SCTE STANDARDS JOURNAL

VOLUME 2 NO. 1 MARCH 2022

SCTE TECHNICAL JOURNAL

SCTE
a subsidiary of CableLabs®

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs

SCTE TECHNICAL JOURNAL

VOLUME 2, NUMBER 1

March 2022

Society of Cable Telecommunications Engineers, Inc.
140 Philips Road, Exton, PA 19341-1318

Table of Contents

Editorial Correspondence: If there are errors or omissions to the information provided in this journal, corrections may be sent to our editorial department. Address to: SCTE Journals, SCTE, 140 Philips Road, Exton, PA 19341-1318 or email journals@scte.org.

Submissions: If you have ideas or topics for future journal articles, please let us know. All submissions will be peer reviewed and published at the discretion of SCTE. Electronic submissions are preferred, and should be submitted to SCTE Journals ,journals@scte.org.

Foreword

Live events are back!!!

After two years of virtual conferences, we can't tell you how good it felt last month to be at our SCTE-Georgia Tech Management Development Program in Atlanta and our SCTE Chapter Leadership Conference in Indianapolis. At both events, there was an air of excitement and relief as attendees networked with peers, engaged in the face-to-face transfer of knowledge, and returned to an environment of – dare we say it? – business as usual.

Day by day, event by event, we're switching on the live programs that make SCTE so valuable to the industry. In the months ahead we'll see the return of our SCTE-Tuck Executive Leadership Program to the Dartmouth College campus, as well as the debut of the SCTE-Cornell University Systems Architecture and Management Program and the SCTE-Cornell University Agile Leadership Program in Ithaca, NY. Not to mention cable technology's crown jewel, SCTE Cable-Tec Expo®, September 19-22 in Philadelphia.

In the meantime, our SCTE Technical Journals continue to provide a breadth of thought leadership that is unparalleled in the industry. This month's Journal includes the following articles:

- *Network Analysis Using Only an In-Phase Channel*, by CableLabs' Lin Cheng and Tom Williams;
- *How Implementing Differential Privacy Can Protect Privacy in the Streaming Video Advertising Market* , by Charter's Srilal Weera;
- *Aging in Place and Telehealth Use Cases from the Cable Operator Perspective,* by Shaw Communications' Clarke Stevens and Duke Tech Solutions' Sudheer Dharanikota;
- *Precision Timing Protocol: Impact of Network Design,* by Charter's Marek Hajduczenia and Nathan Sary;
- *A Secure and Scalable Code Signing System,* by CommScope's Tat Chan and Ting Yao;
- *Obtaining Low Latency in Upstream DOCSIS Transmissions,* by Casa Systems' Hongbiao Zhang, Peter Wolff, and Vishnuvinod Sambasivan; and
- *Maximizing Wi-Fi 6E: How to Embrace the 6 GHz Spectrum – and the Future – Now*, by Plume's Bill McFarland.

We invite you to avail yourself of the knowledge our authors have shared, and to consider contributing to the SCTE Technical Journal. Most of all, we urge you to make plans to join us at SCTE Cable-Tec Expo in September. While virtual learning and events certainly have helped to expand knowledge sharing, there is tremendous value in resuming the personal interactions that all of us have been missing for far too long.

The SCTE Editorial Staff

Network Analysis Using Only In-Phase Channel

A Technical Paper prepared for SCTE by

Tom Williams, Distinguished Technologist, CableLabs
858 Coal Creek Cir
Louisville, CO 80027
t.williams@cabelabs.com
(303)661-9100

Lin Cheng, Lead Architect, CableLabs
858 Coal Creek Cir
Louisville, CO 80027
l.cheng@cablelabs.com
(303)661-9100

Table of Contents

List of Figures

1. Introduction

Network analysis is an important test function for cable networks, as well as for many other applications such as medical, defense, exploration, wireless, testing, and optical transmissions. For high-efficiency data transmissions to occur, complex channel responses must be characterized so that pre- and/or post-equalization can be applied. If a linear time-invariant system and its inverse are both causal and stable, the system is said to be minimum phase. This implies that if just the in-phase response is known, the quadrature response can be computed using the KKR (Kramers-Kronig relations), or vice versa. RF signal paths usually have a property of minimum phase. This leads to a new method to determine the channel's complex response, which is useful for accuracy improvement, cost reduction, and verification that a channel has the minimum-phase property. Cable-specific field tests, including tests for locating cable damage, leakage identification, and testing coaxial shield integrity, are presented.

2. Background

Network analysis is very important to communications systems. For data to be transmitted successfully, a channel's response needs to be estimated so that inter-symbol interference caused by group delay or echoes can be canceled. Equalization is established either by using training symbols or by blind using adaptive equalizer.

Transmitters and receivers work together to measure linear distortion and eliminate it with equalizers. Pre-distortion at the CPE is commonly used in DOCSIS® upstream and post-distortion in downstream.

Channel equalization can be accomplished in either the time domain or the frequency domain. Multicarrier modulation techniques, such as OFDM and OFDMA, commonly are equalized in the frequency domain with a single complex multiplication of each component subcarrier. Single carrier modulation such as 256-QAM commonly uses time domain filter structures, such as FIR (finite impulse response) equalizers. The equalizers are normally programmed with the inverse channel response to cancel linear distortion. That said, either technique can be used for both single and multicarrier equalization.

Cable network plant maintenance includes both transmission tests and reflection tests. Reflection tests are commonly used to locate line damage, with distance computed knowing delay time and velocity of propagation. Cable networks present a challenge in that the plant being tested starts at one physical location and ends at another. That is, a common frequency/phase reference is not available at the receiver. In a lab environment, that is not a problem. Another challenge is that network analyzers usually have high cost and complexity. In this paper, we propose a simplified design that achieves equivalent measurement functions.

3. Kramers-Kronig Relations

In control theory and signal processing, a linear, time-invariant system is said to be minimum-phase if the system and its inverse are causal and stable. An all-pass network where the group delay is non-zero but the magnitude response is constant is an example of a non-minimum phase network. In fiber optic cable, chromatic dispersion results in a phase shift without a field magnitude change, so it is another example of a non-minimum phase.

In system analysis, among other fields of study, a linear time-invariant system is a system that produces an output signal from any input signal subject to the constraints of linearity and time-invariance. These properties apply to many important physical systems, in which case the time domain response $y(t)$ of the system to an arbitrary input $x(t)$ can be found directly using convolution:

$$y(t) = x(t) * h(t) \tag{1}$$

where $h(t)$ is called the system's impulse response and $*$ represents convolution.

Using a Fourier transform, a time domain impulse response can be converted to a frequency domain (ω) channel response. The channel's response can be expressed in the frequency domain as:

$$Y(\omega) = X(\omega) \cdot H(\omega) \tag{2}$$

where $H(f)$ is the channel's complex frequency response,

$$H(\omega) = H_r(\omega) + jH_i(\omega) \tag{3}$$

where H_r is the real part and H_i is the imaginary part. The channels response can be decomposed into real and imaginary components which are related to one another by the KKR. The KKR are not derived here but are presented below as equations (1) and (2). See references 1 for derivations.

$$H_r(\omega) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{H_i(\omega')}{\omega' - \omega} d\omega' \tag{4}$$

$$H_i(\omega) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{H_r(\omega')}{\omega' - \omega} d\omega' \tag{5}$$

where P denotes a Cauchy principal value. So, the real and imaginary parts of such a function are not independent, and the full function can be reconstructed given just one of its parts.

Equations (4) and (5) state that if the real values of a frequency response are known the imaginary values can be computed. Likewise, the imaginary values can be used to compute the real values. One problem is the integration limits in the frequency domain go from negative infinity to positive infinity. No test equipment works over an unlimited frequency range. That problem can be solved if the device being tested is band limited. One can also apply a "synthetic" or virtual bandpass filter wider than the frequency response of interest to the measured data, forcing measured responses to zero outside the measurement band. Another solution is available when the frequency response is periodic, but that can produce a "spectral leakage" characteristic. Likewise, if a frequency response is up or down-tilted due to cable attenuation, inverse tilt can be added to the measured response to allow the calculation to be performed. So, these equations have utility, but the limitations must be recognized.

4. Conventional Network Analysis

The conventional instrument used to obtain a complex channel response is a VNA. Vector network analyzers are test instruments that measure linear distortion by providing a stimulus to a DUT (device under test) such as a black box or a cable network and measuring a result. The test signal is commonly a frequency-sweep signal, and a resulting response is captured, typically as a set of received magnitude and phase values. That is, complex numbers which can also be expressed as their real and imaginary components.

Figure 1 is a simplified block diagram of a VNA wired for a transmission test of a DUT. These precision instruments are relatively expensive and use a sweep (or stepped) CW signal generator to measure complex frequency responses in networks. Complex frequency responses involve measuring in-phase and quadrature voltage samples at many frequencies.

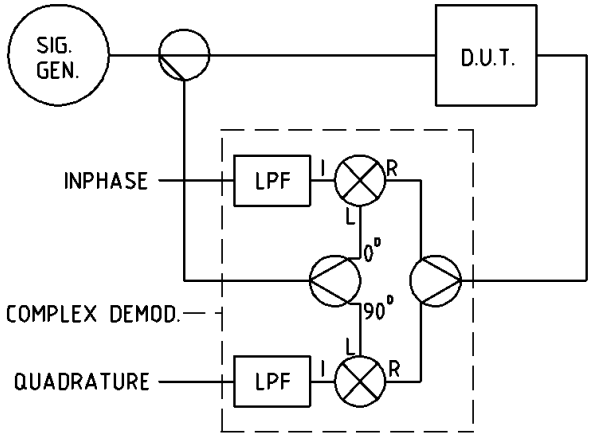

Figure 1 - A simplified block diagram of a VNA performing a transmission test.

The VNA consists of a signal generator that produces a variable frequency test signal, and a receiver using a complex demodulator with in-phase and quadrature voltage outputs that vary with frequency. A sample of the testing signal is fed into the complex demodulator where a splitter produces two outputs, one with 0 degrees phase and one with 90 degrees phase. These reference signals are fed into two double balanced mixers which produce two DC voltages at each test frequency. Two low pass filters (LPF) pass the DC voltages while rejecting noise and harmonics.

Calibration procedures are conventionally used to eliminate test equipment imperfections, such as I-Q imbalance on phase and amplitude.

5. Complex Response Recovery using Real-Only Measurement

DSP techniques allow only the in-phase portion of the complex data to be used to derive a complex impulse response. The complex time response can then be transformed to obtain a complex frequency response.

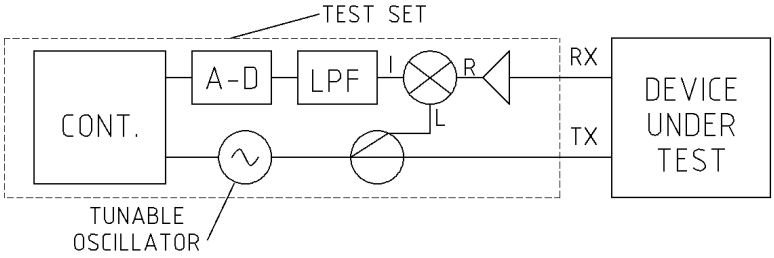

Figure 2 - Block diagram of a test set capable of performing VNA measurements with only in-phase data samples.

Figure 2 is a block diagram of a test set that measures only the in-phase response. It uses a tunable oscillator and one double balanced mixer. The complex demodulator circuit is approximately half of the conventional complex demodulator in Figure 1. In operation, a controller (CONT.) sets a CW frequency at the output of the tunable oscillator; the CW frequency is split and applied to both a double balanced mixer LO port and a DUT. A signal from the DUT is received, optionally amplified, and applied to RF port of the double balanced mixer. The mixer produces an IF signal which consists of a DC (direct current) term and unwanted higher frequency terms, such as interfering signals and harmonics. A lowpass filter removes the high frequency energy. After the lowpass filter settles, the controller orders the analog-to-digital converter (A-D) to take a voltage sample which is stored by the controller. After capturing a number of DC voltage samples at a number of frequencies, the controller has I-only samples to make a plot.

A test of this technique was done using the bandpass filter shown in Figure 3 as the DUT. In this test the drive level to the mixer used in our example was nominally about +7 dBm. The resulting I-only samples are illustrated in Figure 4. In this test, 128 frequency samples (voltages) are taken uniformly between 1 and 26.6 MHz. Note that only in the bandpass region is a non-zero response observed.

Figure 3 – A bandpass filter used for lab tests.

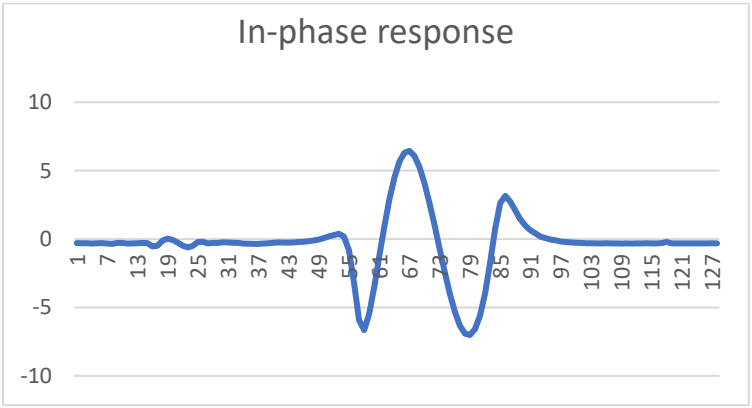

Figure 4 – The in-phase response of a bandpass filter is illustrated.

In a next step the in-phase voltage samples are put into an IFFT structure and converted into the time domain. Zeroes are used for all of the imaginary values. The result of the transform is illustrated in the complex temporal plot illustrated in Figure 5. This complex time plot has a few notable features. One is that the delay is offset from zero time by a few samples. This is due to the time delay associated with the

test leads and the filter's delay. Another is that the samples are conjugate symmetric in time. That is a direct result of using zeroes for all imaginary values in the frequency domain.

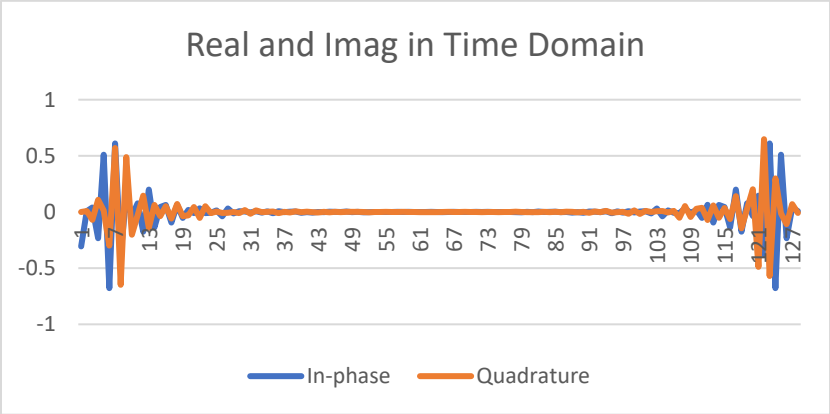

Figure 5 – The complex impulse response associated with the spectral plot of Figure 4.

In the next step, 64 unique values of the 128 total complex time values are transformed with a FFT and converted back into the frequency domain. The result is illustrated in Figure 6.

The voltage amplitude has been halved due to the truncation of 64 time samples, so a correction should be made.

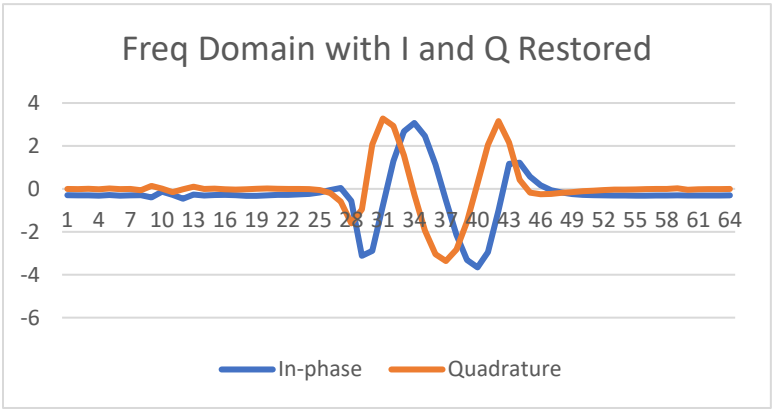

Figure 6 – The FFT of half of the data of Figure 5. Note that imaginary values have been created, but only half as many frequency points are available because of halving the size of the FFT.

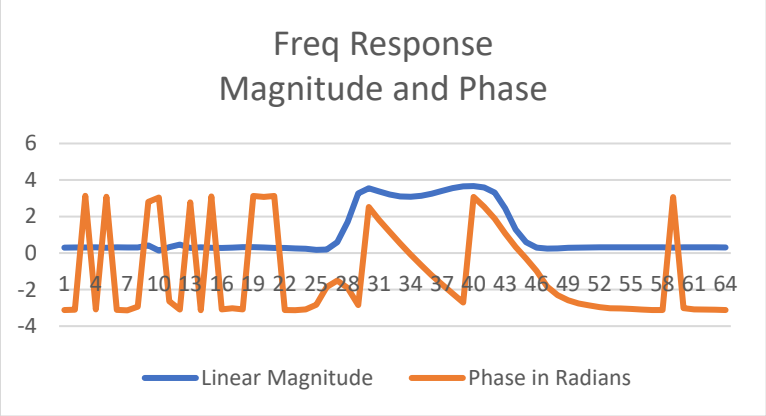

Figure 7 – The complex data of Figure 6 are plotted as magnitude and phase values.

If the I and Q samples in Figure 6 are plotted as magnitude and phase values, the plots of Figure 7 result. Note that phase values are meaningless at frequencies at which there is no significant RF test energy. Thus, magnitude and delay values are rendered.

6. Cable Test Applications

Figure 8 illustrates a drop cable tester. A drop cable is disconnected at the ground block to make an open circuit, and at the tap the cable is connected to the tester of Figure 2. The tester uses a high-quality splitter as a makeshift return loss bridge. This test can produce a TDR plot of voltage vs. distance (time) to reveal cable damage, such as a pirate tap or animal chews. A good cable should have a specified attenuation, and excessive loss may indicate damage or water in the cable. For example, a 100-foot piece of RG-6 cable should have 4.76dB of one-way attenuation at 550 MHz. If it has notably higher round-trip loss, the cable could be damaged.

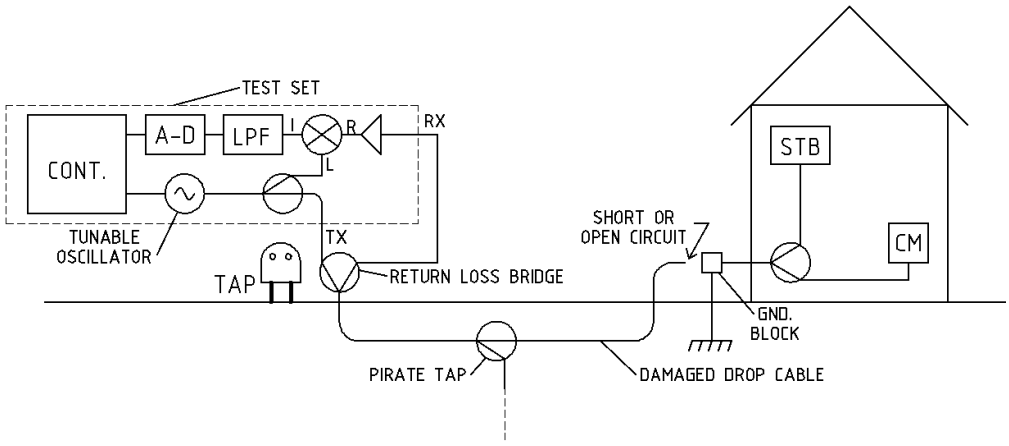

Figure 8 – Drop cable tester.

Figure 9 illustrates a sheath current induction test of house wiring integrity. A transformer is created with a ferrite core by passing the test signal through the core forming a single turn primary winding, and the drop cable's shield forms a secondary winding. A test signal propagates down the outside of the coax into

the house. If there is a shield break, the test current gets inside the coax and travels tack to the test set. Distance to shield break can be estimated. Test frequencies should be in on the order of 5-50MHz. If the drop is buried, the shield integrity test should be performed next the house, as wet soil will absorb the test signal on the outside of the coax.

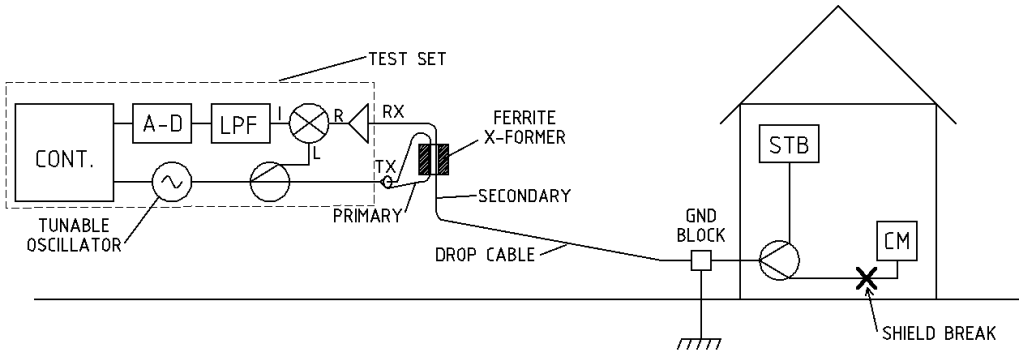

Figure 9 – Sheath current induction test of shield integrity.

Figure 10 illustrates an upstream reflection test for hardline cable damage. A test current is injected into a tap port and a delayed return signal is measured on the seizure screw under the KS port. The seizure screw is probed with a high impedance probe, which typically has 20-30 dB insertion loss. The time delay associated with the cable damage yields a distance, which can speed repairs.

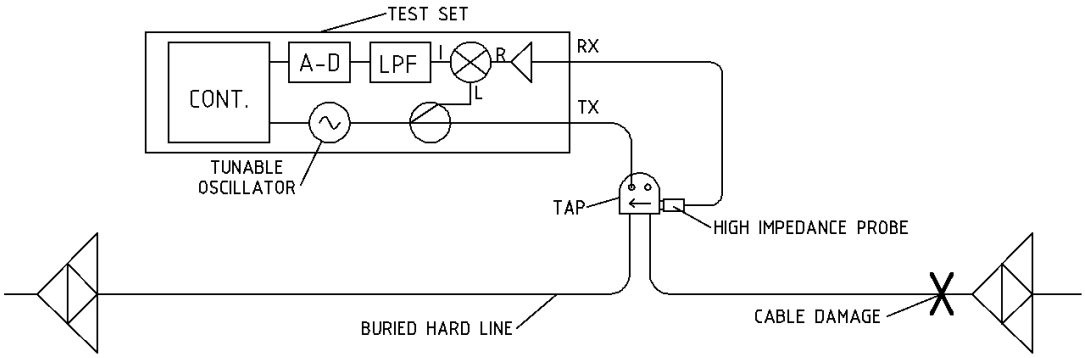

Figure 10 – Hard line cable damage is located by injecting test current into a tap port and measuring a delayed return signal on the seizure screw under the KS port. The seizure screw is probed with a high impedance probe.

Figure 11 is another radiated test of house wiring shield integrity. A test signal is injected into a drop and radiates from a shield break inside the home. An antenna is used to pick up the radiated test signal. This test will also produce a delay value, which can speed troubleshooting. If the shield break is outside the home, a repair can be made when the subscriber is away from home.

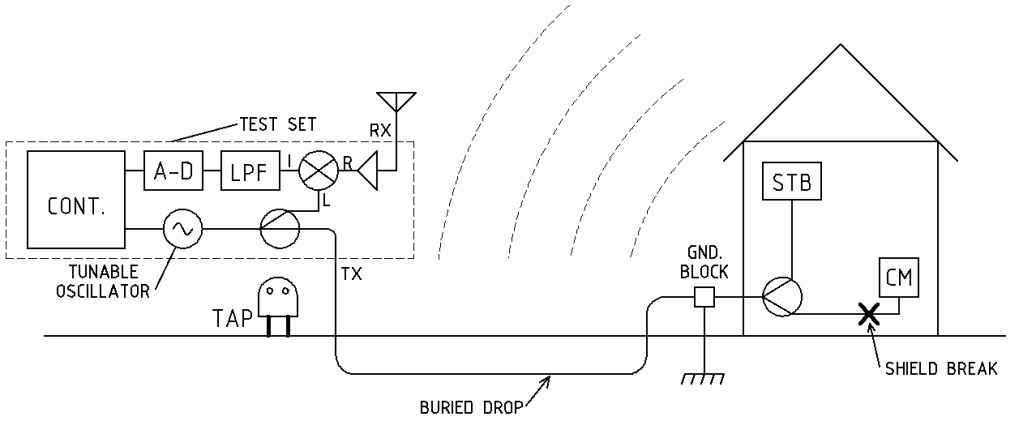

Figure 11 – House leakage testing. This is another radiated test of house wiring shield integrity. A test signal is injected into a drop and radiates from a shield break. An antenna is used to pick up the radiated test signal.

7. Observations and Comments

The processing steps are essentially a method of performing a Hilbert transform on I data to find Q data. Compared with using high-end VNA, the proposed method is exceedingly simple, low cost and works well. The cost of equipment to implement this test is very low, so line technicians and installers can afford to have this test gear to speed their troubleshooting.

Note that this test set can be attached to a return loss bridge (or just a high-quality splitter with good isolation between legs) to make return loss measurements. That is, S-parameters S11 can be obtained as well as S21.

This test method also works for antenna return loss measurements.

Calibration can be used to improve test results. Inherently, the differential errors of I-Q complex demodulators, such as I-Q amplitude and phase imbalance, are eliminated with I-only sampling.

Testing can occur very rapidly as numerically-controlled-oscillators (NCO, a.k.a. direct digital synthesis DDS) can change frequency within a few clock cycles, so frequency jump time is determined by the settling time on the LPF. Note that having a LPF with a long settling time reduces susceptibility to noise but takes longer to test.

Frequency domain sampling needs to obey Nyquist's criteria, with at least two frequency samples per ripple in the frequency domain. If the DUT is a long span of cable, frequency samples must be closer together in frequency than if the cable length is short.

Because a quadrature channel is not needed the working frequency range is higher, potentially extending up to optical frequencies. Thus this idea is applicable for optical tomography medical applications, such as testing retinas, or evaluating cancer cells using frequency tunable lasers.

An amplifier may be used on the transmitted signal. The power amplifier can also provide isolation if there is a mismatch on the TX port.

A related idea is that many NCO designs can be programmed to output a signal with a 90-degree phase shift, so at a test frequency Q values can be measured directly after the I voltage measurement.

8. Conclusions

This paper introduced a low-cost method of achieving VNA functionality using only in-phase measurements to determine complex frequency response. The frequency response can be inverse Fourier transformed to make delay measurements for either reflection or transmission tests. This test method enables portable test equipment that cable technicians can use for plant maintenance, including hard line, drop cables, antenna return loss, and home wiring.

9. Reference

A fast Fourier transform implementation of the Kramers-Kronig relations: Application to anomalous and left handed propagation. Jérôme Lucas, Emmanuel Géron, Thierry Ditchi, et al. Cite as: AIP Advances 2, 032144 (2012)

10. Abbreviations

CW	continuous wave
DOCSIS	Data Over Cable Service Interface Specification
DDS	direct digital synthesis
DSP	digital signal processing
DUT	device under test
IFFT	inverse fast Fourier transform
KKR	Kramers-Kronig relations
LO	local oscillator
LPF	low pass filter
NCO	numerically controlled oscillator
OFDM	orthogonal frequency division multiplexing
OFDMA	orthogonal frequency division multiplexing access
QAM	quadrature amplitude modulation
RF	radio frequency
TDR	time domain reflectometer
VNA	vector network analyzer

How Implementing Differential Privacy Can Protect Privacy in the Streaming Video Advertising Market

A Technical Paper prepared for SCTE by

Srilal Weera PhD, Principal Engineer, Charter Communications
8560 Upland Drive, Englewood, CO 80112-7138
720-699-5079
Srilal.weera@charter.com

Table of Contents

List of Figures

List of Tables

1. Introduction

New consumer privacy laws and regulations are requiring advertisers to re-examine industry practices and develop new, innovative solutions that protect consumer privacy and enable advertisers to provide relevant ads to consumers [[1],[2]]. One solution is to go back to the 'contextual advertising' model, which relies solely on the content of the web page or video. The other is to have a privacy compliant way to use and share consumer data, (subject to appropriate consumer consent and with reasonable limitations and restrictions), to enable relevant ads to be targeted in the streaming video market. The latter option has garnered considerable interest in novel privacy methodologies.

Data anonymizing methods such as IP/MAC hashing are generally considered unbreakable. Improved crypto-hashing algorithms and other security measures (e.g. adding salt) have substantially improved data privacy. Although reversing a hash is practically impossible, current methods are susceptible to dictionary/rainbow and linkage attacks. There are well-publicized cases of re-identifying the de-identified data [[3],[4],[5] ,[6]]. It is claimed that 87% of the United States population could be uniquely identified by just having the individual's ZIP code, date of birth, and gender [7].

