SCTE STANDARDS
JOURNAL

VOLUME 1 NO. 3 DECEMBER 2021

SCTE TECHNICAL JOURNAL

SCTE
a subsidiary of CableLabs®

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

SCTE TECHNICAL JOURNAL

VOLUME 1, NUMBER 3

December 2021

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.

We're looking into the future, and the view is great!

A year ago at this time we were rounding the corner of the first year of our SCTE Standards Program Explorer initiative and looking straight ahead at our new partnership with CableLabs®, two milestones that would help us take SCTE and our operator and vendor partners to new heights.

Having established bona fides in Telehealth and Aging in Place, our Explorer program has turned its focus to the bread-and-butter of Network Reliability and the sizzle of Next-Generation Video and Immersive Experiences. At the same time, we're drawing on the deep technical resources of CableLabs to give our members the knowledge they need to succeed in the future.

For a glimpse of what that looks like, check out our lineup in this latest SCTE Technical Journal, including:

- Deployment strategies for global coverage of quantum key distribution networks, by **CableLabs' Jing Wang** and **Bernardo Huberman**.
- The use of split-key signatures to enhance digital security, by **CommScope's Nicol So** and **Alexander Medvinsky**.
- A deep dive into LoRaWAN and its importance for smart cities infrastructures, by **Charter's Mohamed Daoud**, **Muhammad Khan**, and **Hossam Hmimy**.
- Rethinking of DDoS security in the era of volumetric attacks, by **Nokia Deepfield's Dr. Craig Labovitz, Stefan Meinders, and Alex Pavlovic**.
- Implementation and benefits of federated learning approaches for cable broadband operators, by **CableLabs' Thomas Sandholm** and **Sayandev Mukherjee**.
- A pragmatic approach to extending the DOCSIS® upstream, by **Teleste's Steve Condra, Kari Mäki, Niko Suo-Heikki,** and **Arttu Purmonen**.
- Best practices for operators to use AI/ML to up customer care games, by **IBM's Utpal Mangla** and **Luca Marchi**.
- A novel approach to using Wi-Fi signals to map and locate home devices, by **Comcast's Yonatan Vaizman** and **Hongcheng Wang**.

After almost two years of uncertainty, all of us are looking ahead to the opportunities that await as we move through 2022 to SCTE Cable-Tec Expo – September 19-22 in Philadelphia – and beyond. We are grateful to our Journal authors for giving us the tools we need to drive greater performance and customer satisfaction and invite you to consider sharing your knowledge in a future edition of the SCTE Technical Journal. In the meantime, please accept our best wishes for the holidays and the New Year.

From the SCTE Staff

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Deployment Strategies for the Global Coverage of Quantum Key Distribution Networks

A Technical Paper prepared for SCTE by

Jing Wang, Principal Architect, Next-Gen Systems, CableLabs, SCTE Member
858 Coal Creek CIR
Louisville, CO, 80027
j.wang@cablelabs.com
(303) 661-3337

Bernardo A. Huberman, Fellow & VP, Next-Gen Systems, CableLabs
400 W California Ave
Sunnyvale, CA 94086
b.huberman@cablelabs.com
(669) 777-9040

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

List of Tables

Title **Page Number**

1. Introduction

Modern telecommunication relies on cryptography to protect the security of data traffic, where the confidentiality and integrity of keys become the bottlenecks of the whole system. Today's cryptographic systems can be divided into two categories, symmetric and asymmetric. The security of asymmetric cryptographic algorithms, i.e., public key algorithms, relies on the computational complexities of intractable mathematical problems, e.g., the integer factorization problem (RSA), the discrete logarithm problem (Diffie-Hellman), and the elliptic-curve discrete logarithm problem (ECC) [1]. Solving these problems requires tremendous amounts of computational resources. While not feasible for classical computers, these problems can be solved in polynomial time by a quantum computer running Shor's algorithm [1-2]. To make things worse, increasing the key length does not help, since the required qubit number on a quantum computer only scales linearly with the key length [1]. In 2019, Google claimed to have achieved quantum supremacy [3]; whereas IBM argued that quantum computers will never reign "supreme" over classical computers, but rather will work in concert with them [4].

On the other hand, symmetric cryptographic algorithms, such as AES and SNOW 3G, are considered to be resistant against quantum computers. Although Grover's algorithm does speed up the attacks against symmetric ciphers, increasing the key length can effectively block these attacks [[1], [5]]. In modern communication, symmetric cryptography is only used for encryption and decryption. All other functions, such as signature, authentication, and key exchange, are carried out by asymmetric cryptography. Once sufficiently powerful quantum computers exist, classical cryptography will no longer be safe.

To address the challenges of quantum computing, two technological strategies were developed, post-quantum cryptography (PQC) and quantum key distribution (QKD). PQC, also known as quantum-safe or quantum-proof cryptography, focuses on increasing the computational complexity by inventing new intractable problems [5]. Thanks to its software implementation and full compatibility with existing systems, PQC is considered a good candidate for post-quantum eras, and three rounds of submissions have been organized by the National Institute of Standards and Technology (NIST) [[6],[7],[8]]. It is worth noting that, like their classical counterparts, PQC algorithms also rely on the assumptions of the computational power of attackers, i.e. they are only safe against quantum computers with a certain number of qubits, but may lead to long-term issues due to the ever-growing computational power of quantum computers.

QKD, also known as quantum cryptography, relies on quantum mechanics instead of mathematical assumptions to guarantee the security of keys [[9],[10],[11],[12]]. Instead of computational security, it offers information-theoretic security, i.e., the keys are deemed secure even if the adversary has unlimited computing power. From the original idea of QKD [13]to the first demonstration [14], various protocols [10]and network topologies [[11],[12]] have been reported. It was found later, however, that the absolute security of QKD is only guaranteed for ideal single-photon sources and detectors [15]. The lack of perfect single-photon sources and low detection efficiency create security loopholes, which may be exploited by side-channel attacks.

In a real system, expensive single-photon sources are replaced by weak coherent pulses (WCP), whose photon number follows the Poisson distribution, so there are always pulses containing multiple photons. These multi-photon pulses could become the target of photon-number-split (PNS) attacks. For example, Alice sends qubits to Bob via single-photon pulses. She does not have an ideal single-photon source and uses weak coherent pulses. The eavesdropper, Eve, can develop a strategy to block all single-photon pulses from Alice and divide all multi-photon pulses, keeping half for herself and sending the rest to Bob. In this way, Eve always gets the same information as Bob. To eliminate this loophole, decoy-state protocols were invented to vary the photon number per pulse [16] so that the blocking strategy of the eavesdropper will be revealed.

On the detector side, measurement-device-independent (MDI) protocols were proposed to close all detection loopholes and are immune to side-channel attacks on imperfect detectors [[17], [18]]. In conventional prepare-and-measure protocols, Alice prepares qubits and sends them to Bob, and Bob makes measurements on the received qubits. In MDI protocol, both Alice and Bob prepare random qubits independently, and send them to a third party, Charlie, for Bell-state measurement (BSM). Charlie announces successful BSM results, but he has no access to the qubits sent by Alice and Bob. So Charlie serves as an untrusted relay and can even be controlled by an eavesdropper. The post-selection of successful BSM events entangles the qubits from Alice and Bob, which is why MDI-QKD is equivalent to a time-reversed entangled-photon-pair (EPR) protocol. Therefore, MDI-QKD with the decoy-state method can negate the threats of both imperfect single-photon sources and detection losses.

2. Deployment Strategies of QKD Networks

QKD technologies have grown out of the laboratory and become ready to reach the market [[19], [20]]. In recent years, various demonstrations and field trials have been reported, including terrestrial QKD via optical fibers, free-space QKD on the ground, and free-space QKD to and from aircraft and satellites. Optical fiber-based terrestrial QKD networks include the DARPA quantum network in Boston [[21], [22]], the SwissQuantum network in Geneva [[23], [24]], the SECOQC network in Vienna [[25],[26],[27],[28]], metropolitan QKD networks in Tokyo [29] and Cambridge [30], and the Beijing-Shanghai QKD backbone network in China [31]. It should be pointed out that fiber-based terrestrial QKD is limited by short transmission distances, of usually less than 600 km in the lab and ~100 km in the field. This is because the key rate scales linearly with channel transmittance, which in optical fibers decays exponentially with distance due to photon absorption.

There are several strategies to extend the QKD distance, including quantum repeaters, trusted relays, and untrusted relays. Despite recent advances, a quantum repeater remains infeasible because it requires high-quality quantum memory and complicated local entanglement distillation. Trusted relays can infinitely extend the QKD distance but with the penalty of key leakage since the key information ceases to be quantum at each intermediate node. Untrusted relays, on the other hand, eliminate the possibility of key exposure and seem to be a promising candidate to extend QKD distance. Extensive research effort has been spent on MDI-QKD [[32],[33],[34],[35],[36],[37],[38]] and twin-filed QKD (TF-QKD) [[39],[40],[41],[42],[43],[44],[45],[46]], where Alice and Bob independently prepare random qubits and both send them to the relay node for measurement. Several demonstrations and field trials of time-bin phase-coding MDI-QKD have been reported in China [[32],[33],[34],[35]], featuring a metropolitan scale of less than 200 km between Alice and Bob and a key rate of only several bits per second [33]. A field trial demonstrated fiber distances of 15-30 km from users to the relay node with key rates of 16-38.8 bit/s [35]. More sophisticated three-intensity [36] and asymmetric four-intensity [[37],[38]] decoy-state protocols were proposed to further extend the distance and increase the key rate. The asymmetric four-intensity decoy-state protocol exploited three intensities (vacuum, weak, and signal states) in the X basis, and one intensity in the Z basis, and archived a distance record of 404 km using ultra-low loss optical fibers (0.16 dB/km) with a key rate of only 1.16 bit/s per hour [38].

Both conventional prepare-and-measure and MDI-QKD protocols have their key rates scaling linearly with the channel transmittance η, which decays exponentially with fiber distance. This linear bound severely limits the achievable key rate [39].TF-QKD, on the other hand, overcomes the linear key-rate constraint by matching the phases of two coherent states and encoding key information into the common phase. It has the key rate scaling with the square root of the channel transmittance while keeping the same untrusted relay merit as MDI-QKD [[39],[40],[41],[42],[43],[44],[45],[46]]. Using a practical sending-or-not-sending (SNS) protocol [39], several milestone experiments have been demonstrated to set new distance records for fiber-based terrestrial QKD, e.g. 509 km over ultra-low loss fiber in the lab [43], long-haul field trials over 511 km [44] and 428 km [45], and a dual-band stabilization technique to reduce Rayleigh scattering noise and achieve up to 605 km distance [46].

The point-to-point (P2P) nature of quantum channels and the requirement of dedicated fiber also hamper the wide deployment of terrestrial QKD networks. To enable the coexistence of quantum and classical channels in existing fiber infrastructures, wavelength division multiplexing (WDM) of quantum and classical channels has been investigated [[47],[48]]. Many reported works focus on the interference caused by spontaneous Raman scattering (SRS) from classical channels [[49],[50],[51],[52]]. So far, the coexistence of quantum and classical channels has been demonstrated in backbone [[53],[54]], metro [[55],[56],[57],[58]], and access [[59],[60],[61],[62],[63],[64],[65]] networks.

Due to the low channel loss in space and negligible interference from classical channels, satellite QKD drew significant research interest and has been considered as a promising candidate to provide global coverage of QKD networks [[66],[67]]. The feasibility studies of satellite QKD started in 2002 [[68],[69],[70]]. The first step toward satellite QKD is a ground-based free-space quantum link realized in 1996 [71] with a 150-m indoor path or a 75-m outdoor path. After that, several ground-based free-space QKD links were reported extending the link distance from 75 m to 144 km [[72],[73],[74],[75],[76]]. Then the road toward satellite QKD was paved by the demonstration of dynamic free-space QKD links, including flying transmitters placed on an airborne platform [[77],[78]] and moving quantum receivers placed on a truck [79] or aircraft [[80],[81]]. Most free-space QKD links were investigated as a preliminary study for satellite QKD, but they are not quite practical from the deployment perspective. They require line-of-sight (LoS) connections and are subject to geographical constraints (e.g. landscape and buildings) and environmental influences, such as vibration, adverse weather, and atmospheric turbulence. In real applications, free-space QKD links are used in the last segment of access networks.

Thanks to the low channel loss in space, satellite QKD has achieved distances of more than 1000 km [[82],[83],[84],[85],[86],[87],[88],[89],[90],[91],[92]]. Most reported works focused on low-earth-orbit (LEO) satellites, where a precise acquisition, pointing, and tracking system is required to follow the fast-moving satellite with high angular speed [[82],[83]]. The Micius satellite at ~500 km altitude realized downlink QKD from the satellite to ground over 1200 km [84]. As a trusted relay, it also enables intercontinental quantum-secured communication over 7600 km between China and Austria [[85],[86]]. Although downlink QKD from a satellite to the ground has the potential for higher detection efficiency and higher key rates, it requires more payload on the satellite and is not as flexible as an uplink configuration, where a simple payload of a quantum receiver is placed on a spacecraft. Micius uses a downlink for QKD and entanglement distribution, and it is also compatible with uplink for quantum teleportation [83]. Canada’s satellite plan (QEYSSat) employs an uplink scheme [87], and many works have been done to verify the feasibility of high channel loss [[88],[89],[90]], optical terminal design [91], and noise of single-photon detectors (SPDs) in space [92]. To further simplify the payload on satellite, a corner cube retroreflector with a modulator for polarization encoding is proposed [93]. Besides LEO, medium earth orbit (MEO) satellites [94] and geostationary orbit (GEO) [[95],[96]] provide longer flyover time windows and larger coverage areas, but with the penalty of higher channel loss and lower key rates. Their feasibility is also under investigation. Miniaturization and standardization of satellites are also trends of satellite QKD [[97],[98],[99],[100],[101]].

