Is Intelligence the Answer to Deal with the 5 V's of Telemetry Data?

L. Velasco*, S. Barzegar, and M. Ruiz

Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
e-mail: luis.velasco@upc.edu

Abstract: Telemetry data and big data share volume, velocity, variety, veracity and value characteristics. We propose a distributed telemetry architecture and show how intelligence can help dealing with the 5 V's of optical networks telemetry data.

1. Introduction

Telemetry data, like Big Data, is a collection of data from many different sources and it can be described by means of five characteristics, i.e., the 5 V's: *volume*, *velocity*, *variety*, *veracity*, and *value*, which can be seen as different tiers of a pyramid. At the base, *volume* refers to the size and amount of data that needs to be collected and analyzed. *Velocity* refers to the speed at which data is collected, stored and managed. Volume and velocity together impose requirements that need to be carefully considered, e.g., sometimes it is better to have limited data in real time than lots of data at a low speed. *Variety* refers to the diversity and range of different data types and data sources. *Veracity* is related to the quality, accuracy, trustworthiness of data and data sources and it is the most important factor of all the five V's for business success. Finally, at the very top of the pyramid, *value* refers to the ability to transform data into useful insight.

Some works in the literature have highlighted the need to collect telemetry data into a centralized system (see e.g., [1]) running close to the Software-defined Networking (SDN) controller. This defines a *telemetry pipeline* with basically two elements: *data collectors* that gather measurements from observation points in devices and send them to a *centralized telemetry system* that stores and processes the received data. That design is based on the principle of collecting from the network and storing as much data as possible, in the hope that they can feed network automation systems, e.g., based on Machine Learning (ML) [2]. In this paper, we illustrate with examples the limitation of such telemetry system and highlight requirements that need to be considered when designing an optical network telemetry system.

2. Examples of the 5 V's

In this section, we illustrate each of the 5V's with an example of the optical core network for a national telecom operator. Let us assume a core mesh network with 50 optical nodes, with average nodal degree of 3. Let us assume that each node is connected with any other node in the network through one single optical connection (*lightpath*). Then, the network supports 2,450 unidirectional lightpaths and needs the same number of transmitters (Tx) and receivers (Rx). Finally, let us assume that we can collect telemetry data from Tx, Rx, Optical Amplifiers (OA) in the nodes that compensate for filtering and fiber attenuation (i.e., 300 OAs in total), and from Optical Spectrum Analyzers installed in every optical link in the network (i.e., 150 OSAs in total). Table 1 summarizes the measurements that can be collected from every device and the estimated size.

Table 1. Measurements Data

Device	Measurements	Size (bytes)
Tx	• Laser params, e.g., temperature. • Configuration, e.g., modulation format and symbol rate.	20 20
Rx	• Optical constellation (10,000 IQ symbols). • Receiver parameters, e.g., bit error rate, signal to noise ratio, etc.	80,000 40
OA	• Input power, gain, etc.	20
OSA	• Optical spectrum (C-band, resolution 1 GHz).	16,000

- *Volume:* Let us assume that the measurements in Table 1 are collected every second. Then, for the network described above, this network generates 15.61 terabyte (TB) of data every day, i.e., 5.56 petabytes (PB) every year, that need to be collected, conveyed to the centralized telemetry system, stored in a data lake, and analyzed.
- *Velocity:* Collecting measurements every second imposes additional requirements related to data collection (i.e., in the devices) and data transport to the centralized telemetry system. Starting with the optical devices, those generating measurements of large size (i.e., OSAs and optical receivers) need high-speed data interfaces. E.g., OSAs would require 128 kb/s interfaces, while Rx would require 640 kb/s interfaces. And these speeds are assuming that measurements are generated as a stream of float numbers. However, such measurements, once collected are usually formatted, e.g., as a JSON object, which increases its size. For instance, every node agent collecting local telemetry data,

The research leading to these results has received funding from the HORIZON SEASON (G.A. 101096120) and the MICINN IBON (PID2020-114135RB-I00) projects and from the ICREA Institution.

formatting data and sending them to the centralized telemetry system, would generate around 40 Mb/s, so the centralized system would receive 1.9 Gb/s of data in total.

