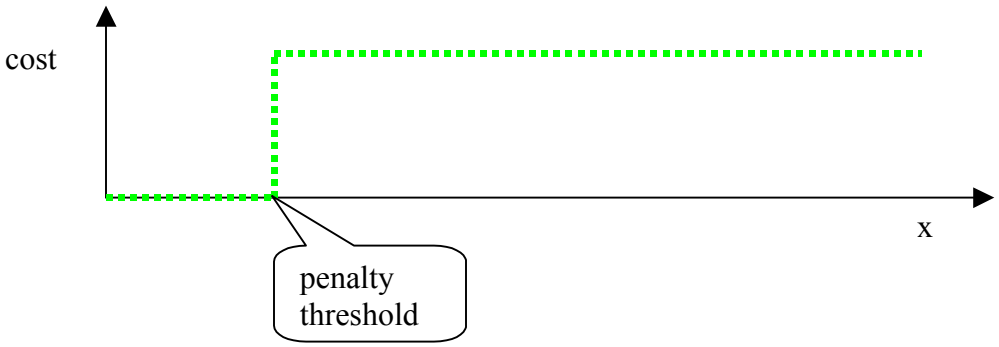
Cost

We now arrive at the actual SLA penalties – termed "cost", so that "p" stands unambiguously for "probability." Before we explain how the graph below arises, let us point out some important features:



- The cost function increases monotonically with x.
- As x approaches worst case, the SLA penalty approaches the penalty budget. This is the sum of all the motivational force built into the SLA. Given this strict limit, one must distribute it wisely among all possible values of x.
- For small x, the cost is zero until a certain "threshold" is crossed.

Contrast this graph with the usual kind of **step function** one sees in SLA's:



In this case, no penalty accrues until one crosses the penalty threshold, at which point the full penalty kicks in. The problem with this is that it provides little motivation to improve service until one nears the penalty threshold. Once performance exceeds the threshold irrevocably, all hope for avoiding the penalty is lost, again failing to motivate improvement.

Obviously, such penalty functions are chosen because they are easier to understand and to implement, but we argue that such excessive simplicity undermines the motivational purpose of SLA's.

The univariate case: how SLA's could work if only one thing mattered

So how do we get the cost function?

To answer this question, consider that there are two kinds of reasons for allocating part of the penalty budget at a particular value of x:

1) There is a high probability that this value of x will occur, so we want the penalty to ramp up there in order to motivate improvement.

2) As x increases in this region, perceived quality (utility) decreases steeply. We want to counter this with an increase in the cost function.

The single-variable SLA equation follows immediately when we recognize that the concentration of SLA motivation at a value of x is none other than the derivative cost'(x) with respect to x:

$$\text{cost'(x)} = k * \text{probability(x)} * \text{utilty'(x)}$$

We will use this equation to understand our SLA experiences.

Again for utility, the apostrophe (') indicates derivative with respect to x.

To solve this differential equation, we need to specify an initial condition in order to set cost equal to zero when x is perfect, say when $x=x_0$. In our example, we said this happened at x=0. In this case, we can solve the equation to obtain:

$$\text{cost}(x) = \int_{x_0}^{x} k * probability(u) * utility'(u) du$$

If we let x get as bad as possible, then cost(x) should approach the entire penalty budget. To make this work, we can adjust k appropriately. Notice that a doubling of k, for example, would double all SLA penalties. So alteration of k gives a very simple way to change the size of the penalty while leaving the remaining SLA machinery in place – so k is a constant proportional to our "volume knob" of penalty budget.

The definitive SLA equation: the multivariate case

In general, several different variables will matter to the customer, in which case x becomes a multidimensional vector, and the **SLA equation** generalizes to:

$$\nabla\text{cost(x)} = k * \text{probability(x)} * \nabla\text{utility(x)}$$

Here, ∇ denotes gradient with respect to x. Again, k is a constant adjusted to fit the penalty budget.

Unfortunately, this more general equation is not always solvable. Also, it's bad enough to try to foist single variable equations on industry, so we would like to reduce the more complicated situation to single-variable cases.

It turns out that this is not so difficult, at least up to approximation. If one is able to separate behavior of interest into "root causes", these tend to be relatively independent. Then, one can consider each on their own. To do this, one must first divide up the total penalty budget into pieces for each root cause, and then solve the separate single variable problems, each with their own utility and probability curves, and each with their own constant k.