All the aforementioned satellite QKD schemes utilize the satellite as a trusted relay. To eliminate the key leakage at the satellite, satellite-to-ground entanglement distribution has been demonstrated [[102],[103],[104],[105]] with a distance record of 1200 km [104]. Before that, free-space entanglement distribution on the ground was studied [[106],[107],[108],[109]] over a distance of more than 100 km in the atmosphere [[107],[108],[109]]. Moreover, free-space MDI-QKD was also demonstrated as an alternative to entanglement distribution [110].

So far, many deployment strategies for QKD have been developed, including terrestrial QKD based on optical fibers, QKD via ground-based free-space links and air-to-ground links, and satellite QKD. Since each method has its strengths and limitations, none of them can achieve global coverage alone. So far as we know, there is no investigation to compare the pros and cons of different deployment technologies. In this paper, we present a literature overview and a comparative study of existing deployment strategies of QKD and compare their pros and cons in terms of channel loss, interference, distance limit, connection topology, deployment cost, and use scenarios. Selection criteria and deployment requirements for different network segments are developed to enable a global coverage of QKD networks, from intercontinental, long-haul, to metro and access networks.

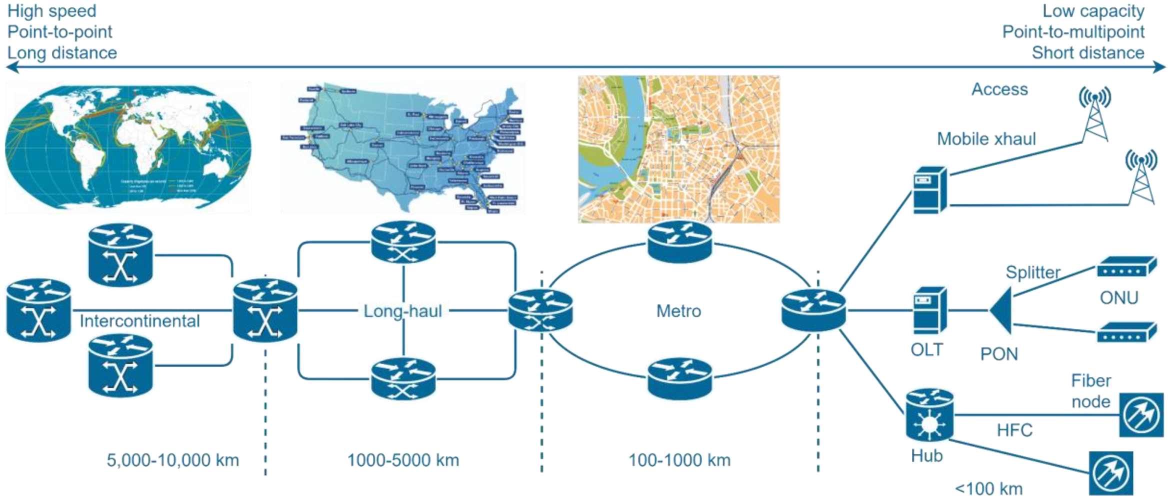


Figure 1 - Global coverage of telecommunication networks, from the intercontinental, long-haul to metro and access networks

Figure 1 shows a global telecommunication network, which can be divided into four segments according to the coverage area, intercontinental (>5000 km), long-haul (1000-5000 km), metro (100-1000 km), and access (<100 km). Each segment features different connectivity topologies. Intercontinental and long-haul networks feature P2P topology; metro networks utilize ring and mesh topologies; access networks have tree or star topologies.

3. Terrestrial QKD via Optical Fibers

Figure 2(a) shows the architecture of a terrestrial QKD link via optical fibers. To avoid the interference caused by SRS noise from classical channels, the quantum channel is ideally deployed in a dedicated dark fiber. In case of fiber deficiency, it can also be deployed in the same fiber with the classical channel using time/wavelength-division multiplexing (TDM/WDM) techniques. There are several techniques to reduce the interference from classical channels, such as spectral filtering before the quantum receiver, temporal filtering (i.e., gated single-photon detectors), and power control of classical data traffic.

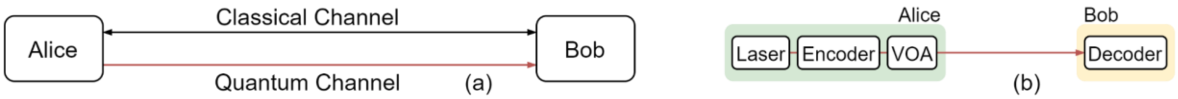


Figure 2 - Terrestrial QKD via optical fibers. (a) To avoid interference from classical channels, the quantum channel is deployed in a dedicated fiber. (b) The setup of a prepare-and-measure QKD link.

Figure 2(b) shows the setup of a prepare-and-measure QKD link. So far, several fiber-based terrestrial QKD networks have been demonstrated, including the DARPA quantum network in Boston [[21],[22]], SwissQuantum network in Geneva [[23],[24]], SECOQC network in Vienna [[25],[26],[27],[28]], metropolitan QKD networks in Tokyo [29] and Cambridge [30], and Beijing-Shanghai QKD backbone network in China [31]. Most of them are based on the prepare-and-measure protocol with limited distances to hundreds of km. The usable distance will be further reduced to ~100 km in real deployments. This is because the achievable key rate scales linearly with channel transmittance, whereas in optical fibers the channel transmittance decays exponentially with distance due to the photon absorption, making fiber-based terrestrial QKD impractical for long-haul applications. For example, with a loss of 0.2 dB/km, a 1000 km fiber introduces a channel loss of 200 dB, which is so high that only 0.3 photons arrive at the receiver per century even if a 10-GHz single-photon source was used at the transmitter.

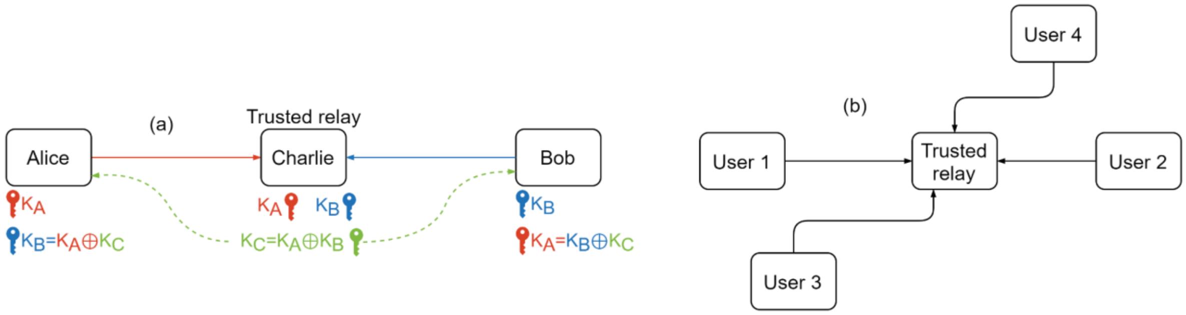


Figure 3 - Trusted relay for terrestrial QKD networks. (a) The operation principles of a trusted relay. (b) A trusted relay node offers compatibility with point-to-multipoint networks

Relay technologies are essential to increase the distance and enhance the coverage area of terrestrial QKD networks. There are two categories of relay technologies, trusted and untrusted, depending on whether or not the relay node has access to the keys. The operation principles of a trusted relay node are shown in Figure 3(a). It connects two neighboring nodes that are too far away from each other to establish a direct

QKD link. The trusted relay node, Charlie, performs QKD with Alice and Bob and obtains keys of KA and KB, respectively. He then makes a parity announcement of KC=KA⨁KB, which is a bitwise parity-check of KA and KB. Since the original keys are independent bit strings, their bitwise parity is a uniformly random bit string, which does not reveal any information about the keys. With the help of KC, both Alice and Bob can then infer the key of each other using the fact that KA⨁(KA⨁KB)=KB and KB⨁(KA⨁KB)=KA. Trusted relays can extend the distance of secure communication without limitation, but with the penalty of key exposure at each relay node. An interesting synergy is that classical fiber cables have repeaters every 100 km for the reamplification, reshaping, and retiming of classical pulses. Trusted relay nodes can be deployed at the same locations as classical repeaters. Since classical repeaters have fixed and public locations, relay nodes collocated with repeaters will be subject to constant surveillance and probing. For example, the Beijing-Shanghai backbone link in China uses 32 trusted relay nodes to divide the overall distance of more than 2000 km into many small segments, each less than 100 km. Moreover, a trusted relay node offers compatibility to the point-to-multipoint (P2MP) network topology, as shown in Figure 3(b).

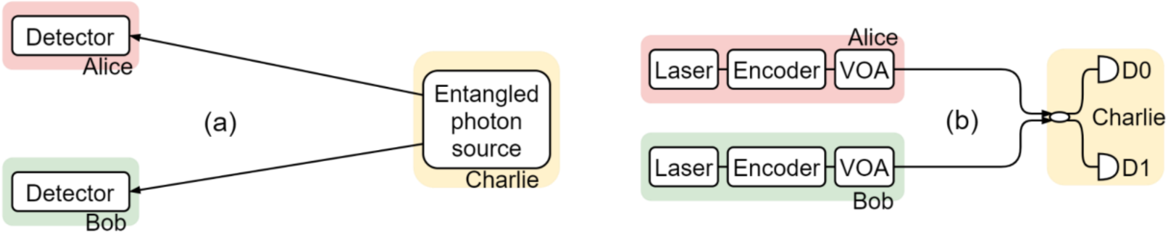


Figure 4 - Untrusted relay for terrestrial QKD networks. (a) Distribution of entangled photon pairs. (b) Measurement-device-independent (MDI) QKD

On the other hand, an untrusted relay eliminates the key leakage at the relay node. It can be implemented by the distribution of entangled photon pairs or MDI-QKD. In either case, the relay node has no information on the keys and could even be an eavesdropper itself. Figure 4(a) shows an entanglement distribution setup, where an entangled photon source at the relay node generates entangled photon pairs (EPR) using a nonlinear crystal or nonlinear fibers. The entangled photons are distributed to two users, who make independent measurements and get correlated results. The relay node is considered secure since the entangled photon source has no access to the exact states of two photons, but the measurement results of two users are always correlated. Figure 4(b) shows an MDI-QKD setup. Two users independently prepare random qubits and send them to the relay node for Bell state measurements (BSM). Although the BSM cannot tell the exact states of two incoming photons, it can tell whether or not the two photons have entangled states. By post-selecting entangled photons from the two users, MDI-QKD is equivalent to a time-reversed EPR protocol. So far, the distance record for MDI-QKD is 404 km using asymmetric four-intensity decoy-state protocol in ultra-low loss optical fibers [38]. The key rate, however, is only 1.16 bit/s per hour, which is orders of magnitude lower than practical requirements.

The key rate of conventional QKD, including prepare-and-measure protocols, entanglement distribution, and MDI-QKD, has linear dependency on the channel transmittance η. Since the channel transmittance decays exponentially with distance in optical fibers, this linear bound severely limits the achievable key rate and distance of terrestrial QKD networks. Recently, a new QKD protocol, twin-field (TF) QKD, was proposed to overcome the linear key-rate constraint. Its setup is almost identical to a phase-encoding MDI-QKD and maintains the same merit of an untrusted relay, with pairs of phase-randomized optical fields generated at two distant locations and combined at a central measuring station. Fields imparted with the same random phase are ‘twins’ and can be used to distill a key. By matching the phases of two coherent states and encoding key information into the common phase, TF-QKD exhibits the same dependence on distance as quantum repeaters, i.e. its key rate scales with the square root of the channel transmittance. Several milestone experiments have been demonstrated to set new distance records of fiber-based terrestrial QKD links [[43],[44],[45],[46]]. It should be noted that in MDI-QKD, the two photons from two users interfere at the relay station. Charlie’s receiver has two-photon interference and records coincidence detections. In TF-QKD, however, two optical fields are sent from both users and Charlie’s receiver has a single-photon interference followed by a single-photon detection event. TF-QKD retains the characteristics of MDI-QKD, whereas gaining extra distance thanks to the square-root dependence of key rate on the channel transmittance.

