SCTE Technical Journal

Volume 6, Issue 1 | February 2026

SCTE Engineering Committee Chair

Bill Warga
Liberty Global

Publication Staff

Derek DiGiacomo
Senior Director, SCTE

Dean Stoneback
Senior Director, SCTE

Margaret Bernroth
Editor, SCTE

Editorial Correspondence: If there are errors or omissions to the information provided in this journal, corrections may be sent to our editorial department at journals@scte.org.

Submissions: Have an idea or topic for a future journal article? Send your submission to journals@ scte.org. All submissions are subject to peer review and publication is at the discretion of SCTE.

A Note from the Editor

Welcome to the Winter 2026 Issue of the SCTE Technical Journal. In this issue, we've sequenced the articles to flow from core network understanding to the experiences those networks enable. We begin with "*Processing OFDM Channel Estimate Data to Expose Secondary Resonant Cavity Peaks*", which strengthens the measurement and diagnostic foundation by showing how channel-estimate data can be processed to reveal otherwise hidden plant behaviors.

From there, we build directly into "*Operationally Aligned Artificial Intelligence (AI): Redefining Intelligence for the Physical Realities of Broadband Networks*," grounding AI in the practical constraints of broadband operations and emphasizing alignment with real-world telemetry, workflows, and outcomes.

Next, the issue transitions from insight to deployable architecture with "*Implementation Strategies for Private Secure Networks.*" This article focuses on practical approaches to creating and operating secure private connectivity, including implementation patterns and lab-based exploration of tunneling and VPN technologies paired with programmatic control and orchestration concepts.

We then widen the lens to the facilities that make modern connectivity possible with "*Renewable Power Recipes for Powering Telecommunication Hub Sites,*" which surveys workable renewable power options and procurement pathways to help operators

We close with the most audience-facing contribution, "*Maximizing Engagement Through QR Codes in Video Advertising: Innovative Solutions for Enhanced Viewer Interaction.*" This piece explores QR codes as an interactive call-to-action mechanism in video workflows and considers practical presentation choices—including reduced-motion variants—that can influence delivery efficiency and viewer response.

Taken together, the issue offers a clear arc: better visibility into network behavior, smarter operational decisioning, secure service implementation, resilient and sustainable infrastructure, and —finally—new ways to connect with audiences.

- Margaret Bernroth

SCTE STANDARDS JOURNAL

SCTE a subsidiary of CableLabs®

Processing OFDM Channel Estimate Data to Expose Secondary Resonant Cavity Peaks in the Time Domain, Including Stitching Two OFDM Channel Estimates

A Technical Paper prepared for SCTE by

Thomas J. Kolze, Ph.D.
Broadcom, Inc.
tom.kolze@broadcom.com

Alexander “Shony” Podarevsky
Promptlink Communications
shony@promptlink.com

Ron Hranac
Retired
rhranac@aol.com

Table of Contents

List of Figures

Introduction

In the early days of proactive network maintenance (PNM), cable operators used single carrier quadrature amplitude modulation (SC-QAM) channel adaptive equalizer or pre-equalizer coefficients to 1) derive the frequency response of the adaptive equalizer and 2) determine the presence of a resonant cavity (a.k.a. "echo tunnel", see [Broadband]) in the cable network and the approximate distance between impedance mismatches at the ends of the resonant cavity. The distance resolution was largely limited to adaptive equalizer tap spacing in the time domain, yielding, for example, the equivalent of about 85 feet per pre-equalizer tap in a 6.4 MHz-wide upstream SC-QAM channel. Some improvement in resolution was possible with parabolic interpolation, but was still limited. Some test equipment can display downstream adaptive equalizer data and taps (and other parameters) to support distance estimates, but the resolution is also limited. With the deployment of wide-bandwidth downstream orthogonal frequency division multiplexing (OFDM) and upstream orthogonal frequency division multiple access (OFDMA) channels used in DOCSIS® 3.1 and later technology, the resolution was significantly improved.

This *SCTE Journal* technical paper describes using multiple OFDM channels (or segments) to "synthesize" an even wider channel in terms of better time domain resolution. The approach follows the same concept as synthetic aperture radar (SAR), which uses multiple small aperture antennas dispersed over a larger area to synthesize a large aperture antenna.

The approach for combining multiple OFDM channels to synthesize the time domain impulse response of a wider channel has been denoted "stitching."

The processing steps for stitching two OFDM channels' channel estimates are listed and explained, followed by presentation of equations defining the processing.

Throughout this paper the term "OFDM channel" means a DOCSIS channel using OFDM technology. The described approach is applied to stitching the time domain impulse response(s) of either downstream OFDM channels or upstream OFDMA channels.

Throughout the paper stitching may apply to contiguous segments of an OFDM channel separated by excluded subcarriers or to actual separate OFDM channels, or a combination of both segments and different channels.

This paper highlights three points to note regarding implementing the stitching process.

1) The phase tilt-versus-time that multiplies the time domain impulse response of all but the lowest frequency segment, depends on the separation frequency of the higher frequency segment and the lowest frequency segment. The wrinkle is that the actual calculation or equation for the separation frequency depends on how the inverse Fourier transform algorithm treats what it uses as "0 Hz."

2) Phase-versus-frequency variation in the reflection coefficient of any significant reflection, such that there is a difference in the phase of a reflection between the two OFDM channels, can be problematic in terms of stitching effectiveness in obtaining finer resolution of secondary peaks of the impulse response. (A linear change in phase-versus-frequency, i.e., a tilt, is not problematic and simply acts as more or less path delay for the secondary peak.)

3) We can always ignore a unit magnitude, complex exponential that simply adds phase tilt-versus-time, when it is a leading coefficient multiplying the entire time domain impulse response.

Regarding #1, for the stitching to work properly, the calculation of the “separation frequency” has to match whether the inverse Fourier transform routine maps 1) the lowest frequency of the segment to 0 Hz; or 2) the center frequency of the segment to 0 Hz.

Regarding #2, the benefit of stitching may not be realized if the phase variation-versus-frequency of a reflection is large.

In addition to the above, this paper provides a short list of the basic steps involved in preparing each individual segment’s frequency response into a time domain impulse response which is suitable for stitching. The basic processing steps for the single OFDM channel estimate processing are listed and explained. Several references for single channel processing are cited. The single channel processing steps described in this paper are modified and updated compared to those references. The updated steps of the basic processing as described in this paper were developed and refined in collaboration with SCTE’s Network Operations Subcommittee Working Group 7 (NOS WG7) and its subgroup that has been working on the original CableLabs MIND^{TM} initiative.

As of the writing of this paper, the CableLabs PNM Working Group has been developing software and methodology for PNM and various network operations use cases including topology discovery and validation. Some of the pre-processing calculations are discussed in this paper. The open-source licensed software is available to members of that working group, but is not yet available publicly. Those interested in learning more about or accessing the software should consider joining the CableLabs PNM Working Group. To become a member, go to https://www.cablelabs.com/working-groups to learn more. If you are already a CableLabs member, you can ask to join by selecting the PNM working group on the CableLabs Causeway site.

The authors are grateful to the members of CableLabs’ PNM Working Group and SCTE’s NOS WG7 and their valuable contributions to the techniques discussed in this paper. Special thanks to James Medlock from Akleza for inspirational ideas in figure generation and baseline data. The authors are also grateful to Jason Rupe for leading the PNM Working Group and many editing sessions, and to Rich Prodan, for valuable feedback during peer review.

Notes on Stitching Time Domain Impulse Response from Separate Channels or Segments

This section explains the pre-processing steps with a deeper explanation of the signal processing engineering involved, extending to methods that allow combining time domain impulse responses. These steps have been demonstrated and tested extensively using channel estimation data, but apply equally to pre-equalization data though one may want to invert the pre-equalization data first so that it represents the channel.

Synthetic aperture radar informs our approach to combining channels (or contiguous complex frequency domain bin data sets). With SAR, the dispersed antennas’ responses are combined to provide a synthetic aperture wider than achieved by any one antenna (describing the approach in a nutshell). Applying this principle to the combining of frequency-separated OFDM channels, each contiguous span of frequencies containing a channel estimate is treated individually with some processing, then generating the time domain impulse response for each, and finally combining those separate time domain impulse responses synthetically.

So, in the following section, first the basic steps for processing a single channel's impulse response are provided. Following the overview of the steps, a few pages are given to additional explanation of the steps and their nuances and justifications (to some degree). The goal of the basic steps outlined is to remove the frequency domain magnitude and phase tilt across the channel, because even a small amount of either tilt (versus frequency) introduces a significant amount of "tilt clutter" in the time domain response, which often obscures important secondary time domain impulse peaks. Then, a thumbnail description of the stitching processing is provided. Finally, details and examples of the stitching processing are given.

Steps for Processing a Single Channel's Impulse Response

Processing separately on each contiguous span of frequency response consists of following steps (the order of #b and #c can be commuted):

a) remove anomalies, for example, such as at band edges.
b) tilt removal of the magnitude in the frequency domain (to reduce "tilt clutter" in the time domain);
c) tilt removal of the phase in the frequency domain, which is identical to shifting the impulse response in the time domain, in an attempt to make the largest time domain impulse response value (the primary impulse peak) occur at time t = 0;
d) apply a window (Hanning), by multiplying the frequency response by the window values.

After applying the window,

e) a second round of magnitude tilt correction;
f) a second phase tilt removal (in the frequency domain);
g) then normalizing the time domain values such that the value at time t = 0 is 1 + *j*0;

Additional optional analysis steps (not needed for stitching):

h) evaluate the time domain impulse response and identify the secondary impulse peaks;
i) evaluate the time domain impulse response in the "full" neighborhood of the secondary impulse peaks, to do fine time resolution of the secondary peaks (optional).

Step g is the same as normalizing the impulse response in the frequency domain to have an average value of 1 + *j*0 because the average of the frequency domain values is exactly the time domain value at time t = 0 (by definition of the inverse Fourier transform).

Also, some refinements of the basic operation or processing on a single channel have been identified since the references cited were published, in particular the need for the second step of tilt adjustment (phase tilt, in the frequency domain, or time shift in the time domain) after application of the window, and a second step of normalization of the value at time t = 0.

A frequency domain plot of an example channel is provided in Figure 1. Each channel is individually processed to accomplish the following:

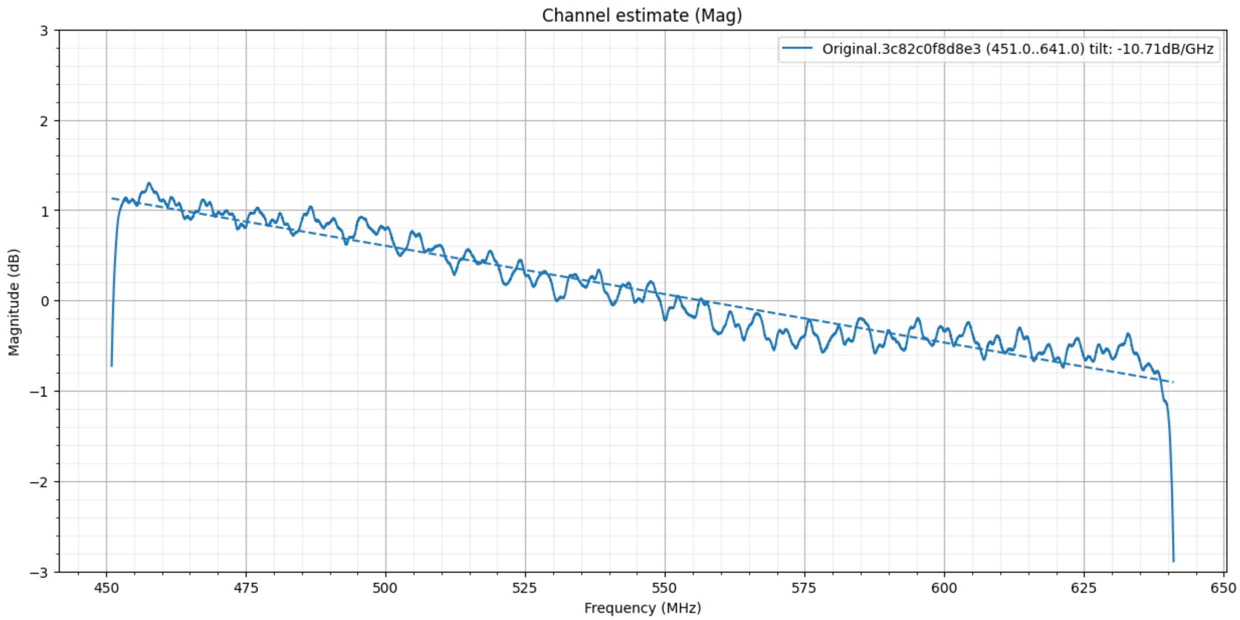


Figure 1 - A frequency plot of a single channel estimate magnitude (dB), before processing.

Channel estimates in the frequency domain often contain anomalies such as magnitude roll-off, bathtub edges in the group delay variation (GDV), artifacts or jumps, etc. It is important to remove identified anomalies before processing data further. More explanation is given in Section 2.2.

If these characteristics are observed, the indicated first step is to remove band edge channel frequency domain response where these deviations are non-negligible. In other words, put "zero" (not 0 dB but actually "zero" in linear value) for the channel impulse response in the frequency domain for those frequencies/subcarriers. This is "narrowing" or reducing the frequency domain of this channel or channel-segment (if there is exclusion band, this is widening the exclusion band for the purpose of this analysis).

If removing an impairment captured in a certain frequency band in the middle of the channel, one must remove the data over the impairment frequencies, then treat each portion of the channel as a single channel, then later to stitch the channels together as described in Sections 2.3 and 2.4.

Remove magnitude tilt.

It is not important to make the "endpoints" of the contiguous span of frequencies "match" or equal each other, because windowing will be applied; the main concern here is removing the magnitude tilt across the middle (frequency domain) of the sampled-channel.

The authors recommend removing tilt using linear magnitude (non-dB) values.

Efficient removal of gross phase tilt across the channel, to shift the primary peak of the time domain impulse response near time t = 0.

Unwrap the phase (see Figure 11) and determine the average phase tilt across the band (i.e., the phase accumulated across the band divided by the bandwidth).

Remove the average phase tilt across the band from the frequency response data.

Figure 2 shows the data from Figure 1 but after phase correction and normalization of the levels. Then Figure 3 shows the data from Figure 2 but after removing the remaining magnitude tilt. Figure 4 combines the plots from Figures 1 though Figure 3 for comparison.

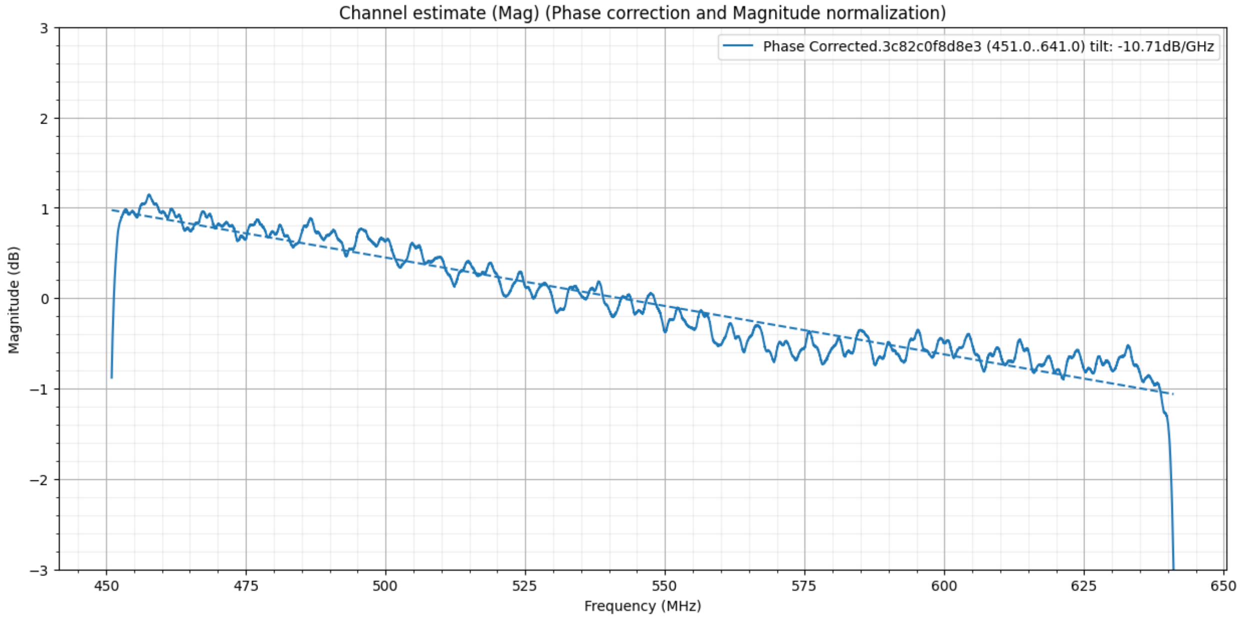


Figure 2 - A frequency plot of the channel estimate magnitude (dB) after phase correction and magnitude normalization.

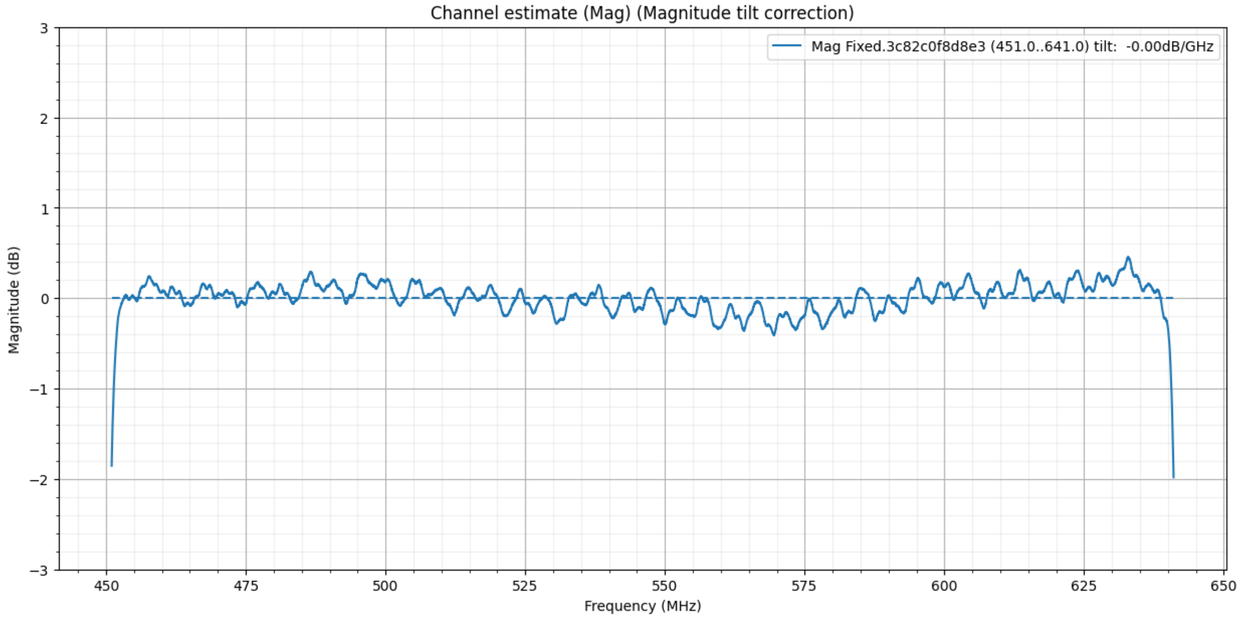


Figure 3 - A frequency plot of the channel estimate magnitude (dB) from Figure 2 but after removal of the magnitude tilt.

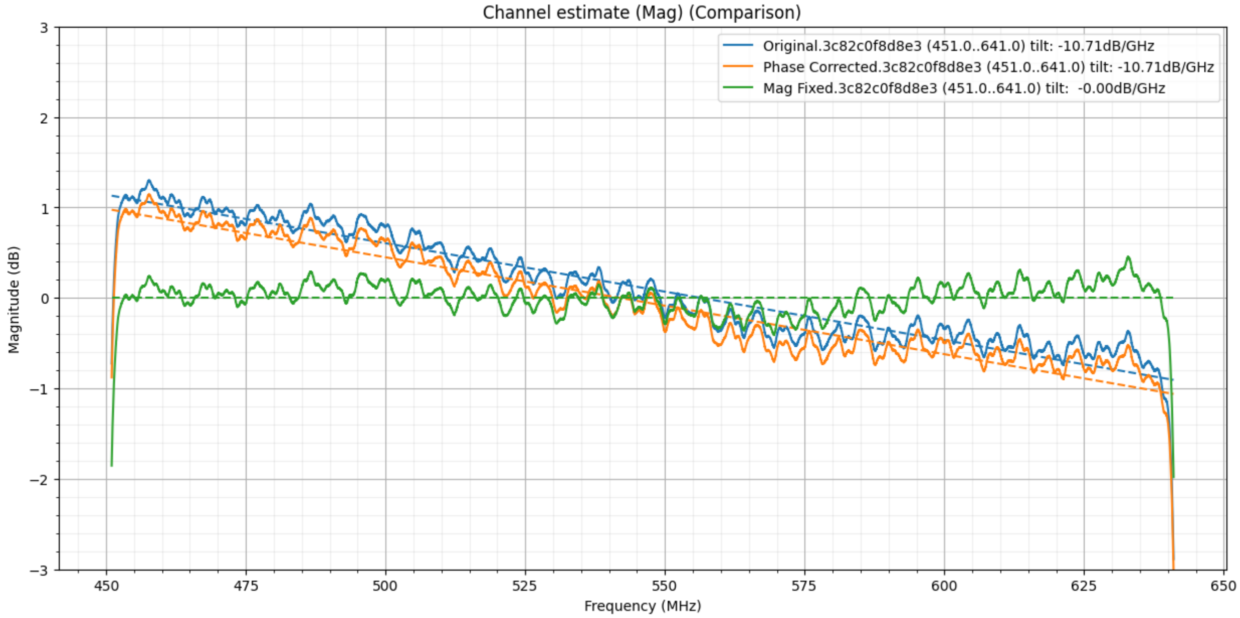


Figure 4 - A combination of Figures 1 through Figure 3 plotted together for comparison.

Figure 5 shows the original magnitude of the impulse response (blue trace), the magnitude of the impulse response after the first phase correction (orange trace), and the magnitude of the impulse response after the first magnitude correction (green trace).

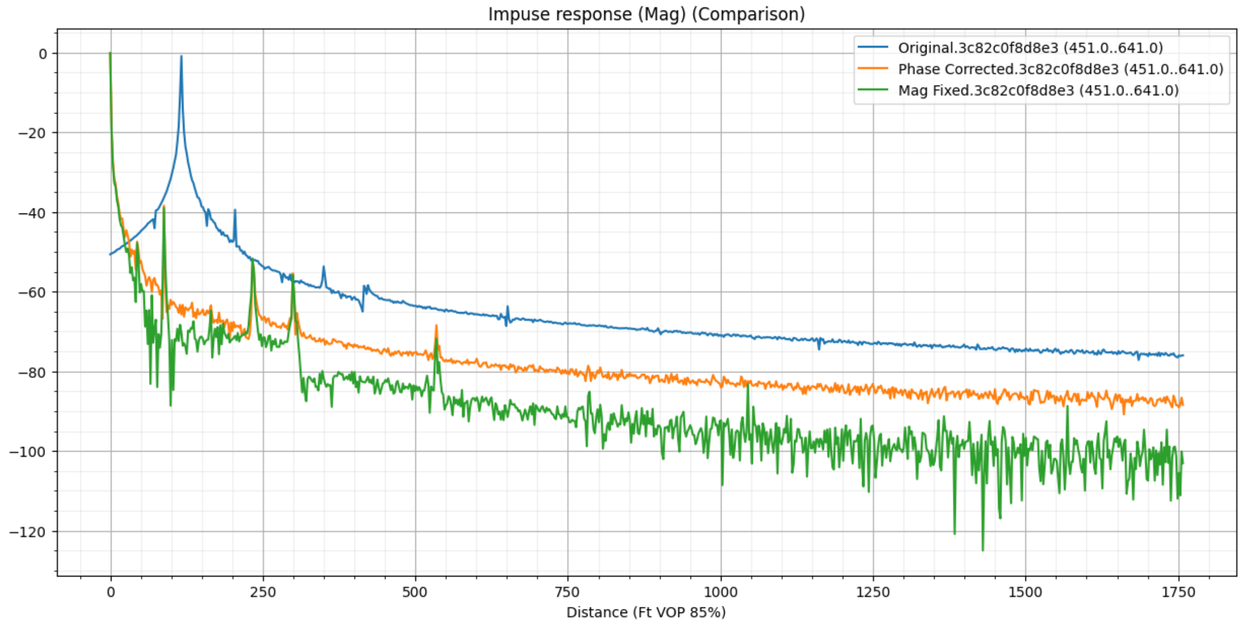


Figure 5 – Magnitude of the time domain impulse response before and after first corrections (displayed in distance instead of time for the first 1750 ft).

Apply the window (multiply the frequency response with the window values).

Window the frequency domain response to mitigate the impact of the abrupt transition from "sampled channel response" in the frequency domain to frequency domain samples with "no channel estimate."

In some implementations, a normalization of the time domain impulse response may have already been applied, and a "gross" time shift has been applied in a previous step. Note that applying the window will impact the time of the peak of the primary impulse, and will possibly impact the value of the peak of the primary impulse; this means another step is required for time shift adjustment and normalization after applying the window.

After applying the window, be aware that the application of the window potentially changes the magnitude tilt of the frequency domain samples. It is important to **re-calculate magnitude tilt and remove it.**

Figure 6 shows the data from Figure 3 after applying a Hanning window, which is the recommended window. When applying the Hanning window, that window should be scaled for the frequency span of the channel estimate after any band edge exclusions. Note that the Hanning window is applied to the frequency domain linear complex values after 1 + *j*0 is subtracted from each frequency sample. This is effectively removing all the energy at the time t = 0 sample in the time domain because of the normalization previously implemented. By doing this, the window in the time domain does not convolve and spread the large primary impulse at time t = 0. After applying the window in the frequency domain the 1 + *j*0 values are added back. Additional details about windowing are beyond the scope of this paper, and could be the topic of a future *Journal* contribution.

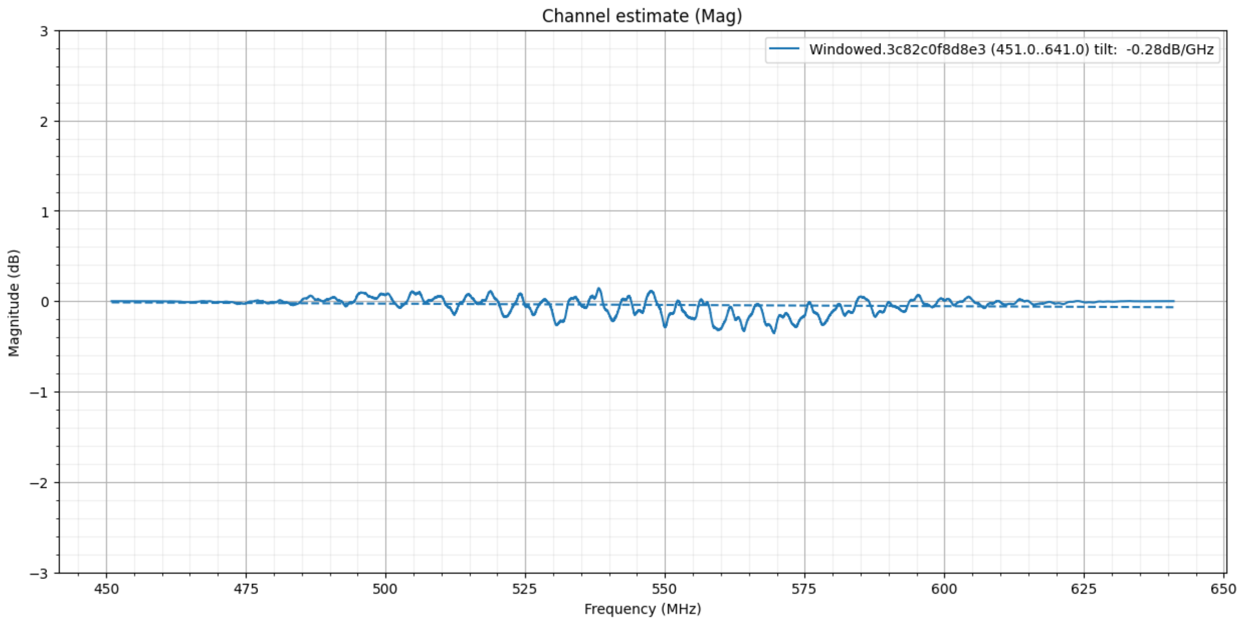


Figure 6 - A frequency domain plot of the data from Figure 3 after application of a Hanning window.

Calculate the time domain impulse response at fine time resolution, near the time t = 0.

To hone in on a precise time value for the primary impulse peak, we could just evaluate the time domain impulse response in the neighborhood of time t = 0, (evaluate a few points around t = 0).

Performing an upsample in the frequency domain, using zero padding with sample size equal to a power of two (2^N), is an optional step which makes it possible to use both fast Fourier transform (FFT) and inverse FFT (IFFT) algorithms in further analysis.

Note: Upsampling using zero padding, is a technique typically employed to make the size of the input sequence equal to a power of two. In the context discussed in this document, the zero padding is performed on a frequency series.

Figure 7 shows the data from Figure 6 after upsampling to a full FFT length (4k or 8k for 50 kHz or 25 kHz subcarrier spacing respectively).

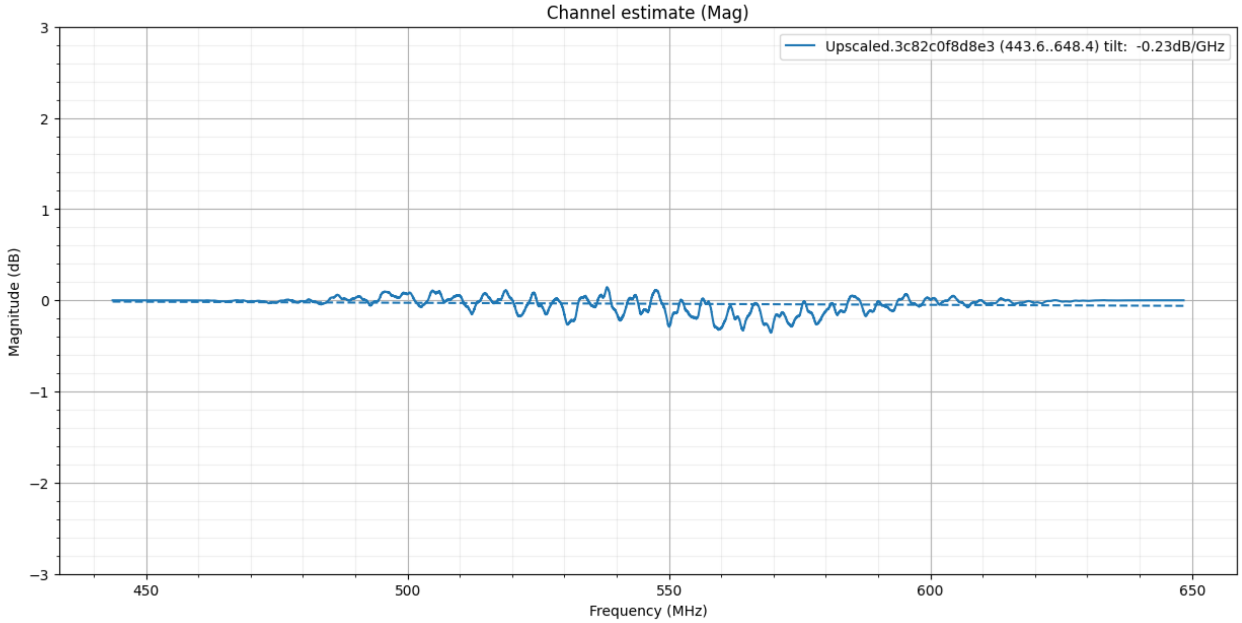


Figure 7 - A frequency domain plot after upsampling the data plotted in Figure 6.

A second magnitude tilt correction should now be applied, using the previous window function for the weighted least squares fitting. (This step is required whether or not upsampling was performed.)

The next step is to precisely locate a near-exact time of the primary peak of the time domain impulse response.

This time offset from t = 0 will determine another phase tilt removal in the frequency domain values, to shift the precise time peak to t = 0. As an alternative method (not discussed in this paper), a weighted least squares fitting can be used to remove the phase tilt.

Apply the indicated phase tilt removal.

Normalize the value of the peak at time t = 0 to 1 + *j*0.

Because the window has already been applied in the frequency domain, the best way to accomplish this normalization is to multiply each frequency domain value by the complex reciprocal of the time domain complex value at time t = 0. The most efficient way to obtain the time domain value at time t = 0 is to average the windowed frequency domain complex values.

Figure 8 shows the data from Figure 7 after an additional magnitude and phase correction.

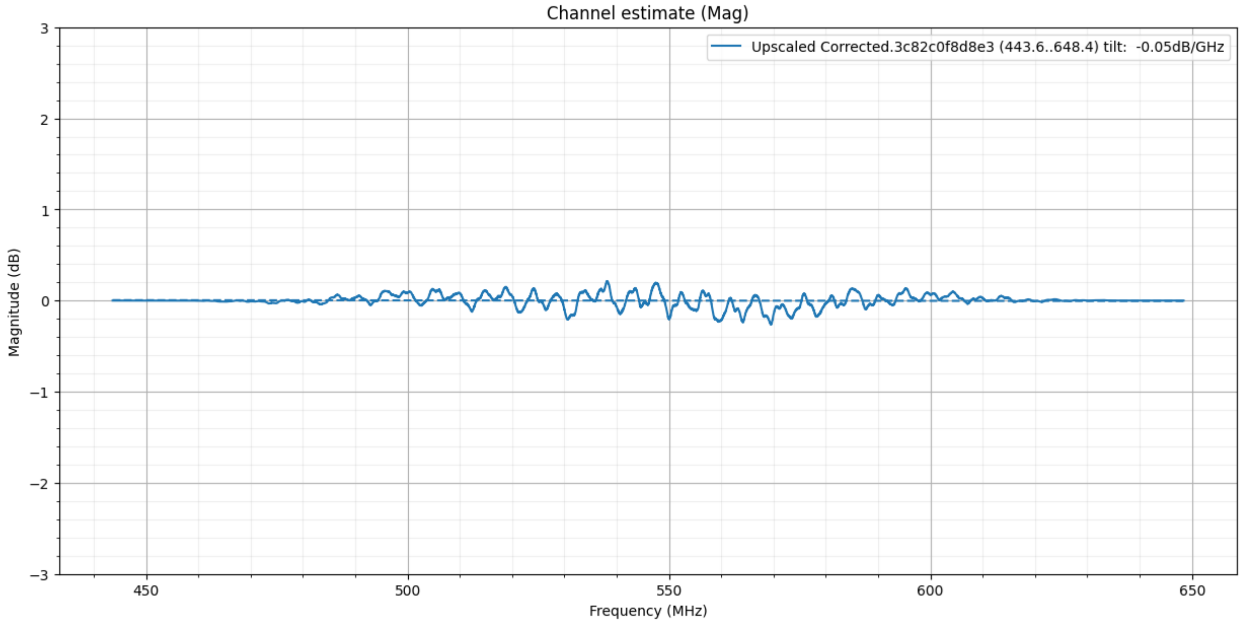


Figure 8 - A frequency plot of the data from Figure 7 after an additional magnitude and phase correction.

Evaluate the time domain impulse response in the “full” neighborhood of the secondary impulse peaks.

If a large number of time domain samples is desired to be computed, the authors recommend using the previously stated optional step of upsampling.

Figure 9 shows an impulse response plot (time domain) of the data from Figure 8 with bin indexes converted to distance for velocity of propagation 85% and limited to the first 300 bins, while Figure 10 shows the same data from Figure 9 with the peaks identified.

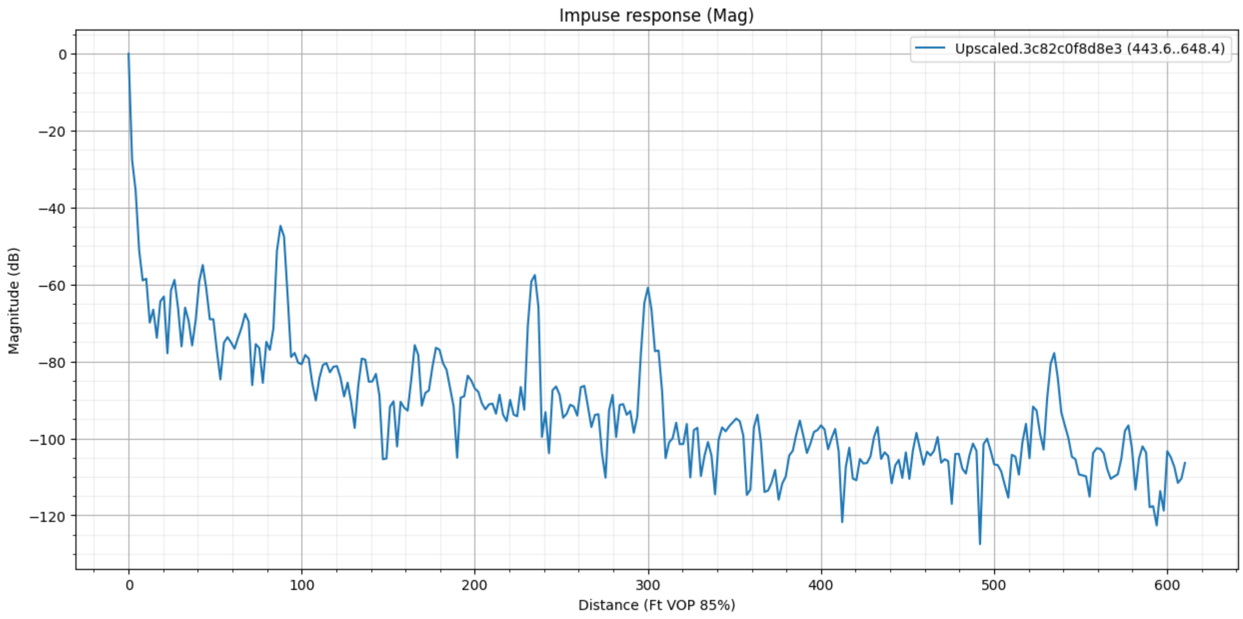


Figure 9 - An impulse response plot (magnitude over distance with velocity of propagation 85%) of the frequency response from Figure 8 limited to first 610 feet (300 bins).

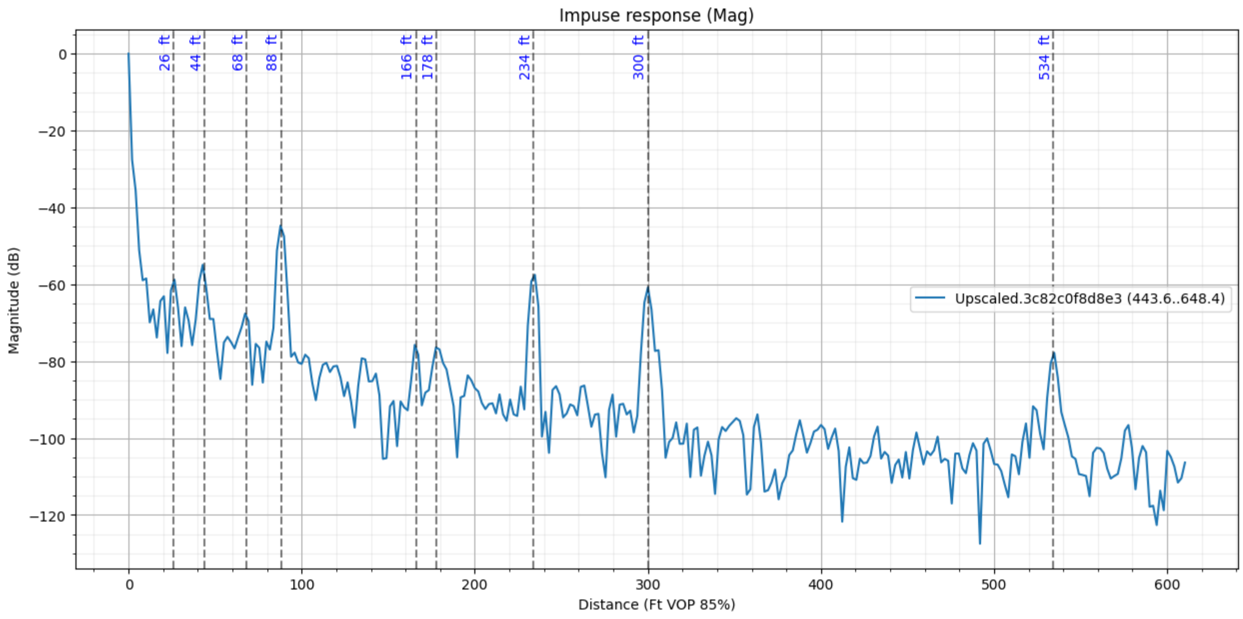


Figure 10 – Same impulse response as Figure 9 with the peaks identified..

Detailed Explanation of Elements of Processing

The first frequency domain steps discussed earlier of tilt removal (magnitude and phase) do not commute with windowing. It is important to be mindful of this. The processing steps for generating a "clean" time domain impulse response, to enhance identification of secondary impulse response peaks due to resonant cavities, listed in Section 2.1 and described here in more detail. “Cleaning” here means reducing the “tilt clutter” occurring in the time domain impulse response which occurs when there is magnitude and/or phase tilt in the frequency domain. These steps are revised from the steps given in the references, with additional detail provided (anomaly removal, band edge exclusion criteria, windowing details, etc.), and with the addition of the second tilt removal and normalization steps.

For step #a in Section 2.1, it is important to identify and remove anomalies in the frequency response, such as band edges(s) with roll-off, “bathtub” edges in GDV or any other anomalies or artifacts. If these characteristics are observed, they need to be mitigated. To address roll-off or band edge distortions, the affected frequencies have to be excluded.

To know how much to remove, one may begin removing at the edge and work toward the center of the channel, checking until removal of more data no longer impacts the results significantly. The “results” to examine are the slope of the tilt (is it changed compared to the previous exclusion amount?) and the residual squared error of the overall “fit” of the tilt to the frequency response (examining both the magnitude and the phase “fits” separately, if least squares fitting is used for the phase tilt).

A second step's (step #b listed in Section 2.1) primary goal is to remove linear magnitude tilt across a significant portion of the processed channel in the frequency domain to reduce “tilt clutter” in the time domain. The removal of the linear magnitude tilt should use linear regression to find best fit with least squared error. Note that magnitude tilt removal needs to be done in linear magnitude (non-dB) values. However, examples of tilt removal using the decibel values have been successful in generating “cleaned” time domain impulse responses.

The next step (step #c listed in Section 2.1) is to remove linear phase rotation (tilt) in the frequency domain across the channel, effectively positioning primary impulse in the time domain near time t = 0. Note that in [Williams] this phase tilt is referred to as phase rotation. When plotting channel estimate in polar coordinates it appears as a rotation in the IQ plane, and speed and direction of the rotation are proportional to phase tilt-versus-frequency. Although the time domain primary peak occurs at a time corresponding to a weighted average of the frequency response phase (i.e., the magnitude values used as “weights” for the phase values at each frequency), it has been shown that using unweighted phase often produces adequate results for positioning the time domain primary impulse near time t = 0. Due to the periodic nature of the phase (±180º), it may have multiple periods across the channel (up to several thousand times) and it is important to unwrap it to calculate corrections.

What we mean by unwrapping: undoing the jumps introduced by the modulo 360° (±180°) function and recovering the continuous phase-versus-frequency characteristic. For more information, see [Williams] and [Wolcott], and the examples in Figure 11 and Figure 12.

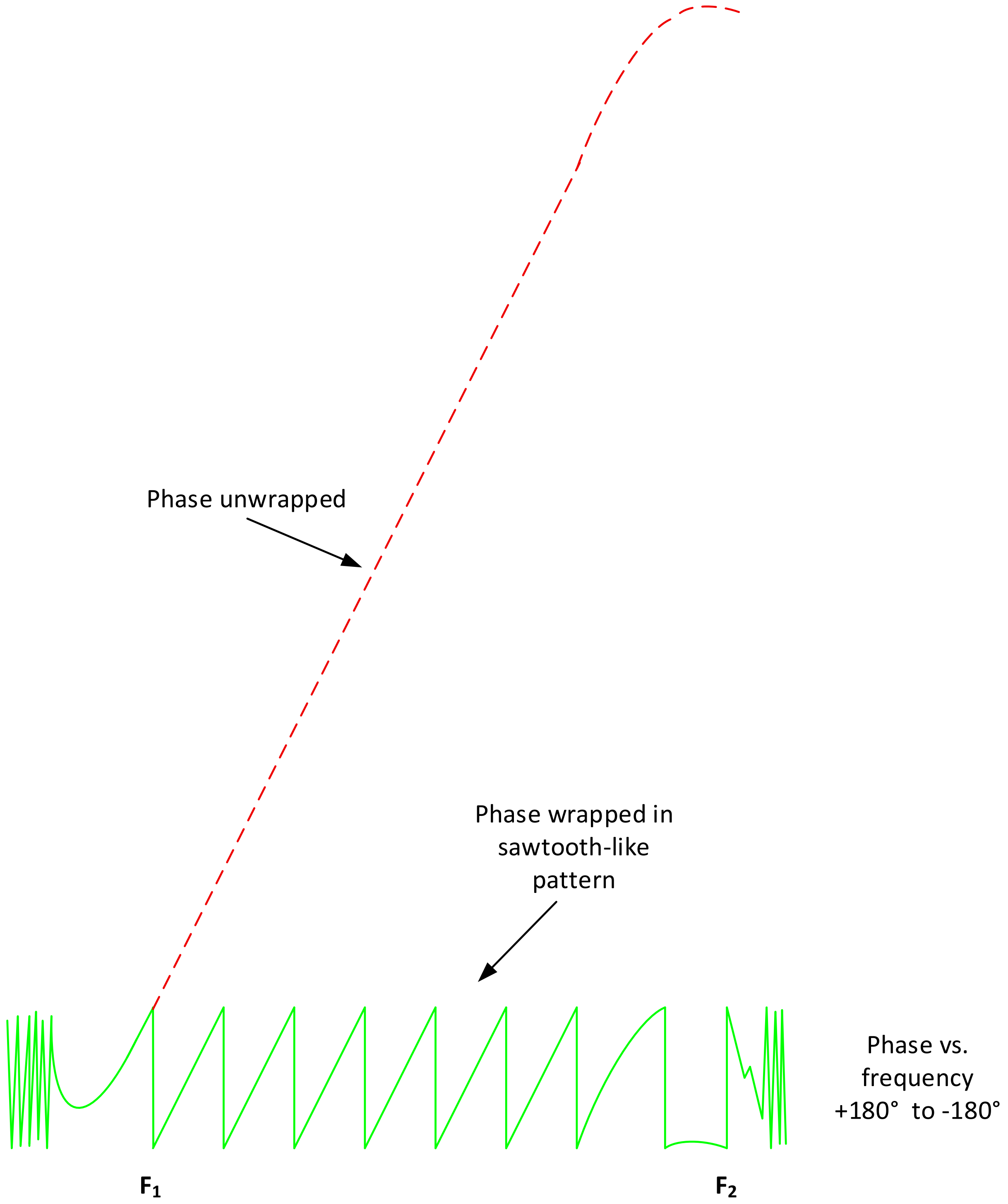

Phase unwrapped
Phase wrapped in sawtooth-like pattern
Phase vs. frequency +180° to -180°
F_1
F_2

Figure 11 - Graphical explanation of phase unwrapping.

