electronics

MDPI

Article

Detection and Localization of Failures in Hybrid Fiber–Coaxial Network Using Big Data Platform

Milan Simakovic and Zoran Cica *

School of Electrical Engineering, University of Belgrade, 11120 Belgrade, Serbia; milanrus@hotmail.com
* Correspondence: zoran.cica@etf.bg.ac.rs; Tel.: +381-11-3218-377

Abstract: Modern HFC (Hybrid Fiber–Coaxial) networks comprise millions of users. It is of great importance for HFC network operators to provide high network access availability to their users. This requirement is becoming even more important given the increasing trend of remote working. Therefore, network failures need to be detected and localized as soon as possible. This is not an easy task given that there is a large number of devices in typical HFC networks. However, the large number of devices also enable HFC network operators to collect enormous amounts of data that can be used for various purposes. Thus, there is also a trend of introducing big data technologies in HFC networks to be able to efficiently cope with the huge amounts of data. In this paper, we propose a novel mechanism for efficient failure detection and localization in HFC networks using a big data platform. The proposed mechanism utilizes the already present big data platform and collected data to add one more feature to big data platform—efficient failure detection and localization. The proposed mechanism has been successfully deployed in a real HFC network that serves more than one million users.

Keywords: big data; failure detection; failure localization; HFC networks; network management

Citation: Simakovic, M.; Cica, Z. Detection and Localization of Failures in Hybrid Fiber–Coaxial Network Using Big Data Platform. *Electronics* **2021**, *10*, 2906. https://doi.org/10.3390/electronics10232906

Academic Editors: Stavros Shiaeles, Bogdan Ghita and Nicholas Kolokotronis

Received: 6 October 2021
Accepted: 22 November 2021
Published: 24 November 2021

1. Introduction

HFC (Hybrid Fiber–Coaxial) networks evolved from traditional cable TV networks to offer their users a broader spectrum of services, i.e., triple-play service (voice, data, video). HFC networks employ DOCSIS (Data over Cable Service Interface Specification) standards. The DOCSIS 3.1 supports 10 Gbps downstream and 1 Gbps upstream [1]. DOCSIS 4.0 will support even higher capacities and full-duplex transmission [2]. Thus, HFC networks can offer high-quality broadband access to their users.

Given the trend of remote work, which has been further increased due to the COVID-19 (Coronavirus Disease 2019) pandemic [3,4], it is of great importance that users are provided high network access availability. Therefore, it is essential for telecom service providers to detect and localize malfunctions and failures in their networks as soon as possible. Thus, an efficient monitoring system needs to be deployed in the network.

The telecom service provider networks typically comprise huge numbers of network devices and links to provide service to large number of users—usually in the order of millions. Network devices and devices at user premises (CPE—Customer Premises Equipment) can provide a lot of useful collected information. However, the amount of data that can be collected is enormous. Thus, big data technologies must be able to collect, process, and store such an amount of data. Today, many telecom service operators deploy big data platforms to exploit the collected data from their networks [5–8].

In this paper, we propose a novel approach to efficiently detect and localize malfunctions and failures in HFC networks using a big data platform. In the remainder of the paper, we refer to our proposed approach as FDLBD—Failure Detection and Localization using Big Data platform. Since big data platforms collect data on a regular basis, the collected data can be used to estimate the health of all devices in the network. However,

some devices in the HFC network, such as amplifiers, cannot directly provide their status, i.e., data cannot be collected from such devices. Our proposed solution can indirectly estimate the health of such devices from the data collected from the devices that can provide their status. By using the network topology information, FDLBD can also localize the problematic devices. In this way, complete automatized monitoring of all devices is achieved. Furthermore, the monitoring costs are reduced, since our approach uses the big data platform that already collects data, and there is no need for additional equipment and capacities for the monitoring purposes. Note that the proposed FDLBD is currently successfully deployed in the real HFC network comprising over one million users. The main contributions of the paper are the following:

- Efficient automatized failure detection and localization in HFC networks.
- Failure detection of devices that cannot be monitored and probed directly.
- Utilization of big data technology for failure detection and localization.

The remainder of the paper is organized as follows. Related work is presented in Section 2. Section 3 briefly covers typical HFC network architecture. In Section 4, we provide a high-level architecture description of the big data platform that FDLBD uses. Section 5 gives detailed description of FDLBD. Finally, Section 6 concludes the paper.

2. Related Work

With the introduction of DOCSIS 3.1, the new cable modem communication standard, the speed of transmission has been significantly increased. This is achieved by using Orthogonal Frequency-Division Multiplexing (OFDM) and higher modulation order. OFDM enables carrier configuration per each cable modem and unlocks flexibility for throughput optimization, which was not the case in the previous standard (DOCSIS 3.0). This opportunity is recognized by [9,10]. Anastasia et al. proposed an algorithm for dynamic carrier configuration for each individual modem based on machine learning, tested it on 10 real case datasets and proved better results [9]. Sumayia et al. carried out similar research but proposed their own algorithm for profile management [10]. Both approaches relied on data collected from the network, proving the importance of the data platform.

Emilia et al. focused on proactive network maintenance. They applied unsupervised machine learning to group cable modems. Based on the similarity of the received signal and cable modem location, they attempted to estimate whether the problem was inside or outside the house [11].

Myung-Sun et al. focused on the self-interference phenomenon in DOCSIS 3.1. In full-duplex communication, signals in different directions distort each other. Since the amount and nature of interference are known in advance, they proposed a self-estimation technique to minimize the impact of the mentioned phenomenon and proved it in laboratory conditions [12,13].

New internet trends show a higher demand in upstream communication. While DOCSIS 3.1 is still the top-notch standard in HFC networks, it still has asymmetrical links. There are several papers that propose an extension of this standard to achieve symmetric communication [14,15]. Brian et al. analyzed a full-duplex extension to DOCSIS 3.1 and gave an overview of challenges grouped by the physical, MAC (Media Access Control) domain, and system layer [14], while Werner et al. proposed a full-duplex standard that attempts to solve all the challenges and simulate the proposed protocol in laboratory conditions [15].

Obviously, most papers in the area of HFC networks focus on improving the performance of communication either on the detection or correction of problems of specific parts/elements of the HFC network. In this paper, we focus on the mechanism to detect and localize points of failure regardless of the network device type.

However, the problem of failure detection and/or localization is important aspect of any communication network technology. Thus, in this section, we also give an overview of the related work regarding failure detection and/or localization in wired communication networks.