The challenge for businesses is how to share consumer data for analytics without compromising privacy. In the wake of the European Union's General Data Protection Regulations ("GDPR") and the state of California's California Consumer Privacy Act ("CCPA"), companies are eager to adopt novel privacy-based solutions. One such technique is "Differential Privacy." Many prominent companies have employed this technology [8]. The 2020 U.S. Census utilized it as well, calling it "the new gold standard in data privacy protection." [9]

Differential privacy (DP) is a statistical technique that permits database querying while protecting identities associated with individual data records. This is done by adding a random noise perturbation to the query response, which obfuscates the user data. It works well in generic scenarios. An example would be when a researcher queries a public database to determine how many watch streaming TV in a locality [10]. The dataflow is one-way in this case. Targeted advertising, however, adds a twist to this model because a return path is needed to serve the ads. Also, the ads must be served to the actual number of consumers who met the query criteria (and not the augmented number after the noise addition). The flip side is that random perturbation masks the consumer identity, thus rendering the ad-campaign ineffective. For example, due to the de-identification, an alcohol ad could be displayed on a minor's viewing screen inadvertently.

Another complication occurs post ad-delivery. It is common business practice to report the actual ad viewership data after an ad is aired. Such ad view metrics are used to measure the performance of an ad campaign as well as for billing reconciliation. The dilemma is that once the actual viewership data are reported then the user identities are revealed; hence the purpose of data obfuscation is lost.

One solution is to add a second stage DP prior to billing and reporting. However, this adds another conundrum: The viewership data reported to the advertiser, (after appropriate consumer consent is received), will now be skewed due to the added noise. On the other hand, a service provider can only bill the advertiser for the actual number of users served. Conversely, the advertisers may not want to pay the full price if, say, 2% of the consumer data reported is inaccurate.

1.1. Differential Privacy – A Brief Overview

Differential privacy (DP) is a powerful privacy tool. When a query is run on a database, it proposes adding a carefully chosen amount of noise/perturbation to the result, masking the user identity. The essence of the algorithm is that not every data record is changed. For example if the query is to find those who subscribe to streaming TV, then some of the user responses (yes/no) are flipped randomly. This gives rise to a new concept called "plausible deniability." Just by looking at specific data, it is not possible to establish if that data is truly associated with a person or randomly generated by the algorithm. The flip side of adding noise is the need to strike a balance between utility and privacy. As more noise is added, the user privacy is enhanced but the utility value of the dataset is decreased. One attribute cannot be improved without diminishing the other.

A simplified example to illustrate the concept of DP is as follows: Assume a database of 10 consumers. In response to the TV viewership question, 7 have answered "Yes," yielding a response set of 7Y / 3N.

Assume the noise margin is set to ± 1. When the query is run, the responses received could be either 6Y / 4N, 7Y / 3N or 8Y / 2N, as shown in Table 1. Given the error margin of ± 1, all 3 responses are equally acceptable answers. Note the 'Y' values of 1,2,3,4,5 and 9 are outside the set error margin.

Table 1 - Differential Privacy – Simplified conceptual example

Consumer ID #	Actual Response	Possible responses (with ± 1 noise added)		
1	Y	Y	Y	N
2	Y	Y	N	Y
3	Y	Y	Y	Y
4	Y	N	N	Y
5	Y	Y	Y	Y
6	Y	Y	Y	Y
7	Y	N	Y	Y
8	N	Y	Y	Y
9	N	N	Y	N
10	N	N	N	Y
Total	**7Y / 3N**	**6Y / 4N**	**7Y /3N**	**8Y / 2N**

The last three columns of the table denote all possible answers for the specified error margin.

1.2. Algorithmic Basis

Differential privacy is based on the premise that adding any one person's individual data to a dataset should not materially change the result of queries run on that data. This is achieved by adding random noise to the data, such that in aggregate the statistical average would remain unchanged. The data assumes a Laplacian distribution and the small random perturbation (epsilon) is called the privacy-budget. It is commensurate with the privacy loss, such that higher the epsilon value, the lesser the privacy.

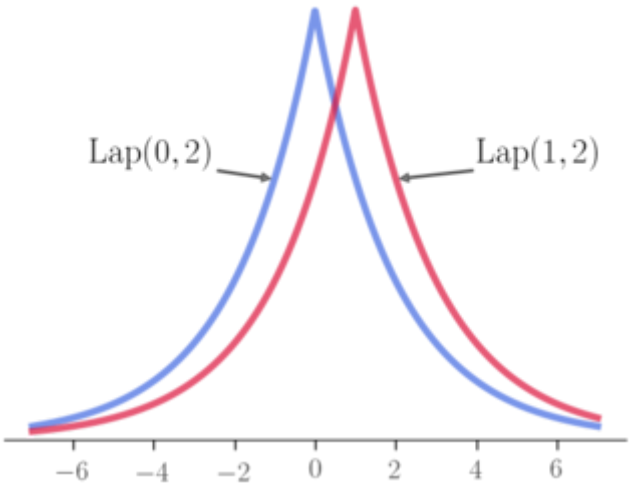

Figure 1 - Two Laplace distributions of neighboring datasets (photo source: Wikipedia)

Given two neighboring datasets D and D' differing by one data record, the randomized function K provides ε-differential privacy when the following probability condition (denoted by 'Pr') is satisfied for all S ⊆ Range(K).

$$Pr[K(D) \in S] \leq exp(\varepsilon) \times Pr[K(D') \in S]$$

ε (epsilon) denotes the privacy loss. Note that "differing by one data record" implies that it would not make a difference whether one individual's data are included in the dataset or not (plausible deniability). The query result would still be within the statistical error margin.

Small ε (more noise) – Better privacy but low utility

Large ε (less noise) – Low privacy but high utility

At ε = 0 ($e^0 = 1$), maximum privacy is achieved though the utility value is diminished. Consider D and D' as two databases: one with and the other without my data records. At the limit when ε = 0, the probabilities *Pr[K(D)]* and *Pr[K(D')]* are equal, i.e. the external querier couldn't learn anything about myself. Thus, while my privacy is protected, from a usage point of view it has no utility value to the person querying the database.

Differential privacy uses Laplace distribution for random sampling, though other statistical distributions such as Gaussian and binomial are also possible. DP can be implemented in two modes. Global DP refers to the case of adding noise to queries on a database of records such as the U.S. census. Local differential privacy is when DP is implemented at the device level. Examples are Apple iPhone or Google Chrome browser (RAPPOR Project) implementations. Note that the same query may result in thousands of records from one database vs. a handful of consumers in another, necessitating an adjustment to the statistical noise accordingly, as the Laplace noise distribution is not linear.

2. Applying DP to Targeted Advertising – Challenges

In spite of its success, applying DP in its current form could disrupt the targeted advertising paradigm. As an example, consider querying a customer database for an ad campaign. DP will modify the query responses to obfuscate user identity. In that process however, there is also the risk for inadvertently misidentify user data. Per FCC Guidelines for Ads, "Broadcasters are responsible for selecting the broadcast material that airs,…including advertisements" [11].

Another issue is related to how the targeted ad-campaigns are run in the industry. The campaign manager module (see section 3) could belong to an external/untrusted entity but embedded within the operator network. The challenge for the service provider would be protecting customer data while sharing it with an external system. (Note: Not applicable if it is a vendor partner with contractual limitations in place). A hybrid solution is proposed to address this issue.

2.1. Ad Verification

After the ad campaign, the advertiser would be eager to know if the ad was correctly delivered, as well as how it was viewed by the intended audience. The ad verification data that are sent back to the advertiser would need to be masked as well, or else the user identities are revealed.

2.2. Billing

The dilemma for billing is that only the actual ad impressions or views can be billed. On the other hand revealing the actual audience would negate the DP model. The proposed solution consists of two parts:

- Ad viewership data supplied to external partners such as advertisers are processed through a second stage DP module (after appropriate consent checks are conducted).
- Billing is based on actual data that can be verified in an audit.

While this seems incongruous at first, it protects the user privacy while billing the advertiser accurately. Consider the previous example: 7Y is the actual value, while 6Y and 8Y are also acceptable values. Thus the advertiser is billed for 7Y (actual), while the viewership reporting could be either 6Y, 7Y or 8Y (plausible).

This would be a new paradigm for targeted advertising billing models. To address the disparity, billing statements may clarify that the reported audience data are in accordance with consumer privacy protection measures.

3. Stages of a Targeted Ad Campaign

Before discussing the role of DP, it is beneficial to understand how targeted ad campaigns work in the streaming video market. There are three stages: database querying, ad delivery and ad reporting (post-campaign).

Queries are run on the consumer database SIS (subscriber information service) by the advertiser/proxy. The goal is to extract consumer profiles that meet the specified criteria. Once a targeted audience is identified, an ad campaign is designed. The campaign terms would generally include the duration, the audience definition, and the terms of payment. The ads are supplied by the advertiser/proxy and are stored on a CDN. Each is assigned a unique URL for retrieval.

During video stream playback the ads are retrieved from the CDN and delivered to the consumer device. The ad campaign manager (CM) determines which ad to display to a target audience based on a multitude of factors. For this purpose, CM also incorporates the ad decision server (ADS) functionality. The ad campaign is scheduled to run at a later time. When an ad break occurs, the default ad is instantly replaced by the targeted ad. Post ad campaign, the viewership and billing data are supplied to the advertiser.

3.1. Workflow Steps for Ad Serving in IP Streaming Video

The following is a high-level description of ad delivery in IP-based TV systems. A description of the steps follows Figure 2.

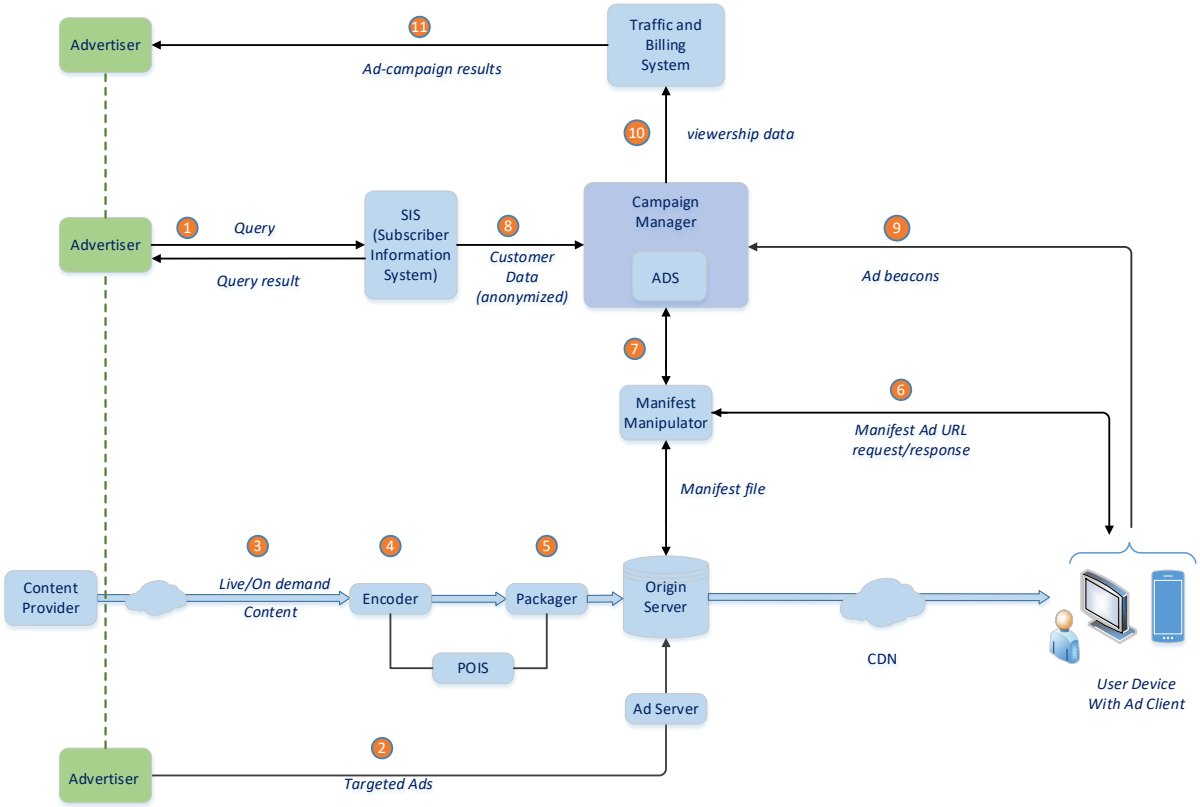

Figure 2 - Ad serving workflow for IP based system

1. Advertiser/proxy queries the consumer database. (Note that consumer opt-outs are honored and excluded from the querying process). The query response contains the records that match the qualifiers defined in the query. The ad campaign is scheduled to run at a later time. When an ad break occurs during that time, the default ad is instantly replaced by the targeted ad.

2. The targeted ads are uploaded to the CDN origin server and assigned a unique URL (uniform resource locator). The URL is supplied to the end user device for retrieving the ad from the CDN.

3. The IP-based video stream is supplied by the content provider to the content distributor. In addition to video content, the stream includes default/baked-in ads.

4. The encoder adds SCTE-35 markers which denote start/end timestamps of ads, enabling replacement of default ads with targeted ads. POIS ensures valid ad break transitions.

5. The packager breaks the content stream into many short segments of few seconds each. The video segments are uploaded to the CDN headend. The packager also creates a manifest, a continuous listing of video and ad segments. The manifest is used by the end user device for playing the video content and ads in a sequential manner. For each entry in the manifest, its URL on the CDN is listed. This enables user devices to locate and download video content and ads.

6. During streaming, the client on the consumer device uses the updated manifest to fetch video segments from the CDN. In targeted advertising though, a check first needs to be made with the ad campaign manager. This is to identify if the end user device is a participant of the targeted ad campaign, in which case the targeted ad is served, instead of the default ad embedded with the video stream.

7. ADS determines which targeted ad to play based on the consumer profile data in the SIS system. When there is an upcoming ad break, the manifest manipulator (MM) sends a placement request to the ADS. Note that the user device identities are masked via cryptographic hashing.

8. SIS contains anonymized consumer profile data required for targeted advertising. It may be a single database with consumer demographic data or multiple systems containing additional data such as TV viewing and web browsing history.

9. As the ad is being displayed, the end user device notifies the ad campaign manager of the progress. These notifications are known as “impression beacons” and are sent for each quartile of the ad display [12], e.g. when 25% of the ad is played.

10. The ad campaign manager supplies ad viewership data to traffic and billing (T&B) systems for further processing.

11. The viewership and billing data are supplied to the advertiser. The advertiser is billed based on the ads served or ads viewed. Served ads are counted based on how many ads were delivered by the ad server. Viewed ads are measured per the guidelines established by industry standards. For instance IAB/MRC ad-viewability metrics define: “A minimum of 50% of the ad is in view for a minimum of one second for display ads or two seconds for video ads.”

3.2. Relation to SCTE-130

SCTE establishes the standards that are widely used in cable TV. The SCTE-130-6 (2020) standard defines the key components displayed in Figure 3. The consumer database is formally defined as subscriber information service (SIS).

Some pertinent excerpts regarding the SIS from pages 12 and 13 of the SCTE 130-6 standard:

- "A subscriber information service responds to queries run against the data gathered from its associated subscriber information data source(s)…
- "Using the basic query mechanism, it can return a list of unique qualifiers that identify subscribers whose characteristics conform to the basic query selection criteria. This kind of query might be used by a campaign management system when determining population counts to meet campaign requirements…
- "The subscriber information service is employed to automate the publishing of audience profile information to the campaign management system…"

ANSI/SCTE 130-6 2020

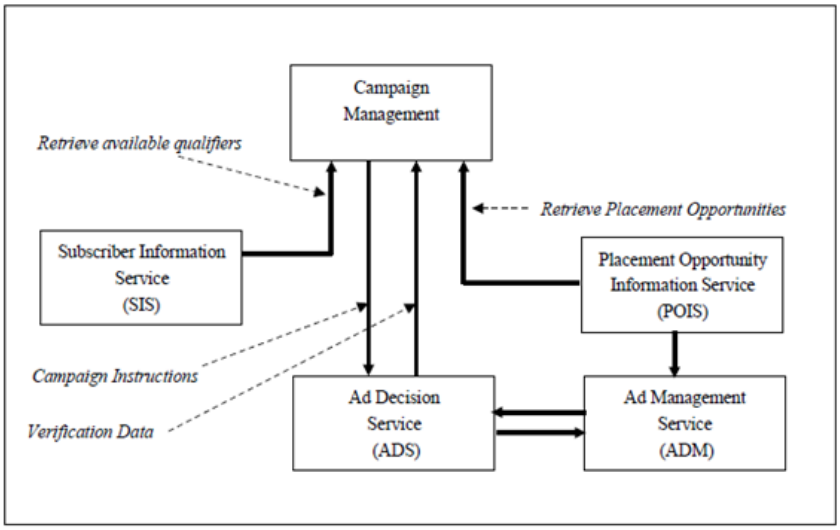

Figure 3 - SCTE-130 connectivity diagram

4. Differential Privacy Implementation for Targeted Advertising

Applying DP to targeted advertising involves several additional components and process steps. Its components are described first.

4.1. Components of the Differential Privacy Enabled Module

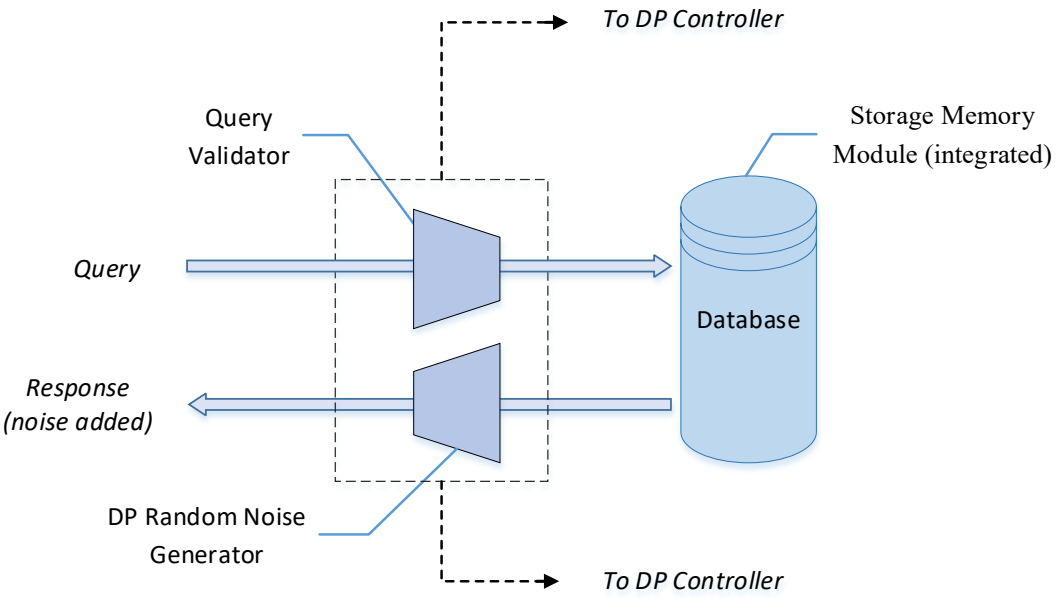

Figure 4 - Differential Privacy Components

1. Statistical/random noise generator – The noise is generated according to a statistical distribution, such as Laplace or the double exponential distribution. The noise is added to the query response (outgoing) from the database.
2. Query validator – Validation protects against queries that are too sensitive and thus could identify individuals. For example, by running coordinated queries with slightly changed criteria, an outsider might be able to make a close guess.
3. Storage memory module – Stores the query and response data, including the parameter values relevant to the algorithm. A return path is maintained to serve ads to the actual audience that met the query criteria. This would also enable recreation of query results at a later time, as may be needed for billing reconciliation.
4. DP controller – The DP controller coordinates parameter settings (such as epsilon values), across multiple instances of DP modules. It also interfaces with the rules engine.

It is to be noted that while differential privacy is the focus of this discussion, the solution described is applicable to other privacy schemes such as k-anonymity and l-diversity.

4.2. Modified Steps for Implementing DP

The steps listed under the section "Ad Serving Workflow" in section 3 are augmented as follows. The primary difference is the addition of enhanced privacy architecture.

When the campaign manager is an external/untrusted entity, then additional privacy safeguards are warranted. This is done via pseudo records. Based on the statistical noise parameters of the query, a commensurate amount of pseudo records are created. For example, if the actual MAC address of the device is "30-56-EC-6F-C4-57", the corresponding SHA-256 hash could be: "b7ed142f1f501eb598e651df9b7802f109635563ee7796cba1a6a2abc9ac95c1." By randomly changing a few alpha-numeric values, pseudo/fake hashes can be created. The fake hashes do not correspond to actual devices. Any ad directed to a pseudo device address is discarded. More importantly, it will not impact the viewership data, also called ad impressions.

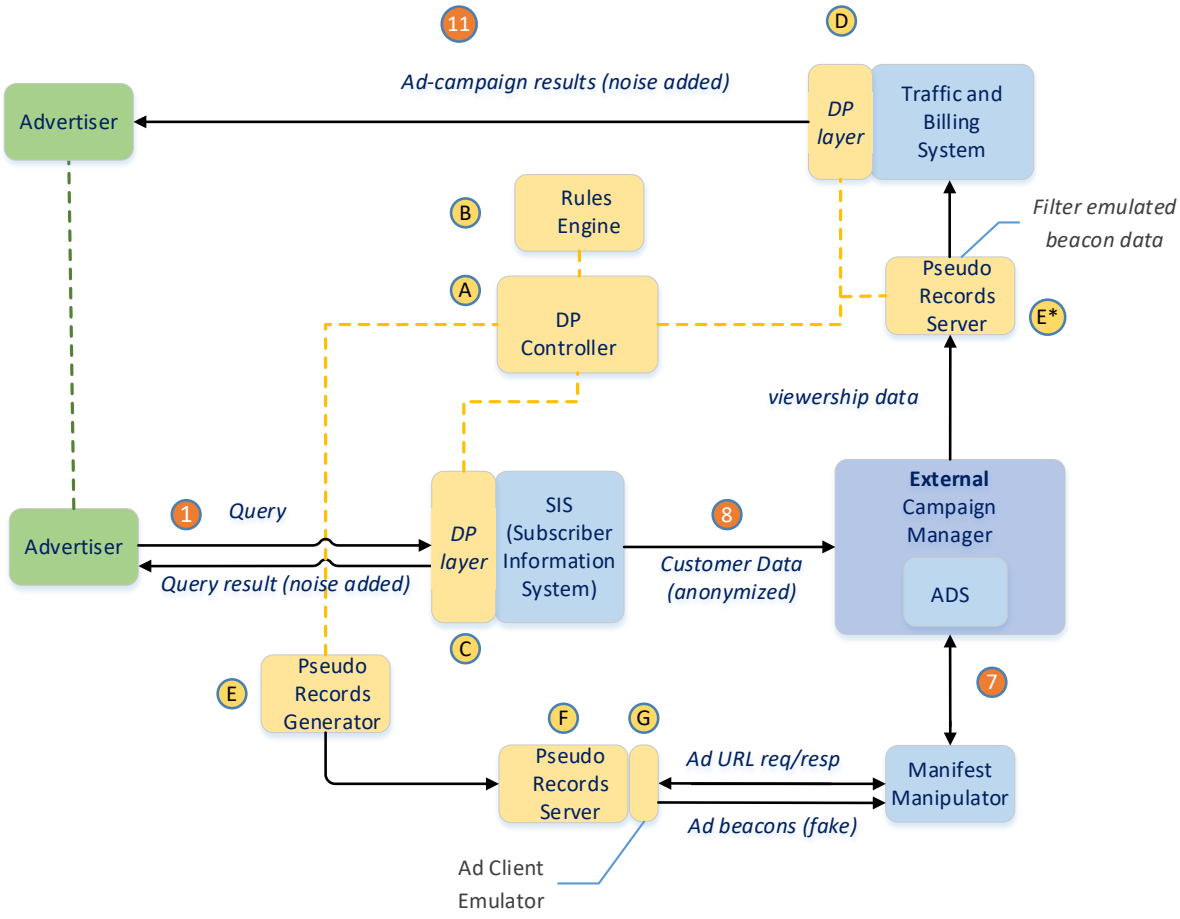

Figure 5 - DP enhanced targeted advertising workflow

Step 1 – The incoming query is first validated for sensitivity. The return response (back to the querier) is passed through the DP layer (**C**). The pseudo records generator (**E**) uses the noise parameter settings to create a commensurate amount of pseudo records. Examples of pseudo records are fake hash addresses. This is facilitated by the DP controller (**A)** in conjunction with the rules engine (**B)**. The pseudo records created are stored in the pseudo records server (F).

Step 7 – During the ad-campaign run stage, the pseudo records server (**F)** will intermittently send fake client requests for ad URLs to the external campaign manager (via the MM). These may look as if they have originated from an actual user device.

Step 8 – The ad campaign manager (CM) checks the validity of the received ad URL requests against the SIS database. This is to confirm that the calling device is in the target audience for the ad campaign. The internal database (SIS) affirms the pseudo record's validity. The CM then supplies the targeted ad URL to the manifest manipulator, which forwards the URL to the fake client on the pseudo records server (**F**). As the ad-requests were not originated from actual client devices, the URLs are discarded by **F**. Note that the ad view metrics (ad impressions) are not changed because the pseudo devices cannot fire quartile beacons, which are measures of ad viewing.

Step 11 – Prior to sending the ad view and billing data back to the advertiser, random noise is injected to obfuscate the user data. This is the function of the DP layer (**D**).

4.3. Additional Enhancements

An additional privacy enhancement would be the ability to emulate the ad client functionality in F. Thus, instead of discarding the targeted ad URL responses received from the CM, it would now be possible to emit simulated ad completion quartile beacons as well. The intent of this is to further mimic the behavior of actual ad clients. When the external campaign manager receives the beacons, those would be added to the total beacon count, as there is no way to distinguish those from the authentic beacons received. Note that this method will not create an erroneous count. The pseudo records server duplicate (**E***) contains records of the fake beacons that were fired previously. As such, when the CM supplies viewership data to traffic & billing, it is first passed through **E***, which removes the fake beacon data. Since the fake beacon counts are now filtered out, the final ad view metrics received by traffic & billing would still be the actual counts.

4.4. Database specific considerations

So far we have considered the database as a monolithic single entity that automatically responds to queries. However, many databases are distributed and transient time considerations are important. Unlike RDBMS databases, the non-relational NoSQL distributed databases prioritize availability rather than consistency. This paradigm is known as "eventual consistency." An example of a NoSQL database is Amazon's DynamoDB which is designed to be available to users without any delay. However, the accuracy of data presented (i.e. consistency), may not be the most current. This could introduce a brief delay for data to be consistent [13]. The implication is that when a query is run on such a database, the data presented may not be the most up to date. This may skew the results of a query response that has the noise added. The DP controller addresses this issue by storing the status, which includes query results and statistical noise parameters. Another related scenario is when data is refreshed or pushed periodically to update a database.

5. Conclusion

Given the advertising industry's desire for privacy enhancement mechanisms, differential privacy offers a plausible solution. Unlike its current applications, targeted advertising adds a complication due to the need for a return path to serve the ads. An implementable solution was presented that tackles this issue and reconciles the billing and ad viewership discrepancies.

6. Abbreviations and Definitions

ADS	ad decision server
CM	campaign manager
DP	differential privacy
LAPLACE DIST.	A continuous probability distribution consists of double exponentials
MM	manifest manipulator
NoSQL	Not only SQL (non-tabular database, different from RDBMS)
OTT	over-the-top
POIS	placement opportunity information system
RDBMS	relational database management system
SIS	subscriber information service
URL	uniform resource locator

Bibliography and References

[1] "*Cookies crumbling as Google phases them out*" BBC News (Jan, 2020) – https://www.bbc.com/news/technology-51106526

[2] *"Apple launches the post-IDFA world to the dismay of advertisers"* Venturebeat (Apr 2021) – https://venturebeat.com/2021/04/21/apple-launches-the-post-idfa-world-to-the-dismay-of-advertisers

[3] *"Keeping Secrets: Anonymous Data Isn't Always Anonymous"* (Mar, 2014) – https://ischoolonline.berkeley.edu/blog/anonymous-data/

[4] *"AOL's disturbing glimpse into users' lives"* (Aug, 2006) – https://www.cnet.com/news/aols-disturbing-glimpse-into-users-lives/

[5] *"NetFlix Cancels Recommendation Contest"* (Mar, 2010) – https://www.wired.com/2010/03/netflix-cancels-contest/

[6] *"New York taxi details can be extracted from anonymised data"*(Jun, 2014) – https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

[7] The Harvard Gazette (Oct, 2011) – https://news.harvard.edu/gazette/story/2011/10/youre-not-so-anonymous . Canadian survey on privacy concerns (2021) – https://www.priv.gc.ca/en/opc-actions-and-decisions/research/explore-privacy-research/2021/por_2020-21_ca

[8] DP in industry – https://www.idownloadblog.com/2017/07/07/apple-differential-privacy-web-browsing-health-data/

[9] 2020 census – https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

[10] "*Query Monitoring and Analysis for Database Privacy*" (Mar, 2016) – https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4795904/

[11] FCC Guidelines for Ads – https://www.fcc.gov/consumers/guides/complaints-about-broadcast-advertising

[12] The current standard for quartile beacon functionality – https://iabtechlab.com/standards/vast

[13] AWS database consistency– https://docs.aws.amazon.com/whitepapers/latest/comparing-dynamodb-and-hbase-for-nosql/consistency-model.html

Aging in Place and Telehealth Use Cases from the Cable Operator Perspective

A Technical Paper prepared for SCTE by

Clarke Stevens, Principal Architect, Emerging Technologies,
Shaw Communications Inc., SCTE Member
1801 N. Broadway, Suite 501
Denver, CO 80202
clarke.stevens@sjrb.ca
720-723-2316

Sudheer Dharanikota, Managing Director,
Duke Tech Solutions Inc., SCTE Member
111 Fieldbrook Ct.
Cary, NC 27519
sudheer@duketechsolutions.com
+1-919-961-6175

Table of Contents

List of Figures

List of Tables

1. Introduction

The COVID-19 pandemic has required people to discover new ways to do things from home over a home network. One area in which this has become obvious is wellness. While many healthcare providers have scrambled to provide remote visits, this only solved part of the problem. What people need is a support infrastructure at home that gives wellness providers the information they need as well as the ability to provide remote assistance. This requires a secure and reliable home network with connected devices to do the observing and controlling. Additionally, connected services beyond the home are needed to log measurements, connect families and wellness providers, and provide notification of significant events.