4. Free-Space QKD

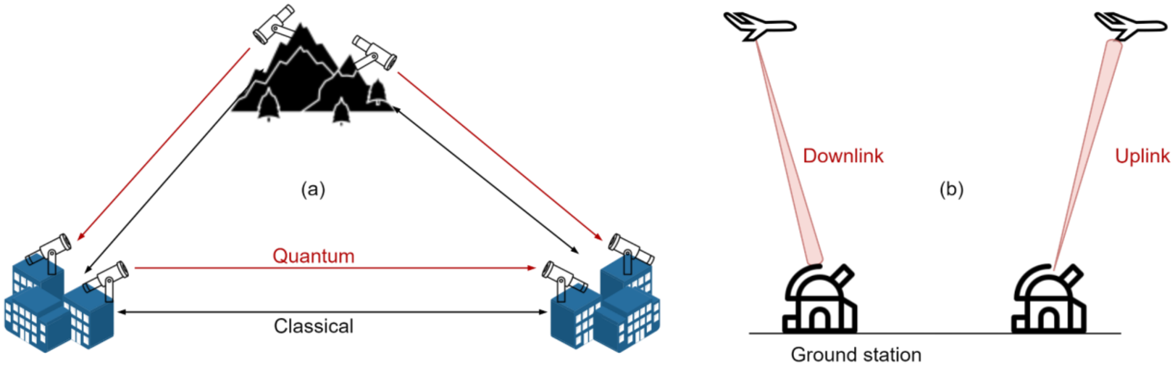


Figure 5 - Free-space QKD. (a) Ground-based free-space QKD links. (b) Free-space QKD links to/from an aircraft

Figure 5 shows the architectures of free-space QKD. Figure 5(a) shows ground-based free-space QKD links. Different from optical fibers, free-space QKD requires LoS connections, and the transmitters and receivers are usually deployed on top of buildings or mountains to avoid obstruction in the path. The associated classical channels could exploit wireless links, e.g. cellular, microwave, or rely on free-space optics as well. Since no fiber trenching is required, free-space QKD features low deployment cost, easy and fast installation, and is an important reinforcement for fiber-based QKD networks owing to its configurational flexibility. The distance record for ground-based free-space QKD is 144 km [76]. Dynamic free-space QKD links to/from an aircraft were investigated as a preliminary step toward satellite QKD and the feasibility of both downlink and uplink configurations have been verified, shown in Figure 5(b). A downlink scheme includes a flying transmitter on an airborne platform and a receiver on the ground [[77],[78]]; an uplink configuration uses a ground-based transmitter and places a quantum receiver on aircraft [[80],[81]]. The downlink scheme has higher detection efficiency, whereas the uplink scheme has a smaller payload on aircraft.

Since the quantum channel is not confined in the waveguide of optical fibers, free-space QKD is subject to environmental influence, such as vibration, adverse weather (fog, rain, cloud), and atmospheric turbulence. Although the atmosphere has lower absorption than optical fibers, only 0.07 dB/km at 2400 m, the channel loss of free-space QKD is not dominated by absorption. Instead, it is determined by diffraction, weather, turbulence, and misalignment. Moreover, free-space quantum channels are subject to decoherence more than those in optical fibers, which further limits the link distance. On the other hand, there is no interference from classical channels in free-space and the coexistence of quantum and classical channels is no longer an issue. Free-space QKD can easily support P2MP topologies, making it a promising candidate for inter-building secure communication in the last few miles of access networks.

5. Satellite as a Trusted Relay

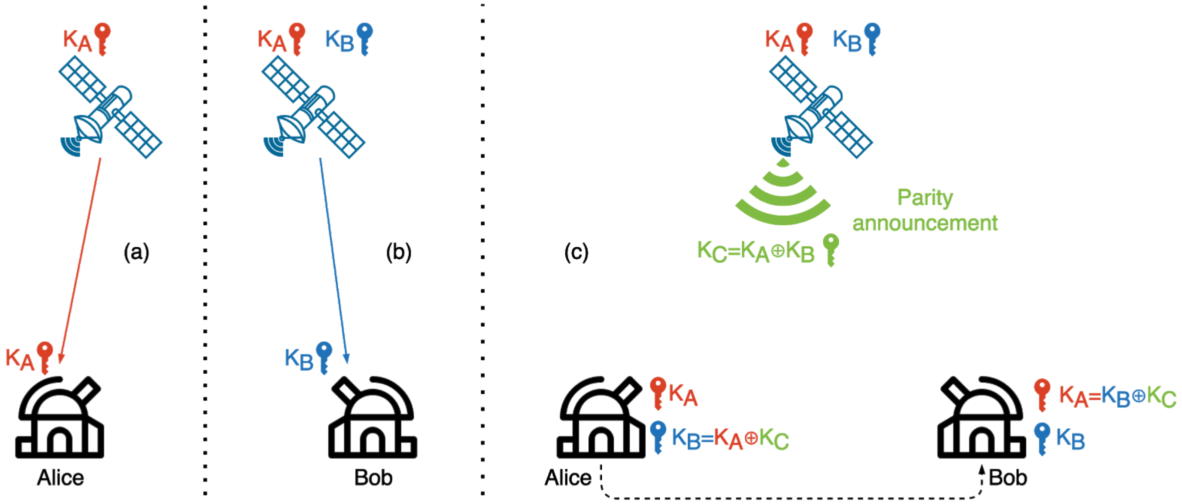


Figure 6 - Satellite as a trusted relay. The satellite first exchanges keys with ground stations, Alice (a) and Bob (b), respectively, then makes a parity announcement (c) so that Alice and Bob can infer each other's keys.

Thanks to the low channel loss in space, negligible interference from classical channels, and reduced environmental influences, satellite QKD can achieve distances of more than 1000 km. It is not limited by terrestrial conditions and can provide coverage for rural areas. Most reported work focuses on LEO satellites with altitudes of less than 900 km, requiring a precise acquisition, pointing, and tracking system to follow the fast-moving satellite. The feasibility of MEO and GEO satellites is also under investigation. Miniaturization and standardization of satellites are also trends of satellite QKD. Figure 6 shows the operation principles of satellite QKD where the satellite is used as a trusted relay. An LEO satellite performs downlink QKD with two ground stations, Alice and Bob, respectively. It then makes a parity announcement, so that Alice and Bob can infer each other's keys. The satellite needs LoS connections with Alice and Bob, but not necessarily at the same time. It can exchange keys with several ground stations one after another as it flies over them. As a trusted relay, any access to the satellite leaks the complete information about keys. The associated classical channels for satellite QKD can rely on terrestrial fibers, microwave, or free-space laser communication in space. For example, most Starlink satellites are currently operating in Ku and Ka bands and can be upgraded to laser communication in the future.

Since the effective thickness of the atmosphere is only ~10 km, the propagation of a quantum channel takes place mostly in vacuum space with negligible absorption and turbulence. Instead of absorption, the channel loss of satellite QKD is determined by beam diffraction and scales quadratically with distance. In comparison, the channel loss of terrestrial QKD is dominated by fiber absorption and scales exponentially with distance. Channels in space also have smaller decoherence than those in the atmosphere or optical fibers. For example, a 600-km optical fiber has a channel loss of 120 dB, whereas a link of the same length in space from satellite to the ground has a loss of only 50 dB given a reasonable aperture size is used at the receiver telescope. This is why satellite QKD can reach much longer distances. Inter-satellite channels have even lower losses due to the absence of atmosphere.

Channel loss in space comes from two sources, beam diffraction and beam spreading beyond the effects of diffraction. Diffraction loss depends on the divergence of the transmitter telescope and the aperture size of the receiver telescope. Further beam spreading arises from wavefront aberrations caused by refractive index inhomogeneities due to atmospheric turbulence. There are two categories of turbulence. Small turbulence induces beam spreading, whereas large turbulent eddies with sizes larger than the beam spot cause beam wandering. A long-term beam spot is a superposition of moving short-term beam spots. The short-term beam size is determined by spreading and the instantaneous beam displacement from the unperturbed position caused by beam wandering. In real applications, the channel loss from a satellite to a ground station is dominated by diffraction, followed by beam spreading. Beam wandering and absorption have negligible contributions to the channel loss.

Satellite QKD has three different schemes, downlink, uplink, and retroreflection. In Figure 7(a), the downlink scheme has the quantum transmitter on a satellite and receiver on the ground. Since the effective thickness of the atmosphere is only ~10 km, the optical beam first propagates through vacuum space where the only channel loss is diffraction, then passes through the atmosphere in the final stage of the path. Due to the diffraction effect, when the beam arrives at the atmosphere, its size has been larger than most turbulent eddies. There is no beam wandering and the beam size is spread slightly by wavefront aberrations caused by turbulence. For the downlink configuration, atmospheric turbulence has a limited impact on the channel loss and beam spreading. For example, the beam size after 1200 km downlink propagation expands to 12 m with diffraction loss of ~22 dB depending on the receiver telescope size [84]. Atmospheric turbulence introduces additional 3-8 dB attenuation, with an overall channel loss of less than 30 dB [84].

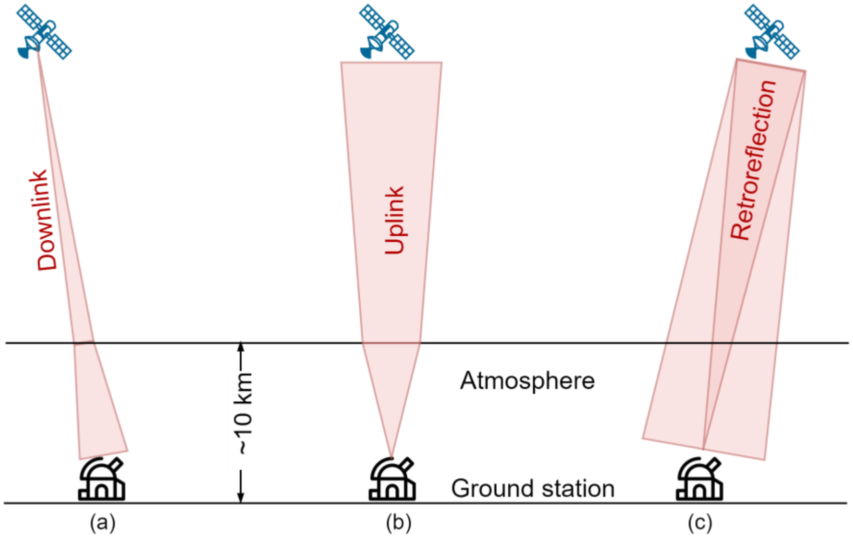


Figure 7 - Downlink (a) and uplink (b) configurations of satellite QKD

In Figure 7(b), an uplink channel first propagates through the atmosphere, where the wavefront aberration induced by turbulence causes significant beam spreading. At 500 km altitude, the beam size of an uplink channel can reach up to 50 m, much larger than any available spaceborne telescope aperture. Downlink channels can exploit large aperture receiver telescopes on the ground, but uplink channels have limited aperture size for receiver telescopes due to the weight and size limit on satellites. Thanks to the strong wavefront aberration, large beam spot, and small aperture size, uplink channels have higher channel loss than downlink ones. For example, a 500 km uplink channel has a loss up to 50 dB; whereas a downlink channel of the same length would have a loss less than 20 dB [88]. Most uplink channels cannot work without the help of the decoy-state technique [88].

Although the downlink scheme has higher detection efficiency and higher key rates, the transmitter setup requires more payload on the satellite and needs more adjustment during operation, which makes the downlink scheme not as flexible as an uplink configuration. The uplink scheme, on the other hand, only needs a simple payload of quantum receivers on the satellite, enabling an easier operation on the satellite. The downlink scheme leaves expensive and delicate SPDs on the ground for better protection, cooling, and maintenance. The uplink scheme has to launch the sensitive SPDs into space, a process involving launch vibration, shock in the flight, extreme temperature, and work under adverse conditions in space. Due to the sunlight, the satellite temperature varies by up to tens of degrees in one orbit, and there is limited electrical power on the satellite for cooling. The only way to dissipate heat is by radiation. To make things worse, most SPDs are avalanche photon detectors (APD), which are sensitive to dark counts caused by ionizing radiation in space. The feasibility of low-noise SPDs on a satellite is under investigation [92]. So far, downlink and uplink schemes are both considered important for future satellite QKD. For example, Micius uses downlink QKD and entanglement distribution, and it is also compatible with uplink for quantum teleportation [83]. Canada’s satellite plan (QEYSSat) employs an uplink scheme [87], and many works have been done to verify the feasibility of high channel loss [[88],[89],[90]], optical terminal design [91], and noise of SPDs in space [92].

In a quantum channel, the qubits are carried by single photons and no amplification is allowed. The only way to increase the signal-to-noise ratio (SNR) is to reduce channel loss and background noise. Thanks to the low loss, downlink channels have larger SNR than uplink ones. In the daytime, the background noise from sunlight makes it difficult to establish a QKD link. One way to improve SNR in the daytime is to use the wavelengths at Fraunhofer lines, i.e. sun absorption lines. At night, background noise is dominated by moonlight and scattered light from human activities, depending on the locations of the ground stations. SNR at night is orders of magnitude higher than it is in the daytime, which is why most satellite QKD works were demonstrated at clear night by downlink channels. There are several techniques to improve the SNR of a free-space quantum link, e.g., reducing the beam size, reducing the field-of-view of the receiver telescope, narrowband spectral filtering before the receiver, and temporal filtering (gating window) of SPDs.