A very detailed survey regarding the network monitoring aspects is given in [16]. An important part of any network monitoring system is efficient fault management that comprises detection and localization [16]. The importance of knowledge about network topology for efficient failure detection and localization is emphasized in [16]. Failure detection and localization in optical networks is hot-topic area because these networks usually represent backbone and core of service provider communication network infrastructure [17–22]. Two types of failures can be observed, hard and soft failures. Hard failures represent complete malfunction of the device or link (for example, fiber cut), while the soft failures represent the degradation of performance (for example, component aging) [17]. Hard failures impact the network performance immediately and are easier to detect. However, the soft failures are important as well, as they gradually degrade the overall network performance. In our paper, the focus is on hard failures, but the big data platform and collected data can be used for soft failures detection as well; however, this aspect is not within the scope of this paper. Typically, machine learning and neural networks are used for soft failure detection in optical networks [17–19]. Similarly, machine learning can be used for failure localization in optical networks as well [20–22]. The proposed machine learning-based solutions require some data (such as power spectrum density [18], bit error rate samples [19], routed lightpaths [20], mean time between failures [21], etc.) from the optical network to perform failure detection or localization.

Failure detection and localization are important for computer networks and data centers as well. Many protocols and mechanisms are defined for fast failure detection and fast rerouting, especially for the transport (core) parts of the network, such as MPLS (MultiProtocol Label Switching) fast re-route mechanisms [23]. A survey of failure diagnostic techniques (localization, detection) in computer networks is given in [24]. Failure diagnostic can be passive or active. Passive techniques rely on monitoring agents deployed on network devices, where monitoring agents can signal various alarms to network management system. Active techniques rely on sending probes across various paths in network to detect and localize the failures in the network. Optimal selection of probes is a major aspect of active techniques [25]. Besides alarms generated by the devices, logs can be used for failure detection as well [16,26]. Logs can not only be processed to detect failures, but to predict failures as well. The main issue is that logs are not structured data, and the log format depends on the vendors but also on the versions of software installed on the network devices [26]. This increases the complexity of log processing and requires updates of log processing tools whenever a device from another vendor is installed in the network or when devices are updated with new firmware versions. Data centers represent computer networks with huge numbers of devices. For the efficient data center performance, it is important to swiftly detect and localize failures. The problem is even harder given the existence of at least two routes between any node pair, and the equal-cost, multi-path routing that is commonly used in data centers [27]. Active probing was proposed in [27] for fast failure detection and localization. There are a few papers that deal with the failure detection and localization in access networks as well. In [28], failure detection solution that uses RADIUS (Remote Authentication Dial-in User Service) protocol was proposed for an xDSL (x Digital Subscriber Line) access network. A failure localization solution for FTTH (Fiber-to-the-Home) networks was proposed in [29].

In case of large networks with enormous number of devices, big data is unavoidable for network monitoring purposes. However, as we stated in Section 1, big data collected from communication networks can be used for many purposes. In [30], a very detailed survey is given on big data usage in wired and wireless networks. Some of the discussed aspects of big data usage include traffic prediction, QoS (Quality of Service) improvement, cybersecurity, network performance optimization, etc. [30]. There are not many papers about big data use for failure detection and localization. Mainly, these papers cover the mobile networks [31,32]. XDR (External Data Representation) data is used for failure detection in mobile networks [31], while in [32], bandwidth trends were analyzed to predict the equipment failures.

HFC networks are addressed in the literature mostly regarding their technical aspects. The focus is on the improvement and optimization of devices, throughput, etc. However, there are no papers that deal with hard failure detection and localization in complete HFC network. Furthermore, big data usage in HFC networks is also not covered adequately in the literature. There are papers that cover big data usage in other communication networks, but HFC networks have their specific properties that differ them from other networking technologies. Additionally, neither of the papers that discuss big data usage in other communication networks cover fast failure detection and localization jointly. We hope that our paper can fill up this research gap.

3. HFC Network Architecture

In this section, we give a brief description of HFC network architecture. The purpose of this section is to present typical devices in HFC networks and determine which of them cannot be monitored directly.

Figure 1 shows a typical segment of an HFC network covered by one CMTS (Cable Modem Termination System). The CMTS distributes services, such as TV, voice, and data, in the downstream direction using optics. Downstream optical signal from CMTS reaches ON (Optical Node) via Hub. Optical signal is converted to an electrical signal in ON. The services, received from CMTS, are then distributed as electrical signals in the downstream direction to the end users. AMPs (Amplifiers) are passive devices that perform non-linear RF (Radio Frequency) amplification. They are needed to cope with coaxial cable attenuation. At the end, the signal reaches CPE via an Access Point (AP). CPEs are typically cable modems or set-top boxes. In the upstream direction, signals from CPEs are frequency multiplexed and sent in the opposite direction through the cable network to an optical node that performs electrical to optical conversion of the signal. Finally, the upstream signal reaches the CMTS via Hub.

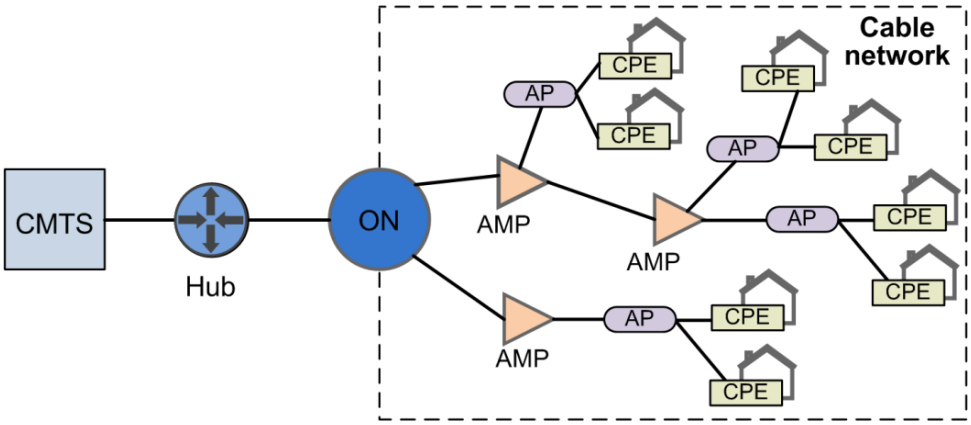

Figure 1. HFC network. CMTS: Cable Modem Termination System, ON: Optical Node, AP: Access Point, CPE: Customer Premises Equipment, AMP: Amplifiers.