For example, some streaming media SLA's we have helped build are based on the five metrics or availability, startup time (i.e. before playback begins), rebuffering behavior, reported client-side packet loss, and other gaps in playback (generally implying a server-based omission of packets in order to budget resources.) It turns out these metrics are not at all independent, but nevertheless, their treatment as such provides a workable approximation.

The case of discrete measurements

Many variables of interest are not continuous, so that calculus does not seem to imply. For example, a single measurement is either available or not available. Given that any measurement service can only take an integral number of measurements, strictly speaking, one cannot use derivatives. However, the *fraction* of unavailable measurements is a number between zero and one specified in principle to an arbitrary degree of precision, depending on how many measurement are used to calculate it. Thus, we may use derivatives as an approximation.

However, some variables are either binary (e.g. a task was or was not performed at a specified deadline) or can take only a few number of possible values (say from one to five, for the number of incidents during some period.) In this case, the equations above will not work, so we resort to the discrete analogue of derivatives, namely differences:

$$d(cost(x)) = k * probability(x) * d(utility(x))$$

In this case, "d" is a discrete difference. Suppose the utilities for zero and one incidents were "10" and "7", respectively. In this case, in the equation above, we would have:

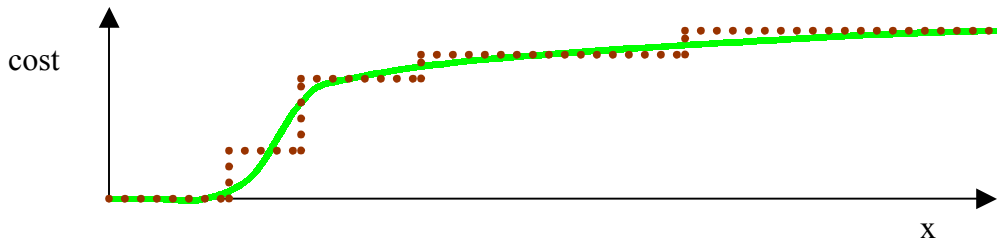
$$d(utility(0)) = 0$$

$$d(utility(1) = 10\text{-}7 = 3,$$

and we would solve for the cost function accordingly, obviously penalizing more for one incident than for zero.

What do you do in real life?

Regrettably – and without justification, we might add – much of the business community is as hostile to calculus as year 2001 venture capitalists are toward start-ups selling pet food over the web. However, we advocate the differential model for the cost equation mainly as a thought experiment. In reality, it is probably too unwieldy to calculate day-to-day penalties. Instead, we propose using a combination of step functions to approximate the ideal cost function. If the practical cost function has cutoffs at know points and always returns an integer percentage of the total penalty budget, one has a more palatable compromise between accuracy and transparency (as shown in brown dotted lines):



We emphasize that motivation is brought by the SLA only in the vicinity of cutoffs (vertical jumps in the graph above), so the most crucial thing is to spread enough of these over the domain of interest.

What the theory tells you to do

As stated, many implications of this theory agree with our experiences in helping to write (and being subjected to) various kinds of SLA's

Below, we discuss a whole "game theory" that both sides can use to their advantage. First, we offer just a few general comments.

Cover your back side

Recall that we call the worse side of the probability ramp its "back side" – mainly in order to allow the memorable title for this section. Suppose you as a customer believe your provider to be near perfect, and accept a probability curve concentrated near x=0. Perhaps this was based on performance benchmarks on unloaded servers, and servers are unable to handle a production user load. Thus the real probability curve is further to the right than expected.

If you have negotiated a penalty factor of 100%, both you and your provider are miserable now. However, if the penalty factor is only 10%, say, your provider may simply take that loss, since the penalty is mostly unavoidable under current circumstances.

This discussion highlights the need to form a conservative probability curve, particularly for the worse scenarios you hope to avoid. This will spread enough of the penalty budget to worse performance that the provider will be incented to move beyond this area and thus reduce penalties very substantially (see this strategy in the section "game theory for providers.")

Capture as much of utility as possible

Although this is not a direct consequence of the theory, it follows from the stated goal fo the theory, namely to provide the optimal influence toward increasing utility subject to the penalty budget limit.

All to often, we encounter situations where a customer use a service with an optimal utility corresponding to $200K per year. Suppose this service is measured by yet another expensive system, but that due to limitations in accuracy, optimal measurement might predict only an annual utility of only $150K. In this case, the customer has a full 25% of utility unaccounted for, and thus there is no incentive for this to be optimized.