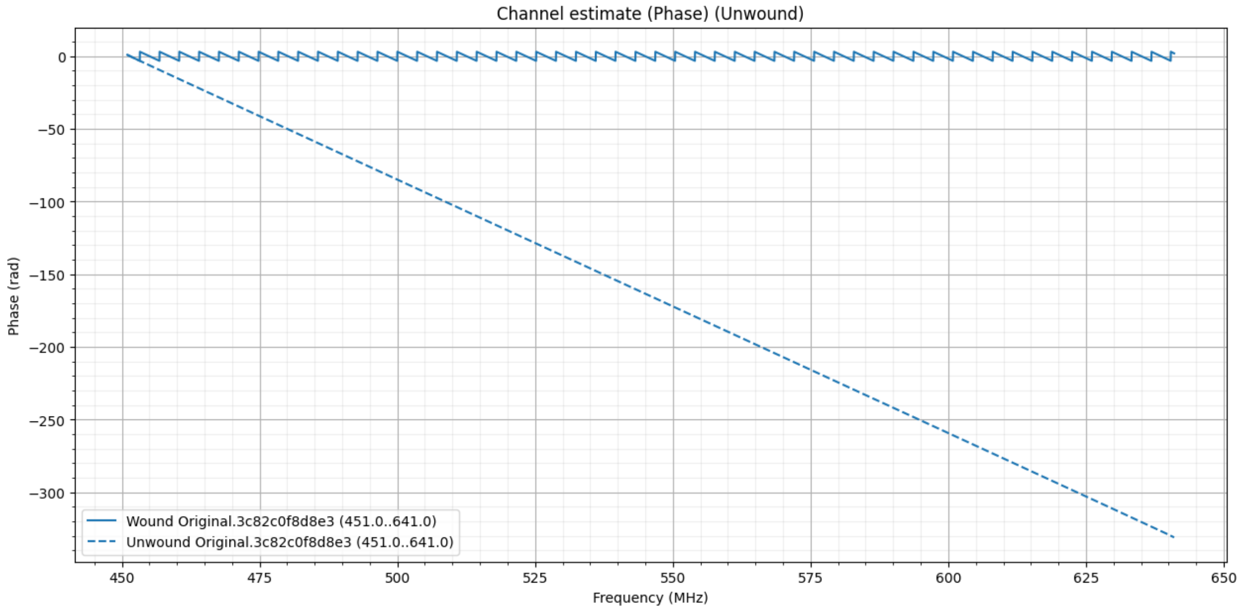


Figure 12 – Example of phase unwrapping from a real cable modem.

We have now completed steps #a, #b, and #c discussed in Section 2.1. If the intention is to eventually apply a window in examining the "cleaned" time domain impulse response, then now is a recommended time to apply the window (step #d listed in Section 2.1). This is prior to fine-tuning the time domain impulse response primary peak at time t = 0, and normalizing the primary peak to 1 + *j*0.

The reason for applying the window now, after removal of magnitude tilt and removal of the tilt of the gross (unweighted) phase tilt versus frequency, before fine-tuning the phase tilt removal (to position the primary peak at time t = 0 precisely) and before normalizing to 1 + *j*0:

> *The application of a window will (potentially/likely) change the precise time of the primary peak and also change the size and angle of the primary peak***.** That is, a normalization to 1 + *j*0 prior to application of a window will result in a value at time t = 0 different from 1 + *j*0 after application of the window! The average of the values in the frequency domain is potentially changed by applying the weighting of the window and then averaging after applying the window.

Note that as described in Section 2.1, we have created a unique method of applying windowing, where the value at time t = 0 is "removed" for applying the window, and then reinserted; see Section 2.1.

After applying the window, it is important to re-calculate magnitude tilt and remove it (steps #d and #e listed in Section 2.1) and normalization of time domain values with time t = 0 equal to 1 + *j*0 (step #f listed in Section 2.1). This is because use of the window potentially changes the magnitude tilt of the frequency domain samples.

After applying the window in the frequency domain, and removing any residual magnitude tilt, calculate the time domain impulse response at fine time resolution, near the time t = 0 (step #g listed in Section 2.1). This should be carried out to precisely locate a near-exact time of the time domain primary peak. This time offset from t = 0 will determine another phase tilt removal in the frequency domain values, to shift the precise time peak to time t = 0.

> A reminder that upsampling frequency domain samples using zero values is optional at this point, since only a few dozen or more fine-time domain values are needed near time t = 0. It is not necessary to use IFFT to get to the time domain, for just the purpose of calculating the time shift needed to place the primary impulse peak at time t = 0.

At this point it is time to normalize the value of the peak at time t = 0 to 1 + *j*0. Because the window has already been applied in the frequency domain, the best way to accomplish this normalization is to multiply each frequency domain value by the complex reciprocal of the time domain complex value at time t = 0.

The complete time domain impulse response may not be needed. It is possible that evaluation in the neighborhood of secondary impulse peaks is sufficient. However, for clustering algorithms, it is probably desirable to be certain of finding all secondary peaks, so the full IFFT of the upsampled frequency response is probably valuable.

Thumbnail Description of the Stitching Processing

To combine the impulse response information from two different OFDM channels, or two segments of a single OFDM channel separated by an interior exclusion band, it has been suggested to leverage the approach used to combine two antennas separated by a distance, to create an antenna pattern with resolution corresponding to a larger, synthetic aperture size.

The basic idea of the steps in synthetic aperture combining of frequency responses of different frequency segments is laid out in this section, and Section 2.4 contains examples that further explain and illustrate the details required.

Section 2.3.2 contains necessary assumptions and conditions required of the network response to beneficially use synthetic aperture combining.

Section 2.4.4 provides guidance on the number of samples to use in upsampling each frequency segment's frequency response with augmented zero values.

Basic Steps in the Stitching Process, i.e., Synthetic Aperture Combining

Multiply the time domain impulse responses by phase versus time tilt, with tilt according to frequency separation of the two frequency segments.

Weight each frequency segment's time domain impulse response values by the number of samples in their frequency response (prior to upsampling). Then perform the weighted average of each time domain sample. This provides the time domain impulse response of the two synthetically combined different frequency segments' time domain impulse responses.

> The approach generalizes to combining multiple separated frequency segments, using the weighting of the number of non-zero samples in each segment.

When stitching more than two contiguous segments, the weighting of the various contiguous segments may be adjusted, rather than have them all be averaged together with unity value at time t = 0 (effectively) and weighted by the number of frequency samples in each frequency segment.

Signal-to-noise ratio (SNR) is one additional possible consideration in weighting.

For example, if one contiguous segment has very low SNR, it may be advised to not include it or perhaps include it with a further reduced weighting factor based on its lower SNR.

Exclusion bands within a single channel are to be used to create separate contiguous spans of frequencies with channel estimates, rather than trying to bridge the excluded samples.

Impaired frequencies, such as those containing suckouts, will be excluded and create separate contiguous segments.

Band edges may be excluded in addressing tilt removal, as mentioned in the preliminary step described in Section 2.2.

If the conditions cited in the next section (Section 2.3.2) regarding negligible uncompensated detrimental frequency dependency of separate frequency segments are satisfied, then the steps for synthetic combining of multiple frequency segments' time domain impulse responses are provided below:

Identify the frequency separation of a pair of segments.

There are two ways of defining frequency separation of segments, and these are explained in the following section, with details.

The application of one or the other definitions of the frequency separation of a pair of segments has roots in the details of subroutines used for computing the inverse Fourier transform. This is explained in Section 2.3.4.

Obtain the time domain impulse response of the two frequency segments (using the method/processing of the previous sections), with both time domain impulse response segments having the same time increment between each time-sample integer index.

This property is satisfied if after upsampling of the frequency domain for the two segments, each segment's frequency domain has the same number of samples.

Necessary Assumptions and Conditions Required of the Network Response to Beneficially use Synthetic Aperture Combining

There are some necessary assumptions or characteristics required of the network response to beneficially combine the time domain responses of multiple frequency segments. These assumptions are listed below:

The time domain impulse response of the network is the same for both (or all) frequency segments which will have their responses combined EXCEPT for:

Magnitude or amplitude tilt versus frequency can be different.

An absolute amount of attenuation (constant across each frequency segment) can be different.

Group delay variation floor (i.e., phase versus frequency tilt) can be different.

An absolute amount of phase shift (constant across each frequency segment) can be different.

The frequency spans of each frequency segment can be different.

If an impedance mismatch which contributes to a resonant cavity in the network (i.e., a secondary peak in the time domain impulse response) has a different reflection coefficient phase value for the frequencies of two frequency segments, then the combining of the secondary peaks that involve that reflection (from that impedance mismatch) will not add coherently in the synthetic combining.

The "vector tips" of the two time domain impulses will be different by the amount of the phase difference between the two frequencies (this difference in phase is AFTER the phase tilt in the time domain is applied, due to the frequency separation of two segments described later in this paper).

Comparison of the complex phase value of the time domain secondary impulses which have the same time delay from the primary impulse, for a pair of frequency segments, will show if there is a phase difference in the reflection coefficients due to frequency dependency. (Again, this is comparing the phase of the same secondary time domain impulse, in two different frequency segments, AFTER applying the phase tilt in the time domain due to the frequency separation of the two segments.)

Evaluation of phase differences of some or all secondary impulses, for different frequency segments, can help show which impedance mismatches might be frequency dependent, and by how much. Evaluation of multiple cable modems' responses should align when a mismatch point is shared with other cable modems.

If there is some certainty of a frequency dependent component in one or more secondary peaks for two or more different frequency segments, then the phases of the secondary peaks could be adjusted for the time domain samples to bring them into alignment for the various frequency segments, prior to applying the synthetic aperture combining.

Impairments which cause severe non-flatness (other than tilt) in amplitude and GDV are problematic in ANY frequency segment. These can hopefully be identified and removed prior to the processing of individual frequency segments. Such impairments which have severe non-flat and different frequency responses in two different frequency segments are detrimental to the synthetic aperture approach.

Guidance on the Number of Samples to Use in Upsampling Each Frequency Segment's Frequency Response with Augmented Zero Values

This section provides guidance on the number of samples to use (after upsampling) for each frequency segment. Each frequency segment should be upsampled to the same value. If there are more than two frequency segments, and all are to be "synthetically aperture" combined, then the number of samples is guided by the lowest frequency segment and the highest frequency segment.

Ideally, the number of samples of each frequency segment after upsampling, *N*, is the same value, and is lower bounded as calculated below, where we are given:

F_{low} equals the center frequency of the lowest active subcarrier of the lowest frequency segment, after excluding any band edge impairments,

F_{high} equals the center frequency of the highest active subcarrier of the highest frequency segment, after excluding any band edge impairments, and

delta_sample_freq equals the difference in frequency between adjacent samples of the frequency responses, i.e., subcarrier frequency spacing in the OFDM channel or segment.

THEN, *N*, the number of samples, should satisfy:

$$N \geq 1 + \left(\frac{F_{high} - F_{low}}{delta\_sample\_freq}\right)$$

Different Ways of Calculating Frequency Separation of Frequency Segments, Depending on Fourier Transform Subroutine Details

Additional notes – lowest frequency of segment versus center frequency of segment:

Depending on the Fourier transform subroutines used, it may be easier – and it is acceptable – to use the lowest frequency of a contiguous set of frequencies in a frequency domain channel estimate as the “0 Hz” frequency, instead of finding a “center frequency” of the segment of the channel estimate (and using the center frequency as the “0 Hz” frequency).

IF the lowest frequency of each contiguous channel estimate segment is treated as 0 Hz in performing the steps to

a) remove tilts;
b) place the primary time domain impulse at index time t = 0;
c) normalize the time t = 0 value to 1 + *j*0;

THEN: use the frequency difference of the lowest frequency of the contiguous channel estimates, compared to the lowest frequency of a higher frequency channel estimate segment, to calculate the phase tilt to be added to the time domain impulse responses of the higher frequency channel estimate segments, prior to averaging the multiplicity of channel estimate segments.

This approach for calculating the frequency separation of two different contiguous sets of frequencies (i.e., using the lowest frequency subcarrier of each contiguous set) is called method 1 and is illustrated with an example in Section 2.4.2.1.

In a second method for calculating the frequency separation of two different contiguous sets of frequencies, the center frequency of the two segments is used to calculate the frequency separation. This is illustrated with an example in Section 2.4.2.2.

The choice of using method 1 or method 2 for calculating the frequency separation is dependent upon details of indexing the subcarriers for the inverse Fourier transform calculation, in particular, whether the 0 Hz subcarrier is the first subcarrier (then use method 1) or the middle subcarrier (then use method 2) in the calculation.

*Each contiguous segment of channel estimates should be zero padded, i.e., upsampled, to the same value***,** and this value should be at least as large as the frequency span of the entire set of channel estimates (the lowest subcarrier center frequency of the lowest frequency segment, to the highest subcarrier center frequency of the highest frequency segment). See Section 2.3.3. Even larger zero padding is better in terms of finer time domain resolution.

A Note on Zero Padding

There are a number of methods for what engineers refer to as zero padding their complex data. However, the method becomes important when stitching together multiple channels. In the previous section we explained two methods for stitching channels together. For method 1, it is important to add zero values to the higher frequencies of the channel. For method 2, one should add an equal number of zeroes to each side of the channel.

Details on Combining Second (and more) OFDM/A Segments’ Response

Additional notes

In the following pages, an example is provided for computing the phase versus time tilt to be “added” to the channel estimates of higher frequency channels, prior to averaging all the time domain impulse responses (one time domain impulse response per frequency domain channel estimate).

The computation of the slope of the phase versus time to “add” to a higher frequency channel’s time domain impulse response is shown using each of the following:

a) using the difference between the lowest frequencies of each channel (method 1);
b) using the difference between the center frequency of each channel (method 2).

It is shown that both approaches yield the same result.

The example also illustrates how much upsampling to provide, at a minimum, for the frequency domain which is spanned by all the channels.

It is worth noting that reduction in the computations can be achieved if one channel’s time domain impulse response is seen to have one or several secondary peaks of interest, and then limiting the computation of the time domain impulse response of the remaining channels to the time domain samples at those secondary peaks and at a few neighboring time values. It is not necessary to evaluate the full time domain impulse response for all channels after the secondary peaks of interest are identified.

Details of Calculating Phase Tilt Versus Time for Second (and more) OFDM Channel(s)

Example illustrating computation of phase versus time tilt for higher frequency channel:

The following pages summarize two approaches for calculating the phase tilt versus time to apply to a second (higher frequency) OFDM channel.

One approach uses the *lowest subcarrier frequency* of the two OFDM channels, and the other approach uses the *center frequency* of the two OFDM channels. Both approaches provide the same result.

It is also shown what the minimum number of frequency domain samples should be used when “upsampling” the frequency domain for each channel (or portion of a channel when a channel has excluded subcarriers).

Some basics leading up to RULE #1:

Unless otherwise indicated, subsequent references to frequency response or impulse response are assumed to be complex-valued.

Consider at first just the lowest frequency channel, and let the channel frequency response be $H(f)$ for all frequencies (e.g., 0 Hz to over 1 GHz for this example).

Let

$$h_1(t) = \frac{1}{bw_1} \int_{f=600\ MHz}^{f=790\ MHz} \left[H(f) e^{j2\pi f t}\right] df$$

where $h_1(t)$ is the Fourier transform of the channel frequency response $H(f)$, and

$$bw_1 = \int_{f=600\ MHz}^{f=790\ MHz} H(f)df$$

and $j = \sqrt{-1}$.

This “normalizes” $h_1(0) = 1$.

Now consider a frequency-shifted function, *Ha(f) = H(f + 600)*, for all *f*.

This function *Ha(f)* is what we will have if we take the channel response from 600 MHz to 790 MHz and treat it in our inverse Fourier transform routine as if it was occupying frequencies from 0 Hz to 190 MHz.

Then let

$$h_1a(t) = \frac{1}{bw_{1a}} \int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f)e^{j2\pi ft}\right]df$$

where the normalization factor

$$bw_{1a} = \int_{f=0\ MHz}^{f=190\ MHz} Ha(f)df$$

This is the response we will get from an inverse Fourier transform of *Ha(f)*, with normalization.

Using a substitution, *Ha(f-600) = H(f)* where the units on 600 are MHz and the time units are microseconds to match, and a change of variables (*f’ = f-600*) with the integral for $h_1(t)$, we see that

$$h_1(t) = \frac{1}{bw_1} \int_{f'=0\ MHz}^{f'=190\ MHz} \left[Ha(f')e^{j2\pi(f'+600)t}\right]df' \ = \frac{e^{2\pi j600t}}{bw_1} \int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f)e^{j2\pi ft}\right]df$$

We can also note that if we use the definition (equation) for finding bw_1, the “normalization” to set $h_1(t)$ to be 1 + *j*0 at time t = 0, to also find the normalization “bw_{1a},” for $h_1a(t)$, we will obtain that $bw_1 = bw_{1a}$ (the same integral is obtained when evaluated at time t = 0).

From this we see that

$$h_1(t) = e^{j2\pi600t}h_1a(t)$$

> *Note: In the analyses that we have used in MIND and other work related to PNM, we have become familiar with working with the time domain impulse response such as $h_1a(t)$, without “seeing” or using the complex-valued phase ramp that multiplies $h_1a(t)$ to yield “the actual” time domain impulse response of the channel (which occupies 600 MHz to 790 MHz). The*

presence of the complex exponential which multiplies $h_1 a(t)$, is there, and useful to "see" $h_1(t)$. But for some use cases such as in this document it isn't particularly needed or useful so is often neglected.

Thus, the important take-away from this is – which we call RULE #1 – *it is ok to neglect a leading multiplication of a time domain impulse response by a complex exponential, which effectively adds a phase ramp versus time to the time domain impulse response.*

Now, let's examine the approaches for stitching together two separate segments of OFDM frequency responses.

Numerical Example Illustrating Two Methods of Calculating the Separation Frequency" of Two OFDM Channels

Consider one OFDM channel (with 50 kHz subcarrier spacing) with 3800 subcarriers, 190 MHz of modulated spectrum (bandwidth), with the first subcarrier centered at 600 MHz; and a second OFDM channel (with 50 kHz subcarrier spacing) with 2000 subcarriers, 100 MHz bandwidth, with the first subcarrier centered at 840 MHz. Thus, the first channel covers 600 MHz to 790 MHz (actually, up to 789.950 MHz highest subcarrier center frequency), and the second channel covers from 840 MHz to 940 MHz (actually, up to 939.950 MHz highest subcarrier center frequency).

1. First Approach (Using SeparationFrequency$_1$):

Use the *lowest frequency of each channel* to compute the phase tilt versus time to apply at the higher frequency channel. This is also termed "method 1" in this paper.

The difference between the lowest frequency subcarriers of the two channels is:

800 MHz – 600 MHz = 200 MHz.

This difference will be called the $SeparationFrequency_1$ (where the 1 indicates this is the first *SeparationFrequency* used for the first approach to differentiate from a different *SeparationFrequency* we will use in the second approach), *and in the case of an inverse Fourier transform subroutine that considers the lowest frequency of a channel as if it is 0 Hz*, in computing the time domain impulse response, this will be the $SeparationFrequency_1$ to use: the difference between the two channels' lowest frequencies.

There are multiple ways of treating the frequency response of an OFDM channel in performing the inverse Fourier transform to find the time domain impulse response of the channel.

If the channel frequency response, *H(f)*, is complex, and covers 600 MHz to 790 MHz (789.950 MHz), and is manipulated by the inverse Fourier transform to find the channel impulse response in the time domain, then:

$$h_1(t) = \frac{1}{bw_1} \int_{f=600\,MHz}^{f=790\,MHz} \left[H(f)e^{j2\pi ft}\right] df$$

where

$$bw_1 = \int_{f=600\ MHz}^{f=790\ MHz} H(f)df$$

This normalizes $h_1(0) = 1$.

Then let *Ha(f) = H(f + 600)*, for *f* from 0 to 190 MHz. Also let

$$h_1a(t) = \frac{1}{bw_{1a}} \int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f)e^{j2\pi ft}\right]df$$

where

$$bw_{1a} = \int_{f=0\ MHz}^{f=190\ MHz} Ha(f)df$$

This normalizes $h_1a(0) = 1$.

We also show, as we did in explaining RULE #1, that

$$h_1(t) = \frac{1}{bw_1} \int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f)e^{j2\pi(f+600)t}\right]df$$

$$h_1(t) = \frac{1}{bw_1} \int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f)e^{j2\pi ft}e^{2\pi j600t}\right]df$$

$$h_1(t) = \frac{e^{2\pi j600t}}{bw_1} \int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f)e^{j2\pi ft}\right]df$$

$$h_1(t) = e^{j2\pi 600t} h_1a(t)$$

And then, we apply RULE #1, and note that we can effectively ignore the leading complex exponential, for the purposes of analysis.

Now we are about halfway done with the example using the low frequency of each frequency segment to calculate the difference frequency of the segments to determine the complex phase ramp in the time domain for stitching two OFDM frequency segments (method 1).

Now we will look at the OFDM channel in our example that occupies 800 MHz to 900 MHz (899.950 MHz).

As with the lowest frequency channel, let *H(f)*, evaluated from 800 MHz to 900 MHz have the time domain function from its inverse Fourier transform, and bw_2 is defined similarly as bw_1:

$$h_2(t) = \frac{1}{bw_2} \int_{f=800\ MHz}^{f=900\ MHz} \left[H(f)e^{j2\pi ft}\right] df$$

where

$$bw_2 = \int_{f=800\ MHz}^{f=900\ MHz} H(f) df$$

Then let *Hb(f) = H(f + 800)*, for *f* from *0 Hz* to *100 MHz*. Then,

$$h_2(t) = \frac{1}{bw_2} \int_{f=0\ MHz}^{f=100\ MHz} \left[Hb(f)e^{j2\pi(f+800)t}\right] df$$

$$h_2(t) = \frac{1}{bw_2} \int_{f=0\ MHz}^{f=100\ MHz} \left[Hb(f)e^{j2\pi ft}e^{2\pi j800t}\right] df$$

$$h_2(t) = \frac{e^{2\pi j800t}}{bw_2} \int_{f=0\ MHz}^{f=100\ MHz} \left[Hb(f)e^{j2\pi ft}\right] df$$

$$h_2(t) = e^{j2\pi 800t} h_2 b(t)$$

Note that now the leading complex exponential term is different than with the lower frequency channel.

Now let's consider the composite of the two OFDM channels together, excluding the frequencies between the modulated spectrum (where we have no channel frequency response information).

$$h(t) = \frac{1}{bw_{total}} \left(\int_{f=600\ MHz}^{f=790\ MHz} \left[H(f)e^{j2\pi ft}\right] df + \int_{f=800\ MHz}^{f=900\ MHz} \left[H(f)e^{j2\pi ft}\right] df \right)$$

where bw_{total} is the integral of *H(f)df* over the range 600 MHz to 900 MHz, excluding 790 MHz to 800 MHz, or

$$bw_{total} = \int_{f=600\ MHz}^{f=790\ MHz} H(f) df + \int_{f=800\ MHz}^{f=900\ MHz} H(f) df$$

Then, let *Ha(f) = H(f + 600),* for f from 0 Hz to 190 MHz, and let *Hb(f) = H(f + 800*), as we did earlier.

As just above, we again obtain

$$h_1(t) = e^{j2\pi 600t} h_1 a(t)$$

and

$$h_2(t) = e^{j2\pi 800t} h_2 b(t)$$

Performing the change of variables and substitutions as done on the individual integrals,

$$h(t) = \frac{1}{bw_{total}} \left(\int_{f=0\ MHz}^{f=190\ MHz} \left[Ha(f) e^{(j2\pi ft} e^{j2\pi 600t} \right] df + \int_{f=0\ MHz}^{f=100\ MHz} \left[Hb(f) e^{j2\pi ft} e^{j2\pi 800t} \right] df \right)$$

This reduces to

$$h(t) = h_1(t) \left[\frac{bw_1}{bw_{total}} \right] + h_2(t) \left[\frac{bw_2}{bw_{total}} \right]$$

$$\begin{aligned} h(t) &= h_1 a(t) e^{j2\pi 600t} \left[\frac{bw_1}{bw_{total}} \right] + h_2 b(t) e^{j2\pi 800t} \left[\frac{bw_2}{bw_{total}} \right] \\ &= \ e^{j2\pi 600t} \left(\left[h_1 a(t) \frac{bw_1}{bw_{total}} \right] + h_2 b(t) e^{j2\pi 200t} \left[\frac{bw_2}{bw_{total}} \right] \right) \end{aligned}$$

Applying RULE #1 to ignore the leading complex exponential term, the result for $h(t)$ is therefore

$$h(t) = \ h_1 a(t) \left[\frac{bw_1}{bw_{total}} \right] + h_2 b(t) e^{j2\pi 200t} \left[\frac{bw_2}{bw_{total}} \right]$$

Note that $h_1 a(t)$ is the time domain impulse response obtained from the inverse Fourier transform of the OFDM channel occupying 600 MHz to 790 MHz, but shifted to occupy 0 Hz to 190 MHz, and $h_2 b(t)$ is the time domain impulse response obtained from the inverse Fourier transform of the OFDM channel occupying 800 MHz to 900 MHz, but shifted to occupy 0 Hz to 100 MHz. *This is the case where the inverse Fourier transform subroutine used treats the lowest frequency of a frequency segment as if it corresponds to 0 Hz.*

The frequency separation to be used for the synthetic aperture approach combining with this type of inverse Fourier transform subroutine is to calculate the difference between the two lowest subcarrier frequencies of the two frequency segments, and use this to form the complex exponential term, which introduces the phase versus time ramp upon the second (highest frequency) OFDM channel's time domain impulse response before performing a weighted average of them. (method 1)

Also note that the weighting terms, bw_1/bw_{total} and bw_2/bw_{total}, correspond to the weighting based on the number of subcarriers in the two frequency responses, prior to upsampling or zero padding. Note that bw_{total}equals the sum of bw_1 and bw_2.

Thus, the synthetic aperture combining equation for this example is generalized to the following.

LET: $SeparationFrequency_1$ equal the lowest frequency of the highest frequency channel of the pair, minus the lowest frequency of the lowest frequency channel of the pair, in units of Hz.

The number of subcarriers (after excluding band edges, prior to upsampling) of the lowest frequency channel and highest frequency channel are respectively N_1 and N_2.

The time domain impulse response of the lowest frequency channel, after frequency down-shifting and processing and normalization, is $h_1 a(t)$, and for the highest frequency channel is $h_2 b(t)$.

THEN: the synthetic aperture combined time domain impulse response, $h(t)$, is given by

$$h(t) = \ h_1 a(t)\left[\frac{N_1}{N_1 + N_2}\right] + h_2 b(t) e^{j2\pi SeparationFrequency_1 t}\left[\frac{N_2}{N_1 + N_2}\right]$$

where the units of $SeparationFrequency_1$ and time, *t*, are matched, such as MHz and microseconds.

2. Second Approach (Using SeparationFrequency₂):

Use the **center frequency of each channel** to compute the phase tilt versus time to apply at the higher frequency channel. This is also termed “method 2” in this paper.

The difference between the center frequency of the subcarriers of the two channels is

(899.950 + 800)/2 – (789.950 + 600)/2 = 155 MHz.

This difference will be called the $SeparationFrequency_2$, and in the case of an inverse Fourier transform subroutine that considers the center frequency of a channel (or contiguous modulated frequency segment) as if it is 0 Hz, in computing the time domain impulse response, this will be the $SeparationFrequency_2$ to use: the difference between the two channels’ (segments’) center frequencies.

There are multiple ways of treating the frequency response of an OFDM channel in performing the inverse Fourier transform to find the time domain impulse response of the channel.

The math for this second approach is similar to the math of the first approach, but the limits of integration are different due to the inverse Fourier transform using the center of the segment as 0 Hz.

The steps are shown below, but streamlined because the steps are so similar to the first example.

If the channel frequency response, $H(f)$, is complex, and covers 600 MHz to 790 MHz (789.950 MHz), and is manipulated by the inverse Fourier transform to find the channel impulse response in the time domain, then

$$h_1(t) = \frac{1}{bw_1} \int_{f=600\ MHz}^{f=790\ MHz} \left[H(f) e^{j2\pi f t}\right] df$$

where

$$bw_1 = \int_{f=600\ MHz}^{f=790\ MHz} H(f)df$$

This normalizes $h_1(0) = 1$.

Then let *Ha(f) = H(f + 695)*, for *f* from *-95 MHz to 95 MHz*. Also let

$$h_1a(t) = \frac{1}{bw_{1a}} \int_{f=-95\ MHz}^{f=95\ MHz} \left[Ha(f)e^{j2\pi ft}\right]df$$

where

$$bw_{1a} = \int_{f=-95\ MHz}^{f=95\ MHz} Ha(f)df$$

This normalizes $h_1a(0) = 1$.

Note that we could compute the actual center frequency of the subcarriers in this case, and that would be 694.975 MHz, and not 695 MHz. This is not a subcarrier center frequency, but is between two subcarrier center frequencies. It is possible that an inverse Fourier transform subroutine would use a subcarrier center frequency at exactly 0 Hz, and this may in many cases be the subcarrier with a center frequency which is closest to, but just above, the actual segment (or channel) center frequency. In this example, the subcarrier at 695.0 MHz center frequency would then be shifted to 0 Hz for the inverse Fourier transform. This explanation is why the value 695 MHz is used in the equations instead of the actual average of the subcarrier center frequencies.

We also show, as we did similar to explaining RULE #1, and the first example, that

$$h_1(t) = \frac{1}{bw_1} \int_{f=-95\ MHz}^{f=95\ MHz} \left[Ha(f)e^{j2\pi(f+695)t}\right]df$$

$$h_1(t) = \frac{1}{bw_1} \int_{f=-95\ MHz}^{f=95\ MHz} \left[Ha(f)e^{j2\pi ft}e^{2\pi j695t}\right]df$$

$$h_1(t) = \frac{e^{j2\pi 695t}}{bw_1} \int_{f=-95\ MHz}^{f=95\ MHz} \left[Ha(f)e^{j2\pi ft}\right]df$$

$$h_1(t) = e^{j2\pi 695t}h_1a(t)$$

And then, we apply RULE #1, and note that we can effectively ignore the leading complex exponential, for the purposes of analysis.

Now we are about halfway done with the example using the center frequency of each frequency segment to calculate the difference frequency of the segments to determine the complex phase ramp in the time domain for stitching two OFDM frequency segments (method 2).

Now we will look at the OFDM channel in our example that occupies 800 MHz to 900 MHz (899.950 MHz).

As with the lowest frequency channel, let *H(f)* evaluated from 800 MHz to 900 MHz have the time domain function from its inverse Fourier transform

$$h_2(t) = \frac{1}{bw_2} \int_{f=800\ MHz}^{f=900\ MHz} \left[H(f) e^{j2\pi ft}\right] df$$

Let *Hb(f) = H(f + 850)*, for *f* from *-50 MHz* to *50 MHz*. Also let

$$h_2b(t) = \frac{1}{bw_{2b}} \int_{f=-50\ MHz}^{f=50\ MHz} \left[Hb(f) e^{j2\pi ft}\right] df$$

where

$$bw_{2b} = \int_{f=-50\ MHz}^{f=50\ MHz} Hb(f) df$$

This normalizes $h_2b(0) = 1$.

Then,

$$h_2(t) = \frac{1}{bw_2} \int_{f=-50\ MHz}^{f=50\ MHz} \left[Hb(f) e^{j2\pi(f+850)t}\right] df$$

$$h_2(t) = \frac{1}{bw_2} \int_{f=-50\ MHz}^{f=50\ MHz} \left[Hb(f) e^{j2\pi ft} e^{2\pi j850t}\right] df$$

$$h_2(t) = \frac{e^{2\pi j850t}}{bw_2} \int_{f=-50\ MHz}^{f=50\ MHz} \left[Hb(f) e^{j2\pi ft}\right] df$$

$$h_2(t) = e^{j2\pi850t} h_2b(t)$$

Note that now the leading complex exponential term is different than with the lower frequency channel.

Now let’s consider the composite of the two OFDM channels together.

$$h(t) = \frac{1}{bw_{total}} \left(\int_{f=600\ MHz}^{f=790\ MHz} \left[H(f) e^{j2\pi ft} \right] df + \int_{f=800\ MHz}^{f=900\ MHz} \left[H(f) e^{j2\pi ft} \right] df \right)$$

where $bw_{total}$is the integral of *H(f)df* over the range *600 MHz* to *900 MHz*, excluding *790 MHz* to *800 MHz*, or

$$bw_{total} = \int_{f=600\ MHz}^{f=790\ MHz} H(f) df + \int_{f=800\ MHz}^{f=900\ MHz} H(f) df$$

Then let *Ha(f) = H(f + 695 MHz)*, for *f* from *-95 MHz* to *95 MHz*, and let *Hb(f) = H(f + 850)*, for *f* from *-50 MHz* to *50 MHz,* as we did earlier.

As just above, we again obtain

$$h_1(t) = e^{j2\pi 695t} h_1 a(t)$$

and

$$h_2(t) = e^{j2\pi 850t} h_2 b(t)$$

Performing the change of variables and substitutions as done on the individual integrals:

$$h(t) = \frac{1}{bw_{total}} \left(\int_{f=-95\ MHz}^{f=95\ MHz} \left[Ha(f) e^{j2\pi 695t} \right] df + \int_{f=-50\ MHz}^{f=50\ MHz} \left[Hb(f) e^{j2\pi 850t} \right] df \right)$$

This reduces to

$$h(t) = h_1(t) \left[\frac{bw_1}{bw_{total}} \right] + h_2(t) \left[\frac{bw_2}{bw_{total}} \right]$$

$$\begin{aligned} h(t) &= h_1 a(t) e^{j2\pi 695t} \left[\frac{bw_1}{bw_{total}} \right] + h_2 b(t) e^{j2\pi 850t} \left[\frac{bw_2}{bw_{total}} \right] \\ &= e^{j2\pi 695t} \left(\left[h_1 a(t) \frac{bw_1}{bw_{total}} \right] + h_2 b(t) e^{j2\pi 155t} \left[\frac{bw_2}{bw_{total}} \right] \right) \end{aligned}$$

Applying RULE #1 to ignore the leading complex exponential term, the result for $h(t)$ is therefore

$$h(t) = h_1 a(t) \left[\frac{bw_1}{bw_{total}} \right] + h_2 b(t) e^{j2\pi 155t} \left[\frac{bw_2}{bw_{total}} \right]$$

Note that $h_1a(t)$ is the time domain impulse response obtained from the inverse Fourier transform of the OFDM channel occupying *600 MHz* to *790 MHz*, but shifted to occupy -95 MHz to 95 MHz, and $h_2b(t)$ is the time domain impulse response obtained from the inverse Fourier transform of the OFDM channel occupying *800 MHz* to *900 MHz*, but shifted to occupy -50 MHz to 50 MHz. This is the case where the inverse Fourier transform subroutine used treats the center frequency of a frequency segment as if it corresponds to 0 Hz.

The frequency separation to be used for the synthetic aperture combining with this type of inverse Fourier transform subroutine is to calculate the difference between the two center frequencies of the two frequency segments, and use this to form the complex exponential term, which introduces the phase versus time ramp upon the second (highest frequency) OFDM channel's time domain impulse response before performing a weighted averaging of them. (method 2)

Also note that the weighting terms, bw_1/bw_{total}, and bw_2/bw_{total}, correspond to the weighting based on number of subcarriers in the two frequency responses, prior to upsampling or zero padding. Note that bw_{total} equals the sum of bw_1 and bw_2.

Thus, the synthetic aperture combining equation for this example is generalized to the following.

LET: $SeparationFrequency_2$ equals the center frequency of highest frequency channel of the pair, minus the center frequency of lowest frequency channel of the pair, in units of Hz.

Number of subcarriers (after excluding band edges, prior to upsampling) of the lowest frequency channel and highest frequency channel are respectively N_1 and N_2.

The time domain impulse response of the lowest frequency channel, after processing and normalization, is $h_1a(t)$, and for the highest frequency channel is $h_2b(t)$.

THEN: the synthetic aperture combined time domain impulse response, $h(t)$, is given by

$$h(t) = h_1a(t)\left[\frac{N_1}{N_1+N_2}\right] + h_2b(t)e^{j2\pi SeparationFrequency_2 t}\left[\frac{N_2}{N_1+N_2}\right]$$

where the units of $SeparationFrequency_2$ and time, *t*, are matched, such as MHz and microseconds.

Zero Padding

Recall that in Section 2.3.5 we explained that the details of the zero padding depend on the indexing of the subcarrier frequencies which is employed in the IFFT calculation, i.e., the 0 Hz subcarrier may be the first indexed subcarrier (method 1) or the 0 Hz subcarrier may be the center (middle) subcarrier (method 2). Refer to Section 2.3.5. For method 1, it is important to add zero values to the higher frequencies of the channel. For method 2, one should add an equal number of zeroes to each side of the channel.

How Much Upsampling for M Channels or Separate Frequency Segments?

IF the lowest frequency of the lowest frequency segment is F_{low}, and the highest frequency of the highest frequency segment is F_{high}, then each of the M frequency segments needs to be upsampled to the same value, and it should be at least as large as given below in the equation for *N*.

As in Section 2.3.3, we also define *delta_sample_freq* as the difference in frequency between adjacent samples of the frequency responses, i.e., subcarrier frequency spacing in our OFDM.

THEN, *N*, the number of samples in each of the M segments, after upsampling, should be equal to or greater than

$$N \geq 1 + \left(\frac{F_{high} - F_{low}}{delta\_sample\_freq}\right)$$

In the example of Section 2.4.2 the subcarrier spacing is *50 kHz* and F_{high} is *899.950 MHz* and F_{low} is *600 MHz*.

The equation is evaluated with F_{high} minus F_{low} equal to *299.950 MHz* and divided by *delta_sample_freq* which is *50 kHz*, which becomes 5,999. Adding 1 to this yields the lower bound for *N* at *6000* frequency domain samples for each of the two segments.

Synthetic Aperture Combining for M Channels or Separate Frequency Segments:

Let the lowest frequency segment be denoted with m = 1, and the next lowest frequency segment with m = 2, and so on, up to the highest frequency segment, m = M.

Then calculate the $SeparationFrequency_{n,k}$ for the separation between segment m = 1 and m = k using approach *n=1* or *2* specified above.

The previous Section (2.4.2) showed the two different ways of calculating the $SeparationFrequency_n$ depending on how the inverse Fourier transform subroutine maps the frequency segment channel response data into 0 Hz. One approach is that a center frequency of a segment maps into 0 Hz for the inverse Fourier transform computation (method 2), and another approach is that the lowest frequency of a segment maps to the 0 Hz value for the subroutine (method 1).

To reduce the notation length, in the following equations we will use $SF_{n,k}$ to denote the $SeparationFrequency_{n,k}$, the separation frequency between the kth channel and the first (lowest frequency) channel. Also, *n* corresponds to the separation frequency method used, either *n* = 1 or *n* = 2, as described in Section 2.4.2.1 and Section 2.4.2.2.

Each of the M frequency segments has N_m subcarriers, for m = 1 to M, after eliminating some band edge subcarriers as previously discussed in the basic steps.

The total number of subcarriers in all of the stitched channels (or segments) is N_{total}:

$$N_{total} = \sum_{m=1}^{M} N_m$$

Given the normalized time domain impulse responses of the M segments, denoted has $h_M(t)$, for m = 1 to M, we compute the synthetic aperture combined impulse response by the following equation.

The synthetic aperture combined time domain impulse response, $h(t)$, is given by

$$h(t) = h_1(t)\left[\frac{N_1}{N_{total}}\right] + h_2(t)e^{j2\pi SF_{n,2}t}\left[\frac{N_2}{N_{total}}\right] \ldots + h_M(t)e^{j2\pi SF_{n,M}t}\left[\frac{N_M}{N_{total}}\right]$$

Conclusions

An approach was described for combining frequency responses from multiple OFDM or OFDMA channels (or their segments) to synthesize the time domain impulse response of a wider channel, with enhanced time domain resolution. That approach is denoted as stitching.

The processing steps for stitching two OFDM segments' frequency responses were listed and explained, followed by presentation of equations defining the processing. Processing steps are the same for stitching two OFDMA segments' frequency responses.

As a prelude to the presentation of the stitching approach, the basic steps were reviewed and explained for processing a single segment's frequency response to obtain a time domain impulse response that is cleaned, to provide a very low amount of tilt clutter. In this context, "cleaning" includes magnitude-versus-frequency tilt removal and phase-versus-frequency tilt removal.

It was shown that the calculation of a separation frequency in the stitching process has to match the indexing of an inverse fast Fourier transform routine that is used. An amount of upsampling was described and a windowing suggestion was provided. Network conditions required for beneficial stitching were discussed, especially for reflection coefficients of important resonant cavities to have a small amount of frequency-dependence between the different channels' occupied frequencies.

Abbreviations and Definitions

Abbreviations

dB	decibel
DOCSIS	Data-Over-Cable Service Interface Specifications
FFT	fast Fourier transform
GDV	group delay variation
GHz	gigahertz
Hz	hertz
i.e.	that is (*id est*)
IFFT	inverse fast Fourier transform
kHz	kilohertz
MHz	megahertz
MIND	Methodology for Intelligent Network Discovery
NOS WG7	Network Operations Subcommittee Working Group 7
OFDM	orthogonal frequency division multiplexing
OFDMA	orthogonal frequency division multiple access
PNM	proactive network maintenance
SAR	synthetic aperture radar
SC-QAM	single carrier quadrature amplitude modulation
SCTE	Society of Cable Telecommunications Engineers
SNR	signal-to-noise ratio

Bibliography and References

[Broadband] “Reflecting on Impedance Mismatches,” R. Hranac, T. Kolze, *Broadband Library*, Fall 2025

[Williams] “OFDMA Predistortion Coefficient and OFDM Channel Estimation Decoding and Analysis,” T. Williams et al, 2021 Fall Technical Forum, SCTE Cable-Tec Expo 2021

[Wolcott] “The Proactive Network Maintenance Comeback OFDM and OFDMA Bring New, Laser-Guided Precision for Plant Fault Distancing,” L. Wolcott et al, SCTE TechExpo24

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Operationally Aligned Artificial Intelligence (AI):

Redefining Intelligence for the Physical Realities of Broadband Networks

A Technical Paper prepared for SCTE by

Allen Maharaj – **Principal Access Network Designer,** Rogers Communications
8200 Dixie Rd
Brampton, ON, L6T 0C1
Allen.maharaj@rci.rogers.com
416-450-4820

Table of Contents

Figures

Tables

Introduction – The Disconnect Between AI Hype and Network Reality

Artificial intelligence has become a dominant theme in broadband operations. Vendors promote predictive analytics, autonomous optimization, and AI-driven decision-making as pathways to simpler operations and improved customer experience with minimal human oversight. Yet this narrative often conflicts with the technical realities of broadband access network architectures. The gap is not philosophical; it is physical. Access networks behave in ways most AI systems were never designed to interpret, and existing maturity models fail to account for that. This paper therefore argues that operational intelligence in broadband must evolve from pattern recognition to causal reasoning, from predicting behavior to understanding why it occurs.

The industry frequently equates AI adoption with AI correctness. An organization may automate workflows and deploy AI features while still operating models that misread network behavior. Until correctness is evaluated independently from adoption readiness, systems will continue to appear mature on paper while remaining unsafe in practice. As AI moves closer to the control plane, even minor misinterpretations of telemetry (measurement data) can cause configuration oscillation, unnecessary dispatches, or degraded customer experience.

Hybrid fiber coax (HFC) broadband access networks are governed by RF physics. Impairment behavior, signal variability, upstream noise, and environmental factors all interact to influence service quality in complex, interdependent ways. The introduction of Low-Latency Data Over Cable Service Interface Specification (DOCSIS LLD) illustrates that responsiveness and user-experience gains come from architectural refinement rather than raw capacity expansion. CableLabs research shows that modulation stability, forward error correction (FEC) performance, and modulation error ratio (MER) can fluctuate rapidly as the physical cable network conditions change, while upstream impairments such as: ingress (unwanted signals leaking into the network), impulse noise, and common path distortion (CPD) rarely resemble the clean datasets used in typical AI training pipelines. Models without domain awareness misinterpret these variations, producing unstable or misleading results.

Current maturity frameworks such as the TeleManagement (TM) Forum AI Maturity Model and Autonomous Networks Levels focus on adoption and workflow integration but do not assess whether AI systems can interpret DOCSIS or HFC telemetry accurately. A network can therefore achieve a high maturity score while deploying AI that misclassifies impairment or destabilizes upstream performance. Accuracy, not adoption, must define maturity.

Because AI now informs fault isolation, maintenance triage, and configuration changes, its reliability is both a technical and ethical requirement. The National Institute Standards and Technology (NIST) AI Risk Management Framework specifies that systems influencing infrastructure must demonstrate reliability, transparency, and harm prevention before gaining operational authority. In broadband, harm includes misdiagnosed impairment, unnecessary field dispatches, and degraded customer experience.

This paper defines operational AI as any system that interprets multi-layer telemetry and produces recommendations or actions affecting plant behavior, service quality, or customer experience. Unlike analytical AI, which provides insights without altering the network, operational AI participates directly in configuration and maintenance decisions, making correctness and reliability mandatory.

This distinction anchors the analysis that follows. Operational AI must align with the physical behavior of the plant and be evaluated for accuracy, not organizational readiness.

Section 2 builds on this foundation by examining why conventional machine-learning approaches struggle to interpret access-network behavior, establishing the technical basis for the requirements defined later in the paper. Together, these sections connect physical interpretation, operational safety, and ethical responsibility, framing the shift toward intelligence that reasons within network physics rather than abstracting away from it.

Background – How AI Systems Interpret Networks

Artificial intelligence used in broadband operations typically relies on statistical learning methods such as classification, anomaly detection, clustering, and short-term forecasting. These approaches assume stable baselines, predictable variation, synchronized sampling, and low measurement noise. In broadband access networks, these assumptions fail because measurement noise is inherent rather than incidental. In enterprise and cloud systems, those conditions generally hold true, allowing statistical methods to perform reliably. In access networks, however, those same assumptions fail immediately because plant behavior depends on the physical principles of RF operation rather than IT-style workload dynamics.

This mismatch between statistical-learning assumptions and the noise-dominated, dynamically varying behavior of DOCSIS and HFC networks is the root cause of most AI misinterpretations in broadband operations. Physical-layer variability, impairment dynamics, and the distributed telemetry introduced through Distributed Access Architectures (DAA) create conditions that differ fundamentally from the environments for which most AI models were designed.

CableLabs documentation for DOCSIS 3.1 and 4.0 shows that modulation stability, MER, FEC behavior, and upstream signal-to-noise ratio (SNR) can vary sharply even under normal plant conditions. These fluctuations are not random; they are predictable outcomes due to the nature of RF and network design. Proactive network maintenance (PNM) research further demonstrates that impairments such as impulse noise, micro-reflections, and common path distortion generate patterns that seem random to statistical models, even though they are predictable within HFC plants. When AI systems treat DOCSIS telemetry as generic numerical data rather than as signals governed by physical constraints, they often reach technically incorrect conclusions.

Telemetry alignment adds another layer of complexity. Even when data is accurate, differences in timing and vantage points distort the relationships models depend on. In DAA architectures, telemetry from Remote PHY Devices (RPDs) and the Cable Modem Termination System (CMTS) frequently diverges because each component operates with different views, timing references, and processing states. These inconsistencies are structural, not incidental. The RPD timestamps PHY-layer measurements according to its local clock, the CMTS manages media access control (MAC)-layer scheduling based on its own timing domain, and cable modems report metrics aligned with customer premise equipment (CPE) timing. Because these are inherently unsynchronized, AI systems that assume temporal alignment encounter data conflicts that appear as noise, delay, or contradiction within the model's input space.

For example, cable modems may report rising upstream error rates that the RPD does not observe, or customer-experience telemetry may indicate degradation even while PHY-layer metrics appear stable due to minislot (a small unit of upstream scheduling time) contention or partial service states. The Broadband Forum's TR-452 performance-measurement framework emphasizes that accurate assessment of service quality requires multi-layer visibility spanning the physical, MAC, and customer-experience planes. When telemetry from any one plane is missing or incomplete, AI systems lose the contextual grounding needed for consistent interpretation.

Understanding these challenges requires examining two key dimensions: the physical behaviors that violate statistical assumptions and the architectural factors that cause AI techniques from enterprise or cloud environments to fail without domain awareness. The following subsections examine these two dimensions in detail, first outlining how DOCSIS and HFC behaviors violate foundational statistical assumptions, and then explaining why AI approaches built for enterprise or cloud environments misread broadband telemetry when domain context is absent.