- *Variety:* Six structured measurements are defined in Table 1, which consist of tuples of individual magnitudes to vectors of related values. Additionally, also logs and other unstructured data from different systems are collected, which require a totally different processing process. All these different data types need to be processed, analyzed and correlated in real time. For instance, analysis of spectrum measurements from nodes in the route of a lightpath, together with analysis of optical constellation in the receiver can be used to identify and localize the cause of a sudden increase of the BER measured in the receiver.
- *Veracity:* Right decisions are made with thorough and correct information. Data can only help if it is clean, i.e., it is accurate, error-free, reliable, consistent, bias-free, and complete. Therefore, some factors that contaminate data are, among others: *i*) meaningless information that distorts the data; *ii*) outliers that make the dataset to deviate from the normal behavior; *iii*) software vulnerabilities that could enable data hijacking; and *iv*) statistical data that misrepresents a particular network resource.
- *Value:* Telemetry data can bring large benefits for network automation but only if they are converted into useful insight. Operators can capture value from telemetry data by: *i*) reducing network margins; *ii*) automating service provisioning; *iii*) improving resource utilization and reducing operational costs; *iv*) extending the working life of network equipment; *v*) detecting soft-failures before they become hard failures; *vi*) simplifying maintenance by finding root cause of failures and scheduling works; and many others.

3. How can Intelligence help and where it should be implemented?

Intelligence can be applied along the telemetry pipeline to reduce the impact of the 5V's. However, let us first challenge some of the assumptions in the previous section, which impact on 5V's factors:

1) *Do all measurements need to be collected with the same periodicity?* The answer is no. For instance, the configuration of the Tx happens at setup time or upon some event, and the temperature of the laser would not significantly change that fast. Then, different measurements should have different collection periodicity, which can be variable, or even being collected asynchronously.
2) *Is it useful to store all the measurements even when no significant changes happen?* Again, the answer is no. However, to determine whether a significant variation in one measurement has happen, some analysis needs to be carried out, and that should be done earlier in the telemetry pipeline, e.g., at the node level.
3) *Can measurements be compressed to reduce bandwidth requirements?* Here, a clear yes. Compression techniques should be explored, which can be either lossy or lossless.
4) *Should telemetry systems be totally centralized?* From the two previous questions, the answer is a clear no. Some processing and data analysis might be needed at the node level. However, such analysis might be orchestrated by some entity running at a centralized level, which can have global network vision.
5) *Where is the correct place for cleaning data?* Data veracity should be checked along the telemetry pipeline and it should be discarded from the main pipeline whenever there is evidence that such data is somehow contaminated. For instance, a sample data that does not follow statistically last measurements can be either an outlier or an anomaly. However, the detection point can be local if it refers to the gain of an amplifier or it needs to be in the centralized system if it requires correlation with other measures, e.g., in the case of spectrum measurements in the route of a lightpath.
6) *Where is the best place to extract value from data?* One might answer: as soon as you can in the telemetry pipeline. If you can detect degradations from the data collected from a network node, why we should wait to do so in the centralized telemetry system? However, sometimes, it is necessary to perform correlation among data collected from different network nodes to extract value from data.

4. Intelligence and telemetry pipeline

Data collected from observation points in the devices (measurements) are typically conveyed to a central system for further analysis. In addition, events generated by applications/platforms (e.g., SDN controllers and management systems) can be used to keep consistency among systems. In this section, we first introduce a telemetry architecture supporting distributed intelligence along the telemetry pipeline and next, we present two techniques to greatly reduce the dimensionality of telemetry data.

1. Telemetry architecture to support intelligence

Fig. 1 presents the reference network architecture with distributed telemetry. An SDN architecture controls a number of optical nodes, e.g., optical transponders and reconfigurable optical add-drop multiplexers, in the data plane. A centralized *telemetry manager* is in charge of receiving, processing and storing telemetry data in a telemetry database (DB). Some data exchange between the SDN control and the telemetry manager is needed, e.g., the telemetry manager needs to access the topology DB describing the optical network topology, as well as the DB describing the lightpaths (these DBs are not shown in Fig. 1).

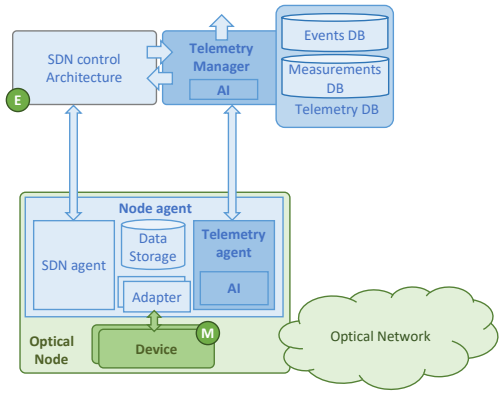

Fig. 1: Network architecture with distributed telemetry

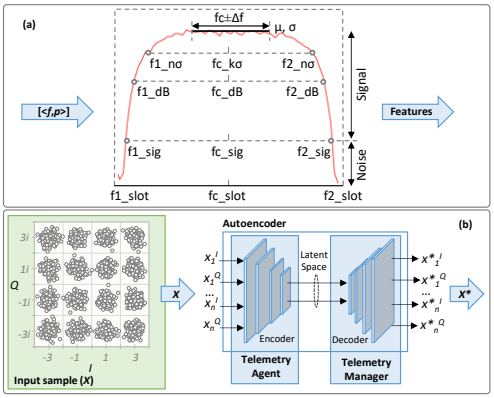

Fig. 2: Feature extraction (a) and autoencoder (b)