Cable operators are well-positioned to provide these services. They can extend their traditional offerings of bandwidth and television service to include aging in place (AIP), telehealth, in home care, and safer alternatives to surgery. This can unleash many inter-industry, revenue-generating opportunities in healthcare by linking healthcare devices, families, and wellness providers into coordinated health communities. In-home care solutions over 10G, highly available access networks can both improve the wellness experience and provide new cable operator revenue streams while fully satisfying stringent HIPAA data security and privacy requirements. The operators already have the core technology and the consumer relationships to naturally provide this type of solution.

This paper will explore use cases in this space and point out the key areas where cable operators can provide a collaborative in-home wellness solution, introduces several use cases that SCTE standards working groups are considering fom the aging in place [11] and telehealth spaces in the wellness industry, and considers use cases common to both areas such as connectivity and analytics infrastructure, as well as use cases specific to AIP and telehealth.

2. The Opportunity

U.S. Healthcare costs are increasing at 5.4% year over year and are estimated to reach $5.5 trillion by 2026 [1]. The U.S. healthcare industry is huge, and policymakers have been concerned by its growth relative to total GDP. The criticism is sometimes characterized by the idea that the U.S. healthcare system is a sick-care system and that boundaries must be broadened to effect positive change on national healthcare. Wellness and social determinants of health are important items for discussion within the healthcare industry. This represents an interesting opportunity for cable operators with strong residential franchise presences.

The healthcare industry has been modernizing its infrastructure intending to control costs and improve the quality of care. Telehealth is one such mechanism that has been gaining adoption. Telehealth played a critical role in virtualizing care during the COVID pandemic. Telehealth has been growing at a yearly rate of ~15% with 2020 seeing a 175x increase in Telehealth adoption mainly due to COVID-19 [2]. This telehealth infusion is driven by increased patient and provider adoption, better reimbursements, and relaxed regulations. Although adoption may slow during COVID case decreases, telehealth benefits are recognized and are here to stay. Telehealth is not just video communications; it also touches on different technological solutions that cable operators have mastered and have been deploying. Healthcare has lagged most industries regarding the virtualization of services. Consider how the retail, finance, and entertainment industries have been transformed by digital technology over the last decade. The potential disruption to healthcare is inevitable. The cable industry not only brings technology and service management but also leadership in building standards-based platforms that can deliver critical cost reductions required to assist the healthcare industry.

The Center for Medicare and Medicaid Services (CMS) projects up to $5.5T healthcare spend in 2026. Of the total spend, we project that $3T can be optimally addressed by virtualizing the care models with telehealth and better connecting the existing wellness and housing sectors to the healthcare industry [3]. $1.3T of this total spend can be addressed and reduced by telecom operators. In this report, we demonstrate some of the needs that cable operators can support using their developing capabilities such as in-home technologies, IoT, broadband communication enhancements, consumer service development, platform standardization expertise, back-office capabilities, and installation and support resources.

Different services are driving healthcare costs that can be addressed by telehealth initiatives [3]. These services include perpetual wellness, aging in place (AIP), communication-enabled medical encounters (CEME), virtual pharmacy, hospital at home (HAH), and remote specialty services. A very conservative analysis in these six segments shows that a U.S. telecom operator can recognize ~$27 billion per year (Figure 1).

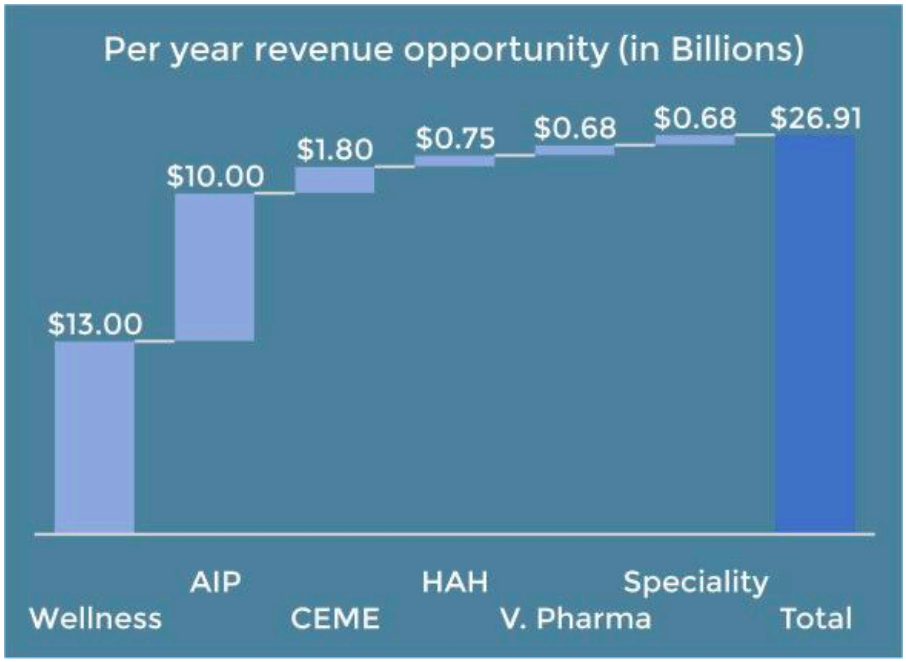

Figure 1 - A conservative estimation of U.S. telecom for wellness opportunity from DTS [3]

In this paper, we will concentrate on three use cases: aging in place; CEME for independent living; and hospital at home. Each of the use cases considered is analyzed from the wellness stakeholders' points of view with clear definition of the problems, how stakeholders interact with one another, how and what to sell to the stakeholders, what is involved in the solutions, what are the opportunities for the stakeholders, and what are the opportunities for the cable operators.

All these opportunities are analyzed against the framework shown in Figure 2 [3], [4]. The cable operator can use this framework to realize their telecom for healthcare (T4H) solutions. The framework includes telehealth patient/consumer home components, the telehealth sensor network infrastructure, and the telehealth hosting back-office infrastructure.

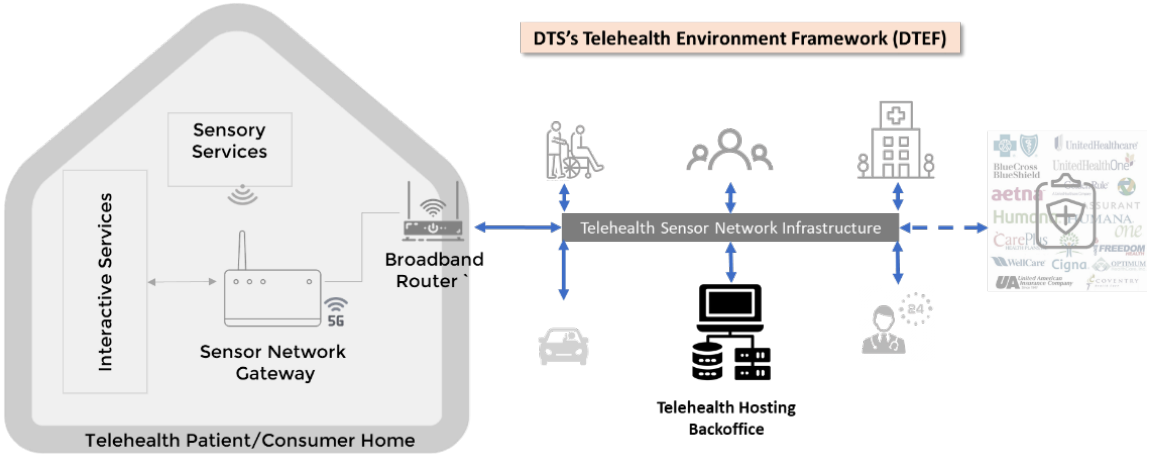

Figure 2 - A telehealth environment framework

COVID-19 has accelerated the rollout of telehealth from an experimental new service to a necessity. It has uncovered a powerful new wellness delivery model that could fundamentally change the cost structure of the entire wellness industry. Because of this, the transition to a virtualized wellness service model is at the top of the wellness strategy stack. We believe the wellness industry transformation goals can be addressed by the cable operators, making this inter-industry collaboration a success. As mentioned previously, it has the potential to mold into a $1.3 trillion opportunity for U.S. telecom operators by 2026 [3].

2.1. Telehealth Concepts

There is too much confusion on the tele-X terminology [5]. The US Department of Health and Human Services (HHS) [6] defines telehealth as "the use of electronic information and Telecommunications technology to support and promote long-distance clinical health care, patient and professional health-related education, and public health and health administration".

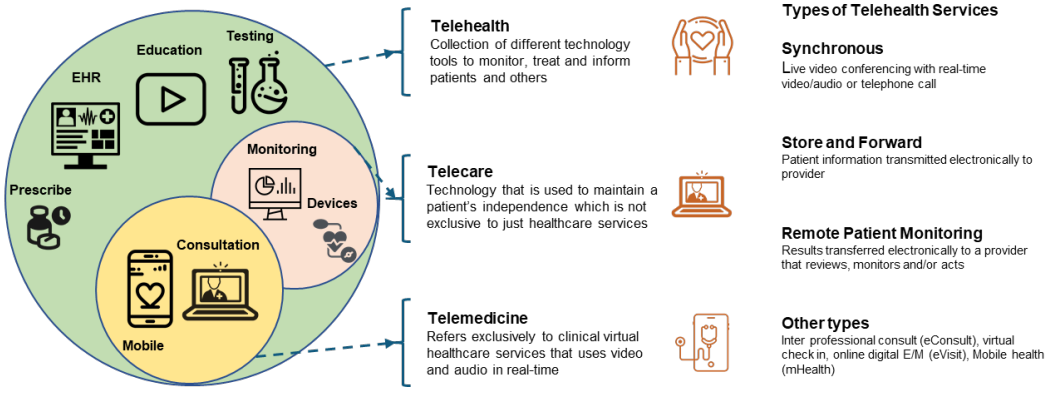

Figure 3 - Untangling the tele-X terminology [5]

Because universal definitions for virtual care have yet to be established, the Federal Communications Commission (FCC) uses different terms and definitions for the tele-X space than HHS [7]. Thus, it is imperative to acknowledge that the industry has yet to land on universal definitions around virtual care. In

this report, we use the definitions presented in the above figure [5] to discuss opportunities for telecoms across various telehealth services.

3. The Stakeholders

Every patient ages uniquely and has unique needs. The solution for effectively addressing the needs of patients will need to be tailored to those requirements. However, there is a sort of continuum of aging that is common. A patient moves from a position of relative independence toward more dependence on a team of supporting stakeholders.

The patient is the focus of this support. Stakeholders in the process of caring for the patient include a community of family and friends such as the patient advocate/power of attorney (POA) and neighbors, and a collection of caregivers including primary care doctors, medical specialists, nurses, therapists, personal care assistants, housekeepers, meal providers, etc. There are also insurance companies, Medicare, and other payers. Many patients want to remain in their own homes with relative autonomy as long as possible. Covid has pushed the limits of some of these ideas. Technology has demonstrated that telecommunications, networking, and automation can fill in gaps that previously required manual or higher-level assistance. Technology solutions can be more convenient, more effective, and less expensive. This is an opportunity for cable operators [8], [9].

Patient or User: The primary stakeholder is the patient. Proper care of the patient is the primary goal. Proper care includes providing that care in compliance with the patient's preferences as far as possible. The stakeholders must have a common plan of care. They must be able to communicate that plan and execute it effectively. Clear means of measuring the effectiveness of the care and communicating that information with all the stakeholders is key.

Community: Perhaps the most important group of stakeholders in the patient's circle are their friends and family. These are the people with the most direct and personal relationships with the patient. In many cases, they are responsible for making decisions about the patient's care. They should be notified of significant events and be in regular communication with the patient and other caregivers.

Caregiver: Caregivers include those responsible for providing patient care. Caregivers include physicians, nurses, therapists, specialists, and other healthcare providers, but also include cleaning services, personal hygiene providers, and mealproviders. They will interact with the patient on a scheduled or routine basis but may also be required to respond to emergency needs. Also, it is important to note that the location of the service could be at home, an independent living facility, or an assisted living facility. Providing the continuum of care for a given user throughout their transitions is very rewarding and increases the stickiness of the user.

Payer or Insurance: Except for the patient's community of family and friends, most stakeholders will need to be paid. Depending on circumstances, this can include personal funds, insurance coverage, and government resources. This stakeholder service can be highly automated to simplify interactions and ensure that caregivers and service providers are promptly compensated while the patient and their designated care advocate(s) clearly understand the process, costs, and expected benefits.

3.1. Infrastructure

Infrastructure is the most important aspect of a remotely-delivered integrated wellness care system. This is what ties all the component pieces together. At the most basic level this includes installing sensors,

actuators, and monitoring equipment. At the higher levels it involves keeping track of all the stakeholders, providing them with the information they need, and providing the tools so they can act on that information.

Wellness-related infrastructure: Wellness infrastructure is the glue that alerts families and caregivers when the patient needs attention. It communicates when care is rendered so the insurance company can pay for the service. It reminds the patient to take medications and schedules visits. This would be a new area for cable operators so it may make sense to partner with or acquire a company that has expertise in this area already.

Technical: Technical infrastructure is where the cable operators have an incumbent advantage. Providing networking to the home is a core competency. Connecting in-home devices (such as Wi-Fi infrastructure, sensor devices, television, and other equipment) inside the home is also routine for cable operators. Extending the connectivity to networked health and monitoring devices is certainly within current skill sets. Setting these monitoring systems up may require wellness care partnerships or additional training. There will be additional security regulatory requirements for the type of traffic transported. There may be an opportunity to provide monitoring, data collection, and data analytics services on behalf of the wellness care stakeholders.

4. Use Cases

Use cases in wellness care are as unique as the patients themselves, so flexibility is very important. However, there are some key categories of service that are common across most use cases. This paper examines three use cases that cover the spectrum from simple assistance for daily living to acute hospital care from home.

4.1. Aging in Place

Aging in place involves passive monitoring with selective assistance for daily tasks. The best aging in place technologies do little or nothing to disrupt a patient's normal activities. For example, a connected pillbox can report which medications a patient takes and when. Falls can be detected by a watch or other item that is worn or can be detected passively by a floor-mounted impact sensor. The point is to provide technical tools and assistance to help the patient do simply what might otherwise be difficult.

The table below maps several services that might be required for a patient aging in place mapped against the various stakeholders who might be involved in delivering that service.

Table 1 – Aging in place services

Service	Community	Caregiver	Payor	Infrastructure
Connectivity	H	L	L	M
Installation	M	L	N/A	M
Communication	M	M	L	M
Monitoring	M	L	N/A	M
Analytics	L	L	L	L
Cable operator opportunity: (L)ow, (M)edium, (H)igh, (N/A) not applicable				

The in-home architecture illustrated in Figure 2 outlines an infrastructure that can provide these services. The data network in the home connects all the relevant devices. It also allows for remote communication

with the patient's community and healthcare providers. Information from the sensors can be logged and analyzed and can be shared during remote visits. All this information travels over the network protected by the highest reasonable levels of data and transport security. The status of each connected device can be monitored with deviations from expected norms reported. Defective equipment can be not only detected but also proactively determined for immediate service or replacement.

To be sure, the data collected has HIPAA security and privacy implications for it to be carefully secured and managed throughout its lifetime. It is expected that the health infrastructure provider will manage this, and the physical security of the network will be the responsibility of the technical service provider. The ultimate success of the system will be indicated by positive health outcomes at a lower total capital and operational cost than current alternatives. With lower costs and improved patient outcomes, this system can be attractive to all stakeholders.

4.2. Independent Living

The next level of use case is independent living [10]. In this circumstance, patients are not able to do some daily living activities without assistance. Patients may have mobility issues around the home or they may have cognitive impairments that require personal assistance. It may be important to have professional home visits scheduled and audio or other reminders provided. Physical access to the home may need to be controlled through a video camera interface and verified credentials. Automated locks can secure the home but allow access for verified visitors. The table below illustrates some of the differences between aging in place and independent living patients.

Table 2 – Independent living services

Service	Community	Caregiver	Payor	Infrastructure
Connectivity	H	L	L	M
Installation	H	M	N/A	H
Communication	H	M	L	H
Monitoring	M	M	N/A	M
Analytics	M	L	L	M
Cable operator opportunity: (L)ow, (M)edium, (H)igh, (N/A) not applicable				

Some of the stakeholder roles may be different, but they fall into the same general categories. Independent living still requires connectivity, infrastructure, and security. While aging in place can get away with mostly passive monitoring, assisted living use cases may require more direct monitoring. This might include blood, urine, and other samples that may need to be performed by a professional with data collection, storage, and analysis capabilities by the envisioned system.

4.3. Hospital at Home

At the end of the spectrum is the hospital at home use case. In this instance, acute care is needed. The patient may be bedridden or otherwise be mobility limited. Maybe complicated procedures need to be regularly performed. If constant in-person professional monitoring is not required, though, the patient will often prefer in-home care where they can be in familiar surroundings and with family more frequently. This alternative when possible also drastically reduces the provider costs. Assistance with meal preparation and personal hygiene may be required. The table below shows how hospital at home services differ from aging in place or independent living use cases.

Table 3 – Hospital-at-Home Services

Service	Community	Caregiver	Payor	Infrastructure
Connectivity	H	L	L	H
Installation	H	H	N/A	H
Communication	H	H	M	H
Monitoring	H	H	N/A	H
Analytics	H	H	M	H
Cable operator opportunity: (L)ow, (M)edium, (H)igh, (N/A) not applicable				

Again, the technical infrastructure is the same. Just the level of service is increased.

5. Services

Now let's look in a little more detail at the technical service categories required by patients in these various use cases.

Secure Connectivity: At the most basic level, any sort of in-home care requires a secure, reliable network. This is already the core service that cable operators provide to homes. What is new is the extension of networking to devices within the home. In general, this has been the responsibility of the homeowner. Any device being managed within the home will have to be verified to work on the network and the security of the data traveling or at rest while under the control of the operator will need to be guaranteed. A service level agreement (SLA) for network availability and quality as well as a detailed HIPAA conformance test will likely be required. Secure connectivity is also required by the other stakeholders on the business side. Depending on the markets addressed by the cable operators, there may be connectivity opportunities here as well.

Accessible Communications: Beyond the basic network, applications will be required for communication. This will include video conferencing to communicate with the community and with caregivers. During Covid, many people became quite familiar with this technology, but older patients will likely need drop/moisture-proof and cleanable equipment design, simpler user interfaces, and pre-populated call lists. Patients may find it easier and more convenient to do video conferencing on the television from their couch.

Communications will also need to extend beyond video conferencing. During medical visits, it may be important to share output from devices like networked blood pressure cuffs or pulse oximeters. This information becomes even more useful if it has been logged over time with significant readings/events highlighted and commented. A video conference may be initatied by sensors noting events that need to be evaluated by medical professionals.

An important part of this communication network is the notification infrastructure. In certain cases a caregiver, doctor, or nurse may need to be notified that conditions have exceeded a certain threshold. Other less critical circumstances (like a change of routine or elapsed time since the bathroom was used or a refrigerator door opened) might trigger an alert to a family member who may discuss the change with the patient. A complex set of rules for guiding who gets notified and when the information is collected, becomes a feature set that cable operator's home care makes apart from more traditional medicine. Monitoring, informing, and control happen automatically and immediately so nobody needs to remember

anything and events are logged for analysis. This information collection is generally inexpensive or even free, decreasing the cost for everyone while improving the level of patient care.

Monitoring: Connecting equipment to a secure network enables remote monitoring. On a basic level, this can be door and motion sensors that record when the patient uses the bathroom. Other instruments may monitor sleep conditions or measure parameters from more involved medical instruments. Authorized caregivers or family might be alerted to a fall or a failure to take medications. Equipment can notify technicians if it fails calibration or needs service. Alerts can be generated if the equipment becomes disconnected and fails to call in on a specified schedule. Data can be securely logged so the circumstances of any event can be placed in context even after it occurs.

Analytics: Collecting accurate time-stamped data is critically important, but often isolated data yields little insight. Predictive and proactive analytics must take information from all sources, look at it in context, and extract the important information. Big data analytics techniques can evaluate massive amounts of unstructured data and determine correlations that might be impossible for humans to discover. More information with better insights obtained more immediately at lower costs means more efficient healthcare and better care for the patient.

6. Cable Operator Solutions

This section examines how this business might make sense for cable operators. Covid forced many people to utilize remote healthcare much sooner than they might have done otherwise. While remote doctor appointments are just one feature of home health care, they introduced people to the concept.

Integration: From the start of the cable industry, cable providers have been involved in integrating and packaging services for simpler consumer consumption. Cable providers collect content from several services and sell it as a single service to consumers. The idea of aggregating wellness services from several stakeholders into a single home service is a natural evolution. The networks that connect stakeholders are provided by cable operators with billing, operations, installation, management, customer service, and other cable skills critical for a successful home wellness care service.

Connectivity Provider: Connectivity is a primary cable product. Cable operators are premium providers of networking services for consumers and businesses. A secure network is a basic infrastructure required for any home wellness care service. Customers already trust this service and are comfortable paying for it every month. Regardless of who provides home healthcare, there is a good chance based on home service penetrations that cable networks underly that service. It is not much of a stretch to think that cable operators can create viable and secure home healthcare services.

Installation Services: Cable operators have fleets of service vehicles and trained installers who regularly install equipment in customer homes. There are few industries more capable of installing networked residential devices. Given the HIPAA requirements, additional training and certification will be required for technicians with this responsibility, but several business models could be used to manage qualified technicians.

Monitoring Services: Monitoring equipment is necessary to keep the network running optimally. Remote monitoring is required to do this at scale. For cable operators, this is business as usual. Extending this service to wellness equipment in the home will require expansion, but it is expanding an existing service rather than introducing a new one. Data collected from equipment can be used to diagnose both individual

hardware devices and the network in general. A wellness portfolio would require many more data models, increased storage capacity, and an improved analytics capability.

Analytics: Analytics involves making sense of data. The cable industry has tremendous storage and computational resources capable of performing these complex analytics. When cable companies don't own the technical and human resources outright, they can get these services through many cloud service providers.

7. Conclusions

These use cases provide some insight into how a common cable operator infrastructures might be leveraged to provide a common home wellness care service that is flexible enough to support a range of use cases from assistive services to the management of complex hospital-at-home healthcare services. Cable operators can leverage their existing competencies in providing network services, installation, managing monthly subscriptions, monitoring systems, analyzing data, and providing customer service to make a legitimate case for operating home healthcare services like aging in place, independent living, and hospital at home. Business conditions, and especially the presence of standards to drive the industry to common and cost-effective solutions, will determine whether cable operators could more successfully offer these services by building their services, partnering with companies in this space, or acquisitions. As pricing pressures move basic networking services towards commoditization, cable operators need to evaluate new business opportunities that maintain and increase profit margins. Home healthcare services offer a promising opportunity.

8. Bibliography and References

[1] IBIS World, *Telehealth Services in the US –Market Size 2005 –2026*, Aug 2020, available here
[2] Oleg Bestsennyy, Greg Gilbert, Alex Harris, and Jennifer Rost, *Telehealth: A quarter-trillion-dollar post-COVID-19 reality?* McKinsey report, May 2020, available here
[3] Duke Tech Solutions market Research, *Telehealth market report – A Telecom based opportunity analysis*, available here
[4] Sudheer Dharanikota, Clarke Stevens, *End to End Telecom for Healthcare Architecture – A Cable Industry Perspective*, 2021 SCTE Expo, available here
[5] Ayarah Dharanikota, Untangling the Tele-x Terms for Telecom Operators, Duke Tech Solutions blog, available here
[6] *What is Telehealth?* Health and Human Services, March 2020, available here
[7] *Telehealth, Telemedicine, and Telecare: What's What?* Federal Communications Commission, available here
[8] Sudheer Dharanikota, Ayarah Dharanikota, Dennis Edens, Bruce McLeod, *Aging in Place Business Case for Cable Operators*, SCTE Journal, June 2021, available here
[9] Sudheer Dharanikota, Ayarah Dharanikota, Dennis Edens, Bruce McLeod, *Telehealth Business Case for Cable Operators*, SCTE Journal, June 2021, available here
[10] *What Is an Independent Living Facility?* Siyanda, Dec 2021, available here
[11] Ian Wheelock, Charles Cheevers, Sudheer Dharanikota, *The Business Case for Aging in Place with Cable Operators*, 2020 SCTE Expo, available here

Precision Timing Protocol (PTP)

Impact of Network Design

A Technical Paper Prepared for SCTE•ISBE by

Marek Hajduczenia, Distinguished Engineer, Spectrum Enterprise
14810 Grasslands Dr | Englewood, CO 80112
marek.hajduczenia@charter.com
+1-720-536-1366

Nathan Sary, Principal Engineer I, DOCSIS Design
14810 Grasslands Dr | Englewood, CO 80112
Nathan.Sary@charter.com
+1-608-286-5274

Table of Contents

List of Figures

1. Introduction

Precise time, phase, and/or frequency synchronization (hereinafter referred to as synchronization) is crucial in a variety of industrial, automation, and control applications, though from the perspective of a service provider, it is typically provided as a service to end customers for the use in:

- Cell tower synchronization, starting with 2G systems onwards,
- Circuit emulation services, in which the clock and frequency information need to be recreated at the edges of the network to maintain all the properties of the emulated Time Division Multiplexing (TDM) circuit; and
- Distributed MAC/PHY architectures, where the control and physical planes are separated and interconnected by a data network, requiring precise timing alignment between remote system elements.

Approaches alternative to Precision Time Protocol (PTP) exist, such as deploying a Global Navigation Satellite System (GNSS) receiver at each network node and obtaining timing and frequency information. (Hereinafter, GNSS is used to indicate any satellite-based timing sources.) However, such approaches are more expensive, require line of sight to the GNSS constellation, and are susceptible to both natural (atmospheric, e.g., rain, fog, snow) and man-made interference (e.g., deliberate jamming, industrial electromagnetic noise). In the majority of cases, distributed clock synchronization becomes necessary.

The process of synchronizing distributed clocks is continuous. A clock is essentially a two-part device, consisting of a frequency source and an accumulator. In theory, if two clocks were set identically and their frequency sources ran at the exact same rate, they would remain synchronized indefinitely. However, in practice clocks are set with limited precision, frequency sources run at slightly different rates, and the rate of a frequency source changes over time and temperature. Most modern electronic clocks use a crystal (usually some form of a quartz) oscillators as frequency sources. The frequency (read analog here) of a crystal oscillator varies due to initial manufacturing tolerance, temperature and pressure changes, and aging. Because of these inherent instabilities, distributed clocks must continually be synchronized to match one another in frequency and phase.

PTP, as defined in IEEE Standard (Std) 1588 (version 3 was published in 2019), provides a method to synchronize time and/or frequency between hosts connected over a data network. PTP is capable of synchronizing multiple clocks to better than 1 microseconds on a data network designed specifically for PTP distribution, addressing even the most stringent applications in existence today, including the emerging 5G wireless networks.

2. PTP Overview

2.1. Phase, Frequency, and Time of Day Synchronization in PTP

PTP supports synchronization of time (phase), frequency, and time of day (ToD) across message-switched networks. These three aspects of PTP operation are commonly confused and used interchangeably, though they do represent different aspects of clock synchronization.

The frequency synchronization (also referred to as *syntonization*) is shown in Figure 1. The local oscillator in the TimeTransmitter (TT) and the local oscillator in the TimeReceiver (TR) have the same frequencies (hence $f_{TR} = f_{TT}$). (Note that this document uses the inclusive terminology for clock naming

per IEEE P1588g, replacing the legacy terms of "master" and "slave"). There may be a certain phase difference (phase delta) between the TT and the TR, i.e., the time when the clock pulses occur are not aligned in time. The frequency synchronization guarantees therefore only the alignment of local oscillator frequency between clocks but does not provide the phase alignment. Frequency synchronization may be achieved using a variety of mechanisms, such as SyncE, GNSS, Network Timing Protocol (NTP), 1 Pulse Per Second (1PPS), E1/T1 interfaces as well as PTP.