DOCSIS and HFC Behaviors That Challenge Conventional AI Models

Broadband access networks exhibit operational behaviors that are routine in the field yet incompatible with the statistical assumptions underpinning most machine-learning models. These behaviors fall into three primary categories: transient impairment events that distort error patterns, frequency-dependent effects that violate smoothness assumptions, and architectural factors that fragment or misalign telemetry streams. Each category disrupts a different modeling assumption, collectively explaining why conventional AI systems struggle to interpret access-network behavior with technical accuracy.

1. **Impulse noise bursts**
   Short, high-energy bursts of impulse noise create abrupt spikes in upstream FEC errors. Models that treat these events as anomalies rather than expected plant behavior generate false alarms and distorted trend forecasts.

2. **Common Path Distortion (CPD)**
   Common Path Distortion (CPD) introduces frequency-dependent distortion in the return path (the upstream direction back to the network) and creates complex cross-channel correlations. Many upstream impairments appear concurrently across multiple channels or orthogonal frequency-division multiple access (OFDMA) subcarriers, but not in uniform patterns. Models that evaluate each channel in isolation fail to capture these relationships, misclassifying correlated impairment as unrelated load variation. The structured interference caused by CPD violates assumptions of smooth, stationary frequency behavior and is often misread as random jitter or instability.

3. **Micro-reflections**
   Micro-reflections generate narrowband, periodic ripple across OFDM carriers, producing structured fluctuations in MER. These patterns violate the assumptions of stationarity and gradual variation embedded in most time-series models. As a result, models that treat MER as a stable or slowly varying signal often misinterpret this ripple as congestion or impairment.

Dynamic Range Window (DRW) effects
Cable modems continuously adjust upstream transmit power to remain within the Dynamic Range Window (DRW). These adjustments occur independently of impairment or traffic load, yet many models interpret power variation as evidence of instability. As a result, DRW-induced changes often lead to incorrect root-cause classification.

4. **Partial service states**
   When a modem remains online but loses one or more channels, throughput and modulation metrics shift even though the device has not failed. Models that rely solely on throughput features often interpret these conditions as normal variation instead of recognizing them as a constrained operational state.

5. **Upstream noise funneling**
   Multiple noise sources can converge at the node, producing apparent correlation between otherwise independent signals. To an unconstrained model, this aggregation can resemble user-driven congestion or gradual load growth, even though its origin is purely physical noise.

6. **Environmental drift**
   Temperature-induced attenuation and moisture-related ingress violate the assumption of a stable baseline. These variations lead to false impairment detection and inconsistent trend interpretation.

7. **Seasonal plant variation**
   Signal characteristics shift markedly across seasonal extremes. Models trained on warm-weather conditions often fail in colder periods because the plant's statistical profile itself changes.

8. **DAA visibility limitations**
   In DAA architectures, the RPD and CMTS observe different parts of the network. When telemetry from one vantage point is incomplete, model inference becomes unstable or contradictory.

9. **Asymmetric telemetry timing**
   Telemetry reports from modems, RPDs, and CMTS components occur at different intervals. Time-series models depend on synchronized sampling; misalignment generates spurious correlations or hides causal events.

10. **Service-group contention**
    High-upload users can influence upstream behavior in ways that resemble impairment. Both conditions reduce upstream modulation efficiency or throughput, but their causes differ. Models often misinterpret legitimate load variation as physical instability.

These behaviors are inherent to broadband access networks and should be recognized as normal operational patterns, not anomalies. Because each violates a different statistical assumption, the resulting AI failures are predictable and traceable to specific physical-layer dynamics. Understanding these patterns is essential before applying AI to any operational workflow.

Why Enterprise and Cloud-Origin AI Techniques Misinterpret Access-Network Telemetry

Enterprise and cloud-origin AI systems are built for environments characterized by predictability, low noise, and complete system visibility. These models assume stable baselines, consistent sampling intervals, synchronized telemetry, and minimal measurement noise. In broadband access networks, these assumptions collapse immediately. Physical-layer impairments, environmental variation, and timing divergence generate data that is noisy, incomplete, and sometimes contradictory. As a result, enterprise-style models often interpret meaningful RF behavior as statistical anomaly and dismiss legitimate impairment signatures as random variation.

Cloud-origin anomaly detection and forecasting techniques rely on the statistical stability typical of IT or cloud workloads. When applied to DOCSIS telemetry, they misinterpret normal RF fluctuations as instability and treat impairment-related signatures as outliers rather than as predictable physical effects. The same issue occurs with enterprise-style clustering and classification, which assume that feature relationships are stationary and primarily user-driven. In the access network, by contrast, those relationships are shaped by physical processes, environmental factors, and RF-path dynamics.

Further limitation arises from assumptions about data completeness. AI models developed for IT or data-center environments expect full visibility into relevant telemetry, but distributed broadband architectures rarely provide it. In DAA deployments, for example, the RPD may report a clean upstream path while cable modems observe error bursts. Customer-experience data may indicate degradation even when PHY-layer metrics appear normal. Models built without awareness of these architectural differences reach incorrect conclusions because they cannot reconcile conflicting or missing signals.

In practice, these limitations cause enterprise-origin AI to misclassify normal DOCSIS behavior as instability. A common example is when an anomaly detector trained on stable IT telemetry misreads a brief MER dip caused by micro-reflections as a service-impacting event. Without correlating the dip against PHY or RPD telemetry, the model triggers alerts that contradict field evidence. This divergence between algorithmic inference and physical reality is a predictable outcome when enterprise-trained AI is applied directly to network environments.

For these reasons, AI systems derived from enterprise or cloud contexts struggle to interpret DOCSIS and HFC behavior reliably. Their core design assumptions do not reflect the operational realities of access networks, leading to inconsistent or incorrect inference unless the model incorporates domain-specific knowledge and cross-layer telemetry.

This challenge becomes even more significant in DAA and virtualized CMTS architectures, where timing differences between RPD PHY measurements and MAC scheduling are structural rather than incidental. These inherent misalignments cannot be resolved by models that lack domain awareness or architectural understanding.

Ethical AI in Network Operations – A Technical Perspective

Ethical discussions in artificial intelligence typically center on bias, fairness, privacy, and transparency. These dimensions remain essential but represent only part of the ethical landscape in broadband operations, where responsibility extends beyond data and into the physical behavior of the network itself. In broadband access network environments, ethics is inseparable from system safety, operational reliability, and the prevention of unintended technical or customer harm. When AI systems influence maintenance, configuration, or service-quality decisions, correctness becomes not only a technical obligation but an ethical one.

The NIST AI Risk Management Framework stipulates that AI systems deployed within critical infrastructure must be demonstrably reliable, transparent, and contextually aligned with their operational domain. Institute of Electrical and Electronics Engineers (IEEE)'s *Ethically Aligned Design* framework complements this by requiring that human operators retain the ability to understand, audit, and override AI behavior. In broadband operations, these are not abstract ideals but engineering imperatives. They translate directly into the requirement for AI systems that interpret telemetry accurately, expose their limitations transparently, and prevent unintentional degradation of plant stability or customer experience.

Ethical Meaning of AI in a Network Engineering Context

Within access networks, ethical responsibility manifests as engineering responsibility, ensuring that AI-driven decisions preserve service quality and protect plant stability. Ethical AI in this context requires:

- preventing unsafe or unstable recommendations that could alter plant behavior
- avoiding automated changes that compromise modulation stability or introduce upstream impairment
- maintaining operator visibility into model reasoning, confidence, and data lineage
- ensuring machine-generated outputs remain verifiable against established engineering indicators
- mitigating overconfidence when telemetry is incomplete, divergent, or ambiguous

These principles expand the traditional view of ethics to include engineering practices that safeguard both the network and its users. Dataset bias represents another ethical dimension frequently overlooked in broadband. Models trained on incomplete, seasonally narrow, or topology-specific data often fail to generalize across the plant. The result is systematic blind spots; either suppressing genuine impairment signatures, overreacting to benign fluctuations, or embedding seasonal and geographic bias into operational logic.

In broadband environments, dataset bias evolves from a technical flaw into an operational hazard, because training distributions rarely capture the full variability of real plant behavior across time, topology, and season.

Ethical Risks Unique to Broadband Network Operations

The ethical dimension deepens when technical misinterpretation translates directly into operational risk. Physical behaviors in DOCSIS and HFC networks are often misread by poorly designed AI systems, and when those misreadings drive decisions, the result is not simply inefficiency but ethical failure. When these misinterpretations inform operational actions, the consequences extend beyond inefficiency; they become both technical liabilities and ethical breaches:

- **Misclassification of impairment:** Impulse noise or micro-reflections may be mistaken for congestion or user-driven load, resulting in incorrect maintenance actions or unnecessary truck rolls.
- **Incorrect configuration changes:** Automated profile adjustments or upstream tuning can destabilize the plant when models fail to account for partial service states or DRW constraints.
- **Overfitting to incomplete telemetry:** Limited or unrepresentative training data can suppress genuine impairment signals, amplify noise, and create operational blind spots where plant degradation goes undetected.
- **Invisible degradation of customer experience:** Optimizing internal KPIs can unintentionally increase latency, retransmissions, or application-level delay. Broadband Forum TR-452 notes that throughput metrics alone cannot detect these degradations.
- **Disproportionate service impact:** If a model deprioritizes certain traffic or usage patterns, it risks inconsistent user experiences, an ethical issue in consumer networks.
- **Automation bias:** Operators may over-trust AI recommendations, particularly under cognitive load during peak events or outages. This overreliance magnifies the operational consequences of incorrect, opaque, or context-blind model reasoning.

These risks underscore that ethical AI in broadband is inseparable from technical accuracy. Without trustworthy interpretation, automation ceases to enhance reliability and instead erodes both operator confidence and customer trust.

Transparency, Explainability, and the Risk of Opaque Control Loops

Explainability is foundational to operational AI. Systems that influence plant behavior must provide operators with sufficient context to understand not only what recommendation was made, but why and under which conditions it was generated. IEEE guidance emphasizes that meaningful oversight is only achievable when the decision pathways of an AI system are transparent, traceable, and fully auditable. In broadband environments, AI often influences critical operational domains such as:

- profile selection
- resource allocation

- impairment classification
- maintenance prioritization

When explainability is absent, these systems form hidden control loops, processes that appear stable until a specific condition triggers unforeseen or harmful actions, violating NIST's traceability and interpretability principles.

This risk is amplified in DAA and CMTS environments, where telemetry sources inherently diverge. When operators cannot identify which inputs the model relied upon or how it reconciled conflicting signals, both operational trust and systemic safety erode rapidly.

Ethical Data Requirements for Operational AI

The reliability of operational AI depends fundamentally on the completeness, consistency, and representativeness of its telemetry inputs. Incomplete or misaligned data creates systemic interpretive errors. Examples include:

- RPD telemetry reporting clean MER while modems register increasing FEC errors
- customer-experience data indicating degradation even as PHY metrics appear stable
- lack of visibility into upstream interference, ingress, or impulse noise
- data gaps during partial-service or degraded-channel states

The GSM Association's AI ethics guidance emphasizes that data integrity and representativeness are prerequisites for responsible deployment. For access networks, this means AI must integrate PHY-layer, MAC-layer, and customer-experience telemetry to ensure representational balance across data planes and prevent structural bias in operational reasoning.

Governance Requirements for AI That Influences Operations

AI systems that generate operational outputs must be governed with the same rigor and discipline applied to other control-plane systems. The European Telecommunications Standards Institute (ETSI) Zero-Touch Network and Service Management (ZSM) framework mandates structured lifecycle governance, version control, drift detection, rollback mechanisms, and defined oversight structures for automation systems; requirements that are equally critical for AI-driven operations.

Effective governance within broadband operations requires structured mechanisms that maintain control, accountability, and recoverability:

- controlled deployment through formal change-management procedures
- safety valves and guardrails to prevent unstable or harmful automation
- human-in-the-loop or human-on-the-loop oversight for higher-risk operational functions
- drift monitoring and trend analysis to ensure sustained reliability under evolving plant conditions

- version tracking, validation checkpoints, and rollback capabilities for rapid operational recovery

Ethical operation depends on ensuring that AI systems cannot act silently, autonomously, or irreversibly in ways that operators cannot trace, audit, or override.

Clarifying the Scope – What This Paper Is Not Arguing

This paper does not argue against the use of AI in broadband networks, nor does it suggest that AI is inherently unsuitable for DOCSIS, HFC, DAA, or CMTS architectures. It does not discourage automation or suggest that vendor initiatives are misguided. Instead, it emphasizes that AI used in broadband must be:

- aligned with the physical behavior of the plant
- supported by complete, multi-layer telemetry
- governed by safeguards that prevent unintended degradation or instability

The goal is not to constrain innovation, but to ensure that AI integration remains responsible, technically rigorous, and operationally safe.

The State of Current AI and Automation Maturity Models – and Their Structural Flaws

The telecommunications industry relies heavily on maturity models to evaluate progress in automation, AI adoption, and operational transformation. These frameworks provide valuable guidance on organizational readiness, data governance, workflow integration, and strategic alignment of AI within the enterprise. Yet their design focuses on structural and procedural maturity rather than on the technical correctness or operational reliability of AI systems interpreting network conditions.

In broadband access networks, where AI systems influence impairment classification, maintenance triage, configuration changes, and customer-experience management, technical accuracy is not optional; it is fundamental. A maturity score reflecting organizational posture rather than operational capability cannot measure how safely or effectively AI functions within the access network. Consequently, an organization may appear highly mature on paper while deploying AI systems that misinterpret network telemetry, amplifying operational risk instead of mitigating it.

Overview of Current Maturity and Automation Models

Several established frameworks attempt to categorize an organization’s progress toward automation and AI integration. Among the most influential are the TM Forum AI Maturity Model, the TM Forum Autonomous Networks Levels (ANL), and a range of vendor-developed readiness frameworks.

TM Forum AI Maturity Model

The TM Forum AI Maturity Model assesses an organization's readiness to adopt, operationalize, and scale AI across its business domains. It evaluates maturity across several dimensions:

- data management
- governance
- workflow integration
- organizational culture
- change management
- strategic alignment

The framework offers a useful enterprise-level overview of AI readiness but does not assess whether AI systems interpret physical-layer behavior accurately or whether model outputs align with the operational realities of our networks.

Autonomous Networks Levels (ANL)

The TM Forum Autonomous Networks Levels (ANL) framework defines the progression of automation from manual operation (Level 1) to full intent-driven autonomy (Level 5). It is structured around:

- automated workflows
- closed-loop control
- self-optimization
- self-healing capabilities
- orchestration depth
- policy-driven decision making

ANL effectively characterizes the sophistication and scale of automation but does not evaluate technical accuracy. A system may reach a high automation level while still making unsafe or incorrect decisions if its underlying AI misinterprets network behavior or lacks domain grounding.

Vendor AI and Automation Readiness Models

Many vendors develop proprietary maturity frameworks that emphasize deployment readiness and ecosystem integration across their product portfolios:

- cloud-native readiness
- orchestration and containerization capabilities

- integration with analytics platforms
- automated workflow coverage
- AI-feature availability

These vendor models highlight deployment and integration complexity but rarely assess whether the underlying AI is technically reliable within network environments. These frameworks measure the maturity of deployment and tooling ecosystems, not the operational accuracy or contextual reliability of the AI systems themselves.

Summary

Across these frameworks, the common thread is organizational and architectural readiness. None of the major models evaluate whether AI can:

- interpret multi-layer telemetry correctly
- operate reliably under real plant conditions
- avoid impairment misclassification
- support stable configuration decisions
- remain robust under drift, noise, or incomplete telemetry

This omission forms the structural foundation of the limitations examined in the following section, where the practical consequences of misaligned maturity scoring are made explicit.

Structural Limitations of Maturity Models in Network Operations

When maturity frameworks are applied to broadband access networks without accounting for physical context or plant behavior, their structural limitations become evident.

1. **Adoption does not indicate technical correctness**
   AI adoption or workflow automation does not guarantee interpretive correctness. A network may achieve a high adoption score while still deploying AI systems that misclassify impairment, generate unstable recommendations, or misread DOCSIS and HFC telemetry.

2. **Frameworks are physically agnostic**
   Current models remain physically agnostic. They overlook RF propagation effects, DOCSIS scheduling dynamics, upstream noise behavior, Dynamic Range Window constraints, and DAA telemetry divergence. These physical factors define how the plant behaves, yet none are represented within the evaluative structure of most frameworks.

3. **Automation depth is not equivalent to safety**
   Higher automation levels, such as those defined in ANL, measure orchestration sophistication and closed-loop autonomy, not interpretive accuracy. In practice, an

incorrect decision can be automated just as efficiently as a correct one, amplifying the consequences of model error rather than reducing them.

4. **Qualitative, checklist-style scoring**
   Maturity assessments rely on qualitative scoring criteria that seldom correlate with measurable operational outcomes such as improved modulation stability, reduced impairment frequency, or verified gains in customer-experience metrics.

5. **No stress-testing under real plant conditions**
   Most frameworks do not evaluate AI performance under real operational stress conditions such as:

   - impulse noise events
   - modulation or MER instability
   - seasonal or temperature-related drift
   - partial service or constrained-channel states
   - DAA visibility divergence and incomplete telemetry
   - asymmetric reporting intervals across CMTS, RPD, and modem domains

   These are precisely the conditions that reveal the difference between statistical robustness and operational safety, the point where ethical and technical responsibility converge.

6. **Telemetry assumptions do not hold in access networks**
   Maturity models implicitly assume synchronized, lossless, and complete data visibility. In broadband access network architectures, telemetry is structurally fragmented across cable modems, RPDs, and CMTS components, each operating within its own timing and processing domain. This fragmentation breaks the temporal and spatial coherence these frameworks assume, rendering their evaluation criteria incompatible with the realities of multi-domain telemetry alignment in broadband access networks.

7. **Explainability and operator validation are not required**
   Systems can achieve high maturity ratings even when operators cannot interpret, audit, or replicate the reasoning behind model outputs. This opacity introduces significant operational risk in domains where physical behavior must be explicitly understood and validated before any action is taken.

8. **Limited lifecycle governance evaluation**
   Most frameworks provide minimal evaluation of lifecycle governance elements such as model drift detection, rollback mechanisms, version control, and continuous degradation

monitoring, controls that are mandatory for sustaining reliability and safety in critical infrastructure environments.

Together, these limitations demonstrate that maturity scoring without technical validation produces a false sense of readiness, quantifying organizational ambition rather than engineering capability. Without grounding in measurable operational correctness, these frameworks risk legitimizing AI systems that are elegant in architecture yet unreliable in action.

Consequences of Misaligned Maturity Scoring

When organizationally oriented or workflow-focused maturity models are applied to access network environments, they create a fundamental disconnect between perceived capability and real operational reliability. This disconnect manifests operationally when a network is labeled “mature” despite deploying AI systems that:

- misinterpret upstream noise as congestion
- incorrectly adjust OFDM or OFDMA profiles
- generate inaccurate impairment classifications
- trigger unnecessary maintenance activities
- fail during partial service conditions
- degrade under seasonal or environmental drift
- issue conflicting recommendations from incomplete telemetry
- reduce customer experience metrics despite appearing stable

These behaviors have been documented in CableLabs PNM Working Group research, SCTE operational guidance, and vendor impairment analyses. The risks compound when such systems are embedded within automated workflows, because incorrect inferences can amplify network instability.

In broadband access networks, maturity cannot be measured by the quantity of automated workflows or the visibility of AI features. True maturity must be defined by the operational accuracy, robustness, and safety of the intelligence interacting with the plant. Without explicit technical validation, maturity ratings measure enthusiasm and integration velocity, not engineering integrity or system trustworthiness.

Requirements for Operationally Aligned AI in Broadband Networks

Sections 2 through 4 established that most AI systems in broadband operations originate from enterprise or cloud environments and were never engineered to interpret DOCSIS or HFC behavior. These models assume statistical stability where none exists, collapsing when faced with non-stationary data, fragmented telemetry, or physics-driven plant behavior. As a result, maturity scores and automation levels routinely exaggerate operational readiness while

concealing the fragility beneath.
AI systems deployed within broadband access network architectures must therefore satisfy engineering requirements defined by the physical and architectural behavior of access networks.

This section formalizes those requirements. Each is grounded in operational reality, supported by industry documentation, and directly tied to the failure modes identified in preceding sections. Together, they constitute an architectural framework for determining whether an AI system is operationally safe, technically sound, and contextually aligned with broadband network behavior. In essence, operationally aligned AI is not a layer of analytics applied to the network, but a reasoning framework built within it; one that interprets cause, not merely correlation.

Domain-Integrated System Understanding

AI systems must interpret telemetry through the architectural and physical logic that defines DOCSIS, HFC, and DAA environments. These models must reason within the constraints of RF physics, DOCSIS scheduling, coding and modulation behavior, and access-topology design, treating the network not as data, but as a living system with structure, timing, and causality. A domain-integrated model should:

- recognize that impulse noise, micro-reflections, and CPD each have distinct signatures.
- differentiate congestion from physical impairment
- account for DOCSIS MAC behavior, including minislot allocation, request-grant cycles, and upstream scheduling
- incorporate the impact of the Dynamic Range Window on upstream power behavior
- identify partial-service states that distort throughput patterns without representing full impairment
- reconcile visibility differences between RPD and CMTS telemetry in DAA

Without domain context, AI treats DOCSIS signals as generic numerical data instead of structured physical-layer information. The result is misinterpretation, false alarms, and unstable recommendations. Domain awareness is therefore the foundation of operationally aligned AI.

Telemetry Completeness and Cross-Layer Visibility

No single telemetry plane can fully describe the behavior of a broadband network. PHY-layer signals, MAC-layer metrics, and customer-experience indicators each reveal distinct dimensions of network health, yet none are complete on their own. A complete telemetry model should include:

- **PHY metrics:** MER, FEC, upstream SNR, pre- and post-equalization coefficients
- **MAC metrics:** minislot contention, MAP behavior, request-grant cycles, partial-service indicators

- **Customer-experience telemetry:** latency, packet loss, application-layer performance, quality of experience (QoE) metrics per Broadband Forum TR-452
- **Architectural telemetry:** RPD-side vs CMTS-side reports, channel maps, OFDMA profiles
- **Environmental and temporal factors:** seasonal variation, temperature-related attenuation, weather-driven ingress

Particularly in DAA, models must reconcile measurements from multiple vantage points. AI limited to CMTS, RPD, or CPE data alone produces incomplete or misleading results. Cross-layer correlation is essential for distinguishing congestion from impairment, RF issues from scheduling behavior, and physical instability from user-driven variation. Essentially, cross-layer visibility is what transforms AI from a statistical observer into an operational participant, capable not just of describing what the network does, but of understanding why it does it.

Robustness Under Real Plant Conditions

AI systems must demonstrate resilience under the very conditions that violate the statistical assumptions most machine-learning methods depend upon. Evaluation must move beyond simulation or lab stability testing and prove reliability under the unpredictable, noisy, and asymmetrical realities of the plant itself.

Stress testing should include:

- impulse-noise bursts and upstream interference
- MER and SNR fluctuations from micro-reflections
- seasonal and temperature-related drift
- upstream noise funneling and aggregation
- partial-service states
- asymmetric telemetry timing
- service-group contention
- weather-related ingress and physical disturbance

Without stress testing, AI systems may appear stable in lab conditions yet fail under operational load. Robustness testing is therefore a mandatory validation step for any operational AI. Reliability in this context is not defined by stability alone but by survival under variance; the system’s ability to retain interpretive accuracy even when telemetry becomes inconsistent, incomplete, or contradictory.

Operator-Aligned Success Metrics

Conventional machine learning (ML) metrics such as accuracy, precision, recall, and F1 describe mathematical performance, not operational safety. In broadband environments, an error is not an abstract statistic; it is an action taken on a live network. These actions carry asymmetric consequences that directly influence customer experience, plant stability, and maintenance workload.

- **False positives** create unnecessary operational activity, corrective actions that destabilize the plant, or oscillating configurations.
- **False negatives** allow genuine impairments to persist, causing long-term degradation and unresolved service issues.

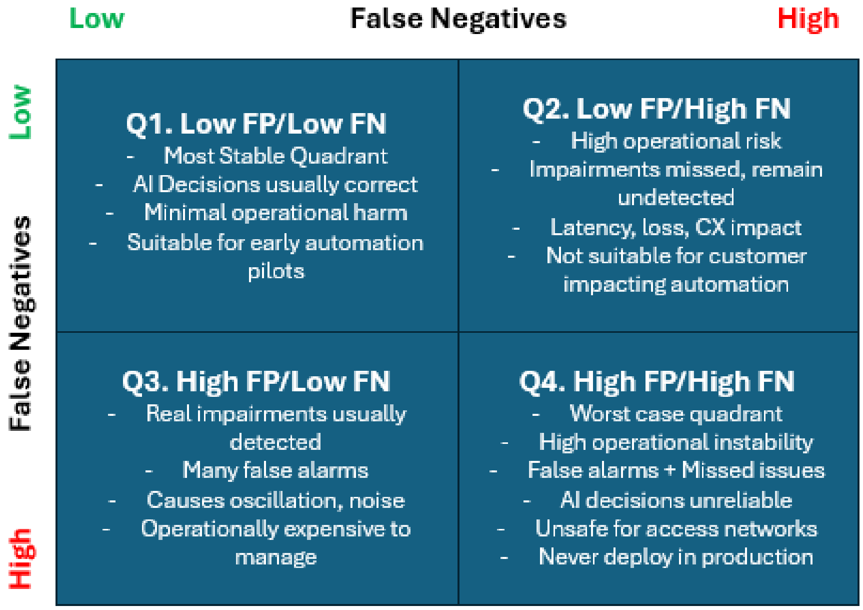
A model with high statistical accuracy may still be unsafe if its residual errors contain many false negatives in noise detection, partial-service recognition, or OFDMA stability. Conversely, a model producing excessive false positives in high-upload scenarios may overwhelm operational systems with alerts or trigger needless dispatches.

To evaluate readiness, operators must assess AI according to the *operational impact* of its errors. The **False-Positive / False-Negative Operational Risk Matrix** provides a conceptual tool for this assessment. It helps engineers determine which models are safe for automation, which require human oversight, and which should remain in shadow evaluation.

Evaluation steps:

1. Characterize error tendencies under real plant conditions (noise bursts, drift, load, ripple, telemetry divergence).
2. Map the pattern to one of four matrix quadrants, dominant FP, FN, or both.
3. Determine deployment suitability by operational risk:
   - *Low FP / Low FN (Quadrant 1):* stable, suitable for controlled automation.
   - *Low FP / High FN (Quadrant 2):* high customer-impact risk, requires tight oversight.
   - *High FP / Low FN (Quadrant 3):* operationally expensive, best for advisory use.
   - *High FP / High FN (Quadrant 4):* unsafe, not for production deployment.

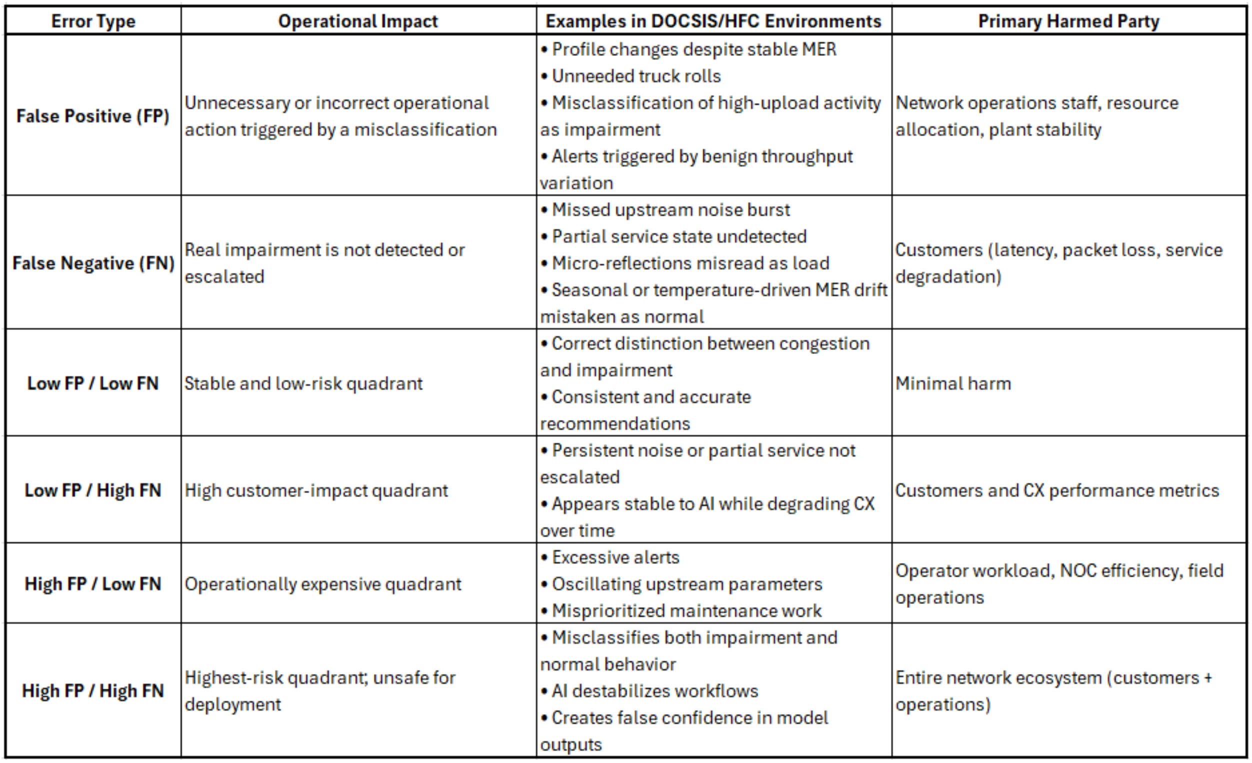
This risk-based framework aligns with NIST AI RMF guidance on context-dependent harm and reflects broadband network operational practice, where the cost of misclassification varies by time, load, impairment, and layer.



	Low — False Negatives — High	
False Negatives	Low	High
Low	**Q1. Low FP/Low FN** - Most Stable Quadrant - AI Decisions usually correct - Minimal operational harm - Suitable for early automation pilots	**Q2. Low FP/High FN** - High operational risk - Impairments missed, remain undetected - Latency, loss, CX impact - Not suitable for customer impacting automation
High	**Q3. High FP/Low FN** - Real impairments usually detected - Many false alarms - Causes oscillation, noise - Operationally expensive to manage	**Q4. High FP/High FN** - Worst case quadrant - High operational instability - False alarms + Missed issues - AI decisions unreliable - Unsafe for access networks - Never deploy in production

Figure 13. False-Positive vs False-Negative Operational Risk Matrix

A conceptual illustration of how FP and FN error patterns influence operational risk in broadband networks.



Error Type	Operational Impact	Examples in DOCSIS/HFC Environments	Primary Harmed Party
False Positive (FP)	Unnecessary or incorrect operational action triggered by a misclassification	• Profile changes despite stable MER • Unneeded truck rolls • Misclassification of high-upload activity as impairment • Alerts triggered by benign throughput variation	Network operations staff, resource allocation, plant stability
False Negative (FN)	Real impairment is not detected or escalated	• Missed upstream noise burst • Partial service state undetected • Micro-reflections misread as load • Seasonal or temperature-driven MER drift mistaken as normal	Customers (latency, packet loss, service degradation)
Low FP / Low FN	Stable and low-risk quadrant	• Correct distinction between congestion and impairment • Consistent and accurate recommendations	Minimal harm
Low FP / High FN	High customer-impact quadrant	• Persistent noise or partial service not escalated • Appears stable to AI while degrading CX over time	Customers and CX performance metrics
High FP / Low FN	Operationally expensive quadrant	• Excessive alerts • Oscillating upstream parameters • Misprioritized maintenance work	Operator workload, NOC efficiency, field operations
High FP / High FN	Highest-risk quadrant; unsafe for deployment	• Misclassifies both impairment and normal behavior • AI destabilizes workflows • Creates false confidence in model outputs	Entire network ecosystem (customers + operations)

Table 1. Operational Consequences of FP vs FN

Examples showing how misclassification patterns manifest in DOCSIS / HFC behavior and affect customer experience and workload.

Explainability and Operator Interpretability

Operators must not only understand but also interrogate AI reasoning. Explainability in this context extends beyond transparency, it is the operator's right to contest the model's conclusions and trace how it arrived there. This interpretive visibility preserves situational awareness and mitigates automation bias. IEEE guidance on transparency and auditability requires that users have clear insight into system decisions.

Operational interpretability demands:

- visibility into the data sources used by the model
- traceability of its reasoning and decision path
- clarity on confidence bounds and uncertainty
- ability to verify outputs against engineering indicators
- mechanisms for operator override without service disruption

These capabilities ensure that AI functions as an extension of engineering judgment rather than a substitute for it. When interpretability is preserved, the operator remains the final arbiter of safety, and the system's intelligence becomes accountable rather than opaque.

Lifecycle Governance, Drift Management, and Safety Controls

AI deployed within broadband infrastructure must operate under the same rigor as any other control-plane system. Governance is not an administrative layer but a technical safeguard, the mechanism through which reliability, traceability, and accountability are enforced. Frameworks such as ETSI Zero-Touch Network and Service Management (ZSM) and NIST AI RMF provide the foundation, but operational implementation determines whether safety is preserved or compromised. Operational governance includes:

- model versioning and change tracking
- drift detection for divergent outputs
- periodic retraining or recalibration across seasonal cycles
- rollback mechanisms for rapid reversion
- validation gates to block unverified models
- integration with existing change-control processes
- safety valves that pause automated actions under uncertainty

These controls ensure that AI cannot act in silence, cannot drift into autonomy, and cannot obscure its influence on the network. When governance is embedded at the architectural level, operational trust is preserved, not because the system is perfect, but because every decision remains visible, reversible, and accountable. Together, the requirements in this section form the engineering foundation of operationally aligned AI. They define not only how AI must behave, but how it must be built, anchored in domain understanding, sustained by cross-layer telemetry, stress-tested under real conditions, and governed through traceable, reversible control. This structure transforms AI from a predictive utility into a disciplined participant within the operational fabric of the network.

Application Examples in Modern Cable Architectures

The requirements defined in Section 5 become tangible when applied to real operational contexts. DOCSIS, HFC, DAA, and CMTS architectures each present conditions that challenge AI systems lacking domain awareness or complete telemetry. These examples demonstrate how physical-layer dynamics, architectural partitioning, and environmental variability collectively determine model behavior, illustrating why operationally aligned AI must integrate the structural, causal, and telemetry-completeness principles established earlier.

The following cases originate from documented field conditions in CableLabs PNM studies, SCTE operational guidance, and vendor analyses. Each represents a verified, real-world scenario that exposes how AI misinterpretation emerges from architectural or physical constraints, not from theoretical error but from systemic mismatch between model assumptions and network physics.

Misclassification of Upstream Utilization Due to Noise

Scenario:
Upstream noise bursts elevate FEC error rates or briefly depress upstream SNR, reducing throughput and altering minislot scheduling behavior. AI systems that rely only on throughput or load statistics often misread this condition as user-driven congestion.

Why it happens:

- The model treats noise-induced throughput reduction as user-driven activity.
- It lacks correlation with PHY-layer metrics showing noise bursts.
- Telemetry sampling is misaligned across CMTS, RPD, and CPE sources.

Required capabilities:

- Cross-layer correlation of upstream FEC, MER, minislot behavior, and throughput.
- Recognition of characteristic noise signatures documented in PNM research.
- Ability to distinguish impairment from legitimate user contention.

Profile-Management Errors and Oscillation

Scenario:
Dynamic profile-management systems can inadvertently reduce downstream or upstream stability when they react to transient or misinterpreted conditions, including:

- Temporary MER dips that prompt unnecessary profile reductions
- Oscillations between modulation states caused by short-term metric fluctuations
- Overcorrections triggered when impairment signatures are inaccurately classified

Why it happens:

- MER fluctuations caused by micro-reflections or ripple are misread as sustained impairment.
- Noise signatures are not contextualized with plant behavior.

Required capabilities:

- Understanding of OFDM modulation stability and ripple characteristics.
- Robustness to short-duration impairment patterns.
- Validation of recommendations against customer-impact metrics.

Proactive-Maintenance Misdiagnosis

Scenario:
Predictive and anomaly-detection models trained on limited or single-plane telemetry often misclassify impairment when encountering complex plant conditions such as:

- partial-service states that artificially reduce throughput
- environmental drift that alters MER or FEC stability
- node-temperature variations that mimic long-term degradation
- DAA visibility gaps that conceal localized upstream noise

Why it happens:

- Training uses single-plane or incomplete datasets.
- The model assumes stationarity across environmental changes.
- Divergence between CMTS-side and RPD-side data cannot be reconciled.

Required capabilities:

- Multi-layer telemetry integration including environmental data.
- Awareness of partial-service and MAC-layer interactions.
- Stress-tested interpretation stable during environmental drift.

Misinterpretation of Service-Group Behavior

Scenario:
In service groups with high-upload users, legitimate usage can mimic impairment symptoms. Reduced modulation efficiency or increased mini-slot contention may result from user-driven load rather than physical plant instability.

Why it happens:

- The model assumes reduced modulation always signals impairment.
- It ignores service-group usage patterns.
- Customer-experience indicators are absent.

Required capabilities:

- Integration of customer-experience telemetry, including latency and application behavior.
- Differentiation between user-driven contention and plant-driven impairment.
- Modeling that avoids overreaction to valid usage variation.

Environmental and Seasonal Effects

Scenario:
Temperature fluctuations, humidity, and moisture ingress alter attenuation, noise susceptibility, and component stability across the coaxial plant. These variations introduce predictable yet complex shifts in signal behavior that many AI systems fail to model.

Why it happens:

- The model assumes baseline stability.
- Training data do not span multiple seasons.
- Environmental inputs are excluded from telemetry.

Required capabilities:

- Seasonal and environmental normalization.
- Drift-aware modeling consistent with NIST guidance on non-stationarity.
- Correlation of environmental and PHY-layer data.

DAA Telemetry Divergence and Interpretation Failures

Scenario:
In DAA architectures, Remote PHY Devices (RPDs) handle physical-layer processing while the virtual CMTS (vCMTS) retains MAC-layer scheduling and control. This architectural split often

creates visibility divergence: a cable modem may record rising upstream error rates while the RPD reports stable MER.

Why it happens:

- AI relies on a single vantage point (RPD, CMTS, or CPE).
- Timing irregularities distort correlation.
- Architectural split complicates root-cause tracing.

Required capabilities:

- Cross-architecture telemetry reconciliation.
- Awareness of RPD and CMTS interpretation differences.
- Inclusion of customer-experience metrics to confirm impact.

Case Example: High-Upload Users Masking Upstream Impairment

Scenario:
In some service groups, a concentration of high-upload users can generate usage patterns that mask underlying impairment. During periods of heavy load, noise bursts overlap with legitimate traffic, causing the AI system to classify the impairment as normal usage rather than physical disturbance.

Why it happens:

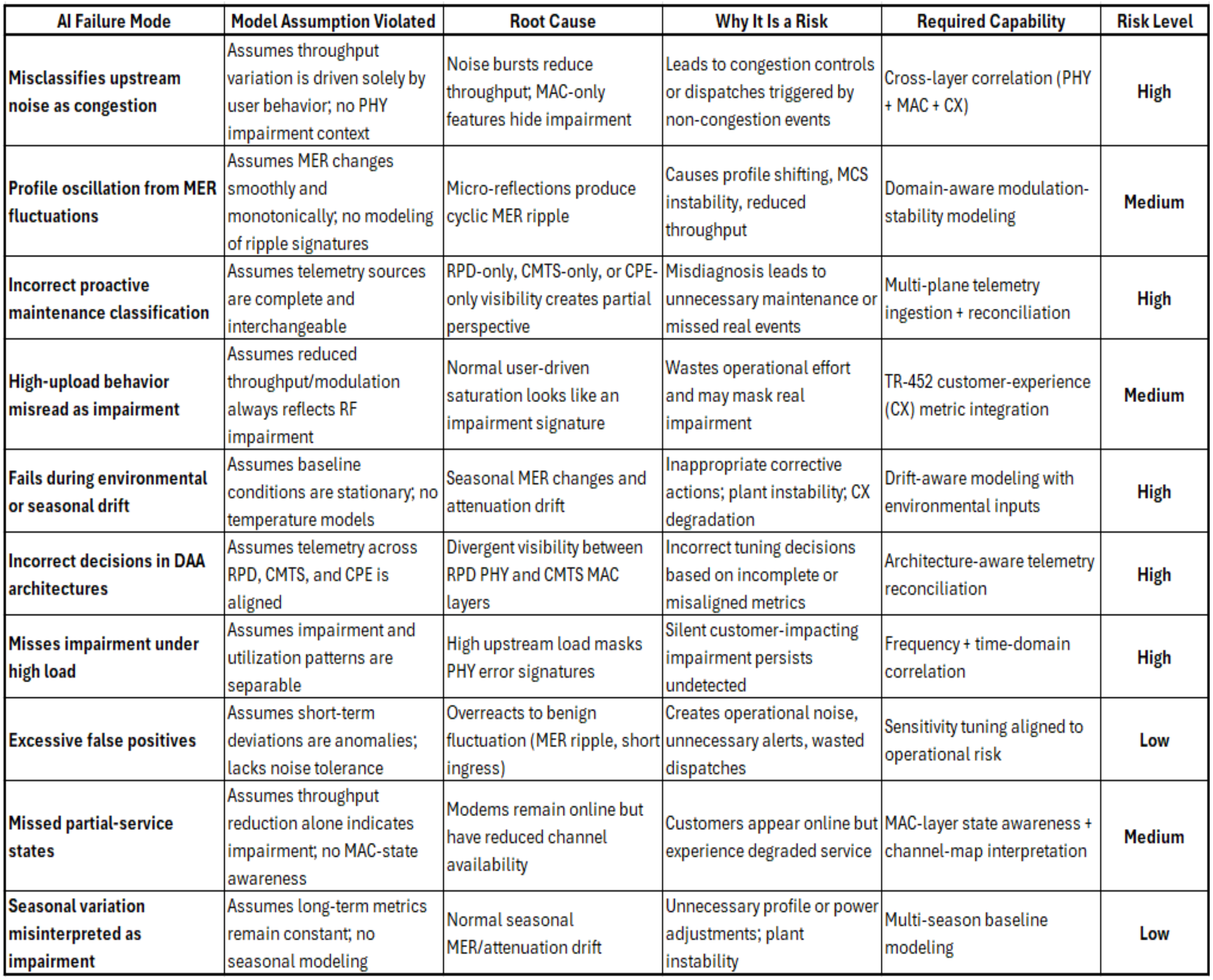
- Utilization and impairment signals overlap.
- Frequency-domain impairment is undetected under heavy load.
- Customer-experience degradation conflicts with misinterpreted PHY metrics.

Required capabilities:

- Correlation across frequency- and time-domains.
- Integration of multi-layer telemetry.
- Differentiation between user-driven and impairment-driven load changes.

Summary – AI Failure Modes, Root Causes, and Required Capabilities

The case studies above illustrate a consistent pattern: when AI models operate without structural awareness of network physics, they fail in predictable and diagnosable ways. The underlying causes – violated statistical assumptions, fragmented telemetry, and lack of cross-layer reasoning – are systemic rather than incidental. Table X consolidates these findings, linking each failure mode to the violated modeling assumption, the physical behavior that produced it, the resulting operational risk, and the capability required to prevent recurrence. This table transforms anecdotal misinterpretations into a structured taxonomy of AI failure, providing operators with a practical diagnostic reference for model validation before deployment.



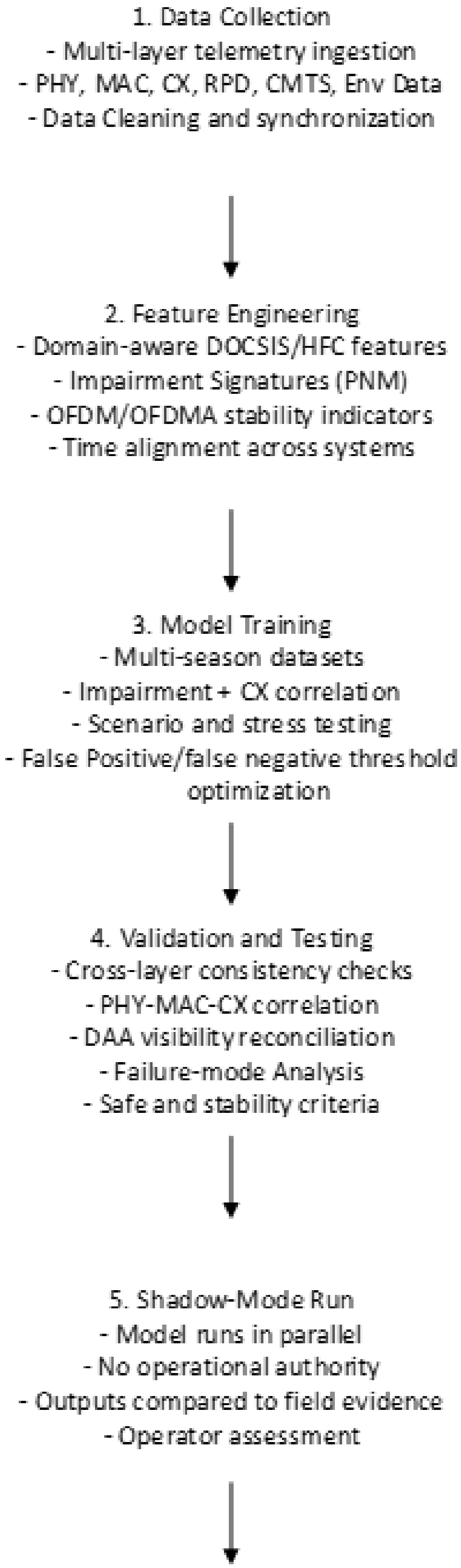
AI Failure Mode	Model Assumption Violated	Root Cause	Why It Is a Risk	Required Capability	Risk Level
Misclassifies upstream noise as congestion	Assumes throughput variation is driven solely by user behavior; no PHY impairment context	Noise bursts reduce throughput; MAC-only features hide impairment	Leads to congestion controls or dispatches triggered by non-congestion events	Cross-layer correlation (PHY + MAC + CX)	**High**
Profile oscillation from MER fluctuations	Assumes MER changes smoothly and monotonically; no modeling of ripple signatures	Micro-reflections produce cyclic MER ripple	Causes profile shifting, MCS instability, reduced throughput	Domain-aware modulation-stability modeling	**Medium**
Incorrect proactive maintenance classification	Assumes telemetry sources are complete and interchangeable	RPD-only, CMTS-only, or CPE-only visibility creates partial perspective	Misdiagnosis leads to unnecessary maintenance or missed real events	Multi-plane telemetry ingestion + reconciliation	**High**
High-upload behavior misread as impairment	Assumes reduced throughput/modulation always reflects RF impairment	Normal user-driven saturation looks like an impairment signature	Wastes operational effort and may mask real impairment	TR-452 customer-experience (CX) metric integration	**Medium**
Fails during environmental or seasonal drift	Assumes baseline conditions are stationary; no temperature models	Seasonal MER changes and attenuation drift	Inappropriate corrective actions; plant instability; CX degradation	Drift-aware modeling with environmental inputs	**High**
Incorrect decisions in DAA architectures	Assumes telemetry across RPD, CMTS, and CPE is aligned	Divergent visibility between RPD PHY and CMTS MAC layers	Incorrect tuning decisions based on incomplete or misaligned metrics	Architecture-aware telemetry reconciliation	**High**
Misses impairment under high load	Assumes impairment and utilization patterns are separable	High upstream load masks PHY error signatures	Silent customer-impacting impairment persists undetected	Frequency + time-domain correlation	**High**
Excessive false positives	Assumes short-term deviations are anomalies; lacks noise tolerance	Overreacts to benign fluctuation (MER ripple, short ingress)	Creates operational noise, unnecessary alerts, wasted dispatches	Sensitivity tuning aligned to operational risk	**Low**
Missed partial-service states	Assumes throughput reduction alone indicates impairment; no MAC-state awareness	Modems remain online but have reduced channel availability	Customers appear online but experience degraded service	MAC-layer state awareness + channel-map interpretation	**Medium**
Seasonal variation misinterpreted as impairment	Assumes long-term metrics remain constant; no seasonal modeling	Normal seasonal MER/attenuation drift	Unnecessary profile or power adjustments; plant instability	Multi-season baseline modeling	**Low**

Table 2. Summary of AI failure modes

The table shows violated model assumptions, operational risks, and required capabilities in DOCSIS / HFC access networks.

