Evolution of Network Robustness and Resiliency in the CIN

Strategies for High Availability in the R-PHY Network

A technical paper prepared for SCTE by John Huang

John Huang
Lead Network Engineer
Cox Communications
john.huang@cox.com

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

1. Introduction

Within North America, Cox Communications maintains one of the largest Converged Interconnect Network (CIN) deployments among service providers (Malla, 2021). The CIN is the component in a distributed access architecture (DAA) that makes Remote PHY (R-PHY) possible. It is essentially the transit layer that connects the Converged Cable Access Platform (CCAP) / Cable Modem Termination System (CMTS) core to the Remote Physical Devices (RPDs) and makes the capabilities offered by DOCSIS 3.1 and beyond, possible.

An increasingly large proportion of Cox's footprint is serviced by R-PHY as more and more nodes are digitalized. Hence, the CIN is foundational to a reliable product. Without a robust and dependable CIN, the increasing advantages and benefits inherently provided by ever-evolving DOCSIS technologies would be compromised. This metro delivery network, therefore, must be as reliable and resilient as possible. The CIN must not only be adaptable to provide increasingly greater levels of bandwidth, but just as important, it must also evolve in its ability to withstand failures of various kinds.

In this paper, we will discuss some of the key high-availability methods and technologies utilized by Cox over the past several years to develop an increasingly resilient CIN network.

2. The Need for Resiliency in the CIN

The obvious major benefit of Remote PHY has been the unprecedented levels of throughput made possible by the de-coupling of the PHY component from the traditional CMTS. However, with this benefit has come a new vulnerability. That is, of course, the introduction of the additional nodes and links that constitute the CIN. Essentially, these are additional points of failure in the end-to-end design of the R-PHY architecture that the traditional DOCSIS environment did not have to deal with. Of course, the components and design of what constitutes the CIN can vary from provider to provider, or even within the same provider's network, but invariably, it introduces new points of failure in the overall architecture. Every piece of the end-to-end design must be operational for services to be optimally delivered. It goes without saying then that having the benefits and advantages offered by R-PHY would be weakened or perhaps even rendered moot if the underlying architecture was deficient.

3. Cox CIN Design

The current Cox CIN design consists of 3 components – 1) RPA or "RPD Aggregation", 2) DPA or "Digital Physical Interface Card (DPIC) Aggregation", and 3) SPINE layer (i.e. HUB aggregation) -- that reside in a leaf-spine architecture as shown in the below diagram (Figure 1). RPDs terminate on the RPA platform in a single-homed manner via the access fiber ring – i.e. each RPD has one uplink to one RPA port. DPAs terminate the connections from the DPIC line card(s). DPA devices are deployed in an active/standby manner to provide node and link redundancy. And finally, SPINE-1 and SPINE-2 are HUB devices that serve as the aggregation point for RPAs and DPAs.

