From Dark to Grey Failures in the Internet

JP Vasseur (jpv@cisco.com), PhD, Cisco Fellow, Engineering Lead - June 2021

Network availability has always been at the heart of networking, but how is "availability" defined? For years, the main focus has been to quickly detect link and path failures due to temporary lack of connectivity. A plethora of control and data plane technologies have been developed to (quickly) detect link/node failures and reroute along alternate paths. The term SLO (Service Level Objective) refers to network availability where network failure is determined by a temporary lack of connectivity between two edges. With the rapid growth of technologies such as the Cloud that heavily rely on the network, application impacts are now expanding the concept of SLA/SLO where the application Quality of Experience (tightly coupled with the user experience) becomes the new metric. In this short paper, we introduce the concept of grey failures (significantly harder to detect and fix), which refers to situations where network connectivity is (at least partially) preserved, but the application SLA/SLO do not meet the application expectations, thus affecting the user experience, a much harder problem to solve.

Detecting (dark) failures: a solved problem

A number of technologies have been developed over the past two decades allowing for the fast detection of failures at multiple layers: lower layers (L1, L2) triggering (when possible) cross layer signaling, L3 using IGP hello packets (at low/medium frequency rates) and/or using routing agnostic Keep-Alive (KA) such as BFD at relatively higher rates. Combining fast failure detection mechanisms with fast rerouting has operated for decades. Needless to mention that upon detecting such dark (connectivity) failures, network recovery technologies allowed for rerouting traffic within few milliseconds using either *protection* (at multiple layers) technologies, such as optical or IP/MPLS FRR, or *restoration* via Optical/ or IP rerouting techniques, where alternate paths are recomputed on the fly. Many strategies arose dealing with the delicate trade-off between fast convergence time (thus reducing the downtime) and stability (oscillations between paths triggering out-of-order packets delivery and jitter to mention a few of the challenges). *Problem (mostly) solved.*

Unfortunately, connectivity is far from being sufficient in a network to meet the SLA of sensitive applications. Various path characteristics come into play that can greatly influence the application experience and (unfortunately) such characteristics are highly application specific. In this paper, we introduce the concept of *Grey* failures where a path exhibits characteristics that preserves connectivity but does not provide sufficient Quality of Service to an application (i.e. does not meet the required application SLA). Grey failures are hard to ***define***, ***detect*** and ***remediate***. Nonetheless, advanced solutions do exist that are briefly described in this paper.

Defining grey failures

The definition of dark failure is quite straightforward: a dark failure refers to a (temporary) lack of connectivity between two edges. In contrast, Grey failures refer to the network phenomenon that impacts application experience although the path is technically "alive". A widely adopted approach that has been used so far consists in using so-called SLA templates. A template is a set of static conditions (thresholds) provided for several network KPI (typically loss, delay and jitter, sometimes throughput) that must be met in order to meet the application SLA. For example, the recommendation for H323 devices connected to Webex meetings are not to exceed 0.5% of pack loss, and 150ms latency. Webex Teams clients could tolerate up to 2% of loss and 300ms of latency. As discussed above, such templates are application-dependent. Furthermore, the type of CODEC, transport layers and other parameters of the network also play a role.

A critical (and often overlooked) question: *how should the network KPI be computed?* Measuring an average packet loss over a long period of time is meaningless. Using a series of percentiles over shorter periods of time combined with max values is without a doubt more efficient but would not allow for capturing short and sporadic phenomenons impacting the user experience. Such (sporadic) issues are often undetectable by using network probes (without introducing unreasonable network overhead). Still the use of such templates is the most common approach so far.

To make matters worse, such grey/application failures are often subjective. For example, the perception of voice quality has been specified using the well-known MoS (Mean Opinion Score) metric that can be computed using disparate algorithms. As a matter of fact, the user perception is subjective and other metrics come into play such as the number of interruptions and their duration. When considering video, the user perception may be even more subjective and highly driven by the application (type of CODEC, rebuffering strategy, …). In other words, a given path characteristic may lead to different Voice/Video user experiences for different users.

Detecting grey failures

Detecting grey failures is challenging but must be solved in order to optimize the user experience and route the traffic in the Internet that maximizes application experience by increasing network SLOs. This means that some grey failures may not be detectable at L1-L4 layers, (sometimes) because of lack of granularity provided by probing (i.e. measuring the loss/delay/jitter every few seconds will not allow for the detection of short, sporadic network phenomenon that can highly impact the user experience especially for network sensitive applications), such as voice and video. A promising approach consists in coupling telemetry from multiple layers, including the application layer thanks to open APIs, whereby application experience is provided by the application experience itself (feedback). Such application feedback can then be correlated with network centric KPI to train ML models capable of detecting such grey failures. Other approaches consist in combining and cross correlating multiple layer input features for the detection of such grey failures.