In reality, it is quit impossible to capture 100% of utility through measurement. Practically, measurement itself has costs, and provides only partial information. However, a stronger theoretical limit applies, in that full measurement would add so much overhead though cost, privacy intrusion, and performance hits, that a sort of Heisenberg uncertainty applies:

measurement overhead * uncertainty in utility > minimum,

for some "theoretical" minimum. Unfortunately, this minimum seems much bigger for Internet-based services than it is for mainframe applications.

If one understands that everything is understood only in approximation, it is much easier to accept the role of our theory as a conceptual approach rather than as a complete recipe.

However, certain important techniques contribute substantially to improved accuracy. This topic itself deserves a paper, and we only summarize some highlights:

- Unavailability (unreliability) must be understood in the context of measurement errors. For example, a measurement service has its own unavailability (whose amelioration follows Heisenberg-like uncertainties as above.) A provider's unavailability is best not *defined* absolutely (since this is not measurable), but rather as a fraction of valid measurements.
- Measurements have limited meaning without an understanding of their confidence intervals and sources of either systematic bias or else merely sampling error.
- The most obvious aggregate (arithmetic average) is numerically unstable for web page download times, and should be substituted with a more robust statistic such as geometric mean. The author has heard an urban legend about this mistake having cost a vendor a $1M SLA penalty! Due to their effective *squaring* of heavy tail distances from the mean, standard deviations of web download times are even much worse, and should be expunged from humanity.
- Internet-based SLA's should weight server availability by the relative importance to users over time (and over location, for geographically distributed servers.) One obvious example is the online brokerage.

Real-world experience

We share some practical advice – understandable with or without the theory.

How SLA's get built

<<TO BE COMPLETED>>

Legal suggestions

We receive frequent requests for boilerplate SLA text. This is particularly difficult for us, since we consider our NDA's with customers to protect the legal language they reveal to us. Such language is highly particular to each contract, and like software code follows a reasoning of its own. We expect common patterns to become more standardized, at which point we might pass them on, presumably upon receiving signed liability waivers.

Such difficulties notwithstanding, we do offer the following grab bag of suggestions, which of course are presented as is and without assumption of any liability.

Legal language differs from usual English. The intent is not to facilitate understanding, but rather to make misunderstanding by a "reasonable person" difficult. Thus, writing legal language resembles trying to write uncrashable program code. This may help justify contortedness of legal language.

Unlike for normal prose or even for technical documentation, succinctness and enjoyable reading are *much* less important than accuracy

Besides expressing intent, one needs to cover all important conceivable possibilities

While lawyers craft language very precisely, their work is difficult if *intent* of SLA clauses has not been made evident prior to their review. This suggests the following sequence of events:

1) Reconciliation between Marketing and Engineering goals
2) Encoding in precise, but maybe not legal, language

3) Legal review

Steps 2 and 3 may both be performed by suitably experienced in-house counsel; otherwise this can become quite expensive. Especially for larger contracts, vendors should seek qualified legal help before placing themselves at risk

Approach document with appropriate standard

- When writing document, ask: “could a reasonable person fail to see this?”, not just “is this true?”
- When reading document, ask: “is guarantee expressed clearly enough to stand up in court?”

Unlike smoothly flowing prose, define a term once, and always use the same boring words to refer to it. Different terminology could allow an opposing party to question whether something else is being referenced.

Avoid any unintended ambiguity.

The outcome of court challenges depends substantially on state law and legal precedents. One may assume unneeded risk by signing away the right for jurisdiction in a familiar court. One must know what laws apply, and how they have been interpreted. This is particularly important if the relevant jurisdiction has a smaller case history, for example, than California.

A legal challenge is generally bad publicity (at best), so only choose provisions you are willing not to contest. This means that legal documents must limit hyperbole much more strictly than marketing materials. Thus, an SLA might be good occasion to work out differences between hopes and reality.

Game theory for SLA’s

Having seen several stalled SLA negotiations, as well as several successful negotiations that survived frustrating obstacles, it becomes crucial to participants in SLA negotiations to arm themselves with ways to facilitate their positions (or stall them, as the case may be.) This is particularly true given that the “games people play” are so characteristic, despite veneers of confidentiality.

The subject of game theory

Game theory evolved within mathematics as a way to describe outcomes obtainable from certain starting positions in games with definite rules and definite desirable outcomes, but with or without an element of chance, e.g. blackjack and chess, respectively. In some cases, very advanced progress has been made (such as a computer world chess champion), but in other cases, fundamental results are still quite elusive (such as whether the game of chess is winnable for white, i.e. whether white is guaranteed a win if he plays perfectly.)