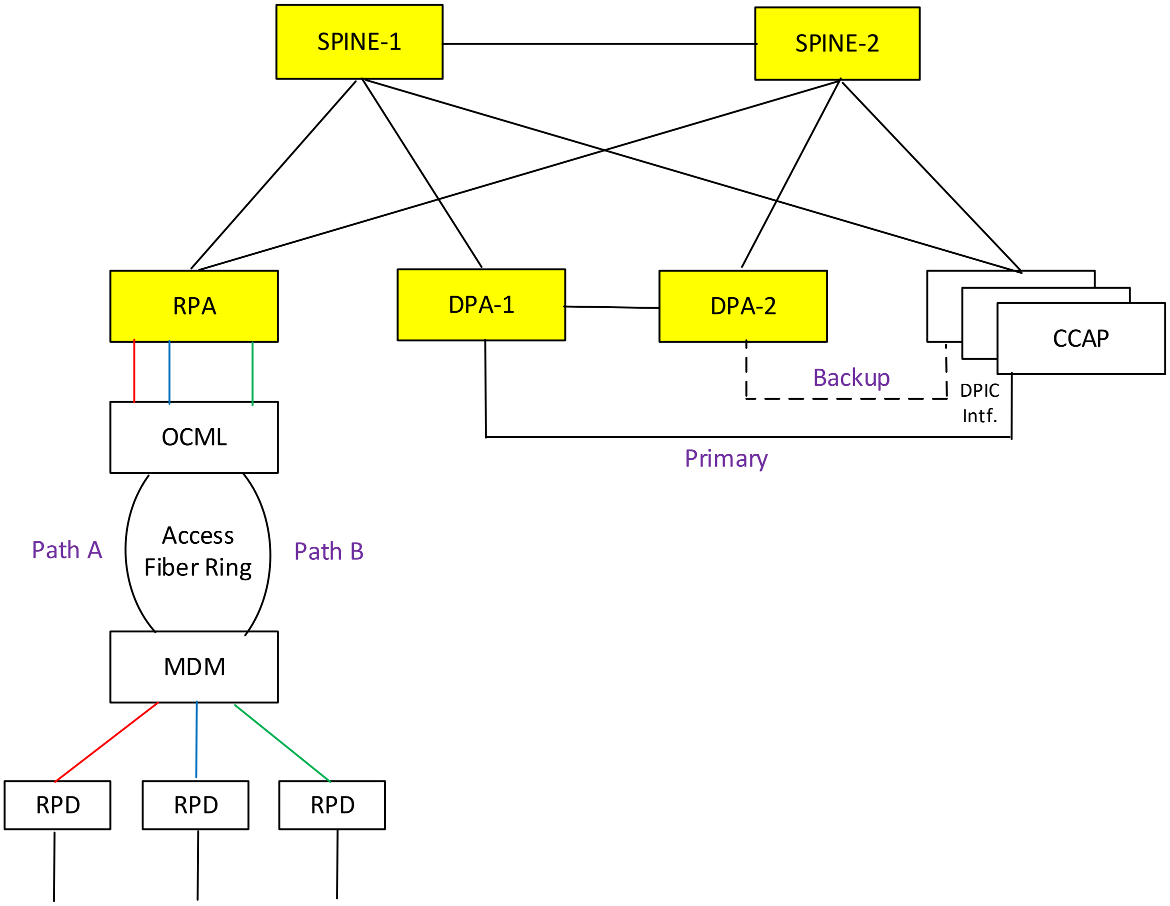

Figure 1 - Cox Communications CIN Topology (Malla, 2021)

With this topology in mind, the engineers at Cox have been regularly performing assessments of the CIN network at large to address areas of vulnerability in a cost-effective manner. Implementing resiliency usually comes at some type of cost, whether monetary or administrative, so consideration must be made to determine if the added benefits outweigh any potential risks or negative consequences. To state it another way, simplicity and redundancy are usually at opposite spectrums, so any new feature or tactic should be considered with great care.

In general, the major aspects of vulnerability that are evaluated are identifying physical areas of risk (i.e. single points of failure, congestion areas, hardware redundancy, etc.), assessing L2/L3 route convergence, and determining the blast radius of any given type of failure.

4. Day 1 Approaches to Resiliency & High Availability

The mitigation and resiliency strategies described in this section have been implemented in the Cox CIN from the time of initial deployment. As many of these are common strategies, we will not go into great detail.

4.1. L1/L2 methods

a. LAG/Port-channel – deployment of leaf-spine interconnects as LAG interfaces to ensure availability even if “x” number of physical links fail. Also, each RPA and each DPA pair are multihomed to the spine layer for node redundancy at the spine. Note the DPAs are indirectly multihomed to the spine rather than each DPA being physically connected to both spine routers.

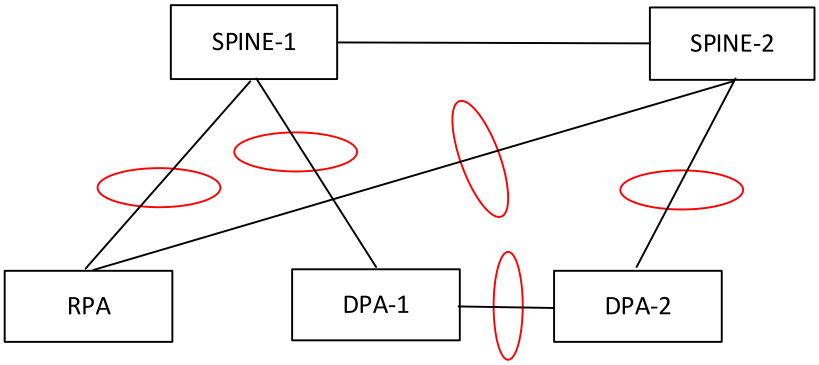

Figure 2 - LAGs in CIN

b. Diverse transport paths for RPA to RPD connections – provides redundancy at layer 1 to mitigate effects of fiber outage. Utilize proprietary OCML (Optical Communications Module Link Extender) and MDM (Mux/Demux Module) devices to deliver DWDM wavelengths over primary and redundant fiber paths.

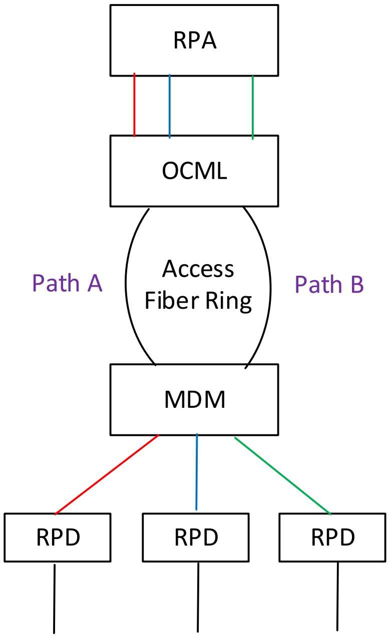

Figure 3 - RPA to RPD redundancy

c. Redundant DPA to DPIC links – as mentioned in section 3, DPA-DPIC connections are terminated and provisioned with primary and backup ports to provide both link and node redundancy. As a result, we supplement the port and line card redundancy that is available on the CCAP with port and node redundancy at the DPA layer.