To further simplify the payload on satellites, a third configuration, retroreflection, was proposed [[93],[94]], as shown in Figure 7(c). It uses an orbiting corner cube retroreflector on a satellite with a modulator to encode polarizations. The single-photon transmitter is realized by corner cube retroreflectors mounted on a satellite. Only the reflected beam from the satellite to the ground is a quantum channel; the laser beam from the ground station to the satellite is a classical channel with strong pulse intensity. This configuration features a compact and low-cost payload on satellite and can be used on not only LEO but also MEO and GEO satellites. The feasibility of single-photon exchange from an MEO satellite using a retroreflection scheme has been verified [94].

6. Satellite as an Untrusted Relay

When a satellite is used as a trusted relay, it has access to all the keys of all ground stations. To avoid the key leakage at the satellite, untrusted relaying is preferred since the eavesdropper gets no information even if it takes full control of the satellite. Figure 8 shows the architecture of satellite QKD with the satellite as an untrusted relay. Figure 8(a) shows entanglement distribution, where an entangled photon source on a satellite sends entangled photons down to two ground stations, Alice and Bob, respectively. Alice and Bob make independent measurements on the incoming photons and get correlated results. Since the entangled photon source has no control over the exact qubits carried by each photon, the satellite has no information of the key. For entanglement distribution, the loss of two downlink channels has to be combined since only photon pairs that both arrive at ground stations can be used for keys.

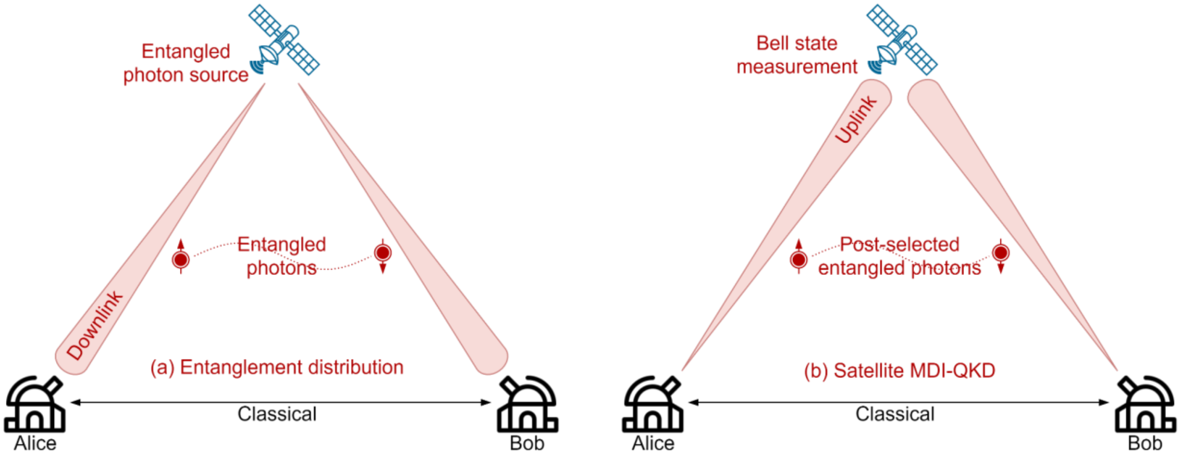


Figure 8 - Satellite QKD with the satellite as an untrusted relay. (a) Entanglement distribution from a satellite. (b) Free-space MDI-QKD to a satellite

As an alternative, Figure 8(b) shows satellite MDI-QKD, where two ground stations independently prepare random qubits and send them via uplink channels to a satellite for BSM. Satellite MDI-QKD is equivalent to a time-reversed entanglement distribution protocol. The BSM can only tell whether or not the two photons are entangled, but it cannot tell the exact states of two incoming photons. The loss of two uplink channels has to be combined since only photon pairs that both arrive at the satellite can be used for keys. Due to the high loss of uplink channels, there is no demonstration of satellite MDI-QKD so far. But the feasibility study of free-space MDI-QKD has been reported on the ground over 19.2 km [110], well beyond the effective thickness of the atmosphere (~10 km).

Unlike trusted relaying, untrusted relaying requires simultaneous LoS connections from the satellite to both ground stations, which limits the separation distance between ground stations. For a given altitude of the satellite, wider separation between ground stations makes lower slant angles and longer propagation in the atmosphere, which leads to higher channel loss. The current distance record for entanglement distribution is ~1200 km, achieved by an LEO satellite Micius of China [104].

7. Deployment Strategies for Global Coverage

Table 1 lists the pros and cons of different deployment strategies of QKD networks, including fiber-based terrestrial QKD, free-space QKD including ground-based and ground-to-air schemes, satellite QKD with the satellite used as a trusted or untrusted relay. Terrestrial QKD via optical fibers suffers from high channel loss and short distance but offers compatibility with existing fiber infrastructure and P2MP topologies. Since the quantum channels are confined in fiber waveguides, terrestrial QKD networks can operate all day in adverse environments, such as background light, weather, and vibration. Without relays, a single span of fiber-based QKD can reach ~100 km in the field, only suitable for metro and access networks. Trusted relaying can extend the distance of fiber-based QKD unlimitedly with the penalty of key leakage at each relay node. An interesting synergy is that classical fiber cables also have repeaters every 100 km. Trusted relay nodes can be deployed at the same locations as classical repeaters. Since classical repeaters have fixed and public locations, relay nodes collocated with repeaters will be subject to constant surveillance and probing. In contrast, satellite QKD using a satellite as a trusted relay is more secure because the satellite and quantum links are moving fast, making side-channel attacks difficult.

Table 1 - Pros and cons of fiber-based terrestrial QKD, free-space QKD, and satellite QKD

Deployment strategies	Fiber-based terrestrial QKD	Ground-based Free-space QKD	Ground-to-air free-space QKD	Satellite QKD (trusted relay)	Satellite QKD (untrusted relay)
Attenuation	Fiber absorption	Diffraction Turbulence Weather Absorption	Same as ground-based QKD plus Misalignment Vibration	Diffraction Turbulence Weather	Diffraction Turbulence Weather
Interferences from classical channels	Spontaneous Raman scattering noise	No	No	No	No
Channel loss	High, scale exponentially with fiber length	High, scale exponentially with distance		Low Scale quadratically with distance	
Distance	~100 km in fields without relay unlimited distance with trusted relay MDI-QKD: 404 km in the lab [38] 200 km in fields [33] TF-QKD: >500 km in lab [[43], [44], [46]] 428 km in fields [45]	144 km in experiments [76] <10 km in real fields	96 km in experiments [78]	Satellite to the ground distance over 1200 km [84] Unlimited distance between ground stations	1200 km between ground stations for a 500-km altitude LEO satellite longer for MEO/GEO
Compatibility to P2MP topology	P2MP	P2MP	P2P at a time	P2P at a time	satellite to two ground stations
Line-of-sight	No	Yes	Yes	Yes	Simultaneous LoS with both ground stations
Time window	whole day	Only night need special care for daytime operation		Short window in the clear night	
Deployment	Low cost Dedicated fiber or reuse existing ones	Low cost Simple and fast installation No fiber trenching		Expensive and slow Synergy with satellite laser communication in space	
Application scenarios	Metro, access	last few miles of access networks		Long-haul	

Ground-based free-space QKD requires LoS connections, and the transmitters and receivers are usually deployed on top of buildings or mountains to avoid obstruction in the path. It supports P2MP topology and can handle the coexistence of quantum and classical channels without interference. These features make it suitable for the last few miles of access networks among buildings. Although the atmosphere has lower absorption, the channel loss of free-space QKD is dominated by diffraction, adverse weather, and atmospheric turbulence. The distance record of ground-based free-space QKD is 144 km [76], but in real deployments, the usable distance will be less than 10 km for practical key rates. Since no fiber trenching is required, ground-based free-space QKD features low deployment cost, fast and easy installation, and serves as an important reinforcement for fiber-based QKD networks. The ground-to-air free-space QKD shares the same pros and cons of ground-based counterparts plus the additional channel loss caused by misalignment and vibration due to the movement of the aircraft. We do not include the applications of airborne free-space QKD here as it was mainly investigated as a preliminary step towards satellite QKD.

Compared with terrestrial and free-space QKD, satellite QKD features low channel loss and long distances. The downlink scheme from satellite to the ground has higher detection efficiency and higher key rates thanks to the lower loss and less turbulence-induced wavefront abbreviation. But it requires more payload on the satellite and needs more adjustment during operation. The uplink channels are more flexible, since it only needs a simple payload of quantum receivers on the satellite, enabling an easier operation on the satellite. On the other hand, the downlink scheme leaves expensive and delicate SPDs on the ground for easiest maintenance; the uplink scheme, on the other hands, launches the sensitive SPDs into space, which have to go through the launch vibration, shock in the flight, extreme temperature, and work under adverse conditions in space.

Satellite QKD requires LoS connections between the satellite and ground stations and only works at night due to the background noise from sunlight during the daytime. To reduce the channel loss, LEO satellites are preferred, but low altitude leads to the fast movement of the satellite, a small coverage area, and a short flyover time window for each ground station. MEO satellites at higher orbit provide wider coverage and longer flyover time, but with the penalty of higher channel loss and lower key rate [94]. To choose an appropriate altitude, a trade-off must be made between the coverage area and time window versus channel loss and key rate. An extreme example is a GEO satellite, which has an operational time window of the whole night but with a long path length of 35,786 km [[95],[96]].

There is a strong synergy between satellite QKD and classical satellite communication. For example, space communication also exploits LEO satellites at an altitude of 300-1000 km. Starlink plans to launch thousands of satellites at altitudes of 350-580 km. Although these satellites are using microwave communication in Ku and Ka bands, most of them are equipped with optical transceivers for future upgrades to laser communication. By adding quantum transmitters onboard, these satellites can be used as a trusted relay for QKD in space. Since quantum transmitters for most prepare-and-measure protocols only consist of commercial off-the-shelf devices, this upgrade will not significantly increase the satellite cost. The beam acquisition, tracking, and pointing systems designed for laser communication in space can also be reused by quantum channels. Satellite QKD covers long-haul networks, and by using the satellites as a trusted relay, its secure distance can be extended unlimitedly.

The scheme with a satellite as an untrusted relay shares the same pros and cons with trusted relays but eliminates the key leakage at satellites. It requires simultaneous LoS connections from the satellite to both ground stations, which limits the separation between two ground stations. The distance record of entanglement distribution from a satellite is 1200 km, which can be employed for long-haul networks, but not long enough for intercontinental connections. Figure 9 shows the deployment strategies for global coverage of QKD networks, from the intercontinental, long-haul, metro to access networks.

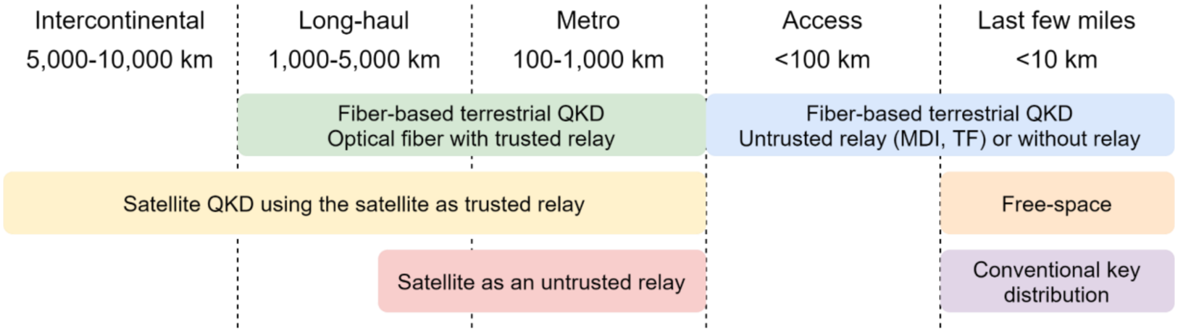

Intercontinental
5,000-10,000 km
Long-haul
1,000-5,000 km
Metro
100-1,000 km
Access
<100 km
Last few miles
<10 km
Fiber-based terrestrial QKD
Optical fiber with trusted relay
Fiber-based terrestrial QKD
Untrusted relay (MDI, TF) or without relay
Satellite QKD using the satellite as trusted relay
Free-space
Satellite as an untrusted relay
Conventional key distribution

Figure 9 - Deployment strategies for global coverage of QKD networks

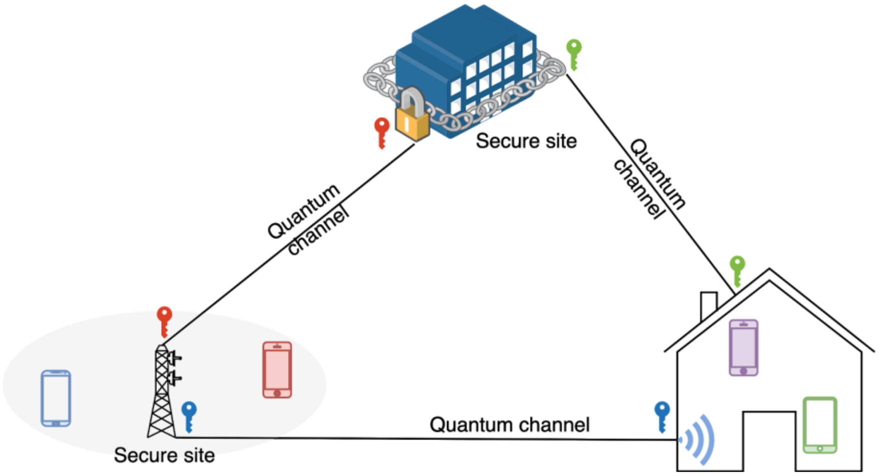


Figure 10 - Hierarchical key delivery over the last few miles in access networks

It should be noted that not all user devices are equipped with optical terminals for fiber or free-space optics connections. Radio access has been and will continue to be used extensively in the last few miles of access networks. In these cases, keys have to be distributed wirelessly in a classical way to user devices. Figure 10 shows a hierarchical key delivery architecture. Several secure sites, e.g. bank buildings, business campuses, government offices, are connected by satellite, fiber-based or free-space QKD links, so the keys are delivered in an absolutely secure way among these secure sites. Within each secure site, however, the keys are distributed wirelessly to mobile users using PQC algorithms. This involves a trade-off between security and mobility because it is not feasible to connect all devices with optical fibers or free-space optics. We thus have to leverage the ubiquity and flexibility of radio access technologies in the last few miles.