The term game theory is applied loosely (and somewhat comically) to situations like SLA’s where optimal play for each party depends in characteristic ways on the “hand they have been dealt.”

In each case, we outline common situations, and suggest possible actions.

Strategies for the customer

A customer's goal is to maximize value obtained

SLA's are a tool for doing this, but SLA penalties are often misconstrued as a game outcome rather than a strategy. Certainly the existence of an appropriate and cost-effective SLA is a desirable outcome, since it correlates with better product value. However, it is quite short sighted to bury the provider in penalties, since providers in this case will not be motivated to continue the game. One cannot keep winning games when the other player walks out.

Thus, the customer's sight should remain on the expected (utility + SLA cost), not just the cost. This expectation should take into account realistic understanding of the circumventability and terminability of the SLA. Even under our assumption of a rigorous SLA that cannot be faked, once the provider's main motivation becomes the evasion of the SLA, the provider's mind share has been lost from the cooperation the SLA was supposed to facilitate.

Again, cost is a way of inducing a higher expected utility. If this goal becomes lost, what is the point of the contract in the first place?

Inertia is not the customer's friend

As we have pointed out, providers are reluctant to assume responsibility for uncertainty outside of their perceived domain of control. Also, human nature disinclines providers toward giving up even a penny of revenue. Also the pace of technological change in IT has diminished very substantially, and market consolidation is leaving a shrinking choice between providers. Thus, the ball is most frequently in the customer's court to seek out and drive the most favorable SLA's. If this does not happen at the time of contract negotiation, it is highly unlikely to occur later.

Benefits and difficulties of customer-driven SLA's

A customer-driven SLA is much likely to reflect the customer's utility function. Since the customer's utility function most accurately reflects the perceived value of the product, this implies an inherently more *accurate* SLA. Accuracy, while perhaps uncomfortable for the provider, tends to tune performance to improve contracted service levels, so a customer-driven SLA leads somewhat ironically to a more competitive offering.

We have seen cases where the initial discomfort with this process has in fact led to later similar provider offerings to other customers. Thus, a customer-driven SLA may in the long term benefit the provider, and may certainly be advertised as such.

On the other hand, customer-driven SLA's are not easily reusable with other providers, and thus generally occur only for large contracts, and mainly for large customers. This process generally has higher friction than the tweaking of a provider-offered SLA. This usually requires a longer negotiation, but generally with a superior outcome if successful.

If the power imbalance between provider and customer is too extreme, the provider may buckle under and end up with a risk disproportionate to their foreseeable revenues. In this case, the cooperation between customer and provider has failed, and the provider will seek ways to evade the penalties, such as by installing fakeable metrics. In this case, the

customer fails to derive expected value from the SLA, and so has not played their position effectively.

Strategies for the provider

A provider's goal is to maximize revenues

In cases of SLA's a short term mentality often considers the penalties independently as lost revenues. Instead, providers should understand an SLA as a package, which includes motivations for improved performance. Generally, qualitatively new SLA's are uncomfortable in the short term, but facilitate improved performance in the long term. This value should not be underestimated.

Avoiding excessive costs

On the other hand, in cases of difficulty in achieving higher performance levels, or when competitors do not make similar offerings, partial concessions may go a long way toward satisfying customers. I have seen powerful customer's SLA initiatives put to rest indefinitely by statements of "SLO's" (service level objectives, "like" SLA's , except without penalties.) Reassurance by mere statement of intentions may be quite naïve, but perhaps this improves over prior conditions!

Getting ahead of the curve

The simplest strategy that can pay off royally for the provider is a targeted interaction with the probability curve. A critical part of a provider's preparation for SLA negotiation consists of a thorough understanding of the probability distribution of their performance. However, intelligent providers engineer into the SLA product plan that ability to "outgrow" the currently accepted probability curve. In other words, while insisting on a conservative curve today, they plan for optimizations that improve their expected performance, so that the expected penalties are quite low. At the same time, the "teeth" in the SLA appear menacing enough that the customer is willing to pay a premium for the service.

This is standard strategy for insurance providers, but so far surprisingly undiscovered by many Internet service providers.

Marketing

Without intending any disrespect, marketing may be defined as the exercise of increasing the perceived utility of a product. We have seen instances in industry of grossly exaggerated benefits of so-called SLA's, which even in a preliminary reading reveal gigantic loopholes. As the Internet industry becomes more sophisticated, and as decisions are being made by more seasoned workers, we would expect increases in the accuracy of how SLA's are advertised. However, the SLA and its marketing are two different things, especially in the provider-driven case. Thus, any perceived benefits of an SLA are certainly advertised, which becomes all the easier in cases where guarantees are rigorous.