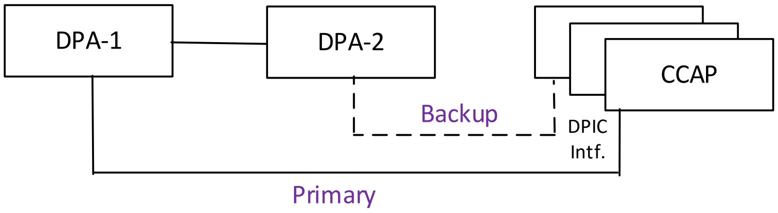

Figure 4 - DPA to DPIC redundancy

d. Redundant hardware – leveraged with the goal of consuming minimal physical footprint, provide hardware redundancy where possible.
   1. RPAs – redundant fans, power supplies
   2. DPAs – redundant chassis, fans, power supplies

4.2. L3/L4 methods

a. Border Gateway Protocol (BGP) multihoming – utilize the common practice of BGP multihoming with optimized timers for quick convergence. Allows routing updates and information to be readily available even in the event of a BGP process failure or node failure at the spine layer. Also, use appropriate attributes – e.g. local preference – for deterministic traffic flows.

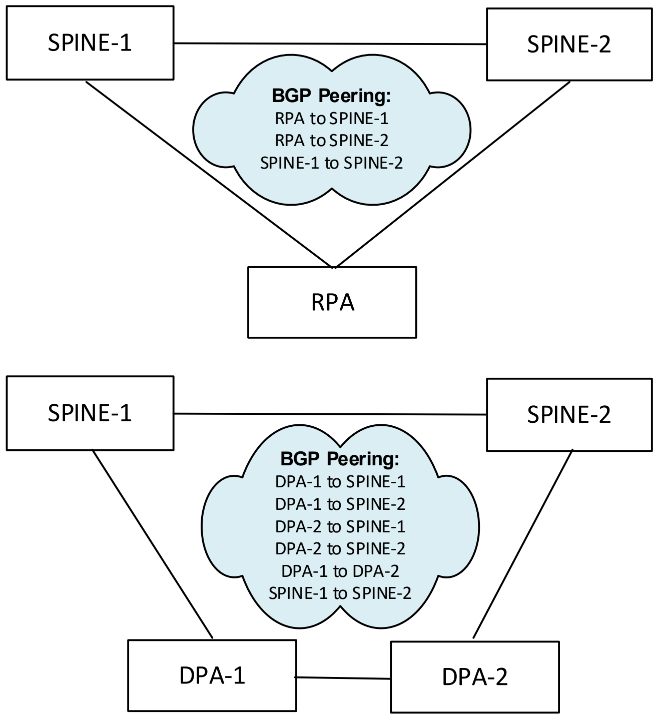

Figure 5 - BGP multihoming design

b. BGP error handling – another recommended best-practice is to ensure BGP neighbor state is preserved even in the event of a malformed update being received. Although the implementation procedures may vary from vendor to vendor, error handling and error tolerance is something that typically needs to be manually enabled.

c. Routing resiliency for Generic Control Plane (GCP) traffic – when considering the routing design of the CIN network, it is crucial to pay special attention to GCP communication, as this is the critical "underlay" protocol that prevents RPDs from having to reinitialize. If there is any unicast summarization occurring in the network, be mindful to ensure the reachability to RPD and DPIC prefixes is maintained even in various failure scenarios.

   For example, consider the scenario below, where unicast routes at the bottom layer -- where the RPA and DPA leaf nodes reside – rely on the availability of the unicast aggregate to be advertised from the Tier 1 route reflectors (RRs) to the Tier 2 routers. GCP reachability could be compromised if for some reason the core layer is unable to advertise the unicast aggregate to the spine layer. This can happen, for instance, when all BGP adjacencies between the core and the spine layers simultaneously go down.