Blueprint for Deploying Safe, Ethical, and Operationally Valid AI

Broadband access networks demand AI systems capable of interpreting multi-layer telemetry, reasoning within DOCSIS and HFC physics, and maintaining reliability under conditions that routinely break conventional machine-learning assumptions. Building on the technical, ethical, and operational foundations established earlier, this blueprint defines a governance-driven framework for deploying AI that directly influences maintenance, configuration, and customer-experience outcomes across DOCSIS, HFC, DAA, and CMTS environments. It unifies guidance from NIST AI RMF, ETSI ZSM, Broadband Forum TR-452, and CableLabs operational documentation into a repeatable lifecycle that ensures AI remains transparent, auditable, and operationally safe.



1. Data Collection
- Multi-layer telemetry ingestion
- PHY, MAC, CX, RPD, CMTS, Env Data
- Data Cleaning and synchronization

2. Feature Engineering
- Domain-aware DOCSIS/HFC features
- Impairment Signatures (PNM)
- OFDM/OFDMA stability indicators
- Time alignment across systems

3. Model Training
- Multi-season datasets
- Impairment + CX correlation
- Scenario and stress testing
- False Positive/false negative threshold optimization

4. Validation and Testing
- Cross-layer consistency checks
- PHY-MAC-CX correlation
- DAA visibility reconciliation
- Failure-mode Analysis
- Safe and stability criteria

5. Shadow-Mode Run
- Model runs in parallel
- No operational authority
- Outputs compared to field evidence
- Operator assessment

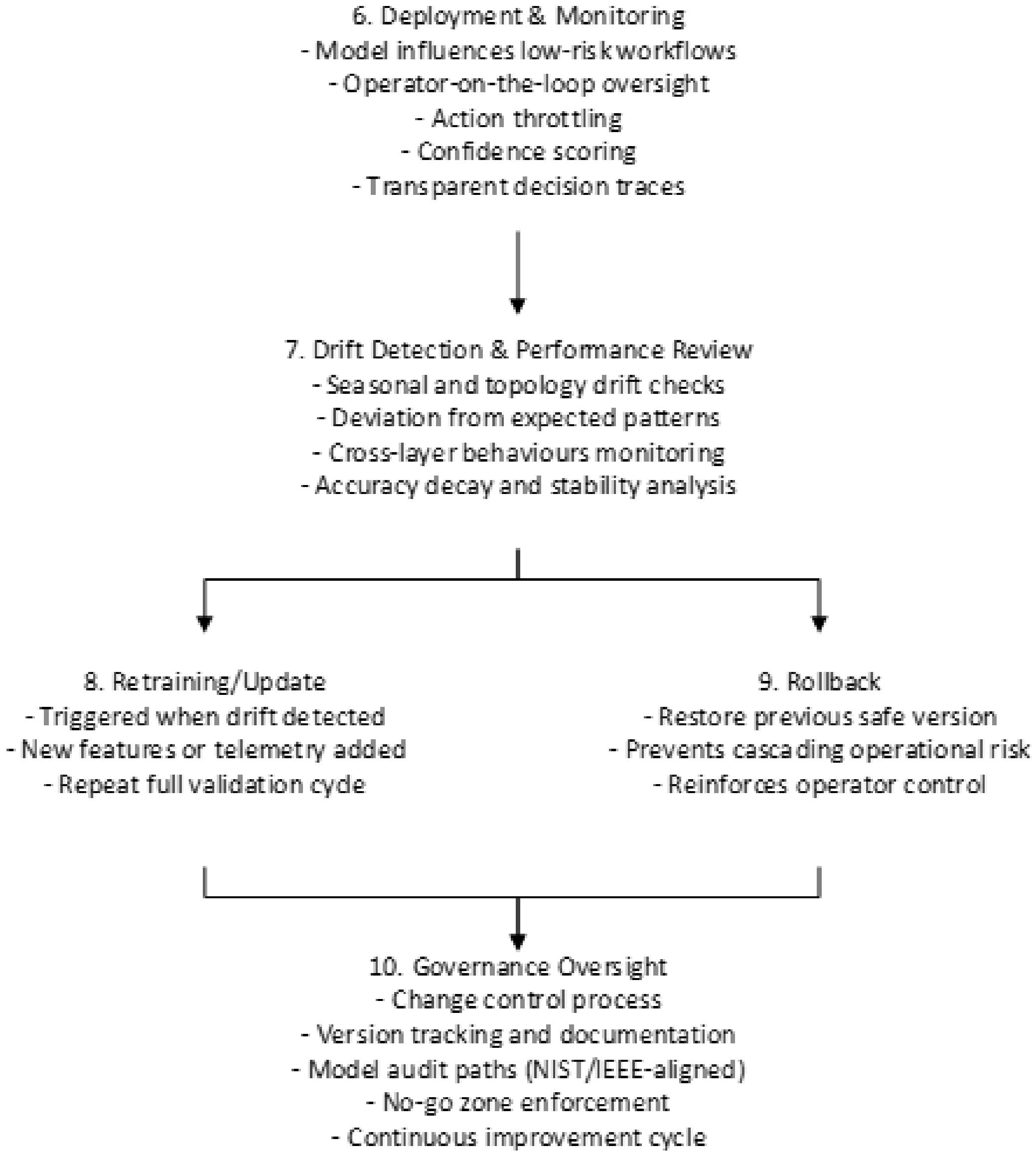

6. Deployment & Monitoring
- Model influences low-risk workflows
- Operator-on-the-loop oversight
- Action throttling
- Confidence scoring
- Transparent decision traces
7. Drift Detection & Performance Review
- Seasonal and topology drift checks
- Deviation from expected patterns
- Cross-layer behaviours monitoring
- Accuracy decay and stability analysis
8. Retraining/Update
- Triggered when drift detected
- New features or telemetry added
- Repeat full validation cycle
9. Rollback
- Restore previous safe version
- Prevents cascading operational risk
- Reinforces operator control
10. Governance Oversight
- Change control process
- Version tracking and documentation
- Model audit paths (NIST/IEEE-aligned)
- No-go zone enforcement
- Continuous improvement cycle

Figure 14. Operational AI Lifecycle and Governance Loop

A high-level model illustrating the continuous governance cycle that sustains AI safety and reliability in broadband operations. It shows how data acquisition, model evaluation, drift monitoring, operator oversight, and controlled automation interact as an integrated feedback system, ensuring that AI remains transparent, stable, and aligned with operational intent.

Architectural Requirements for Operational AI

Operationally aligned AI must rest on a network-aware architecture that mirrors both the physical and logical realities of the access plant. The design must capture how telemetry, timing, and topology interact across DOCSIS, HFC, DAA, and virtualized environments to ensure AI operates as a coherent extension of the network rather than as an external analytic layer.

1. **Multi-layer telemetry ingestion**
   No single data plane provides a complete view of network health. Operationally aligned AI **MUST** merge telemetry across all functional layers to preserve causal continuity between physical behavior and customer experience. Integration should include:

   - PHY layer: MER, SNR, FEC rates, OFDM profile parameters, and pre/post-equalization coefficients.
   - MAC layer: minislot contention, request-grant cycles, scheduling behavior, and partial-service signals.
   - Customer experience layer: latency, loss, application-responsiveness, and QoE metrics defined in TR-452.
   - Architectural context: RPD-side and CMTS-side telemetry in DAA and virtualized CMTS (vCMTS) environments.
   - Environmental factors: temperature, humidity, weather, and seasonal influence on signal attenuation.

2. **Time alignment and normalization**
   All telemetry sources **MUST** operate within a synchronized temporal framework. Misaligned timestamps distort causality, producing artificial correlations or hiding true dependencies. Time normalization therefore functions not as preprocessing, but as a prerequisite for valid causal reasoning and model stability.

3. **Domain-aware feature engineering**
   Features **MUST** be engineered within the physical and protocol constraints that define DOCSIS and HFC behavior. Feature construction should capture how signal, scheduling, and topology interact so that the model's variables mirror operational reality rather than abstract data trends. Examples include:

   - Impairment-signature extraction based on PNM characteristics and spectral fingerprinting.
   - Modulation-stability indicators derived from OFDM and OFDMA profile dynamics.
   - Upstream noise-window modeling aligned with request-grant cycles and minislot timing.
   - Dynamic Range Window (DRW) modeling that reflects power-adjustment logic and physical-layer constraints.

4. **Architecture-aware correlation**
   Distributed Access Architectures (DAA) fragment network visibility between multiple control and processing layers. To maintain causal accuracy, AI **MUST** correlate telemetry across the RPD, CMTS or vCMTS, and CPE before inference. Architectural correlation ensures the system recognizes when identical behaviors originate from different sources; for example, distinguishing PHY-level disturbance from MAC scheduling or transport delay. Only after reconciling these perspectives can an AI model safely infer plant state or customer impact.

5. **Data survivability and completeness**
   AI systems **MUST** remain functional under data loss, telemetry gaps, or architectural blind spots. Data survivability refers not only to redundancy but to continuity of inference when visibility is degraded. Models must use fallback logic, temporal interpolation, and confidence weighting to maintain stability during partial observability. In broadband networks, the absence of data is itself a signal, it must be detected, contextualized, and incorporated into decision logic rather than ignored.

Data Acquisition Strategy

Reliable AI performance begins with the integrity and representativeness of its data foundation. In broadband operations, the quality of AI outcomes is inseparable from the diversity, temporal breadth, and physical realism of the data used to train and validate the system. A model's judgment is only as broad as the conditions it has witnessed; therefore, datasets must reflect the plant's full operational range, across time, topology, and environment, to avoid narrow inference and blind operational zones.

1. **Seasonal coverage**
   Training data **MUST** include multiple seasons to capture drift, attenuation shifts, and temperature effects documented by SCTE.

2. **Diverse impairment patterns**
   Datasets **MUST** include examples of impulse noise, micro-reflections, CPD, ingress, upstream noise funneling, intermittent impairments, and DAA visibility divergence.

3. **Customer-experience alignment**
   Data **MUST** include latency, application performance, and QoE indicators per TR-452, not only throughput or counter data.

4. **Labeling grounded in engineering validation**
   Labels define truth within an AI system, and in broadband operations, that truth **MUST** be anchored in engineering evidence. Labeling should be validated through spectrum traces, PNM correlation, or technician-confirmed field outcomes, not through statistical clustering or assumption-based inference. When labels are wrong, every layer of the model inherits that

error. Unverified labeling is therefore not merely a data issue but a propagation of technical falsehood into operational logic.

5. **Continuous data-quality evaluation**
   Telemetry **MUST** be monitored for integrity, alignment, and completeness to prevent spurious model behavior. Reliable data acquisition is not a one-time achievement but a continuous act of calibration between the physical network and its digital reflection. Every dataset is a snapshot of reality filtered through measurement constraints, timing offsets, and environmental variance. Operationally aligned AI must therefore treat data as a living system, audited, updated, and revalidated as the plant evolves. This dynamic approach ensures that model fidelity grows with network maturity rather than diverging from it.

Integration with Service-Management Plans (SMPs) and KPI Frameworks

Operational AI **MUST** exist within the same structural discipline that governs every other element of broadband assurance. Service-Management Plans and key performance indicator (KPI) frameworks provide the scaffolding that links data interpretation to operational action, ensuring that AI outputs reinforce, rather than bypass, established control mechanisms. In this context, AI is not an external intelligence but an analytical participant in the SMP process, its insights must map directly to:

- Service Management Plans
- HFC or DOCSIS performance indicators
- impairment-detection processes
- proactive-maintenance workflows
- customer-impact thresholds
- escalation procedures

KPI alignment
AI outcomes **MUST** be evaluated against metrics such as:

- modulation stability
- upstream reliability
- noise and interference behavior
- customer-experience metrics
- service-group performance patterns

Integrating AI into the SMP and KPI ecosystem transforms metrics from passive indicators into active validation tools. Every AI-generated recommendation should trace back to a measurable outcome within these frameworks, whether modulation stability, noise containment, or latency consistency. This closed feedback loop allows AI performance to be judged by its operational consequences rather than statistical fit, creating a self-correcting ecosystem where the network, the data, and the intelligence evolve together.

Guardrails for Autonomous and Assisted Actions

Any AI that participates in operational or configuration decisions must function inside clearly defined safety boundaries. These guardrails preserve the network's stability, maintain operator trust, and ensure that automation complements rather than overrides human judgment. The intent is not to slow innovation but to establish the same rigor applied to other control-plane systems, ensuring that autonomy never exceeds the limits of verified understanding.

1. **Human-in-the-loop (high-impact actions)**
   Operator authorization remains mandatory for any action capable of altering plant state or customer experience. These interventions require explicit human oversight not because AI lacks potential accuracy, but because its reasoning **MUST** remain accountable to engineering judgment. Approval is mandatory for:

   - OFDMA profile changes
   - upstream power adjustments
   - channel or profile migrations
   - DAA timing or synchronization modifications
   - maintenance dispatch decisions initiated from AI indicators

2. **Human-on-the-loop (medium-impact actions)**
   Operators maintain continuous situational awareness over reversible or low-risk actions. In this mode, AI may act autonomously within predefined thresholds, but every action **MUST** remain observable, interruptible, and reversible. This supervisory framework balances agility with accountability and preserves operational transparency. Examples include:

   - issue prioritization
   - advisory recommendations
   - trend and anomaly analysis
   - telemetry correlation and cross-layer data mapping

3. **Action throttling**
   AI **MUST** regulate both the frequency and cumulative scope of its operational actions to avoid oscillation, cascading reversals, or unintended amplification effects. Throttling criteria should include rate limits, confidence thresholds, and situational cooldowns that pause

automation during unstable or conflicting telemetry conditions. This ensures that corrective intent does not transform into instability through overcorrection.

4. **Safety criteria**
   When telemetry diverges, confidence degrades, or conflicting data emerges between layers, the AI **MUST** default to a verifiable safe state. A safe state is defined as one where no autonomous control-plane actions can alter network behavior without operator confirmation. This rule aligns with NIST AI RMF principles of traceability and harm prevention and ensures that uncertainty triggers human review rather than automated escalation.

5. **Change-control alignment**
   All AI-driven modifications **MUST** flow through formal change-management procedures identical to those governing human-initiated network actions. This includes documentation, pre-approval, rollback validation, and post-change review. Integration into the existing Service Management Plan ensures that automated decisions are traceable, reversible, and subject to the same accountability as traditional engineering interventions.

Lifecycle Management and Drift Controls

AI models are not static tools; they evolve alongside the network itself. As plant conditions, traffic patterns, and environmental influences shift, model performance naturally drifts. Without structured lifecycle management, this drift introduces silent degradation, eroding accuracy, stability, and trust. Therefore, lifecycle management in broadband AI **MUST** explicitly follow NIST AI RMF and ETSI ZSM principles, embedding continuous validation, traceability, and safety checkpoints throughout the model's operational lifespan.

1. **Drift detection**
   Drift detection **MUST** operate as a continuous validation process rather than a periodic audit. Monitoring should capture not only statistical deviation but also contextual and causal drift across network layers. Systems must track:

   - performance decline in inference accuracy or correlation with field evidence
   - seasonal deviation from training distributions caused by temperature or ingress variation
   - topology or hardware changes that alter signal propagation or telemetry alignment
   - emergence of new impairment patterns not represented in baseline data (e.g., new ingress sources, plant upgrades, spectrum reallocation)
   - sustained divergence between AI predictions and technician-confirmed outcomes

Detected drift must trigger automated alerts, quarantine of affected models, and escalation through change-control procedures before reactivation.

2. **Versioning and rollback**
   Every AI model **MUST** exist within a controlled versioning system that preserves full lineage, training data, feature definitions, parameter settings, and deployment history. Each version must be traceable to its validation record and operational performance period. Rollback procedures must be pre-tested and executable in real time, ensuring that a prior stable version can be restored immediately if instability or drift is detected.

   This approach mirrors software release management and ensures that AI, once operational, remains subject to the same reliability and accountability standards as network firmware or configuration updates.

3. **Scheduled retraining and validation**
   Retraining **MUST** be scheduled according to both temporal and behavioral triggers. Seasonal cycles, topology changes, and large-scale weather events alter plant behavior in ways that require model recalibration. Retraining should incorporate:

   - updated telemetry from all network layers, including new impairment patterns or modulation behaviors
   - representative data from high-variance conditions (temperature extremes, ingress fluctuations, holiday traffic loads)
   - cross-validation using customer-experience and field-confirmed data to ensure operational accuracy

Validation **MUST** confirm that retrained models improve interpretive stability across PHY, MAC, and CX domains. Retraining without empirical validation risks embedding bias or false correlations rather than resolving them.

4. **Shadow-mode deployment**
   All new models **MUST** operate in shadow mode prior to any active control or recommendation authority. During this phase, model output is logged and compared against real network outcomes, technician observations, and established baseline models.

   Verification criteria include:

   - statistical alignment with validated models under stable plant conditions
   - consistent interpretation of impairment, load, and modulation behavior during stress events
   - absence of unsafe recommendations when telemetry is incomplete or noisy

Exit from shadow mode requires documented validation through change control, with approval from both engineering and operations stakeholders. This ensures that activation follows empirical verification, not assumption.

Required Operator Competencies

AI augments rather than replaces human expertise. Safe and effective use of operational AI requires engineers with cross-domain fluency, capable of interpreting network behavior, validating model logic, and intervening safely when automation diverges from physical reality.

Minimum competency areas:

- deep understanding of DOCSIS PHY and MAC-layer operation, including scheduling, modulation, and telemetry alignment
- ability to interpret impairment signatures, ripple effects, and upstream instability within RF physics context
- working knowledge of AI fundamentals, feature engineering, model drift, and interpretability limits
- methods to test, validate, and benchmark AI outputs against engineering indicators
- practical understanding of governance, change control, and rollback protocols
- situational awareness of automation bias and safe human-in-the-loop procedures

Organizations should establish a formal certification track for "Operational AI Engineers," combining SCTE broadband engineering standards with foundational AI literacy. Continuous cross-disciplinary training ensures that operational oversight evolves alongside model sophistication and network complexity.

Defining "No-Go Zones" for Autonomous AI

Certain domains within access networks must remain restricted from autonomous AI control to prevent unintended or unsafe outcomes. These "no-go zones" represent areas where physical dynamics, telemetry incompleteness, or architectural coupling introduce unacceptable operational risk.

Category A – High-Risk Technical Controls (Direct Physical Impact)

- OFDM/OFDMA profile changes during unstable modulation or temperature drift conditions.
- Upstream power or equalization adjustments near Dynamic Range Window thresholds.
- Remote PHY timing or MAP-interval modifications under load or noise fluctuation.

Category B – Conditional Controls (Require Multi-Layer Validation)

- Channel or service-group migrations during partial-service or noise-affected states.
- Node-split recommendations based on statistical utilization without PHY or CX corroboration.
- Fault-isolation triggers or maintenance dispatch based solely on algorithmic inference.

Category C – Forbidden Contexts (Incomplete Observability)

- Any AI-initiated action under conditions of missing, unsynchronized, or conflicting telemetry.
- Scenarios where CMTS, RPD, and CPE data diverge without validated reconciliation.

Actions within these domains require explicit human authorization and cross-layer verification. Establishing these safety classifications formalizes operational boundaries, ensuring that AI remains an augmenting partner, not an uncontrolled actor, within the network environment.

Together, these architectural, procedural, and governance safeguards form the foundation of operationally aligned AI. They establish a disciplined framework where machine intelligence coexists with human expertise under verifiable control, ensuring that automation improves reliability rather than replacing understanding. In broadband networks, where every signal reflects a physical reality, safety is not achieved through abstraction but through alignment: between data and domain, between automation and accountability, and between intelligence and integrity.

Model Architecture for operationally aligned AI

Broadband access networks operate within the laws of physics, topology, and time. Their behavior follows causal chains, not random correlations. Each modulation shift, noise burst, and load variation has a reason traceable through structure and sequence. For AI to function safely within such environments, its architecture must mirror this reality. Intelligence designed for operational domains must reason in the same dimensional space the network itself occupies; structural, causal, temporal, and intentional, rather than merely detect patterns that correlate without understanding why.

Conventional AI architectures, neural networks, transformer-based language models, and ensemble predictors, excel at finding correlations across vast data but lack the structural awareness needed for engineering domains. They recognize patterns without understanding what generates them. In broadband networks, that distinction is critical: prediction without structure becomes guesswork, and automation without reasoning becomes risk.

Operationally aligned AI demands a model architecture that reflects how the network *actually behaves*: structured, directional, temporal, and guided by intent. Each layer of this architecture contributes a distinct form of understanding, structure defines where things are, causality defines why they interact, time defines when behaviors emerge, and intent defines what should be done about them. The following subsections outline the five model classes that together form a coherent reasoning stack capable of true operational intelligence.

Knowledge Graph: Structural Representation of the Network

The Knowledge Graph (KG) forms the structural backbone of operationally aligned AI. It encodes the physical, logical, and functional relationships among every element of the access network, nodes, amplifiers, taps, modems, CMTS resources, service-group associations, and the configuration and historical behavior of each.

The KG captures adjacency, lineage, capacity limits, and equipment dependencies, forming an internal "map" of the network. This spatial intelligence allows the AI to reason with awareness of physical context, understanding how one element's behavior affects another across the topology. Without this structural grounding, higher-level reasoning collapses into correlation. The KG therefore provides the spatial and logical foundation that transforms predictive analysis into operational understanding.

Directed Acyclic Graph: Causal Structure of System Behaviour

While the KG defines *what* is connected, it does not explain *how* behavior flows through those connections. That function belongs to the Directed Acyclic Graph (DAG). The DAG encodes causal relationships among variables, how load influences modulation stability, how temperature drives plant drift, how voltage ripple alters upstream MER, and how impairments propagate through broadband access domains.

The DAG is directional and acyclic, meaning it represents systems where cause always precedes effect. By enforcing that structure, it prevents the model from mistaking correlation for causation. The DAG explicitly encodes the operational dependencies that experienced engineers recognize, how and why network events unfold in sequence and with consequence.

Structural Causal Model: The Reasoning Engine

The Structural Causal Model (SCM) forms the reasoning core of the architecture, operating on top of the DAG and KG. It performs causal inference, evaluating competing hypotheses, quantifying causal effects, identifying confounding conditions, performing counterfactual reasoning, and managing uncertainty within operational limits.

Unlike statistical models that infer similarity, the SCM tests explanation. It applies the structural equations defined in the DAG to evaluate which causal pathway best fits observed behavior. This mirrors the reasoning process of skilled operators, hypothesis, verification, and causal validation, yielding explanations that are interpretable, verifiable, and operationally meaningful, grounded in the measurable physics of the network rather than abstract correlation.

State Model: Temporal Behaviour and System Memory

Networks are dynamic systems whose behaviour evolves across time. Daily load cycles, temperature and seasonal shifts, local events, and maintenance windows all change how signals present on the plant. The state model supplies temporal coherence. It preserves baselines, tracks recurrent impairments, separates short-lived anomalies from persistent drift, and maintains the memory required for reliable interpretation.

This temporal dimension is essential for accurate causal reasoning. Without it, the SCM cannot differentiate chronic impairments from transient ones, nor can it identify when behaviour departs meaningfully from historical norms. The state model ensures that conclusions are grounded in both present and past behaviour.

Operator Intent Model: Alignment With Practical and Ethical Priorities

Operator intent defines how technical observations are transformed into operational meaning. Two identical events can carry different implications depending on service-group importance, customer segment, time sensitivity, or the potential harm of inaction.

The operator intent model captures this contextual intelligence. It encodes reliability priorities, policy thresholds, acceptable risk, and definitions of both healthy and degraded states. By translating causal conclusions into actions that align with operator judgment and network policy, it ensures that AI reasoning remains anchored in real operational priorities, preventing false escalation, missed events, or inappropriate intervention.

How the Models Interlock to Form Operationally Aligned AI

- The knowledge graph provides the structural map of the network.
- The DAG defines how cause flows through that structure.
- The SCM uses both to evaluate competing hypotheses and isolate the most probable root cause.
- The state model adds temporal grounding, linking present behaviour to historical patterns and drift.
- The operator intent model converts conclusions into actions that respect operational priorities and ethical boundaries.

Layered together, these components transform AI from a statistical observer into an operational reasoning system. They allow the model to reason within the network, explaining why behaviour occurs, how it evolved, and what actions are most appropriate in context.

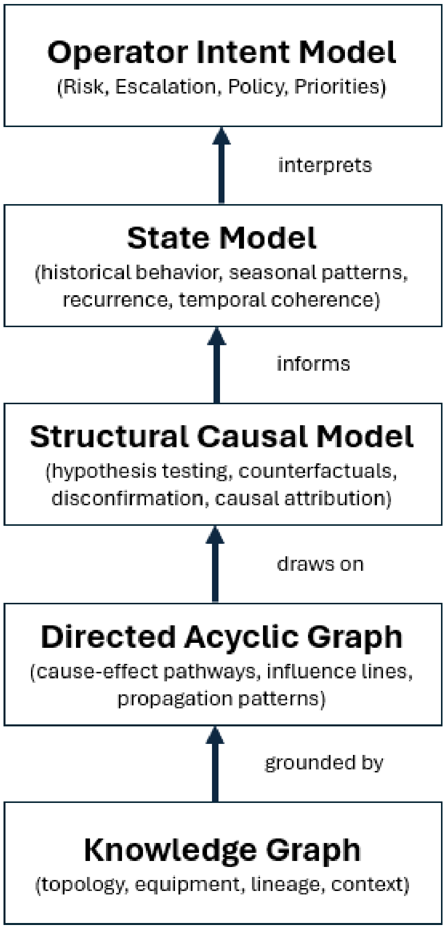

Operator Intent Model
(Risk, Escalation, Policy, Priorities)
interprets
State Model
(historical behavior, seasonal patterns, recurrence, temporal coherence)
informs
Structural Causal Model
(hypothesis testing, counterfactuals, disconfirmation, causal attribution)
draws on
Directed Acyclic Graph
(cause-effect pathways, influence lines, propagation patterns)
grounded by
Knowledge Graph
(topology, equipment, lineage, context)

Figure 15. Proposed model stack for operationally aligned AI

The model stack for Operationally Aligned AI integrates structural, causal, temporal, and intent-based reasoning. Knowledge Graphs provide topological grounding, DAGs encode causal pathways, SCMs perform hypothesis testing, state models provide temporal context, and operator intent ensures alignment with operational priorities.

Example: The Model Stack Diagnosing an Upstream Noise Event

1. **Event Trigger**

   At 19:42, the system detects a rise in upstream noise on an OFDMA channel serving Node X1.

The raw telemetry alone cannot determine whether the impairment is meaningful, temporary, or severe.
This is where the model stack activates.

2. **Layer-by-Layer Reasoning**

   - **Layer 1: Knowledge Graph (KG)**

The KG immediately grounds the event in the network's structure.

From the KG, the system knows:

- Node X1 is fed by Power Supply P1
- The affected OFDMA channel is shared by 412 modems
- A legacy amplifier on the northern leg has a history of thermal drift
- This segment sits on aerial plant exposed to weather
- The CMTS involved is using a known sensitive profile for mid-split return

The KG gives location, dependencies, and equipment lineage.

The system now knows *where* the problem is and *what assets* are involved.

- **Layer 2: Directed Acyclic Graph (DAG)**

  The DAG provides the **causal structure** required to interpret the event.

  The DAG includes relationships like:

  - Temperature → Amplifier Gain Drift
  - Gain Drift → Upstream MER Degradation
  - MER Degradation → Increased Corrected Errors
  - Power Ripple → MER Instability
  - External Ingress → Noise Floor Rise
  - Load Spikes → CMTS Return Sensitivity
  - Weather → Aerial Plant Movement → Ingress Susceptibility

  The DAG does not decide what happened.
  It defines *what can* happen and *how causal chains behave*.

  The DAG provides the candidate causal pathways.

- **Layer 3: Structural Causal Model (SCM)**

  Using the KG (structure) and DAG (causal map), the SCM begins reasoning.

  It evaluates multiple hypotheses:

  1. **External ingress from aerial plant**
  2. **Thermal drift on legacy amplifier**
  3. **Power supply fluctuation on P1**
  4. **CM-generated noise**

5. **Return sensitivity issues on CMTS profile**
6. **Weather-induced movement causing intermittent shielding gaps**

The SCM assigns likelihood scores based on:

- observed MER patterns
- behaviors from the past week
- the exact modulation failures
- the time alignment with known patterns
- temperature and load at the moment of failure
- whether a cyclic pattern matches previous impairments

The SCM *disconfirms* incorrect paths.
For example:

- The pattern does not match CM-generated ingress.
- The modulation pattern does not align with CMTS return sensitivity issues.
- There is no recent maintenance on this leg that would cause misalignment.

The SCM converges on two likely causes:
thermal drift and **external ingress**.

- **Layer 4: State Model (Temporal Context)**

  The state model now checks:

  - Is this behaviour recurring?
  - Is it seasonal?
  - Does it correlate with time-of-day or temperature?
  - Has Power Supply P1 behaved inconsistently this month?

  The model finds:

  - A similar pattern occurred at 20:00 on two prior warm evenings.
  - The northern leg amplifier has a known drift signature when the temperature drops quickly at sunset.
  - The impairment aligns with a known cyclical pattern.

  This temporal dimension sharpens the reasoning:

  The most likely cause is thermal drift in the legacy amplifier on the northern leg.

- **Layer 5: Operator Intent Model (Interpretation and Action)**

  The system now frames the diagnosis according to operator intent.

  Operator intent includes:

  - prioritizing nodes with high business customers
  - reducing repeat impairments
  - escalating issues that match chronic behaviour
  - lowering tolerance for impacts during peak hours

  The intent model determines:

  - This is a **chronic impairment**, not a one-off anomaly
  - It occurs during peak evening usage
  - It is linked to customer-affecting symptoms (bufferbloat, video call instability, gaming latency)

  Therefore, the recommended classification is:

  Severity: Medium-High
  Action: Scheduled maintenance with near-term dispatch
  Interpretation: Expected recurrence within 24 hours if unaddressed
  Operator Note: Replace legacy amplifier prone to thermal instability

Final Output (What the System Reports)

Root Cause: Thermal gain drift on legacy amplifier (Northern leg, Node X1).
Confidence: 82 percent
Contributing Factors: Evening temperature drop, equipment age, historical drift pattern.
Customer Impact: High during peak hours, intermittent service degradation.
Recommended Action: Schedule amplifier replacement. Consider temporary gain stabilization adjustments if recurrence accelerates.

Why This Example Matters

This illustrates how:

- the **KG grounds** the event structurally,
- the **DAG maps** all plausible causal pathways,
- the **SCM reasons** through competing explanations,
- the **State Model contextualizes** behaviour in time,
- the **Operator Intent Model** determines appropriate action.

No LLM, RAG model, or statistical predictor could produce this sequence or explanation.

This is **operationally aligned AI in action.**

Why This Architecture Works and Others Do Not

Pattern-based AI models, including deep neural networks, language models, and retrieval-augmented systems, excel at pattern recognition but not at reasoning. They identify correlations without understanding structure, intent, or time. They lack causal grounding, cannot maintain a coherent internal state, and have no mechanism to verify whether their conclusions align with network physics or operator priorities.

The reasoning stack formed by the KG, DAG, SCM, state model, and intent model overcomes these limits. It mirrors how engineers diagnose faults: tracing structure, inferring cause, referencing history, and acting within policy. This architecture fuses physical, temporal, and ethical dimensions into a transparent reasoning system, an AI that does not just predict outcomes but understands the network it serves.

Summary

Operationally aligned AI demands model classes that embody structure, causality, time, and intent. Together, knowledge graphs, DAGs, SCMs, state models, and operator intent models form an intelligence layer that understands the network rather than merely observing it.

This architecture transforms AI from a passive analytic tool into an active reasoning framework; one that explains behaviour, respects operator context, and maintains the reliability expected of critical infrastructure. It provides the conceptual and technical foundation for transparent, verifiable, and operationally trustworthy AI in broadband networks.

Future Outlook – What the Industry Must Build Next

As broadband networks evolve, artificial intelligence will move from advisory analysis to active participation in operational control. The shift toward DOCSIS 4.0, deeper fiber reach, and virtualized CMTS architectures, combined with rising expectations for latency, reliability, and customer experience, will demand AI that interprets telemetry as fluidly as engineers do. Future networks will not only process more data; they will require reasoning systems capable of distinguishing physical cause from statistical coincidence.

Meeting that challenge demands a pivot in focus. The industry must advance beyond adoption and maturity metrics toward the deeper test of operational validity: how reliably an AI system interprets, explains, and acts within the physics of DOCSIS and HFC plants. Only through technical reliability, transparency, and governed execution can automation scale without compromising the stability of critical infrastructure.

AI That Reasons Within Network Physics

The next generation of operational AI must reason within the physics of the access network rather than around it. Systems built for broadband access network environments must encode the causal principles that engineers already use intuitively, how energy, noise, modulation, and scheduling interact across layers.

Such AI **MUST**:

- recognize impairment signatures documented in CableLabs PNM research
- model the causal links between modulation stability and OFDM profile behavior
- account for upstream noise and FEC performance as coupled phenomena
- interpret DOCSIS MAC scheduling as a determinant of throughput and latency; and
- reconcile DAA-specific telemetry divergence to preserve a unified view of plant health

Embedding these principles transforms AI from pattern observer to causal participant—an analytic system that operates inside network physics and sustains stability through understanding rather than approximation.

Industry-Wide Benchmark Datasets

A major obstacle to operationally aligned AI is the absence of standardized, multi-operator datasets that reflect the realities of broadband plant behavior. To evaluate reliability and generalization, the industry requires shared data encompassing:

- PHY-layer impairment signatures across modulation types.
- seasonal and environmental variation.
- topological and architectural diversity, including DAA and vCMTS perspectives; and
- cross-correlations between physical metrics and customer-experience indicators.

A federated benchmark of this kind would allow operators and vendors to test models against identical physical conditions, quantify robustness under drift, and establish common performance baselines. Such datasets are not simply a research convenience; they are the foundation of engineering reproducibility and trust in AI-driven operations.

Auditable AI Behavior for Critical Control Systems

As automation extends deeper into control systems, auditability becomes the defining safeguard between intelligence and automation risk. Operators must be able not only to view what an AI decided but understand why, which inputs, assumptions, and confidence levels shaped each conclusion.

Operationally aligned AI therefore requires:

- complete traceability of data flow and decision logic.
- visibility into every telemetry source contributing to an inference.
- quantified confidence bounds and explicit uncertainty disclosure.
- transparent causal error analysis rather than statistical post-mortems; and
- human-readable explanations that link outcomes to measurable network behavior.

These properties anchor AI within the principles of transparency and accountability defined by NIST and IEEE. In broadband operations, where decisions directly influence physical plant stability, traceability is not a compliance feature; it is the mechanism that preserves operator trust and prevents silent failure within automated loops.

Toward an Industry Standard for Operational AI Certification

The next step in operational maturity is an **industry certification framework** that evaluates AI not by marketing claims but by demonstrable performance under plant conditions. Such a framework must test whether an AI system can remain accurate, explainable, and stable when confronted with the realities of broadband access networks.

Certification should require:

- stress testing under impairment, drift, and visibility-divergence scenarios.
- validation using cross-layer and cross-architecture telemetry.
- seasonal and environmental reproducibility.
- explicit drift-resistance and retraining governance.
- architecture-aware benchmarking that includes DAA and vCMTS perspectives.
- explainability, auditability, and causal traceability standards; and
- alignment with TR-452 and related customer-experience metrics.

Such certification would finally separate AI systems engineered for operational reliability from those optimized for analytics or sales demonstration, establishing a measurable threshold for trust in AI that touches the control plane.

Ethical Obligations as Networks Become More Autonomous

As AI gains authority within operational workflows, ethics evolves from policy guidance into an engineering discipline. Every decision pathway, recommendation, configuration, or autonomous correction, carries potential impact on customers, plant stability, and public trust. Ethical design therefore demands that AI systems:

- prevent unintended or disproportionate harm to customer experience.
- preserve operator control and interpretability at all times.
- prioritize safety, reliability, and explainability over automation speed.
- operate within proven governance and rollback structures.
- defer to human expertise when uncertainty exceeds acceptable thresholds; and
- demonstrate consistent performance under degraded or high-stress conditions.

Responsible deployment is not risk-elimination but risk architecture, the deliberate engineering of transparency, predictability, and proportional safeguards into every layer of automation.

A Collaborative Path Forward

The evolution of AI in broadband operations will depend on deliberate collaboration across the ecosystem; operators, vendors, standards bodies, and research communities working from a shared technical foundation. Meaningful progress requires:

- common testing environments and federated data repositories
- standardized impairment, drift, and topology libraries
- formalized telemetry and data-schema definitions
- open, reproducible benchmarking frameworks
- unified performance metrics grounded in operational accuracy; and
- joint development of certification and governance standards.

CableLabs, SCTE, Broadband Forum, NIST, IEEE, and ETSI have already begun constructing the scaffolding for this collaboration. Extending that work to operational AI will ensure that automation evolves as a trusted extension of engineering judgment, transparent, resilient, and aligned with the physics and ethics of broadband systems.

Conclusion

Artificial intelligence can transform broadband network operations, but only when its design and evaluation remain aligned with the physical and architectural realities of access network systems. Current maturity frameworks emphasize organizational readiness, workflow integration, and governance posture, useful indicators of adoption, but not of operational safety or technical validity. None reveal whether AI truly understands the plant.

Broadband access networks are non-stationary systems shaped by incomplete telemetry, asymmetric visibility, and dynamic noise environments. Their impairment signatures differ fundamentally from those encountered in enterprise or cloud data domains. Without domain awareness, cross-layer correlation, and stress testing under real conditions, AI risks misclassifying impairment, misinterpreting upstream dynamics, initiating unnecessary corrections, or silently degrading customer experience. These risks grow as AI becomes embedded within closed-loop control and autonomous workflows.

This paper defined **operationally aligned AI** as a framework grounded in engineering discipline. It requires:

- multi-plane telemetry integration
- domain-aware interpretation of network behavior
- robustness validation under live plant conditions

- operator-aligned performance metrics
- explainability and interpretability
- lifecycle governance and drift management; and
- clearly defined operational safety boundaries.

Together, these principles form the criteria by which AI can be judged technically suitable for broadband operations.

The case studies in Section 6 demonstrated how domain-agnostic AI introduces predictable failure modes, while the blueprint in Section 7 established a repeatable lifecycle for safe deployment across DOCSIS, HFC, DAA, and CMTS architectures. Supported by standards from NIST, ETSI, IEEE, and the Broadband Forum, these models create a foundation for reliable, transparent automation rooted in engineering truth.

As the broadband industry advances toward DOCSIS 4.0, deeper virtualization, and greater automation, success will depend on systems that reason within network physics, operate with full auditability, and meet standardized certification for operational reliability.

Broadband networks are critical infrastructure. Ensuring that AI functions safely and ethically within them is not merely a technical requirement, it is a moral imperative. By shifting focus from maturity scoring to operational correctness, the industry can realize AI that strengthens reliability, supports engineers, and enhances customer experience while preserving the stability and integrity of the access network.

Abbreviations and Definitions

Abbreviations

AI	artificial intelligence
ANL	Autonomous Networks Levels (TM Forum)
CM	cable modem
CM-SP-PHYv3.1	CableLabs specification identifier for DOCSIS 3.1 PHY (CM-SP-PHYv3.1)
CMTS	Cable Modem Termination System
CPD	common path distortion
CPE	customer premises equipment
CX	customer experience
DAA	distributed access architecture
DAG	Directed Acyclic Graph
DOCSIS®	Data Over Cable Service Interface Specification
DRW	dynamic range window

ETSI	European Telecommunications Standards Institute
FEC	forward error correction
FN	false negative
FP	false positive
GS	Group Specification (ETSI document type, e.g., GS ZSM 002)
GSMA	GSM Association
HFC	hybrid fiber coax
IEEE	Institute of Electrical and Electronics Engineers
IT	information technology
KG	knowledge graph
KPI	key performance indicator
LLD	Low-Latency DOCSIS
LLM	large language model (mentioned as language models)
MAC	media access control
MAP	MAP message (DOCSIS upstream scheduling; not expanded in draft)
MER	modulation error ratio
ML	machine learning
NIST	National Institute of Standards and Technology
OFDMA	orthogonal frequency-division multiple access
OFDM	orthogonal frequency-division multiplexing
PHY	physical layer
PNM	proactive network maintenance
QoE	quality of experience
RAG	retrieval-augmented generation
RF	radio frequency
RMF	Risk Management Framework (as in NIST AI RMF)
RPD	remote PHY Device
SCM	structural causal model
SCTE	Society of Cable Telecommunications Engineers
SMP	service-management plan
SNR	signal-to-noise ratio
TR-452	Broadband Forum Technical Report 452 (Performance Measurement and Application Layer Testing)

TM	TM Forum (TeleManagement Forum; the industry body)
vCMTS	virtual Cable Modem Termination System
ZSM	Zero-touch Network and Service Management (ETSI)

Definitions

action throttling	Rate limits/cooldowns/confidence thresholds to prevent oscillation or cascading changes from overcorrection.
analytical AI	AI used for insights/analysis that does not directly alter network configuration or maintenance decisions (contrasted with operational AI).
cross-layer visibility	Use of PHY, MAC, and customer-experience telemetry together so the model can distinguish impairment vs contention and understand "why" behavior occurs.
data survivability	Ability to remain functional when telemetry is missing/gapped; includes fallback logic, confidence weighting, and treating missing data as a signal.
Directed Acyclic Graph (DAG)	Model component encoding directional causal pathways (cause precedes effect) to prevent mistaking correlation for causation.
domain-integrated system understanding	Model understanding that incorporates DOCSIS/HFC/DAA behavior (RF physics, scheduling, impairment signatures) rather than treating telemetry as generic data.
explainability / operator interpretability	The operator can trace inputs, reasoning path, uncertainty, and verify outputs against engineering indicators; includes ability to override safely.
False-Positive / False-Negative Operational Risk Matrix	A decision aid that maps FP/FN tendencies to operational risk and helps decide whether a model is safe for automation, advisory use, or not deployable.
human-IN-the-loop	Operator authorization required for high-impact actions that can alter plant state or customer experience (e.g., profile changes, dispatch).
human-ON-the-loop	AI may act within defined thresholds for medium/low-risk actions, but actions remain observable, interruptible, and reversible under operator supervision.
Knowledge Graph (KG)	Model component encoding the network's structural/topological relationships (assets, dependencies, lineage) to ground reasoning in physical context.

lifecycle governance / drift management	Controls for versioning, change tracking, drift detection, retraining validation, rollback, and guardrails so models remain reliable over time.
no-go zones	Operational domains explicitly restricted from autonomous AI control due to high risk or incomplete observability (require human authorization and cross-layer verification).
operational AI	Any system that interprets multi-layer telemetry and produces recommendations or actions that affect plant behavior, service quality, or customer experience.
operationally aligned AI	AI built to reason within broadband network physics and architecture (cause-and-effect), not just detect correlations/patterns.
operator intent model	Component that encodes operational priorities/policies and risk tolerances so identical technical events can be acted on differently depending on context.
operator-aligned success metrics	Evaluation based on operational impact of errors (e.g., false positives vs false negatives) rather than ML-only metrics like accuracy/F1.
robustness under real plant conditions	Demonstrated accuracy under live-network stressors (noise bursts, drift, partial-service states, timing asymmetry, telemetry divergence), not only in lab/simulated data.
safe state	A state where no autonomous control-plane actions can alter network behavior without operator confirmation.
shadow mode	New models run in parallel first; outputs are logged and compared to real outcomes before the model is allowed to recommend/act operationally.
state model	Temporal component that maintains system memory/baselines, separates transient events from persistent drift, and provides time context for inference.
structural causal model (SCM)	Reasoning core that performs causal inference and hypothesis testing over the causal structure, including counterfactual reasoning and uncertainty handling.

Bibliography and References

1. CableLabs, DOCSIS 3.1 Physical Layer Specification (CM-SP-PHYv3.1). Louisville, CO: CableLabs, 2019.

2. CableLabs, DOCSIS 4.0 Physical Layer Specification. Louisville, CO: CableLabs, 2020.

3. CableLabs, Distributed Access Architecture (DAA) Specification. Louisville, CO: CableLabs, 2018.
4. CableLabs, Proactive Network Maintenance (PNM): Use Cases and Architecture. Louisville, CO: CableLabs, 2020.
5. CableLabs, DOCSIS Profile Management Application: Technical Guidance. Louisville, CO: CableLabs, 2020.
6. CableLabs, DOCSIS 4.0 Technology Overview and Roadmap. Louisville, CO: CableLabs, 2021.
7. SCTE, Operational Practices for HFC Networks. Exton, PA: Society of Cable Telecommunications Engineers, 2019.
8. Broadband Forum, TR-452: Performance Measurement and Application Layer Testing. Broadband Forum, 2021.
9. TM Forum, AI Maturity Model Toolkit. TM Forum, 2021.
10. TM Forum, Autonomous Networks Framework (ANL): Levels and Architecture. TM Forum, 2022.
11. ETSI, Zero-Touch Network and Service Management (ZSM) Architecture (GS ZSM 002). Sophia Antipolis: ETSI, 2020.
12. National Institute of Standards and Technology (NIST), Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg, MD: NIST, 2023.
13. IEEE Standards Association, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. IEEE, 2019.
14. GSMA, AI Ethics in Telecommunications: Guidelines and Principles. GSMA, 2021.
15. D. Amodei et al., “Concrete Problems in AI Safety,” arXiv preprint arXiv:1606.06565, 2016.
16. Cisco Systems, Upstream Noise in HFC Networks: Causes and Mitigation. San Jose, CA: Cisco Systems, 2018.

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Implementation Strategies for Private Secure Networks

A technical paper prepared for the SCTE Technical Journal

A Technical Paper

Rohith Kumar Punithavel, Senior Software Engineer, Emerging Technology, Charter Communications
6399 S Fiddler’s Green Circle
Greenwood Village, CO 80111
RohithKumar.Punithavel@charter.com

Vishal Chopade, Senior Software Engineer, Emerging Technology, Charter Communications
6399 S Fiddler’s Green Circle
Greenwood Village, CO 80111
Vishal.Chopade@charter.com

Sagar Panchal, Network Engineer IV, Emerging Technology, Charter Communications
6399 S Fiddler’s Green Circle
Greenwood Village, CO 80111
Sagar.Panchal@charter.com

Table of Contents

List of Figures

List of Tables

Introduction

Data has become one of the most valuable assets, and its effective utilization plays a crucial role in decision-making, growth, innovation, profitability, and competitive advantage. With advancements in communication networks, enterprises can efficiently and seamlessly transmit and receive data. As data transactions and their criticality continue to rise, the demand for secure and reliable networks has grown among enterprise customers who need to transmit confidential and sensitive information securely.

This paper explores the implementation of Network-as-a-Service (NaaS) strategies to establish private and secure networks between wireless access points and enterprise customers. At Charter Communications, we conducted an end-to-end study in a controlled lab environment utilizing Generic Routing Encapsulation (GRE), Internet Protocol Security (IPSec), and Virtual Private Network (VPN) technologies, along with Application Programming Interfaces (APIs) for remote interaction with components in the architecture. The lab demonstration covered key aspects such as authentication, authorization, accounting, customer premises equipment (CPE) management, traffic segregation, and tunnel management with the help of APIs. The paper concludes with insights into challenges and best practices that contribute to the successful deployment of secure private networks.

Overview of Private Networks

A private network is a secure, isolated system used in businesses, homes or cloud environments to safeguard sensitive information by restricting access to authorized devices and users. In a private network, a single physical network logically segments into multiple separate virtual networks, each functioning independently. Compared with open public access networks, private networks are designed to reduce exposure to external threats through controlled ingress, segmentation, and monitoring.