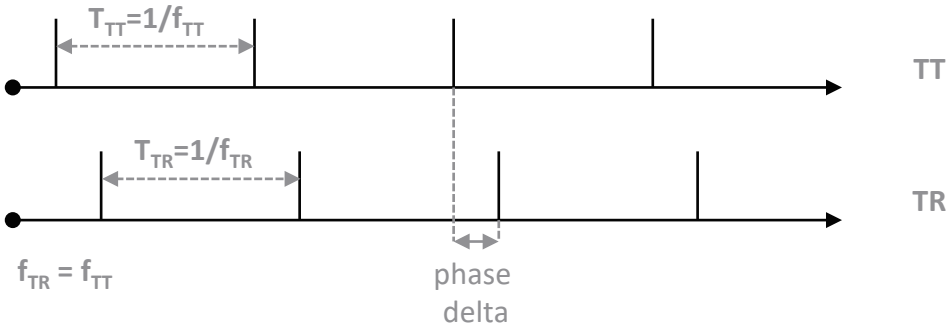

Figure 1 - Frequency synchronization

The phase time synchronization is shown in Figure 2, where at a specific moment of time the oscillators at the TT and the TR become phase aligned. The phase synchronization typically provides frequency synchronization as well, i.e., both oscillators start operating at different frequencies and drift over time until the next phase synchronization event. Note that phase synchronization (even though it is commonly referred to as time synchronization) is not the same as ToD synchronization and the frequency synchronization is optional. Phase synchronization may be achieved using a variety of mechanisms, such as GNSS, NTP, 1PPS interface as well as PTP.

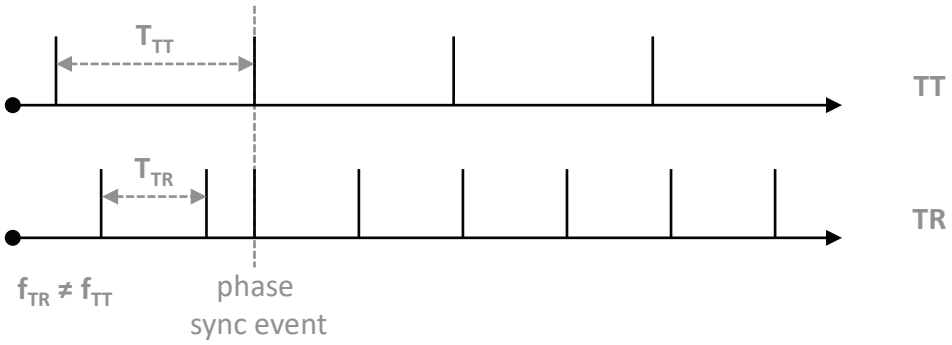

Figure 2 - Phase synchronization

The ToD synchronization (sometimes referred to as *time synchronization*) transfers the current time of day information from the TT to the TR, while not performing any phase and frequency alignment. Effectively, the TT tells the TR that when the given synchronization message is received, the time of day has a specific value. The ToD synchronization is shown in Figure 3. ToD synchronization may be achieved using a variety of mechanisms, such as GNSS, NTP, as well as PTP.

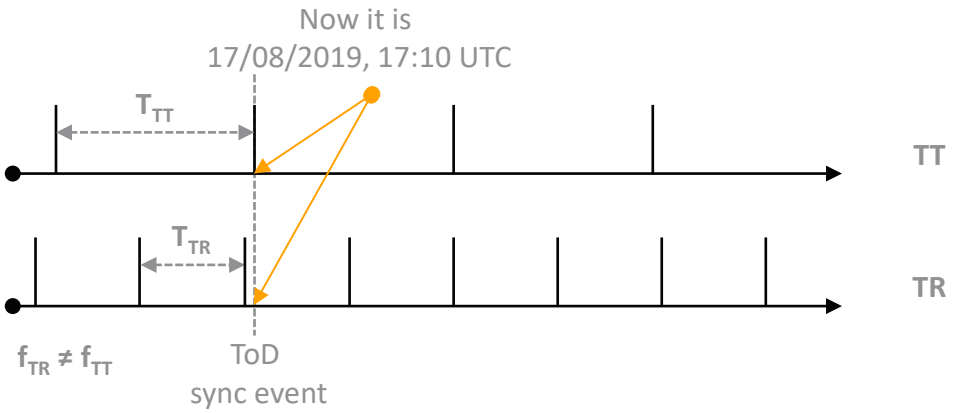

Figure 3 - ToD synchronization

The combination of phase, frequency, and ToD synchronization achievable via PTP is shown in Figure 4, where at some point of time the TT aligns with at least some of the aspects of the TR operation. For example, phase and frequency alignment between the Leader and the TRs may be achieved and then combined with the ToD on the TR being aligned with the TT. For illustration purposes, the phase and frequency synchronization event is shown separate from the ToD synchronization event. This follows a typical PTP operation mode, in which phase and frequency are aligned first, before the ToD is transferred between the TT and the TR.

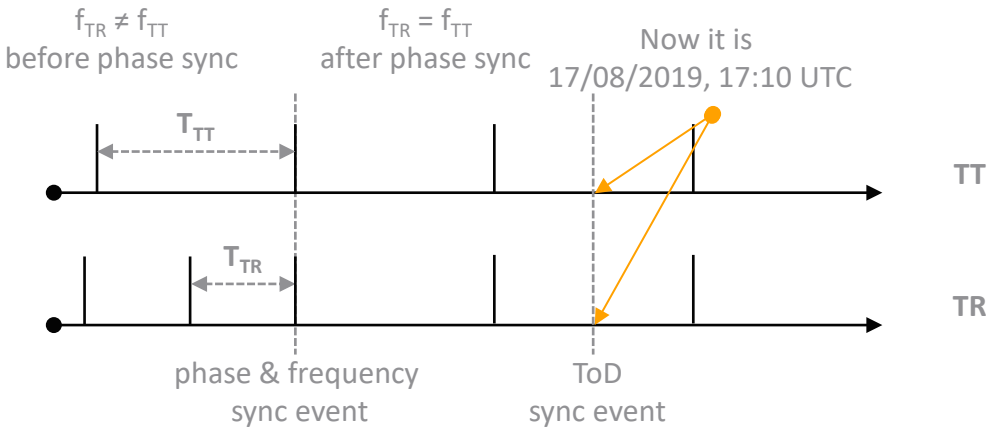

Figure 4 - Phase, Frequency, and ToD synchronization

Some applications require only frequency synchronization, with the primary examples being TDM circuits and TDM circuit emulation as well as synchronous Ethernet. The phase alignment in these particular applications is not required, as long as both ends of the circuit are frequency locked.

Wireless applications, including 4G LTE as well as the upcoming 5G networks, require not only the frequency but also the phase alignment to properly perform coordinated multipoint transmission as well as interference avoidance to achieve the enhanced inter-cell interference coordination.

2.2. PTP Synchronization Process

A heterogeneous network of clocks is a network containing clocks of varying characteristics, such as the origin of a clock's time source, and the stability of the clock's frequency. PTP provides a fault tolerant method of synchronizing all participating clocks to the highest quality clock in the network. IEEE Std 1588 defines a standard set of clock characteristics and defines value ranges for each. By running a distributed algorithm, called the Best TimeTransmitter Clock Algorithm (BTCA), each clock in the network identifies the highest reference quality clock with the best set of characteristics. Each TR listens to the TT Announcement messages containing TT properties, including clock class, reference source, expected precision level (variance, accuracy), configurable clock priority value, etc. Based on such Announcement messages from several TT, each TR may then select the TT with the best properties and designate it as its local Grand TimeTransmitter (GTT).

The highest-ranking reference clock (elected due to its stability, clock properties, etc.) is commonly referred to as the GTT and is used to synchronize TRs. If the current GTT is removed from the network, or if its characteristics change in a way such that it is no longer the best clock from the perspective of the given TR, the BTCA provides a way for the TR to automatically determine the new best reference clock. The BTCA provides a fault tolerant, and administrative-free way of determining the reference clock used as the time source for the entire network.

PTP supports a number of different transport options, with Internet Protocol (IP) unicast and IP multicast being the most commonly deployed ones. For the following functional description, IP unicast communication is assumed. TRs synchronize to the selected TT using bidirectional unicast communication (see Figure 5). The TT periodically issues a Sync message containing the timestamp value of the local clock within the TT when the given Sync message left the TT (the value of "100?" in Figure 5). The TT may also optionally issue a Follow Up message containing the timestamp for the Sync message (the value of "100!" in Figure 5). The use of a separate Follow Up message allows the TT to accurately timestamp the Sync message in devices where the departure time of a message cannot be known accurately beforehand. For example, the collision detection and random back off mechanism of Ethernet communication prevents the exact transmission time of a message from being known until the message is completely sent without a collision being detected, at which time it is impossible to alter the message's content. Other examples include lack of hardware support for PTP in the given device, requiring a software-based timestamping, resulting typically in very high timestamping jitter.

A TR receives the Sync message from the TT and timestamps the message's arrival time using its own local clock. The difference in the Sync message's departure timestamp (carried in the message itself) and the Sync message's arrival timestamp is the aggregate of the TR's timing offset relative to the TT time and the network propagation delay (transmission time between the TT and the TR). By adjusting its clock by the offset measured at this point, the offset between the TT and TR may be reduced to the network propagation delay only.

PTP operates under the assumption that the network propagation delay is symmetric. That is, the delay of a message sent from the TT to the TR is the same as the delay of a message sent from the TR to the TT. Under this assumption, the TR can calculate, and compensate for the network propagation delay. The delay calculation is accomplished by the TR issuing the Delay Request message which is time stamped on departure from the TR (the value of "108" in Figure 5). The Delay Request message is received by and timestamped by TT, and the arrival timestamp is sent back to the TR in a Delay Response message (the value of "108, 112" in Figure 5). The difference in these two timestamps is the round-trip network

propagation delay. The one-way network propagation delay is then calculated by dividing the round-trip network propagation delay by two.

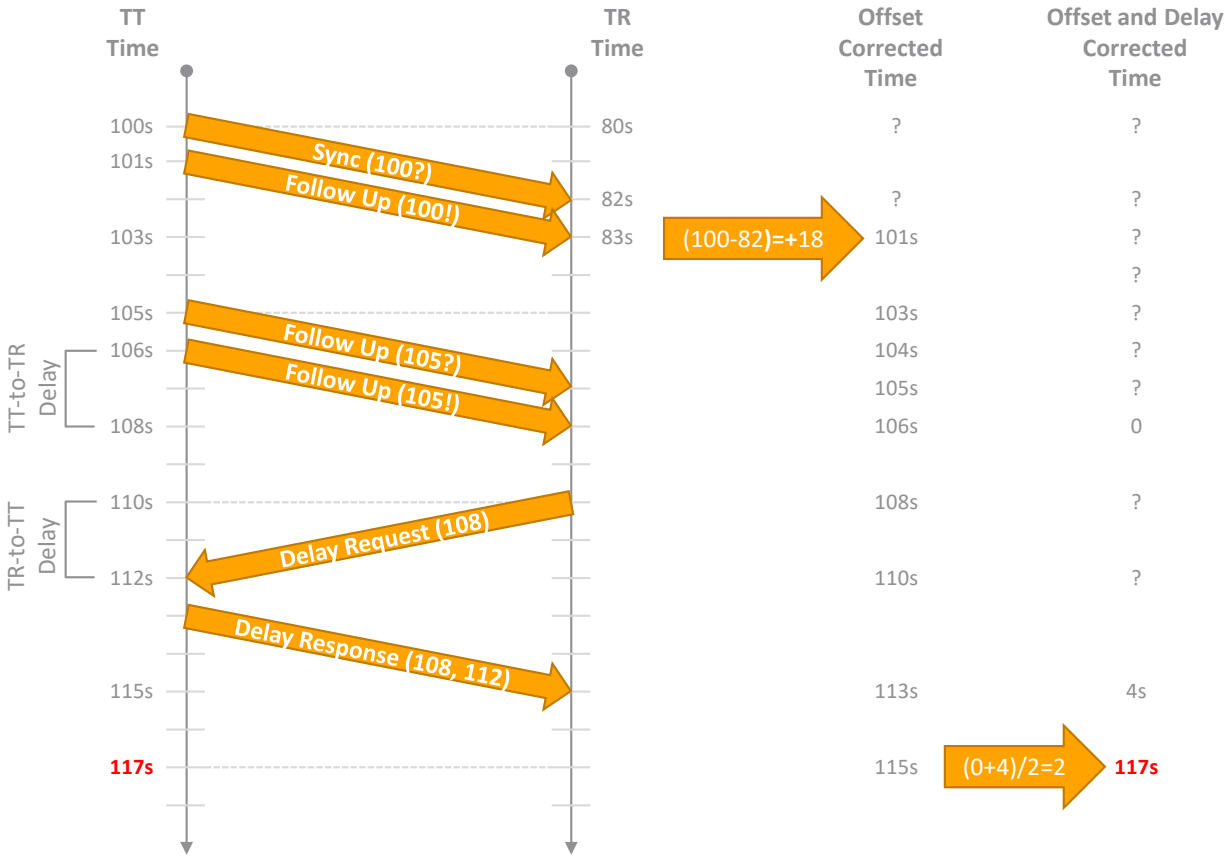

Figure 5 - PTP synchronization process

By sending and receiving the aforementioned synchronization messages, the TRs can measure the offset between their TR and the TT. TRs can then adjust their local clocks by the calculated offset value to match the time of the TT. The IEEE Std 1588 does not include any standard implementation for adjusting a clock; it merely provides a standard protocol for exchanging the synchronization messages, allowing devices from different manufacturers, and with different implementations to interoperate.

Note that the IEEE Std 1588 does not mandate the reference timestamping interfaces in any implementation, as well as does not provide recommendations on the hardware-based versus software-based timestamping process. The highest precision PTP-enabled devices perform hardware-assisted timestamping as close to the actual transmission medium (fiber, coax, twisted-pair, air waves) as possible to eliminate any variable propagation delays from the timestamp calculations.

3. Impact of Network Design on PTP Precision

For the proper PTP operation, the network delay must be symmetric, i.e., the time it takes a PTP message sent by the TT to reach the TR and the time it takes a PTP message sent by the TR to reach the TT, are equal and the same. This assumption is true to a certain degree, depending on the network design, ability

to control the data path through the network, etc.. There are several sources of network delay asymmetry found in networks which are described in more detail in the following sections, including their potential impact on PTP operation. Individual sources of PTP imprecision can be divided into first order (network routing asymmetry, optical transport asymmetry, and in-hub cabling asymmetry), second order (message serialization delay) and third order (wavelength propagation asymmetry) sources, in the order of decreasing impact on the resulting PTP precision.

3.1. Logical Path Assymetry

One of the most common sources of asymmetric delay is associated with the logical (L2/L3) path asymmetry, where the selection of the logical path through the network differs in the ingress and egress directions. An example of such a logical path asymmetry is shown in Figure 6, where the forward path and return path have substantially different number of hops.

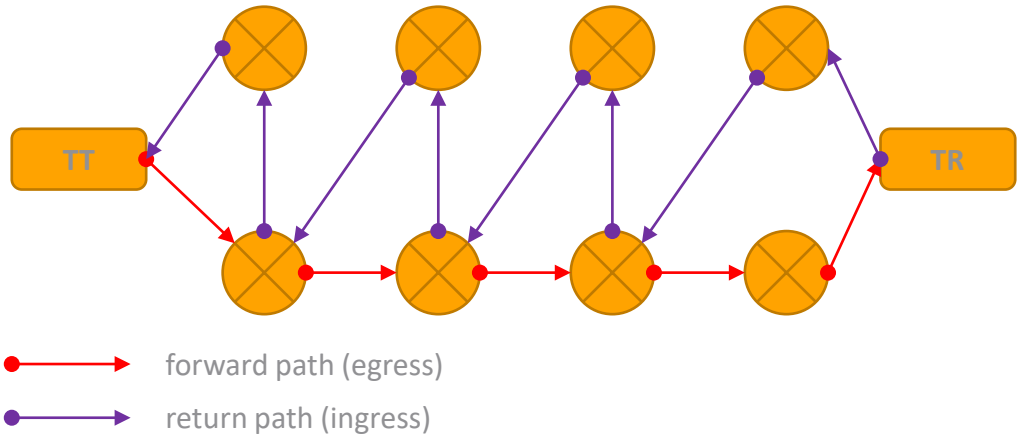

Figure 6 - Logical path ssymmetry

The logical path asymmetry is caused by the operation of the underlying routing protocols, path selection, availability of multiple paths to reach the given destination, path congestion in the given direction, different path cost (weight) between return and forward paths, etc. The logical path asymmetry may be caused by the operation of layer 2, layer 3, or any other logical layer protocols, for example – multiprotocol label switching (MPLS) traffic engineering (TE), resource reservation protocol (RSVP)-TE, etc., forming transport paths through the network optimal in terms of bandwidth, capacity, and/or availability. These protocols do not take delay symmetry into account. Even though latency may be minimized in ingress and egress directions, the latency symmetry may not be maintained, resulting in different forward and backward delays between TT and TR instances, as shown in Figure 6.

To minimize the logical path asymmetry in the network, some form of a path-aware forwarding must be employed, where the path selection decision is taken not only on the link bandwidth, capacity, and/or availability but also on the resulting path delay, something that the existing class of traffic engineering protocols do not take into account. In the case of most non-PTP optimized network designs, the logical path asymmetry is the primary source of PTP imprecision and cannot be disregarded even in the case of first order approximation of path delay symmetry for most common network designs. When the network design prevents the asymmetric routing operation from taking place, this particular delay asymmetry component may be disregarded.

3.2. Physical Path Asymmetry

The physical path asymmetry is the result of different physical link lengths between connected devices, as shown in Figure 7. In this case, the forward and return paths use different cables between interconnected devices, resulting in a different path delay, affecting PTP calculations. In the example shown in Figure 7, the forward path of 120 km and return path of 125 km results in the path delay asymmetry of around 24.47 µs, provided the use of G.652 SMF fiber cable with the refractive index of 1.4677, with 1310 nm center wavelength optical carrier.

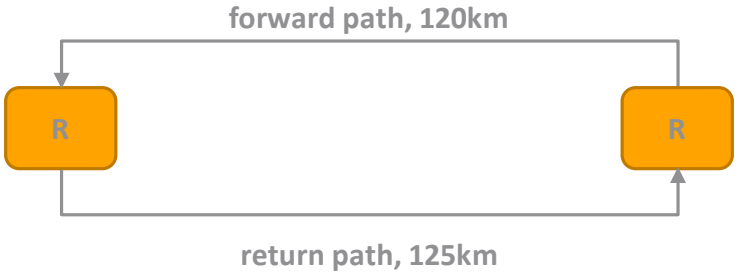

Figure 7 - Physical path asymmetry

The aforementioned physical path asymmetry is relatively easy to detect at the network design and deployment time, when the forward and return path can be measured directly or indirectly, using for example a high precision Optical Time Domain Reflectometer (OTDR). A PTP-optimized network design must account for the potential physical path asymmetry and opt to use the same physical cable between interconnected devices, at best – the very same fiber strand – and operate using Dense Wavelength Division multiplexing (DWDM)over the very same fiber pair, with DWDM channels spaced as close as possible to minimize the impact of any wavelength propagation asymmetry on PTP precision.

The situation becomes more complicated when the forward and return paths use link aggregation (LAG) bundles to increase capacity and/or provide redundancy between interconnected devices, as shown in Figure 8. The two bottom LAG members follow the same primary path (120 km length) but the top LAG member follows a slightly different path (125 km length) and was added to the LAG bundle between devices after the primary path has been constructed and for various reasons (e.g., lack of dark fiber in the previous sheath, etc.) had to follow a different physical path.

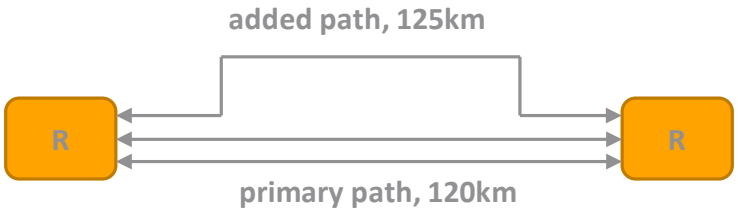

Figure 8 - Physical path asymmetry in a LAG bundle

In the case of any LAG implementation, the specific member within the LAG bundle used to transmit a data message is selected based on the output of a vendor-specific hashing algorithm. Implementation details vary from vendor to vendor and also between platforms, relying on some combination of L2 and

L3 data message fields. PTP messages are subject to the very same LAG bundle member selection. As shown in Figure 8, if the PTP messages flow using one of the bottom LAG members in ingress direction but take the top LAG member in egress direction, even though both devices are connected using a single LAG bundle, the 5 km of physical path difference results in the path delay asymmetry of around 24.47 µs, provided the use of G.652 SMF fiber cable with the refractive index of 1.4677, with 1310 nm center wavelength optical carrier.

3.3. In-hub Cabling Asymmetry

Within the hub, individual devices are typically interconnected using a number of patch cables to fiber patch panels, as shown in Figure 9, where a single router (R) is connected to two different fiber patch panels (A and B) using dedicated uplink patch cables. While efforts are made to use standardized patch cable lengths, there are certain variations in the length of individual cable batches due to the manufacturing process tolerances and post-manufacturing processing. Effectively, the side A and side B patch cables may be of a slightly different length, even if the patch cables from the very same production batch are used.

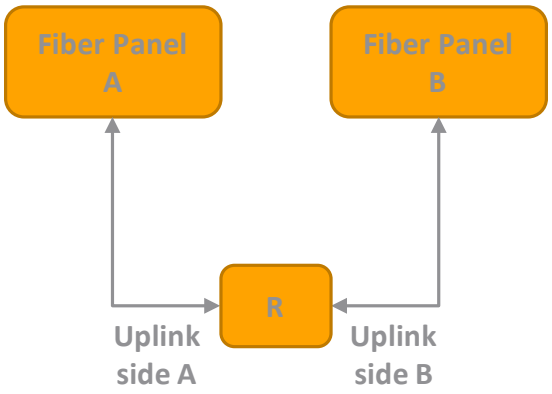

Figure 9 - Single patch cable uplink situation

Furthermore, patch cables get replaced as part of regular maintenance processes, resulting at times in the use of different patch cable lengths, becoming effectively one of the sources of the optical transport asymmetry. Consider the example of using 10 m side A and 20 m side B G.652 SMF fiber patch cables: Assuming that standard 1310 nm center wavelength is used (resulting in the refractive index of 1.4677), the effective path latency is 48.96 ns for side A and 97.91 ns for side B, respectively, resulting in 48.96 ns path asymmetry when (for example) PTP messages reach the router R alongside A and leave it alongside B (asymmetric routing). Each cm of a patch cable length difference results effectively in 0.05 ns of added path delay asymmetry.

The patch cable length control becomes especially crucial in the case of LAG bundles, as shown in Figure 10, where the router R is connected to two different fiber patch panels (A and B) using dedicated uplink patch cable bundles, implementing LAG connections on side A and side B. In the case of any LAG implementation, the specific member within the LAG bundle used to transmit a data message is selected based on the output of a vendor-specific hashing algorithm. Implementation details vary from vendor to vendor and also between platforms, relying on some combination of L2 and L3 data message fields. PTP messages are subject to the very same LAG bundle member selection. Effectively, this means that some PTP messages may traverse one LAG bundle member, while other PTP messages traverse other LAG

bundle members. From the logical-layer perspective, LAG does not introduce any message-dependent latency, though the underlying cable plant may, especially when individual LAG member cables are of different length.

Using the same G.652 SMF fiber patch cable configuration example, assume that one of the LAG bundle members is 10 m long and another one – is 20 m long, resulting in 48.96 ns latency asymmetry when operating across a single LAG bundle connecting the given device to the rest of the network. Such added latency is very hard to troubleshoot since it requires validation of the cabling plant comprising individual physical and logical connections, something that cannot be done easily from within the device itself and under live traffic conditions.

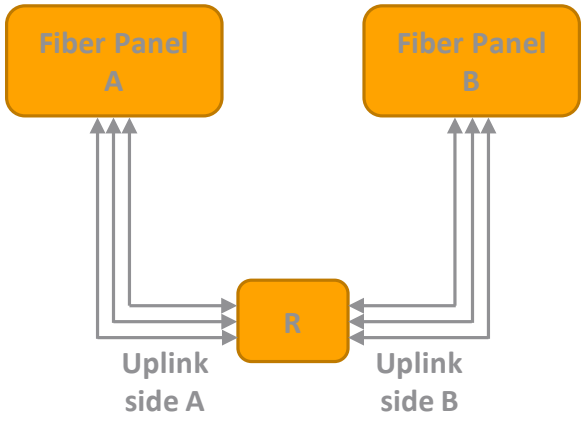

Figure 10 - Multiple patch cable uplink situation

3.4. Message Serialization Delay

The message serialization delay (also referred to as a transmission delay) is incurred at any intermediate network elements in the data path between the TT and TR instances, when PTP messages are stored-and-copied between links and network element buffers. This includes the copying over internal links in such network elements, including router backplanes/switching fabrics.

The message serialization delays come into play when PTP messages traverse any intermediate network elements, being switched between ingress and egress interfaces on the way from the TT and TR instances. As the PTP messages are placed into interface queues, they may suffer from additional delay associated with the queue occupancy at the given moment of time, having to wait for their time to be transmitted on the egress interface. Given that such ingress / egress queues exist at several different locations within any intermediate network elements, the aggregate message serialization delay observed for PTP messages is non-zero, and depends on the interface data rate(s), the size of the message already in the egress queue and subject to transmission, message size distribution, etc. The actual message serialization delay can be assessed only in statistical terms, and it is at best equal to the half of the average message size serialization delay at the given interface speed. For example, if the given 10GE egress interface supports Jumbo frames of 9000 B, with the average messages size of around 1500 B, a typical serialization delay is on the order of 1.2 µs, as shown in the table below. With the average message sizes increasing, the serialization delay observed by the PTP messages increases as well.

Packet size	64 kbps	1 Mbps	10 Mbps	100 Mbps	1 Gbps	10 Gbps
64 B	8 ms	0.512 ms	51.2 µs	5.12 µs	0.512 µs	51.2 ns
512 B	64 ms	4.096 ms	409.6 µs	40.96 µs	4.096 µs	409.6 ns
1500 B	187.5 ms	12 ms	1.2 ms	120 µs	12 µs	1.2 µs
9000 B	1125 ms	72 ms	7.2 ms	720 µs	72 µs	7.2 µs

To address some of the negative aspects of the serialization delay in the intermediate network elements, it is recommended that the Transparent Clock (TC) function is enabled where possible, allowing to compensate for any residency time within the intermediate devices, including the message serialization delay. The TC function within a specific network node measures and adjusts for packet delay when traversing the given network node. Effectively, the TC function measures the residency time for PTP packets within the given network node (when transiting between the ingress and egress interfaces) and adjusts the correction field within the PTP messages, allowing the TR to account for residency times when performing the necessary calculations.

3.5. Wavelength Propagation Asymmetry

Even though light waves traverse the fiber medium at a very large velocity, this velocity is not infinite and what is more – the propagation velocity is wavelength (frequency, or light color) dependent. For example, considering an ideal point-to-point (P2P) link, where one direction uses wavelength L_{down} (with the associated effective refraction index N_{down}) and the other direction uses wavelength L_{up} (with effective refraction index N_{up}), the delay asymmetry due to the difference in light propagation velocity between both directions is given by the following simple equation $0.5 – N_{down} / (N_{up} + N_{down})$. Using a 500 km long transport link and assuming the use of 1529 nm and 1560 nm DWDM channels, the resulting delay asymmetry is equal to roughly 103 ns, with the delay asymmetry of 0.00011.

In the case of ideally symmetric transmission channels, the delay asymmetry is equal to zero. The larger the wavelength difference between both wavelengths on the given link, the larger the delay asymmetry, though it is typically requires accounting for a third order PTP imprecision. Note that in most network implementations for short and medium reach links using standard dual-fiber IEEE 802.3 Ethernet optics, the wavelengths in both directions (forward and back) are the same, hence the resulting delay asymmetry may be simply neglected. The emerging use of bidirectional links (single fiber strand used for both communication directions) will bring back the need to take wavelength propagation asymmetry into design consideration when using PTP over such links.

4. Deployment Recommendations and Conclusions

The best PTP network designs reduce the number of hops between GTT and TRs, primarily to minimize any of the sources of path delays asymmetry covered in the previous section. Where the large number of intermediate network segments between the GTT and TRs cannot be avoided, network planners should consider taking advantage of the following options.

- Enable, where possible, the Transparent Clock (TC) function on any of the L2/L3 intermediate network elements, to minimize the impact of device residency times, packet serialization, inter-device jitter, etc.
- Deploy, where possible, the Boundary Clock (BC) as either a function on the L2/L3 intermediate network element or a dedicated hardware. The BC segments the PTP clocking domain, decreasing the number of hops, and improving jitter tolerance. Using BC also eliminates the need

for installation of additional GNSS receivers at a larger number of locations, lowering the overall cost of deploying timing infrastructure into the network.

- Use any available traffic engineering mechanisms to avoid the following:
  - Any links using different wavelengths on the same fiber strand (bidirectional PHYs);
  - Any LAG links with individual members of varying lengths; and
  - Any paths with asymmetric routing / switching.
- Inspect patch panel connections for fiber length deviations and normalize the lengths, where possible.
- Avoid mixing media for network element connections, i.e., do not mix copper and fiber uplinks from the same device.
- Avoid mixing interface speeds for network element connections, i.e., do not use 1G and 10G optics for uplink purposes from the same device.

Given the predominant use of LAG links in production networks today, it is likely not possible to avoid the traversal of any LAG links altogether. However, the selection of PTP-aware hashing algorithms directing all PTP traffic to one and the same LAG member in both directions may eliminate some of the potential negative impact of the LAG links on the PTP operation.