HFC network operators can collect data from the HFC network devices. However, not all devices support data collection. AMPs, APs, and ONs are not capable of supporting data collection. On the other hand, CMTSs and CPEs do support data collection. Depending on the supported protocols, there are several ways to collect data from CMTSs and CPEs. For example, SNMP (Simple Network Management Protocol), IPDR (Internet Protocol Detail Record), and FTP (File Transfer Protocol) can be used for data collection purposes on CMTSs and CPEs.

Figure 1 shows that the cable network part has a tree topology. This means that from each user there is exactly one path to the CMTS. This information can be used to indirectly estimate the health of other devices in the tree, such as ONs, AMPs, and APs. For example, if the data collection process shows that all CPEs under some APs are not available, most likely that AP is down.

4. Big Data Platform

Telecom service providers can collect enormous amounts of data from their networks. Information extracted from the collected data can be used for various purposes. For example, data can be used to obtain information for business planning, efficient network expansions, better offers to their customers, performance monitoring, and performance optimization. Obviously, there are numerous possibilities for how to use collected data. However, traditional methods for processing and storing collected data are not powerful enough. For this reason, big data technologies are used to manage collection, processing and storing of these enormous amounts of data.

In this section, we present the big data platform that FDLBD utilizes for failure detection and localization purposes. Figure 2 shows the overall big data platform architecture that comprises the data collection layer and big data cluster. The data collection layer is responsible for data collection, while the big data cluster is responsible for data processing and storage. Information obtained from collected data is consumed by data consumers (call centers, dashboards, alarm systems, etc.).

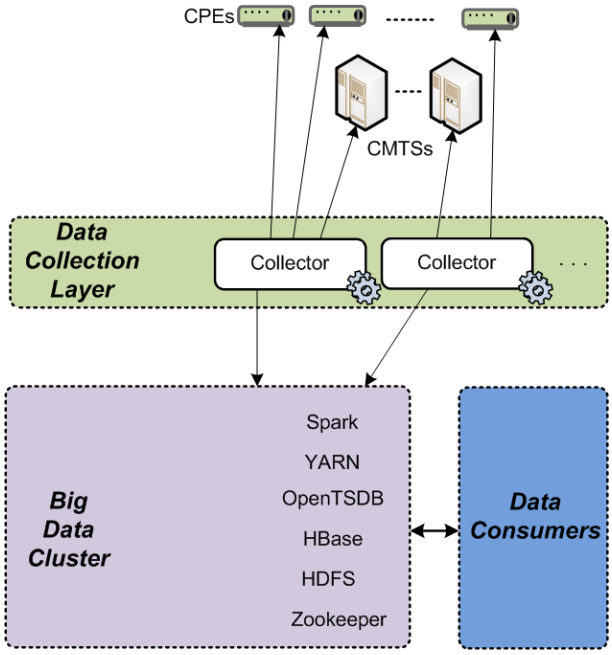

Figure 2. Big data platform.

The data collection layer collects data from CPEs and CMTSs in HFC network. SNMP is used for data collection. Given the large number of CPEs, the collection period from CPEs is set to 1 h. On the other hand, the number of CMTSs is significantly lower than the number of CPEs. Additionally, CMTSs should be constantly online and highly responsive. Furthermore, CMTSs contain a lot of useful data, where some data refer to end-users themselves. For all these reasons, the collection period for CMTSs is set to range from one to five minutes, depending on the importance of the particular data metric that is collected. In case of troubleshooting, corresponding troubleshooting mechanisms can decrease the collection period to the order of seconds for the corresponding (inspected) device.

Collected data are sent to the big data cluster. More precisely, data are sent to the OpenTSDB (Open Time Series Database) for data aggregations, and the aggregation results are used immediately by the data consumers. At the same time, a copy of the same data is stored to the HDFS (Hadoop Distributed File System) for later batch processing that is conducted using the spark. Batch processing results are stored in the HDFS and can be used by the OpenTSDB for data aggregation purposes and furthermore, for data consumption by data consumers.

5. Failure Detection and Localization Based on Big Data

One common problem related to telecom service providers is network failure. Network failures are not rare and happen for a variety of reasons. Power failure, optical cable being cut, and equipment failure are only some of the problems that occur on a daily basis. There are two challenges that telecom service providers need to overcome: failure detection and failure localization. To overcome these challenges in the case of HFC networks, we propose the FDLBD mechanism. In this section, we first give an in-depth description of the FDLBD mechanism. Then, we provide more detailed descriptions of failure detection and failure localization parts of the FDLBD mechanism.

An important requirement is that any failure needs to be detected as soon as possible. Failures are monitored on the CMTS level. CMTS is the centralized device for one part of the network, which is responsive for querying, and, therefore, the most suitable. We use the metric cdxCmtsCmRegistered [33] in FDLBD. Metric cdxCmtsCmRegistered shows the number of online and active modems per MAC domain. Since the number of active modems constantly changes, the failure threshold needs to be defined to avoid false detections. We define the following rule: a failure is detected on one CMTS and one MAC domain when the current collected number of active modems is 15% less than the average of twenty previous collections for that CMTS and MAC domain. This rule has been tested in practice on a real HFC network. The practice has shown excellent results for failure detection while minimizing false detections.

After the failure is detected, the failure localization part of the FDLBD is performed. Failure localization helps the HFC network operator to efficiently solve the failure problems. First, the FDLBD mechanism tries to establish a connection with every CPE device that belongs to the problematic CMTS MAC domain. Information on whether the connection is established or not is correlated with network topology. FDLBD traverses from the bottom (CPE) to the top (ON) of the tree topology and seeks the top device in the network hierarchy under which all modems are unavailable. The found device is the point of failure.

Figure 3 shows an example of failure localization. Red color shows that all CPEs under such devices are offline. Green color shows that all CPEs under such devices are online. Yellow color shows that there are both online and offline CPEs under such devices. Tree topology makes it easier to detect the root of the problem. It is the top device under which all CPEs are offline. In the given example, the amplifier AMP_ng2 is the point of failure. AMP_as218 is not the point of failure because there are online CPEs beneath it.

5.1. Failure Detection

The detection part of the FDLBD mechanism should be constantly observing to the number of online users and triggering the failure localization in case of sudden drops. The number of online users is not constant and fluctuates with time. Figure 4 shows the number of online users in one network part for one day.