Evolution and Necessity of Private Networks

The concept of private networks dates back to the telegraph era when dedicated physical circuits ensured privacy through complete physical separation. Today, private networks have evolved into sophisticated, virtualized infrastructures, offering greater flexibility while maintaining strong security controls. As networking technologies matured, the industry shifted from circuit-switched to packet switched models, introducing the concept of shared infrastructure. In packet switched networks, data is divided into packets and transmitted dynamically across shared paths. This evolution dramatically improved scalability and resource utilization but introduced inherent challenges in maintaining the strict isolation and control that dedicated lines once provided. Multiple virtual network segments can be created in a private network, allowing independent operation. Each virtual network segment can be tailored to meet the unique requirements of a specific service, application, or enterprise, offering custom bandwidth and security policies. This innovation allows organizations to enjoy the benefits of a dedicated private network, such as guaranteed quality of service (QoS), traffic isolation, and enhanced security, without the complexity and cost of deploying separate physical infrastructure [5]. Virtual partitioning has become essential in delivering performance predictability and compliance in mission-critical operations, especially as demands for real-time data processing and secure connectivity increase. Virtual local area networks (VLANs) have long empowered network administrators to organize devices logically, regardless of physical location. By isolating network traffic, VLANs enhance security and manageability through simplified segmentation. Similarly, virtual private networks (VPNs) create secure, encrypted connections over public networks, enabling users to remotely access enterprise systems safely. VLANs and VPNs demonstrate how logically distinct traffic domains can operate securely over shared physical infrastructure, delivering the flexibility and protection required in modern networking environments.

As industries increasingly adopt technologies like the Internet of things (IoT), automation, and real-time analytics, the volume and sensitivity of data generated continue to grow. This trend underscores the rising importance of private networks, which provide secure, high-performance connectivity for critical and sensitive operations. Isolated, encrypted environments serve as the backbone for safeguarding this data. Complete control over traffic flow, authentication, and encryption is no longer optional, but a core requirement. Through virtual segmentation and software-defined policy enforcement, private networks are transforming into essential building blocks of today's digital ecosystems. They combine the security of dedicated lines with the scalability and efficiency of shared systems.

Use cases like smart grids, industrial automation, and remote healthcare monitoring depend on uninterrupted, reliable, and secure connectivity, free from public internet exposure. This is especially true in the example of remote monitoring of medical devices, where hospitals rely on connected systems to track patient vitals after major procedures. These devices generate highly confidential health data that must be securely and instantly transmitted to medical teams for timely analysis and intervention. Private networks ensure this data remains protected from general internet traffic, reducing the risks of data breaches, unauthorized access, and service disruption. Moreover, private networks offer advanced control over data routing, user authentication, and policy enforcement, helping organizations meet stringent regulatory and operational requirements [3].

Vulnerabilities in End-to-End Data Transfer

Data vulnerability during network transmission remains a critical concern in modern digital infrastructures, particularly as reliance on connected devices and distributed systems grows. A primary challenge arises from using public or shared networks, where data may be exposed to interception, tampering, or unauthorized access due to insufficient isolation. Weak authentication mechanisms, such as reliance on user-generated credentials, further exacerbate the risk of creating exploitable entry points for malicious actors. Additionally, legacy systems with inconsistent or outdated encryption protocols undermine the confidentiality and integrity of transmitted data, making them susceptible to compromise. Human error in manual configuration processes also contributes to security gaps, often introducing misconfigurations that expose systems to attack. These vulnerabilities collectively highlight the need for comprehensive security architectures incorporating encrypted, policy-based routing, automated authentication tied to verified device identities, and standardized encryption protocols enforced across all communication layers. The private network implementations in this paper leverage network-level authentication, automated routing, and end-to-end encryption to reduce the risk of data breaches and system compromise. This architecture offers a significantly more secure and controlled environment than conventional solutions.

Network as a Service using API

Application programming interfaces (APIs) are functions or interfaces that accept defined requests, enabling seamless exchanges of functionality between software entities. In today's digital environment, APIs are fundamental building blocks of any digital application, allowing applications to access data securely and reliably. Now, API traffic accounts for most of the overall internet traffic on the network [8]. One of the advantages of using an API is that it gives access to data from various independent services or sources. APIs have adopted event-driven architectures, aligning with their shift toward being cloud-native. APIs are crucial in building interoperable and scalable systems by abstracting underlying complexity. With growing security and privacy concerns, the APIs are built with authentication and authorization, ensuring resilience in a dynamic web ecosystem.

Another benefit of APIs is related to monetization. Developers and organizations can expose APIs and charge their public users based on usage. One such use case is when a network operator exposes network capabilities such as location services, monitoring, or management programmatically using APIs that are available for external consumption, referred to as network as a service (NaaS). NaaS transforms telco networks into programmable service platforms, allowing network integration with third-party applications and facilitating direct and open interactions. NaaS APIs from operators offer controllability and auditability when exposing capabilities to third-party applications. Controllability will enable providers to enforce policies and regulate access through mechanisms like Role-Based Access Control (RBAC), while auditability ensures all interactions are logged for traceability and non-repudiation. From the customer's perspective, NaaS delivers a consumer-like experience: it enables third-party developers and enterprises to access and integrate network capabilities on demand via APIs, offering greater flexibility, scalability, visibility, and control to support next-generation digital services [4].

Private Secure Network Concepts Overview

Router and OpenWrt

A router is a packet-switched networking device that forwards traffic between networks (for example, between a local area network (LAN) and a wide area network (WAN)). It selects a next hop based on routing information (such as a routing table) and packet headers, enabling devices on one network to reach destinations on another. In many home deployments, an ISP-managed gateway combines routing, firewalling, DHCP/DNS services, and Wi-Fi access point functionality to provide shared internet access to multiple devices. Wireless routers have service set identifiers (SSIDs). These enable users to find and connect to the wireless network broadcasted by the router [2].

OpenWrt is an open-source, Linux-based operating system for embedded devices, particularly routers. It provides a writable file system and a package ecosystem that enables substantially more customization than typical vendor firmware. In this work, OpenWrt is used to implement VLANs/bridging, firewall policy, VPN functions, and policy-based routing (PBR). These capabilities allow a managed access point to support both residential and enterprise connectivity while preserving logical isolation. OpenWrt transforms a standard router into a highly configurable networking device [7].

Policy Based Traffic Segregation

PBR is a powerful networking technique that enables traffic routes based on a wide range of attributes beyond the traditional destination IP address. These attributes may include source IP, MAC address, device type, user identity, or traffic type. By leveraging PBR, network administrators can implement fine-grained control over how different types of traffic are handled and routed. In enterprise-connected environments, specific devices generate high-priority or sensitive data that must be treated differently from general internet traffic.

Traditional routing methods, which apply uniform rules to all traffic, are often inadequate in such scenarios. PBR allows for the creation of custom routing rules that direct enterprise-bound traffic through secure, policy-enforced paths, often via encrypted tunnels. To further support this separation, network segmentation mechanisms such as logical identifiers, tagging protocols, or traffic classification systems can mark enterprise traffic at the access layer. When combined with PBR, these mechanisms ensure that sensitive traffic is logically and operationally isolated from unmanaged or personal traffic, such as that generated by household devices like smartphones, tablets, or smart home appliances. This layered approach is essential for maintaining security, performance consistency, and regulatory compliance in multi-tenant or mixed-use environments. It enables service providers and enterprises to enforce differentiated handling policies, ensuring that enterprise traffic remains distinct, protected, and appropriately routed throughout the network.

Authentication and Authorization for Network Access

The ISP-managed gateway offers a service through a second SSID that offloads data traffic onto Wi-Fi for mobile customers. It is widely available and can be used by third-party organizations. The primary purpose of the second SSID is to ensure consistent access to this feature from any location within the ISP's service coverage area. A mechanism is required to authenticate customers using third-party enterprise services and devices to enable this access securely.

One practical approach uses AAA (Authentication, Authorization, and Accounting) systems. AAA is built on RADIUS (Remote Authentication Dial-In User Service) protocol. RADIUS is a network protocol that defines rules and conventions for communication between network devices. It provides centralized Authentication, Authorization, and Accounting for users to connect and access the network. Authentication is a process that validates a user's identity by matching their credentials with those on the server. If the credentials match, the user gains access to the network. Authorization is the process of granting or denying permissions based on information sent by the Network Address System (NAS). The RADIUS server manages the authorization policy while the NAS enforces it. Accounting records user resources consumed during network sessions, including system time, data sent, and received. The RADIUS session process involves a remote user connecting to a Network Access Server (NAS) device, initiating login, and initiating NAS conversations. The server communicates with the NAS via a shared secret mechanism, using the UDP protocol for authentication and accounting. The NAS sends an Access-Request containing user information, authentication credentials, and requested service to the RADIUS server, which is then processed and verified against network authentication services. Validation results are sent to the NAS in three forms: Access-Reject, Access-Challenge, or Access-Accept. The RADIUS client initiates the accounting process by sending an Accounting-Request Start and Stop message. The data from accounting sessions, including billable information and reports, are stored.

In modern architectures, RADIUS requests are proxied from one server to another to manage AAA for various services, including third-party enterprise services. This architecture allows network operators to delegate authentication responsibilities while maintaining centralized policy enforcement and accounting. Third-party enterprise customers utilize AAA servers to enforce device-related attributes, certificates, customer tiers, or user roles. RADSEC, defined as RADIUS over TLS, is employed to enhance the security of RADIUS protocol. RADSEC encrypts RADIUS traffic, safeguarding sensitive information like credentials and attributes from cyberattacks or interception [6].

Network Tunnels

Network tunnels are a technique utilized in networking to encapsulate one network protocol within another. One of the most used tunneling protocols is Generic Routing Encapsulation (GRE) tunnels, which allow the transmission of diverse network layer protocols across point-to-point links. By creating a virtual internet connection, GRE tunnels facilitate direct communication between geographically separated networks. By encapsulating original data packets inside new IP packets, GRE enables smooth data exchange between remote networks, making them as if they are locally connected. VPN tunnels are commonly used for secure communications. VPN tunnels add cryptographic protection (confidentiality and integrity) in addition to encapsulation.

Centralized API Controller

APIs act as a bridge connecting customer equipment and provisioned services, but raise two important questions: How does a client know which endpoints to interact with? How do you control usage and access to data in the APIs? This is where a centralized management tool called an API gateway emerges. An API gateway authenticates API requests from clients, processes them based on policies, and routes them to appropriate services. It acts as an entryway to access backend systems or data. API gateways can be deployed at the network edge, on-premises, or in the cloud. Network providers use APIs to manage customer premises equipment and its capabilities remotely. They commonly implement capabilities such as security policy (e.g., authentication, authorization, access control), routing policy (e.g., load balancing, health checks, error handling, routing, rate limiting), and observability policy (e.g., logging, tracing, metrics).

Since most internet traffic now consists of API traffic, the complexity and volume of APIs have increased, making a centralized controller necessary for effective, reliable, and secure management. A centralized API gateway acts as a governance layer to enforce rules, regulations, and compliance. Because it offers a single point of control, businesses can implement consistent access rules, monitor usage patterns, identify irregularities, and implement changes throughout the API ecosystem. Each service may use its authentication, throttling, or logging strategy without centralized control, resulting in fragmentation, uneven user experiences, and security flaws. API gateways abstract functions and transform endpoints into manageable, secure, scalable, and observable digital platforms for simplicity [1].

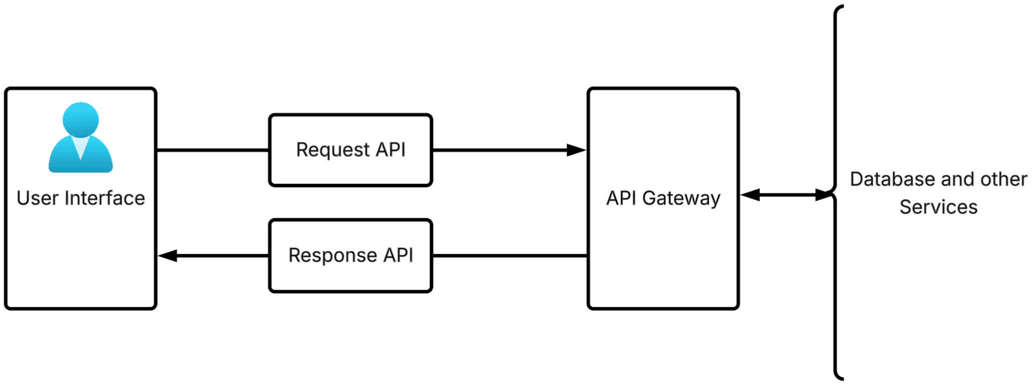


Figure 16 Working of API and API Gateway

Implementation 1: End-to-End VPN

Architecture Description

The setup is logically divided into three segments: Enterprise, ISP Network, and Home. Within this setup, the Internet Service Provider (ISP) manages the Wi-Fi access point located in the home segment. Management is facilitated through well-defined APIs, enabling external control and configuration. The ISP establishes an agreement with the enterprise to support enterprise use cases, granting them authorized access to these APIs. This collaboration encompasses a structured, multi-party coordination process involving the following key steps:

- Device Onboarding Request: Enterprise devices initiate a provisioning request directed to the ISP network administration. This request typically includes ISP and enterprise customer account associations and authorization information necessary for access.
- Enterprise-ISP Coordination: The enterprise ensures a functional VPN server. Upon receiving the onboarding request, the enterprise uses APIs to use the Wi-Fi access points at customer homes. This interaction facilitates the establishment of secure connectivity between home and enterprise resources.

The ISP will facilitate home Wi-Fi access points to communicate with the authentication server. Also, the ISP will be responsible for managing authentication flow, API feature enhancement, and providing API access to the enterprise. At home, the Wi-Fi access point runs on OpenWrt firmware. The firmware is configured to support dual wireless access through network segmentation. This configuration enables a hybrid network environment where residential users maintain standard internet access while enterprise devices can securely connect to corporate networks through the same physical infrastructure. This orchestrated process ensures controlled and secure integration of enterprise devices into home-based ISP infrastructure while maintaining role-based access, accountability, and policy enforcement.

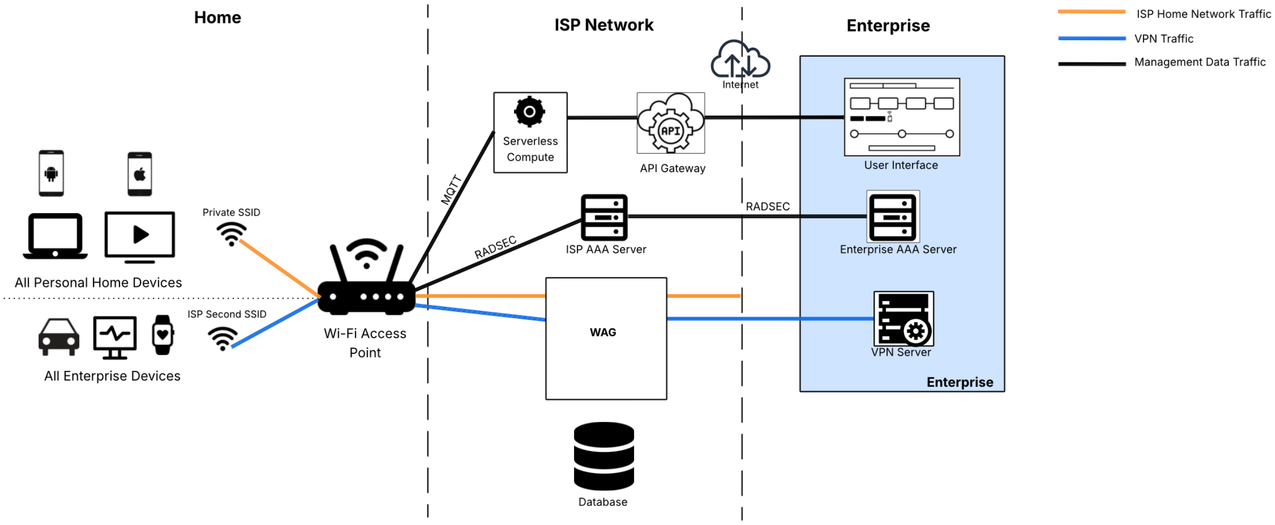


Figure 17 Implementation 1: End-to-End VPN Architecture

The architecture diagram illustrates the end-to-end architecture for the implementation, divided into three segments: Enterprise, ISP Network, and Home.

In the home segment, the Wi-Fi Access Point broadcasts two SSIDs: one dedicated to home internet customer personal use and another for enterprise connectivity. The Wi-Fi access point runs a custom software module that receives MQTT messages and executes corresponding actions. These MQTT messages are triggered by API calls authorized by the API gateway managed by the ISP. The API calls are initiated through APIs accessible to the enterprise via the user interface. The Wi-Fi Access Points are also integrated with the ISP's AAA servers. The ISP's AAA servers authenticate devices connected via the second SSID. These AAA servers in the ISP network can function as a primary authentication authority or proxy, forwarding requests to the enterprise-owned AAA. The AAA supports various authentication methods like EAP-TLS, EAP-TTLS, OCSP, and other standard EAP-based protocols, ensuring flexible and secure device authentication. The AAA also maintains the accounting data for this solution, which the ISP uses to bill the enterprise. The end-to-end AAA connectivity between the Home and ISP network and between ISP and enterprise networks are secured using RADSEC (RADIUS over TLS), which ensures encrypted, authenticated, and integrity-protected communication.

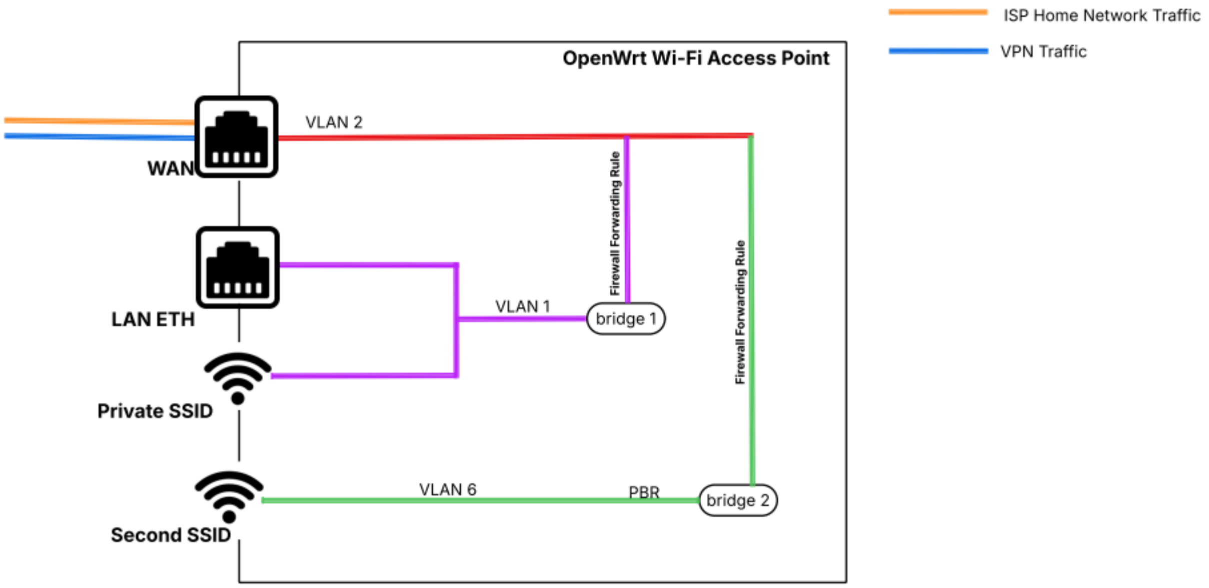


Figure 18 Configuration Inside an OpenWrt Router

The OpenWrt system implements VLAN-based network isolation for WAN and LAN. In the current system, WAN is assigned as VLAN-2 and LAN home network as VLAN-1. On the Wi-Fi Access Point, the second SSID traffic is segregated using a dedicated VLAN-6, which is assigned a separate bridge and subnet. The LAN network has two separate subnets for VLAN-1 and VLAN-6, respectively. The created VPN interface is associated with this bridge to route traffic securely through the VPN tunnel using the WireGuard protocol.

Appropriate firewall zones and rules are configured to allow forwarding to the WAN and provide DHCP, DNS, and other network services for devices within VLAN-6. In addition, PBR is implemented to route enterprise traffic appropriately. This approach effectively separates enterprise traffic from regular home traffic. MAC-based policy routing can be applied to support various enterprise clients. This evolution allows for flexible, fine-grained traffic control, ensuring secure enterprise connectivity alongside traditional home network usage without compromising logical separation or security.

Network as a Service APIs

In this private network implementation, we have developed and deployed twelve well-defined APIs to manage and control the entire solution, from access management to resource provisioning. Two of these APIs are dedicated solely to enterprise–ISP communications, handling tasks such as onboarding and offboarding enterprise customers and devices. The remaining ten APIs are exposed to the enterprise as service-provisioning interfaces, yet their actions ultimately take effect on Wi-Fi access points. We ensure a secure and modular architecture by isolating enterprise-ISP and ISP-access point APIs. The ten APIs are to:

- check device health,
- check VPN status,
- create a second SSID,
- apply configuration policies (VLAN, bridge, firewall rules, VPN interface, PBR),
- create VLAN,
- one-click construction of end-to-end configuration,
- one-click deconstruction of end-to-end configuration,
- delete VLAN,
- remove configuration policies (VLAN, bridge, firewall rules, VPN interface, PBR), and
- delete second SSID.

These APIs facilitate automated coordination between enterprise systems and ISP infrastructure, enabling dynamic end-to-end process creation and management.

Sequence Diagram

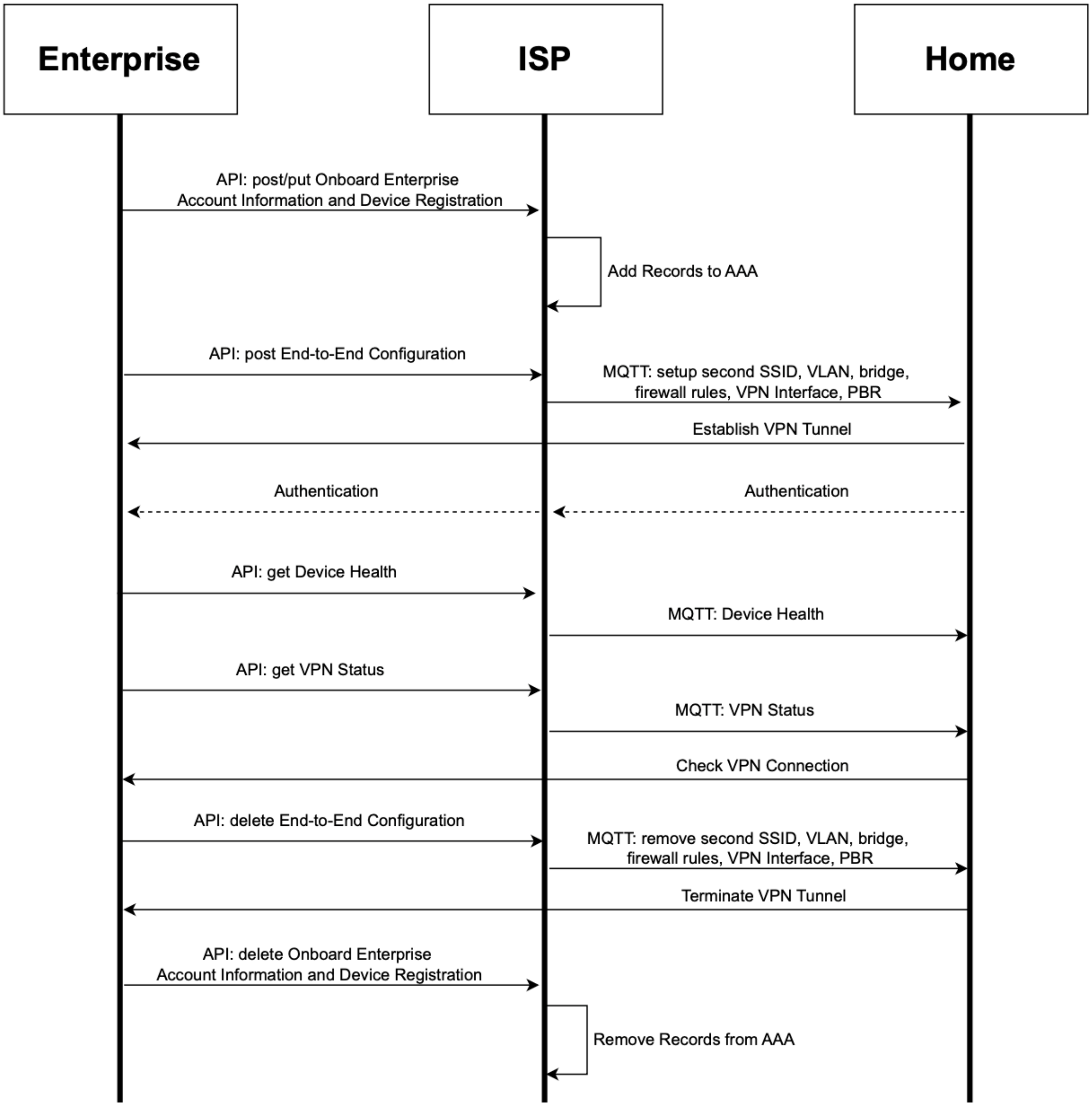

Enterprise
ISP
Home
API: post/put Onboard Enterprise Account Information and Device Registration
Add Records to AAA
API: post End-to-End Configuration
MQTT: setup second SSID, VLAN, bridge, firewall rules, VPN Interface, PBR
Establish VPN Tunnel
Authentication
Authentication
API: get Device Health
MQTT: Device Health
API: get VPN Status
MQTT: VPN Status
Check VPN Connection
API: delete End-to-End Configuration
MQTT: remove second SSID, VLAN, bridge, firewall rules, VPN Interface, PBR
Terminate VPN Tunnel
API: delete Onboard Enterprise Account Information and Device Registration
Remove Records from AAA

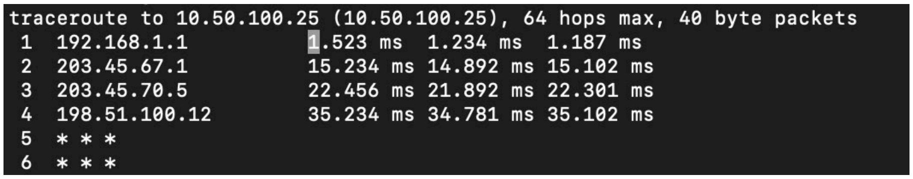
Figure 19 Implementation 1: End-to-End VPN Sequence Diagram

Lab Test Results

Traceroute analysis is a diagnostic method used to understand the network path of data packets between source and destination. This test demonstrates the end-to-end behavior of Implementation 1. The enterprise resource server IP is 10.50.100.25 and VPN tunnel is 100.200.1.1.

Test case 1: Enterprise Connectivity via Private SSID

The device connected to private SSID attempting to reach the enterprise resource at 10.50.100.25.



```
traceroute to 10.50.100.25 (10.50.100.25), 64 hops max, 40 byte packets
 1  192.168.1.1          1.523 ms  1.234 ms  1.187 ms
 2  203.45.67.1          15.234 ms 14.892 ms 15.102 ms
 3  203.45.70.5          22.456 ms 21.892 ms 22.301 ms
 4  198.51.100.12        35.234 ms 34.781 ms 35.102 ms
 5  * * *
 6  * * *
```

Figure 20 Traceroute Results for Test Case 1

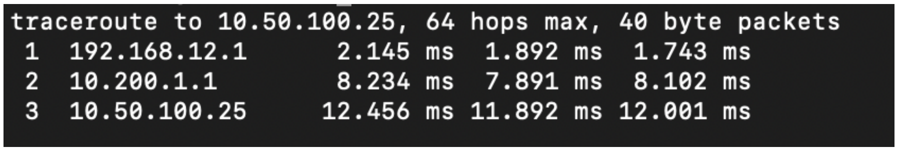
Table 3 Traceroute Information for Test Case 1

IP	Description
192.168.1.1	router-vlan1
203.45.67.1	ISP-Gateway
203.45.70.5	Next Hop
198.51.100.12	Next Hop

The user device connected to private SSID fails to reach the enterprise resource. The traceroute shows packets traveling through the standard internet path (ISP gateway at 203.45.67.1, backbone routers) but timing out when attempting to reach the private enterprise network.

Test Case 2: Enterprise Connectivity via Second SSID

The device connected to second SSID attempting to reach the enterprise resource at 10.50.100.25.



```
traceroute to 10.50.100.25, 64 hops max, 40 byte packets
 1  192.168.12.1       2.145 ms  1.892 ms  1.743 ms
 2  10.200.1.1         8.234 ms  7.891 ms  8.102 ms
 3  10.50.100.25      12.456 ms 11.892 ms 12.001 ms
```

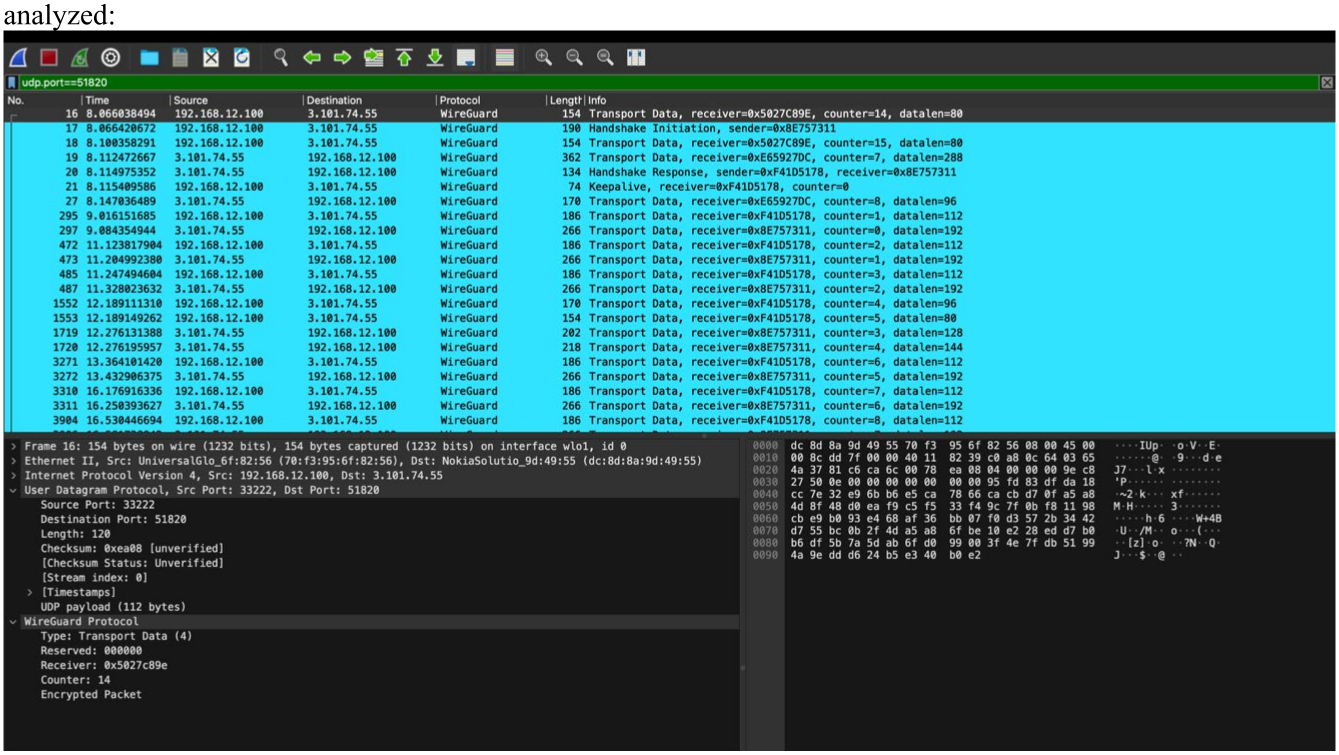
Figure 21 Traceroute Results for Test Case 2

Table 4 Traceroute Information for Test Case 2

IP	Description
192.168.12.1	router-vlan6
10.200.1.1	wireguard-tunnel
10.50.100.25	enterprise-server

The enterprise device connected to the second SSID successfully establishes connectivity through the WireGuard tunnel. The traceroute shows a direct path from the VLAN-6 gateway (192.168.12.1) through the VPN tunnel endpoint (10.200.1.1) to the enterprise server (10.50.100.25). The encrypted connection

and consistent response times reflect a secure and efficient performance. The traceroute results demonstrate the effectiveness of network isolation and VPN connectivity in the dual-SSID configuration.
Further, network packets are captured between the enterprise device and enterprise resources are analyzed:



```
udp.port==51820

No.   Time          Source           Destination      Protocol    Length Info
   16 8.066038494   192.168.12.100   3.101.74.55      WireGuard   154 Transport Data, receiver=0x5027C89E, counter=14, datalen=80
   17 8.066420672   192.168.12.100   3.101.74.55      WireGuard   190 Handshake Initiation, sender=0x8E757311
   18 8.100358291   192.168.12.100   3.101.74.55      WireGuard   154 Transport Data, receiver=0x5027C89E, counter=15, datalen=80
   19 8.112472667   3.101.74.55      192.168.12.100   WireGuard   362 Transport Data, receiver=0xE65927DC, counter=7, datalen=288
   20 8.114975352   3.101.74.55      192.168.12.100   WireGuard   134 Handshake Response, sender=0xF41D5178, receiver=0x8E757311
   21 8.115409586   192.168.12.100   3.101.74.55      WireGuard    74 Keepalive, receiver=0xF41D5178, counter=0
   27 8.147036489   3.101.74.55      192.168.12.100   WireGuard   170 Transport Data, receiver=0xE65927DC, counter=8, datalen=96
  295 9.016151685   192.168.12.100   3.101.74.55      WireGuard   186 Transport Data, receiver=0xF41D5178, counter=1, datalen=112
  297 9.084354944   3.101.74.55      192.168.12.100   WireGuard   266 Transport Data, receiver=0x8E757311, counter=0, datalen=192
  472 11.123817904  192.168.12.100   3.101.74.55      WireGuard   186 Transport Data, receiver=0xF41D5178, counter=2, datalen=112
  473 11.204992380  3.101.74.55      192.168.12.100   WireGuard   266 Transport Data, receiver=0x8E757311, counter=1, datalen=192
  485 11.247494604  192.168.12.100   3.101.74.55      WireGuard   186 Transport Data, receiver=0xF41D5178, counter=3, datalen=112
  487 11.328023632  3.101.74.55      192.168.12.100   WireGuard   266 Transport Data, receiver=0x8E757311, counter=2, datalen=192
 1552 12.189111310  192.168.12.100   3.101.74.55      WireGuard   170 Transport Data, receiver=0xF41D5178, counter=4, datalen=96
 1553 12.189149262  192.168.12.100   3.101.74.55      WireGuard   154 Transport Data, receiver=0xF41D5178, counter=5, datalen=80
 1719 12.276131388  3.101.74.55      192.168.12.100   WireGuard   202 Transport Data, receiver=0x8E757311, counter=3, datalen=128
 1720 12.276195957  3.101.74.55      192.168.12.100   WireGuard   218 Transport Data, receiver=0x8E757311, counter=4, datalen=144
 3271 13.364101420  192.168.12.100   3.101.74.55      WireGuard   186 Transport Data, receiver=0xF41D5178, counter=6, datalen=112
 3272 13.432906375  3.101.74.55      192.168.12.100   WireGuard   266 Transport Data, receiver=0x8E757311, counter=5, datalen=192
 3310 16.176916336  192.168.12.100   3.101.74.55      WireGuard   186 Transport Data, receiver=0xF41D5178, counter=7, datalen=112
 3311 16.250393627  3.101.74.55      192.168.12.100   WireGuard   266 Transport Data, receiver=0x8E757311, counter=6, datalen=192
 3904 16.530446694  192.168.12.100   3.101.74.55      WireGuard   186 Transport Data, receiver=0xF41D5178, counter=8, datalen=112
```

```
> Frame 16: 154 bytes on wire (1232 bits), 154 bytes captured (1232 bits) on interface wlo1, id 0
> Ethernet II, Src: UniversalGlo_6f:82:56 (70:f3:95:6f:82:56), Dst: NokiaSolutio_9d:49:55 (dc:8d:8a:9d:49:55)
> Internet Protocol Version 4, Src: 192.168.12.100, Dst: 3.101.74.55
v User Datagram Protocol, Src Port: 33222, Dst Port: 51820
    Source Port: 33222
    Destination Port: 51820
    Length: 120
    Checksum: 0xea08 [unverified]
    [Checksum Status: Unverified]
    [Stream index: 0]
  > [Timestamps]
    UDP payload (112 bytes)
v WireGuard Protocol
    Type: Transport Data (4)
    Reserved: 000000
    Receiver: 0x5027c89e
    Counter: 14
    Encrypted Packet
```

```
0000  dc 8d 8a 9d 49 55 70 f3  95 6f 82 56 08 00 45 00   ····IUp· ·o·V··E·
0010  00 8c dd 7f 00 00 40 11  82 39 c0 a8 0c 64 03 65   ······@· ·9···d·e
0020  4a 37 81 c6 ca 6c 00 78  ea 08 04 00 00 00 9e c8   J7···l·x ········
0030  27 50 0e 00 00 00 00 00  00 00 95 fd 83 df da 18   'P······ ········
0040  cc 7e 32 e9 6b b6 e5 ca  78 66 ca cb d7 0f a5 a8   ·~2·k··· xf······
0050  4d 8f 48 d0 ea f9 c5 f5  33 f4 9c 7f 0b f8 11 98   M·H····· 3·······
0060  cb e9 b0 93 e4 68 af 36  bb 07 f0 d3 57 2b 34 42   ·····h·6 ····W+4B
0070  d7 55 bc 0b 2f 4d a5 a8  6f be 10 e2 28 ed d7 b0   ·U··/M·· o···(···
0080  b6 df 5b 7a 5d ab 6f d0  99 00 3f 4e 7f db 51 99   ··[z]·o· ··?N··Q·
0090  4a 9e dd d6 24 b5 e3 40  b0 e2                     J···$··@ ··
```

Figure 22 Packet Analysis for Implementation 1: End-to-End VPN

In the packet capture, the 3.101.74.55 is a public IP of enterprise VPN endpoint. From further analysis we can see that the packets are encrypted and use the WireGuard Protocol. The Source/Destination address uses the default WireGuard port 51820.

Lab Test Conclusion

The implementation enables enterprise devices to establish secure, authenticated connections to enterprise resources using the WireGuard VPN tunnel. VPN-based access controls combined with VLAN segmentation effectively restrict unauthorized access. Encrypted traffic ensures data confidentiality and integrity, as validated by packet capture analysis showing that payloads are fully encrypted within the tunnel. This layered approach enforces strict network isolation.

The current VPN implementation is successful but extending it to support multiple concurrent tunnels presents significant performance challenges. The router's CPU incurs computational overhead for encryption and decryption operations, and without dedicated hardware acceleration, performance degradation occurs under high throughput scenarios or managing multiple VPN connections.

Implementation 2: GRE Tunnel and VPN

The previously discussed implementation challenges arose concerning scalability when supporting multiple VPN connections on a single Wi-Fi access point. To address these issues, the VPN functionality has been shifted to the Wireless access gateway (WAG). This transition significantly reduces the load on Wi-Fi access points, thereby enhancing overall system efficiency.

Architecture Description

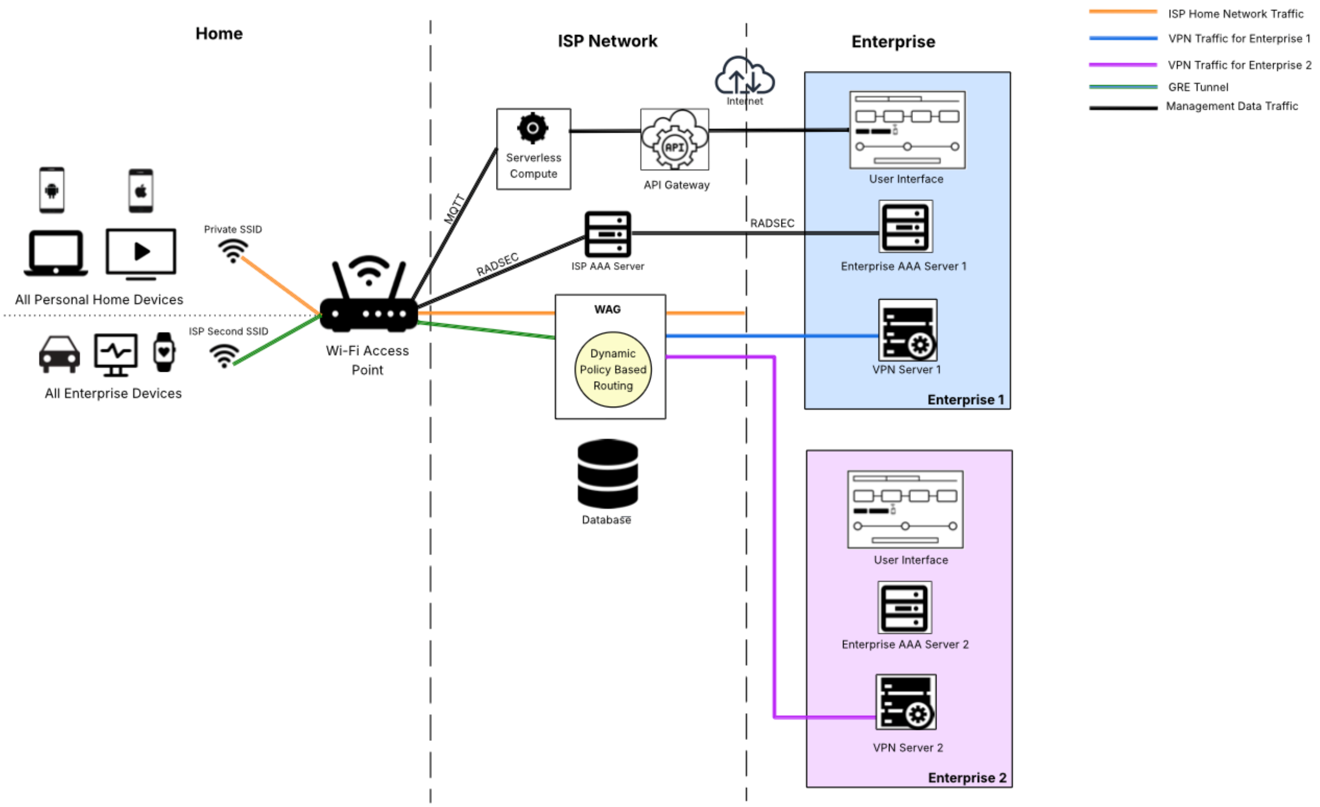


Figure 23 Implementation 2: GRE Tunnel and VPN Architecture

This architecture effectively accommodates multiple enterprise organizations through integration with an ISP-managed BNG. The system distinctly manages devices from separate enterprises, such as Device-1 from Enterprise-1 and Device-2 from Enterprise-2, utilizing AAA-based authentication and MAC address-driven subnet assignments.

The architecture provides a secure and scalable solution for routing enterprise-specific traffic across an ISP network by leveraging a secondary SSID, GRE tunnel, and VPN connections to individual enterprises. Each ISP-managed access point broadcasts a secondary wireless network that, while visible, is restricted to authorized devices. These devices, typically enterprise devices, must be pre-registered or possess valid certificates to gain network access. Upon connection, their MAC addresses or certificates are validated to ensure authorization for enterprise-specific traffic routing. Once authorized, device traffic is securely directed along a designated path toward the respective enterprise endpoint. A persistent GRE tunnel between the customer router and the WAG encapsulates all verified enterprise traffic, eliminating the complexity of per-device tunnels and simplifying overall network management. At the WAG, traffic is decapsulated and forwarded via pre-established VPN tunnels to the appropriate enterprise network.

Routing decisions rely on allocated subnet assignments or enterprise-specific certificates, ensuring clear separation and integrity of data. Enterprise device traffic is efficiently managed and forwarded using dynamic PBR, while unauthorized device traffic is restricted. PBR enforces bidirectional, device-level access control while AAA systems handle session-level authorization and detailed accounting. This setup supports comprehensive usage monitoring, enabling accurate enterprise billing. Enterprises have access to user interface for onboarding devices, configuring VPN settings, and monitoring network traffic. Additionally, API-based interfaces streamline integration with existing software systems, simplifying troubleshooting processes and lifecycle management.

The architecture maintains a comprehensive connection with devices. It promptly identifies disconnects due to power outages or signal interruptions and notifies enterprise administrators through event-driven APIs. This capability significantly enhances operational visibility and control, enabling proactive management and detailed session tracking for enterprise software teams.

Sequence Diagram

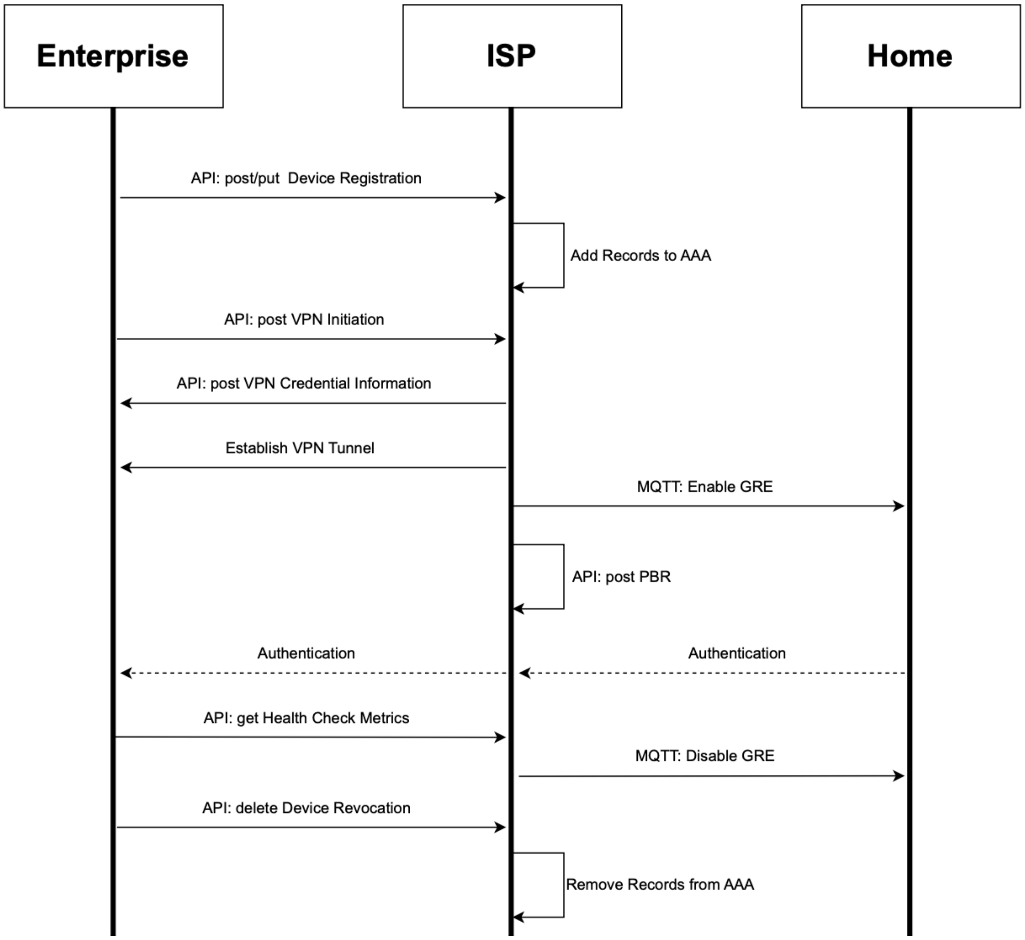


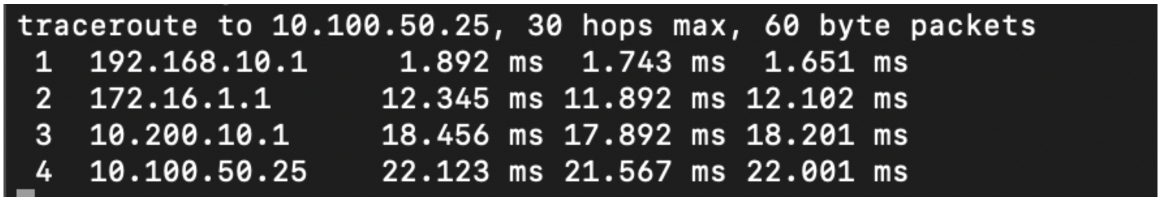
Figure 24 Implementation 2: GRE Tunnel and VPN Sequence Diagram

Lab Test Results

Traceroute analysis is a diagnostic method used to understand the network path of data packets between source and destination. This test is used to demonstrate end-to-end implementation. The enterprise-1 resource server IP is 10.50.100.25 and VPN-1 tunnel is 100.200.10.1. Enterprise-2 resource server IP is 10.150.75.30 and VPN-2 tunnel is 100.200.20.1.