In this hierarchical architecture, two different levels of security-as-a-service (SaaS) are provided: absolute security over long-distance among secure sites; and computational classical security over a short distance within each site. Once the mobile users get the keys, they can use these keys to encrypt their wireless communication. They can even roam away from the secure site and continue the secure communication as soon as they still possess the keys. Once they consume all the keys, they have to return to a secure site to fetch new keys. It should be noted that PQC and QKD do not necessarily compete with each other. Instead, they should work in an orchestrated way to complement each other. For example, PQC could exploit the keys delivered by QKD to enhance its security, while QKD can employ PQC for authentication, which cannot be handled by QKD itself.

8. Conclusions

To date, many deployment strategies of QKD networks have been demonstrated, but none of them provides global coverage of QKD networks. A comparative study on the pros and cons of various deployment strategies is still missing. In this paper, the state-of-the-art deployment technologies of QKD networks, including fiber-based terrestrial QKD, free-space QKD, and satellite QKD, are compared in terms of channel loss, interference, distance limit, connection topology, deployment cost, and application scenarios. Instead of competing with one another, these different deployment strategies will work in an orchestrated way to complement each other and enable a global coverage of QKD networks, from intercontinental, long-haul, to metro and access networks.

Given its compatibility with P2MP topology and ~100-km distance limit without relay, fiber-based terrestrial QKD is suitable for metro and access networks. With the help of a trusted relay, the QKD distance can be extended unlimitedly to cover long-haul networks, where the relay nodes are collocated with classical fiber repeaters. Ground-based free-space QKD is limited to 10 km due to diffraction, weather, and atmosphere turbulence, and is suitable for the last few miles among buildings in access networks. Satellite QKD features low channel loss, high key rates, and long distances more than 1000 km. By utilizing satellites as trusted relays, the QKD distance can be extended infinitely and can be used for intercontinental, long-haul, and metro networks. Furthermore, satellite QKD is not restricted by terrain conditions and can reach rural underserved areas without difficulty. On the other hand, using an LEO satellite as an untrusted relay requires simultaneous LoS connections from the satellite to both ground stations, where the separation between the ground stations is limited by the altitude of the satellite. MEO and GEO satellites feature longer time windows and larger coverage areas, but with the penalty of higher channel loss and lower key rate.

9. Abbreviations

APD	avalanche photon detector
BSM	Bell-state measurement
ECC	elliptic curve cryptography
EPR	entangled photon pair
GEO	geostationary orbit
LEO	low-earth-orbit
LoS	line-of-sight
MEO	medium-earth-orbit
MDI	measurement-device-independent
NIST	National Institute of Standards and Technology
P2P	point-to-point
P2MP	point-to-multipoint
PNS	photon-number-split
PQC	post-quantum cryptography
QKD	quantum key distribution
SaaS	security-as-a-service
SNR	signal-to-noise ratio
SNS	sending-or-not-sending

SPD	single-photon detector
SRS	spontaneous Raman scattering
TDM	time-division multiplexing
TF-QKD	twin-filed QKD
WCP	weak coherent pulses
WDM	wavelength division multiplexing

10. References

[1] P. W. Shor, “Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer,” SIAM Journal on Computing, vol. 26, no. 5, pp. 1484-1509, 1997.

[2] E. Grumbling, M. Horowitz, (eds.), “Quantum Computing: Progress and Prospects,” The National Academies Press, Washington, DC, https://doi.org/10.17226/25196.

[3] F. Arute, K. Arya, R. Babbush, “Quantum supremacy using a programmable superconducting processor,” Nature, vol. 574, pp. 505-510, October 2019.

[4] IBM Research Blog, “On ‘Quantum Supremacy’” 22 October 2019.

[5] D. J. Bernstein, “Introduction to post-quantum cryptography,” In: D. J. Bernstein, J. Buchmann, E. Dahmen (eds) Post-Quantum Cryptography. Springer, Berlin, Heidelberg, 2009.

[6] L. Chen, S. Jordan, Y.-K. Liu, et. al., "Report on Post-Quantum Cryptography," NISTIR 8105, 04/28/2016.

[7] G. Alagic, J. Alperin-Sheriff, D. Apon, et. al., "Status Report on the First Round of the NIST Post-Quantum Cryptography Standardization Process," NISTIR 8240, 01/31/2019.

[8] G. Alagic, J. Alperin-Sheriff, D. Apon, et. al., "Status Report on the Second Round of the NIST Post-Quantum Cryptography Standardization Process," NISTIR 8309, 07/22/2020.

[9] C. H. Bennett, “Quantum Cryptography: Uncertainty in the Service of Privacy,” Science, vol. 257, no. 5071, pp. 752-753, 1992.

[10] N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, “Quantum cryptography,” Reviews of Modern Physics, vol. 74, no. 1, pp. 145-195, Mar 2002.

[11] V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, et al., “The security of practical quantum key distribution,” Reviews of Modern Physics, vol. 81, no. 3, pp. 1301-1350, Sep 2009.

[12] H. K. Lo, M. Curty, K. Tamaki, “Secure quantum key distribution,” Nature Photonics, vol. 8, pp. 595-604, 2014.

[13] C. H. Bennett and G. Brassard, “Quantum cryptography: public key distribution and coin tossing,” Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing, Bangalore, India, 10-12 December 1984, pp. 175-179.

[14] C. H. Bennett, F. Bessette, G. Brassard, et al., “Experimental quantum cryptography,” Journal of Cryptology, vol. 5, pp. 3-28, 1992.

[15] D. Gottesman, Hoi-Kwong Lo, N. Lutkenhaus, and J. Preskill, "Security of quantum key distribution with imperfect devices," Quantum Information & Computation, vol. 4, no. 5, pp. 325-360, Sep 2004.

[16] H.-K. Lo, X. Ma, K. Chen, “Decoy State Quantum Key Distribution,” Physical Review Letters, vol. 94, no. 23, pp. 230504, Jun 2005.

[17] H.-K. Lo, M. Curty, and B. Qi, “Measurement-device-independent quantum key distribution,” Physical Review Letters, vol. 108, 130503, Mar 2012.

[18] Z. Tang, Z. Liao, F. Xu, et al., “Experimental demonstration of polarization encoding measurement-device-independent quantum key distribution,” Physical Review Letters, vol. 112, 190503, May 2014.

[19] J. Qiu, “Quantum communications leap out of the lab,” Nature, vol. 508, pp. 441-442, April 2014.

[20] M. Peev, "Why do I believe that Quantum Key Distribution (QKD) is Finally About to Reach Telecom Markets and Grow Out of Its Present Exotic Standing?," Optical Fiber Communications Conference (OFC) 2019, paper W4D.3.

[21] C. Elliott, “Building the quantum network,” New Journal of Physics, vol. 4, 46.1-46.12, Jan 2002.

[22] C. Elliott, A. Colvin, D. Pearson, et al., “Current status of the DARPA quantum network,” Proc. SPIE 5815, Quantum Information and Computation III, May 2005.

[23] P. Eraerds, N. Walenta, M. Legré, et al., “Quantum key distribution and 1 Gbps data encryption over a single fibre,” New Journal of Physics, vol. 12, 063027, June 2010.

[24] D. Stucki, M. Legré, F. Buntschu, et al., “Long-term performance of the SwissQuantum quantum key distribution network in a field environment,” New Journal of Physics, vol. 13, 123001, December 2011.

[25] A. Poppe, M. Peev and O. Maurhart, “Outline of the SECOQC quantum-key-distribution network,” International Journal of Quantum Information, vol. 6, no. 2, pp. 209-218, 2008.

[26] M. Peev, T. Länger, T. Lorünser, et al., “The SECOQC Quantum-Key-Distribution Network in Vienna,” Optical Fiber Communication Conference 2009, paper OThL2.

[27] M. Peev, A. Poppe, O. Maurhart, et al., “The SECOQC Quantum Key Distribution Network in Vienna,” 35th European Conference on Optical Communication, Vienna, Austria, 2009, paper 1.4.1.

[28] M. Peev, C. Pacher, R. Alléaume, et al., “The SECOQC quantum key distribution network in Vienna,” New Journal of Physics, vol. 11, 075001, July 2009.

[29] M. Sasaki, M. Fujiwara, H. Ishizuka, et al., “Field test of quantum key distribution in the Tokyo QKD Network,” Optics Express, vol. 19, no. 11, pp. 10387-10409, 2011.

[30] J. F. Dynes, A. Wonfor, W. W. -S. Tam, et al., “Cambridge quantum network,” Nature Partner Journals (NPJ) Quantum Information, vol. 5, article number 101, 2019.

[31] Q. Zhang, F. Xu, Y.-A. Chen, et al., “Large scale quantum key distribution: challenges and solutions,” Optics Express, vol. 26, no. 18, pp. 24260-24273, 2018.

[32] Y. Liu, T.-Y. Chen, L.-J. Wang, et al., “Experimental measurement-device-independent quantum key distribution,” Physical Review Letters, vol. 111, 130502, Sep 2013.

[33] Y.-L. Tang, H.-L. Yin, S.-J. Chen, et al., “Measurement-device-independent quantum key distribution over 200 km,” Physical Review Letters, vol. 113, 190501, Nov 2014.

[34] Y.-L. Tang, H.-L. Yin, S.-J. Chen, et al., “Field test of measurement-device-independent quantum key distribution," IEEE Journal of Selected Topics in Quantum Electronics, vol. 21, no. 3, pp. 116-122, May-June 2015, Art no. 6600407.

[35] Y.-L. Tang, H.-L. Yin, Q. Zhao, et al., “Measurement-device-independent quantum key distribution over untrustful metropolitan network,” Physics Review X, vol. 6, 011024, Mar 2016.

[36] X.-B. Wang, “Three-intensity decoy-state method for device-independent quantum key distribution with basis-dependent errors,” Physical Review A, vol. 87, 012320, January 2013.

[37] Y.-H. Zhou, Z.-W. Yu, and X.-B. Wang, “Making the decoy-state measurement-device-independent quantum key distribution practically useful,” Physical Review A, vol. 93, 042324, April 2016.

[38] H.-L. Yin, T.-Y. Chen, Z.-W. Yu, et al., “Measurement-Device-Independent Quantum Key Distribution Over a 404 km Optical Fiber,” Physical Review Letters, vol. 117, 190501, November 2016.

[39] M. Lucamarini, Z. L. Yuan, J. F. Dynes, et al., “Overcoming the rate-distance limit of quantum key distribution without quantum repeaters,” Nature, vol. 557, pp. 400-403, 2018.

[40] X. Ma, P. Zeng, and H. Zhou, “Phase-Matching Quantum Key Distribution,” Physical Review X, vol. 8, 031043, 2018.

[41] X. T. Fang, P. Zeng, H. Liu, et al., “Implementation of quantum key distribution surpassing the linear rate-transmittance bound,” Nature Photonics, vol. 14, pp. 422-425, 2020.

[42] X.-B. Wang, Z.-W. Yu, and X.-L. Hu, “Twin-field quantum key distribution with large misalignment error,” Physical Review A, vol. 98, 062323, December 2018.

[43] J.-P. Chen, C. Zhang, Y. Liu, et al., “Sending-or-Not-Sending with Independent Lasers: Secure Twin-Field Quantum Key Distribution over 509 km,” Physical Review Letters, vol. 124, 070501, February 2020.

[44] J. P. Chen, C. Zhang, Y. Liu, et al., “Twin-field quantum key distribution over a 511 km optical fibre linking two distant metropolitan areas,” Nature Photonics, vol. 15, pp. 570–575, 2021.