We have seen well-known companies offer quite dramatic SLA's that upon closer scrutiny promise very little. In this paper, we concentrate less upon advertisement wording (which sometimes bears little resemblance to reality) and focus instead upon contracts between two competent parties.

Competitive landscape and version control

The author has been accused of being biased toward the customer perspective in the inevitably different customer-provider standoff. In fact, providers often want to avoid rigorous guarantees, and there is large un-met customer demand in this area. However, we would be remiss to providers if we failed to point out that few customers today receive SLA coverage of anything resembling their entire utility.

Providers may use two techniques to delay gratifying customer demand. First, they may analyze the marketplace to find that competitors also fail to offer rigorous SLA's. In fact, a provider may have special arrangements with larger accounts, so very cautious market research is warranted. A provider may want to assume responsibility for just slightly more risk than competitors.

Alternately, given an assessment of how fast the market is moving toward rigorous guarantees beyond the firewall, a provider may want to ramp up SLA coverage in stages, so as to allow for a big learning curve, and so as not to be exposed to more risk than competitors.

We believe that willingness to cover beyond the firewall will catch on, although we find significant trends decelerating this change.

Game theory within providers

As seen by the customer, the provider may appear as one entity. However, almost any of the involved parties within the provider will tell you otherwise. We contrast typical different points of view toward SLA's, as well as suggested responses to common situations.

SLA's are typically initiated as a managerial level, even though their practicality will typically rest on the shoulders of operations and engineering groups, their public image will be portrayed by marketing, and their actual statement will be overseen either by legal counsel, or else by a stand in of varying aptitude.

Each of these parties ma bring different expectations and goals:

- Manager (possibly in sales): cement deal and then pass on execution to others
- Marketer: offer what customers are perceived to want
- Operations: avoid taking responsibility for difficult guarantees
- Legal counsel: protect client from risk

Often, an agreement established between these parties requires so much work as to provide extensive value. On the other hand, deals made without agreement between these parties can lead to severe problems, and undermine not just the SLA but underlying products as well. The difficulty of this process is a good measure of the efficiency within an organization.

Strategies for the third party (e.g. consultant)

Above all, a third party often has the privilege to arbitrate between two very different parties. While a disinterested third party can be of great value, typically, one of the interested parties is the primary contact for the consultant. The consultant's role thus differs greatly depending on the primary client.

Along with the client comes the attitude toward the SLA: providers may want no teeth, and customers may want indemnification (reimbursement for lost revenues.) As a consultant enters into a role of SLA negotiation, it is critically important to have credibility to both sides, and to understand (and not foreclose) the differing attitudes of both sides.

Why the world continues to go on despite these conflicting interests and behaviors

Just as SLA's have become prevalent in more established forms of business, and just as insurance is widespread in industrialized countries, the benefits to long-term provider performance and to diminished customer apprehension should counteract cultural obstacles to SLA adoption.

Web services

<<<NOTE: CONFIDENTIALITY AGREEMENTS PREVENT US FROM DISCLOSING CERTAIN CONTENT FOR THIS TOPIC AS OF 9/11/2001, BUT WE SHOULD BE ABLE TO FILL THIS IN WELL BEFORE THE CONFERENCE IN A REVISED PAPER. BESIDES ADDING MATERIAL IN TE SECTIONS NOTED, THE REVISED PAPER WILL ALSO REPLACE MORE GENERAL EXAMPLES BELOW WITH CONCRETE EXAMPLES FOR TYPES OF PRODUCTS.>>>

After most of the anticipated revenue from Internet services has failed to materialize, certain large vendors are launching new architectures to facilitate a complexity of web interactions on par with the complexity of application behavior on a single machine. By virtue of Internet communication lying at the core of these so-called "web services", the interactivity across machines and over networks becomes much richer. As a by-product of this richness, the uncertainty and heavy-tailed performance of the Internet become problematic in qualitatively new ways.

The new products

<<<TO BE ADDED>>>

New problems of measurement

Besides heightening the usual Internet uncertainties, we survey new issues that reach critical mass with web services.

Boundaries between overlapping entities

Much of the richness of Internet experience derives from a user's ability to follow as many as dozens of links out of a given web page. With web services, this becomes all the more multidimensional, since servers' behaviors are determined not just by a target page, but also by the dynamic content that depends on multiple user and/or environmental variables.