Test case 1: Device-1 Connectivity to Enterprise-1 via Second SSID

Device-1 connected to second SSID attempting to reach Enterprise-1 resource at 10.100.50.25.



```
traceroute to 10.100.50.25, 30 hops max, 60 byte packets
 1  192.168.10.1     1.892 ms  1.743 ms  1.651 ms
 2  172.16.1.1      12.345 ms 11.892 ms 12.102 ms
 3  10.200.10.1     18.456 ms 17.892 ms 18.201 ms
 4  10.100.50.25    22.123 ms 21.567 ms 22.001 ms
```

Figure 25 Traceroute Results for Test Case 1

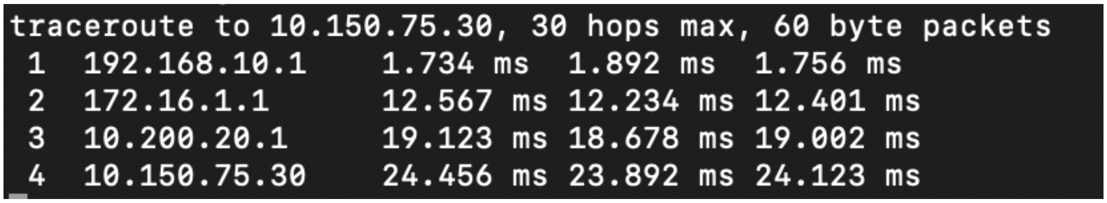
Table 5 Traceroute Information for Test Case 1

IP	Description
192.168.10.1	router ssid2
172.16.1.1	WAG tunnel endpoint
10.200.10.1	enterprise-1 vpn gw
10.100.50.25	enterprise-1 server

At home, Device-1 from Enterprise-1 connects to the second SSID and successfully establishes connectivity to Enterprise-1 via both the GRE and WireGuard tunnels. The traceroute results show a path from the router's second SSID endpoint (192.168.10.1) to the WAG tunnel endpoint (172.16.1.1). From there, the path continues to the VPN-1 tunnel endpoint (10.200.10.1) and then to the Enterprise-1 server (10.50.100.25). This defined path to the Enterprise-1 network provides lower hop counts, and consistent response times reflect secure and efficient performance.

Test case 2: Device-2 Connectivity to Enterprise-2 via Second SSID

Device-2 connected to second SSID attempting to reach Enterprise-2 resource at 10.150.75.30.



```
traceroute to 10.150.75.30, 30 hops max, 60 byte packets
 1  192.168.10.1    1.734 ms  1.892 ms  1.756 ms
 2  172.16.1.1      12.567 ms 12.234 ms 12.401 ms
 3  10.200.20.1     19.123 ms 18.678 ms 19.002 ms
 4  10.150.75.30    24.456 ms 23.892 ms 24.123 ms
```

Figure 26 Traceroute Results for Test Case 2

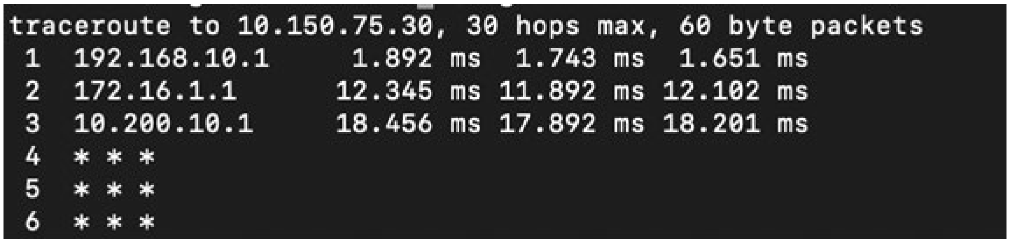
Table 6 Traceroute Information for Test Case 2

IP	Description
192.168.10.1	Wireless access ssid2
172.16.1.1	WAG tunnel endpoint
10.200.20.1	enterprise-2 vpn gw
10.150.75.30	enterprise-2 server

At home, Device-2 from Enterprise-2 connects to the second SSID and successfully establishes connectivity to Enterprise-2 via both the GRE and WireGuard tunnels. The traceroute results show a path from the router’s second SSID endpoint (192.168.10.1) to the WAG tunnel endpoint (172.16.1.1). From there, the path continues to the VPN-2 tunnel endpoint (10.200.20.1) and then to the Enterprise-2 server (10.150.75.30). Analysis of the defined network path to Enterprise-2 reveals decreased hop counts and consistent response times. These metrics collectively prove the network's secure and efficient operation.

Test case 3: Device-1 Connectivity to Enterprise-2 via Second SSID

Device-1 connected to second SSID attempting to reach Enterprise-2 resource at 10.150.75.30.



```
traceroute to 10.150.75.30, 30 hops max, 60 byte packets
 1  192.168.10.1     1.892 ms  1.743 ms  1.651 ms
 2  172.16.1.1      12.345 ms 11.892 ms 12.102 ms
 3  10.200.10.1     18.456 ms 17.892 ms 18.201 ms
 4  * * *
 5  * * *
 6  * * *
```

Figure 27 Traceroute Results for Test Case 3

Table 7 Traceroute Information for Test Case 3

IP	Description
192.168.10.1	router ssid2
172.16.1.1	WAG tunnel endpoint
10.200.10.1	enterprise-1 vpn gw

Device-1 cannot access Enterprise-2 resources, demonstrating effective cross enterprise isolation through routing policies. A traceroute shows a path from the router's second SSID endpoint (192.168.10.1) to the WAG tunnel endpoint (172.16.1.1). From there, the path continues to the VPN-1 tunnel endpoint (10.200.20.1) but timing out when attempting to reach the Enterprise-2 network (10.150.75.30).

Lab Test Conclusion

The GRE and WireGuard mediated architecture effectively supports secure, isolated connections for multiple enterprise devices to their respective enterprise resources. Each enterprise operates within its isolated network environment, enabled by dedicated VPN tunnels and distinct subnet allocations. Further, IPSec can be used to encrypt the tunnel between the wireless access point and WAG. This implementation demonstrates enterprise scalability and provides quantifiable benefits such as configurable QoS, consistent throughput, and other necessary network parameters. However, it introduces single points of failure, and without standardization, enterprise tools and hardware can complicate diagnostic efforts.

Conclusion

Securing data is a fundamental requirement in modern digital architecture, but protecting data in transit requires continual adaptation as threat vectors evolve. In an ISP environment, private secure tunnels can isolate enterprise traffic from general subscriber traffic while it traverses shared infrastructure.

This paper presented two implementations of private network connectivity within a network-as-a-service (NaaS) model. Implementation 1 establishes an end-to-end VPN from an in-home wireless access point to the enterprise network. Implementation 2 uses a GRE tunnel from the in-home wireless access point to an ISP-managed wireless access gateway (WAG), and then a VPN from the WAG to the enterprise.

Both solutions rely on the same core building blocks: (1) software modules on the access point to execute configuration actions, (2) RADIUS-based authentication and accounting, (3) policy-based routing (PBR) to segregate and steer traffic, and (4) centralized API orchestration to provision SSIDs, VLANs, and tunnel configuration. A dual-SSID design separates consumer and enterprise devices. Traffic from each SSID maps to a distinct VLAN and is then steered by PBR rules using attributes such as source identifiers (for example, MAC addresses) and authentication-derived device context.

A centralized API gateway decouples orchestration from device messaging by converting RESTful API requests into MQTT messages. This provides a single point for access control and input validation while allowing the MQTT broker layer to scale or be upgraded independently, without requiring changes to client integrations.

The two implementations differ primarily in where encryption is performed. Implementation 1 provides end-to-end encryption and integrity protection from the device to the enterprise, but it places the cryptographic workload on the access point. Lab results indicated that multiple concurrent VPN sessions can significantly reduce throughput on general-purpose access point hardware without acceleration. Implementation 2 reduces the processing burden on the access point by using a persistent GRE tunnel to the WAG and terminating enterprise VPNs at the WAG, which improves scalability across multiple enterprises. The trade-off is additional architectural complexity and operational risk, including the WAG becoming a potential single point of failure.

Several enhancements could further improve these approaches: increasing access-point compute capacity (or enabling hardware acceleration where available), expanding NaaS APIs to support event-driven policy updates and diagnostic workflows, and introducing software-defined WAN overlays that steer traffic based on real-time performance and application policy to optimize quality of service in distributed, multi-tenant deployments.

Acknowledgement

The authors would like to express their sincere gratitude to Dr. Hossam Hmimy (Sr. Director, Charter Communications), for invaluable guidance, providing critical technical input, meticulous review, task planning, progress tracking, which helped the research and trials stay on track and meet key milestones.

Abbreviations

AAA	authentication, authorization, and accounting
API	application programming interface
BNG	broadband network gateway
CPU	central processing unit
GRE	generic routing encapsulation
IoT	internet of things
IP	internet protocol
IPSec	internet protocol security
ISP	internet service provider
MAC	media access control
MQTT	message queuing telemetry transport
NaaS	network-as-a-service
OpenWrt	OpenWrt (open-source Linux distribution for routers)
PBR	policy-based routing
QoS	quality of service

RADIUS	remote authentication dial-in user service
RBAC	role-based access control
REST	representational state transfer
SSID	service set identifier
VLAN	virtual local area network
VPN	virtual private network
WAG	wireless access gateway
WAN	wide area network
Wi-Fi	wireless fidelity
WireGuard	WireGuard (VPN protocol)

Bibliography & References

[1] API gateways. Microsoft Learn.

[2] Cloudflare. (2024). What is a router? | Router definition. Cloudflare.com.

[3] Cybersecurity in Medical devices: quality system considerations and content of premarket submissions - Guidance for industry and Food and Drug Administration staff | FDA. (n.d.).

[4] Jose Ordonez-Lucena and Felix Dsouza. 2022. Pathways towards network-as-a-service: the CAMARA project. In Proceedings of the ACM SIGCOMM Workshop on Network-Application Integration (NAI '22). Association for Computing Machinery, New York, NY, USA, 53–59.

[5] Kadia, H. (2024, March 19). The Evolution of Private Wireless Networks: An In-depth Exploration into the Past, Present, and Future. TeckNexus.

[6] Network RADIUS SARL. (2014). The FREERADIUS Technical Guide.

[7] OpenWrt Documentation.

[8] The Hacker News. (2024). APIs Drive the Majority of Internet Traffic and Cybercriminals are Taking Advantage.

SCTE STANDARDS JOURNAL

Renewable Power Recipes for Powering Telecommunication Hub Sites

A Technical Paper prepared for SCTE by

Isaac Bua, Zach Taylor, Duncan McNamara, Queen Okon

Villanova University
800 East Lancaster Ave.
Villanova, PA 19085
610-519-4500

Table of Contents

List of Figures

List of Tables

Introduction

In recent years, the U.S. has seen a spike in electricity consumption for the first time since the early 2000s [1]. This increase in demand is approximately 2% annually and is predominantly occurring in the commercial and industrial sectors [1]. Driving this increase is the widespread creation of data centers, particularly those focused on utilizing artificial intelligence (AI) [2]. The most concerning part of this nationwide increase in electric demand is that large-scale providers are beginning to fall short of their target energy reserves, foreshadowing a potential energy crisis in the coming years if more generation sources are not added to the grid rapidly [2]. This is of particular concern to the telecommunication and cable provider industry, where staying online is paramount to their business model, even in the smallest sites where allowable downtime is less than one hour per year [3].

However, potential solutions exist in the form of small-scale renewable power options that can be implemented at a localized level. This enables critical facilities to be less dependent on the existing power grid, helping to make each facility more resilient while also alleviating the climate crisis by switching to sustainable options. The biggest issue with localized renewable energy is that there is not one renewable power option or even a combination of power options that exists that can be implemented at every facility. Rather, an analysis of what power options are the most ideal is considered in this paper when looking at three case study states.

Project Goal and Scope

The overarching goal of this project was to identify established and emerging renewable energy options for powering what the Society of Cable Television Engineers (SCTE) classifies as class D hub sites under SCTE standard 226 [3]. The project seeks to analyze both renewable power generation and battery energy storage systems (BESS) that go beyond the current industry standard at these facilities. The three states in which these power generation and storage options were considered in were Pennsylvania, New York, and Arizona to capture a range of climatic, regulatory, and grid conditions. The feasibility of each power and storage strategy was further assessed using social, technological, economic, environmental, and political (STEEP) framework to evaluate real-world implementation potential.

Hub Site Compared to Data Centers

While telecommunication hub sites and data centers share the need to process digital information with near-zero downtime, they diverge significantly in scale, function, and infrastructure requirements. Data centers are on a much larger scale, often ranging from 5 MW all the way to 100MW in the amount of power they need to stay up and running [4]. Additionally, the average footprint of a data center is around 40 acres or just under 2 million square feet [5]. Whereas hub sites that fall within the class D classification from SCTE 226 are around 625 square feet on average and have a power draw between 10 and 50 kW, several orders of magnitude smaller than a data center, both in power and size.

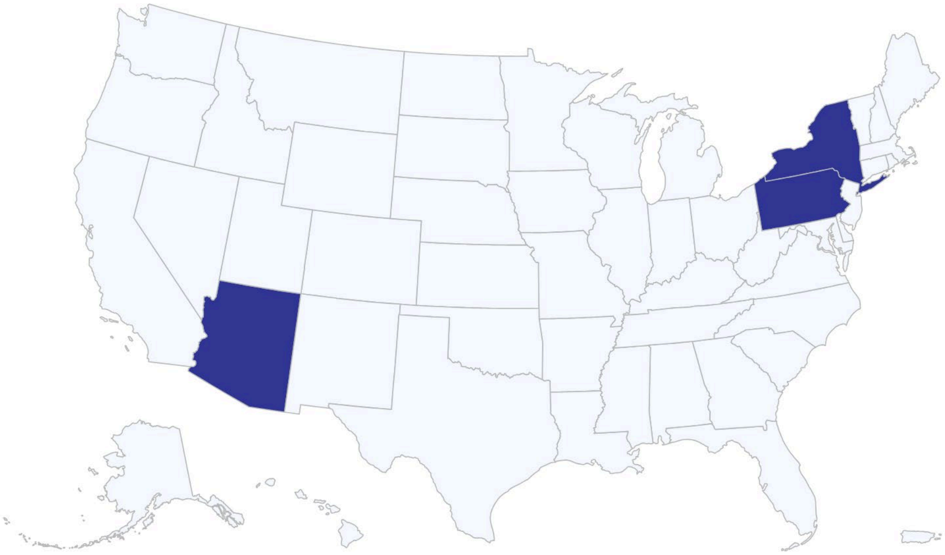
Despite the large amount of power being consumed by data centers, they have been a prime target for efforts to implement renewable power options in the digital information industry. The biggest shortcoming, however, is that multiple renewable options are either too intermittent or not compatible in the locations where data centers are present to fully run a facility off of renewable power. Instead, efforts at implementing renewable power focus on optimization of renewable energy options while still maintaining the uptime of the data center [6]. Examples often detailed the use of advanced algorithms that would make use of power from solar panels when it was available to cut back on power sourced from the grid; other models would also try to predict how much power needed to be sourced, not only at present but for subsequent days [6]. Most articles considered both solar and wind as the two main options for renewable power options in the efforts to decrease the use of fossil fuel power for data centers [6-8].

Existing Challenge of Hub Sites in Industry

The primary existing challenge associated with telecommunication hub sites when considering the implementation of localized renewable power is that each hub site varies greatly in location. One facility may be in a small room on the top of a penthouse apartment in the downtown of a busy city, while another may be tucked into the basement of the high-rise across the street. Others could be in an extra room within a shopping mall, while others could be a standalone building in a more rural location. This challenge is exacerbated when considering hub sites across multiple states, especially with site-specific renewable power options for the more site-specific options like wind, solar, hydro, and geothermal. This makes the average hub site quite different from the classic data center, which is often a large, standalone building with lots of usable spaces surrounding it for implementing on-site renewable power generation options.

Discussion of the Case Study States

As shown in Figure 1, the case study states Pennsylvania (PA), New York (NY), and Arizona (AZ) were selected not only for their diverse meteorological conditions, which favor different renewable energy technologies, but also for the identified variation in state-specific STEEP (Social, Technological, Environmental, Economic, and Political) factors influencing the deployment and perception of these technologies. A critical factor in this selection was the differences in population density across the three states, which made the analysis more tailored to city-scale contexts characterized by high levels of telecommunications activity, data traffic, and energy demands. Finally, part of the decision was reinforced by the results of a previous RISE project [9], which revealed a higher potential of renewable energy deployment and adoption in these states.



Case Study States: NY, PA & AZ

Source: Renewable Power Recipes for Powering Telecommunication Hub sites • SCTE-RISE Student Team, Villanova University

Figure 28 - Map of the U.S. Showing Location of the Chosen Case Study States, Pennsylvania, New York, and Arizona

Governing Assumptions

At the start of the project, the team conducted a site visit to one of the SCTE member data center research facilities in Pennsylvania, where key assumptions were devised to suit the scale of hub sites. The following assumptions were made to govern and inform the design and implementation strategies for each technology, as well as to vet if some technologies would best fit or fail the design constraints.

- 625 square feet of space available to work with in total
- 10 kW power draw, a small-scale facility that can fit technologies with smaller capacities
- 50 kWh needed for battery backups, considering a daily 5-hour backup power needs,
- Stability/reliability first, as these sites need to have a higher uptime as much as possible, rather than sustainability first
- On-site power generation as opposed to off-site generation to eliminate transmission challenges such as cost and associated power losses,
- Avoid the use of renewable energy credits and power purchase agreements

Renewable Power Options in the U.S.

An initial list of ten renewable and low-carbon options was considered for the hub sites, but half were quickly removed due to a lack of feasibility or a mismatch in scale. These options are shown in Table 1, and the remaining viable options are listed at the bottom of the table.

Table 8 Common Renewable and Low-Carbon Power Sources

Source	Suitable For a Hub Site?
Wind	No, not viable within the case study states [10]
Hydro	No, the scale of the power generation facility is too large
Nuclear	No, the scale of the power generation facility is too large
Natural Gas with Carbon Capture	No, the scale of the power generation facility is too large
Geothermal	No, the scale of the power generation facility is too large
Solar	Yes
Fuel Cells	Yes
Green Hydrogen	Yes
Biogas	Yes
Battery Energy Storage Systems (BESS)	Yes

The main disqualifier for the sources no longer considered was the fact that the infrastructure needed to generate power from them is too large, or that they generated far too much power when looking at the assumption made about the size of the average hub site. For power options like nuclear and geothermal, the scale of power generated was determined to be large [11-12]. On the small end of these facilities is 100 kW generation, with most being closer in size to several MW, all the way up to a GW [11-12]. Most of the hydropower in the U.S. is based in the Pacific Northwest and is also heavily reliant on the existence of prebuilt dams, often federally funded ones [13]. This specific geolocation does not line up with the case study states chosen, and the need to create an entire dam eliminates this option. Utilizing carbon capture at a natural gas power plant is only an economically viable option for a power plant larger than 100MW, which is far larger than the power needed for a hub site [14]. Discussions with SCTE revealed that these facilities would be far too large for powering a hub site in addition to trying to fit one within the 625 square foot threshold.

The other disqualifier, which was specific to wind, is that the three case study states studied are not in a viable zone of the country to generate power from wind [10]. To generate a meaningful amount of power for a utility scale requires constant wind speeds around 6 meters per second [10]. These winds can be reached in the three states, but not reliably [10].

After eliminating the first five options, the remaining options cover both power generation options and battery back-ups, and an in-depth analysis of each option is discussed in the following section.

Viable Power Options

Solar

Solar photovoltaic (PV) systems represent a mature and increasingly cost-effective renewable energy option for hub sites and network facilities. This white paper evaluates the practical performance and limitations of rooftop solar PV deployment using standardized design assumptions and the National Renewable Energy Laboratory (NREL) PVWatts® modeling tool version 8.5.2 [15]. Three representative U.S. locations within the states of New York, Pennsylvania, and Arizona are analyzed to illustrate how geography influences annual energy output and emissions avoidance. The results demonstrate both the value and constraints of rooftop solar as part of a broader energy strategy for telecommunications infrastructure.

Solar energy is one of the most widely deployed renewable energy technologies in the United States, benefiting from decades of technological advancement, declining capital costs, and improved modeling accuracy. For telecommunications operators, solar PV offers a pathway to reduce grid dependence, hedge against future electricity price volatility, and lower carbon emissions at network facilities. However, solar energy production is inherently variable and dependent on site-specific factors, making a realistic performance assessment essential before deployment.

To ensure comparability across locations, a standardized rooftop solar configuration was defined based on industry norms and site-visit feedback. Rooftop availability was assumed to be limited to a 25 ft × 25 ft footprint, reflecting common structural and clearance constraints observed at existing hub facilities. Using standard commercial solar panel dimensions (approximately 3.5 ft × 6 ft), this available area supports the installation of 28 photovoltaic panels, as illustrated in Figure 2.

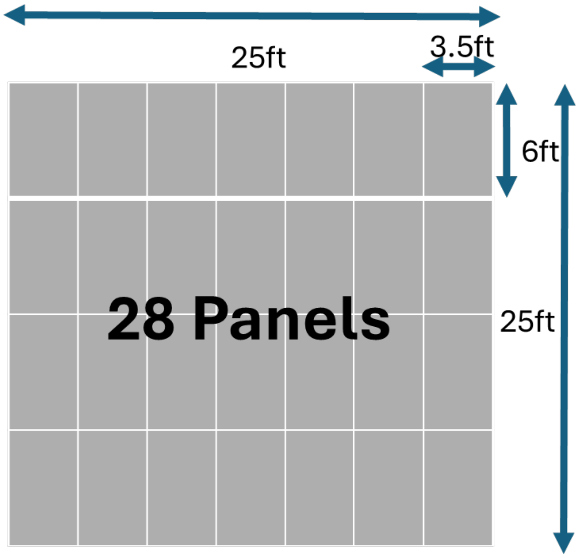


Figure 29 - Ideal Rooftop Solar Panel Installation for a Hub Site

The selected module technology was crystalline silicon, which remains the dominant panel type in commercial and utility-scale installations due to its balance of efficiency, durability, and cost. The array was modeled as a fixed, roof-mounted system with a 20-degree tilt angle, reflecting a practical compromise between optimal solar geometry and typical rooftop construction constraints. The array azimuth was assumed to be south-facing (180°), consistent with standard design practice in the Northern Hemisphere.

System losses—including inverter inefficiency, wiring losses, soiling, and mismatch—were modeled at approximately 14%, aligned with industry-accepted default values. A DC-to-AC size ratio of 1.2 and an inverter efficiency of 96% were applied to reflect common commercial system designs.

Energy generation was calculated using the PVWatts® Calculator, a widely adopted modeling tool developed by NREL. PVWatts integrates long-term solar irradiance data from the NREL database with system-specific parameters to estimate annual energy production. The model follows a transparent calculation sequence that includes:

1. Solar geometry
2. Plane-of-array irradiance
3. Cell temperature
4. DC power generation
5. AC power conversion
6. Annual energy output

This approach allows for consistent comparison across locations while accounting for regional differences in solar resource availability and ambient conditions.

Three locations were selected to represent a range of solar resource conditions commonly encountered by U.S. telecommunications operators:

- Rochester, New York – representing a northern climate with moderate solar irradiance
- Exton, Pennsylvania – representing a mid-Atlantic climate with balanced seasonal exposure
- Tempe, Arizona – representing a high-irradiance southwestern climate

Each location was assumed to support the same physical system configuration, allowing differences in output to be attributed primarily to geographic and climatic factors.
The modeled annual energy production results demonstrate the strong influence of location on solar PV performance. The Arizona site produced the highest annual energy output, followed by Pennsylvania and New York. This outcome aligns with expectations based on regional solar irradiance patterns.

- Rochester, NY: approximately 12,052 kWh/year, avoiding 1.32 metric tonnes of CO_2
- Exton, PA: approximately 13,121 kWh/year, avoiding 3.55 metric tonnes of CO_2
- Tempe, AZ: approximately 17,472 kWh/year, avoiding 5.58 metric tonnes of CO_2

These results highlight that even with identical system designs, annual energy generation can vary by more than 40% between northern and southwestern U.S. locations. For telecommunications operators with geographically distributed assets, this reinforces the importance of site-specific modeling rather than relying on generalized assumptions. This analysis was conducted as part of a broader evaluation of renewable energy options for telecommunications hub sites. The goal was not to optimize a single site for maximum output, but rather to establish a repeatable, industry-representative methodology that could be applied across multiple geographic regions using consistent assumptions.

While rooftop solar provides measurable energy and emissions benefits, it also presents fundamental technical limitations. Solar generation is intermittent by nature, dependent on daylight availability and weather conditions. Peak energy production often does not align with peak facility demand, particularly for network sites that operate continuously.

As a result, energy storage becomes a critical enabling technology for maximizing the value of solar PV in telecommunications applications. Battery systems allow excess daytime generation to be stored and deployed during periods of low or no solar output, improving self-consumption and resilience. However, storage systems introduce additional capital cost, operational complexity, and lifecycle considerations that must be incorporated into deployment decisions.

Space constraints also limit the absolute contribution of rooftop solar at hub sites. Even in high-irradiance regions, a 25 ft × 25 ft array represents a supplemental energy source rather than a full replacement for grid power. Solar PV should therefore be viewed as a component of a diversified energy strategy rather than a standalone solution.

Fuel Cells

Fuel cells convert the chemical energy of a fuel directly into electricity and heat by electrochemical reactions, avoiding combustion and its associated efficiency and emissions constraints. Stationary systems typically operate on hydrogen or hydrogen-rich fuels and provide high-efficiency DC power with low local emissions, low noise, and high modularity [16–18]. These characteristics align well with SCTE Class-D hub sites (≈ 10kW, 240kWh/day plus 50kWh battery), where space is limited (625ft^2), reliability is paramount, on-site generation is preferred, and the use of RECs and PPAs is to be avoided

Proton-exchange-membrane fuel cells (PEMFCs) use a proton-conducting polymer electrolyte and operate at 50-80°C with high-purity hydrogen at the anode and air at the cathode. Commercial systems typically achieve 40-60% electrical efficiency at practical part-loads and can start and respond to load changes in seconds due to their low thermal mass [16–18]. Telecom-oriented PEMFC products are available in the 1-10kW range as outdoor cabinets feeding 48V DC buses and have been deployed at thousands of base stations and central offices as extended-runtime backup, maintaining batteries at float charge under normal conditions [18, 23–26]. Solid-oxide fuel cells (SOFCs) employ an oxygen-ion–conducting ceramic electrolyte and operate at 600-800 °C. These temperatures allow internal reforming of light hydrocarbons, so SOFCs can operate on pipeline natural gas, LPG, biogas, or hydrogen with limited external fuel processing [19–21]. Electrical efficiencies of 50-60% are typical, and total fuel-to-useful-energy efficiencies around 80% are achievable in combined-heat-and-power (CHP) configurations [19–22]. High operating temperature and thermal inertia imply start-up times of tens of minutes, making SOFCs better suited to continuous prime-power or long-duty-cycle operation than to purely standby backup roles. For a representative 10kW Class-D facility, a PEMFC+battery architecture is naturally configured as a fast-start backup or N+1 resilience layer on top of the grid. An SOFC-based system is instead positioned as a high-efficiency prime-power plant running continuously on natural gas or biogas, with the battery system providing ride-through and transient support [18–22, 25, 26]. Both technologies integrate cleanly with the DC power architecture used in telecom environments.

Empirical performance data from U.S. telecom applications show that fuel-cell systems can meet stringent reliability requirements. A multi-year NREL program with hundreds of 4-6kW PEMFC backup units across 15 states documented a 99.5% successful-start rate over more than 1500 demand events; most failures were linked to balance-of-plant and fuel logistics rather than stack failure [23]. Cost-of-ownership analyses indicate that for backup duration $\geq 72h$, fuel-cell systems can be competitive with diesel generators on a levelized cost of-reliability basis, especially when noise and local-emissions constraints are binding [24, 25]. A broader NREL assessment of distributed-energy-resource reliability concludes that modern fuel-cell systems match or exceed diesel gen-sets for start reliability, shifting the main risk to fuel availability and hydrogen system integrity [27]. Relative to small diesel gen-sets, fuel cells offer higher efficiency and decouple power from energy. PEMFCs generally achieve 40-60% electrical efficiency, while SOFCs exceed 50% and can surpass 80% total efficiency with CHP [19–22]. Small reciprocating engines of similar sizes commonly operate at 30-35% efficiency [18, 22].

Because fuel-cell run-time is limited by fuel storage rather than by battery capacity, multi-day autonomy can be obtained without the space penalty of very large battery banks or diesel tanks. Telecom demonstrations have recorded multi-day runtimes using compressed hydrogen and effectively unlimited run-time for natural-gas-fired systems [23–26]. Packaged 5-20kW units typically occupy only a few square meters and meet urban noise limits without additional acoustic treatment [18, 23], which is compatible with the 625ft^2 constraint. The main barriers to wider deployment are economic, infrastructural, and regulatory. Installed costs for stationary fuel-cell systems remain high, in the 2400-5500 USD/kW range for natural-gas systems [22], with small PEMFC telecom units at still higher specific costs [22, 24]. Hydrogen for PEMFCs is expensive and logistically constrained. Delivered prices of 5-15 USD/kg translate into electricity prices that may exceed commercial tariffs unless low-cost hydrogen or high-value resilience services are available [24, 25]. On-site electrolysis improves autonomy but ties hydrogen cost to local electricity rates, which vary significantly across the three states [42,43]. In addition, stationary fuel-cell installations must comply with NFPA 853, and hydrogen-fueled systems must meet NFPA 2 and NFPA 55 requirements as adopted through the International Fire Code and state building codes [28–31]. NREL code-and-standards reviews highlight permitting complexity and AHJ familiarity as key non-technical barriers, particularly at constrained urban sites [28, 29].

- ***State-level STEEP assessment and implications for fuel cells***

Performing a STEEP assessment highlights how New York (NY), Pennsylvania (PA), and Arizona (AZ) differentially shape fuel-cell viability for SCTE Class-D sites.

Socially and technologically, all three states can benefit from the low noise and negligible local emissions of fuel cells, and PEMFC/ SOFC technologies are commercially mature [16–22]. New York has the strongest track record, driven by NYSERDA programs and participation in a Northeast hydrogen hub [32–34]. Pennsylvania has moderate experience through industrial and institutional projects associated with the Appalachian Regional Clean Hydrogen Hub [35–37]. Arizona's fuel-cell activity has been more limited and exploratory [38, 39].

Economically and environmentally, New York's comparatively high commercial tariffs (≈ 0.18 USD/kWh) [42] and aggressive CLCPA driven decarbonization [43] make high-efficiency on-site generation more attractive and improve the lifecycle performance of electrolytic hydrogen relative to PA and AZ. Historic NYSERDA programs (PON 2157 and the Clean Energy Fund Stationary Fuel Cell Program) have provided substantial capacity and performance incentives, and a recent $3.7 million solicitation targets clean-hydrogen-based fuel-cell resources [42, 43]. Pennsylvania supports fuel cells as Tier I resources under AEPS and offers grants and loans through the ACE Program [35–37], but these instruments are most effective at a larger scale. Arizona's Renewable Energy Tax Incentive Program primarily targets large manufacturing and data-center investments, while the Hydrogen Fuel Production Support statute created a study framework but no direct small-scale deployment incentives [38–40].

Politically and regulatorily, New York again ranks strongest, with a decade-plus history of fuel-cell-specific programs and relatively clear code integration of NFPA 2 and 853 [28, 32–34].

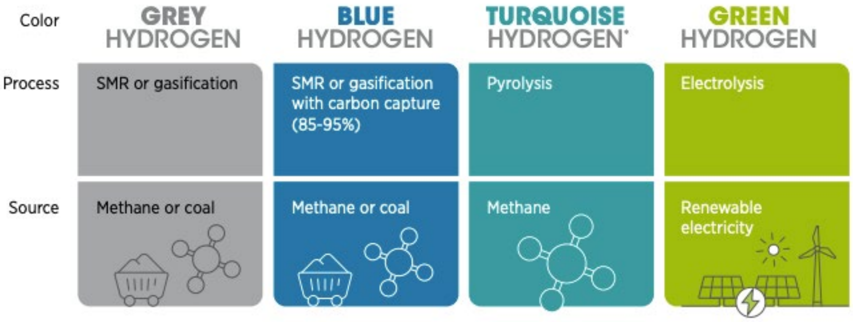
Pennsylvania provides a supportive but more generic framework under AEPS/ACE and IFC adoption, as permitting outcomes for hydrogen systems are more site- and AHJ-dependent [29, 35–37]. Arizona has initiated hydrogen-policy discussions but offers limited direct incentives, and some local fire-code guidance adopts conservative interpretations of NFPA 2/853 that can complicate siting at compact telecom locations [38–41].

For SCTE's power "recipes," this suggests that PEMFC+battery architectures are most compelling in New York and selected Pennsylvania sites where hydrogen logistics and incentives are favorable. Across all three states, fuel cells provide a high-reliability, small-footprint option that aligns with the project's emphasis on stability and on-site generation, but their relative weight in the overall recipe should be highest in New York, moderate in Pennsylvania, and more selective in Arizona.

Green Hydrogen

Hydrogen is an invisible gas, but depending on the type of production used, different color names are assigned to hydrogen. Green hydrogen, grey hydrogen, blue hydrogen, and turquoise hydrogen. These are essentially color codes or nicknames used within the energy industry to differentiate between the types of hydrogen. Grey hydrogen, for example, is currently the most common form of hydrogen production. Grey hydrogen is created from natural gas, or methane, using steam methane reformation, but without capturing the greenhouse gases made in the process. Alternatively, blue hydrogen is produced using the same process, but it includes the use of carbon capture and storage (CCS) to trap and store the carbon dioxide released in the process. Blue hydrogen is sometimes described as 'low-carbon hydrogen', as the steam reforming process doesn't avoid the creation of greenhouse gases [44].

In this project, we focused on using green hydrogen, which is made by using clean electricity from surplus renewable energy sources, such as solar or wind power, to electrolyze water. Electrolyzers, the device used in the electrolysis process, use an electrochemical reaction to split water into its components of hydrogen and oxygen, emitting zero-carbon dioxide in the process. As illustrated in Figure 3, green hydrogen currently makes up a small percentage of the overall hydrogen because production is expensive. Just as energy from wind power has reduced in price, green hydrogen will come down in price as it becomes more common [45].



Color	GREY HYDROGEN	BLUE HYDROGEN	TURQUOISE HYDROGEN*	GREEN HYDROGEN
Process	SMR or gasification	SMR or gasification with carbon capture (85-95%)	Pyrolysis	Electrolysis
Source	Methane or coal	Methane or coal	Methane	Renewable electricity

Figure 30 - Selected Shades of Hydrogen and Their Production Methods [45]

- ***How Green Hydrogen is produced***

Electrolysis is a promising option for carbon-free hydrogen production from renewable and nuclear resources. Electrolysis is the process of using electricity to split water into hydrogen and oxygen. This reaction takes place in a unit called an electrolyzer. Electrolyzers can range in size from small, appliance-size equipment that is well-suited for small-scale distributed hydrogen production to large-scale, central production facilities that could be tied directly to renewable or other non-greenhouse-gas-emitting forms of electricity production. Like fuel cells, electrolyzers consist of an anode and a cathode separated by an electrolyte. Different electrolyzers function in different ways, mainly due to the different types of electrolyte material involved and the ionic species it conducts [46].

In a polymer electrolyte membrane (PEM) electrolyzer, the electrolyte is a solid specialty plastic material. The unit has both an anode and a cathode end. Water reacts at the anode to form oxygen and positively charged hydrogen ions (protons). The electrons flow through an external circuit, and the hydrogen ions selectively move across the PEM to the cathode. At the cathode, hydrogen ions combine with electrons from the external circuit to form hydrogen gas. This process within the PEM occurs between 70 and 90 °C, as shown in Figure 4, and summarized below [47].

- Anode Reaction: $2H_2O \rightarrow O_2 + 4H^+ + 4e^-$
- Cathode Reaction: $4H^+ + 4e^- \rightarrow 2H_2$

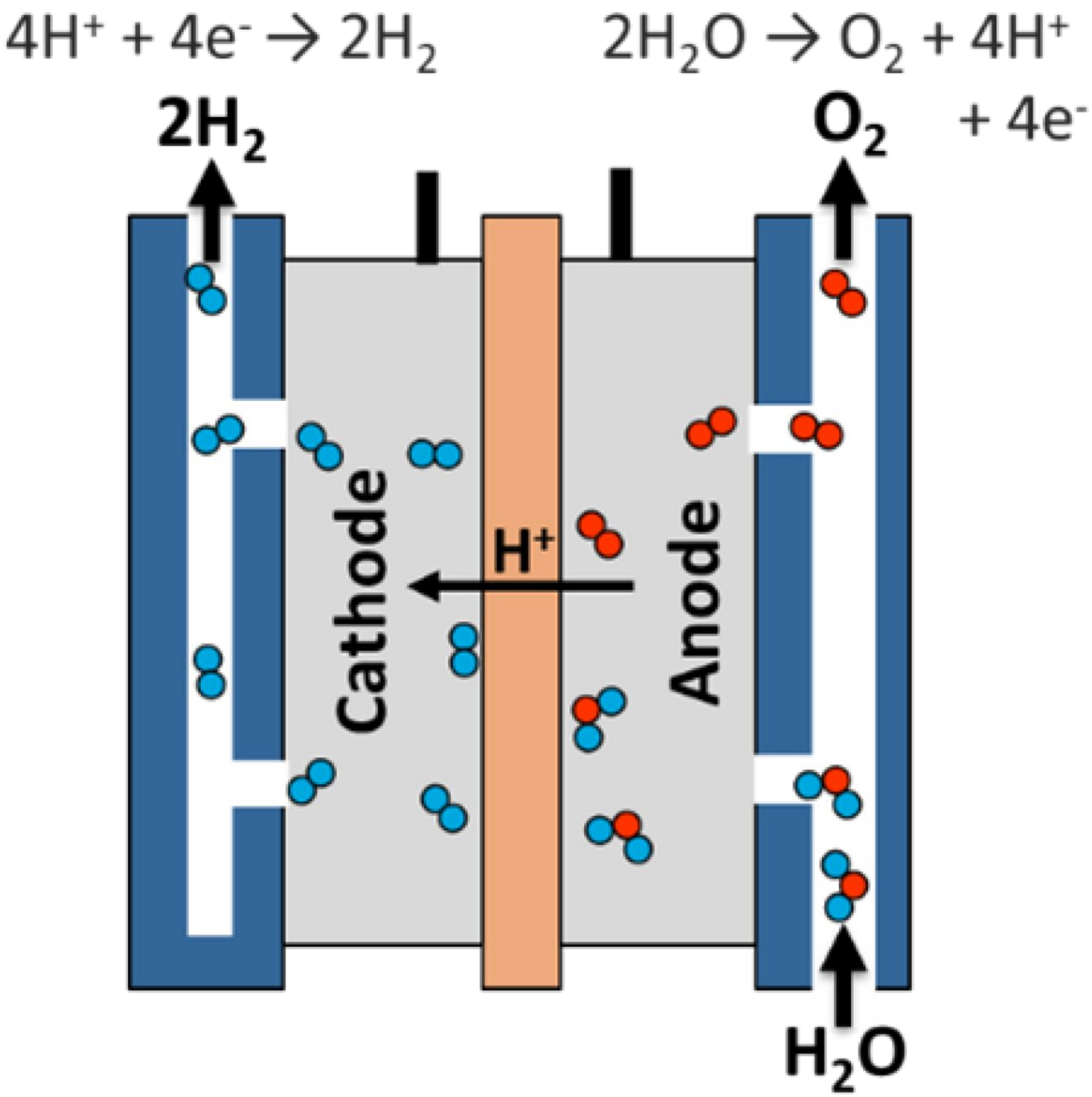

$4H^+ + 4e^- \rightarrow 2H_2$
$2H_2O \rightarrow O_2 + 4H^+ + 4e^-$
$2H_2$
O_2
Cathode
H^+
Anode
H_2O

Figure 31 - Polymer Electrolyte Membrane Electrolyzer, for Hydrogen Production [47]

- ***Key drivers in Green Hydrogen production today***

There have been several waves of interest in hydrogen in the past. These were mostly driven by oil price shocks, concerns about peak oil demand or air pollution, and research into alternative fuels. For green hydrogen, there are several key factors driving its adoption globally:

- Low variable renewable energy costs since the major cost driver of green hydrogen is the cost of electricity.
- Government objectives for net-zero energy systems
- Broader use of hydrogen. Previous waves of interest in hydrogen were focused mainly on expanding its use in fuel cell electric vehicles (FCEVs). As illustrated in Figure 5, in contrast, the new interest covers many possible green hydrogen uses across the entire economy, including the additional conversion of hydrogen to other energy carriers and products, such as ammonia, methanol, and synthetic liquids. These uses can increase the future demand for hydrogen and can take advantage of possible synergies to decrease costs in the green hydrogen value chain, shown in Figure 5 [47].

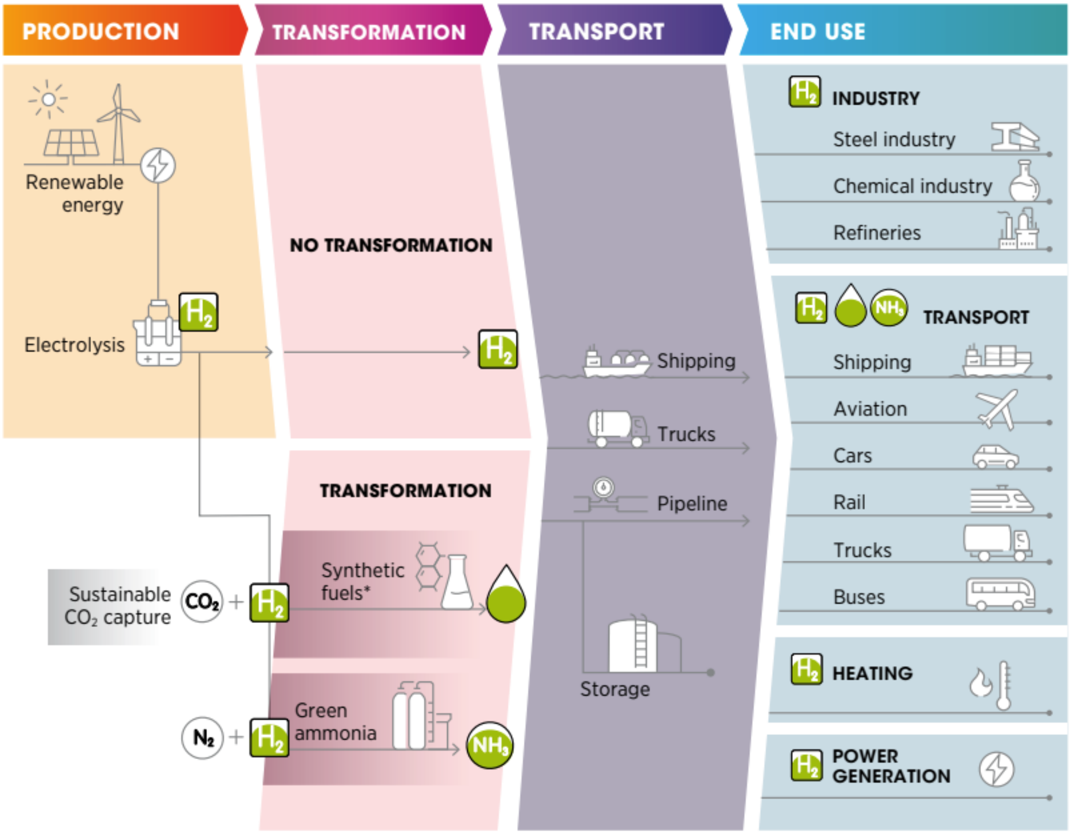


Figure 32 - Green Hydrogen Production, Conversion, and End Use Across the Energy System [45]

- ***Barriers to uptake of Green Hydrogen***

- High Production costs

The levelized cost of clean hydrogen (LCOH) produced from renewable electricity using currently available PEM electrolyzer technology and various renewable energy sources is approximately \$5 to \$7 per kilogram (kg) (in 2022 USD and without subsidies). These values are based on a range of PEM electrolyzer installed capital costs (average of \$2,000/kW) using various references, real-world data, and analytical models at low manufacturing volumes; renewable electricity costs of roughly $0.03/kWh; and capacity factors ranging from approximately 50 to 75%. Additional cases with higher renewable costs and lower capacity factors (e.g., solar) can result in higher LCOHs.

- Lack of dedicated infrastructure

Hydrogen has to date been produced close to where it is used, with limited dedicated transport infrastructure. There are only about 5000 kilometers (km) of hydrogen transmission pipelines around the world (Hydrogen Analysis Resource Center, 2016), compared with more than 3 million km for natural gas. There are 470 hydrogen refueling stations around the world (AFC TCP, 2020), compared with more than 200,000 gasoline and diesel refueling stations in the United States and the European Union. Natural gas infrastructure could be repurposed for hydrogen (IRENA, IEA, and REN21, forthcoming), but not all regions of the world have existing infrastructure.

- Energy losses.

Green hydrogen incurs significant energy losses at each stage of the value chain. About 30-35% of the energy used to produce hydrogen through electrolysis is lost. In addition, the conversion of hydrogen to other carriers (such as ammonia) can result in 13-25% energy loss, and transporting hydrogen requires additional energy inputs, which are typically equivalent to 10-12% of the energy of the hydrogen itself. Using hydrogen in fuel cells can lead to an additional 40–50% energy loss. The total energy loss will depend on the final use of hydrogen. The higher the energy losses, the more renewable electricity capacity is needed to produce green hydrogen [49].

Biogas

Biogas is primarily a mixture of methane and carbon dioxide produced through the bacterial decomposition of organic waste materials in the absence of oxygen, a process known as anaerobic digestion. The primary feedstocks for biogas systems include livestock manure, municipal solid waste, wastewater biosolids and primary sludge, food loss and waste, and food production residuals (see **Figure *6***). These bio-wastes present a major challenge for economies worldwide, as several million tons of organic waste are disposed of in landfills each year. However, these materials can instead be diverted from landfills and converted into usable energy [50, 51, 54].