Network planners also need to take into account the PTP scaling numbers. Each GTT and BC support only a specific number of client clocks with particular PTP message rates, requiring the design in which additional GTT and/or BC instances are added on-demand, to increase the overall capacity of the PTP infrastructure. This is especially critical for the emerging microcell and picocell mobility applications, in which the number of TRs increases exponentially compared to the previous generations.

Individual TRs typically support the negotiation of PTP message rates in the function of the precision requirements. Effectively, even though the GTT / BC may be configured to support the maximum message rates, a TR may request lower message rates meeting its local precision requirements. The rule of thumb is that the higher the required precision, the higher the resulting PTP message rates need to be supported (specifically, the Delay Request and Delay Response rates). To avoid the overload of the GTT and/or BC elements due to the aggregate message rates, it is recommended that individual TRs be configured with the Delay Request and Delay Response message rates meeting their precision requirements.

Finally, it is also important to properly validate and test all of the equipment participating in the transfer of PTP messages in the network. The use of carrier class hardware, software, and management components can provide the highest level of performance, reliability and serviceability of the resulting PTP infrastructure. The critical PTP elements, such as GTT and BC, should be deployed with redundant power supplies, redundant clock modules, and where needed – with dual GNSS receivers to avoid any potential lightning exposure and minimize the down time. Properly designed traffic class of service (CoS) policies need to be deployed on all intermediate network elements to minimize jitter and delay, prioritizing PTP traffic over other (non-critical) service frames.

5. Abbreviations and Definitions

5.1. Abbreviations

1PPS	1 Pulse Per Second
BC	Boundary Clock

CoS	Class of Service
DWDM	Dense Wavelength Division Multiplexing
TR	TimeReceiver
GTT	Grand TimeTransmitter
GNSS	Global Navigation Satellite System
HCR	Hub Core Router
HS	Hub Site
IP	Internet Protocol
L2	Layer 2
L3	Layer 3
LAG	Link Aggregation
TT	TimeTransmitter
LTE	Long Term Evolution
MAC	Media Access Control
MPLS	Multi Protocol Label Switching
NTP	Network Time Protocol
OTDR	Optical Time Domain Reflectometer
P2P	Point to Point
PHY	Physical Layer
PTP	Precision Time Protocol
RSVP	Resource Reservation Protocol
SFP	Small Format Pluggable
SMF	Single Mode Fiber
SyncE	Synchronous Ethernet
TC	Transparent Clock
TDM	Time Division Multiplexing
TE	Traffic Engineering
ToD	Time of Day

6. Bibliography and References

IEEE Std 1588, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, version 3, published 2019

IEEE Std 802.3, IEEE Standard for Ethernet, published 2018

RFC 5905, Network Time Protocol Version 4: Protocol and Algorithms Specification, June 2010

G.8275, ITU-T Recommendation on Architecture and requirements for message-based time and phase distribution, published 2017

G.8275.1, ITU-T Recommendation on Precision time protocol telecom profile for phase/time synchronization with full timing support from the network, published 2016

G.8275.2, ITU-T Recommendation on Precision time protocol telecom profile for time/phase synchronization with partial timing support from the network, published 2016

G.8265.1, ITU-T Recommendation on Precision time protocol telecom profile for frequency synchronization, published 2014

7. Acknowledgments

The authors would like to thank the following reviewers for their valuable input and comments received during the development of this document: Ryan Tucker, James Graves, Thomas Dudley, Justin Hattaway, and Matt Rocker.

A Secure and Scalable Code Signing System

A Technical Paper prepared for SCTE by

Tat Chan, Distinguished Systems Engineer, CommScope, SCTE Member
6450 Sequence Dr.
San Diego, CA 92121
tat.chan@commscope.com
+1 858 404 3252

Ting Yao, Software Engineering Director, CommScope, SCTE Member
6450 Sequence Dr.
San Diego, CA 92121
ting.yao@commscope.com
+1 858 404 3084

Alexander Medvinsky, Engineering Fellow, CommScope, SCTE Member
6450 Sequence Dr.
San Diego, CA 92121
sasha.medvinsky@commscope.com
+1 858 404 2367

Table of Contents

List of Figures

1. Introduction

Code signing is a cryptographic mechanism to ensure the authenticity of software such that only legitimate code can be executed on a target system. It is one of many tools necessary to prevent malicious actors and attackers from running modified or arbitrary code of their choice. Code signing systems are designed and built to protect the code signing keys and manage the signing operations and permissions.

In this paper we discuss a code signing system with multiple layers of physical and network security that addresses the security issues associated with managing and protecting code signing keys. We further discuss fine-grained multi-level access and permission control for company/user accounts, signing key access, and product-specific configurations. We also describe tracking and reporting features that enable tracing of the source(s) of signed malicious code and enable detection of abnormal signing activities as well as various mechanisms that could mitigate the risks.

2. Code Signing System

2.1. Importance of Code Signing

Integrity-protected software typically goes through four stages from creation to execution as illustrated in the following figure. It includes software development, software build, code signing, and execution. The software development system includes the software source control system. It allows developers to design and implement software based on functional requirements. The software build system takes the source code developed, compiles and combines it to generate a software package to be ready for signing. The software development system and build system may be combined as a single system but there are also cases where they are maintained and hosted separately. The resulting software package is then submitted to the code signing system for signing. Either the code signing system or the software build system generates the final signed package, which is then deployed to the target execution environment. In the target execution environment, the signature of the software package is verified before it is accepted for execution.

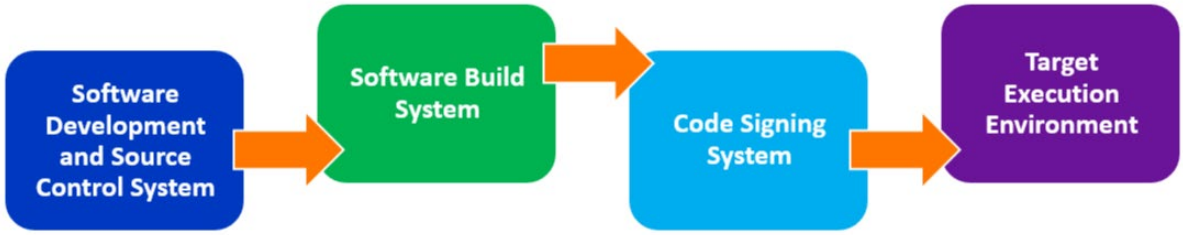

Figure 1 - Four stages of software development and execution

All four stages must be secured to prevent attackers from compromising the integrity of the software. Adversaries typically attack the weakest link in the system. In this paper we are going to discuss a secure code signing system in more detail; however, the security of the other stages must also be adequate. In other words, having a secure code signing system without protecting the software build process or target

execution environment will only provide a false sense of security. Software development life cycle (SDLC) addresses the first and second stages shown above. To secure the SDLC, the readers are referred to the secure software development framework (SSDF) recommended by NIST [1].

In December 2020, a group of suspected nation-state cyber attackers managed to compromise the SolarWinds software development platform and injected SUNBURST malware into properly signed Orion products, allowing unauthorized access to networks of a wide range of government and private sector institutions [2]. It is interesting to note that in this case the infected software package was legitimately signed. This incident demonstrates that securing a software development and build system is as important as securing a code signing system.

Shortly after the discovery of the SolarWinds SUNBURST attack, another malware named SUPERNOVA was also reported [3]. In this case, however, the malicious DLL module was not part of a signed SolarWinds package. The Supernova DLL is a web shell which ended up inside the platform after hackers exploited a critical vulnerability in product installations reachable over the public web. The SUPERNOVA malware is designed to masquerade as a legitimate SolarWinds web service and is able to execute commands that are sent to it by the attackers. Since the DLL module was not signed, it is only allowed to execute if the target environment does not enforce signature verification. Such unsigned malware attacks are blocked if code signature verification is enforced. It is therefore important to ensure that only properly signed software is accepted in the target platform. It is also important that the operating system (OS) is hardened at the target system to prevent modification of files after installation or deployment.

Software running on any electronic devices or platforms is designed and developed for a specific purpose, but unauthorized and concealed modifications to that software can cause significant damage or harm to its user or other parties which may be targeted by an adversary. Some of the numerous examples of applications that have been recently compromised include:

- Mobile banking applications [12], [13];
- A video streaming client application [14];
- Applications that are intended to provide a security function such as anti-virus or password manager [15], [16].

It is crucial that these applications remain intact after they are released by their publishers, not modified, replaced, or corrupted by anyone intentionally or unintentionally. To ensure that only legitimate software or scripts are executed on a device or a platform, a digital signature is generated over an executable using a private signing key. The signature and the executable are packaged together for delivery to the target platform. In order to ensure that the code is not modified, this digital signature is verified on the target platform using the corresponding verification key.

There are in general two main approaches where signed code can be utilized: secure code download and secure boot. With secure code download, when a code image is downloaded to a device, its signature is checked before the device will accept the download and save it locally into persistent storage. This prevents any code download that is not signed properly with the correct signing key. When the device reboots, however, the code signature may not be checked again. For devices with silicon that do not support secure boot, limiting physical access to the device and hardening against remote installation of unauthenticated software (including of course, implementing secure code download) may be the best option available. Some older cable modem products fall into this category.

However, sophisticated attackers can and do often find ways to bypass code authentication and install malicious software. A few of the many examples that allow malicious software to be installed into IoT (Internet of Things) and mobile devices without physical access are described in [4], [5] and [6]. An example where physical access or proximity is initially required for hacking a smart TV device is described in [7]. Malicious software can enable attackers to take control of a device, access device storage for any sensitive information, and explore both internal and external device interfaces in order to launch further attacks on other connected devices.

The limitation of secure code download is the lack of integrity verification, authentication, or authorization of software that has already been installed on a device. If an attacker gains access to the device and manages to modify the code responsible for verifying the code download signature, then any code can be downloaded and accepted. Alternatively, if a code signature is only checked after a download, an attacker with access to the device may be able to overwrite persistent storage with a malicious unsigned code image. Secure boot, on the other hand, enforces that a device is booted from trusted authenticated software all the way from boot code, to platform, OS, and applications.

In an ideal scenario, a device with secure boot will start up from hardware-protected boot code. For instance, the boot code may be programmed into read-only memory that is not modifiable unless someone physically swaps out the memory module. The boot code includes verification code that will load and verify the next boot stage, using an embedded verification key that cannot be modified. From then on, each boot stage is responsible for verifying the next software layer before executing it. This forms a strong chain of trust for all the software running on the device.

Code signing normally utilizes asymmetric cryptography, such as RSA or elliptic curve algorithms. There is a private signing key and a public verification key. The signer responsible for signing the code owns the signing key, while the verification key is to be used by the verification software. The verification key needs to be integrity-protected on the target platform, so that it cannot be modified or replaced with another key. The signing key needs to be kept confidential and stay at the signing authority. It must not be exposed to the target execution environment. If it is compromised, the attacker can use it to sign any code and defeat the purpose of code signing altogether. The verification key, on the other hand, must be integrity-protected at the target execution environment to prevent modification. Commonly, a signed application is bundled with a corresponding signature verification key and a chain of digital certificates. In those cases, target execution environment must be secured from modifications to the list of root CA certificates which are trusted to validate signed code images.

2.2. Scalable Code Signing Architecture

As explained above, it is of utmost importance to protect the code signing key. If the code signing key is stored as a data file in a developer's laptop, it is susceptible to exposure and can be used to forge signatures of malicious files that will pass validation checks. Multiple copies of the signing key may be created, and it will be impossible to determine valid signatures as compared to forged signatures of any given version of the code. Even if the signing keys are well protected, code signing infrastructure with inadequate physical, network, or system security may be susceptible to attacks. If attackers manage to hack into the code signing system, even though they may not be able to extract the keys, they may still be able to direct the system to sign malicious code as they wish. For example, in 2012 Adobe reported that one of its code signing certificates had to be revoked, as hackers managed to gain access to a build server to access Adobe's code-signing infrastructure to sign malware [8]. Adoble had to revoke the CVC and issue a new one for affected products. There have been similar attacks against certificate authories (CAs),

[9], [10], [11]. Therefore the risk of attackers compromising a code signing system should not be underestimated.

To design a secure code signing system, one must address many security issues. In the following, we explore one possible design of such systems. The code signing system proposed here utilizes a cluster of centralized servers with a set of hardware security modules (HSMs) and multiple layers of physical and network security. All code signing and encryption keys are protected by HSMs. All network devices and physical hosts are hardened according to the latest security guidelines of the industry. Regular and extensive network scanning, and penetration testing should be in place to minimize any remaining vulnerabilities.

A conceptual view of this code signing system is shown in the figure below. The system consists of components that manage the organizations, users, and signing configurations (as explained further in this paper). Crypto operations of the code signing system are performed within HSMs which host and protect the code signing and encryption keys. To address different signing needs and use cases, the system should provide both a human GUI (graphical user interface) as well as machine and API-based interfaces. It should also maintain transaction logs of all signing operations that were performed for traceability and accountability.

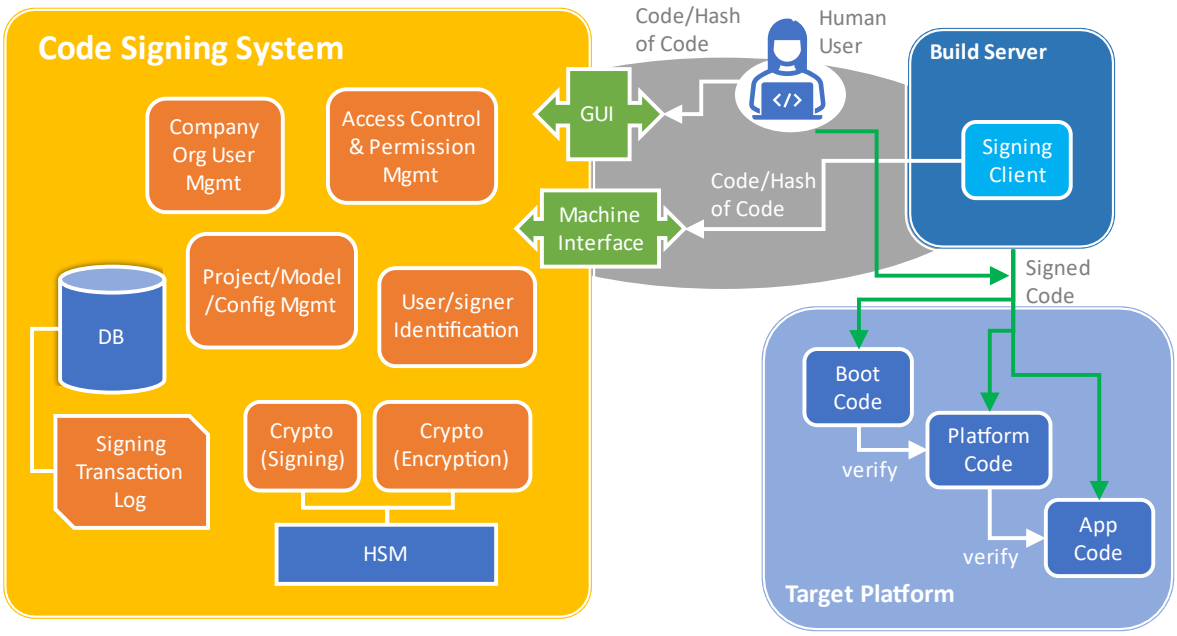

Figure 2 – Conceptual view of a code signing system

On the target device or platform, signed code images of various stages in the software stack are verified in sequence to ensure authenticity. For example, boot code is verified using hardware-protected firmware; boot code then verifies the platform code. The platform code in turns verifies the applications – either individually after each one is launched or as a signed bundle, e.g., when a read-only file system is mounted.

To support developers and software build systems around the world, the code signing system should be designed with reliability in mind and therefore should include failover and high availability features, ensuring that the code signing system is available whenever it is needed. It is not uncommon for one company to include development teams in multiple time zones – for example in North America, Europe/Middle East and the Far East – requiring the code signing system to be available around the clock. High availability is particularly crucial for automated software builds. A build server operated by a large device manufacturer may be running overnight with tens or even hundreds of software builds for numerous product lines and customers that all require a code signature. A geographically diverse disaster recovery and business continuity plan should be in place to further ensure system recoverability in case of natural disasters and other emergency scenarios.

The same code signing system may also provide users with a code encryption capability which provides additional protection against reverse engineering and static code analysis. Code encryption may be combined with code signing into one operation or both may be supported by separate configurations.

2.3. User Interfaces

An effective code signing system must support multiple interfaces for code signing, including interactive browser-based interfaces and automated machine-to-machine (M2M) interfaces. With the browser-based interface, registered users can log into the code signing system with any web browser. A code signing system should provide a simple GUI for users to submit code for signing and/or encryption against a specified configuration that has been prepared by a system admin ahead of time. Users should only be allowed to perform signing operations for which they are authorized. The same browser-based interface may also provide users with administrator and managerial privileges to perform administrative and other functions, such as granting or removing signing privileges to users for each signing configuration that they manage.

To support scenarios such as nightly software builds, a code signing system should provide automated M2M interfaces, eliminating the need for human intervention. For best security practice, M2M interfaces should use hardware-based credentials for client authentication, such as hardware crypto tokens (e.g. USB crypto token) or TPMs. For example, a unique private key and certificate are issued to the hardware token and provided to the user for installation into a client machine. Alternatively, a client key pair is created on the client TPM, and a corresponding client certificate is issued and loaded accordingly. The private key/certificate in the client can then be used to set up a secure session (such as TLS) with the code signing server with mutual authentication. The use of hardware-based protection for client authentication ensures that the client credential cannot be copied. An adversary would have to take physical possession of that token or the code signing client in the TPM case in order to gain unauthorized access to the code signing system, which should result in revocation of the corresponding certificate. In addition, the code signing client should be hardened and patched frequently to remain secure. Otherwise, attackers may exploit vulnerabilities in the client and submit malicious code to the code signing server.

In some cases, the use of hardware-based credentials may not be feasible or required, such as for a cloud-based code signing client running in a DOCKER container. For such cases, the client credentials may be protected using other mechanisms, such as key obfuscation techniques. Cryptographic keys and corresponding functions implementing a cryptographic algorithm can be mathematically transformed using a class of techniques called whitebox cryptography [17].

Depending on the respective security policy, a signing operation may be allowed or disallowed on the interactive GUI-based interface or M2M interface. For instance, boot code is typically the first layer of a

software stack and is more security sensitive than the rest. Boot code is usually protected using a hardware protection mechanism. For example, it can be part of the ROM firmware or BIOS that is not modifiable. Alternatively, secure systems-on-chips (SOCs) come with one-time-programmable (OTP) memory in which a root public key hash can be programmed. ROM code or a specialized security processor on the same SOC verifies a root public key based on this hash, and then verifies boot code with the validated root public key. If boot code is somehow compromised, all the subsequent layers of a software stack cannot be trusted. Also, the frequency of boot code updates is significantly lower compared to higher layers of the software stack, such as platform or application code. A policy may be in place to allow boot code signing for a product only on the GUI interface by human users, and not on the M2M interface. This way, the boot code signing activities can be traced to the specific human user, rather than a code signing client machine. Applications that are updated frequently may be allowed to be signed via both GUI and M2M interfaces.

Automated machine-to-machine interface to a code signing system requires setup and security hardening on a client machine. Even when it is permitted, individual users may still require GUI-based access when a code signing machine with M2M access is not yet set up or when the effort to set it up is deemed unnecessary.

2.4. Access and Permission Control

Access and permission management of the code signing system may be addressed using different approaches. An example of a role-based permission control model is described in this section. In this simple example, three user roles are defined within the code signing system: administrators, managers, and users as shown in Figure 3.

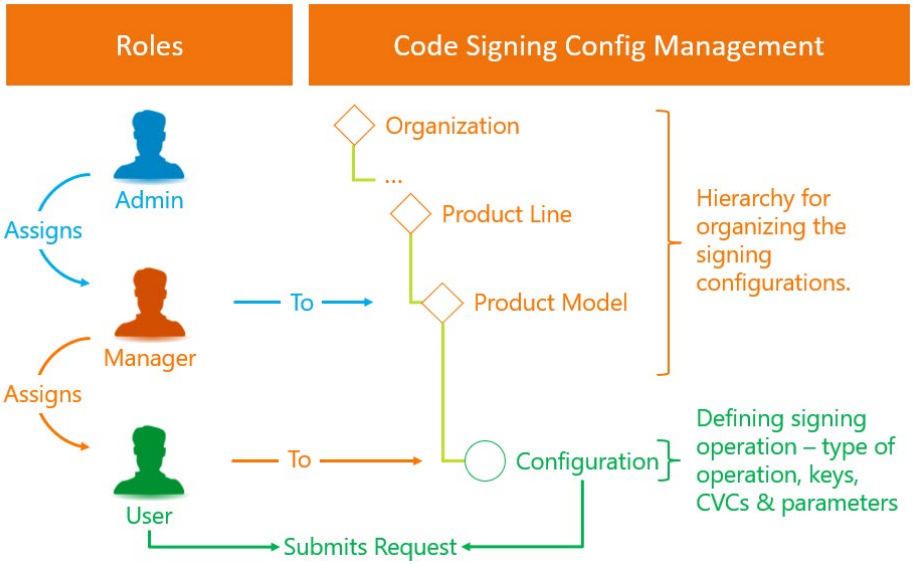

Figure 3 – Role-based access and permission management

Administrators are responsible for managing the code signing keys and certificates in the system, including the generation, deletion, replacement, and renewal of these components. They work with the customers or users of the system to determine the type of code signing that is required. For example, a

product may have boot code that utilizes a SOC vendor proprietary signing format, whereas the platform code adopts a standard PKCS#7 code signature format. The boot code may require just a single key pair, whereas the latter requires a code signing certificate (CVC) hierarchy to be defined and generated. While the signing keys are protected and kept confidential at the code signing server, the corresponding verification keys (or certificate trust chain as in the case of PKCS#7 signatures) are provided to the user for embedding in the target platform so that verification can be performed.

Administrators are also responsible for specifying the code signing format, keys used, and any associated parameters for a code signing project. These parameters can be defined in a code signing configuration in the system. An example of a code signing configuration is shown in the following figure.

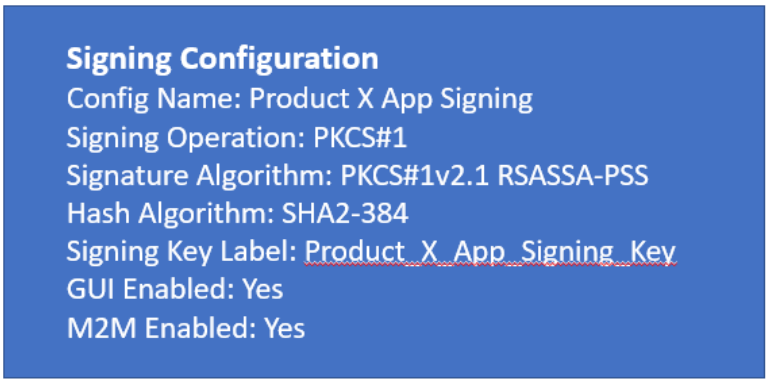

Figure 4 – An example signing configuration

This example signing configuration is defined for signing applications for "Product X". The signing operation is PKCS#1 with signature algorithm being PKCS#1v2.1 RSASSA-PSS. The hash algorithm parameter specifies that SHA2-384 is used. The signing key label points to the signing key in the HSM. For this configuration, both GUI-based and M2M-based accesses are allowed. Other types of operations may require more or less parameters. For example, a PKCS#7 signing operation requires the code signing certificate to be specified, and there may be other attributes such as the signing time.

Code signing configurations are typically defined at the beginning of a project. As shown in Figure 3, a hierarchy structure can be used to organize the code signing configurations under different organizations, product line, product model, etc. Occasionally, there may be a need to update some of the configuration parameters, e.g., incrementing the security version number.[1] These changes are handled by administrators and are not editable by users with other roles. This prevents any intentional or unintentional modifications of the signing parameters and enforces strict control of the signing configurations.

Managers assign or revoke user signing privileges to code signing configurations for users within their organization. Typically, a team lead or a project manager is most familiar with the software development tasks assigned to each team member and his or her corresponding code signing needs. Therefore,

[1] Security version number typically instructs an execution platform to block previous signed software releases labeled with a numerically smaller security version, as they may contain critical security vulnerabilities that were fixed in the latest release.

administrators can assign them as the managers for the respective code signing configurations, so that they can manage user permissions accordingly.[2]

Typical users are developers or project team members who submit code for signing to the code signing system under the signing configuration they are authorized to use. Parameters needed for a signing operation are already defined within a code signing configuration. A user only needs to select a pre-defined configuration and submit the code that needs to be signed. A user is only shown the signing configurations that he/she is authorized for. This not only enables fine-grained access control for different vendors, groups, products, and projects, but also makes it simple and straightforward for users to sign code.

This example of role-based permission control allows for simple and effective user permissions and signing configuration management.

2.5. Flexible Key Management

In a code signing system that follows best security practice, as described in this paper, all signing and encryption keys are protected by HSMs. To ensure sufficient security and to provide an audit trail for key generation, keys are typically generated in multi-person audited events referred to as key ceremonies.

If there are multiple HSMs used, e.g. one set of HSMs is used for production while another set of HSMs is used for disaster recovery, the keys generated and loaded to one set of HSMs need to be loaded to the other set as well. This can be achieved either with a key ceremony, or alternatively by using secure backup/restore features of the HSMs. All these activities, while secure, are very labor-intensive and time-consuming.

2.5.1. Creating New Code Signing Keys and Configurations

To address this operations intensity and reduce operational overhead, an advanced key management feature can be implemented in the code signing system to handle key generation, export, backup, and import. An administrative function may allow an administrator to create new keys as session objects in an HSM.[3] In Figure 5, an admin issues a command to generate a new signing key in step 1. The HSM generates the new key as a session object in step 2. The newly generated key is then wrapped by the HSM (meaning encrypted) with a global wrapping key (step 3) and stored in the code signing system's database (step 4). This database can be automatically replicated to a disaster recovery site, for example, eliminating the need to back up and restore HSMs. After new keys are generated, the system provides a way to allow the administrators or authorized users to download the corresponding verification keys (step 5). The user can then embed the verification key in the target environment, such that signed code can be verified accordingly. If a CVC is involved, and the CVC CA is not part of the code signing system, the admin will download a certificate signing request (CSR) in step 5. The CSR can then be submitted to the

[2] Managers of a code signing system might not be managers in the company. The actual project or team manager might delegate the role of a code signing system manager to team leads or other trusted or more senior people on his team.

[3] A session object typically lives in the volatile memory of an HSM and is cleared out on a reset or when the current session with the HSM is closed. Session objects can be created and destroyed much more frequently than persistent HSM objects since they do not use up limited persistent storage in an HSM and do not wear out this persistent storage which is typically provided as flash memory.

CVC CA such that the CVC can be issued (step 6). After that, the admin uploads the CVC back to the code signing system in step 7 to be associated with the signing key.

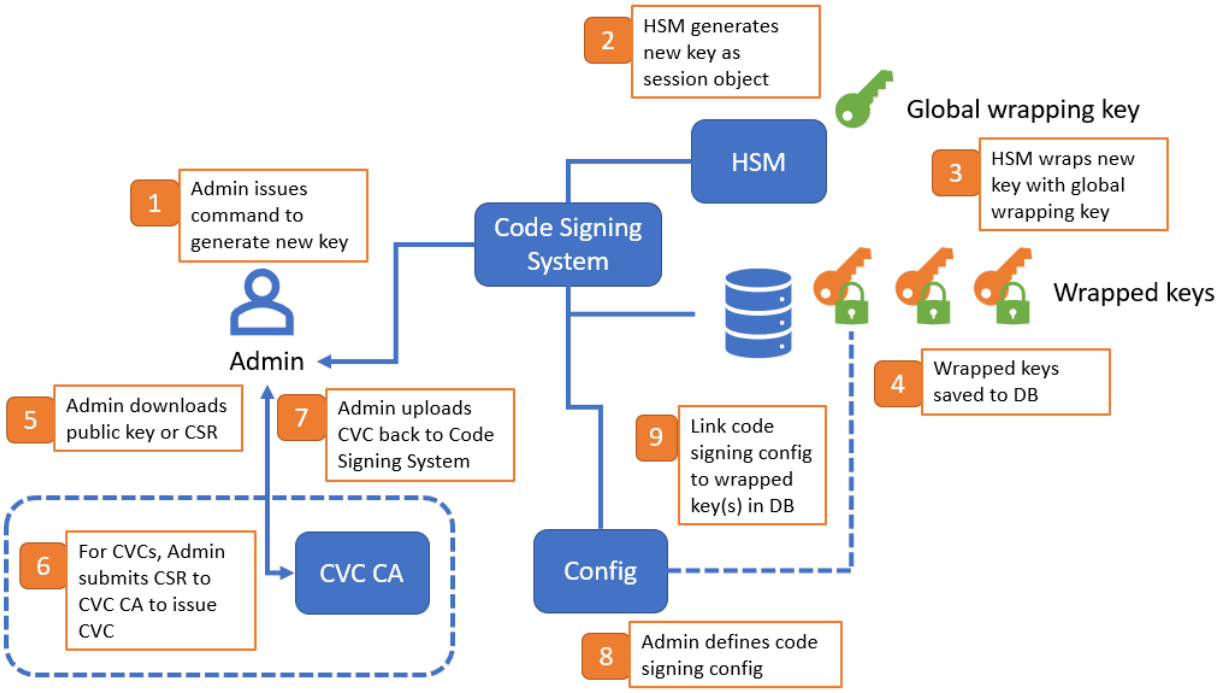

Figure 5 – Creating new code signing keys and configurations

When defining a code signing configuration, the system should allow an administrator to specify keys from the HSM, as well as wrapped keys stored in the database. This is illustrated in steps 8 and 9 of Figure 5. Note that this may be implemented in a way that is transparent to the administrator while a code signing configuration is defined. A code signing configuration can be defined by referencing the signing key the same way, regardless of whether it is permanently loaded to the HSM or wrapped and stored in the database.