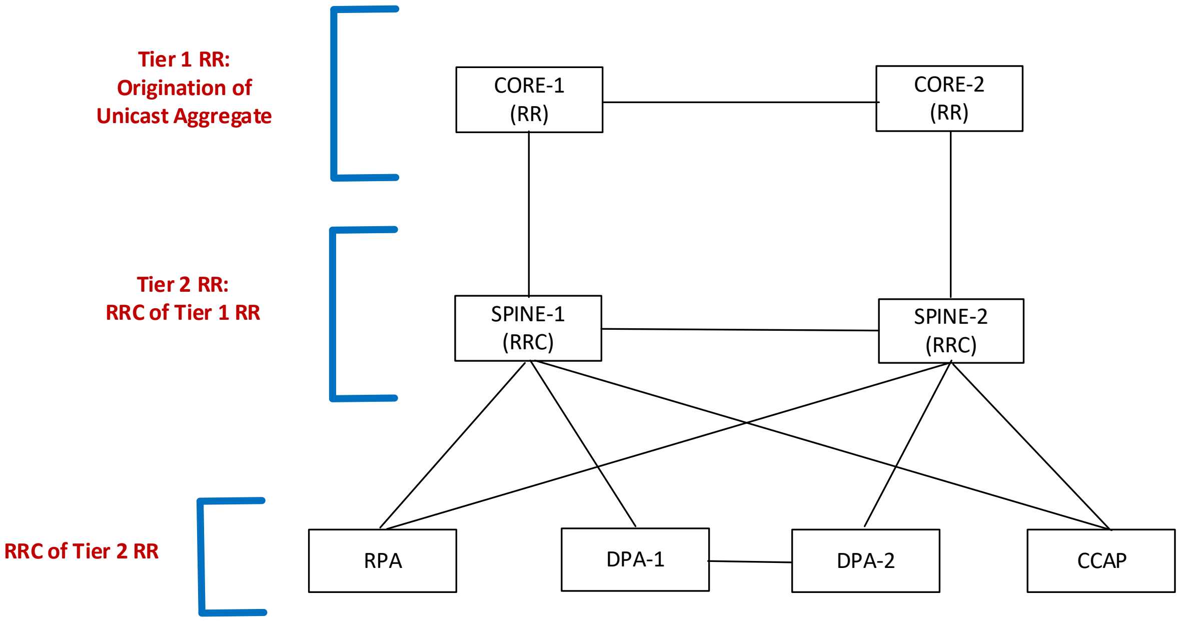

Figure 6 - Route Design

The solution to this vulnerability that has been implemented at Cox is to have the spine layer "leak" the specific RPD and DPIC prefixes that are advertised northbound from the RPAs and DPAs, respectively. As a result, in steady state, RPA and DPA nodes would receive both the aggregate route as well as the specific prefixes in its unicast table. In the failure state mentioned above, even if the aggregate were to disappear, the advertisement of the RPD and DPIC prefixes would be maintained, thereby keeping GCP communication alive.

d. Routing optimization for Precision Time Protocol (PTP) traffic – due to the sensitive nature of PTP and its low tolerance for jitter, it is important to ensure symmetry between the PTP clients (i.e. CCAPs, RPDs) and the PTP clock source. In Cox's R-PHY domain, this means symmetry should be maintained between each of the below entities:

1. Between CCAP and PTP source
2. Between RPD and PTP source
3. Between PTP boundary clock and PTP grandmaster

The above has been achieved in the Cox environment simply by adding each of the clock components, whether they are client or source, into IS-IS, the IGP used in the Cox metro network. Technically, regarding RPDs, the RPD prefixes themselves are not added to the IGP, but rather, the RPA nodes which serve as the next hop for RPDs, must be present in the IGP. This essentially accomplishes the same objective with optimal efficiency.

By having all the necessary components in the IGP, bidirectional symmetry is achieved between each segment of the PTP domain as each participant follows the least-cost path to the destination.

e. QoS – prioritization of traffic across various services. To ensure the delivery of priority traffic during times of congestion, it is critical to continually assess utilization throughout in the CIN for all types of services. It may be necessary, for example, to modify QoS parameters such as the "committed information rate" (CIR) and/or the "peak information rate" (PIR) buffer values, or perhaps even to assign traffic to additional queues, if available.

   At a minimum, the RPA and DPA platform should have the ability to classify and distinguish multicast and unicast traffic into separate queues. Furthermore, both multicast and unicast traffic should each have multiple queues available to isolate best-effort and priority traffic. In this way, the appropriate amount of bandwidth and buffering can be allocated to each queue.

5. Day 2 Approaches to Resiliency & High Availability

As of this writing, it has now been over seven years since Cox Communications first deployed R-PHY with its CIN architecture. As with all technologies, a refresh and reevaluation are periodically needed. From the initial deployment of R-PHY in the Cox network, the CIN has proven to be stable and resilient. But, at the same time, we have learned from various events that further issues needed to be addressed and additional tactics implemented to improve upon the stability and resiliency of the CIN.

The items covered in this section cover some of the relatively significant measures we have implemented in recent years to achieve increased high availability.

5.1 GCP timeout

As mentioned earlier, GCP is a critical component to maintaining availability in an R-PHY network. GCP is the protocol used for managing remote devices and it essentially keeps RPDs online; without GCP reachability between the CCAP core and any given RPD, the RPD would go offline and must reinitialize. Depending on the number of RPDs and controllers in the network, this reinitialization process can consume a significant amount of time, in the order of multiple hours in highly dense deployments. Therefore, it is imperative to keep GCP communication alive amid various CIN and general network-related failure events.

To this end, we collaborated with our partners in access engineering to allow the RPD timeout or threshold to be something that could be increased and manually set via CCAP configuration. The initial default setting was in the magnitude of seconds, and it has since been increased to the present value of ~1 minute. Although most link/node/protocol failures should converge in a matter of seconds (at worst), there have proven to be other variables at play that could result in an actual GCP unreachable state of significantly greater duration.

5.2 PTP design

PTP is another foundational protocol that maintains the underlay infrastructure and stability of the R-PHY environment. An unreliable PTP network could result in a very significant outage potentially affecting an extremely large blast radius.

Due to the criticality of PTP communication, Cox recently completed a redesign of its PTP architecture to address some vulnerabilities in the previous design and thereby, make the infrastructure much more resilient. In the previous design, the PTP infrastructure consisted of a 3-tier hierarchy, as shown in Figure 7.