Understanding path characteristics that may impact application experience is of the utmost importance. Advanced studies have been made to compute a broad series of path characteristics such as the entropy, Welch spectral density, some measures of statistical dispersion, so as to determine a set of path metrics that may impact the application experience. Such systems are complex involving a number of network and layer components.

For illustrative purposes, figures 1, 2 and 3 show a series of paths dynamics represented in a 3D space (loss, delay, jitter) with very different dynamics over time (time density plot is shown on the right) in a simplified

3-D space (loss, delay, jitter). This highlights that measuring an average network KPI does not reflect the actual path characteristics. Note that such trajectories may be observed in higher dimension spaces potentially after dimensionality reduction.

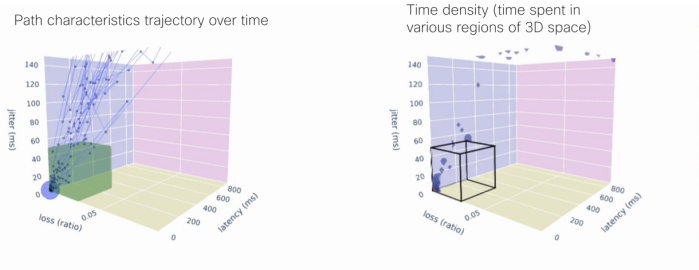

Figure 1 – Example of a path dynamics (high entropy, chaotic)

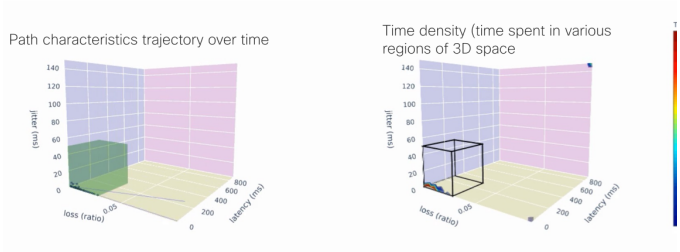

Figure 2 – Path dynamics (unstable (high standard deviation), low entropy (not chaotic))

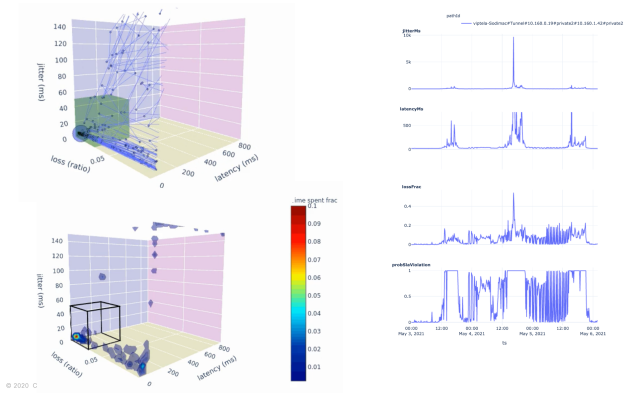

Figure 3 – Path dynamics (non-seasonal, high standard deviation)

The details of such analysis and computation of path characteristics and their influence on application are beyond the scope of this document.

Another approach consists in detecting changing conditions. For example, the system could detect that the proportion of time not meeting some application SLA may suddenly change (see figure 4), a situation the user is very sensitive to.

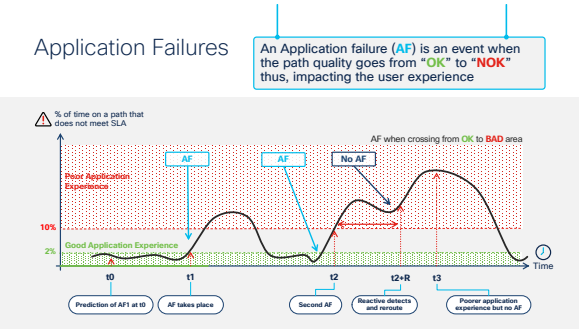

Figure 4 – detecting "Changing conditions"

Are grey failures that common? Yes, they are. For the sake of illustration, figure 1, 2 and 3 shows raster plots with both dark and grey failures on several (real life) paths. A series of paths are monitored along time, each grey dot refers to a grey failure whereas dark failures are shown in red (each raw corresponds to a different path in the network). Several scenarios can be observed:

- Reliable paths: rare occurrences of grey and dark failures,
- Paths with mostly dark failures but a low number of grey failures: usually indicative of well dimensioned path rarely affected by grey failures,
- Paths with high rates of Grey (not dark) failures (potential issue with dimensioning),
- Paths suffering from both dark and grey failures.

Figure 5 shows a network with approximately 3,000 paths (Y-axis, each row corresponding to a different path) and the raster plot shows the number of dark (red) failures with temporary lack of connectivity and grey failures with application experience impact (no loss of connectivity). The blue histogram represents the distribution of traffic onto such paths. During that period of time, 4,794 dark failures (tunnel down) occurred and there were 34,908 grey failures.