2.5.2. Using Wrapped Code Signing Keys

In Figure 6, an end user submits a code signing request (step 1) and if a specified configuration involves a wrapped key that is not already unwrapped to the HSM, the system will automatically retrieve the wrapped key from the database and unwrap it to the HSM as a session object (step 2). The code signing request can then be processed (step 3). When finished, the code signing system can delete the session object from the HSM (step 4). In general, the code signing system can implement a policy that best fits the usage pattern and it should be transparent to the code signing system users. In one extreme, all wrapped keys are unwrapped when the server starts up, and they will stay in the HSM as long as the server is running. This assumes that the HSM has sufficient capacity to host all the wrapped keys. In another extreme, a key is only unwrapped when it is needed, and will be removed as soon as the operation is completed. This, however, may incur a significant overhead and impact the overall code signing

performance and turnaround time. In practice, an intermediate policy may be adopted to strike a balance between the two extremes.

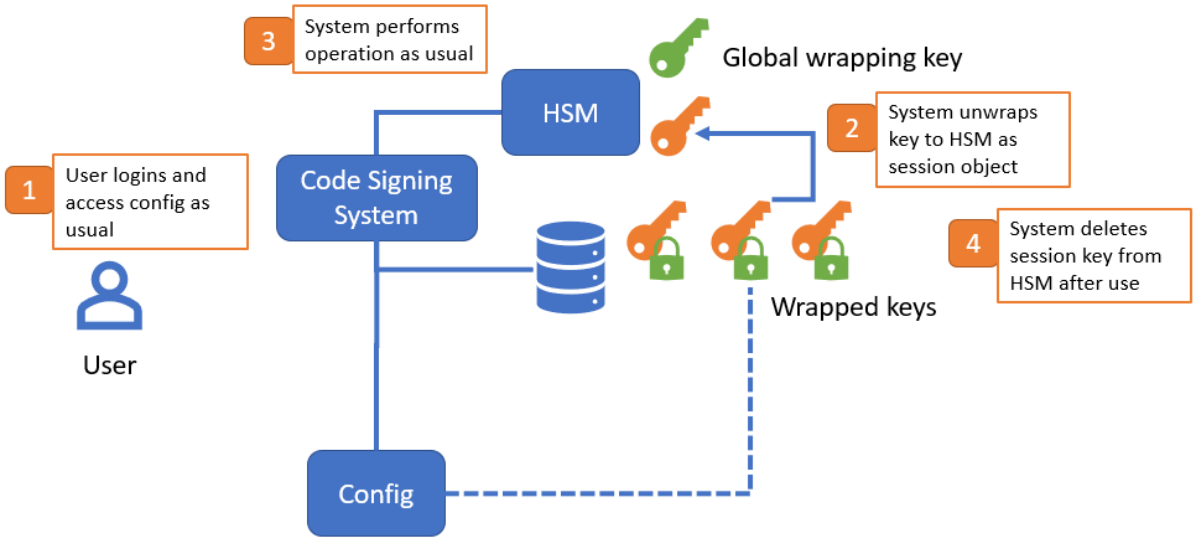

Figure 6 – Using wrapped keys to sign code

To allow for easy migration of keys from one system to another (e.g. from production to DR), the system should provide a feature to export one or more of the wrapped keys, along with any metadata (such as key type, key label, etc.) as a file. The wrapped keys are encrypted by the global wrapping key that is shared between the production and DR system. An additional encryption layer for transport security may be added on top. Each wrapped key file is then copied to the other code signing system, imported into its database, and made available from specified configurations. The exported key files can also be included in a secure backup of a code signing system.

2.6. Aberrant Use Detection and Management

Like any online system, a code signing system may also be subject to attack or abuse. If a hacker were somehow able to access the system (for example, by stealing an authorized user's credential), malicious code may be signed and treated as legitimate code.

The system can use various mechanisms to detect, prevent, and mitigate aberrant user activities. The user's transaction log as shown in Figure 2 is used to measure user behavior. Via the company/org/user management module in Figure 2, administrators can assign values to certain thresholds (such as maximum number of daily transactions) per user contract, or customer contract in case a customer company to the system has multiple users. The system can also rely on user activity characteristics such as user ID, transaction time, input file hash, operation type, configuration, and visiting IP address to determine a user's activity patterns based on their past activity captured in the transaction logs.

A background service can run periodically to compare the user's most recent activities with their respective activity patterns. If the differences are within the allowed range, the service will determine and

record the new activity pattern within the latest time period, i.e., within a sliding time window. Otherwise, the service will send an administrator an alert and perform pre-defined actions based on the activity pattern. This includes a temporary lock or an indefinite lock on the user account. For example, if a human user's working hours pattern shows 9am-5pm PST, a 2-3am PST login session and code signing transactions may indicate abnormal use and therefore an administrator would be notified.

During each user session, thresholds such as maximum number of daily transactions can be checked. When the current daily transaction number for a human user is approaching the threshold (e.g., 80% of the maximum), Captcha can be enabled to detect if the user is using scripts or not. When it reaches the daily maximum value, the user can be blocked from any further transactions for a specific configuration or for all configurations. This temporary lock is automatically removed the next day.

3. Conclusions

In this paper we discussed the need for code signing and the importance of securing a code signing system. We explored best practice for such a code signing system with multiple layers of physical and network security. We further described a possible fine-grained multi-level access and permission control for company and user accounts, signing key access and code signing configurations. We described additional features the system can offer, including flexible key management, and detecting and addressing abnormal signing activities. These guidelines and recommendations can be useful to anyone designing and building a secure code signing system.

4. Abbreviations and Definitions

4.1. Abbreviations

API	application programming interface
CA	certificate authority
CVC	code verification certificate
DLL	dynamic-link library
DR	disaster recovery
IoT	Internet of Things
GUI	graphical user interface
HSM	hardware security module
M2M	machine-to-machine
OS	operating system
OTP	one-time-programmable
QA	quality assurance
RSA	Rivest–Shamir–Adleman
SDLC	software development life cycle
SOC	system-on-chip
SSDF	secure software development framework
TLS	Transport Layer Security protocol
TPM	Trusted Platform Module

5. Bibliography and References

[1] Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities, https://csrc.nist.gov/publications/detail/sp/800-218/archive/2021-09-30, Sep 30, 2021.
[2] Highly Evasive Attacker Leverages SolarWinds Supply Chain to Compromise Multiple Global Victims With SUNBURST Backdoor, https://www.fireeye.com/blog/threat-research/2020/12/evasive-attacker-leverages-solarwinds-supply-chain-compromises-with-sunburst-backdoor.html.
[3] SUPERNOVA: A Novel .NET Webshell, https://unit42.paloaltonetworks.com/solarstorm-supernova/.
[4] Malicious Apps Bypass Security Tools to Steal Data, https://duo.com/blog/malicious-apps-bypass-security-tools-to-steal-data, Jan 20, 2016.
[5] Hacking Team's evil Android app had code to bypass Google Play screening, https://arstechnica.com/information-technology/2015/07/hackingteams-evil-android-app-had-code-to-bypass-google-play-screening/, July 16, 2015.
[6] Unsecured IoT: 8 Ways Hackers Exploit Firmware Vulnerabilities, https://www.darkreading.com/risk/unsecured-iot-8-ways-hackers-exploit-firmware-vulnerabilities/a/d-id/1335564, Aug 27, 2019.
[7] Hacking an Android TV in 2 minutes, https://medium.com/@drakkars/hacking-an-android-tv-in-2-minutes-7b6f29518ff3, Nov 17, 2019.
[8] Adobe Releases Security Bulletin About Code Signing Certificate, https://codeverge.com/grc.security/adobe-releases-security-bulletin-about-code-sign/1667553, Sep 28, 2012.
[9] Mongolian certificate authority hacked eight times, compromised with malware, https://therecord.media/mongolian-certificate-authority-hacked-eight-times-compromised-with-malware/, Jul 1, 2021.
[10] The real security issue behind the Comodo hack, https://therecord.media/mongolian-certificate-authority-hacked-eight-times-compromised-with-malware/, Apr 5, 2011.
[11] Final Report on DigiNotar Hack Shows Total Compromise of CA Servers, https://threatpost.com/final-report-diginotar-hack-shows-total-compromise-ca-servers-103112/77170/, Oct 31, 2012.
[12] 2FA App Loaded with Banking Trojan Infests 10K Victims via Google Play, https://threatpost.com/2fa-app-banking-trojan-google-play/178077/, Jan 27, 2022.
[13] BRATA Android Trojan Updated with 'Kill Switch' that Wipes Devices, https://threatpost.com/brata-android-trojan-kill-switch-wipes/177921/, Jan 25, 2022.
[14] Disney+ fans without answers after thousands hacked, https://www.bbc.com/news/technology-50461171, November 19, 2019.
[15] Nightmare week for security vendors: Now a Trend Micro bug is being exploited in the wild, https://therecord.media/nightmare-week-for-security-vendors-now-a-trend-micro-bug-is-being-exploited-in-the-wild/, April 22, 2021.
[16] Zoho ManageEngine Password Manager Zero-Day Gets a Fix, Amid Attacks, https://threatpost.com/zoho-password-manager-zero-day-attack/169303/, September 9, 2021.
[17] Dynamically Addressing the Gap of Software Application Protection without Hardware Security, Rafie Shamsaasef and Aaron Anderson, 2019.

Obtaining Low Latency in Upstream DOCSIS Transmissions

A Technical Paper prepared for SCTE by

Hongbiao Zhang, Architect Wireline Solutions, Casa Systems
100 Old River Rd
Andover, MA01810
Hongbiao.zhang@casa-systems.com
978 688 6706

Peter Wolff, VP Wireline Solutions Architecture, Casa Systems
100 Old River Rd
Andover, MA01810
Peter.wolff@casa-systems.com
978 688 6706

Vishnuvinod Sambasivan, Principal Engineer SQA, Casa Systems
100 Old River Rd
Andover, MA01810
Vishnuvinod.sambasivan@casa-systems.com
978 688 6706

Table of Contents

List of Figures

List of Tables

1. Introduction

Latency is gaining attention with cable operators, as it plays an important role in user experiences for many modern applications. In the Data-Over-Cable Service Interface Specifications (DOCSIS) community, the Low Latency DOCSIS (LLD) project was initiated in early 2018, with an attempt to significantly decrease the latency experienced by packets in downstream or upstream service flows, as they traverse DOCSIS links. It is expected that by adopting Low Latency DOCSIS, the latency observed in cable networks would approximate that of fiber with no infrastructure investment, so as to increase customer retention and target niche market such as gaming and mobile backhaul. The details of Low Latency DOCSIS can be found in the DOCSIS 3.1 MULPI Specification [1], where a collection of new features is introduced. These features include those specific to DOCSIS hardware, such as tightened upstream scheduling cycles, as well as those generic to all access networks, such as enhanced queue management. In the latter case, the study is originated from the Internet Engineering Task Force (IETF) community, with the practice fed back into the same.

A significant part of end-to-end latency in the DOCSIS system is attributed to the upstream path, especially the upstream media acquisition process. The lengthy media acquisition time is associated with the request-and-grant based mechanism that is used between the cable modem termination system (CMTS) and the cable modem (CM) before any upstream data can be transmitted by the latter. This mechanism exists because of how DOCSIS has evolved: As opposed to fiber, data transmission in cable networks started with low data rates in the upstream direction, with a great number of users sharing the same upstream channel, and with varying ranging distances associated with the users. These attributes suggest that upstream scheduling design would favor bandwidth utilization instead of latency, and the CM and CMTS must use an accurate byte count in the calculation. Subsequently, various enhancements were adopted, for example, DOCSIS 1.1 [2] allows an upstream scheduling type associated with a service flow to be specified, so that certain services such as voice and video can avoid or reduce the request-grant cycle. Also, DOCSIS 3.0 [3] allows concurrent requests to be transmitted, in a way that requests for additional bytes can be sent before the previous requests are granted.

In Low Latency DOCSIS [1], a new scheduling type is introduced, namely proactive grant service (PGS). It is used for best effort traffic in the upstream. In a nutshell, it allows a certain amount of traffic to take advantage of proactive grants without going through the request-grant cycle, thus reducing media acquisition latency. This improvement in latency comes at a cost of bandwidth utilization. With the advance of new technologies, cable operators can deploy increasingly higher bandwidth, along with smaller populations of users on an upstream channel. Meanwhile, there are also increasingly more stringent latency requirements for many new applications. So, it seems PGS is well-suited for future networks and their applications. This paper explores the effects of using PGS, as well as the tradeoff between latency and bandwidth utilization.

The rest of this paper is organized as follows: Section 2 summarizes various sub-features of LLD and their deployment dependencies, for both upstream and downstream. Section 3 discusses PGS and presents a series of test results using different parameters, followed by comparison and analysis of the test results. Finally, section 4 concludes the paper.

2. Low Latency DOCSIS: Features and Dependencies

In access networks such as cable networks, latency is primarily affected by media acquisition time (in the upstream) and congestion (in both upstream and downstream). To improve media acquisition time, [1] introduces two sub-features of LLD, namely MAP size and turnaround time reduction, and PGS.

[1] places a lot more emphasis on congestion avoidance and control, using mechanisms such as active queue management (AQM) and qual queue. Specifically, it introduces low latency low loss scalable throughput (L4S) forwarding by combining a queue-building service flow and a non-queue-building service flow, each with its own queue and separate instance of AQM algorithm. As an example, the AQM algorithms used for the queue-building service flow and the non-queue-building service flow could be active queue management-proportional integral controller enhanced (AQM-PIE) and immediate active queue management (IAQM), respectively. Weighted round robin (WRR) scheduling and coupled AQM are performed between the two. In addition, classification and queue protection are used to prevent queue-building traffic from interfering with non-queue-building traffic.

L4S is an end-to-end mechanism. Taking full advantage of L4S requires user applications to start adopting non-queue-building transport protocols such as Data Center Transmission Control Protocol (DC-TCP) and utilizing explicit congestion notification capable transport (with encoding 01), or ECT(1), for classification [4]. The table below lists the dependencies of all LLD sub-features.

Table 1 - LLD Dependencies

LLD Feature		CMTS Upgrade	CM Upgrade	User App Upgrade
Upstream PGS		YES	NO*	NO
MAP Size & Turnaround Time Reduction		YES	NO	NO
DS Dual Queue	WRR & Rate Shaping	YES	NO	NO
	AQM-PIE (Classic Queue)	YES	NO	NO
	IAQM (LL Queue)	YES	NO	YES**
	Coupling	YES	NO	YES**
	Queue Protection	YES	NO	NO
US Dual Queue	WRR & Rate Shaping	YES	NO	NO
	AQM-PIE (Classic Queue)	NO	YES	NO
	IAQM (LL Queue)	NO	YES	YES**
	Coupling	NO	YES	YES**
	Queue Protection	NO	YES	NO

* Some legacy CMs do not transport unknown TLVs and will not work with the feature

** User applications have to adopt non-queue-building transport protocols and utilize ECT(1) to take full advantage of L4S.

3. Proactive Grant Service: Experiments and Analysis

Proactive grant service represents an upstream scheduling type of a service flow that allows the CMTS to proactively allocate upstream transmission opportunities, i.e., grants, to the service flow without

transmission requests being received from this service flow. The amount and frequency of grants being proactively allocated are determined by the service flow parameters, including guaranteed grant rate (GGR), guaranteed grant interval (GGI), and guaranteed request interval (GRI). PGS can be used for any best-effort internet traffic. In a nutshell, PGS allows a certain amount of traffic to take advantage of proactive grants without going through the request-grant cycle, thus reducing the average media acquisition latency. When the proactive grants received by a CM are not enough to service the total bytes present in the CM's buffer, the CM could send additional requests on the same service flow. In other words, the CM could use a combination of proactive grants and reactive grants to satisfy the user's needs.

This reduction in latency comes at a cost of bandwidth utilization, i.e., as transmission opportunities are granted to a CM without request, they could be wasted when there are no bytes ready in the CM's buffer.

It is suggested that a CMTS could predict the instantaneous traffic rate of a service flow and adjust the proactive grant rate to match that [1]. We notice that this instantaneous prediction and adjustment is highly impractical. Instead, we propose that the CMTS designates a proactive grant rate based on estimation of the average traffic rate. As is demonstrated later, the average latency is insensitive to the instantaneous fluctuation of the traffic load if the incoming traffic rate does not exceed the maximum sustained traffic rate (MSR). We present experimental results and provide analysis in the subsequent sub-sections.

3.1. Experiment Setup

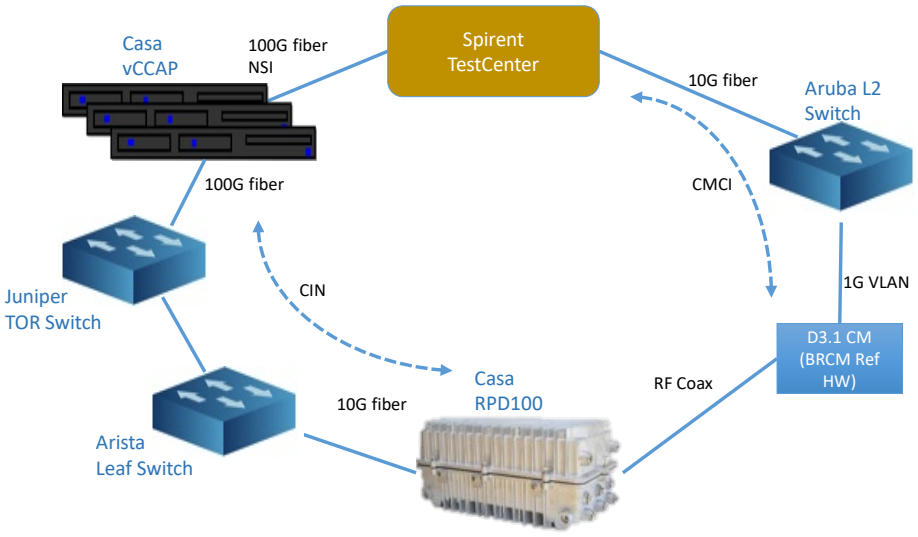

Figure 1 - Experiment Setup

Figure 1 depicts the experiment setup that we used for PGS testing. In the test cases below, we configure an orthogonal frequency division multiple access (OFDMA) channel with MAP size of 1 ms. One upstream service flow with PGS scheduling type is created, whose key parameters are as follows:

- MSR = 100 Mbps
- GGI = 1 ms
- GGR = 0%, 10%, 20%, 30% and 100% of MSR
- AQM is enabled, with latency target = 10 ms

In each test case, we measure the average delay and jitter, as well as how much granted bandwidth is wasted (in % of the total granted bandwidth) as associated with a series of traffic rate values. We repeat the experiments with different GGR levels (as percentage of the MSR) in separate test cases.

3.2. Recording of Results

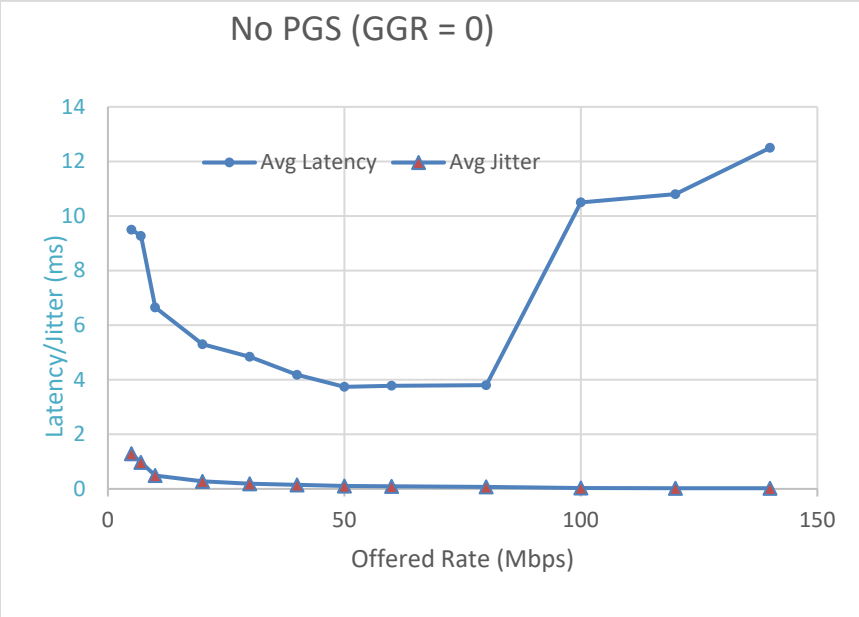

Figure 2 - Test Case 1: No PGS

Test Case 1: When GGR = 0, i.e., PGS is disabled, we obtain the average latency and the average jitter against the offered rate. Results of this test case are illustrated in Figure 2. As shown in the figure, latency starts at ~9.5 ms with a minimum offered traffic rate. When the offered traffic rate increases, latency goes down to <4 ms. The reason is that with a minimum offered traffic, the CM needs to send explicit requests in the contention region before it sends data, resulting in a relatively long latency. When the offered traffic increases, more and more piggyback requests can be used. When the offered traffic rate further increases to a level beyond the MSR, the CM's queueing buffer is quickly filled up, in which case the latency increases significantly. Since AQM is enabled, packets are selectively dropped when the AQM latency target is reached.

In this experiment, with 0 proactive grants allocated, there is no wasted bandwidth, therefore the wasted bandwidth plot is omitted.

This test case sets an upper bound of PGS latency for the subsequent tests, which is roughly 3.74 ms.

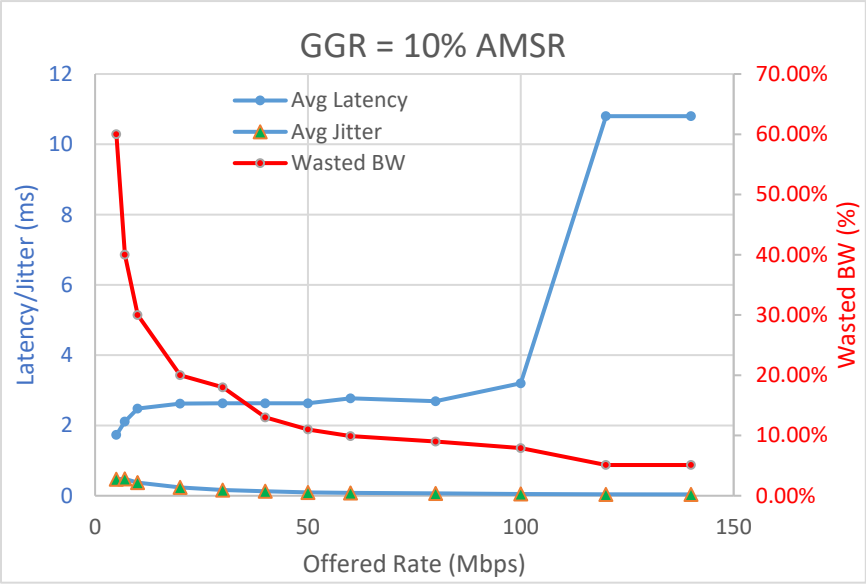

Figure 3 - Test Case 2: GGR = 10% MSR

Test Case 2: When GGR = 10 Mbps, or 10% of MSR, we obtain the average latency, the average jitter, and the wasted bandwidth against the offered rate. Results of this test case are illustrated in Figure 3. As we can see, when the offered rate is 10 Mbps (100% of GGR), the average latency is 2.48 ms (with an average jitter of 0.375 ms). When the offered rate increases to 80 Mbps (800% of GGR), the latency increases to 2.69 ms (with an average jitter of 0.066 ms). Accompanying the increase of latency, the wasted bandwidth decreases from 30% to 9%.

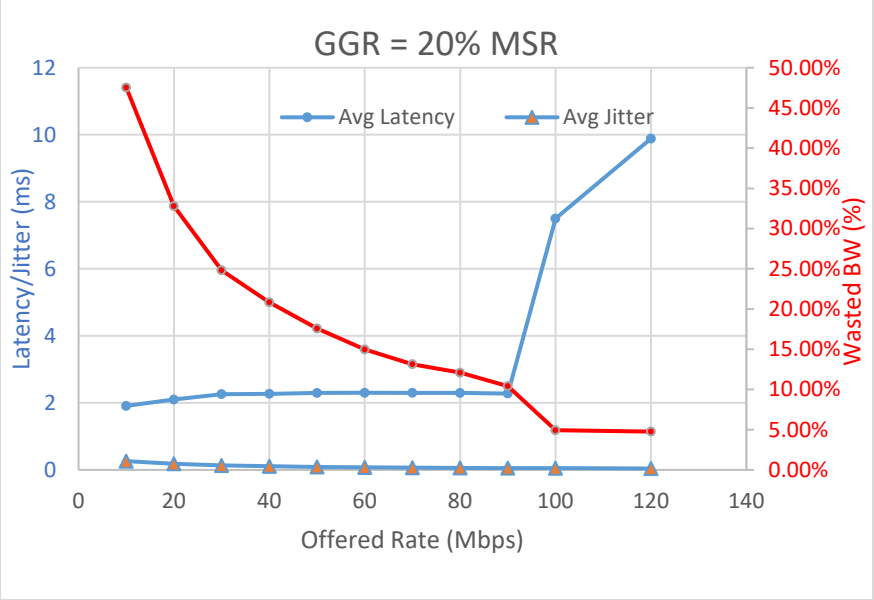

Figure 4 - Test Case 3: GGR = 20% MSR

Test Case 3: When GGR = 20 Mbps, or 20% of MSR, we obtain the average latency, the average jitter, and the wasted bandwidth against the offered rate. Results of this test case are illustrated in Figure 4. As can be seen, when the offered rate is 20 Mbps (100% of GGR), the average latency is 2.10 ms (with an average jitter of 0.183 ms). When the offered rate increases to 80 Mbps (400% of GGR), the latency increases to 2.295 ms (with an average jitter of 0.06 ms). Accompanying the increase of latency, the wasted bandwidth decreases from 32.79% to 12.09%.

Figure 5 - Test Case 4: GGR = 30% MSR

Test Case 4: When GGR = 30 Mbps, or 30% of MSR, we obtain the average latency, the average jitter, and the wasted bandwidth against the offered rate. Results of this test case are illustrated in Figure 5. As is shown, when the offered rate is 30 Mbps (100% of GGR), the average latency is 1.847 ms (with an average jitter of 0.123 ms). When the offered rate increases to 90 Mbps (300% of GGR), the latency increases to 2.069 ms (with an average jitter 0.052 ms). Accompanying the increase of latency, the wasted bandwidth decreases from 37.77% to 12.23%.

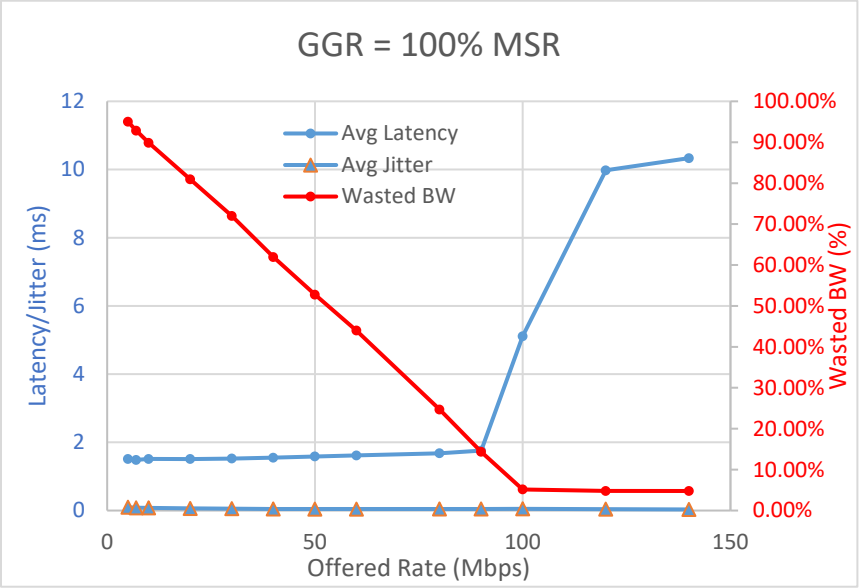

Figure 6 - Test Case 5: GGR = 100% MSR

Test case 5: When GGR = 100 Mbps, or 100% of MSR, we obtain and illustrate the average latency, the average jitter, and the wasted bandwidth against the offered rate. Results of this test case are illustrated in Figure 6. As we can see, latency is kept <2 ms if there's no congestion. Also, wasted bandwidth drops almost linearly with the offered traffic rate.

This test case sets a rough lower bound of PGS latency for the experiments above, which is ~1.5 ms. Increasing GGR further wouldn't lower this value much more, as shown in our experiments (omitted).