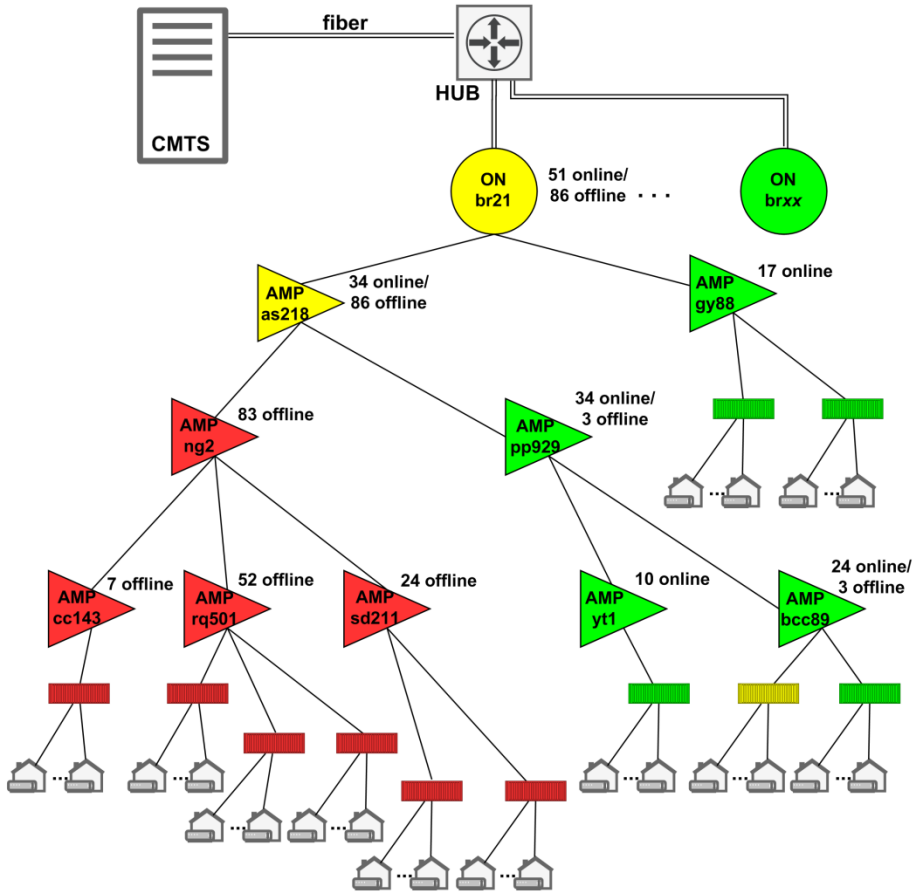

Figure 3. Failure localization example.

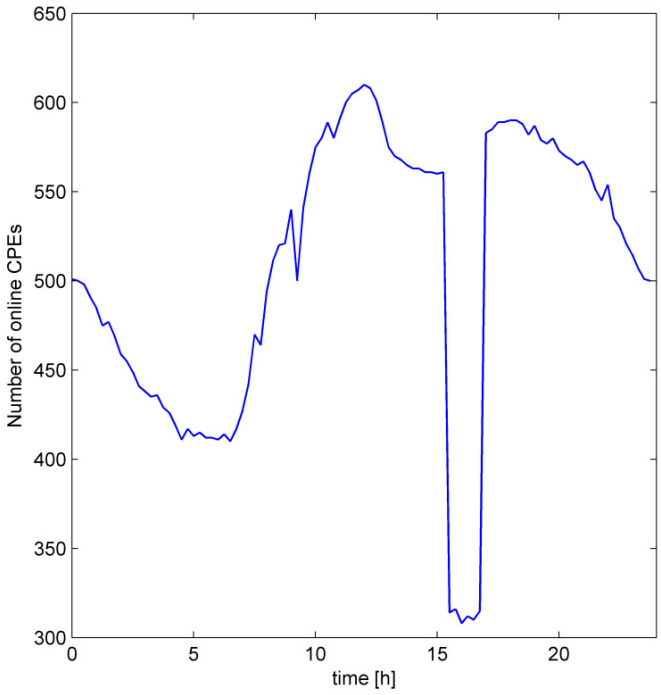

Figure 4. Number of online users.

Obviously, the number of online users goes up and down depending on the period of the day. In Figure 4, it can also be noticed that the number of online users has dropped sharply, which represents an example of an outage of corresponding network part. To detect such cases, we propose a failure detection part of FDLBD as follows. First, the average number of active users N_{onl_avg} in the last K iterations is calculated:

$$N_{onl\_avg} = \frac{1}{K}\sum_{i=1}^{K} N_{onl}[T_{CURR} - i], \tag{1}$$

where T_{CURR} is the current measurement period, and $N_{onl}[j]$ is the number of active users in measurement period j. K can be set to the desired value depending on the particular network behavior. For the real HFC network, where FDLBD is applied and a one minute collection period occurs, we have determined that $K = 20$ gives the best results. However, we suggest testing prior to the deployment of the FDLBD mechanism to determine the optimal value for K in the corresponding HFC network.

We define failure if the following condition is met:

$$Failure = \frac{N_{onl}[T_{curr}]}{N_{onl\_avg}} \leq 1 - Threshold\_d, \quad 0 < Threshold\_d < 1, \tag{2}$$

where *Threshold_d* is, as the name suggests, the threshold for failure detection. The practice in the real HFC network has shown that *Threshold_d* = 0.15 gives optimal results. Note that Equations (1) and (2), as well as suggested parameter values (K and *Threshold_d*), have been devised through experiments on the real HFC network data and by comparing algorithm results with real failures in the network. For the HFC network we tested, the proposed approach detects over 95% of real network failures. Our approach can be used in other HFC networks. However, depending on the social environment where HFC network is deployed, clients' behavior (nation culture, habits, working hours, etc.) can be different, and thus cdxCmtsCmRegisteredtime series might differ. For this reason, it is necessary to tune parameters to obtain the best results. Therefore, to get the best possible results using our proposed algorithm, tests should be conducted prior to FDLBD deployment to determine the optimal *Threshold_d* value for the corresponding HFC network. Our parameters are tuned for one cable operator in Europe.