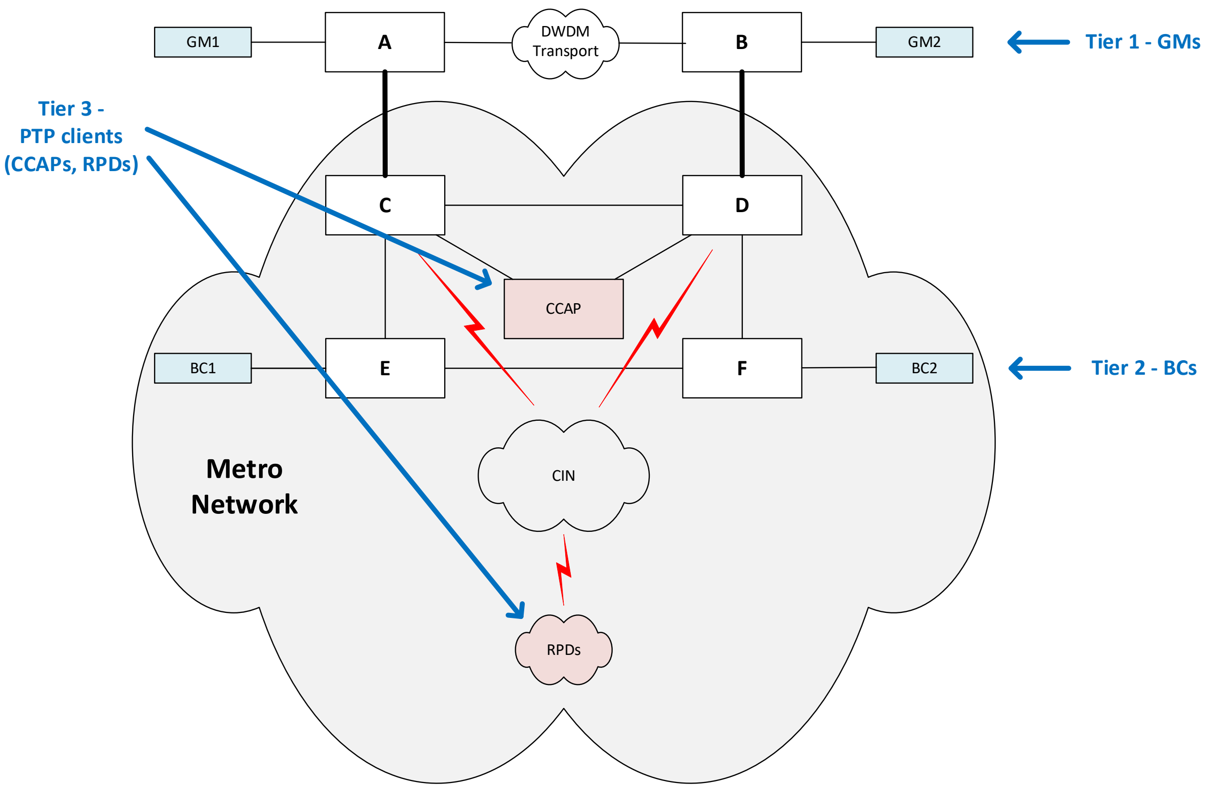

Figure 7 - PTP Old Design

The components of the 3-tier PTP architecture encompassed the following -- grandmaster (GM) clocks at tier 1, boundary clocks (BC) at tier 2, and the PTP client nodes at tier 3. In the R-PHY environment, the 2 main categories of PTP clients are CCAP routers and RPDs. Please note none of the core routers in the metro actively participate in PTP (i.e. they are not transparent clocks).

With the above topology in mind, Router A and Router B, in some cases, can be physically separated at significant distance from one another and utilize DWDM transport. Thus, this poses an inherent risk to the integrity of PTP. Namely, BC to GM communication is critical, and the reliability of that communication is dependent upon the stability of the network infrastructure, such as the links and nodes that reside between each BC and each GM. Link or node failures could result in degraded quality of the clock source and cause service impact due to RPDs and modems going offline.

To address this risk, a new PTP architecture has been implemented at Cox, where the former boundary clocks have now been transformed into hybrid clocks, each having its own GPS antenna as a directly connected clock source. A hybrid clock essentially serves as both a GM and a BC. It is a GM because it has its own local clock source (i.e. GPS), but it is still a BC of the original GMs if the local antenna were to fail. With this design, any network link or node failure on node A, B, C or D should not impact the integrity and consistency of the timestamp on the BC. The new design is now essentially a 2-tier hierarchy versus the former 3-tier model. The 3-tier model would only apply in the unlikely event that the GPS antennas of both hybrid clocks failed.

In this new design, even if the primary hybrid clock were to fail, the backup hybrid clock should have the same timestamp as the primary clock as they are both located in the same site/building. Also, since all wiring to both hybrid clocks are confined to local transport within the site/building, there is no impact due to varying transport distances. The new design is shown in Figure 10.