Every node in the data plane is locally managed by a node agent, which translates the control messages received from the related SDN controller into operations in the local node and exports telemetry data collected from observation points (labeled M) enabled in the optical nodes. In addition, events can be collected from applications and controllers (labelled E). *Telemetry agents* run inside node agents and provide the needed services for intelligent algorithms based on Artificial Intelligence (AI) techniques to process collected telemetry measurements.

2. *Dimensionality reduction*

Let us now present two examples of processing that can be carried out in the telemetry agent and help to reduce the impact of both volume and velocity of telemetry for optical spectrum and IQ constellations, which are by far the cases where collected samples are larger.

A simple but effective dimensionality reduction technique is supervised feature extraction. As an example, in our previous work [3], we proposed a module to pre-process the optical spectrum of a signal, i.e., an ordered list of frequency-power (<*f*, *p*>) pairs (see Fig. 2a). After equalizing power, the module characterizes the mean (μ) and the standard deviation (σ) of the power around the central frequency ($fc \pm \Delta f$), as well as a set of primary features computed as cut-off points of the signal with the following power levels: *i*) equalized noise level, denoted *sig* (e.g., -60dB + equalization level); *ii*) a family of power levels computed with respect to μ minus $n\sigma$, denoted $n\sigma$ (e.g., 3 and 5σ); and *iii*) a family of power levels computed with respect to μ minus a number of dB (e.g., -3 and -6 dB), denoted *dB*. Each of these power levels generates a couple of cut-off points denoted $f1_{(\cdot)}$ and $f2_{(\cdot)}$. In addition, the assigned frequency slot is denoted $f1_{\text{slot}}$, $f2_{\text{slot}}$. Then, the input list with 50 <*f*,*p*> pairs (i.e., 400 bytes) representing the spectrum of a 50GHz channel is processed to generate a set of 13 values (i.e., 52 bytes) that can be easily transformed into value, e.g., for failure detection and identification, in the telemetry agent or the manager.

Another technique to effectively reduce data dimensionality are autoencoders (AE), i.e., a type of neural network with two components: the *encoder*, which maps input data into a lower-dimensional *latent* space, and the *decoder*, which gets data in the latent space and reconstructs the original data back. Assuming the case of a 16-QAM signal, the AE takes as input *k* IQ symbols (e.g., 10,000), i.e., $[x_1^I, x_1^Q, \ldots x_k^I, x_k^Q]$, from the received constellation sample and generates the latent space $Z=[z_1, \ldots, z_L]$, where the size of Z is significantly lower than that of X (see Fig. 2b). In this case, the encoder might run in the telemetry agent and exchange Z for every input sample X with the decoder running in the telemetry manager. The decoder then reconstructs the constellation sample with high accuracy and the sample is stored and analyzed. Reconstruction can be performed also in the telemetry agent, e.g., for veracity checking purposes like detecting outliers and/or anomalies. In [4], we trained the AE for the maximum compression that produces a reproduction error in the decoder lower than 2%, which results in vectors Z of size 32, and therefore, achieving 625:1 compression ratio.

5. Conclusions

The 5V's of telemetry data have been examined and illustrated in the context of optical networking. In the view of intelligence can help to deal with such characteristics, a distributed architecture has been proposed that extend the telemetry pipeline by supporting intelligence close to the observation points as well as in the centralized system. Examples to help reduce the requirements of optical spectrum and constellations telemetry have been shown. Finally, just to mention that the architecture and techniques in Section 4 are demonstrated in [5].

References

[1] L. Velasco *et al*., "Mon. and Data Analytics for Optical Networking: Benefits, Architectures, and Use Cases," IEEE Netw. Mag., 2019.
[2] D. Rafique *et al*., "ML for Optical Network Automation: Overview, Architecture and Applications," IEEE/OSA JOCN, 2018.
[3] A. P. Vela *et al*., "Soft Failure Localization during Commissioning Testing and Lightpath Operation [Invited]," IEEE/OSA JOCN, 2018.
[4] L. Velasco, P. González, and M. Ruiz, "An Intelligent Optical Telemetry Architecture," in Proc. OFC, 2023
[5] P. González *et al*. "Distributed Architecture Supporting Intelligent Optical Measurement Aggregation and Streaming Event Telemetry," OFC demo session, 2023.