Note that the similar threshold approach can also be used to detect situations when a problem is resolved. This is very useful in situations when a problem is resolved by itself. For example, electricity loss for a few minutes can cause such behavior. In the case of integration with fault management service, these two events can correlate and automatically close previously opened corresponding issue. The condition that detects resolution is:

$$Resolution = \frac{N_{onl}[T_{curr}]}{N_{onl\_avg}} \geq 1 + Threshold\_r, \quad 0 < Threshold\_r < 1, \tag{3}$$

where *Threshold_r* is the threshold for problem resolution detection. The only limitation it has is that the problem may happen before condition (3) is applied. In other words, problem resolution detection is activated only after the failure is detected. Note that FDLBD deploys the previously described problem resolution detection as well.

Table 1 shows one example of detection and resolution of one failure that lasted a few minutes. We have only provided relevant measurements ($N_{onl}[T_{curr}]$, N_{onl_avg}) in Table 1 because the data sets collected from the real HFC network are confidential, and we are not able to disclose them. However, the values provided in Table 1 can help to demonstrate our algorithm. The red row shows the moment where *Threshold_d* is violated; thus, in this moment failure is detected because the *Failure* value dropped below 0.85. If the failure lasted more than K minutes, then at one point *Threshold_d* would not be triggered according to (2) because N_{onl_avg} would be averaged to a lower number of active users. However,

this is not an issue because threshold triggering does not have to be continuous since the first capture of threshold violation will cause failure detection. In the example provided in Table 1, we also demonstrate the possibility of detecting resolution. The green row shows the moment when resolution of the problem is detected because *Resolution* value is greater than 1.15. This example shows that failure resolution can be detected even when failures only last several minutes.

Table 1. Failure detection and resolution example.

T_{curr}	$N_{onl}[T_{curr}]$	N_{onl_avg}	$N_{onl}[T_{curr}]/N_{onl\_avg}$
1584659988	417	455.55	0.915377017
1584660048	413	451.15	0.915438324
1584660108	415	446.9	0.928619378
1584660168	412	442.6	0.930863082
1584660228	120	424.05	0.282985497
1584660288	121	405.85	0.298139707
1584660348	121	388.15	0.311735154
1584660408	118	370.2	0.318746623
1584660468	429	368.2	1.165127648
1584660528	427	366.6	1.164757229
1584660588	442	365.95	1.207815275

5.2. Failure Localization

Once the failure is detected for a particular MAC domain, the failure localization is triggered. The main goal of this part is to detect the network device (AP, AMP, or ON) that is the root cause of the problem. Prerequisites for failure localization are topology mapping information (for example, comma-separated relation between parent and child network element) as well as the network element to which the CPE device is connected (for example, modem FF:3B:2A:8C:AE:DC is connected to AP_cc43.) In addition, mapping between MAC domains and CPE is mandatory (this can be populated querying CMTS metrics for docsIf3CmtsCmRegStatusMdIfIndex, docsIfCmtsCmStatusIpAddress, and docsIfCmtsCmStatusMacAddress [34]) and IP (Internet Protocol) address for each CPE device.

The localization algorithm can be split into several steps:

1. Ping all CPE devices that are connected to the problematic MAC domain.
2. Join ping results by the topology and calculate the percentage of offline CPE devices.
3. If there are 100% of offline modems, go one level up in the topology hierarchy and repeat the second step.
4. Repeat the third step until the percentage of offline CPE devices decreases. This means that the algorithm reached out of the root-cause element.

Because most of the CPE devices are constantly turned on, this algorithm gives almost no false positives in practice (theoretically, it would be possible that devices are simultaneously turned off on purpose) and precisely detects the failed network element.

We developed the failure localization algorithm by exploring the possible data that can be obtained from a CMTS and the ones that the network operator has in inventory system. The algorithm was tested and verified in the real HFC network. Real data sets used for algorithm development and verification are confidential and we are not able to disclose them. However, the same results can be achieved by generating appropriate test data sets, creating different case scenarios (no offline devices, some devices are offline, all devices are offline) and running the algorithm on top of it.

The following example shows the execution of the failure localization algorithm for the topology given in Figure 3. The given example is artificial and simplified but

reflects the behavior of the real network. Failure detection discovers a sudden drop in number of active users and triggers failure localization. First, mapping between the MAC domain and CPE (MAC address and IP address) is correlated with metrics (docsIf3CmtsCmRegStatusMdIfIndex, docsIfCmtsCmStatusIpAddress, and docsIfCmtsCmStatusMacAddress) gathered from CMTS. The example of the mapping:

(1) MAC_domain1, 10.0.1.17, FF:3C:2A:1C:FE:AA
(2) MAC_domain1, 10.0.1.22, FF:1B:A7:C9:D9:34
(3) MAC_domain1, 10.0.1.63, FF:AE:FE:2B:98:99
(4) MAC_domain1, 10.0.1.64, FF:B4:93:64:E7:AE
(5) ...
(6) MAC_domain1, 10.0.1.211, FF:9E:FC:98:86:F1
(7) MAC_domain2, 10.0.1.214, FF:DA:E3:5A:ED:79

Given the fact that such information slowly changes over time, it is collected periodically and used directly to reduce the localization time. The topology mapping shows a relation between the CPE MAC address and the network element to which CPE is connected first in the tree topology:

(1) FF:3C:2A:1C:FE:AA, AMP_cc143
(2) FF:1B:A7:C9:D9:34, AMP_rq501
(3) FF:AE:FE:2B:98:99, AMP_sd211
(4) FF:B4:93:64:E7:AE, AMP_yt1
(5) ...
(6) FF:9E:FC:98:86:F1, AMP_bcc89

In step 1 of the failure localization algorithm, all CPE devices that are connected to the given MAC domain are pinged. Modems that respond (records with "ttl" and "time") are online, while others are considered offline. An example of pinging result is as follows:

(1) ping 10.0.1.17, Request timeout for icmp_seq 0
(2) ping 10.0.1.22, Request timeout for icmp_seq 0
(3) ...
(4) ping 10.0.1.63, Request timeout for icmp_seq 0
(5) ...
(6) ping 10.0.1.64, 64 bytes from 10.0.1.64: icmp_seq=0ttl=64 time=0.053 ms
(7) ping 10.0.1.214, 64 bytes from 10.0.1.214: icmp_seq=0ttl=252 time=22.557 ms

In step 2, the algorithm joins results from step 1 with the network topology and calculates the percentage of offline modems. An example of step 2 results is as follows:

(1) AMP_cc143, 100
(2) AMP_rq501, 100
(3) AMP_sd211, 100
(4) AMP_yt1, 0
(5) AMP_bcc89, 11.11
(6) AMP_ng2, 100
(7) AMP_pp929, 8.10
(8) AMP_as218, 71.67
(9) AMP_gy88, 0
(10) ON_br21, 62.7

In steps 3 and 4, the algorithm travels bottom-up through the topology and attempts to find the top element under which all modems are offline. In this example, which is based on network topology in Figure 3, AMP_ng2 is the root cause.