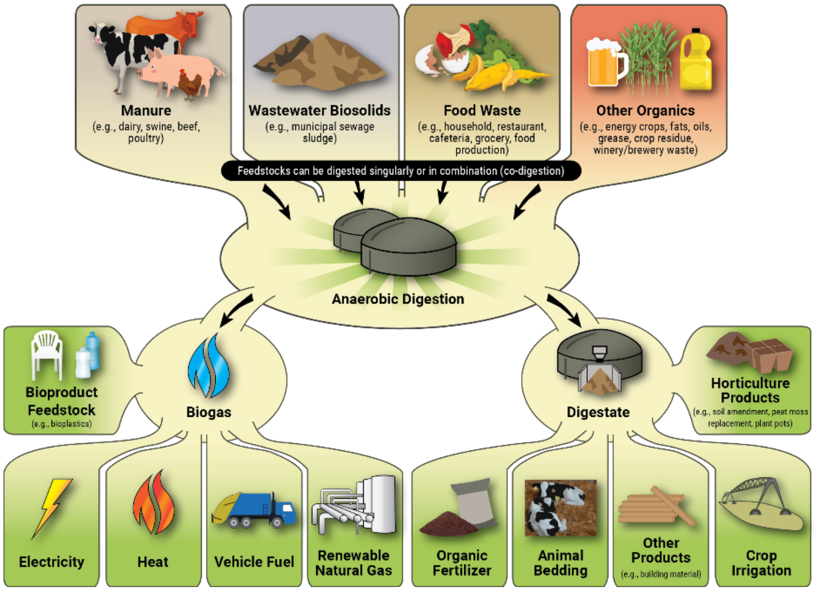

Manure
(e.g., dairy, swine, beef, poultry)
Wastewater Biosolids
(e.g., municipal sewage sludge)
Food Waste
(e.g., household, restaurant, cafeteria, grocery, food production)
Other Organics
(e.g., energy crops, fats, oils, grease, crop residue, winery/brewery waste)
Feedstocks can be digested singularly or in combination (co-digestion)
Anaerobic Digestion
Bioproduct Feedstock
(e.g., bioplastics)
Biogas
Digestate
Horticulture Products
(e.g., soil amendment, peat moss replacement, plant pots)
Electricity
Heat
Vehicle Fuel
Renewable Natural Gas
Organic Fertilizer
Animal Bedding
Other Products
(e.g., building material)
Crop Irrigation

Figure 33 - The Flow of Feedstocks through the Anaerobic Digestion System [51]

Anaerobic digestion systems can process either a single feedstock or multiple organic feedstocks through co-digestion. Biogas systems capture methane that would otherwise escape into the atmosphere and utilize it to generate energy, including electricity, thermal energy, and vehicle fuel. In addition to energy production, biogas systems generate valuable non-energy byproducts, such as nutrient-rich soil amendments, pelletized and pumpable fertilizers, and feedstocks for plastics and chemical production. A complete anaerobic digestion system includes infrastructure for managing organic waste and producing biogas, as well as energy-conversion equipment, such as combined heat and power (CHP) systems. Unlike intermittent renewable energy sources such as wind and solar, biogas provides a continuous and reliable energy supply with a high-capacity factor. This flexibility and reliability make biogas systems an important renewable energy asset [50, 51].

The three main types of anaerobic digesters are stand-alone digesters, on-farm digesters, and digesters located at wastewater treatment plants. Stand-alone digesters typically accept feedstocks from one or more external sources in exchange for a tipping fee, with food waste being the primary feedstock. Digesters designed for food waste can also co-digest other organic materials, including yard waste, manure, and wastewater solids. On-farm digesters, as the name implies, are located on agricultural farms and convert farm-generated organic waste into energy. Digesters at wastewater treatment plants, also referred to as Water Resource Recovery Facilities (WRRFs), are used to treat wastewater solids while producing biogas. The digesters at WRRFs are subject to U.S. Environmental Protection Agency (EPA) biosolids regulations under Title 40 of the Code of Federal Regulations, Part 503 [52, 53].

Beyond improving energy efficiency within data centers, biogas can play a critical role in addressing the rapidly increasing energy demands of data center infrastructure. Specifically, biogas-based energy systems can mitigate the escalating electricity demand associated with growing data center loads by providing a stable and locally sourced power supply [54].

One example of biogas utilization for powering telecommunication and data center infrastructure in rural areas involves partnerships between telecommunication companies and family-owned farms. In such arrangements, farm waste is converted into energy to power telecommunications head ends, network hubs, or larger data centers. For instance, AG-Grid Energy, a Pennsylvania-based company, commissioned an anaerobic digestion system in October 2024 in partnership with Lent Hill Dairy Farm, forming the special-purpose entity Lent Hill Ag-Grid Digital LLC (LHAG Digital). This project supplies electricity to a 1 MW crypto data mining unit by off-taking power from Lent Hill Ag-Grid's anaerobic digesters and CHP system. The system employs a co-digestion process using approximately 45,000 gallons per day of food waste delivered to the farm, along with manure generated by 4,000 cows, as shown in **Figure** *7*[55].

Figure 34 - **Lent Hill Ag-Grid Digital LLC 1MW Capacity Anaerobic Digestion System [54]**

New Battery Energy Storage System Options

The last power option considered is already in place at hub sites, which are BESS's. BESS are used at hub sites in the event of an upset condition to serve as a temporary source of power while backup generators come online to power the facility until the upset condition has ended. However, if due to multiple failures I.e both grid and backup generator failure, then the batteries need to be able to last about five hours or until the generators can be made functional. With the assumption that the hub site is drawing about 10 kWs of power this means the batteries need to supply at least 50kWh of power in case of an upset condition. When speaking with an employee of an SCTE member company, it was established that most BESSs at hub sites utilize lead-acid batteries as they are extremely stable and have a low risk of causing a fire. This project investigated two new potential battery options for BESS that have been gaining traction over the past few years.

- ***Lithium Ion***

The first option considered was the use of lithium-ion batteries, which have become increasingly popular as a form of long-term energy storage due to their high energy density, ranging from 150 to 190 Wh/kg [56]. Some lithium batteries can even have a density of up to 330 Wh/kg, which is over four times that of a typical lead-acid battery at around 75 Wh/kg. Additionally, Lithium-ion batteries do not contain toxic elements like lead or cadmium and have a low self-discharge rate [57]. All these factors have led lithium-ion batteries to become very popular over the last few decades. The global demand for lithium batteries is expected to increase seven times between 2022 and 2030, with 80% of this market being driven by electric vehicle battery creation [58]. In addition to their use as batteries for electric vehicles, lithium-ion batteries are used as BESS for the United States power grid, making up most of the battery storage that exists on the U.S. power grid [59]. Lithium-ion batteries are not without their drawbacks, with the most concerning one for this project being the risk of thermal runaway happening in a lithium-ion cell. Thermal runaway is when a lithium-ion cell enters a self-heating state that is uncontrollable and can result in a fire [60]. Thermal runaway can be caused by several factors, but some examples are the external temperature of the surrounding environment, an overcharge of the battery, or a short circuit within the battery cell [60]. Once these fires start, they can be extremely hard to extinguish due to the self-oxidizing nature of the lithium salts used in the batteries [61].

- ***Zinc Bromine***

The other battery option examined was zinc bromine, which has been around since the 1970s but only recently gained in popularity due to the base materials being more earth-abundant than lithium-ion batteries [62]. Zinc bromine batteries do not have commercial variants with the high energy density of lithium ion, with current batteries having an energy density of between 40-70 Wh/kg [62-63]. There have been trial runs of zinc bromine batteries that can get up to 150 Wh/kg, but it is unknown when these will be available on the market [64]. In addition to being made of earth-abundant materials, zinc bromine batteries have an extremely low risk of thermal runaway due to the batteries using a water-based electrolyte [62, 65]. Where zinc bromine batteries struggle is with the growth of dendrites on the anodes, which can potentially degrade the performance of the battery or cause a short circuit [62]. Zinc-bromine batteries at an industrial level are still relatively new, but they are poised to enter the market as a cheaper, if less energy-dense option compared to lithium-ion [62, 66].

Power Capacities Within Governing Assumptions

In Table 2, the maximum energy that can be generated per technology has been listed. This considers what is feasible given the constraints laid out in the governing assumptions and does not consider a state-by-state analysis yet. Furthermore, these scenarios are looked at from an ideal lens i.e for solar, if the hub site has access to roof-mounted solar in Arizona, then it would be feasible to generate up to 18% of the needed power this way. This ideal case will not always happen, as one hub site is not like another, and there will be lots of variability in the numbers presented below when applying these technologies to a real-world site.

Table 9 Maximum Power Output Achievable for an Ideal Hub Site

Power Source	**Arizona**	**New York**	**Pennsylvania**
Solar- Upper Bound	18%	14%	13%
Biogas- Upper Bound	Up to 100% *Depending on feedstock availability*	Up to 100% *Depending on feedstock availability*	Up to 100% *Depending on feedstock availability*
Fuel Cell + Green Hydrogen – Upper Bound	100%	100%	100%

Analysis of Viable Options

This section examines the power options above first from a STEEP perspective. Across all the states, the STEEP category representing the biggest challenge and the greatest opportunity for each technology is displayed in Tables 3 and 4. Analysis was done strictly on a state-by-state basis, so if there was an overarching federal loan or policy affecting a technology, it was disregarded. This analysis helps identify major areas where these technologies still need to improve and how they could potentially hinder a hub site's operation.

Next, the STEEP assessment was broken down into underlying categories in section with each category given a score of 1-5 for each technology in each state. This examined the trends for each technology in each state to assess if it is progressing positively (5) or negatively (1). Figure 8 showcases these scores in each of the states. Lastly, these scores were regrouped into their overarching categories, and working with SCTE, each category was assigned a weight on how important it would be to a hub site. These final readiness scores are shown in two heat maps for power generation options and BESS options, respectively.

STEEP Greatest Opportunities and Challenges of Power Options

Table 10 Greatest STEEP Opportunity for Each Power Option

Technology	**Greatest Opportunity**
Solar	**Economic:** Once installed, solar panels have a low maintenance cost over their lifetime, and with a life span of about 10-15 years, leading to a low cost after the initial investment
Biogas	**Technology:** Waste heat from the hub site can be fed into digesters for anaerobic digestion. Helping to promote a circular economy model.

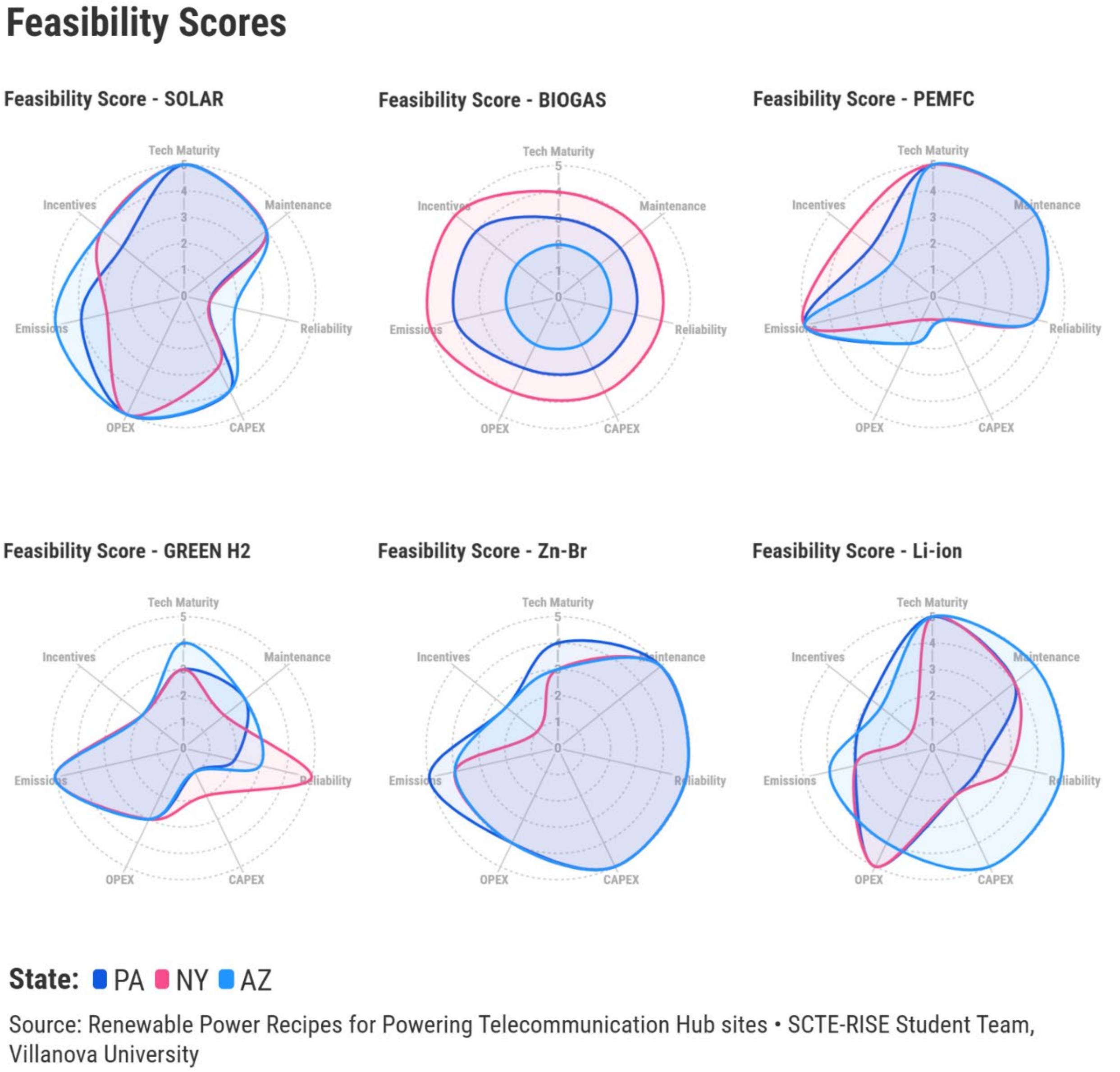
Fuel Cells	**Technology:** Direct electrochemical conversion yields high electric efficiency, and CHP-ready heat recovery lifts total utilization to 80–90% in modular, low-maintenance distributed power with ultra-low local pollutants (potentially net-zero with green fuels).
Green Hydrogen	**Technology:** Energy density of hydrogen is equal to 33.6 kWh/kg of usable energy per kg, versus diesel, which only holds about 12–14 kWh per kg Low water intensity requiring around 20-30 L/kg Hydrogen
BESS	**Economic:** Between governmental loans and a strong U.S.-based supply chain, zinc bromine batteries are primed to be a domestic economic driver, especially in the state of Pennsylvania

Table 11- Greatest STEEP Challenge for Each Power Option

Technology	Greatest Challenge
Solar	**Technology**: Solar is an intermittent power source which makes it hard for it to meet the demands of the headend
Biogas	**Economic**: High upfront capital costs, and unlike other renewable energy sources such as solar, biogas systems require regular specialized maintenance. Smaller telecom sites may also struggle to justify the investment unless the system is integrated with a nearby industrial or agricultural facility
Fuel Cells	**Economic**: Levelized cost of clean hydrogen produced from today’s low volume PEM is around \$5 to \$7/kg-H_2 at an installed capital cost of \$2,000/kW and RE sources with CF - 50-75%
Green Hydrogen	**Technology**: Energy-intensive, electrolyzer system electricity usage is around 57 kWh/kg
BESS	**Social/Political**: Public uncertainty over BESS is very high within the U.S. after several fires involving these power storage systems, leading to legislation focusing on banning BESS

Weighted Heat Maps

The feasibility weighting for the heat maps was implemented in two stages to keep the final “readiness” score on the same 1-5 scale while still reflecting SCTE’s priorities. First, each feasibility category score was constructed from its underlying sub-categories by taking a simple average: Technical Feasibility was the mean of tech maturity, maintenance, and reliability; Economic Feasibility was the mean of CAPEX and OPEX; Environmental Benefit was represented by the emissions score; and Policy & Incentives was represented solely based on the availability and rigor of the various policies. This produced a 1–5 category score for each technology in each state (e.g., a technical score of 3.7 in PA would reflect the average maturity/maintenance/reliability assessments for that technology in PA). Figure 8 below shows the scores for the seven underlying categories in an unaggregated format across the three case study states.


Feasibility Scores

Feasibility Score - SOLAR

Feasibility Score - BIOGAS

Feasibility Score - PEMFC

Feasibility Score - GREEN H2

Feasibility Score - Zn-Br

Feasibility Score - Li-ion

Source: Renewable Power Recipes for Powering Telecommunication Hub sites • SCTE-RISE Student Team, Villanova University

Figure 35 - Feasibility Scores of Renewable Power and BESS Options for Hub Sites in Pennsylvania, New York, and Arizona

Next, the seven scores were grouped back into the four overall categories and then all four category scores were combined into a single overall readiness value using the agreed weights--35% technical, 30% economic, 20% environmental, and 15% policy–so the total is a weighted average (not a sum) and therefore remains interpretable on the original 1-5 scale. One exception was applied for the PEMFC option: because PEM fuel cells are intrinsically dependent on hydrogen as their fuel, the underlying sub-category scores were assessed for both PEMFC and green hydrogen and then averaged together before forming the category-level scores and applying the feasibility weights, ensuring the heat-map result reflects the combined practicality of “PEMFC + hydrogen supply” rather than treating the fuel cell in isolation.

The same scoring system was used again to assess the two different BESS battery options. Since BESS systems focus more on backup power storage rather than generation, they were scored separately from the three power generation options to avoid skewing the heat map.

Power Generation Heat Map

From the heat map on power generation systems, several trends can be seen. The first is that across the board, solar appears as the best option due to high-tech maturity, low maintenance, and low operating cost. However, its reliability will always be an issue, and from Table 2, solar will never be able to fully power the hub site, making it more of a supplementary power system. Biogas has the largest fluctuations across the state due to its reliance on specific feed stocks to achieve the power required for a 10 kW hub site. Sourcing the ideal feedstock will be reliant not only on if the hub site is in proximity to a farm that can provide but also on what type of feedstock is being produced at that farm. Additionally, biogas digesters require special maintenance over their life time to maintain, which is another hindrance when compared to a low-maintenance technology like solar. Fuel cells + greenhydrogen lands somewhere in the middle due to each part of this technology being reliant on the other. While PEMFCs are a known quantity and would be effective at generating the amount of power needed for the hub site, the sourcing of green hydrogen may prove tricky. This is due to robust green hydrogen supply systems not existing in the three case study states, bringing down the effectiveness of the combination of fuel cells + green hydrogen. emissions. Another challenge is that both fuel cells and green hydrogen have high capital and operational expenses. Where this combination technology shines is in its drastic reduction in emissions by burning a clean fuel like hydrogen produced from renewable energy.

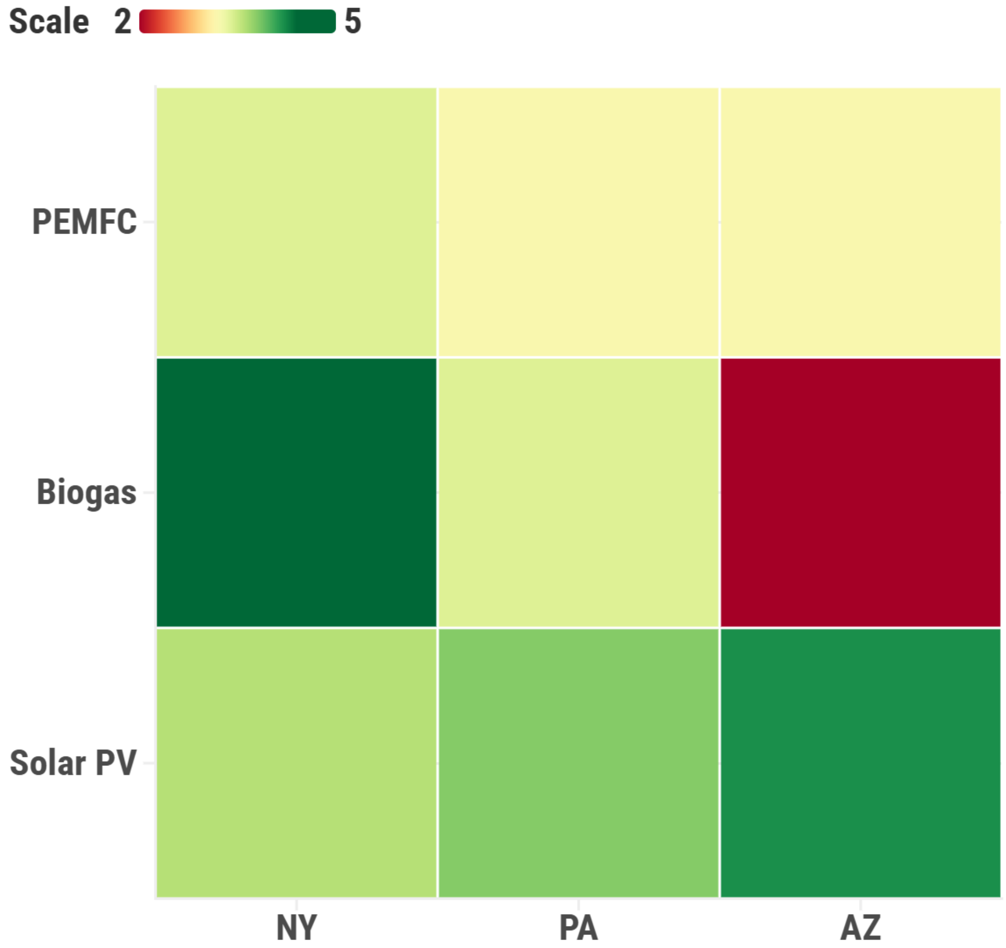


Figure 36 - Feasibility Heat Map of Renewable Power Generation Sources for Hub Sites in New York, Pennsylvania, and Arizona

BESS Heat Map

When looking at the heat for the two BEES options in Figure 10, the more desirable option is the zinc bromine batteries across all three states. A driving factor in this is their low chance of experiencing thermal runaway and causing a fire when compared to lithium-ion batteries. Due to how critical it is for the hub sites to stay up and running, any potential risk for a loss of uptime, including from a fire, was scored poorly. Furthermore, when a hub site completely switches over to batteries as the sole source of power, the BESS only focuses on maintaining power to the server racks and not to the cooling system, which further increases the risk of a fire with lithium-ion batteries. This is especially important to consider in a climate like Arizona, where temperatures are known to exceed 115 degrees Fahrenheit at some points of the year [67].

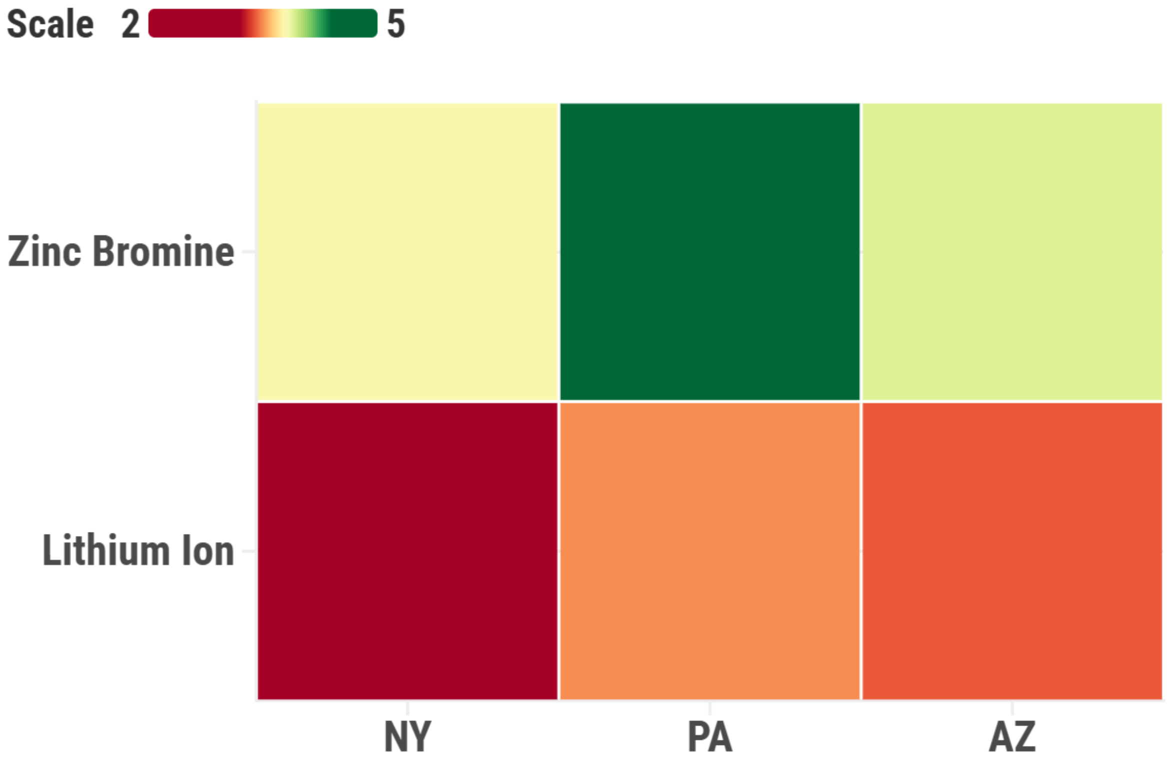


Figure 37 - Feasibility Heat Map for Lithium Ion and Zinc Bromine BESS for a Hub Site in New York, Pennsylvania, and Arizona

Another key factor was the fact that the production of zinc bromine batteries can be achieved utilizing an existing U.S. based supply chain for many of the key components. Whereas with lithium-ion batteries, the US is not a major producer and will be dependent on procuring lithium through imports. This could prove particularly challenging with the current U.S. tariffs. There were also more concerns with the safe recycling and reuse of lithium-ion batteries when compared with zinc bromine batteries. Pennsylvania stands as the best option for testing the implementation of a zinc bromine BESS due to the amount of state-level funding given to creating a zinc bromine manufacturing facility within the state.

Conclusions

The assessment of various energy production and storage technologies yielded several key conclusions:

- Fuel Cells
  Fuel cells emerged as a modular, low-footprint option viable across all three case study states. While the technology demonstrated high resilience and reliability regardless of location, its primary limitation remains fuel sourcing. However, when paired with green

hydrogen, fuel cells offer a sustainable solution capable of meeting industrial energy demands at this scale.

- Solar Photovoltaics
  Solar electricity generation faces significant hurdles, primarily due to the relatively low solar irradiance in the selected states except for Arizona. Furthermore, the substantial spatial footprint required to generate sufficient power exceeds the available site parameters defined for this project.
- Biogas Systems
  While biogas systems proved robust and capable of meeting high energy loads, their feasibility is heavily dependent on the specific digester type and feedstock availability. Since different digesters require precise operating temperatures and specific organic inputs, implementation can be challenging in locations where these resources are inaccessible.
- Battery Storage
  Zinc-bromine batteries were identified as a more reliable alternative to lithium-ion counterparts. Despite having a lower energy density, zinc-bromine batteries exhibit minimal risk of thermal runaway, making them a particularly favorable option for deployment at hub sites.

Recommendations

1. Trial runs of fuel cells and biogas options at hub sites at locations with favorable conditions as discussed.

2. Continue integration of solar panels, which can help reduce overall emissions and reduce the reliance on existing power grids.

3. Investigate the use of zinc-bromine batteries in hub sites in Pennsylvania. Due to the proximity of a zinc bromine battery manufacturer, Pennsylvania represents the best state for implementation.

Next Steps

There are several next steps to expand upon the research of this project. The first is performing a comparative analysis of PEMFC vs SOFC. While SOFCs were discussed in this project, the evaluation for fuel cells was done on PEMFC due to their lower emissions and use of green hydrogen. SOFCs represent a more reliable option since they are fuel flexible and use either green hydrogen or traditional fossil fuels for power. However, this comes with the tradeoff of having more emissions when using fossil fuels, warranting further study. The next option for future work would be to scale up the existing technologies to work for facilities with higher power draws. While this project focused on Class D hub sites that were only 10 kW, these facilities can be up to 50 kW in size. Focus should be spent on how this will impact the effectiveness of the three power generation options, especially solar, as the maximum potential has already been reached within the 625 square foot allotment. Lastly, the existing power options should be evaluated in other high-density case study states, and further investigation should be conducted to determine if there are options for other renewable power sources to be used. While wind was not considered for the three case study states chosen, it could be a potential option in a state like Texas due to the favorable geographic location for wind turbines [10]. Other high population states that could be considered include California and Florida. These states come with new challenges as well in the form of extreme weather from hurricanes in Florida and the potential for earthquakes in California.

Abbreviations and Definitions

Abbreviations

SCTE	Society of Cable Telecommunications Engineers
BESS	Battery Energy Storage System
STEEP	Social, Technical, Economic, Environmental, Political
PEMFC	Proton Exchange Membrane Fuel Cell
SOFC	Solid Oxide Fuel Cell
NFPA	National Fire Protection Association
NREL	National Renewable Energy Laboratory
CHP	Combined Heat and Power

Bibliography and References

[1] M. Schipper and T. Hodge, “After more than a decade of little change, U.S. electricity consumption is rising again - U.S. Energy Information Administration (EIA),” Eia.gov, 2025. https://www.eia.gov/todayinenergy/detail.php?id=65264

[2] IER, “The Grid Reliability Crisis Is Here,” IER, Aug. 11, 2025. https://www.instituteforenergyresearch.org/the-grid/the-grid-reliability-crisis-is-here/ (accessed Dec. 14, 2025).

[3] “Energy Management Subcommittee SCTE STANDARD SCTE 226 2025 Cable Facility Classification Definition Specification.” Accessed: Dec. 14, 2025. [Online]. Available: https://wagtail-prod-storage.s3.amazonaws.com/documents/SCTE_226_2025_p4RNwb8.pdf

[4] Data Center Knowledge, “Data Center Power: Fueling the Digital Revolution,” *www.datacenterknowledge.com*, Mar. 22, 2024. https://www.datacenterknowledge.com/energy-power-supply/data-center-power-fueling-the-digital-revolution

[5] J. Roundy, “The increasing concern of data center land acquisition,” *Search Data Center*, 2025. https://www.techtarget.com/searchdatacenter/feature/The-increasing-concern-of-data-center-land-acquisition

[6] S. Kwon, “Ensuring renewable energy utilization with quality of service guarantee for energy-efficient data center operations,” *Applied Energy*, vol. 276, p. 115424, Oct. 2020, doi: https://doi.org/10.1016/j.apenergy.2020.115424.

[7] T. Yang, Y. Hou, Y. C. Lee, H. Ji, and A. Y. Zomaya, “Power Control Framework for Green Data Centers,” *IEEE Transactions on Cloud Computing*, vol. 10, no. 4, pp. 2876–2886, Oct. 2020, doi: https://doi.org/10.1109/tcc.2020.3022789.

[8]Y. Zhang, Y. Wang, and X. Wang, “GreenWare: Greening Cloud-Scale Data Centers to Maximize the Use of Renewable Energy,” *Middleware 2011*, pp. 143–164, 2011, doi: https://doi.org/10.1007/978-3-642-25821-3_8.

[9] SCTE Technical Journal, Vol. 5, No. 1, SCTE, 2025, p. 184. Available: https://wagtail-prod-storage.s3.amazonaws.com/documents/SCTE_Journal_V5N1.pdf

[10] U.S. Energy Information Administration, “Where Wind Power Is Harnessed - U.S. Energy Information Administration (EIA),” *Eia.gov*, Mar. 30, 2022. https://www.eia.gov/energyexplained/wind/where-wind-power-is-harnessed.php

[11] U.S. Energy Information Administration, “How many nuclear power plants are in the United States, and where are they located? - FAQ - U.S. Energy Information Administration (EIA),” *Eia.gov*, 2016. https://www.eia.gov/tools/faqs/faq.php?id=207&t=21

[12] J. C. Robins *et al.*, “2021 U.S. Geothermal Power Production and District Heating Market Report,” *www.osti.gov*, Jul. 13, 2021. https://www.osti.gov/biblio/1808679

[13] U.S. Energy Information Administration, “Where hydropower is generated - U.S. Energy Information Administration (EIA),” *Eia.gov*, Apr. 20, 2023. https://www.eia.gov/energyexplained/hydropower/where-hydropower-is-generated.php

[14] “Carbon capture for natural gas-fired power generation: An opportunity for hyperscalers | Carbon Direct,” *Carbon-direct.com*, 2025. https://www.carbon-direct.com/insights/carbon-capture-for-natural-gas-fired-power-generation-an-opportunity-for-hyperscalers

[15] “PVWatts Calculator,” *Nrel.gov*, 2019. https://pvwatts.nrel.gov/index.php

[16] N. A. A. Qasem and G. A. Q. Abdulrahman, “A Recent Comprehensive Review of Fuel Cells: History, Types, and Applications,” International Journal of Energy Research, vol. 2024, no. 1, p. 7271748, Jan. 2024, doi: 10.1155/2024/7271748.

[17] FCHEA, “Fuel Cell Basics,” FCHEA. Accessed: Dec. 06, 2025. [Online]. Available: https://fchea.org/learning-center/fuel-cell-basics/

[18] DoE, “DOE Technical Targets for Fuel Cell Systems for Stationary (Combined Heat and Power) Applications,” Energy.gov. Accessed: Dec. 05, 2025. [Online]. Available: https://www.energy.gov/eere/fuelcells/doe-technical-targets-fuel-cell-systems-stationary-combined-heat-and-power.

[19] M. T. Mehran et al., “A comprehensive review on durability improvement of solid oxide fuel cells for commercial stationary power generation systems,” Applied Energy, vol. 352, p. 121864, Dec. 2023, doi: 10.1016/j.apenergy.2023.121864.

[20] O. Corigliano, L. Pagnotta, and P. Fragiacomo, “On the Technology of Solid Oxide Fuel Cell (SOFC) Energy Systems for Stationary Power Generation: A Review,” Sustainability, vol. 14, no. 22, p. 15276, Nov. 2022, doi: 10.3390/su142215276.

[21] D. Roy, S. Samanta, S. Roy, A. Smallbone, and A. P. Roskilly, “Technoeconomic and environmental performance assessment of solid oxide fuel cell-based cogeneration system configurations,” Energy, vol. 310, p. 133145, Nov. 2024, doi: 10.1016/j.energy.2024.133145.

[22] DoE, “Quadrennial Technology Review 2015 Omnibus,” U.S. Department of Energy, 2015. Accessed: Dec. 06, 2025. [Online]. Available: https://www.energy.gov/quadrennial-technology-review-2015, ch. 4Q

[23] J. Kurtz, G. Saur, and S. Sprik, “Hydrogen Fuel Cell Performance as Telecommunications Backup Power in the United States,” NREL/TP--5400-60730, 1260144, Mar. 2015. doi: 10.2172/1260144.

[24] J. Kurtz, G. Saur, S. Sprik, and C. Ainscough, “Backup Power Cost of Ownership Analysis and Incumbent Technology Comparison,” NREL/TP-5400-60732, 1163435, Sept. 2014. doi: 10.2172/1163435.

[25] Z. Ma, J. Kurtz, and J. Eichman, “Fuel Cell Backup Power Unit Configuration and Electricity Cost in Telecommunication Applications,” NREL, Golden, CO, Tech. Rep. NREL/TP-5400-67408, 2017.

[26] Z. Ma, J. Eichman, and J. Kurtz, “Fuel Cell Backup Power System for Grid Service and Microgrid in Telecommunication Applications,” Journal of Energy Resources Technology, vol. 141, no. 6, p. 062002, June 2019, doi: 10.1115/1.4042402.

[27] J. Marqusee and A. Stringer, “Distributed Energy Resource (DER) Reliability for Backup Electric Power Systems,” NREL/TP--7A40-83132, 1964053, MainId:83905, Mar. 2023. doi: 10.2172/1964053.

[28] C. W. Blake and C. H. Rivkin, “Stationary Fuel Cell Application Codes and Standards: Overview and Gap Analysis,” NREL/TP-560-49165, 990103, Sept. 2010. doi: 10.2172/990103.

[29] NREL, “Stationary and Portable Fuel Cell Systems Codes and Standards Citations (Brochure).” 2013. [Online]. Available: https://www.nrel.gov/docs/fy14osti/57944.pdf

[30] R. Krumm and J. Wiley, “NFPA 2 – v2023 Hydrogen Facilities,” presented at the PowerGen 2024, New Orleans, LA, Jan. 23, 2024.

[31] NFPA, “NFPA 853 Installation of Stationary Fuel Cell Power Plants.” Accessed: Dec. 06, 2025. [Online]. Available: https://h2tools.org/fuel-cell-codes-and-standards/nfpa-853-installation-stationary-fuel-cell-power-plants.

[32] NYSERDA, “Renewable Portfolio Standard,” Department of Public Service. Accessed: Dec. 06, 2025. [Online]. Available: https://dps.ny.gov/renewable-portfolio-standard.

[33] NYSERDA, “Current Solicitations and Funding Opportunities, PONs, RFPs, RFQs, and RFQLs,” NYSERDA. Accessed: Dec. 05, 2025. [Online]. Available: https://www.nyserda.ny.gov/Funding-Opportunities/Current-Funding-Opportunities.

[34] NYSERDA, “$3.7 Million Is Available To Support Clean Hydrogen-Based Fuel Cell Resources In New York,” NYSERDA. Accessed: Dec. 09, 2025. [Online]. Available: https://www.nyserda.ny.gov/About/Newsroom/2025-Announcements/2025-06-09-NYSERDA-Announces-3-Million-Available-To-Support-Clean-Hydrogen-Resources

[35] PAPUC, AEPS Act, vol. 1672, No. 213. 2004, p. 12. Accessed: Dec. 06, 2025. [Online]. Available: https://www.puc.pa.gov/filing-resources/issues-laws-regulations/aeps-act/

[36] PAPUC, “Alternative Energy Resource Types & Tiers,” Pennsylvania Alternative Energy Portfolio Standard Program. Accessed: Dec. 06, 2025. [Online]. Available: https://pennaeps.com/alternative-energy-resource-types-tiers/

[37] Pennsylvania Department of Community and Economic Development, “Alternative and Clean Energy (ACE) Program Guidelines,” 2016 (and updates).

[38] DoE, “Alternative Fuels Data Center: Hydrogen Fuel Production Support.” Accessed: Dec. 09, 2025. [Online]. Available: https://afdc.energy.gov/laws/13021

[39] Arizona Corporation Commission, “Renewable Energy Standard and Tariff | Arizona Corporation Commission,” Arizona Corporation Commission. Accessed: Dec. 09, 2025. [Online]. Available: https://www.azcc.gov/utilities/industry-types/electric/renewable-energy-standard-and-tariff

[40] Arizona Commerce Authority, “Renewable Energy Tax Incentive Program (RETIP),” program overview, 2023.

[41] UpCodes, “Chapter 12 Energy Systems: Arizona Fire Code 2018.” Accessed: Dec. 09, 2025. [Online]. Available: https://up.codes/viewer/arizona/ifc-2018/chapter/12/energy-systems#1205

[42] EIA, “Average Price of Electricity to Ultimate Customers by End-Use Sector Table 2.10.,.” [Online]. Available: https://www.eia.gov/electricity/annual/html/epa_02_10.html

[43] Clean Air Task Force, “Regulatory Framework for Hydrogen in the U.S.,” Clean Air Task Force. Accessed: Dec. 09, 2025. [Online]. Available: https://www.catf.us/resource/regulatory-framework-hydrogen-us/

[44] IEA, “The future of hydrogen,” *IEA*, Jun. 2019. https://www.iea.org/reports/the-future-of-hydrogen

[45] IRENA, “Hydrogen,” *www.irena.org*, 2022. https://www.irena.org/Energy-Transition/Technology/Hydrogen

[46] National Grid, “The hydrogen colour spectrum | National Grid Group,” *www.nationalgrid.com*, Feb. 23, 2023. https://www.nationalgrid.com/stories/energy-explained/hydrogen-colour-spectrum

[47] U.S. Department of Energy, “Hydrogen Production: Electrolysis,” *Energy.gov*, 2024. https://www.energy.gov/eere/fuelcells/hydrogen-production-electrolysis

[48] D. Hfto and A. Gilbert, “Campbell Howe (DOE Loan Program Office) Peer Reviewed by: Neha Rustagi,” Strategic Analysis, Inc, 2024. Available: https://www.hydrogen.energy.gov/docs/hydrogenprogramlibraries/pdfs/24005-clean-hydrogen-production-cost-pem-electrolyzer.pdf

[49] IRENA, “GREEN HYDROGEN: A GUIDE TO POLICY MAKING,” 2020. Available: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2020/Nov/IRENA_Green_hydrogen_policy_2020.pdf

[50] U.S. Department of Agriculture, U.S. Environmental Protection Agency, U.S. Department of Energy, “Biogas Opportunities Roadmap: Voluntary Actions to Reduce Methane Emissions and Increase Energy Independence,” August 2014

[51] U.S. Environmental Protection Agency, “How Does Anaerobic Digestion Work?”. Accessed: Dec. 15, 2025. [Online]. Available: https://www.epa.gov/agstar/how-does-anaerobic-digestion-work

[52] U.S. Environmental Protection Agency, “Sewage Sludge Laws and Regulations”. Accessed: Dec. 15, 2025. [Online]. Available: https://www.epa.gov/biosolids/sewage-sludge-laws-and-regulations

[53] U.S. Environmental Protection Agency, “Types of Anaerobic Digesters”. Accessed: Dec. 15, 2025. [Online]. Available: https://www.epa.gov/anaerobic-digestion/types-anaerobic-digesters

[54] Aruchamy D, Sharma H, Chinthakunta N, KN R, Chengappa S, Sundararam S., “Biogas Powered Data Centers: A Study”, Green IT: Go Green for Sustenance.:57

[55] AG-GRID Energy “Lent Hill AG-GRID DIGITAL”. Accessed: Dec. 15, 2025. [Online]. Available: https://aggridenergy.com/project/lent-hill-digital/

[56] K. Araujo, “Battery Comparison of Energy Density - Cylindrical and Prismatic Cells,” *Epectec.com*, 2020. https://www.epectec.com/batteries/cell-comparison.html

[57] University of Washington, “Lithium-Ion Battery,” *Clean Energy Institute*, 2020. https://www.cei.washington.edu/research/energy-storage/lithium-ion-battery/

[58] “Topic: Lithium-ion battery industry worldwide,” *Statista*, 2025. https://www.statista.com/topics/2049/lithium-ion-battery-industry/?srsltid=AfmBOoqeuD3HTEscw9wpUO9lXBJ2TFjAyD9alqMwHRdolsJRCKagCGPi#topicOverview (accessed Dec. 15, 2025).

[59] “Most utility-scale batteries in the United States are made of lithium-ion - Today in Energy - U.S. Energy Information Administration (EIA),” *www.eia.gov*. https://www.eia.gov/todayinenergy/detail.php?id=41813

[60] “What Causes Thermal Runaway? - UL Research Institutes,” *UL Research Institutes*, Aug. 20, 2021. https://ul.org/research-updates/what-causes-thermal-runaway/

[61] “How Do You Put Out A Lithium-Ion Fire?,” *Thompson Safety*, 2025. https://thompson-safety.com/articles/lithium-ion-battery-fire (accessed Dec. 15, 2025).

[62] A. Mahmood, Z. Zheng, and Y. Chen, “Zinc–Bromine Batteries: Challenges, Prospective Solutions, and Future,” *Advanced Science*, vol. 11, no. 3, Nov. 2023, doi: https://doi.org/10.1002/advs.202305561.

[63] Eos , “Eos Z3 Zinc-powered aqueous liquid battery module,” Eos Energy. Accessed: Dec. 15, 2025. [Online]. Available: https://www.eose.com/wp-content/uploads/2023/05/eos_productsheet_Z3_050223.pdf

[64] “A High-Performance Aqueous Zinc-Bromine Static Battery,” *iScience*, vol. 23, no. 8, p. 101348, Aug. 2020, doi: https://doi.org/10.1016/j.isci.2020.101348.

[65] “Technology - Eos Energy Enterprises,” *Eos Energy Enterprises*, Aug. 14, 2024. https://www.eose.com/technology/ (accessed Dec. 15, 2025).

[66] “Flow v. Lithium-Ion Batteries for Energy Storage,” *Battery Technology*, 2021. https://www.batterytechonline.com/batteries/flow-vs-lithium-ion-batteries-for-energy-storage (accessed Dec. 15, 2025).

[67] N. US Department of Commerce, “2021 Climate Year in Review for Phoenix, Yuma, and El Centro,” *www.weather.gov*. https://www.weather.gov/psr/yearinreview2021

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Maximizing Engagement Through QR Codes In Video Advertising:
Innovative Solutions for Enhanced Viewer Interaction

A Technical Paper prepared for SCTE by

Ramachandran Elumalai, Principal Software Engineer II, Charter Communications
6360 S Fiddlers Green Cir, Greenwood Village, CO 80111
ramachandran.elumalai@charter.com

Vipul Patel, Vice President, Charter Communications
14810 Grasslands Dr, Englewood, CO 80112
vipul.patel@charter.com

Table of Contents

List of Figures

List of Tables

Abstract

The rapid adoption of QR codes during the pandemic, with marketing and advertising industries experiencing a 323% increase in QR code usage from 2021 to 2024 [2], has created significant opportunities for video advertising integration. This paper presents a technical framework for integrating QR codes within video advertising using MPEG DASH, focusing on four implementation methods:

1. displaying QR codes in the final seconds of video ads,
2. optimizing them for smaller ad windows with enhanced audio features,
3. presenting them on pause screens in video-on-demand platforms, and
4. delivering location-aware dynamic QR Codes.

A novel approach converts QR codes with sponsor imagery and audio into 1 frame-per-second (fps) video segments, achieving approximately 90% reduction in bandwidth compared to standard 30fps content while maintaining seamless playback. The system architecture leverages Kafka-based microservices, containerized deployment on Kubernetes, and cloud storage to enable scalable, personalized QR code generation. Analysis of viewer engagement metrics indicates that these QR code integrations can significantly enhance viewer interaction [1]. This research provides cable operators and content providers with actionable implementation guidelines for bridging video advertising with digital engagement platforms.

Quick Reference Guide

This paper uses a few streaming terms that may be new if you don't work in video delivery. Here's a quick guide.

- **QR code in video**: A scannable code shown on-screen that sends a viewer to a web page (for example, a product page or coupon).
- **MPEG-DASH (DASH)**: A common way to deliver streaming video. Instead of one big video file, the video is delivered as many small pieces as possible.
- **Segments**: The small video files the player downloads one after another (often a few seconds each).
- **Manifest / MPD**: A small "playlist" file (called an **MPD**) (Media Presentation Description) that tells the player which segments to download and when.
- **Period**: A section of the stream. In this paper, one period can be "normal video," and another period can be "QR code display."
- **Frame rate (fps)**: How many images are shown each second. Normal video is often **30 fps**. A QR code is mostly static, so showing it at **1 fps** can still work while using less data.
- **Bitrate**: How much data is delivered per second. Lower bitrate usually means smaller files and less network usage (but too low can reduce visual quality).
- **Player**: The software (app, TV, browser) that reads the manifest and plays the video.
- **Server / CDN**: Systems that store video segments and deliver them quickly to viewers.
- **Bandwidth:** The amount of data transferred over time. Reducing bandwidth saves network costs and improves playback on slower connections.
- **Why 1 fps matters here**: During the QR-code window, the picture doesn't need to change often. Delivering fewer frames can reduce delivered data while keeping the QR code readable long enough to scan.

Detailed Implementation of Use Cases

This section presents four distinct use cases for integrating QR codes into video advertising workflows, each addressing specific viewer contexts and technical requirements. These implementations leverage the 1fps video technique to achieve significant bandwidth reduction while maintaining QR code scannability. Each use case includes the business scenario, technical solution, step-by-step implementation guidance, MPD configuration approach, and measurable benefits.

Use Case 1: End-of-Video QR Code Integration

Scenario

Video advertisers need to increase engagement and measurability of traditional 30-second TV commercials and digital video ads.

Solution

This solution involves seamlessly switching to a very low frame rate (1fps) video segment during the last 5-8 seconds of the ad. Frame rate refers to how many images (frames) are displayed per second—standard video uses 30 fps to create smooth motion. However, since a QR code is a static image that doesn't require motion, encoding it at just 1 fps (one frame per second) dramatically reduces the amount of data transmitted while keeping the QR code visible and scannable for the full duration. During this period, the QR code is embedded directly into each video frame, ensuring it remains visible and synchronized with the video timeline. The entire process maintains a single, continuous video stream using multi-period DASH.

Implementation Steps

- ***Video Encoding***

The dual period encoding approach optimizes bandwidth utilization while maintaining compatibility with standard DASH players and existing cable infrastructure.

- First 22 seconds: Encode at standard 30fps, requiring bandwidth of approximately 1-3 Mbps for video + 128 kbps for audio
- Final 8 seconds (23s to 30s): Encode at 1fps, reducing video bandwidth to approximately 150-300 kbps (85-92% reduction)
- Critical: Use identical codec parameters (H.264 baseline profile, level 4.0) for both periods to ensure compatibility

- ***Segment Structure***

The segment architecture ensures seamless period transitions and optimal cache efficiency across content Delivery Networks (CDN).