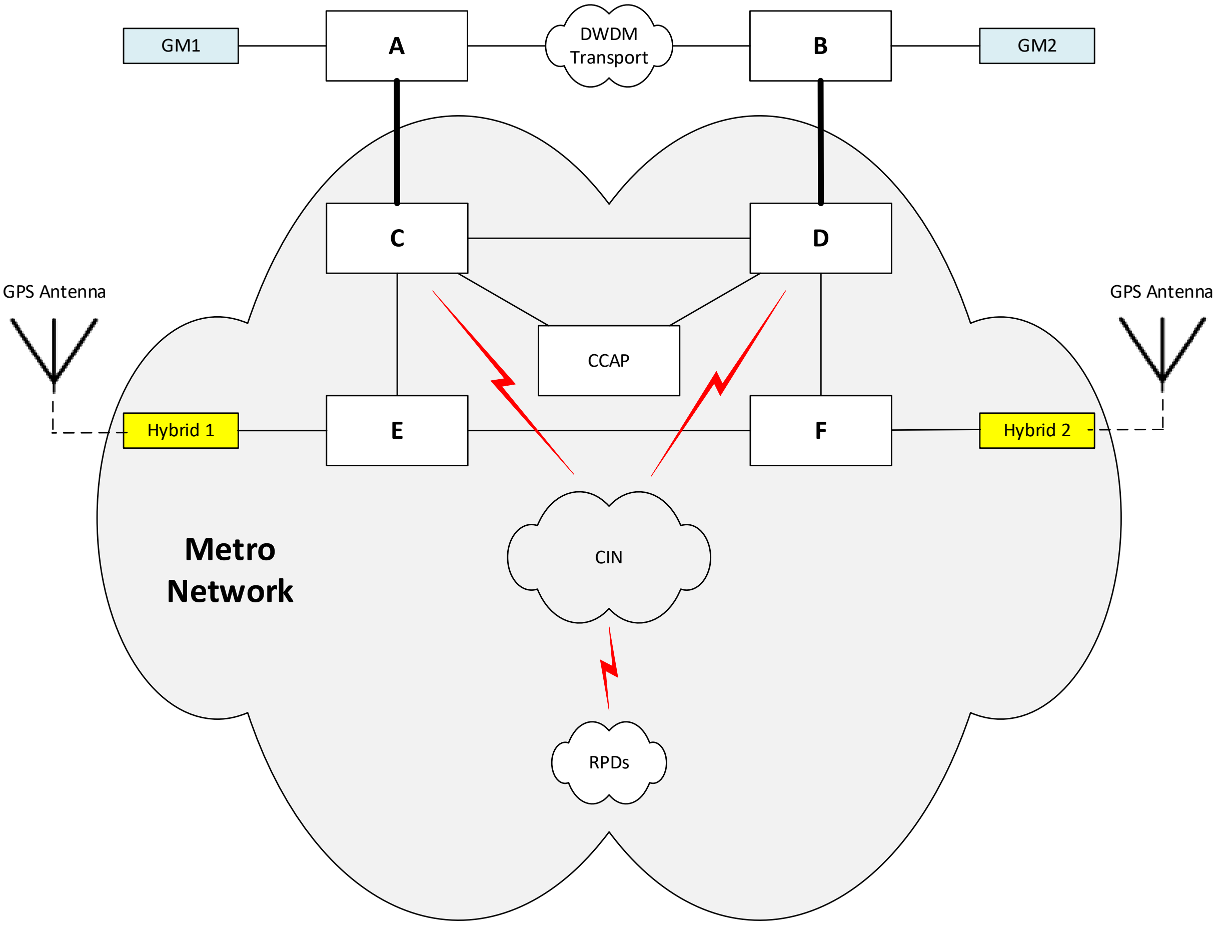

Figure 8 - New PTP Design

5.3 Proactive Network Health (PNH)

Proactive Network Health essentially describes the collection, analysis, and ultimate application of key data and metrics to provide an intelligent means for predicting and proactively acting upon irregular network events. In this way, certain outage events can be averted or at the very least, mitigated.

Cox has been able to improve our PNH capabilities via the use of streaming telemetry. The traditional means for obtaining metrics was via SNMP, which is a UDP-based "pull" model. This refers to the fact that a server must poll for the data it wishes to receive, and the action of transferring that data is initiated by the server. The server is the active party in the transaction. In contrast, streaming telemetry utilizes a "push" model, where data is actively sent by the monitored object (i.e.

the client) towards the monitoring system (i.e. the server). This results in more efficient transmission of data.

Currently, Cox is monitoring the following metrics via streaming telemetry:

a. CPU utilization
b. CPU memory
c. Switch Fabric memory
d. Route Processor memory
e. QoS buffering
f. QoS drops

These key indicators or metrics are given appropriate thresholds, whereupon if breached, an alert is sent to appropriate internal stakeholders. This allows them to evaluate the situation and proactively resolve or mitigate the situation in a timely manner.

5.4 Segment Routing

Segment routing (SR) is a forwarding mechanism that utilizes the concept of “source-based routing”, meaning the path to the destination is encoded in the packet header as “segments”. This is advantageous compared to traditional MPLS for several reasons. One, it is much less complex, as it does not require LDP or RSVP-TE; rather, it utilizes IGP extensions, so no new protocol is required. Second, it removes state from the network, as the path state is now moved to the packet header rather than on all the individual routers along the forwarding path.

At Cox, we have in recent years deployed SR-MPLS at the metro spine layer, as well as on many service layer routers. Since we have the infrastructure for SR-MPLS already implemented, it would behoove us to enable segment routing in the CIN. Again, since SR simply utilizes extensions on the existing IGP (i.e. IS-IS), it inherently supports both IPv4 and IPv6. Therefore, the fact that the Cox CIN is comprised solely of IPv6 prefixes poses no additional challenges to SR itself. In contrast, in a traditional MPLS environment, label allocation and distribution for IPv6 prefixes would require a completely different protocol, such as LDPv6.

The ultimate benefit of SR lies in the fast convergence that it provides, especially when utilizing TI-LFA (Topology Independent Loop-free Alternate). With TI-LFA, sub-50ms failover and repair can be achieved in the event of failure to the primary path. This is accomplished through the use of a pre-calculated backup path, which is essentially the MPLS equivalent to FRR (fast reroute).

The failover and use-case for SR is shown in the below diagrams (Figures 11 & 12) for upstream unicast traffic, and for control traffic such as GCP and PTP.