5.3. Comparison

We compared the proposed FDLBD to other existing solutions in this section. However, direct numerical comparison is not completely possible because the existing solutions in current literature are designed for the network technologies other than HFC networks. Each network technology has some implementation differences which make it difficult to

perform direct comparison of different solutions. Thus, we have chosen the solutions that mainly target access networks [28,29] and a solution that exploits log records for failure detection [26], as this can be used in most of the network technologies. We also compared FDLBD with PNM (Proactive Network Monitoring), as presented in [11], as it also works with HFC networks and performs problem detection and localization. Finally, we would like to point out that FDLBD observes failures from the end user perspective, which is obvious from the workflow description of FDLBD given previously in Section 5. This property is important, as it is directly correlated to users' QoE (Quality of Experience).

A failure detection solution for xDSL access-aggregation network is presented in [28]. xDSL access networks are very similar to HFC networks in sense that both use tree network topology. User equipment (CPE) typically establishes PPP (Point to Point Protocol) sessions with a broadband remote access server. RADIUS is used to collect information about the sessions (for example, starts and endings of the sessions). The logs from the RADIUS servers are then used for failure detection. The approach is very similar to our proposed approach. The threshold is defined, and if the percentage of PPP session disconnections in defined time interval exceeds this threshold, then a failure is detected. Threshold is defined as a percentage of the active users on card in DSLAM (Digital Subscriber Line Access Multiplexer). The threshold is proposed to be slightly lower than 100%. The value of the threshold is high, given that the DSLAM is the last step towards the user premises in the access network—the downstream direction. This threshold requires an extensive search through the history of RADIUS logs, which represents significant processing burden. For this reason, an additional threshold is used. This threshold is set to 30–40% of the smallest card capacity and is used to monitor the complete network. Once this threshold is triggered, the threshold based on active users on card is inspected to avoid unnecessary processing for failure detections. This approach can be used for failure localization purposes (for example, to detect failure of some access switch), but this is not discussed in detail in [28]. We will refer to the solution proposed in [28] as DMPF (Detection of mass PPP failures) because that name is used by the author (Zych). When compared to FDLBD, the main downside of the DMPF is the fact that big data is not used. Zych confirms that the amount of RADIUS logs is huge given the great number of sessions in network, especially taking into account that the history of RADIUS logs must also be kept. This huge amount of data is not easy to process, and this is the main issue of DMPF, which would probably benefit from the big data technologies. Furthermore, details regarding failure localization are not presented in [28]. Finally, one potential downside of DMPF is the fact that PPP needs to be used in the access networks. Some xDSL access networks use DHCP (Dynamic Host Configuration Protocol), thus, in such cases, DMPF cannot be used.

Logs for failure detections are also used in [26]. IP networks are considered in [26], but the approach can be generalized to other networks. Logs are collected from the devices in the network by the NMS (Network Management System). By processing logs, failures can be detected and even predicted. The main problem in this approach is the log structure. Logs can have different formats in case of devices from different vendors, same devices that have different firmware versions, different types of devices, etc. Additionally, logs represent unstructured data. All of this leads to difficulties in log processing, and frequent updating is required whenever a new log structure needs to be added (for example, devices from new vendors are added to the network). Additionally, in the case when there are devices that do not produce any logs or responses (such as AMPs in HFC networks), there is no answer to how the state of these devices can be reconstructed and if it is even possible to do so. Approaches that use performance metrics collected by the devices would be a better approach since such data are already structured, and many network devices are able to collect such metrics. The use of structured data would decrease the data processing requirements. This is an important advantage of FDLBD compared to log-based approaches.

Failure detection and localization in FTTH networks were discussed in [29]. TDM-PON (Time Division Multiplexing Passive Optical Network) was considered. The paper

focused on localization of fiber problems and ONU (Optical Network Unit) failures. In practice, OTDRs (Optical Time Domain Reflectometer) are used by the technicians at ONU sites to send probes in order to detect fiber cuts. However, this is a time-consuming process; thus, the authors proposed automatization of the process by adding SREs (Switchable Reflective Elements). SREs are controlled remotely from the control center and can be configured (switched) to a reflective state. In this way, probing of the optical branches can be automatized and the failures localized. At the end of the paper, the centralized failure detection system is presented, but there are not many details regarding how the failure detection system actually detects failures. Obviously, this solution is highly focused to FTTH networks and cannot be generalized to other types of networks. Additionally, it detects and localizes failures only in the PON part of the network, and not in the aggregation part of the network like FDLBD and DMPF.

PNM for HFC networks is proposed in [11]. The idea of a PNM mechanism is based on identification of weak parts of the network that would fail in the recent future so they can be proactively fixed. The PNM mechanism uses spectral data collected by the Full-Band Capture tool, detects potential problems, and groups devices that exhibit the same problem. Moreover, the PNM mechanism makes a distinction between whether the problem is ingress or egress. PNM uses unsupervised machine learning with a k-means algorithm. On the other hand, FDLBD detects and localizes the problem from the moment it appears. Problem localization is performed by querying the CPEs directly and using a device correlation based on topology information enables FDLBD to precisely calculate problem position. Machine learning models always have space for false results, which is a downside of the PNM approach. We could conclude that PNM and FDLBD are rather complementary tools that could be used together in the network—FDLBD for efficient hard failures detection and PNM for efficient soft failures detection. In cases of hard network failures, FDLBD would detect such failures and trigger failure localization, while PNM would be used for overall network improvement by resolving the potential problems (soft failures) proactively.

A summary of the comparison of FDLBD to other solutions is given in Table 2. Note that under 'Adaptive' we consider the possibility of using the solution in other network technologies. Under 'Coverage' we consider the detection and localization coverage of all devices in the network.