[45] H. Liu, C. Jiang, H.-T. Zhu, et al., “Field Test of Twin-Field Quantum Key Distribution through Sending-or-Not-Sending over 428 km,” Physical Review Letters, vol. 126, 250502, June 2021.

[46] M. Pittaluga, M. Minder, M. Lucamarini, et al., “600-km repeater-like quantum communications with dual-band stabilization,” Nature Photonics, vol. 15, pp. 530-535, 2021.

[47] P. D. Townsend, "Simultaneous quantum cryptographic key distribution and conventional data transmission over installed fibre using wavelength-division multiplexing," Electronics Letters, vol. 33, no. 3, pp. 188-190, 1997.

[48] B. Qi, W. Zhu, L. Qian, and H.-K. Lo, “Feasibility of quantum key distribution through a dense wavelength division multiplexing network,” New Journal of Physics, vol. 12, 103042, 2010.

[49] M. S. Goodman, P. Toliver, R. J. Runser, et al., “Quantum cryptography for optical networks: a systems perspective,” The 16th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS) 2003, paper ThEE1, vol. 2, pp. 1040-1041.

[50] N. A. Peters, P. Toliver, T. E. Chapuran, et al., “Dense wavelength multiplexing of 1550 nm QKD with strong classical channels in reconfigurable networking environments,” New Journal of Physics, vol. 11, 045012, April 2009.

[51] T. E. Chapuran, P. Toliver, N. A. Peters, et al., “Optical networking for quantum key distribution and quantum communications,” New Journal of Physics, vol. 11, 105001, October 2009.

[52] N. A. Peters, P. Toliver, T. E. Chapuran, et al., “Quantum communications in reconfigurable optical networks: DWDM QKD through a ROADM,” Conference on Optical Fiber Communication (OFC) 2010, paper OTuK1.

[53] L.-J. Wang, K.-H. Zou, W. Sun, et al., “Long-distance copropagation of quantum key distribution and terabit classical optical data channels,” Physics Review A, vol. 95, no. 1, pp. 012301, 2017.

[54] Y. Mao, B.-X. Wang, C. Zhao, et al., “Integrating quantum key distribution with classical communications in backbone fiber network,” Optics Express, vol. 26, no. 5, pp. 6010-6020, 2018.

[55] W. Chen, Z. Han, T. Zhang, et al., “Field experiment on a “star type” metropolitan quantum key distribution network," in IEEE Photonics Technology Letters, vol. 21, no. 9, pp. 575-577, May 2009.

[56] S. Wang, W. Chen, Z. Yin, et al., “Field test of wavelength-saving quantum key distribution network,” Optics Letters, vol. 35, no. 14, pp. 2454-2456, July 2010.

[57] S. Wang, W. Chen, Z. Yin, et al., “Field and long-term demonstration of a wide area quantum key distribution network,” Optics Express, vol. 22, no. 18, pp. 21739-21756, September 2014.

[58] T.-Y. Chen, J. Wang, H. Liang, et al., “Metropolitan all-pass and inter-city quantum communication network,” Optics Express, vol. 18, no. 26, pp. 27217-27225, 2010.

[59] K. A. Patel, J. F. Dynes, I. Choi, et al., “Coexistence of high-bit-rate quantum key distribution and data on optical fiber,” Physical Review X, vol. 2, no. 4, 041010, November 2012.

[60] B. Fröhlich, J. F. Dynes, M. Lucamarini, et al., “A quantum access network,” Nature, vol. 501, pp. 69-72, 2013.

[61] K. A. Patel, J. F. Dynes, M. Lucamarini, et al., “Quantum key distribution for 10 Gb/s dense wavelength division multiplexing networks,” Applied Physics Letters, vol. 104, no. 5, 051123, 2014.

[62] I. Choi, Y. Zhou, J. F. Dynes, et al., “Field trial of a quantum secured 10Gb/s DWDM transmission system over a single installed fiber,” Optics Express, vol. 22, no. 19, pp. 23121-23128, 2014.

[63] B. Fröhlich, J. F. Dynes, M. Lucamarini, et al., “Quantum secured gigabit optical access networks,” Scientific Reports, vol. 5, article number 18121, 2015.

[64] J. F. Dynes, W. W-S. Tam, A. Plews, et al., “Ultra-high bandwidth quantum secured data transmission,” Scientific Reports, vol. 6, article number 35149, 2016.

[65] L.-J. Wang, L.-K. Chen, L. Ju, et al., “Experimental multiplexing of quantum key distribution with classical optical communication,” Applied Physics Letters, vol. 106, no. 8, 081108, 2015.

[66] R. Bedington, J. M. Arrazola, A. Ling, “Progress in satellite quantum key distribution,” Nature Partner Journals (NPJ) Quantum Information, vol. 3, article number 30, 2017.

[67] I. Khan, B. Heim, A. Neuzner and C. Marquardt, “Satellite-Based QKD,” Optics and Photonics News, Feb 2018.

[68] J. G. Rarity, P. R. Tapster, P. M. Gorman, and P. Knight, “Ground to satellite secure key exchange using quantum cryptography,” New Journal of Physics, vol. 4, 82.1-82.21, 2002.

[69] C. Bonato, A. Tomaello, V. D. Deppo, et al., “Feasibility of satellite quantum key distribution,” New Journal of Physics, vol. 11, 045017, 2009.

[70] A. Tomaello, A. Dall'Arche, G. Naletto, and P. Villoresi, “Intersatellite quantum communication feasibility study", Proc. SPIE 8163, Quantum Communications and Quantum Imaging IX, 816309, September 2011.

[71] B. C. Jacobs and J. D. Franson, "Quantum cryptography in free space," Optics Letters, vol. 21, no. 22, pp. 1854-1856, 1996.

[72] W. T. Buttler, R. J. Hughes, P. G. Kwiat, et al., “Free-space quantum key distribution,” Physical Review A, vol. 57, no. 4, pp. 2379-2382, April 1998.

[73] W. T. Buttler, R. J. Hughes, P. G. Kwiat, et al., “Practical Free-Space Quantum Key Distribution over 1 km,” Physical Review Letters, vol. 81, no. 15, pp. 3283-3286, October 1998.

[74] R. J. Hughes, J. E. Nordholt, D. Derkacs, et al., “Practical free-space quantum key distribution over 10 km in daylight and at night,” New Journal of Physics, vol. 4, 43.1-43.14, 2002.

[75] C. Kurtsiefer, P. Zarda, M. Halder, et al., “Quantum cryptography: a step towards global key distribution,” Nature vol. 419, pp. 450, 2002.

[76] T. Schmitt-Manderbach, H. Weier, M. Fürst, et al., “Experimental Demonstration of Free-Space Decoy-State Quantum Key Distribution over 144 km,” Physical Review Letters, vol. 98, 010504, January 2007

[77] S. Nauerth, F. Moll, M. Rau, et al., “Air-to-ground quantum communication,” Nature Photonics, vol. 7, pp. 382-386, 2013.

[78] J. Y. Wang, B. Yang, S. K. Liao, et al., “Direct and full-scale experimental verifications towards ground-satellite quantum key distribution,” Nature Photonics, vol. 7, pp. 387-393, 2013.

[79] J.-P. Bourgoin, B. L. Higgins, N. Gigov, et al., “Free-space quantum key distribution to a moving receiver,” Optics Express, vol. 23, no. 26, pp. 33437-33447, 2015.

[80] C. J. Pugh, S. Kaiser, J.-P. Bourgoin, et al., "Airborne demonstration of a quantum key distribution receiver payload," Conference on Lasers and Electro-Optics Europe & European Quantum Electronics Conference (CLEO/Europe-EQEC), 2017.

[81] C. J. Pugh, S. Kaiser, J.-P. Bourgoin, et al., “Airborne demonstration of a quantum key distribution receiver payload,” Quantum Science and Technology, vol. 2, no. 2, 024009, June 2017.

[82] J. Yin, Y. Cao, S.-B. Liu, et al., "Experimental quasi-single-photon transmission from satellite to earth," Optics Express, vol. 21, no. 17, pp. 20032-20040, 2013.

[83] J. Pan, “Quantum science satellite,” Chinese Journal of Space Science, vol. 34, no. 5, pp. 547-549, 2014.

[84] S. Liao, W. Cai, W. Liu, et al., “Satellite-to-ground quantum key distribution,” Nature, vol. 549, pp. 43-47, 2017.

[85] S. Liao, W. Cai, J. Handsteiner, et al., “Satellite-Relayed Intercontinental Quantum Network,” Physical Review Letters, vol. 120, 030501, 2018.

[86] T. Scheidl, J. Handsteiner, D. Rauch, R. Ursin, “Space-to-ground quantum key distribution,” Proceedings vol. 11180, International Conference on Space Optics (ICSO) 2018, Chania, Greece.

[87] T. Jennewein, J. P. Bourgoin, B. Higgins, et al., “QEYSSAT: a mission proposal for a quantum receiver in space,” Proc. SPIE 8997, Advances in Photonics of Quantum Computing, Memory, and Communication VII, 89970A, February 2014.

[88] E. Meyer-Scott, Z. Yan, A. MacDonald, et al., “How to implement decoy-state quantum key distribution for a satellite uplink with 50-dB channel loss,” Physical Review A, vol. 84, 062326, December 2011.

[89] J.-P. Bourgoin, E. Meyer-Scott, B. L. Higgins, et al., “A comprehensive design and performance analysis of low Earth orbit satellite quantum communication,” New Journal of Physics, vol. 15, 023006, February 2013.

[90] J.-P. Bourgoin, N. Gigov, B. L. Higgins, et al., “Experimental quantum key distribution with simulated ground-to-satellite photon losses and processing limitations,” Physical Review A, vol. 92, 052339, November 2015.

[91] H. Podmore, I. D'Souza, D. Hudson, et al., “Optical Terminal for Canada's Quantum Encryption and Science Satellite (QEYSSat),” IEEE International Conference on Space Optical Systems and Applications (ICSOS) 2019.

[92] M. Yang, F. Xu, J.-G. Ren, et al., "Spaceborne, low-noise, single-photon detection for satellite-based quantum communications," Optics Express, vol. 27, pp. 36114-36128, 2019.

[93] G. Vallone, D. Bacco, D. Dequal, et al., "Experimental Satellite Quantum Communications," Physical Review Letters, vol. 115, 040502, July 2015.

[94] D. Dequal, G. Vallone, D. Bacco, et al., "Experimental single-photon exchange along a space link of 7000 km," Physical Review A, vol. 93, 010301, January 2016.

[95] K. Günthner, I. Khan, D. Elser, et al., "Quantum-limited measurements of optical signals from a geostationary satellite," Optica, vol. 4, pp. 611-616, 2017.

[96] Y. A. Chen, Q. Zhang, T. Y. Chen, et al., "An integrated space-to-ground quantum communication network over 4,600 kilometres," Nature, vol. 589, pp. 214-219, 2021.

[97] T. Jennewein, C. Grant, E. Choi, et al., "The NanoQEY mission: ground to space quantum key and entanglement distribution using a nanosatellite," Proc. SPIE 9254, Emerging Technologies in Security and Defence II; and Quantum-Physics-based Information Security III, 925402, October 2014.

[98] D. K. L. Oi, A. Ling, J. A. Grieve, et al., "Nanosatellites for quantum science and technology," Contemporary Physics, pp. 25-52, 2016.

[99] R. Bedington, X. Bai, E. Truong-Cao, et al., "Nanosatellite experiments to enable future space-based QKD missions," EPJ Quantum Technology, vol. 3, article 12, 2016.

[100] D. K. L. Oi, A. Ling, G. Vallone, et al., "CubeSat quantum communications mission," EPJ Quantum Technology, vol. 4, article 6, 2017.

[101] H. Takenaka, A. Carrasco-Casado, M. Fujiwara, et al., "Satellite-to-ground quantum-limited communication using a 50-kg-class microsatellite," Nature Photonics, vol. 11, pp. 502-508, 2017.

[102] K. Boone, J.-P. Bourgoin, E. Meyer-Scott, et al., "Entanglement over global distances via quantum repeaters with satellite links," Physical Review A, vol. 91, 052325, May 2015.

[103] Z. Tang, R. Chandrasekara, Y. C. Tan, et al., "Generation and Analysis of Correlated Pairs of Photons aboard a Nanosatellite," Physical Review Applied, vol. 5, 054022, May 2016.

[104] J. Yin, Y. Cao, Y.-H. Li, et al., "Satellite-based entanglement distribution over 1200 kilometers," Science, vol. 356, no. 6343, pp. 1140-1144, 2017.

[105] A. Villar, A. Lohrmann, X. Bai, et al., "Entanglement demonstration onboard a nano-satellite," Optica, vol. 7, no. 7, pp. 734-737, 2020.

[106] C.-Z. Peng, T. Yang, X.-H. Bao, et al., "Experimental Free-Space Distribution of Entangled Photon Pairs Over 13 km: Towards Satellite-Based Global Quantum Communication," Physical Review Letters, vol. 94, 150501, April 2005.

[107] R. Ursin, F. Tiefenbacher, T. Schmitt-Manderbach, et al., "Entanglement-based quantum communication over 144 km," Nature Physics, vol. 3, pp. 481-486, 2007.