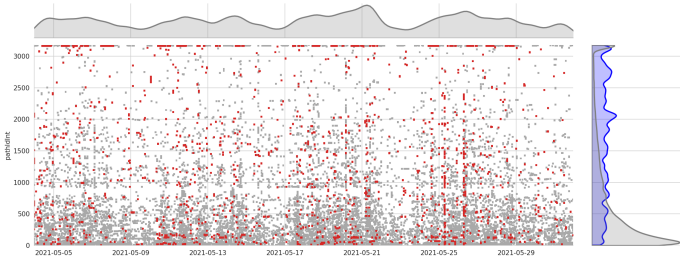

Figure 5 – Set of paths affected by dark and grey failures – majority of grey failures, traffic evenly distributed across paths, fairly even distribution of dark/grey failures across paths.

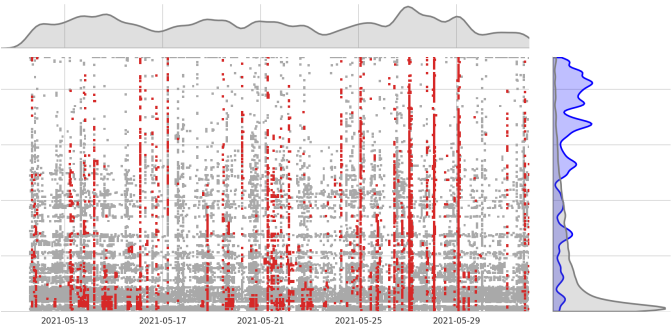

Figure 6 – Set of paths affected by dark and grey failures – large proportion of paths with dark/grey failures, few paths with low number of dark failures but higher number of grey failures along paths carrying most traffic (blue)

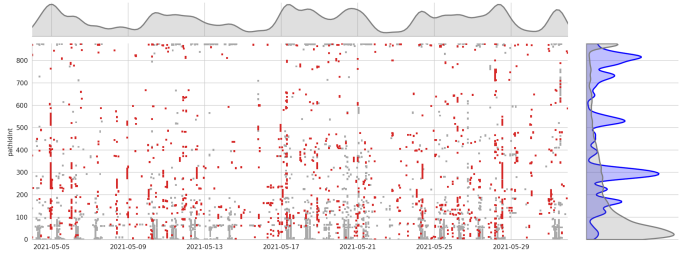

Figure 7 – Very different profile: 40% of failures are dark failures (network nicely provisioned), most traffic carried onto paths with few dark failures but low number of grey failures, just few tunnels with high proportion of grey failures (not carrying high volume of traffic).

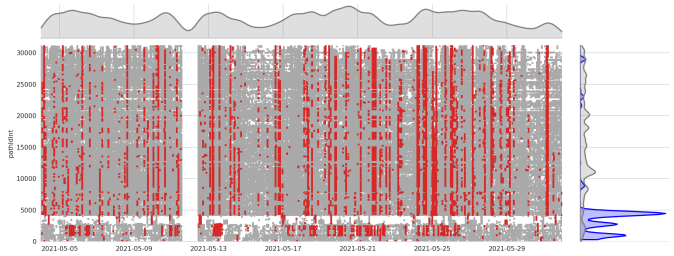

Figure 8 – Network with large proportion of dark/grey failures, 90% are grey failures, traffic highly unevenly distributed and strong correlation between paths impacted by dark failures.

Remediating grey failures

The ultimate goal is to remediate grey failures using either a reactive or a proactive approach (see the paper [predictive-Internet]). Remediation means (potentially proactively) rerouting traffic on a per application basis using an alternate path offering higher application experience. Knowing whether the alternate will offer higher SLO brings by itself a whole range of challenges.

Conclusion

Detecting and remediating dark failures (lack of connectivity) has drawn a lot of attention to the Internet community. Although such failures are still common, a wide range of technologies allowed for high network availability. A high focus must be given to Grey failures that do impact application experience. Such failures are hard to define, detect and remediate but a plethora of novel approaches have emerged to lead drastic improvements of application experience across the Internet. Such approaches rely on the use of significantly more advanced technologies using different metrics and a cross-layer approach to broaden the spectrum of grey failures the new Internet will have to detect and remediate. It is worth mentioning that cross-layer signaling where application feedback would be provided by the application itself (layer-7 KPI) would dramatically improve the application experience, potentially allowing for true application driven routing across the Internet, and will be discussed in a further document.

Acknowledgement

I would like to express my real gratitude to several key contributors I have been working with for a number of years: Gregory Mermoud, Vinay Kolar, Jurg Diemand, Mukund Raghuprasad are few who had a major contribution.

References

[internet-dynamics] “Large-scale Internet Path modelling and applications” September 2020

[predictive-internet] “Towards a Predictive Internet”, June 2021