Because FDLBD covers hard failures detection and localization, most of the solutions considered in this section are dedicated to hard failures and capable for both failure detection and localization. However, failure localization in DMPF is not elaborated on in [28]. Additionally, not much detail is given for failure detection in SRE-based solution in [29]. PNM is the only solution dedicated to soft failures, but on the other hand, PNM is the only solution that targets HFC networks besides FDLBD. FDLBD and log-based solutions are the most adaptive because they rely on SNMP and logs that are commonly used in most of the networking technologies and devices. Of course, in both solutions some adaptations would have to be made since metrics and log structures would differ between networking technologies. In the case of networks that contain loops, additional adaptations should be made in FDLBD localization technique. For this reason, we placed medium and thigh grade for FDLBD in the 'Adaptive' column. DMPF is also adaptive because it is similar to FDLBD, but somewhat less than FDLBD because it relies on RADIUS. On the other hand, SRE-based solutions and PNM are highly focused on specific networking technologies, which makes them unadaptable to other networking technologies. FDLBD, DMPF and PNM are able to cover all devices in the network; thus, their coverage is high. The SRE-based solution has great coverage of the PON part of the network but does not consider aggregation part of the network. Coverage of devices that do not produce logs is not explained in [26]. Depending on the information collected from the logs of other devices, it might be possible to reconstruct the health of devices that do not produce logs but that highly depend on types of information contained in logs.

Table 2. Comparison summary.

	Network Technology	Failure Types	Detection/ Localization	Adaptive	Coverage
FDLBD	HFC	hard	both	medium/high	high
DMPF [28]	xDSL	hard	both	medium	high
Log-based [26]	IP	hard	both	high	medium
SRE-based [29]	TDM-PON	hard	both	low	medium
PNM [11]	HFC	soft	detection	low	high

6. Conclusions

Big data technologies are being introduced to many large service provider networks. The main reason for this trend is the possibility of extracting valuable information from data available and collected from the network. The main purpose is typically to improve business decisions and strategies. However, since collected data are at our disposal, the data can be used for other purposes as well. This paper provides one example in which collected data can be used for efficient automatized hard failures detection and localization. As a consequence, overall network availability and performance are improved, and users are more satisfied. Our proposed approach was confirmed by the successful deployment of FDLBD in real HFC network. Thus, we can conclude that the utilization of big data platform dedicated to collecting and processing data from the network can be increased by adding new functionalities that expand the big data platform's portfolio of possibilities. Additionally, the comparison given in Section 5.3 shows that some other existing solutions could benefit from big data technologies, especially in case of very large networks. Finally, our predictions are that big data technologies will play one of the major roles in most of the large service provider networks in near future.

FDLBD is efficient for tree topology networks. However, in case of adaptation to other network technologies where topologies other than trees are used (e.g., ones that contain loops), our proposed localization solution would have to be modified to support these other topologies as well. Thus, our future work will include the consideration of other network topologies. Furthermore, FDLBD covers hard failures, but soft failures are important too, as they also impact the overall network performance. An important part of our future work will be dedicated to the detection of weak spots in the network that degrade the network performance. In this case, data collected in longer period of time will be evaluated to detect and predict the performance degradation of the devices so they can be timely replaced or repaired. In this way, efficient performance monitoring and optimization of the network will be achieved. These are the two main directions of our future work, but we will also look into other aspects of using big data, such as efficient planning of network expansions and optimization of users' QoE.

Author Contributions: Conceptualization, M.S. and Z.C.; methodology, M.S. and Z.C.; software, M.S.; validation, M.S. and Z.C.; formal analysis, M.S.; investigation, Z.C.; resources, M.S.; data curation, M.S.; writing—original draft preparation, M.S. and Z.C.; writing—review and editing, Z.C.; visualization, Z.C.; supervision, Z.C.; project administration, M.S.; funding acquisition, M.S. and Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nguyen, T.T.; Nguyen, H.H.; Salt, J.E.; Berscheid, B. Zero-CP OFDM for DOCSIS-Based CATV Networks. *IEEE Trans. Broadcast.* **2021**, *65*, 727–741. [CrossRef]
2. Schnitzer, J.; Prahladan, P.; Rahimzadeh, P.; Humble, C.; Lee, J.; Lee, J.; Lee, K.; Ha, S. Toward Programmable DOCSIS 4.0 Networks: Adaptive Modulation in OFDM Channels. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 441–455. [CrossRef]