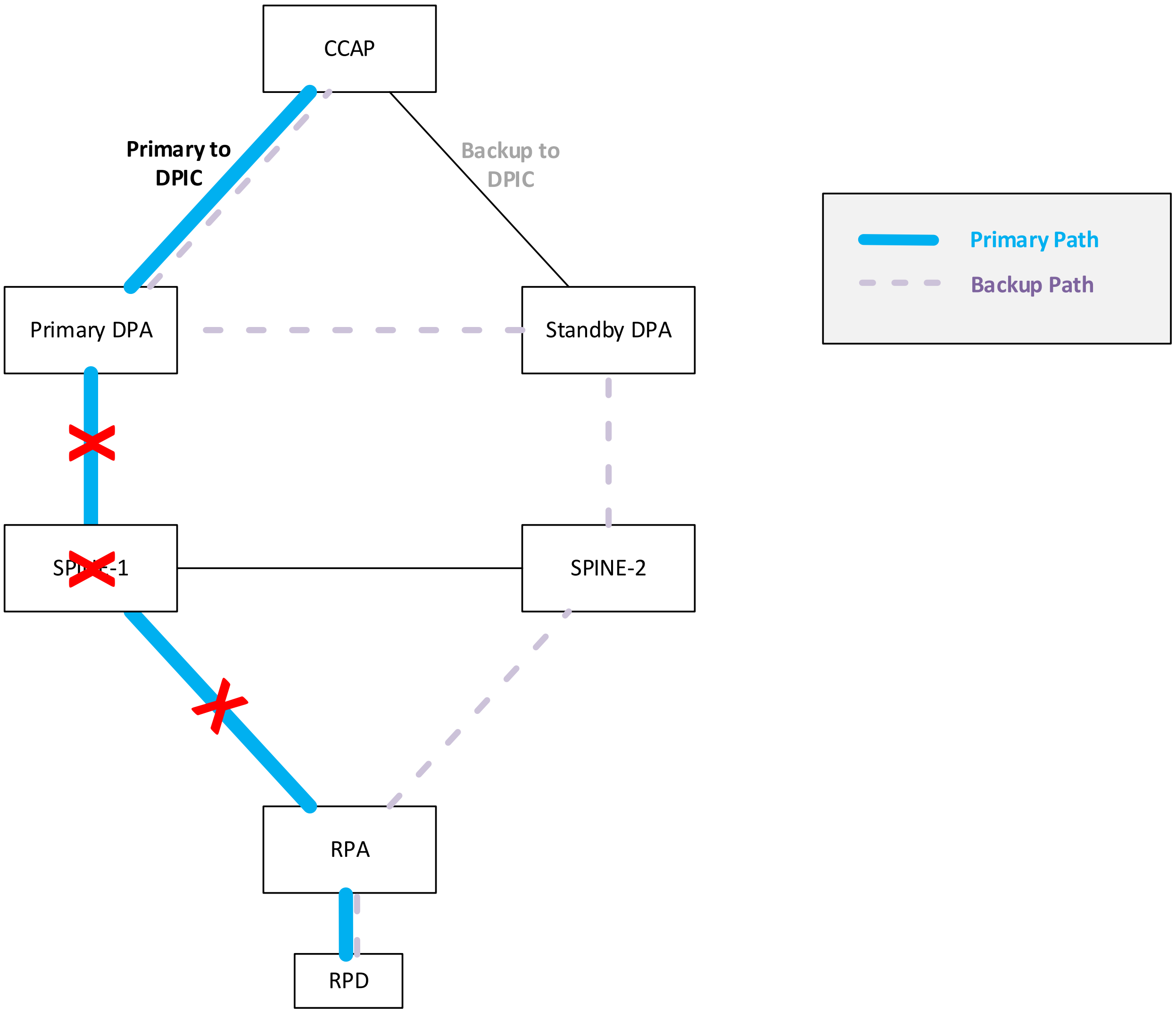

Figure 9 - SR-MPLS with Upstream Unicast and GCP

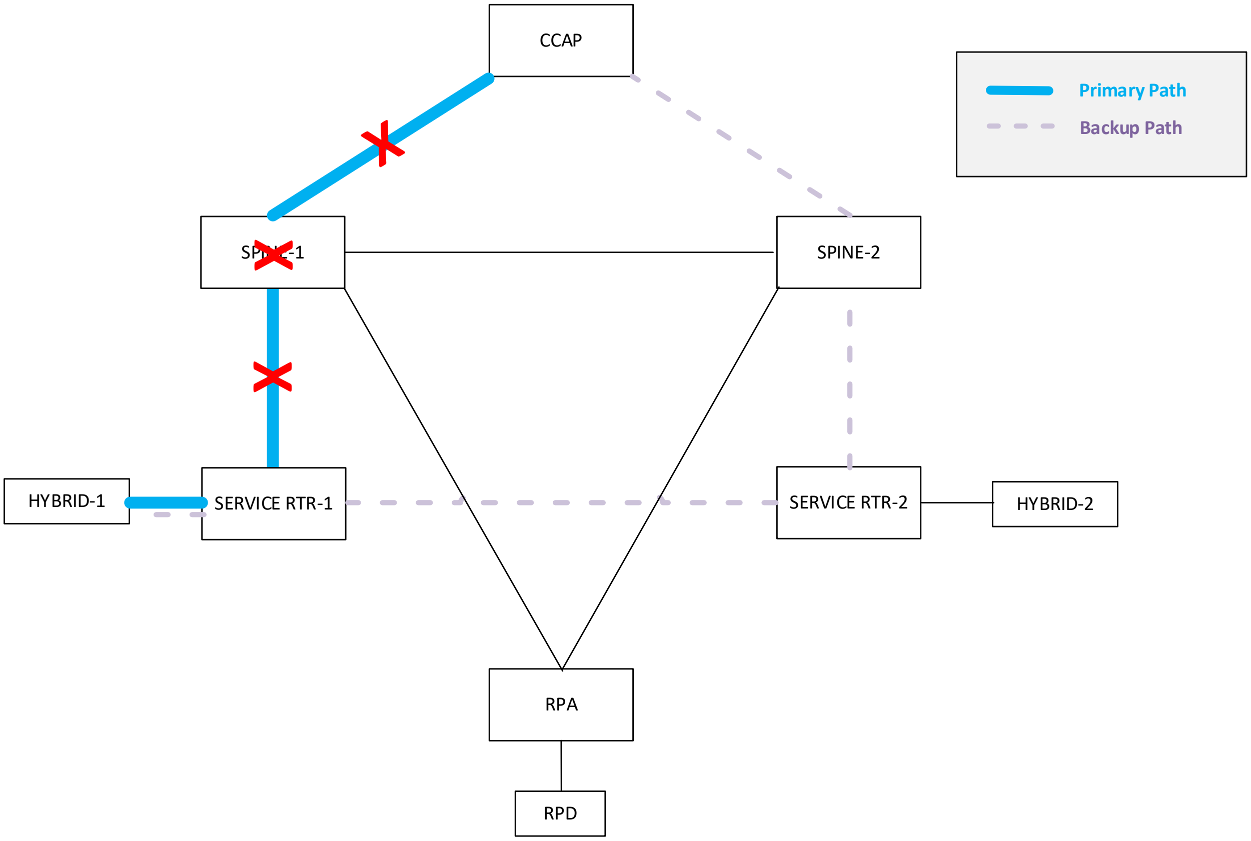

Figure 10 - SR-MPLS with PTP

It should be noted, the benefits of SR-MPLS would apply only to unicast traffic, not multicast. Unicast prefixes can be label-switched via SR, while at this point, multicast prefixes cannot. This is why, as the above diagrams illustrate, the benefits of SR are limited to any and all unicast traffic, such as control traffic as well as upstream unicast traffic. Even with this limitation, its implementation undoubtedly makes the CIN environment more resilient overall.

6. Conclusion

As of this writing, approximately 75% of Cox's residential footprint has been moved from the traditional analog CMTS platform to Remote PHY. And, of course, the expectation is that this number will continue to rise. The obvious impact of this steady increase is that the reliance on the CIN is heavier than ever before, which in turn means the integrity and resiliency of the CIN is more crucial than it has ever been. Ensuring high availability is of paramount importance as the stakes rise, as any given type of outage event will likely impact more and more customers.

From the time of initial deployment, Cox has utilized many of the well-known industry best practices, such as link, protocol, and hardware redundancy, where applicable. This has served us well over the years. However, with the ever-increasing stakes, we have, in recent times, assessed some additional measures, most of which have now been implemented into our production environment and proven to be highly successful. A major accomplishment we have recently integrated into our network has been solidifying the two major components for R-PHY integrity, which are GCP and PTP. This has been achieved by increasing the GCP keepalive value to make RPD