3.3. Comparison and Analysis

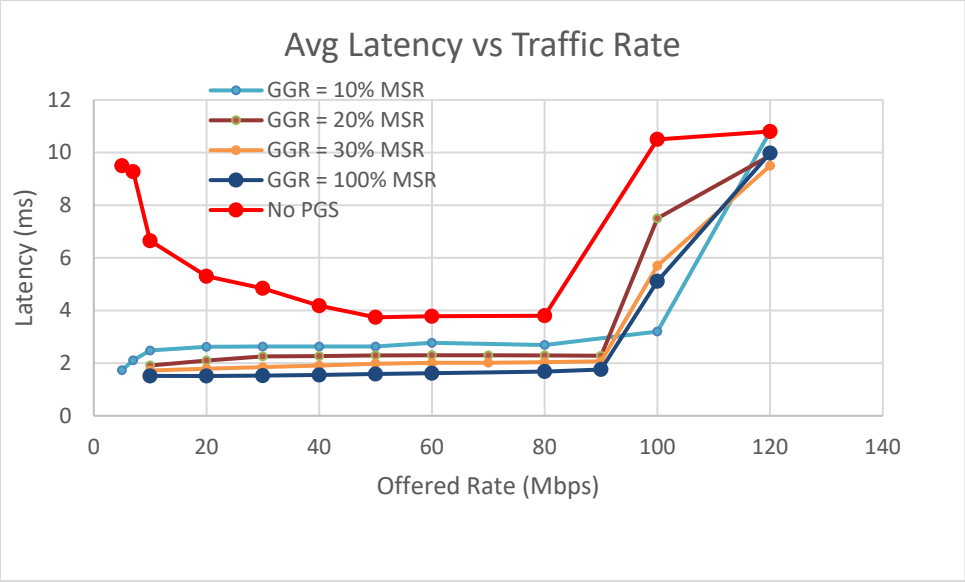

Figure 7 - Average Latency vs Traffic Rate

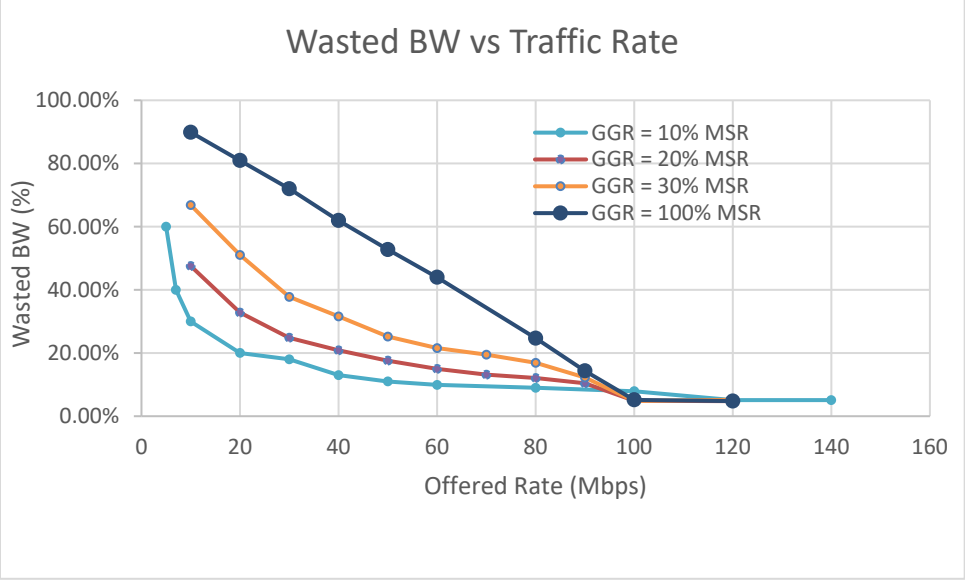

Figure 8 - Wasted Bandwidth vs Traffic Rate

Figure 7 and Figure 8 display the average latency and the wasted bandwidth respectively against the offered rate and compare the effects of various GGRs. As we can see from these figures:

- PGS significantly reduces the mean latency. This is especially true when the offered traffic load is minimal.
- When GGR increases, latency improves a little with the tradeoff in channel utilization.

- The latency differences between GGR = 10 Mbps (shown in the light blue line in Figure 7) and GGR = 100 Mbps (shown in the dark blue line in Figure 7) do not exceed 1 ms.
- When the offered rate increases, latency degrades slightly while channel utilization improves significantly.

These observations suggest that the introduction of PGS improves the performance of average latency, with an acceptable average jitter. Increasing GGR further results in additional improvement in latency; however, when we compare the four plots with non-zero GGRs in Figure 7, we can see that the differences among them is less than 1 ms. Therefore, we can select an appropriate GGR value that is equal to only a percentage of the MSR, e.g., 10% - 30%, yet sufficient to achieve a satisfactory average latency. The smaller the value, the better is the bandwidth utilization. It is possible to strike an ideal balance between low latency and bandwidth utilization. Besides, as we can see from tracing each plot in Figure 7, latency is insensitive to fluctuation of the offered traffic rate if the offered load is kept under the MSR limit. This means that the GGR value, once selected, doesn't need to change constantly to match the very dynamic traffic bursts.

4. Conclusions

Our experiments show that we can achieve low mean latency and jitter even when the instantaneous traffic rate goes several times over the proactive grant rate. In other words, in PGS operations, latency is insensitive to the fluctuation of the offered traffic rate, if the traffic rate doesn't exceed the MSR limit. Therefore, we can choose the proactive grant rate as only a percentage of the estimated traffic rate, which strikes an ideal balance between low latency and bandwidth utilization.

We've demonstrated that it is neither practical nor necessary to adjust the proactive grants to match the instantaneous traffic load. That said, it is certainly possible to estimate the long-term average traffic rate and make adjustment to the proactive grant rate accordingly, also in the long-term. For example, such estimation could be based on measurement conducted by the CMTS periodically, e.g., at per-second interval, or based on signaling information from an external entity.

On top of that, operators could utilize aggregate service flow (ASF) with dual queue/service flow (SF), with delay-sensitive (and non-queue-building) traffic classified into the LLD service flow, and with other traffic classified into the classic service flow. In addition, WRR scheduling between the two ensures that latency sensitive services get prompt treatment, while bandwidth is shared between the two service flows. In the upstream, PGS could be used for the LLD service flow. By allocating the proactive grant rate equal to only a percentage of the estimated low latency (LL) SF rate, we can achieve a desired latency for delay-sensitive traffic, as well as a desired bandwidth utilization for the aggregated service flow. Meanwhile, queue management mechanisms could be deployed independently. IAQM with dual-queue coupling could be enabled later when user applications adopt non-queue-building transport protocols.

With the combination of all LLD features, it is possible to achieve the performance close to if not better than that of fiber optics in upstream transmissions.

5. Abbreviations and Definitions

5.1. Abbreviations

ASF	aggregate service flow
AQM	active queue management

CM	cable modem
CMTS	cable modem termination system
DC-TCP	Data Center Transmission Control Protocol
ECT(1)	explicit congestion notification capable transport (with encoding 01)
DS	downstream
IAQM	immediate active queue management
GGI	guaranteed grant interval
GGR	guaranteed grant rate
GRI	guaranteed request interval
IETF	Internet Engineering Task Force
L4S	low latency low loss scalable throughput
LL	low latency
LLD	Low Latency DOCSIS
Mbps	megabits per second
ms	millisecond
MSR	maximum sustained traffic rate
OFDMA	orthogonal frequency division multiple access
PIE	proportional integral controller enhanced
PGS	proactive grant service
SF	service flow
TCP	Transmission Control Protocol
TLV	type-length-value
US	upstream
WRR	weighted round robin

6. Bibliography and References

1. *Data-Over-Cable Service Interface Specifications DOCSIS® 3.1, MAC and Upper layer Protocols Interface Specification*, CM-SP-MULPIv3.1-I21-201020
2. *Data-Over-Cable Service Interface Specifications DOCSIS 1.1, Radio Frequency Interface Specification,* CM-SP-RFIv1.1-C01-050907
3. *Data-Over-Cable Service Interface Specifications DOCSIS® 3.0, MAC and Upper Layer Protocols Interface Specification,* CM-SP-MULPIv30-171207
4. *K. De Schepper, B. Briscoe, Ed., G. White, DualQ Coupled AQMs for Low Latency, Low Loss and Scalable Throughput (L4S),* draft-ietf-tsvwg-aqm-dualq-coupled-21, Feb 2022

Maximizing Wi-Fi 6E: How to Embrace the 6 GHz Spectrum—and the Future—Now

A Technical Paper prepared for SCTE by

Bill McFarland, Chief Technology Officer, Plume Design, Inc.,
290 California Ave. #200
Palo Alto, CA 94306
bill@plume.com
1-844-69-75863

Table of Contents

List of Figures

List of Tables

1. Introduction

In the last few years, Wi-Fi has become an essential component of the global digital infrastructure. As the backbone of the digital economy, Wi-Fi connectivity is embedded into the way we live and work. But the technology is just getting started.

With Wi-Fi 6 adoption now in full force, the industry expects the technology's improvements to fuel new innovation, especially in areas such as the Internet of Things (IoT), augmented and virtual reality, and high-definition video streaming. As a result, the global economic value of Wi-Fi will increase from $3.3 trillion in 2021 to $4.9 trillion by 2025, the Wi-Fi Alliance estimates (1).

One of the drivers behind the growth in economic value is the 6 GHz spectrum allocation to unlicensed use, which will reduce congestion and significantly boost speed. Wi-Fi 6E extends the Wi-Fi 6 improvements such as increased capacity and throughput to the 6 GHz band—positioning this new generation of Wi-Fi to meet the surging demand for connectivity and more advanced use cases.

For communications service providers (CSPs), these developments bring an opportunity to grow portfolios with new offerings for the smart home. Developing services that capitalize on the full potential of Wi-Fi 6E require redefinition of business models. Traditional models that bind software with hardware cannot keep up with the rapid changes in smart home technology and digital lifestyles. As consumers look to upgrade to the new generation of Wi-Fi, they'll expect CSPs to deliver not only on the higher throughput of Wi-Fi 6E but also on improved quality of experience (QoE).

2. The Growing Momentum Behind 6 GHz

Wi-Fi adoption has exploded in the past decade. With this continuing expansion, preventing congestion has gained more urgency. The industry has been advocating for additional band allocation for some time—and in April 2020, United States regulators blazed the trail by opening up 1,200 MHz of spectrum in the 6 GHz band to unlicensed use.

Since then, many regulatory bodies have followed suit. Some, like those of South Korea, Canada, Chile, and Brazil, opened the full 1,200 MHz of spectrum. Others, including the European Union, have made available only the lower 500 MHz of that spectrum. In all, more than 40 countries have allocated 6 GHz to unlicensed use and many others are considering it (2).

In opening the 6 GHz spectrum, the U.S. Federal Communications Commission (FCC) said it envisioned “new innovative technologies and services that will deliver new devices and applications to American consumers (3).” Other countries expressed similar sentiments, with the United Kingdom’s communications regulator Ofcom stating that “further development of Wi-Fi services has the potential to deliver significant benefits for U.K. consumers and businesses (4).”

Even before regulators took action, chipset makers were announcing the development of Wi-Fi 6 chips for mobile devices in early 2020 (5). The rest of the industry reacted with much enthusiasm to Wi-Fi 6E, calling it “a game-changer” for consumers and enterprises, a “major step forward,” and “the best thing to happen to Wi-Fi since its inception (6).”

In January 2021, the Wi-Fi Alliance introduced interoperability certification for Wi-Fi 6E devices as part of its Wi-Fi CERTIFIED 6 program. As of mid-November 2021, the alliance has certified about 40 different devices, primarily access points (APs), phones, and routers (7).

Other notable certifications include:

· The first 8K TV (Samsung), banking on interest from gaming enthusiasts.

· Windows 11, enabling OEMs to deliver new Wi-Fi 6E-ready Windows PCs.

· Enterprise-grade APs that bring Wi-Fi 6E to the commercial sector.

The momentum behind Wi-Fi 6E opens the door for CSPs to develop and begin offering services that take advantage of the 6 GHz band availability. But there's a lot more to it than simply offering gigabit speeds. In the complex, increasingly congested home network, Wi-Fi 6E presents multiple decisions that need to be made and are bound to present problems—the solution will require advanced network management.

Table 1 below shows the spectrum that is available in the 6 GHz band. The large number of frequency channels, support for very wide channel bandwidths (160 MHz), and large overall quantity of spectrum all make use of the 6 GHz band appealing.

Table 1 – Wi-Fi 6E = Wi-Fi 6 + 6 GHz Spectrum

<table>
<tr><th></th><th colspan="2">US</th><th colspan="2">E.U.</th></tr>
<tr><th></th><th>5 GHz band</th><th>6 GHz band</th><th>5 GHz band</th><th>6 GHz band</th></tr>
<tr><td>Total bandwidth for Wi-Fi (MHz)</td><td>560</td><td>1200</td><td>380</td><td>500</td></tr>
<tr><td>Number of 160 MHz channels</td><td>3</td><td>7</td><td>2</td><td>3</td></tr>
<tr><td>AP Tx power no AFC (dBm EIRP)</td><td rowspan="4">30/36</td><td>27</td><td rowspan="4">23/30</td><td>23</td></tr>
<tr><td>AP Tx power with AFC (dBm EIRP)</td><td>36</td><td>NA</td></tr>
<tr><td>STA Tx power no AFC (dBm EIRP)</td><td>21</td><td>23</td></tr>
<tr><td>STA Tx power with AFC (dBm EIRP)</td><td>30</td><td>NA</td></tr>
<tr><td colspan="5">9 dB more Tx power in 5 GHz band = 2.8x distance in free space, ~2x in home</td></tr>
</table>

2.1. Industry Forecasts

- More than 338 million Wi-Fi 6E devices were expected to enter the market in 2021 (8).
- It is estimated that close to 20% of all Wi-Fi 6 device shipments will support 6 GHz by 2022 (8).
- By 2025, 41% of the 5.2 billion Wi-Fi 6 products forecasted to ship will be Wi-Fi 6E devices (8).

3. Key Wi-Fi 6E Benefits for Customers

Wi-Fi 6E inherits the benefits of Wi-Fi 6. Wi-Fi 6 has tremendous potential to transform the home network. Viewed as a new era of Wi-Fi connectivity, Wi-Fi 6 brought many transformative improvements to the 20-year-old technology.

Designed to better handle different types of traffic simultaneously from different users, Wi-Fi 6 has introduced key changes such as:

- More than 2x higher throughput in environments with low congestion.
- Improved power savings capabilities.
- Greatly improved efficiency when supporting large numbers of devices requiring modest throughput (e.g., IoT devices).
- Improved quality of service (QoS) through time and frequency reservations.

The new features and capabilities are welcome improvements for connected environments, including the smart home—which now handles not only more connected devices but also more bandwidth-hungry applications.

Wi-Fi 6E extends the low latency, high capacity, and gigabit speeds to the 6 GHz spectrum, adding up to seven super wide 160 MHz channels and 14 channels that are 80 MHz wide. In addition to greater capacity and performance, Wi-Fi 6E also brings less interference due to the large number of channels available, few deployed devices, and elimination of legacy modes of operation.

3.1. Top Use Cases

As the Wireless Broadband Alliance (WBA) notes, Wi-Fi 6 and Wi-Fi 6E together provide "more capacity than all the other Wi-Fi bands put together and deliver connections with speeds equivalent to the new advanced 5G mobile (9)." This will unleash a flurry of innovation for products and services that rely on bandwidth-intense connectivity.

The primary use cases that Wi-Fi 6E will support include:

- Broader adoption of IoT, both by consumers and enterprises.
- Multigigabit video streaming, such as 4K and 8K video.
- New generations of augmented reality and virtual reality technology.

While a good deal of innovation spurred by Wi-Fi 6E will take place in the commercial and industrial sectors, consumer adoption won't be far behind. Some industry players forecast that connected home applications—including smart appliances, security and video surveillance, and home automation—will represent nearly half of all the machine-to-machine connections (used by IoT devices) by 2023 (10).

In addition to preparing for the rapid rise of connected home devices, CSPs need to be ready for the smart home evolution. One such evolution is the self-optimizing, adaptive home of the future—which learns from and adapts to consumers' lifestyle patterns and behaviors.

4. The Opportunities for CSPs

Wi-Fi 6 and Wi-Fi 6E are expected to bring drastic changes for consumers and enterprises alike. These developments are exciting, but they will challenge CSPs' ability to ensure service delivery supports the new use cases.

Consider, as one example, extended reality, which includes augmented reality, virtual reality, and mixed reality. One estimate showed the market size was $18.6 billion in 2019 and will grow at a compound annual growth rate of 48.3% in the next decade (11).

While the COVID-19 pandemic boosted the market due to more virtual meetings and training conducted from home, it's the gaming industry that's expected to grow the fastest between 2020 and 2030.

Already, an estimated 58.9 million US consumers use VR, and 93.3 million use AR at least once a month, representing about 18% and 28% of the population, respectively (12).

Although the lack of in-person experiences during the pandemic boosted some of those numbers, new trends, such as the metaverse, will continue to push the market toward innovation and consumers toward adoption. CSPs can ride the popularity of these kinds of trends by offering upgrades and new services that improve the QoE even as consumers add these bandwidth-hungry applications to their already congested home network. By leveraging the power of cloud computing and artificial intelligence—and uncoupling service delivery from hardware—CSPs can optimize the home network connectivity and enable their customers to benefit fully from the speed and capacity of Wi-Fi 6E.

5. Wi-Fi 6E Implementation Requirements

Wi-Fi 6E shares the 6 GHz spectrum with point-to-point microwave links. The US has about 100,000 microwave links that are used by public safety agencies, mobile carriers, and commercial entities (13).

To avoid interference with nearby microwave systems, Wi-Fi 6E devices must either operate at a low power level or implement an automatic frequency control (AFC) system.

5.1. AFC

By avoiding a frequency channel used by nearby microwave links, the AP and its clients can operate at a high power level, enjoying full range and data rates. The AFC system is complicated, requiring cloud interaction and management to communicate with a smart controller that could look up the FCC database, factor in geodata, calculate interference, and deliver instructions back to the AP. This is illustrated in the block diagram presented in Figure 1.

However, the FCC isn't expected to approve any AFC systems before the third quarter of 2022. That means Wi-Fi 6E devices will have to operate on low-power transmission for at least the first year.

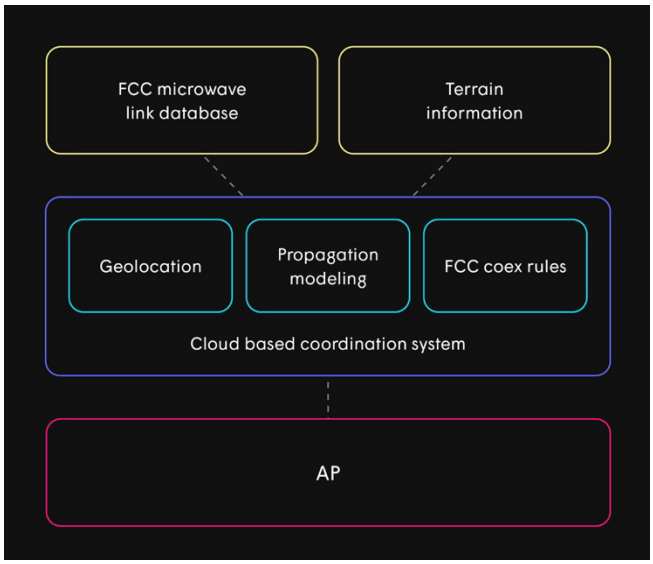

Figure 1 – The AFC system

5.2. Low Power Mode

Operating at approximately 1/10th the power levels allowed with AFC (10 dB less transmitted power), this mode produces a weak signal, which can't transmit as far or at as high a data rate for a given distance compared to 5GHz Wi-Fi. Compensating for the significantly shorter range requires an optimization system that can figure out multi-AP configurations with many parameters that are optimal for the specific environment.

For either AFC or low-power mode, the control system needs to consider the client types, loads, and capabilities before allocating the network's radio resources. The 6 GHz band can be a great benefit, but if used incorrectly, it can degrade performance. An effective network needs to know where, when, and for which links the 6 GHz band should be used.

6. Managing Wi-Fi 6E

As noted earlier, Wi-Fi 6E extends Wi-Fi 6 capabilities to the 6 GHz spectrum. These include, among others:

- 160 MHz channel bandwidth – The wider the channel used, the higher the data rate, and thus the data speed. Technically this is not a new feature, but few Wi-Fi 5 devices supported channel bandwidth greater than 80 MHz.
- OFDMA – A marquee feature, uplink and downlink orthogonal frequency division multiple access, or OFDMA, improves efficiency and capacity by subdividing the channel into smaller frequency allocations (resource units) transmitted from one AP in parallel.
- Resource unit reservations – The smaller frequency slices with which OFDMA operates can get allocated to particular clients, guaranteeing QoS to those clients.
- Target wake time (TWT) – By scheduling and reserving specific times for each client to be awake, this feature improves battery life for devices that are transmitting only occasionally or at a low-duty cycle.

- BSS color – Basic service set (BSS) color allows more efficient airtime usage between overlapping networks that are on the same frequency channel. It does this by assigning a unique identifier ("color") to each network. This identifier is transmitted in the beginning of each packet header. Devices that begin to receive a packet are therefore quickly able to determine if the packet is from a device within their own network, or from a neighboring network. If from a neighboring network, and at a relatively low signal strength, the rules allow the AP observing the packet with the BSS color indication to transmit on top of that packet. The reasoning is that the transmissions in both networks should succeed in parallel given they are separate networks and interfering at low signal levels. This prevents the network in one home from having to constantly defer, waiting for transmissions in the neighboring home to end.

While these features provide various advancements, they also have drawbacks and limitations. To achieve full potential, avoid interference, operate efficiently, and maintain QoS, Wi-Fi 6 APs require more rather than less optimization than previous Wi-Fi generations, and performance will depend on the system controlling it. Wi-Fi 6E, in turn, needs even more sophisticated management than Wi-Fi 6, particularly in two categories:

- Topology – The selection of frequency channels, channel bandwidths, and interconnections between APs.
- Steering – The selection of the AP to which each Wi-Fi client should connect, and the frequency band for that connection.

6.1. Topology Management

Without proper management, the 6 GHz band can quickly turn into a burden rather than a benefit. To compensate for the lower signal strength, many homes will require 6E extenders for extra coverage, and there will be a large number of potential configurations. With so many potential configurations, a lot can go wrong—and configuring each environment for optimal performance is not a simple task.

6.1.1. The Wrong Way to Configure an Environment for Optimal Performance

6.1.1.1. Not Allowing for Extenders

Without AFC, the 6 GHz band will need to use a low-power mode, which has a significantly shorter range than the 5 GHz band. Figure 2 shows an example of the problem, in which the TV, which is distant from the single AP in the home, cannot connect with sufficient signal strength due to the low power limitations of the 6 GHz band. The optimum solution is to bring 6E extenders to the homes that need them. To achieve that, a system must be able to:

- Identify homes with coverage problems.
- Intelligently manage homes where extenders are added.

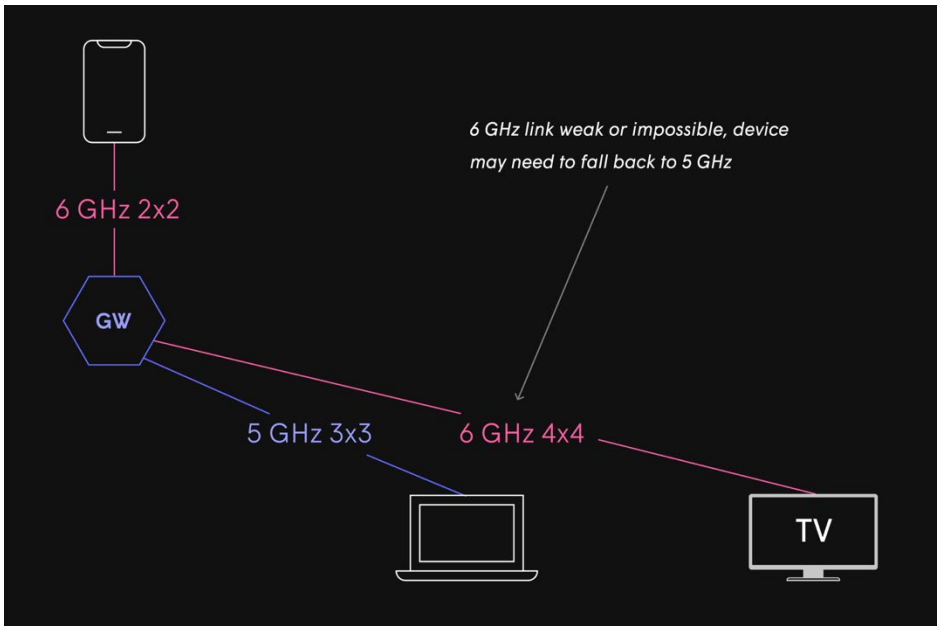

Figure 2 – Need for extenders in the 6 GHz band

6.1.1.2. Locking the Backhaul to 6 GHz

It seems logical to use the 6 GHz band for the backhaul between APs in the home. But, simply placing all backhaul links on 6 GHz is not so wise. Why?

- Even more extenders in homes might need to be deployed for 6 GHz than for 5 GHz Wi-Fi.
- As data travels through the multiple hops in a home, all on the same 6 GHz channel, the self-interference between hops divides the throughput down by the number of hops.

Figure 3 shows an example of using only the 6 GHz band throughout the backhaul, and the self-interference problem that it creates. A given AP has only a single 6 GHz radio that can operate on only one 6 GHz channel. When all backhaul hops are in the 6 GHz band, the same channel must be used in the backhaul everywhere.

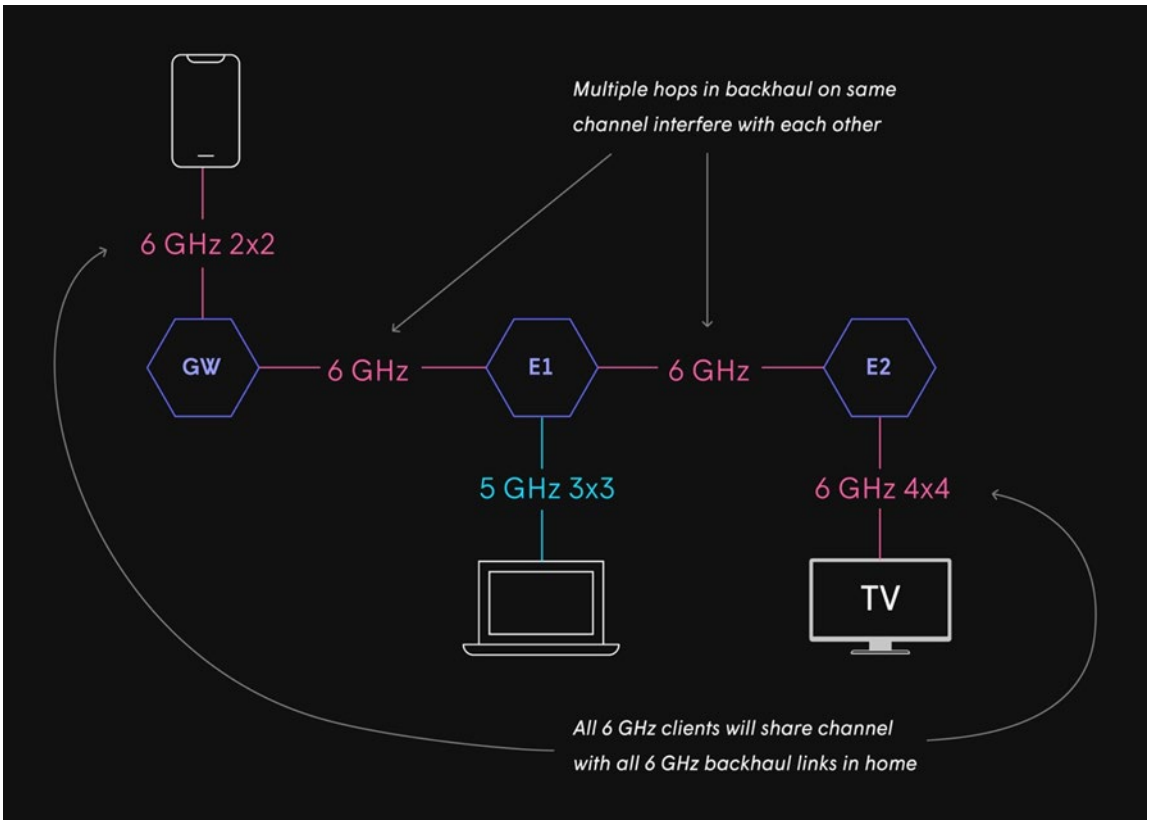

Figure 3 – Self interference from all-6 GHz backhaul

6.1.1.3. Locking The Backhaul To 5 GHz

Fixing all the hops in the 5 GHz band incurs some of the same problems. In particular, as there are more hops, self-interference again reduces throughput. Figure 4 demonstrates the self-interference that arises both on the backhaul links and on the links to 5 GHz clients. There are two additional downsides to this approach:

- Having the backhaul loaded only on a 5 GHz channel makes the system more vulnerable to interference from neighbors.
- The wide 5 GHz channels all require radar detection and radar events that are more common than imagined. Along with real radars, a variety of interference scenarios, including from overlapping Wi-Fi networks, can trigger what looks like radar, sending the network scrambling to get out of the way and disrupting service.