3. Russell, A.; Frachtenberg, E. Worlds Apart: Technology, Remote Work, and Equity. *Computer* **2021**, *54*, 46–56. [CrossRef]
4. Das Swain, V.; Saha, K.; Abowd, G.D.; De Choudhury, M. Social Media and Ubiquitous Technologies for Remote Worker Wellbeing and Productivity in a Post-Pandemic World. In Proceedings of the IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) 2020, Atlanta, GA, USA, 28–31 October 2020; pp. 121–130. [CrossRef]
5. He, Y.; Yu, F.R.; Zhao, N.; Yin, H.; Yao, H.; Qiu, R.C. Big Data Analytics in Mobile Cellular Networks. *IEEE Access* **2016**, *4*, 1985–1996. [CrossRef]
6. Garcia, A.J.; Toril, M.; Oliver, P.; Luna-Ramirez, S.; Garcia, R. Big Data Analytics for Automated QoE Management in Mobile Networks. *IEEE Commun. Mag.* **2019**, *57*, 91–97. [CrossRef]
7. Zhang, Y.; Cheng, X.; Xu, L.; He, X.; Guan, J.; Chao, K.; Song, C. A Novel Big Data Assisted Analysis Architecture for Telecom Operator. In Proceedings of the IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS) 2019, Shenyang, China, 21–23 October 2019; pp. 611–615. [CrossRef]
8. Jia, Y.; Chao, K.; Cheng, X.; Xu, L.; Zhao, X.; Yao, L. Telecom Big Data Based Precise User Classification Scheme. In Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI) 2019, Leicester, UK, 19–23 August 2019; pp. 1517–1520. [CrossRef]
9. Gaydashenko, A.; Ramakrishnan, S. A Machine Learning approach to maximizing Broadband Capacity via Dynamic DOCSIS 3.1 Profile Management. In Proceedings of the 18th International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 341–345. [CrossRef]
10. Abedin, S.; Ben Ghorbel, M.; Hossain, J.; Berscheid, B.; Howlett, C. A Novel Approach for Profile Optimization in DOCSIS 3.1 Networks Exploiting Traffic Information. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 578–590. [CrossRef]
11. Gibellini, E.; Righetti, C. Unsupervised Learning for Detection of Leakage from the HFC Network. In Proceedings of the ITU Kaleidoscope: Machine Learning for a 5G Future (ITU K), Santa Fe, Argentina, 26–28 November 2018; pp. 1–8. [CrossRef]
12. Baek, M.; Song, J.; Jung, J. Design and Performance Verification of Time-Domain Self-Interference Estimation Technique for DOCSIS 3.1 System with Full Duplex. In Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, Valencia, Spain, 6–8 June 2018; pp. 1–4. [CrossRef]
13. Baek, M.; Song, J.; Kwon, O.; Jung, J. Self-Interference Cancellation in Time-Domain for DOCSIS 3.1 Uplink System with Full Duplex. *IEEE Trans. Broadcast.* **2019**, *65*, 695–701. [CrossRef]
14. Berscheid, B.; Howlett, C. Full Duplex DOCSIS: Opportunities and Challenges. *IEEE Commun. Mag.* **2019**, *57*, 28–33. [CrossRef]
15. Coomans, W.; Chow, H.; Maes, J. Introducing Full Duplex in Hybrid Fiber Coaxial Networks. *IEEE Commun. Stand. Mag.* **2018**, *2*, 74–79. [CrossRef]
16. Lee, S.; Levanti, K.; Kim, H.S. Network Monitoring: Present and Future. *Comp. Netw.* **2014**, *65*, 84–98. [CrossRef]
17. Shu, L.; Yu, Z.; Wan, Z.; Zhang, J.; Hu, S.; Xu, K. Dual-Stage Soft Failure Detection and Identification for Low-Margin Elastic Optical Network by Exploiting Digital Spectrum Information. *J. Light. Technol.* **2020**, *38*, 2669–2679. [CrossRef]
18. Lun, H.; Zhuge, Q.; Fu, M.; Wu, Y.; Liu, Q.; Cai, M.; Zeng, X.; Hu, W. Soft failure identification in optical networks based on convolutional neural network. In Proceedings of the European Conference on Optical Communications (ECOC 2019), Dublin, Ireland, 22–26 September 2019; pp. 1–3. [CrossRef]
19. Shahkarami, S.; Musumeci, F.; Cugini, F.; Tornatore, M. Machine-Learning-Based Soft-Failure Detection and Identification in Optical Networks. In Proceedings of the Optical Fiber Communications Conference and Exposition (OFC 2018), San Diego, CA, USA, 11–15 March 2018; pp. 1–3. [CrossRef]
20. Mayer, K.S.; Soares, J.A.; Pinto, R.P.; Rothenberg, C.E.; Arantes, D.S.; Mello, D.A.A. Soft Failure Localization Using Machine Learning with SDN-based Network-wide Telemetry. In Proceedings of the European Conference on Optical Communications (ECOC 2020), Brussels, Belgium, 6–10 December 2019; pp. 1–4. [CrossRef]
21. Panayiotou, T.; Chatzis, S.P.; Ellinas, G. Leveraging statistical machine learning to address failure localization in optical networks. *J. Opt. Commun. Netw.* **2018**, *10*, 162–173. [CrossRef]
22. Li, Z.; Zhao, Y.; Li, Y.; Rahman, S.; Yu, X.; Zhang, J. Demonstration of Fault Localization in Optical Networks Based on Knowledge Graph and Graph Neural Network. In Proceedings of the Optical Fiber Communications Conference and Exposition (OFC 2020), San Diego, CA, USA, 8–12 March 2020; pp. 1–3. [CrossRef]
23. Gray, W.; Tsokanos, A.; Kirner, R. Multi-Link Failure Effects on MPLS Resilient Fast-Reroute Network Architectures. In Proceedings of the International Symposium on Real-Time Distributed Computing (ISORC 2021), Daegu, Korea, 1–3 June 2021; pp. 29–33. [CrossRef]
24. Dusia, A.; Sethi, A.S. Recent Advances in Fault Localization in Computer Networks. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 3030–3051. [CrossRef]
25. Qiao, Y.; Qiu, X.; Cheng, L.; Meng, L. A Methodology Used to Optimize Probe Selection for Fault Localization. In Proceedings of the IEEE Global Telecommunications Conference GLOBECOM 2010, Miami, FL, USA, 6–10 October 2010; pp. 1–5. [CrossRef]
26. Kimura, T.; Watanabe, A.; Toyono, T.; Ishibashi, K. Proactive failure detection learning generation patterns of large-scale network logs. In Proceedings of the International Conference on Network and Service Management (CNSM 2015), Barcelona, Spain, 9–13 November 2015; pp. 1–5. [CrossRef]

27. Tan, C.; Jin, Z.; Guo, C.; Zhang, T.; Wu, H.; Deng, K.; Bi, D.; Xiang, D. Netbouncer: Active device and link failure localization in data center networks. In Proceedings of the USENIX Conference on Networked Systems Design and Implementation, Boston, MA, USA, 26–28 February 2019; pp. 599–613.
28. Zych, P. Network failure detection based on correlation data analysis. *Int. J. Electron. Commun.* **2017**, *77*, 27–35. [CrossRef]
29. Ab-Rahman, M.S.; Chuan, N.B.; Safnal, M.H.G.; Jumari, K. The overview of fiber fault localization technology in TDM-PON network. In Proceedings of the International Conference on Electronic Design, Penang, Malaysia, 1–3 December 2008; pp. 1–5. [CrossRef]
30. Hadi, M.S.; Lawey, A.Q.; El-Gorashi, T.E.H.; Elmirghani, J.M.H. Big data analytics for wireless and wired network design: A survey. *Comput. Netw.* **2018**, *132*, 180–199. [CrossRef]
31. Chih-Lin, I.; Liu, Y.; Han, S.; Wang, S.; Liu, G. On Big Data Analytics for Greener and Softer RAN. *IEEE Access* **2015**, *3*, 3068–3075. [CrossRef]
32. Sahni, A.; Marwah, D.; Chadha, R. Real time monitoring and analysis of available bandwidth in cellular network-using big data analytics. In Proceedings of the International Conference on Computing for Sustainable Global Development (INDIACom 2015), New Delhi, India, 11–13 March 2015; pp. 1743–1747.
33. Simakovic, M.; Masnikosa, I.; Cica, Z. Performance monitoring challenges in HFC networks. In Proceedings of the TELSIKS 2017, Nis, Serbia, 18–20 October 2017; pp. 385–388. [CrossRef]
34. MIB Files Repository. Available online: https://www.circitor.fr/Mibs/Mibs.php (accessed on 27 September 2021).