deinitialization much less likely and by collapsing our PTP infrastructure to make it much less prone to asymmetry or jitter and any PTP related service degradation. On top of this, we have also significantly improved our PNH capabilities, which now allow for better predictability of service-impacting events and enable us to act more proactively. Finally, segment routing implementation is another substantial resiliency measure that is not far off. SR-MPLS together with TI-LFA will allow for extremely fast convergence in the event of a node or link failure in the network.

These strategies have and will continue to allow Cox to improve upon the already high level of resiliency we have experienced in years past, and they will position us to accommodate the continual and increasing transition of customers and services into the digital R-PHY environment in the coming future.

Abbreviations

BC	Boundary Clock
BGP	Border Gateway Protocol
CIN	Converged Interconnect Network
CCAP	Converged Cable Access Platform
CMTS	Cable Modem Termination System
DAA	Distributed Access Architecture
DPA	DPIC Aggregation Router
DPIC	Digital Physical Interface Card
FRR	Fast Reroute
GCP	Generic Control Protocol
GM	Grandmaster
IGP	Interior Gateway Protocol
IS-IS	Intermediate System to Intermediate System
LAG	Link Aggregation Group
MDM	Mux/Demux Module
MPLS	Multiprotocol Label Switching
OCML	Optical Communications Module Link Extender
PNH	Proactive Network Health
PTP	Precision Time Protocol (IEEE 1588)
RPA	RPD Aggregation Router
RR	Route Reflector
RRC	Router Reflector Cluster
QoS	Quality of Service
RPD	Remote PHY Device
R-PHY	Remote PHY
SR	Segment Routing
TI-LFA	Topology Independent Loop Free Alternate

Bibliography & References

Modernizing Cox Communication's Access and Aggregation Network Infrastructure for Remote PHY Deployment, Deependra Malla; 2021 SCTE CableLabs and NCTA