More critically, multiple web services might operate simultaneously on the same browser "real estate." In fact, web services may intersect with each other, so that SLA's may have very unclear starting points.

Influence between multiple levels

Web services allow more highly tiered offerings, where involvement may not be restricted to adjacent tiers. For example, suppose provider A offers a customer referral service and contracts to refer customers to provider B. Part of B's value derives from content originating from provider C, but housed within B's offering. In this case, B contracts with C, but C's performance (i.e. the value to A's customers) is of direct relevance to A.

While such phenomena already occur in web transactions, they assume a higher relative importance for web services. This becomes especially true because in web services, the number of "C's" may be higher than ever before. Somehow, A desires a reliable way to obtain suitable usage knowledge about B's customers, but without intruding on B's business and without having to believe everything B says. Also, while B and C may negotiate with each other for their contract, such negotiation may fail to reflect the relative value of these two parties, if the user traffic for both is driven primarily through A, and thus depends on the nature of users who are persuaded to jump to B.

It seems difficult to design accurate SLA's in such cases without reporting by a trusted neutral third party. This in turn requires a trusted mechanism for information passage beyond immediate business partners.

Time asynchrony

Suppose provider A requests a service from provider B. Assume that from provider A's point of view, thirty seconds elapse. However, provider B claims that the request was serviced within ten seconds. Who is lying?

In fact, both providers may be telling the truth, as measured according to their own clocks. Internet latency is the culprit, and as usual, provider A does not want to take responsibility for it, and provider B wants performance to be measured as he experiences it.

One might attempt to reconcile the times by agreeing to send along clock information at each of the four steps: request-sent, request-received, response-sent, and response-received. Unfortunately, it still seems difficult to agree on a sufficiently accurate time standardization – GPS is certainly one possible approach.

It seems that the only workable solution right now is to build in a certifiable timing mechanism (certifiable in the sense of security mechanisms), but then only to legitimize differences of times measured at one location. In this case, two way Internet latency would be treated for SLA purposes as the difference:

$$\text{provider A lag} - \text{provider B lag}$$
$$= (\text{response-received} - \text{request-sent}) - (\text{response-sent} - \text{request-received}).$$

Note that one might suspect each party of stretching (or shrinking) the truth. This raises interest in synthetic measurement and/or embedded time reporting by the provider of the service infrastructure.

Finally, consider a situation where more than two parties are involved, and each provider's function is not just "embedded" within the next higher. In this case, there is a strong desire to approximate "absolute time", which raises suspicion if conducted at any of the primary interested parties.

Too many "moving parts:" p(A ^ B) = p(A) * p(B)

The successful completion of a web-based transaction, say of logging in and entering a data for one participant in a group activity, is a remarkable success. Consider that chains of up to dozens of routers, web servers, and databases may have been involved. With web services, this becomes much more complicated, since those most failure-prone operations now become distributed. For example, a centralized groupware application (perhaps serviced by redundant expensive databases) now gives way to dozens of databases flying around the world on people's laptops.

Suppose that twenty people's calendars each have failure rates of 2% (very optimistic, in terms of current Internet performance.) In this case, the failure rate of the attempt to schedule them together is

$$1 - (1 - 2\%)^{20} = 33\%.$$

One case see that one's expectations will have do decrease even further than from mainframe to Internet levels. Instead, SLA's for such a service would have to base expectations on specified numbers of users – not the level of detail at which management decisions typically occur.

A glimpse of emerging solutions

<<<TO BE FILLED IN>>>

Conclusion: the state of the SLA industry in 2001

Several factors have interacted in 2001 to build critical mass for rigorous SLA's for Internet-related services:

- A profound change in business culture toward cost-justification
- A new willingness of service providers to act as insurance providers in sharing risk with customers
- A greater representation of mainframe expectations among Internet-service providers with a resulting effort to help drive Internet performance closer to mainframe expectations.

However, SLA's past the provider's firewall require a significant change, so that even if this change is almost inevitable (as the author believes), it will be prolonged. Further prolonging this change are the increased consolidation and conservatism in industry.

However, we feel that the suggested method of analysis in terms of utility, probability, and cost provides a common starting point for negotiations, wherever these may lead. We predict substantially new kinds of SLA negotiations on the horizon, especially with some of the more complex Internet-based services facing imminent release.

References

<<<TO BE FILLED IN>>>