[108] X.-M. Jin, J.-G. Ren, B. Yang, et al., "Experimental free-space quantum teleportation," Nature Photonics, vol. 4, pp. 376-381, 2010.

[109] J. Yin, J.-G. Ren, H. Lu, et al., "Quantum teleportation and entanglement distribution over 100-kilometre free-space channels," Nature, vol. 488, pp. 185-188, 2012.

[110] Y. Cao, Y.-H. Li, K.-X. Yang, et al., "Long-Distance Free-Space Measurement-Device-Independent Quantum Key Distribution," Physical Review Letters, vol. 125, 260503, December 2020.

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Enhanced Digital Signature Service Using Split-key Signatures

A Technical Paper prepared for SCTE by

Nicol So, Sr. Staff Systems Engineer, CommScope, SCTE Member
101 Tournament Dr,
Horsham, PA 19044
nicol.so@commscope.com
+1 215 323 1149

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

List of Tables

1. Introduction

To prevent unauthorized software from being introduced into a user's computers or embedded devices, there are many existing standards that require executable images to include a digital signature. The device or OS platform validates a code image before it is executed or, in other cases, before it is downloaded and installed on the device. Some well-known standards of signed code images include Android APK [1], Microsoft Authenticode [2] and DOCSIS 3.1 cable modem software image signatures [3]. There are many more standards-based and proprietary code signing formats, some of which are verified in hardware (e.g., in the boot ROM).

A common method of code signing is to use a command-line tool together with a private key file. The process may be initiated by a developer who builds the software or by an automated software build workflow. For example, Android APK can be signed with Android Studio and binaries for Microsoft platforms can be signed with SignTool.exe. A chip vendor that supports boot code signature verification would typically provide a device manufacturer with their own command-line tool for signing boot code and it may be combined with their linker.

However, there are security concerns with this approach. Code signing keys may be stolen from development environments. Some well-known incidents involving compromised code signing private keys include those detailed in [8], [9], and [10]. For this reason, it is good practice to protect code signing keys in hardware security modules (HSMs), which provide strong protection against disclosure of the keys. Multiple companies offer commercial code signing tools and online services with hardware-based protection of private keys.

While protecting the confidentiality of code signing keys addresses one aspect of security, it is also important to ensure that only authorized individuals can exercise the protected signing key to sign code, and can do so only according to policy. A code signing service should support permissions and access control management so that an administrator with appropriate privileges can authorize specific individuals or teams of individuals to use specific signing keys based on their areas of responsibility.

A software development organization that uses a third-party code signing service is, in doing so, trusting the signing service not to sign any software using the organization's signing key without authorization. Note that the signing service has the technical ability to sign any code using its subscribers' signing keys; it is the security controls employed by the service that prevent unauthorized code signing from happening. Code signing services can themselves be targets of, and can even fall prey to, cyberattacks. This concern is validated by recent news of high-profile incidents in which providers of security technologies had themselves become victims of cyberattacks, adversely affecting users of their security products [11], [12], [13].

This paper introduces a code signing service architecture in which a subscriber's code signing key is split into shares controlled separately by the code signing service and the subscriber. If the subscriber's share is compromised, the code signing service will continue to protect the other half of the signing key and will prevent unauthorized code to be signed. At the same time, the code signing service is prevented from signing code in the subscriber's name without the subscriber's participation. If the code signing service becomes compromised in a cyberattack, the attacker will not have possession of the subscriber's share of a code signing key and will still be unable to sign unauthorized software releases.

2. Solution Outline

We will discuss applying split-key digital signature to enhance the security of code signing using an example system. In our example system, a trusted code signing service (TCSS) operates a secure infrastructure and performs code signing for its subscribers. A subscriber in this context may be an organization, but it could also be an individual. A subscriber owns a public-private key pair, which is used to sign and authenticate code (data objects) published by the subscriber. A subscriber has one or more authorized users, who are trusted to exercise the subscriber's private code signing key to produce code signed by the subscriber. A trusted third party (TTP) is a party trusted by a subscriber to generate a code signing key pair for the subscriber. Such trust may be established by technical and non-technical means, such as contractual guarantees, certification, and audits. A TTP should be viewed as a role—trusted key generation may not be the only service it provides.

Figure 1 illustrates the process of generating a split signing key in our example system.

In the description that follows, we use the notation $E(K, m)$ to denote a data object m encrypted using a key K. Cryptographic keys are denoted using symbols of the form K_{XYZ}, where the subscript XYZ is meant to be suggestive of the ownership or the purpose of the key. Where there is no confusion, we do not explicitly state which key of a key pair is used in an operation. For example, if K is a public-private key pair, it is obvious from the context that it is the public key of K that is used in computing $\mathrm{E(K, m)}$. When discussing the two shares of a split signing key (more specifically the split private key of the signing key), we will use the convention that subscript 1 is associated with an authorized user, whereas subscript 2 is associated with the TCSS. Where it is necessary to distinguish between the public and private keys of a key pair K, we will use K^{PUB} and K^{PRIV} to refer to the public and private keys respectively.

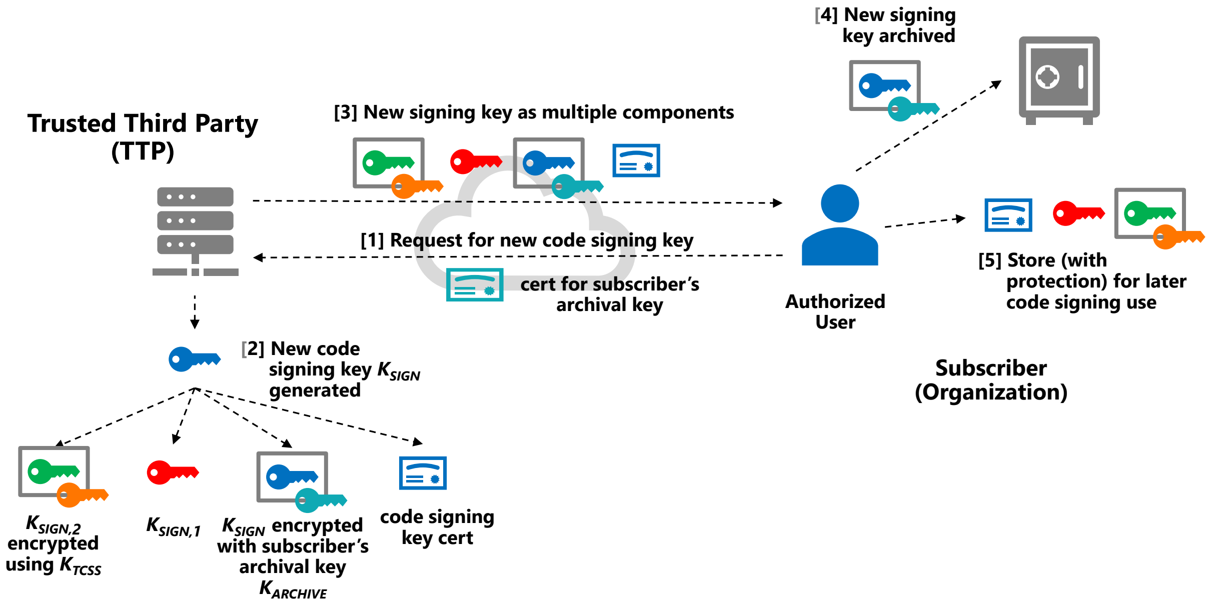

[4] New signing key archived
Trusted Third Party (TTP)
[3] New signing key as multiple components
[1] Request for new code signing key
cert for subscriber's archival key
Authorized User
[5] Store (with protection) for later code signing use
Subscriber (Organization)
[2] New code signing key K_{SIGN} generated
$K_{SIGN,2}$ encrypted using K_{TCSS}
$K_{SIGN,1}$
K_{SIGN} encrypted with subscriber's archival key $K_{ARCHIVE}$
code signing key cert

Figure 1 – Generation of a split signing key

In the key generation process shown in Figure 1, a TTP generates a code signing key K_{SIGN} and then splits it into 2 shares, namely $K_{SIGN,1}$ and $K_{SIGN,2}$. $K_{SIGN,2}$ is encrypted with the (public) key K_{TCSS} of the

TCSS. K_{SIGN} is encrypted with the (public) key $K_{ARCHIVE}$ of the subscriber organization for secure offline archival. The creation of this encrypted archival copy is optional, but some subscribers may find such archiving desirable. The (complete) signing key K_{SIGN} is normally kept offline and not available for code signing, to reduce the security risk of disclosure. $K_{SIGN,1}$, $E\left(K_{TCSS}, K_{SIGN,2}\right)$ and $E(K_{ARCHIVE}, K_{SIGN})$ are returned to the client used by the authorized user.

For clarity, the keys shown in Figure 1 are color-coded as follows: the complete signing key K_{SIGN} is in blue, the share $K_{SIGN,1}$ is in red, the share $K_{SIGN,2}$ is in green, and K_{TCSS} is in orange.

The TTP also provides the public key K_{SIGN}^{PUB} of the key K_{SIGN}. A code signature that is generated with K_{SIGN} may be validated by anyone using the public key K_{SIGN}^{PUB}. In the figure, $\mathrm{K}_{\mathrm{SIGN}}^{\mathrm{PUB}}$ is provided in the form of a digital certificate, as is common practice, although that is not always necessary. It is assumed that communications between TTP and the client are encrypted and authenticated, for example using HTTPS.

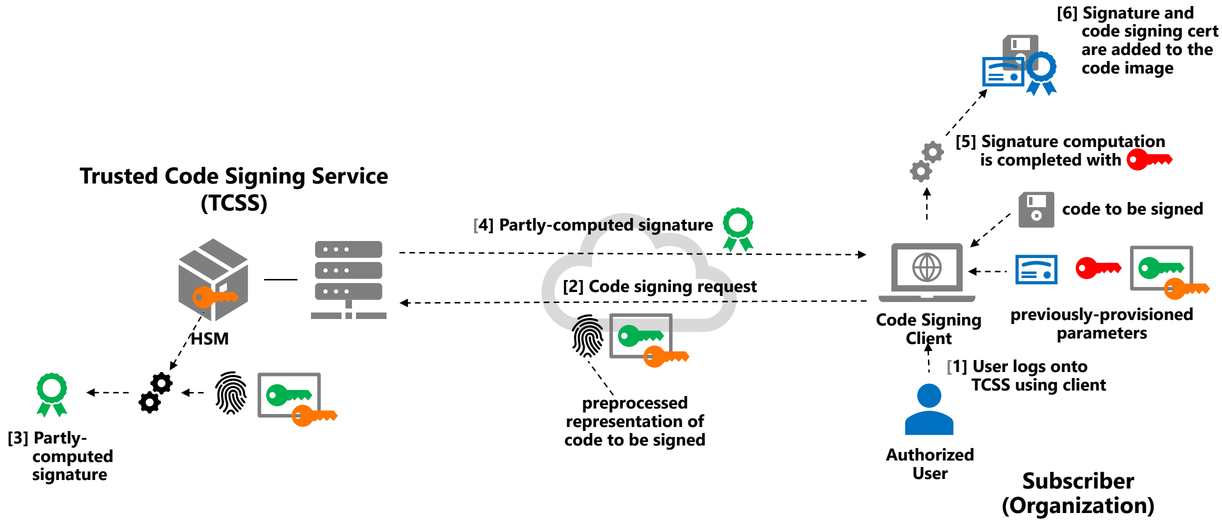

[6] Signature and code signing cert are added to the code image
[5] Signature computation is completed with
code to be signed
Trusted Code Signing Service (TCSS)
[4] Partly-computed signature
[2] Code signing request
Code Signing Client
previously-provisioned parameters
HSM
[1] User logs onto TCSS using client
[3] Partly-computed signature
preprocessed representation of code to be signed
Authorized User
Subscriber (Organization)

Figure 2 – Overview of two-step code signing

Once a client is provisioned with all of the above, the authorized user is able to use an established TCSS account to initiate code signing. Figure 2 illustrates the following code signing sequence:

1) The authorized user logs into the TCSS using a client application and establishes a secure authenticated session with the TCSS. This can be done in many different ways and implementation details are outside the scope of this paper.
2) Using the client, the authorized user submits to the TCSS a code signing request consisting of a preprocessed representation of the to-be-signed software image and $E\left(K_{TCSS}, K_{SIGN,2}\right)$, which was obtained from the TTP before. The preprocessed representation here may contain a cryptographic digest of the software image, possibly with added padding and randomness.
3) The TCSS unwraps the encrypted $K_{SIGN,2}$ inside an HSM. In other words, $E\left(K_{TCSS}, K_{SIGN,2}\right)$ is provided to the HSM, which decrypts it with the private key K_{TCSS}^{PRIV} to recover and use $K_{SIGN,2}$ inside the HSM's security perimeter. $K_{SIGN,2}$ is then used to generate a partly-computed code signature.

4) The partly-computed code signature generated using $K_{SIGN,2}$ is returned to the client application.
5) The client application is now able to compute a full code signature using this partly-computed signature and $K_{SIGN,1}$. Details of the underlying mathematics are described in the next section.
6) Finally, the client application combines the generated code signature, the corresponding code verification certificate, and the code image to create the final signed code image.