- Normal Period: 11 segments x 2 seconds each = 22 seconds total
- QR Period: 4 segments x 2 seconds each = 8 seconds total

- Shared Initialization: Both periods must use identical initialization segments for MSE compatibility

- ***Frame Rate Transition***

The DASH player transitions between 30fps and 1fps segments at the 22-second mark. This is managed through multi-period MPD structure with shared initialization segments.

- ***Audio CTA***

At the 22-second mark, audio continues seamlessly with optional call-to-action overlay. Audio maintains consistent 128 kbps throughout both periods. For example, a phrase like “Scan now for your discount” can effectively prompt viewer action.

MPD Configuration

The MPD file is crucial for defining how these different video segments are presented. Refer the section 6.1 for the sample MPD structure related to this use case.

Benefits

This approach offers significant advantages, including an approximate 92% reduction in bandwidth usage during the QR code display, perfect synchronization between the video and the QR code, and full compatibility with standard DASH players.

Use Case 2: Small-Format Video Ad QR Code Integration

Scenario

This use case focuses on optimizing QR code integration for smaller ad windows with enhanced audio features and picture-in-picture advertisements where screen real estate is limited.

Solution

The entire 15-second ad runs at 1fps with centrally positioned, oversized QR codes designed for small viewing windows, combined with enhanced audio cues to drive engagement when visual space is constrained.

Implementation Steps

- ***Video Encoding***

Encode the 15-second ad at 1fps for ultra-low bandwidth consumption. QR codes are embedded centrally in every frame, sized to ensure optimal readability in constrained picture-in-picture.

- ***Design Optimization***

Implement high-contrast QR codes with medium-level error correction positioned against simplified backgrounds with minimal visual distractions. This design approach ensures maximum readability and scannability even when the ad window is significantly reduced in size compared to traditional video advertisements.

- ***Audio Enhancement***

Layer multiple audio elements including persistent cues ("Scan now for exclusive offers"), directional prompts ("Look for the QR code"), company jingles for brand recognition, and urgency messaging ("Limited time offer") throughout the entire 15-second duration to drive engagement when visual attention may be divided across multiple screen elements.

MPD Configuration

The MPD (Media Presentation Description) file is crucial for defining how these different video segments are presented. Refer the section 6.2 for the sample MPD structure related to this use case.

Benefits

This optimized approach delivers several key benefits, including ultra-low bandwidth consumption (approximately 90% reduction vs standard ads), persistent QR code visibility throughout the entire 15-second ad duration, enhanced audio-driven engagement to compensate for limited visual space, and compatibility with picture-in-picture advertising formats.

Use Case 3: VOD Pause Screen QR Code Integration

Scenario

Video-on-demand (VOD) viewers frequently pause content, creating unique engagement opportunities that traditional advertising approaches fail to leverage effectively.

Solution

This solution relies on the DASH player to detect when a user pauses the video. Upon detection, the player plays the ad stream, which is a 1fps ad stream with QR code and ad sponsor image.

Implementation Steps

- ***Pause Detection***

Implement JavaScript within the DASH player to accurately detect and respond to pause events initiated by the user. Refer section 6.3.1 for more info on player changes.

- ***Play Ad Stream***

During pause, the DASH player plays the ad stream, which is a 1fps ad stream with QR code and ad sponsor image.

- ***MPD Configuration***

The MPD (Media Presentation Description) file is crucial for defining how these different video segments are presented. Refer the section 6.3.2 for the sample MPD structure related to this use case.

Benefits

This approach provides an effective way to monetize VOD content during pause states.

Use Case 4: Location-Aware Dynamic QR Codes in Video Advertising

Scenario

This advanced use case aims to deliver personalized QR codes to viewers based on their geographical location, enabling highly targeted and relevant engagement.

Solution

This solution involves server-side processing to generate unique, location-specific QR codes as video in 1fps.

Implementation Steps

- ***Location Detection***

The first step is to accurately identify the viewer's location. This can be achieved using various methods, such as IP geolocation or by utilizing device location APIs (with user consent).

- ***QR Generation***

Once the location is determined, the system generates QR codes that contain location-specific data. For example, a QR code might direct a user to the nearest retail store or provide information relevant to their local area.

- ***Dynamic MPD***

A unique MPD file is created for each viewer. This MPD is dynamically generated to include the location-specific QR code segments, ensuring that the correct personalized content is delivered.

- ***CDN Caching***

To ensure fast delivery and reduce server load, the generated QR code 1fps videos are cached on a CDN. This allows for rapid retrieval by viewers, regardless of their geographical proximity to the origin server.

- ***Privacy Compliance***

Given that location data is sensitive personal information, strict privacy compliance measures must be in place. Privacy requirements are state-specific in the United States, and operators must comply with applicable state privacy regulations that classify geolocation data as protected personal information. Cable operators can leverage existing privacy controls already deployed for targeted advertising to manage location-based QR code consent and preferences.

- ### *MPD Example*

The MPD structure for location-aware ads would include representations that point to the dynamically generated, location-specific QR code segments. Refer the section 6.4 for the sample MPD structure related to this use case.

Benefits

This use case demonstrates how video advertisers can leverage location data to deliver highly relevant, personalized experiences that significantly improve conversion metrics.

Technical Architecture

This architecture presents a scalable, distributed system engineered to produce dynamic QR code videos at 1fps, achieving approximately 90% reduction in bandwidth compared to standard 30fps video content. The system leverages a combination of real-time data processing, containerized microservices, and cloud storage to deliver personalized video content efficiently with integrated sponsor imagery and audio.

Components and Workflow

Client

The client component serves as the primary interface for QR code video generation workflows. It validates input parameters including QR data format, video duration constraints, and sponsor image specifications, then generates unique UUIDs for request tracking and resource isolation across concurrent operations.

- Initiates the process by providing QR data (e.g., URL, duration, audio settings, UUID) and sponsor image.
- Publishes requests to the qr-gen topic and listens to the output topic for results.
- Queries the S3 bucket (/qr-gen/{UUID}/video-gen) to retrieve the final 1fps video.

Kafka with Multiple Topics

Serves as the messaging backbone, handling multiple topics including:

- qr-gen: Initial request topic for QR code generation.
- image-merge: Topic for merging sponsor images with QR codes.
- video-gen: Topic for video generation tasks.
- output: Final output topic where processed video data is published.

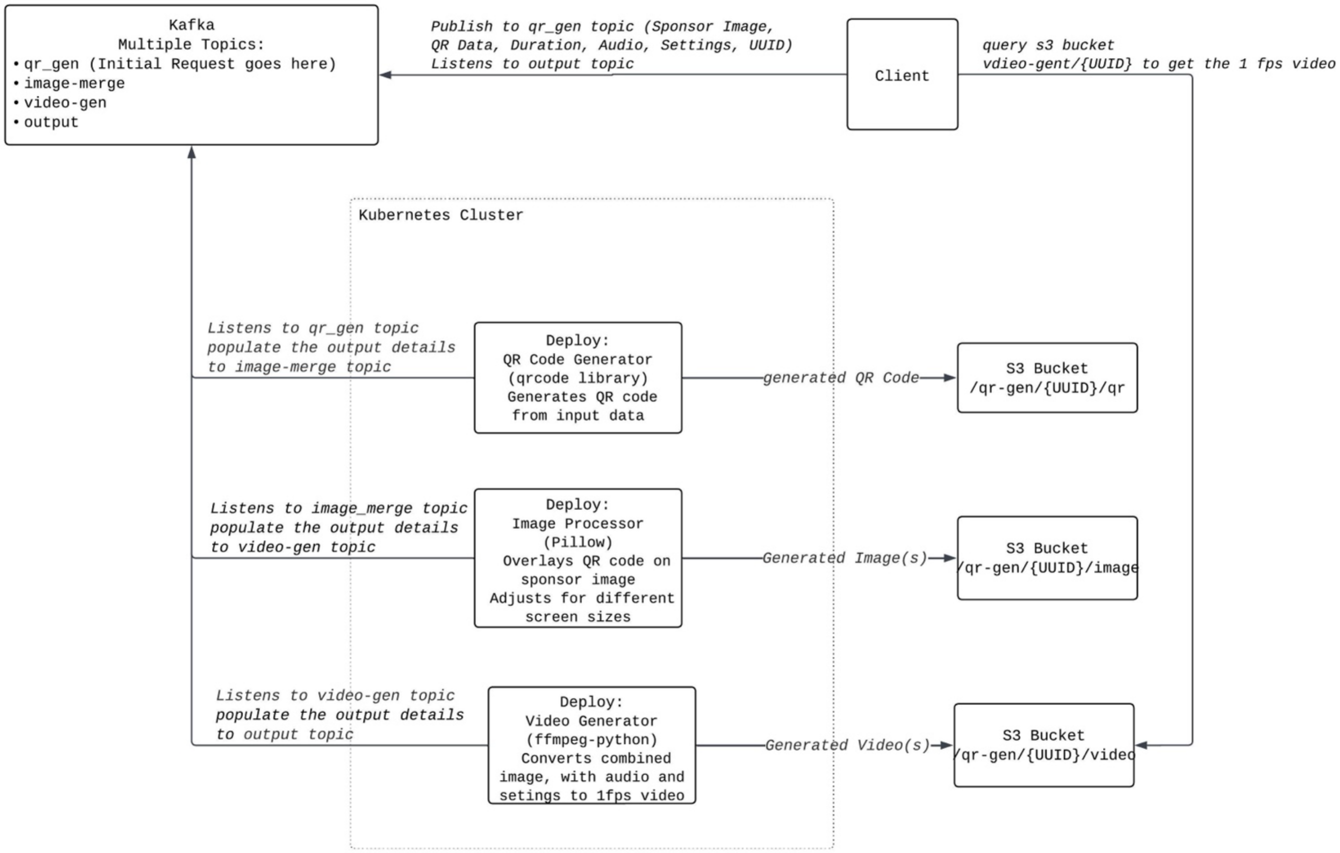


Figure 38: System Architecture Diagram

S3 Buckets

Table 12: S3 Bucket Unified Storage Solution Design

S3 Bucket Path	Description
/qr-gen/{UUID}/qr	Stores generated QR code(s) and listens to the qr-gen topic to populate.
/qr-gen/{UUID}/image	Holds generated image(s) (QR code + sponsor image).
/qr-gen/{UUID}/video	Contains generated video(s)

Kubernetes Cluster

Manages deployment and orchestration of microservices. Ensures high availability and fault tolerance.

Deployed Services

Table 13: Kubernetes Service Deployment details

Service Name	**Description**
QR Code Generator	Generates QR codes from input QR data and publishes to the qr-gen topic.
Image Processor	Adjusts sponsor images and overlays QR codes, adapting to different screen sizes. Publishes results to the image-merge topic.
Video Generator	Combines the merged image with audio and settings to produce a 1fps video, reducing size by 92%. Publishes the final video to the video-gen topic and the output topic.

Data Flow

The system implements an event-driven architecture with asynchronous processing stages, ensuring scalability and fault tolerance as illustrated in Figure 1. The workflow follows a sequential message-passing pattern through Kafka topics:

- ***Initial Request Processing***

- The client sends a request with QR data and sponsor image to the qr-gen topic.
- Client simultaneously listens to the output topic for completion notifications.

- ***QR Code Generation Stage***

- The QR Code Generator listens to the qr-gen topic and processes requests.
- Generates QR code using the open source qr-generator library.
- Stores generated QR code in S3 path: /qr-gen/{UUID}/qr
- Publishes completion event to trigger the next stage.

- ***Image Composition Stage***

- The Image Processor retrieves the QR code and sponsor image from S3.
- Merges images using Python's Pillow library with screen size adjustments.
- Stores composite image in S3 path: /qr-gen/{UUID}/image
- Publishes to image-merge topic to signal completion.

- ***Video Generation Stage***

- The Video Generator fetches the merged image and integrates audio settings.
- Generates 1fps video using Python ffmpeg library, achieving 89% size reduction.
- Stores final video in S3 path: /qr-gen/{UUID}/video
- Publishes completion notification to both video-gen and output topics.

- ***Client Retrieval***

- Client receives notification from output topic.
- Queries S3 path /qr-gen/{UUID}/video to access the final 1fps video.

This event-driven design ensures horizontal scalability, fault tolerance through Kafka's retry mechanisms, and UUID-based request isolation for concurrent processing.

Validation and Implementation Results

The proposed architecture was validated through a working proof of concept demonstration using a live DASH implementation. The POC demonstrates seamless transition from 30fps normal video content to 1fps QR code segments.

- ***Video Segment Analysis:***

- 30fps video segments (normal content): Average 1,588 KB per 2-second segment
- 1fps video segments (QR code display): Average 174 KB per 2-second segment
- Measured bandwidth reduction: 89.0%
- Segment duration consistency: 2 seconds maintained across all segments
- QR code scan success rate: 100% across tested devices and browsers

- ***Technical Validation:***

- DASH manifest transitions: Seamless period switching without buffering interruptions
- Player compatibility: Successful playback in Chrome, Safari, and Firefox
- Audio continuity: Uninterrupted audio track throughout video and QR segments

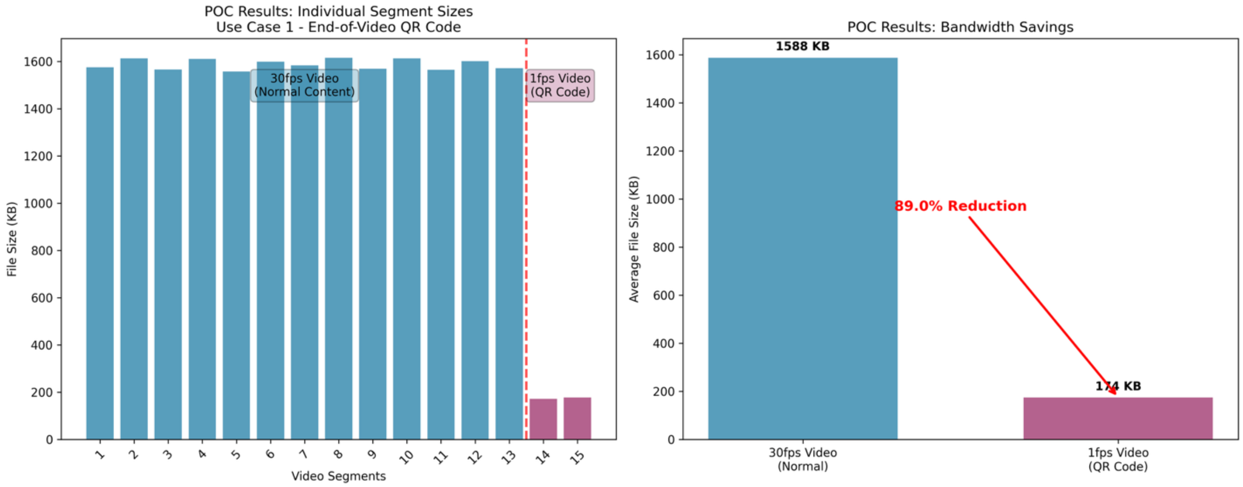


Figure 39: POC Bandwidth Savings Results

The measured results confirm the theoretical bandwidth reduction during QR code display periods, demonstrating the technical feasibility and performance benefits of this innovative approach. This modular architecture provides a robust framework for generating personalized QR code videos at 1fps, making it suitable for marketing, advertising, and interactive media applications with easy integration capabilities for additional features as needed.

User Interaction Data Collection

The QR code video advertising system enables comprehensive user engagement tracking through multiple data collection mechanisms.

QR Code Analytics:

- Unique tracking parameters embedded in QR codes capture scan events, timestamps, and device information.
- Landing page analytics measure post-scan user behavior and conversion rates.

Video Player Integration:

- DASH player events track QR code display timing and user interaction patterns.
- Engagement metrics include scan-to-appearance ratios and viewing completion rates.

Privacy Compliance:

- User consent management for data collection.
- Anonymized analytics while preserving campaign effectiveness measurement.
- Regulatory compliant data handling and retention policies.

This framework provides advertisers with measurable ROI data while maintaining privacy standards, enabling optimization of QR code placement, timing, and campaign performance.

Conclusion

This paper has presented a comprehensive architecture for integrating QR codes into video advertising, leveraging the capabilities of DASH/MPD technologies. Four distinct use cases have been detailed, providing in-depth implementation steps for each. These detailed examples offer a robust starting point for advertisers and developers aiming to create highly efficient and engaging video advertising campaigns.

The POC implementation confirmed the 1fps video technique achieves 89% bandwidth reduction during QR code display segments, seamless integration with existing MPEG DASH infrastructure, and successful QR code scanning across multiple device types. These results demonstrate that the proposed solution is not only theoretically sound but also practically implementable in real-world video advertising environments.

This novel approach to frame rate optimization for static content delivery represents a significant advancement in bandwidth-efficient video advertising, with potential applications extending beyond QR codes to other static overlay scenarios in streaming media.

Abbreviations and Definitions

Abbreviations

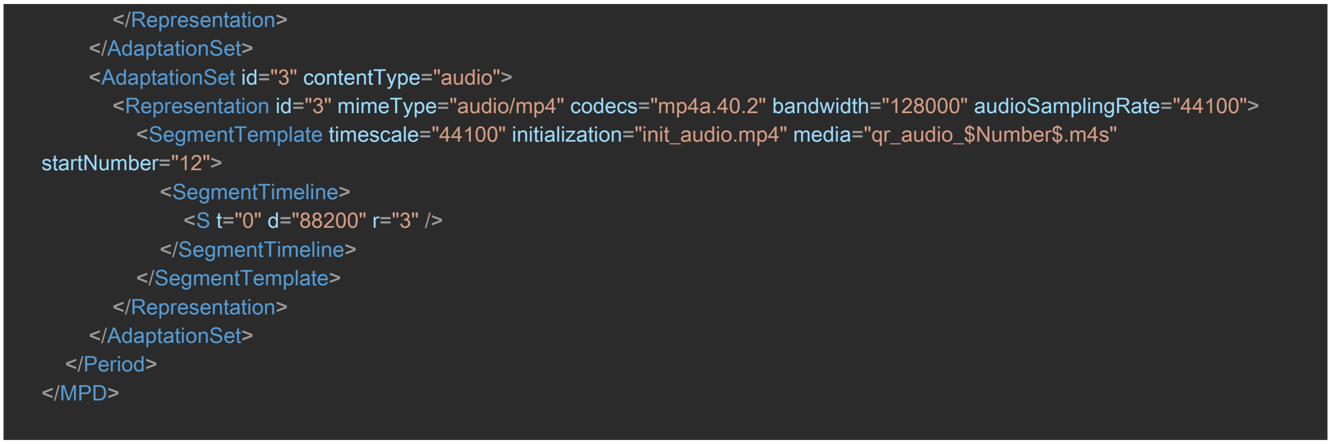
QR	Quick Response (code)
fps	frames per second
S3	Simple Storage Service
UUID	Universally Unique Identifier

MPD	Media Presentation Description
TV	Television
DASH	Dynamic Adaptive Streaming over HTTP
VOD	Video on Demand
CDN	Content Delivery Network
ffmpeg	Fast Forward Moving Picture Experts Group
POC	Proof of Concept
MPEG	Moving Picture Experts Group
API	Application Programming Interface
CTA	Call To Action
MSE	Media Source Extensions
ROI	Return On Investment
H.264	Advanced Video Coding standard

Sample MPD References

Use Case 1: End-of-Video QR Code Integration

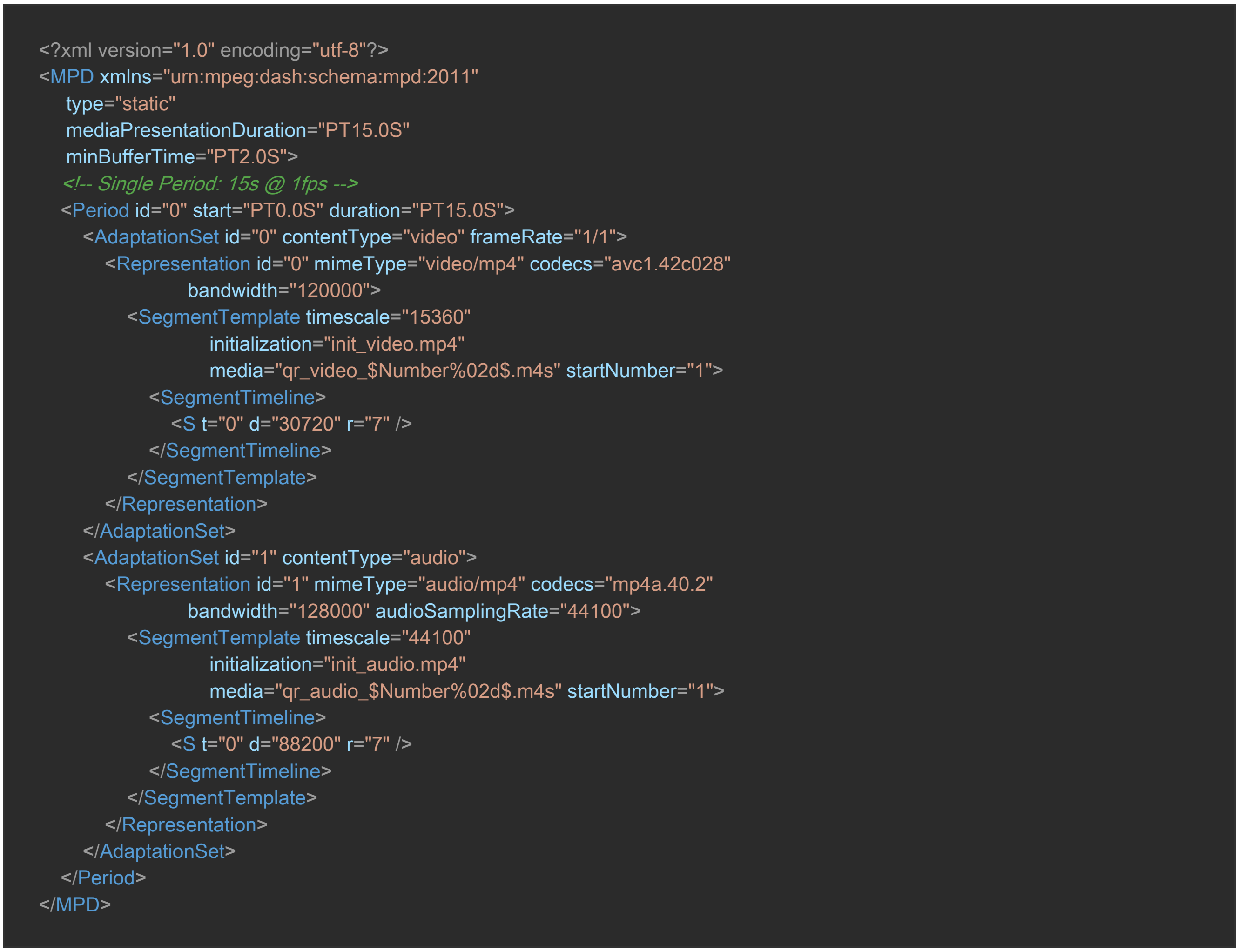
```
<?xml version="1.0" encoding="utf-8"?>
<MPD
  xmlns="urn:mpeg:dash:schema:mpd:2011" type="static" mediaPresentationDuration="PT30.0S" minBufferTime="PT4.0S">
  <Period id="0" start="PT0.0S" duration="PT22.0S"> <!-- Normal Period: 22s @ 30fps -->
    <AdaptationSet id="0" contentType="video" frameRate="30/1" maxWidth="1920" maxHeight="1080">
      <Representation id="0" mimeType="video/mp4" codecs="avc1.42c028" bandwidth="2000000" width="1920" height="1080">
        <SegmentTemplate timescale="15360" initialization="init_video.mp4" media="normal_video_$Number%02d$.m4s"
startNumber="1">
          <SegmentTimeline>
            <S t="0" d="30720" r="10" />
          </SegmentTimeline>
        </SegmentTemplate>
      </Representation>
    </AdaptationSet>
    <AdaptationSet id="1" contentType="audio">
      <Representation id="1" mimeType="audio/mp4" codecs="mp4a.40.2" bandwidth="128000" audioSamplingRate="44100">
        <SegmentTemplate timescale="44100" initialization="init_audio.mp4" media="normal_audio_$Number%02d$.m4s"
startNumber="1">
          <SegmentTimeline>
            <S t="0" d="88200" r="10" />
          </SegmentTimeline>
        </SegmentTemplate>
      </Representation>
    </AdaptationSet>
  </Period>
  <Period id="1" start="PT22.0S" duration="PT8.0S"> <!-- QR Period: 8s @ 1fps -->
    <AdaptationSet id="2" contentType="video" frameRate="1/1" maxWidth="1920" maxHeight="1080">
      <Representation id="2" mimeType="video/mp4" codecs="avc1.42c028" bandwidth="200000" width="1920" height="1080">
        <SegmentTemplate timescale="15360" initialization="init_video.mp4" media="qr_video_$Number$.m4s"
startNumber="12">
          <SegmentTimeline>
            <S t="0" d="30720" r="3" />
          </SegmentTimeline>
        </SegmentTemplate>
```



```
            </Representation>
        </AdaptationSet>
        <AdaptationSet id="3" contentType="audio">
            <Representation id="3" mimeType="audio/mp4" codecs="mp4a.40.2" bandwidth="128000" audioSamplingRate="44100">
                <SegmentTemplate timescale="44100" initialization="init_audio.mp4" media="qr_audio_$Number$.m4s"
startNumber="12">
                    <SegmentTimeline>
                        <S t="0" d="88200" r="3" />
                    </SegmentTimeline>
                </SegmentTemplate>
            </Representation>
        </AdaptationSet>
    </Period>
</MPD>
```

Figure 40: Use Case 1 - End-of-Video QR Code Integration – Sample MPD xml

Use Case 2: Small-Format Video Ad QR Code Integration

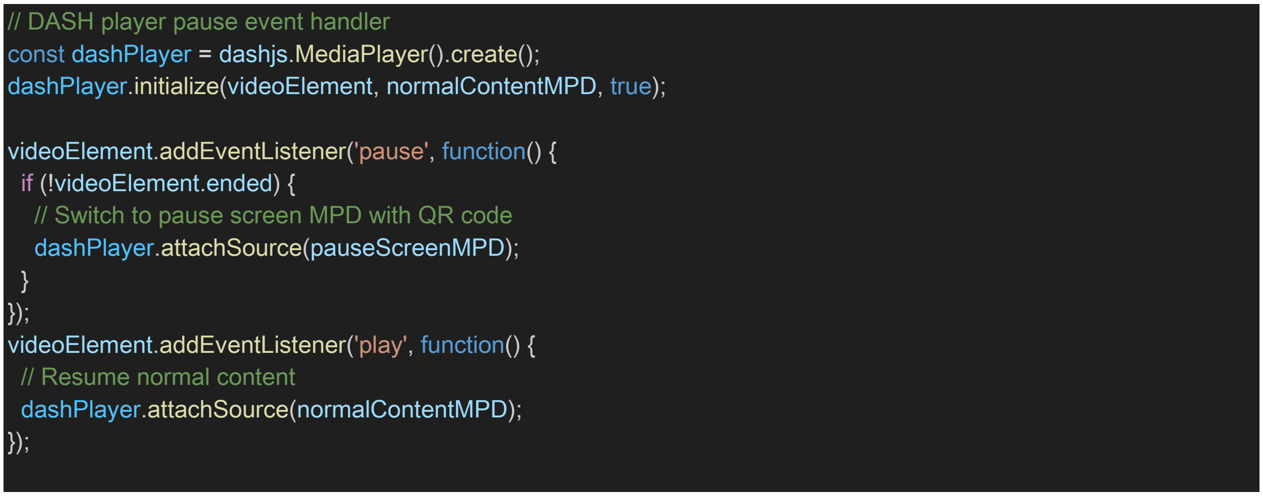


```
<?xml version="1.0" encoding="utf-8"?>
<MPD xmlns="urn:mpeg:dash:schema:mpd:2011"
     type="static"
     mediaPresentationDuration="PT15.0S"
     minBufferTime="PT2.0S">
    <!-- Single Period: 15s @ 1fps -->
    <Period id="0" start="PT0.0S" duration="PT15.0S">
        <AdaptationSet id="0" contentType="video" frameRate="1/1">
            <Representation id="0" mimeType="video/mp4" codecs="avc1.42c028"
                           bandwidth="120000">
                <SegmentTemplate timescale="15360"
                               initialization="init_video.mp4"
                               media="qr_video_$Number%02d$.m4s" startNumber="1">
                    <SegmentTimeline>
                        <S t="0" d="30720" r="7" />
                    </SegmentTimeline>
                </SegmentTemplate>
            </Representation>
        </AdaptationSet>
        <AdaptationSet id="1" contentType="audio">
            <Representation id="1" mimeType="audio/mp4" codecs="mp4a.40.2"
                           bandwidth="128000" audioSamplingRate="44100">
                <SegmentTemplate timescale="44100"
                               initialization="init_audio.mp4"
                               media="qr_audio_$Number%02d$.m4s" startNumber="1">
                    <SegmentTimeline>
                        <S t="0" d="88200" r="7" />
                    </SegmentTimeline>
                </SegmentTemplate>
            </Representation>
        </AdaptationSet>
    </Period>
</MPD>
```

Figure 41: Use Case 2 - Small-Format Video Ad QR Code Integration – Sample MPD xml

Use Case 3: VOD Pause Screen QR Code Integration

Player Changes

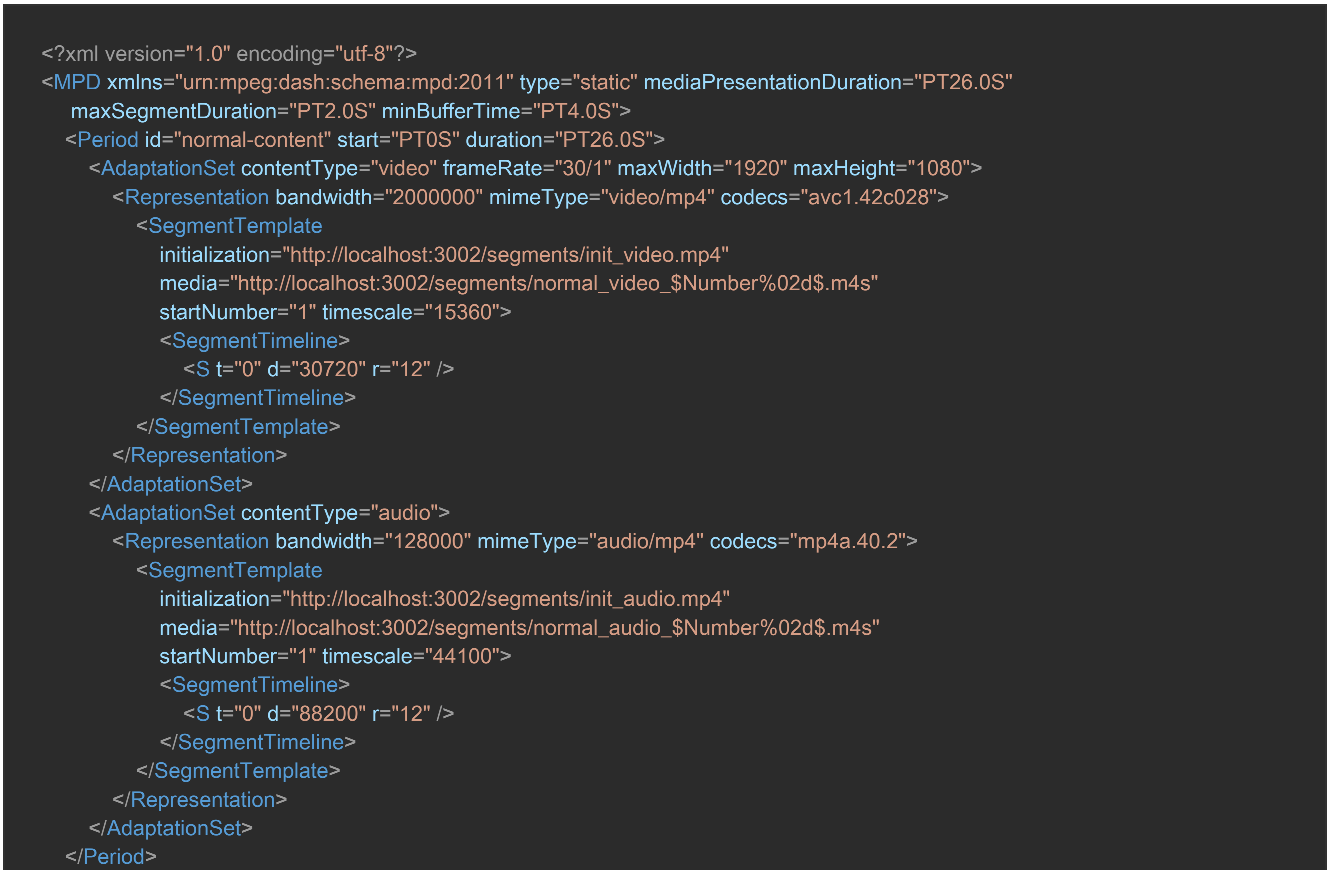


```
// DASH player pause event handler
const dashPlayer = dashjs.MediaPlayer().create();
dashPlayer.initialize(videoElement, normalContentMPD, true);

videoElement.addEventListener('pause', function() {
  if (!videoElement.ended) {
    // Switch to pause screen MPD with QR code
    dashPlayer.attachSource(pauseScreenMPD);
  }
});
videoElement.addEventListener('play', function() {
  // Resume normal content
  dashPlayer.attachSource(normalContentMPD);
});
```

Figure 42: Use Case 3: Pause Detection - Dash Player Changes

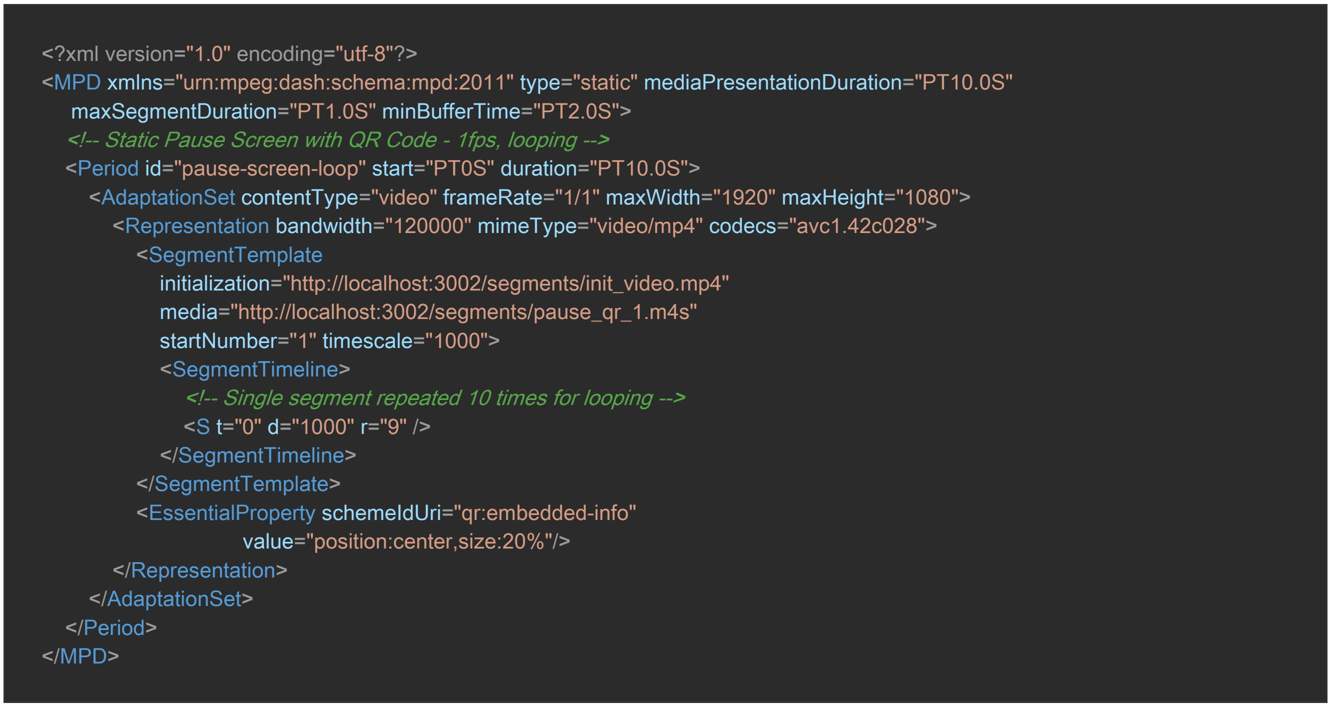
Sample MPD Structure



```
<?xml version="1.0" encoding="utf-8"?>
<MPD xmlns="urn:mpeg:dash:schema:mpd:2011" type="static" mediaPresentationDuration="PT26.0S"
    maxSegmentDuration="PT2.0S" minBufferTime="PT4.0S">
  <Period id="normal-content" start="PT0S" duration="PT26.0S">
    <AdaptationSet contentType="video" frameRate="30/1" maxWidth="1920" maxHeight="1080">
      <Representation bandwidth="2000000" mimeType="video/mp4" codecs="avc1.42c028">
        <SegmentTemplate
          initialization="http://localhost:3002/segments/init_video.mp4"
          media="http://localhost:3002/segments/normal_video_$Number%02d$.m4s"
          startNumber="1" timescale="15360">
          <SegmentTimeline>
            <S t="0" d="30720" r="12" />
          </SegmentTimeline>
        </SegmentTemplate>
      </Representation>
    </AdaptationSet>
    <AdaptationSet contentType="audio">
      <Representation bandwidth="128000" mimeType="audio/mp4" codecs="mp4a.40.2">
        <SegmentTemplate
          initialization="http://localhost:3002/segments/init_audio.mp4"
          media="http://localhost:3002/segments/normal_audio_$Number%02d$.m4s"
          startNumber="1" timescale="44100">
          <SegmentTimeline>
            <S t="0" d="88200" r="12" />
          </SegmentTimeline>
        </SegmentTemplate>
      </Representation>
    </AdaptationSet>
  </Period>
```

```
</MPD>
```

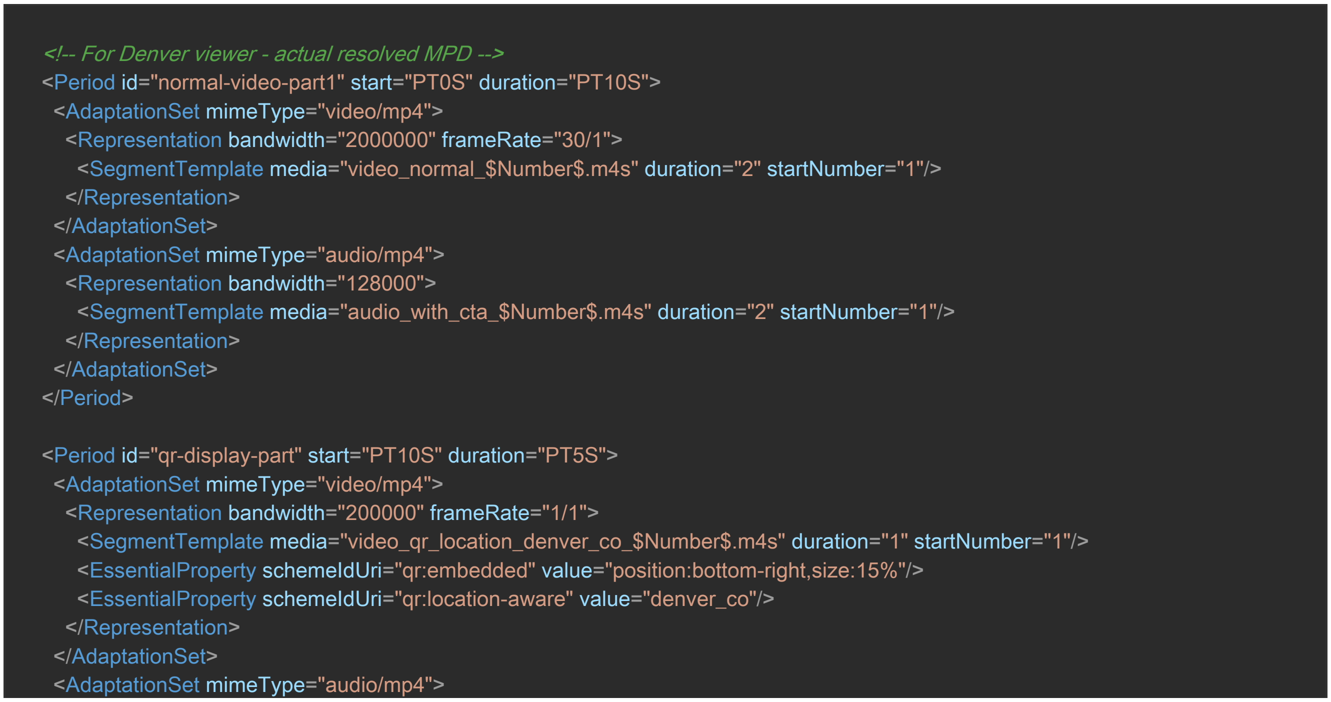
Figure 43: Use Case 3 - Normal Content – Sample MPD xml



```
<?xml version="1.0" encoding="utf-8"?>
<MPD xmlns="urn:mpeg:dash:schema:mpd:2011" type="static" mediaPresentationDuration="PT10.0S"
   maxSegmentDuration="PT1.0S" minBufferTime="PT2.0S">
  <!-- Static Pause Screen with QR Code - 1fps, looping -->
  <Period id="pause-screen-loop" start="PT0S" duration="PT10.0S">
    <AdaptationSet contentType="video" frameRate="1/1" maxWidth="1920" maxHeight="1080">
      <Representation bandwidth="120000" mimeType="video/mp4" codecs="avc1.42c028">
        <SegmentTemplate
          initialization="http://localhost:3002/segments/init_video.mp4"
          media="http://localhost:3002/segments/pause_qr_1.m4s"
          startNumber="1" timescale="1000">
          <SegmentTimeline>
            <!-- Single segment repeated 10 times for looping -->
            <S t="0" d="1000" r="9" />
          </SegmentTimeline>
        </SegmentTemplate>
        <EssentialProperty schemeIdUri="qr:embedded-info"
                      value="position:center,size:20%"/>
      </Representation>
    </AdaptationSet>
  </Period>
</MPD>
```

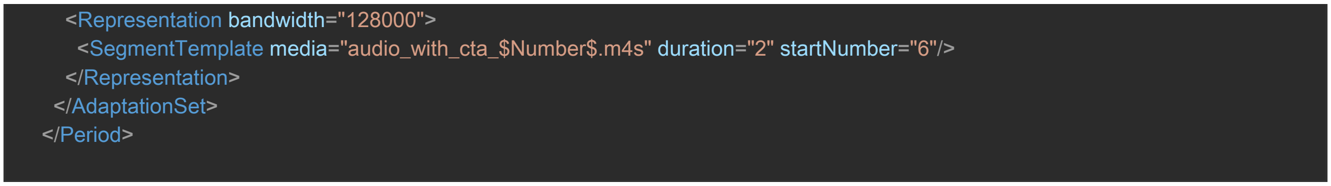
Figure 44: Use Case 3 - Pause Screen – Sample MPD xml

Use Case 4: Location-Aware Dynamic QR Codes in Video Advertising



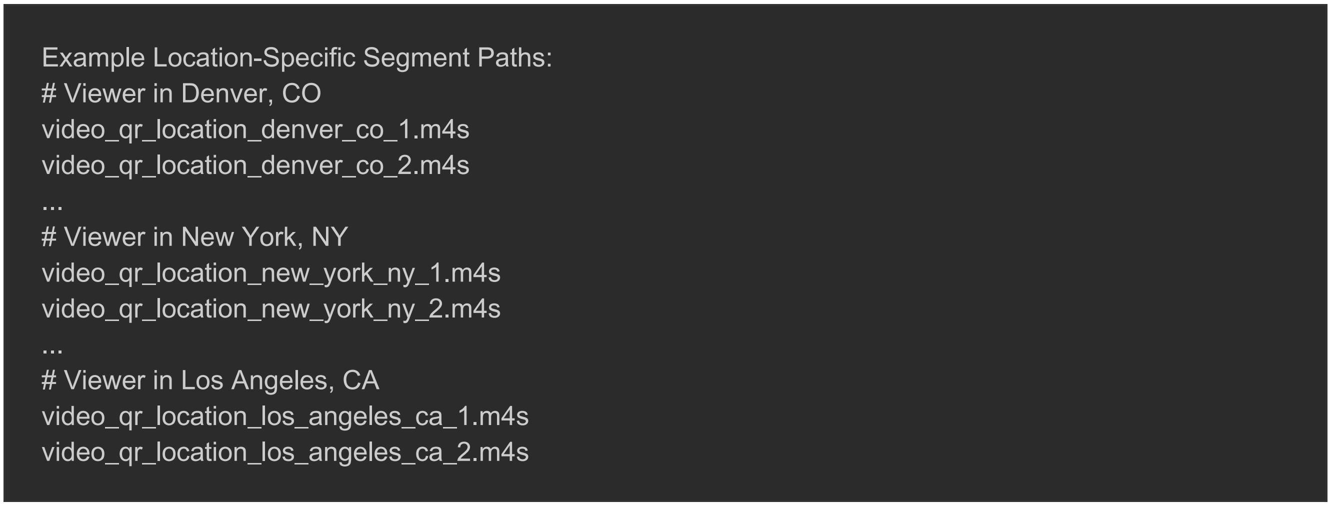
```
<!-- For Denver viewer - actual resolved MPD -->
<Period id="normal-video-part1" start="PT0S" duration="PT10S">
  <AdaptationSet mimeType="video/mp4">
    <Representation bandwidth="2000000" frameRate="30/1">
      <SegmentTemplate media="video_normal_$Number$.m4s" duration="2" startNumber="1"/>
    </Representation>
  </AdaptationSet>
  <AdaptationSet mimeType="audio/mp4">
    <Representation bandwidth="128000">
      <SegmentTemplate media="audio_with_cta_$Number$.m4s" duration="2" startNumber="1"/>
    </Representation>
  </AdaptationSet>
</Period>

<Period id="qr-display-part" start="PT10S" duration="PT5S">
  <AdaptationSet mimeType="video/mp4">
    <Representation bandwidth="200000" frameRate="1/1">
      <SegmentTemplate media="video_qr_location_denver_co_$Number$.m4s" duration="1" startNumber="1"/>
      <EssentialProperty schemeIdUri="qr:embedded" value="position:bottom-right,size:15%"/>
      <EssentialProperty schemeIdUri="qr:location-aware" value="denver_co"/>
    </Representation>
  </AdaptationSet>
  <AdaptationSet mimeType="audio/mp4">
```



```xml
      <Representation bandwidth="128000">
        <SegmentTemplate media="audio_with_cta_$Number$.m4s" duration="2" startNumber="6"/>
      </Representation>
    </AdaptationSet>
  </Period>
```

Figure 45: Use Case 4 - Location-Aware Dynamic QR Code – Sample MPD xml



```
Example Location-Specific Segment Paths:
# Viewer in Denver, CO
video_qr_location_denver_co_1.m4s
video_qr_location_denver_co_2.m4s
...
# Viewer in New York, NY
video_qr_location_new_york_ny_1.m4s
video_qr_location_new_york_ny_2.m4s
...
# Viewer in Los Angeles, CA
video_qr_location_los_angeles_ca_1.m4s
video_qr_location_los_angeles_ca_2.m4s
```

Figure 46: Use Case 4 - Location Specific Segment file download snapshot

Bibliography and References

[1] VideoWeek, “The pandemic sparked a renaissance for QR codes in advertising,” Jul. 19, 2021. Available: https://videoweek.com/2021/07/19/the-pandemic-sparked-a-renaissance-for-qr-codes-in-advertising/.

[2] QRCodeChimp, "QR Code Statistics for 2025: Usage, Trends, Forecasts, and More," 2024. Available: https://www.qrcodechimp.com/qr-code-statistics/ .

SCTE
a subsidiary of CableLabs®

858 Coal Creek Circle Louisville, CO 80027 | 800.542.5040 | **scte.org**