A Proposal for Next Generation Network Requirements

Jason W. Rupe, Ph.D.

14900 Conference Center Drive
Suite 280
Chantilly, VA 20151
703-955-7770

Introduction

The intent of a service level agreement (SLA) is to provide a level of assurance to the customer about the features and performance of a service which they purchase from the carrier or provider. Today, most SLAs are focused only on the network capabilities.

An operational level agreement (OLA) is similar to an SLA, but internal to an organization, and usually works bi-directionally. Like an SLA, an OLA is ultimately linked to the features and performance of the services sold to customers. When done well and effectively, the operations are improved, which enables better services and better costs.

It is reasonable and best to manage the network to optimally enable the delivery of the SLA, and continuously improve on it as well. This framework proposes an approach to do so. It is based on the service performance instead of the network performance, ensuring better alignment with the services and transactions being delivered today.

The life of a service

A service is intended to meet its requirements all the time, but realistically will fail to do so at times. Network capabilities work to minimize this time in diminished states. Some mechanisms for doing this include redundancy, protection, restoration, load sharing, and repair. These terms have specific meanings within network management and technology, but the terms have bled into the services discussion.

However, a more careful translation of these ideas across the value chain is not only possible, but warranted. With this careful translation, we can better design solutions to customers' uses. By first recognizing that service protection and transaction protection are not necessarily the same as network protection switching, we can begin the discussion of how to translate.

Start with recognizing the states of any transaction or communication service. There are the states of fully functional, and completely failed, plus any number of states in between. For a particular transaction, it is mostly in a working state. But something in the value chain can fail, putting it into one of many transitional states. Depending on how long the transition takes, if it impacts the transaction adversely, the transitional state could be considered to be a protection, a recovery, a restoration, or a repair.

- A protection might be considered to have no loss of the transaction state, and no substantial impact to service.
- A recovery might be the case of partial loss of state, but partial recovery of the transaction so that some back tracking is needed, but the transaction can continue after the recovery process completes.

14900 Conference Center Drive
Suite 280
Chantilly, VA 20151
703-955-7770

- A restoral might happen from a complete loss of state of the transaction, but the transaction can begin again immediately or quickly.
- A repair might happen when there is loss of the ability to continue the transaction, and it can't be restarted for some period of time, so once the repair of the capability is complete, new transactions can be initiated, and the given transaction can attempt again.

Likewise, there are equivalent service state definitions we could create. As these are each transitional states, the model they form can be used to indicate the transaction or service state over time as it changes. It can be defined so that each transition state only returns to the working state when the transaction or service resumes as defined by the transition state, or thought to transition from bad state to worse state until the transaction or service resumes as defined by the last transition state.

Based on what has been outlined, a service provider can define transaction or service states, with translations as described above for tracing network events to service and transaction events. We can achieve at least two key advantages: 1) define network capabilities in these terms so that users can decide for themselves whether a service can meet the needs of their transactions or services, and 2) define translation functions that allow reporting of network performance in terms of transactions and services for useful criteria.

Network Capabilities

To support this framework and enable users to select products that will serve them well, the SLAs must be defined in terms of a combination of traditional performance measures and the translation functions that allow connection of the service and transaction requirements they have (but may not articulate) to the service behavior in related terms. There are options for doing this that will be clearer than the general term *translation function*, and useful solutions will be easy to understand.

Option 1: Specify an intended transaction type and recommended architecture. This approach allows the seller to offer a product defined by a precise performance level that makes sense to that market, with meaningful service or transaction state probabilities and durations. This option might appear as a statement of latency, jitter, and other performance guarantees to be brought into the definition of service availability, along with tiers of guarantees for outage duration. The guaranteed level of performance helps define the service availability, and that guaranteed availability relates to the intervals of outage duration provided as well. The outage breakdown would be defined in groups according to service or transaction impact. The following duration results may be offered as an example.

- Protected events, lasting 50ms or less, will happen less than 5 times per year per circuit.
- Recovery events, lasting 2 seconds or less but more than 50ms, will happen less than 3 times per year per circuit.

14900 Conference Center Drive
Suite 280
Chantilly, VA 20151
703-955-7770

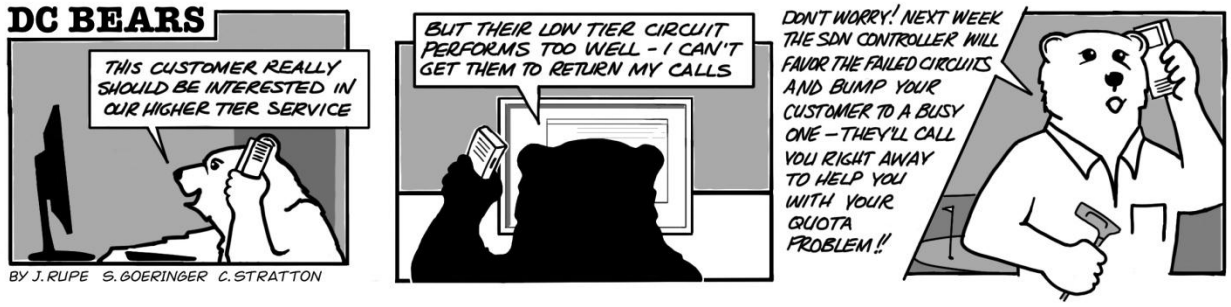
- Restoral events, lasting 20 seconds or less but more than 2 seconds, will happen less than 2 times per year per circuit.
- Repair events, lasting 5 hours or less but more than 20 seconds, will happen less than 1 time per year per circuit.