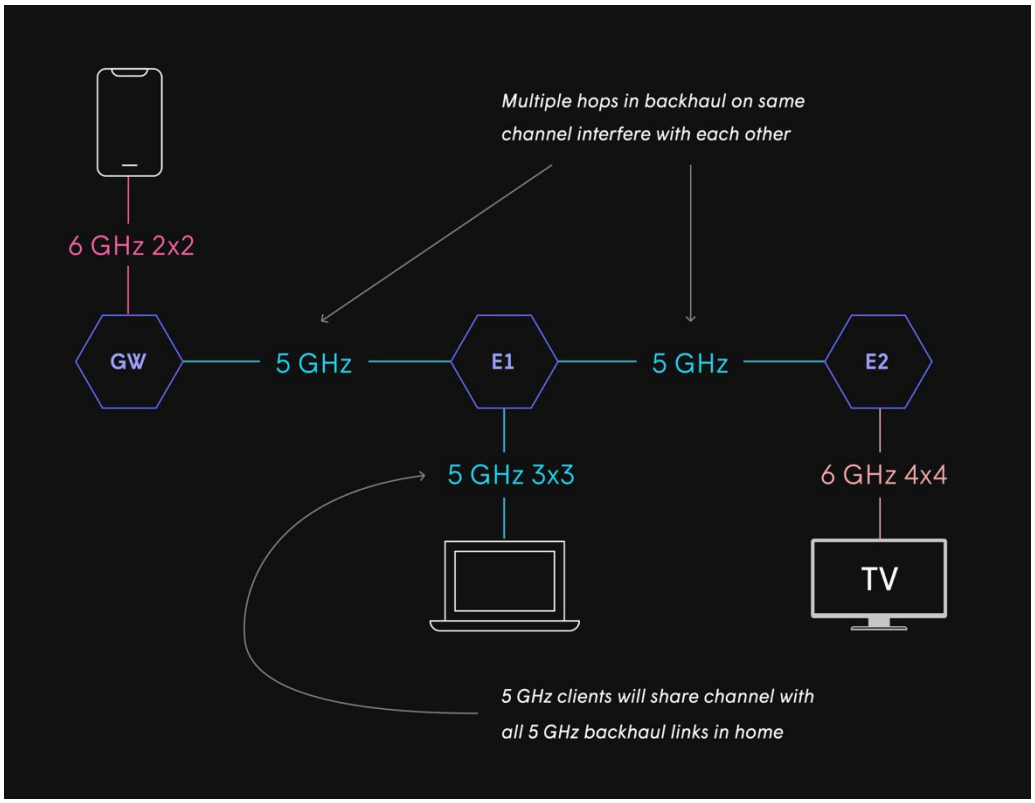

Figure 4 – Two additional downsides to locking the backhaul to 5 GHz

6.1.1.4. Barring Clients On Backhaul Channels

Many of today's multi-AP systems strictly segregate the connections between APs (backhaul) and connections to customer devices (fronthaul). This means that:

- If 6 GHz is being used for the connection between two APs, clients will not be able to connect to either of those APs in the 6 GHz band.
- The same problem occurs when using 5 GHz channels to connect APs in a segregated system.

Figure 5 shows graphically the difficulty that client devices experience when they are not allowed to connect on the same channel/radio that is being used for a backhaul connection. In this example, the cell phone on the left side of the drawing has two poor choices: connect to the nearby AP using the more congested 5 GHz band, or connect to an AP at the far end of the house using the transmit power-limited 6 GHz band. A better solution would be to allow the cell phone to connect to the nearby AP on the 6 GHz band, while the backhaul connection to that AP also operates on that band.

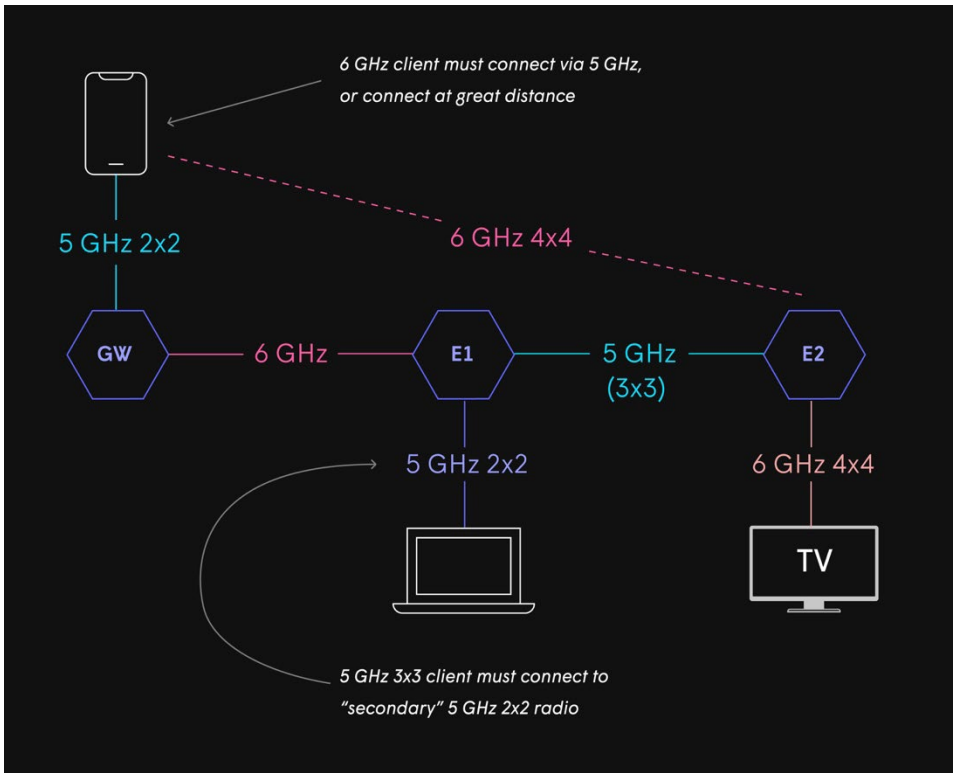

Figure 5 – Why barring clients on backhaul channels won't work

6.1.2. The Right Way to Configure an Environment for Optimal Performance

So, what's the solution? Using a mixture of 5 GHz and 6 GHz band links for the backhaul, and allowing clients to connect on the same channel as the backhaul when appropriate. This results in:

- Reduced self-interference.
- Greater flexibility in avoiding interference and radar events.
- Clients attaching to any of the APs using their best band of operation.

Figure 6 is based on the same home shown previously, but applies the optimal configuration. The best way to achieve this optimal configuration is with sophisticated cloud controls. A cloud-based system can quickly search the huge space of potential configurations and choose the one that's best for the environment.

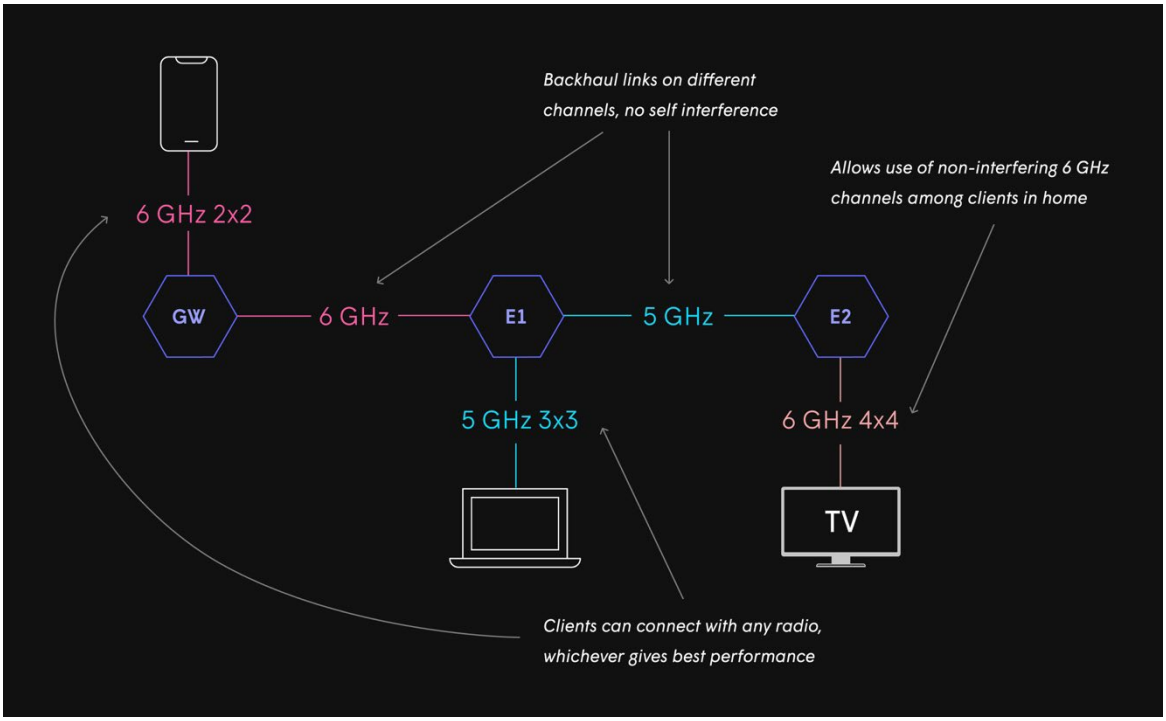

Figure 6 – A flexible and intelligent solution

6.2. Client Steering

The purpose of steering is to connect client devices on the band that will maximize their performance, regardless of whether the home has one or multiple APs. Although many clients can make the connection decision autonomously, historically the selection algorithms have not worked well. Steering clients from the infrastructure side is generally complicated. But prior to Wi-Fi 6E, at least the decision of which band the client should connect on was relatively simple. Typically, 5 GHz has a huge advantage in throughput and interference over 2.4 GHz. With 6 GHz, this decision gets trickier because there are more choices, and the race between 5 GHz and 6 GHz is often a close one. For example, 6 GHz might have less interference, but because of the lower allowed transmit power it might achieve a lower data rate.

The decision about the best option for the highest throughout must take into consideration a variety of factors, including:

- Supported Tx power levels in the two bands on both the client and AP.
- Traffic load in each of the bands.
- Interference levels in the bands.
- The multiple input/multiple output (MIMO) configuration of both the client and AP in each band.
- The condition and channel of all of the hops that may be required to traverse from the AP to the Internet, as subsequent backhaul hops may interfere with the hop from the client to the AP.

Figure 7 shows two important aspects of the optimal topology. First, unnecessary repeating (“hops”) should be avoided if the distances are short. Second, client devices often should *not* connect to the nearest AP, as that may have poorer net performance than connecting to a more distant AP.

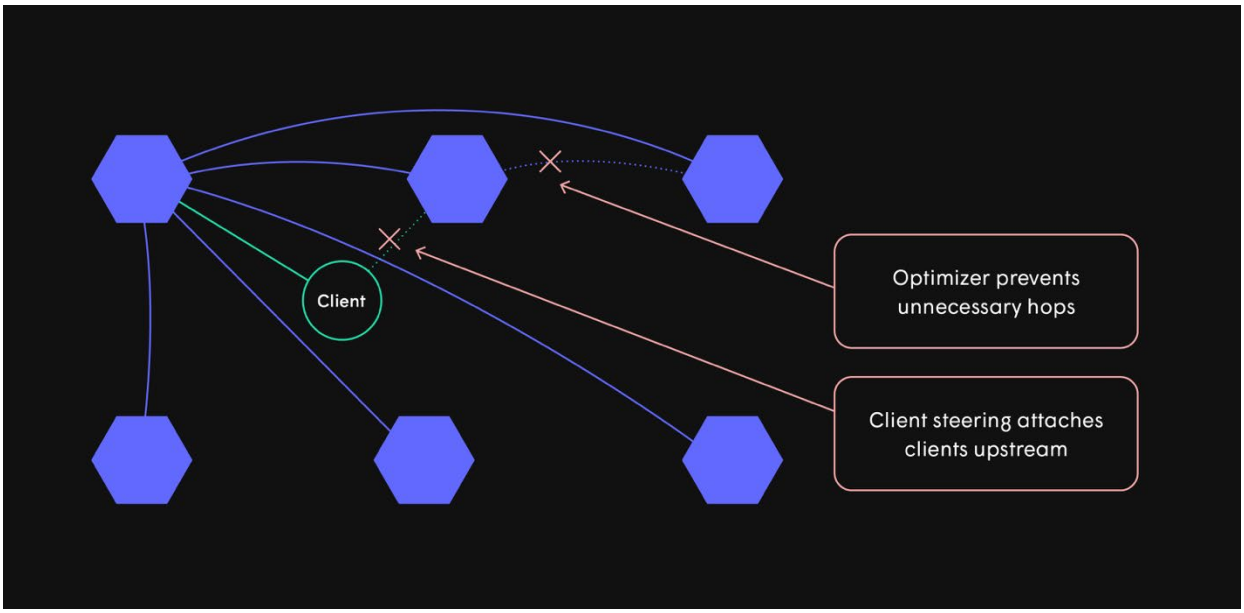

Figure 7 – Client steering overview

Even more complications arise in a home that has a mix of Wi-Fi 6E and legacy devices because Wi-Fi 6E requires stronger WPA3 security whereas previous Wi-Fi devices may have supported only WPA2. When devices roam or are steered among systems with inconsistent security (e.g., from WPA3 to WPA2), customers typically experience significant disruption in service. This would especially be the case when streaming video (whether they're watching content or are in the middle of a business video conference).

Solving this generational problem requires proper topology configuration and appropriate steering decisions. And, as with topology management and band steering, optimal usage of the capabilities of different generation devices in a mixed network can be achieved with an intelligent platform that analyzes the different capabilities and applies dynamic controls and rigorous optimization to the home network.

7. Future-Proofing Now With Intelligent Management

While topology management and band steering are critical to the initial launch of Wi-Fi 6E to the subscribers' homes, CSPs should plan now for augmenting with additional management capabilities in the future.

By planning for additional capabilities now – including creating a central management system – CSPs can make the right choices that will futureproof their service delivery as Wi-Fi 6 becomes more sophisticated. When the FCC eventually approves AFC systems, CSPs will want to obtain the higher transmit power levels. As noted earlier, this will inherently involve a cloud-control system to communicate with and configure the APs.

One mistake is to think of Wi-Fi 6E as something that comes after Wi-Fi 6. In reality, it's a checkpoint along the path of Wi-Fi 6. Features such as OFDMA, BSS color, TWT, and time/frequency slot reservations for high-priority traffic are in the early days of implementation—and all of them need centralized management. Wi-Fi 6E networks will benefit from the management of these same features.

There's some discussion in the industry about whether it makes more sense to postpone any upgrades until the expected arrival of Wi-Fi 7 in 2024. While it may be tempting to not commit to any changes until then, riding out the developments for the next few years would be a mistake that could leave CSPs behind. If you do so, you'll be left behind and will need to catch up later.

Additionally, implementing a centralized system now will make 2.4 GHz and 5 GHz bands work better—revitalizing your existing infrastructure. Wi-Fi 6E adoption will not result in the immediate replacement of legacy devices. Not only do those devices still have a long life in the field, but the silicon shortage will delay your ability to roll out upgrades to your entire customer base. In the meantime, a software upgrade that can achieve a significant portion of the benefits of Wi-Fi 6E could be very attractive.

Figure 8 shows an example of OFDMA aware steering. A home with four APs and eight OFDMA capable clients is shown. If the clients are distributed throughout the home, they might naturally connect two to each AP. However, with only two devices connected to a given AP, the gain from transmitting to multiple clients in the same transmission with OFDMA is modest. Another approach would be to attach all eight clients to the first AP, but this might cause some clients to connect over a great distance, degrading performance for the entire network. The grouping shown might be optimal in this case, balancing the need for short transmission paths with the ability to aggregate OFDMA clients onto a single AP.

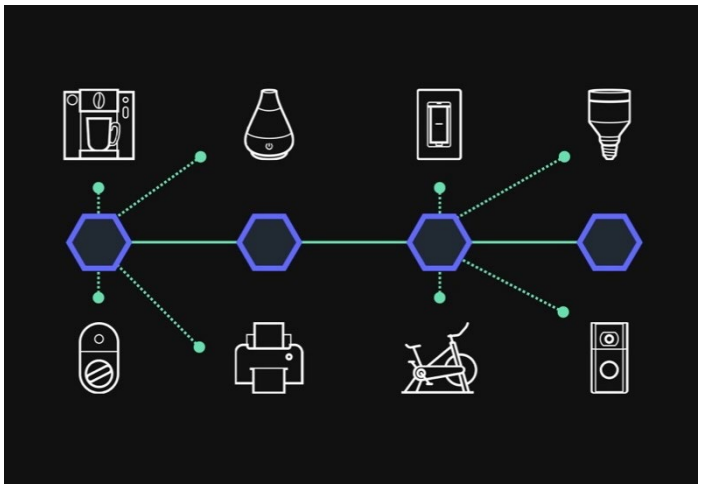

Figure 8 – OFDMA Aware Steering

Figure 9 shows a conceptual diagram of BSS coloring, described earlier in the paper. Cells with different colors (identifiers in the preambles of their packets) can transmit at the same time successfully if the signal strength between them is sufficiently low, improving efficiency.

Figure 9 – BSS Color

Figure 10 shows the organization of high and low priority traffic across time and frequency on two separate APs in the same network. Wi-Fi 6, with or without the 6 GHz frequency bands, allows the potential for reserving resource units (time/frequency slices) for high priority traffic. However, these allocations need to be carefully scheduled across APs that may be operating on the same frequency channels. Here, times and frequencies reserved for high priority traffic are non-overlapping. Low priority traffic will naturally defer to the higher priority traffic when low and high priority traffic is allocated in the same time/frequency slots.

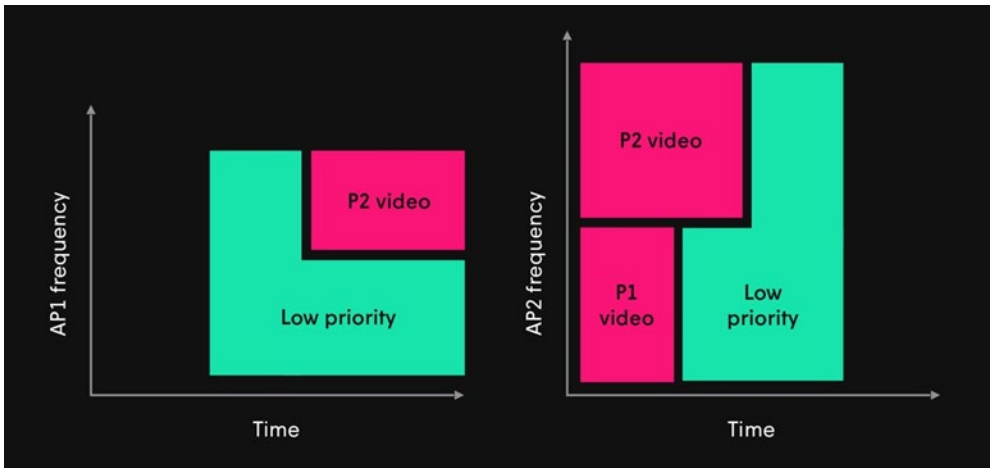

Figure 10 – Time Alignment

Two other benefits of revitalizing the infrastructure now are radar event avoidance and interference mitigation in dense environments such as multiple dwelling units (MDUs). While MDU optimization and radar event avoidance can't replace all the benefits of the 6 GHz spectrum, they can substantially improve the performance of both legacy networks and 6E networks in the 2.4 GHz and 5 GHz bands.

Table 2 illustrates the need for radar event avoidance. It lists the observed rate of radar events across homes and APs by frequency channel, expressed as events per hour, observed over one (1) month. Radar events are unevenly distributed and can be avoided with sufficient intelligence.

Table 2 – Long Term Radar Rate per DFS Channel per AP (Aggregate Learning From July 2021)

Location ID	(AP) Node ID	52	56	60	64	100	104	108	112
58bbb63798c6eff642b1d1a2	EM7F60018C	0.00	0.00	0.00	0.00	0.01	0.01	0.01	0.01
5fbd3b8b66b7453aa7b48808	EM7F300045		0.10	0.10		0.00	0.00	0.01	0.00
5f933d571f6ecf4457c695f0	EM7F300049	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EM7F600065	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	EM7F600083	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00
	EM7F600097	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5f9f8d1a4a576e734d231590	EM83A00143				0.49			0.04	0.04
57f3193f3ce0d96ed09d0217	EM7F60014C	0.00	0.00	0.00		0.06	0.06	0.06	0.21
5f920576bfddfc59ee6dd846	EM7F60007B			0.00	0.00			0.00	
	EM7F600024	0.01		0.00	0.00	0.00		0.00	0.00
	EM7F600054					0.00	0.00	0.00	0.00
	EM7F600064	0.00	0.00	0.00	0.00	0.03	0.03	0.03	0.10
57f489f83ce0d96ed09d6570	EM7F6000C1		0.00	0.00	0.00	0.00	0.00	0.00	
	EM7F300006	0.00	0.00	0.00	0.00				
	EM7F300040	0.03	0.03	0.05	0.00		0.00	0.00	
	EM7F600119		0.00	0.00		0.00			
60480b07328add3166edda7d	EM7F30001D	0.00	0.01	0.01				0.00	0.00
	EM7F30004D		0.00	0.00	0.00	0.00	0.00	0.00	0.00

Location ID	(AP) Node ID	52	56	60	64	100	104	108	112
	EM7F300001	0.02	0.07	0.07	0.06				
60391ea536d7902c9a44a93c	EM7F6000C6	0.01	0.01	0.01	0.01	0.00	0.00	0.02	0.02
	EM7F6000O8	0.00	0.02	0.03	0.01	0.02	0.04	0.03	0.04
	EM7F60012B	0.00	0.01	0.01	0.00	0.00	0.01	0.01	0.00

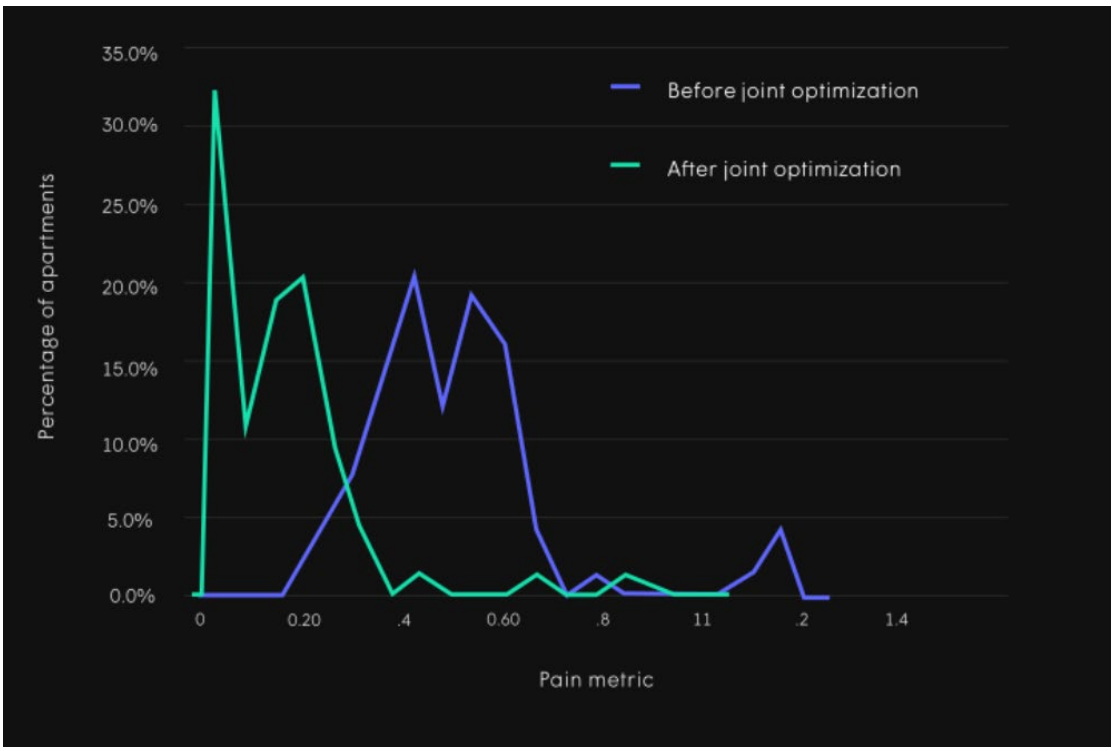

Figure 11 – Percentage of apartments Vs. pain metric

Figure 11 shows a sample result from optimizing frequency channels across an entire apartment complex. The pain metric factors both the level of interference and traffic load in a particular apartment. Pain indexes above 0.5 result in stuttering in video streaming and teleconferencing. The improvement that comes from joint optimization across an entire apartment complex can be seen in the shift of the histogram from higher pain levels to lower pain levels.

8. Intelligent Network Management

An intelligent management platform, powered by the open-source, silicon-to-cloud OpenSync framework creates an agnostic approach that allows multiple devices from different device makers to coexist as part of a home ecosystem.

This platform provides Wi-Fi operational enhancements that optimize the home network's connectivity to achieve better speed and capacity. The cloud-based, AI-driven algorithms learn from data collected across millions of networks and clients to identify the best steering techniques, predict interference, and perform complicated analyses to apply dynamic controls and rigorous optimization to networks with multiple APs or in MDUs. Using AI, the system can determine when to apply 160 MHz bandwidth channels where they are most needed and avoid them where they are not. The system factors interference, loads, and device types. When considering what channel bandwidth to assign to each AP, the platform knows the history and current set of clients and loads that are present in the network at each access point. OpenSync-enabled APs, coupled with the cloud, will address the demand for the additional intelligence necessary to get the most out of the latest generation Wi-Fi specifications and client devices.

9. Conclusion: Getting Ahead of the Curve

Although Wi-Fi 6E adoption is in its early stages, CSPs should act now. In today's extremely competitive market where connectivity speed is no longer a differentiator, customers want a lot more from their providers. As emerging technologies such as AR/VR and 8K video streaming take off, consumers will expect to benefit from the full potential of Wi-Fi 6E. Providers that plan smartly and get ahead of the curve will emerge as the market leaders.

10. Abbreviations

AFC	automatic frequency control
AI	artificial intelligence
AR	augmented reality
AP	access point
BSS	basic service set
CSPs	communications service providers
dB	decibel
dBm	decibel milliwatt
DFS	dynamic frequency selection
EIRP	effective (or equivalent) isotropic radiated power
FCC	Federal Communications Commission
GHz	gigahertz
ID	identification

IoT	Internet of Things
MDU	multiple dwelling unit
MHz	megahertz
MIMO	multiple input multiple output
OEM	original equipment manufacturer
OFDMA	orthogonal frequency division multiple access
PC	personal computer
QoE	quality of experience
QoS	quality of service
SCTE	Society of Cable Telecommunications Engineers
STA	station
TV	television
TWT	target wake time
Tx power	transmitted (or transmit) power
VR	virtual reality
WBA	Wireless Broadband Alliance
WPA	Wi-Fi Protected Access

11. Bibliography and References

1. Global Economic Value of Wi-Fi, 2021-2025; Wi-Fi Alliance
https://www.wi-fi.org/download.php?file=/sites/default/files/private/Global_Economic_Value_of_Wi-Fi_2021-2025_202109.pdf

2. *Quarterly update: Wi-Fi 6E devices driving technology innovation*; Wi-Fi Alliance
https://www.wi-fi.org/beacon/the-beacon/quarterly-update-july-2021-wi-fi-6e-devices-driving-technology-innovation

3. *FCC Opens 6 GHz Band to Wi-Fi and Other Unlicensed Uses*; US Federal Communications Commission
https://www.fcc.gov/document/fcc-opens-6-ghz-band-wi-fi-and-other-unlicensed-uses-0

4. Improving spectrum access for Wi-Fi; UK Ofcom
https://www.ofcom.org.uk/__data/assets/pdf_file/0036/198927/6ghz-statement.pdf

5. *Broadcom Announces World's First Wi-Fi 6E Chip for Mobile Devices*; Broadcom
https://investors.broadcom.com/news-releases/news-release-details/broadcom-announces-worlds-first-wi-fi-6e-chip-mobile-devices

6. *Wi-Fi Alliance® delivers Wi-Fi 6E certification program*; Wi-Fi Alliance
https://www.wi-fi.org/news-events/newsroom/wi-fi-alliance-delivers-wi-fi-6e-certification-program

7. Data aggregated from the Wi-Fi Alliance, last accessed November 12, 2021
https://www.wi-fi.org/product-finder-results?sort_by=certified&sort_order=desc&certifications=1335

8. *Wi-Fi 6 shipments to surpass 5.2 billion by 2025*; Wi-Fi Alliance
https://www.wi-fi.org/beacon/the-beacon/wi-fi-6-shipments-to-surpass-52-billion-by-2025

9. *Wi-Fi 6E Trials*; Wireless Broadband Alliance, last retrieved November 12, 2021
https://wballiance.com/wi-fi-6e-trials/

10. *Cisco Annual Internet Report (2018–2023) White Paper*; Cisco
https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

11. *Extended Reality (XR) Market Research Report*; Prescient & Strategic Intelligence
https://www.psmarketresearch.com/market-analysis/extended-reality-xr-market-insights

12. *US Virtual and Augmented Reality Users 2021*; eMarketer
https://www.emarketer.com/content/us-virtual-augmented-reality-users-2021

13. *What you should know about Wi-Fi 6 and the 6-GHz band*; Test & Measurements Tips
https://www.testandmeasurementtips.com/what-you-should-know-about-wi-fi-6-and-the-6-ghz-band/

SCTE
a subsidiary of CableLabs®

140 Philips Road, Exton, PA 19341-1318 | 800.542.5040 | scte.org