Option 2: Specify a general tiered structure on downtime frequency and duration. This option is the same as above, except it is not linked to specific transaction or service types, but has a general performance definition in terms of latency, throughput, etc., and has an outage breakdown like that above.

Option 3: For a generalized transaction or service category as assumed in the previous option, this option can simplify so that the overall total time in each outage breakdown category is defined. For example, a restoral event might be defined to last less than or equal to 30 seconds and greater than 5 seconds, and the total time a circuit is in that category can be defined to be less than or equal to two minutes for the entire year.

Option 3 is a great place to start when transitioning to an approach like this. A model-based analysis of the given network architecture, or analysis of network failure data when available, can help to set initial levels.

Other options are also possible. When defining alternate options, while there is a desire for the simple, fewer criteria on service or transaction availability or outage breakdown categories is not always best. The number of these categories should be aligned to differentiate between modes of failure for the service or transaction, and not reduced to a level that breaks that alignment. Fewer criteria is not necessarily a simpler approach, given calculations like these are no longer manually done.



Examples

The basic case example will be a completely fictional one, artificially complex to fully use as much of the framework as possible. Imagine a communications product with a latency requirement restricting the communication path to be less than or equal to L. The bandwidth of the product is defined to be BW, but to function as intended it must be greater than or equal to some level B<BW. Note that B can be a

14900 Conference Center Drive
Suite 280
Chantilly, VA 20151
703-955-7770

function of demand for the service when considered at a small interval of time, or set for an SLA over a larger amount of time. BW was set to be sufficient for the service requirements at least 99% of the time (via excellent capacity management).

This particular communication product is created to enable video service. At the frame rate, considering defects that are noticeable, any loss greater than 100ms will be noticed by the user. Further, there will be frame freezing and buffering if the signal is lost more than 2 seconds. Any loss of communication for more than 30 seconds will require the user to reestablish the link. And any outage longer than 1 minute will result in a user not being able to establish communication through the service interface. The intervals are defined as

1. 0<protection<=100ms,
2. 100ms<recovery<=2 seconds,
3. 2 seconds<restore<=30 seconds,
4. 30 seconds<reconnect<=1 minute,
5. 1 minute< repair.

Service Level Agreement Example

Given the two end points of the communication product, through modeling and historic data analysis, the network is expected to deliver the L and B performance required 99.98% of the time. Further, we analyze the model and data to determine that the five categories above have the following SLA that can be offered.

1. No more than 50 protection events per year.
2. No more than 20 recovery events per year.
3. No more than 10 restoration events per year.
4. No more than 5 reconnection events per year.
5. No more than 2 repair events per year.

These SLAs are all defined to be broken with less than a 95% probability when the network is performing as intended, and the SLA is met.

Network Performance Example

To assure performance against the SLA offered, the circuit is monitored and data analyzed to determine the actual performance. The event data are collected from network event capture, network alarms, and repair ticketing systems. The effort here amounts to binning event data into the five categories of service impact.

The network performance is captured monthly, and a running 12 month measure is provided each month for each category. Overall availability is provided, with the performance limits of L and B considered when defining failure from both the overall availability measure as well as the five bins.

14900 Conference Center Drive
Suite 280
Chantilly, VA 20151
703-955-7770

Note that, if one bin's count is exceeded, it is possible that by moving one or more events into a longer duration bin that the SLA is then met, so reclassification of events into worse categories than actually experienced may be warranted. This practice should be allowed, as it will incent the right behavior of reducing the duration of certain types of outages.

Tailored SLAs and Network Responses via Software Defined Networking and Network Function Virtualization

While a given network solution configuration can be offered to support many different applications, the ordering process allows customers to reveal their use case intent and willingness to pay for certain service behavior support. The same solution can be offered with several different SLA outage intervals to choose from based on use intent. But with some capabilities enabled by Software Defined Networking (SDN), and network functions that could be offered dynamically through Network Function Virtualization (NFV), network products could potentially be managed to specific SLAs offered to specific customers. Based on the actual performance of a customer's product, the network can compensate to adjust service and correct an otherwise poor trend. Like offering to correct an order in a restaurant, and offer something free for the customer's trouble, the network could be programmed to implement similar policies based on SLAs offered, which can be a key differentiator for some service providers.

14900 Conference Center Drive
Suite 280
Chantilly, VA 20151
703-955-7770