An advantage of this technique is that a code signing client can use a secure signing service (the TCSS) that uses $K_{SIGN,2}$ only inside an HSM. The TCSS can provide other security advantages, including fine-grained access control to different code signing keys and secure audit logging.

At the same time, the subscriber maintains control of its half of the signing key, $K_{SIGN,1}$, which is required to compute a full code signature. This way, TCSS does not have to be totally trusted – it is not able to sign code in the subscriber's name (i.e. using the K_{SIGN}) without the client's involvement, since the TCSS has access to neither the complete private key K_{SIGN} nor $K_{SIGN,1}$.

The following section provides mathematical details for split-key code signing based on the RSA cryptosystem.

3. Detailed Explanation for RSA-Based Split-Key Signing

3.1. A Quick Review of the Basic RSA Algorithm

RSA-based algorithms are commonly used for public-key encryption and digital signatures. Practical RSA algorithms (see [7], for example) are based on, but not identical to the "textbook" RSA algorithm. Because of security considerations, practical RSA algorithms have additional pre- and post-processing steps. The core of these RSA-based algorithms is the modular exponentiation operation

$$\mathrm{RSA}(\mathrm{m}, \mathrm{e}, \mathrm{n}) = \mathrm{m}^{\mathrm{e}} \bmod \mathrm{n}, \tag{1}$$

where m is the input to be processed, n is a modulus, and e is an exponent. The same operation is used both in the generation and verification of RSA signatures. In RSA key generation, a public-private key pair is generated together by the key owner, or a party or mechanism trusted by the key owner. An RSA private key (d, n) consists of a modulus $n = p\,q$, which is a product of two large primes p and q, and an exponent d. The factorization of n (i.e., the values of p and q) is kept secret by the key owner. The public key corresponding to (d, n) is (e, n) where e is an exponent such that $(\mathrm{m}^{\mathrm{e}})^{\mathrm{d}} \equiv \mathrm{m} \pmod{\mathrm{n}}$. More relevantly computation-wise, d and e satisfy

$$\mathrm{e} \cdot \mathrm{d} \equiv 1 \left(\bmod\ \phi(n)\right), \tag{2}$$

where $\phi(n)$ is Euler's totient function. In RSA, where $n = pq$, $\phi(\mathrm{n}) = (\mathrm{p} - 1)(q - 1)$. In order for (2) to be satisfied, both e and d must be coprime with $\phi(n)$. The exponents e and d form a matched pair. Note that either e or d can be decided first, with the other derived using equation (2).

In RSA signature schemes, the private key of a key pair is the signing key. The public key is the verification key.

In practical RSA-based signature algorithms, the to-be-signed data object undergoes some preprocessing before the RSA operation is applied. The preprocessing typically includes reducing the data object to a digest using a cryptographic hash function and combining the digest with deterministic and random padding. Similarly, in signature verification, postprocessing is performed after the RSA operation. For the purpose of split-key RSA operations, these pre- and postprocessing steps can be ignored, as they do not involve the RSA keys, private or public.

3.2. Additive Splitting

In additive splitting of an RSA key, the private exponent d used in the signing operation is split into two shares d_1 and d_2, where

$$\mathrm{d}_1 + \mathrm{d}_2 \equiv d \left(\mathrm{mod}\ \phi(n)\right). \tag{3}$$

With the split, the RSA operation m^d mod n can be computed as $\left(m^{d_1} \bmod n\right)\left(m^{d_2} \bmod n\right) \bmod\ n$. This computation can be carried out in two parts: one by a party with knowledge of d_1 and another by a party with knowledge of d_2.

3.2.1. Key Generation

To make RSA signing a split-key operation, different shares of the private exponent need to be held by different entities, so that compromising one of the entities will not give an attacker the ability to forge signatures. A possible exception to this arrangement is when the private key is archived in heavily protected and controlled offline storage.

There are different ways in which key generation can be accomplished. If a TTP is available, the TTP can generate the signing key in two shares, securely deliver the shares to a TCSS and the subscriber organization's authorized user, and securely erase its copy afterward. Alternatively, the TCSS can generate the signing key and securely deliver one share to the subscriber organization's authorized user. As in the case of key generation by a TTP, the TCSS securely erases the share of the private key for which it is not a custodian.

Yet another possibility for key generation is for the subscriber organization to generate the key pair, split the private key into two shares, upload one share to the TCSS and give custody of the other to an authorized user. All copies of the complete private key are securely erased afterward.

If the subscriber organization desires to keep a complete copy of the private key in archive, that can be achieved by encrypting a copy of the private key during key generation using an archive key belonging to the subscriber organization.

3.2.2. Key Splitting Operation

To split a private exponent d into d_1 and d_2 additively, d_1 may be chosen uniformly at random from $\{2, \ldots, \phi(n) – 2\}$. d_2 can then be determined based on equation (3). Note that d_2 as the additive inverse of d_1 modulo $\phi(n)$ always exists. Note also that because d is odd but $\phi(\mathrm{n}) = (\mathrm{p} - 1)(\mathrm{q} - 1)$ is even, one of d_1 or d_2 is even, and therefore not a valid exponent for normal RSA operation. However, that is not a problem because (d_1, n) and (d_2, n) don't need to function like normal RSA private keys.

After the private key is split, one share, (d_1, n), is given to the authorized signer. Another share, (d_2, n), is placed under the control of the TCSS. The TCSS can maintain a database of the private key shares entrusted to it by its subscribers. Alternatively, d_2 can be encrypted using a key owned by the TCSS. The encrypted d_2 is given to the authorized signer along with (d_1, n). When the TCSS needs to perform an operation with d_2 in response to a signing request from the authorized signer, the encrypted d_2 can be included in the request.

3.2.3. Signing Operation

To sign a data object m', the authorized signer first performs preprocessing on the data to produce m, a preprocessed representation of m', and sends it to the TCSS.

The TCSS computes a partly-computed signature by performing the RSA operation on the preprocessed representation using its share of the subscriber's signing key

$$\mathrm{s}' = \mathrm{m}^{\mathrm{d}_2} \bmod n, \tag{4}$$

which the TCSS then sends to the authorized signer.

The authorized signer computes the final signature s as

$$\mathrm{s} = \mathrm{s}' \cdot \left(\mathrm{m}^{\mathrm{d}_1} \bmod \mathrm{n}\right) \bmod \mathrm{n}. \tag{5}$$

The authorized signer can verify that the signature is valid by checking that

$$\mathrm{s}^{\mathrm{e}} \bmod \mathrm{n} = \mathrm{m}. \tag{6}$$

Variation: As a variation, the RSA operations in the procedure above can be performed in the opposite order. In that case, the authorized signer computes a partly-computed signature on the preprocessed representation m as

$$\mathrm{s}' = \mathrm{m}^{\mathrm{d}_1} \bmod \mathrm{n}. \tag{7}$$

The partly-computed signature s', together with m, is then sent to the TCSS, which computes the final signature s as

$$\mathrm{s} = \mathrm{s}' \cdot \left(\mathrm{m}^{\mathrm{d}_2} \bmod \mathrm{n}\right) \bmod \mathrm{n}. \tag{8}$$

The final signature s is sent to the authorized signer. To catch unexpected computation errors, the final signature can be verified by the TCSS or the authorized signer, or both.

3.3. Multiplicative Splitting

In multiplicative splitting of an RSA key, the private exponent d used in the signing operation is split into two shares d_1 and d_2, where

$$\mathrm{d}_1 \cdot \mathrm{d}_2 \equiv \mathrm{d} \left(\mathrm{mod}\ \phi(\mathrm{n})\right). \tag{9}$$

With the split, the RSA operation $m^d \bmod n$ can be computed as $\left(\mathrm{m}^{\mathrm{d}_1} \bmod \mathrm{n}\right)^{\mathrm{d}_2} \bmod \mathrm{n}$ or $\left(\mathrm{m}^{\mathrm{d}_2} \bmod \mathrm{n}\right)^{\mathrm{d}_1} \bmod \mathrm{n}$.

3.3.1. Key Splitting Operation

One way to find a pair of d_1 and d_2 that satisfy equation (9) is to choose a random d_1 from among $\{3, \ldots, \phi(\mathrm{n}) - 1\}$ that is coprime with $\phi(n)$. In that case, exactly one solution exists for d_2, which can be computed using the extended Euclidean algorithm.

Using this strategy, sometimes it may take more than one try to find a d_1 that is coprime with $\phi(n)$. If the factors p and q are chosen to be safe primes (i.e. $p = 2p' + 1$ and $\mathrm{q} = 2\mathrm{q}' + 1$, for some primes p' and q'), then choosing d_1 at random is almost guaranteed to succeed on the first try.

3.3.2. Signing Operation

As in the case of additive splitting, to sign a data object m', the authorized signer first preprocesses m' to produce a preprocessed representation m and sends it to the TCSS.

The TCSS computes a partly-computed signature s' by performing the RSA operation on the preprocessed representation using its share of the subscriber's signing key d_2

$$\mathrm{s}' = m^{d_2} \bmod n, \tag{10}$$

which the TCSS then sends to the authorized signer.

The authorized signer computes the final signature s as

$$\mathrm{s} = (s')^{d_1} \bmod \mathrm{n}. \tag{11}$$

The authorized signer can verify that the signature is valid by checking that

$$\mathrm{s}^{\mathrm{e}} \bmod \mathrm{n} = \mathrm{m}. \tag{12}$$

Variation: As a variation, the RSA operations in the procedure above can be performed in the opposite order. In that case, the authorized signer computes a partly-computed signature on the preprocessed representation m as

$$\mathrm{s}' = m^{d_1} \bmod n. \tag{13}$$

The partly-computed signature s' is then sent to the TCSS, which computes the final signature s as

$$\mathrm{s} = (s')^{d_2} \bmod \mathrm{n}. \tag{14}$$

The final signature s is sent to the authorized signer. To catch unexpected computation errors, the final signature can be verified by the authorized signer.

4. Experimental Results

We performed some experiments to compare standard RSA signature performance (with optimization based on the Chinese remainder theorem (CRT)) against the performance of a two-step digital signature approach with a private key that has been split either additively or multiplicatively, as described earlier. The experiments were executed in the following computing environment:

Table 1 – Benchmark Computing Environment

Operating system	Windows 10 64-bit
CPU	Intel® Core™ I7-10610U
RAM	32 GBytes
Cryptographic library	C# BouncyCastle

As expected, the overhead of key splitting is insignificant when compared to the time it takes to generate a random RSA keypair.

On the other hand, there is a noticeable decrease in performance with the two-step digital signatures because CRT-based optimizations are not available. Exploiting the CRT requires knowledge of the prime factors of n (p and q), which would allow the full private exponent to be computed from the public exponent, defeating the purpose of key splitting. Forgoing CRT-based optimizations, we instead used BouncyCastle library functions based on Montgomery's optimization for modular exponentiation.

While this two-step digital signature approach incurs a performance overhead, we consider it acceptable, especially for signing software images, for at least the following reasons:

1. A software image is often signed after a code build, which often involves orders of magnitude more computation than the public-key operations in code signing. This also means code signing is not a very frequently repeated operation for the same software.
2. In code signing, a software image is first reduced to a digest using a cryptographic hash function. The process is likely to be computationally more expensive than the public-key operations involved.
3. Code image signing is normally not expected to be a fast real-time operation. An overhead on the order of 1 second is generally not noticeable.
4. In our proposal, the TCSS's share of a split private key is protected by and used only inside an HSM. HSMs generally have hardware acceleration and very good performance for public-key operations. An increase in execution time of a fast operation by a small multiple is not a big overhead.

4.1. Additive Split Benchmarks

Table 2 – Additive Split Performance Measurements

Operation	Average execution time / ms		
	2048-bit	3072-bit	4096-bit
Key split	0.019	0.028	0.040
Standard CRT-based digital signature	30.479	97.410	290.090
Two-step digital signature[1]	207.100	703.553	2166.583

4.2. Multiplicative Split Benchmarks

Table 3 – Multiplicative Split Performance Measurements

Operation	Average execution time / ms		
	2048-bit	3072-bit	4096-bit
Key split	3.239	6.179	12.678
Number of retries (for the key split)	3.306	3.276	3.561
Standard CRT-based digital signature	30.479	97.410	290.090
Two-step digital signature[1]	206.855	699.414	2159.088

4.3. Benchmark Results Summary

In our experiments, two-step signature computations were very slightly slower with additive key splitting than with multiplicative splitting. This is probably because with additive splitting, an additional modular multiplication is required. The performance difference between the two approaches is insignificant in practical terms.

In our experiments with multiplicative key splitting, we used a simple generate-and-test strategy for selecting d_1. On average it took somewhere between 3 to 4 tries to find a d_1 that is relatively prime to $\phi(n)$. In any case, both multiplicative split and additive split have negligible performance overheads when compared to the computation in RSA keypair generation.

Based on our experimental measurements, we do not see a reason to favor one key splitting approach over the other on performance grounds. Both approaches are viable and are very similar performance-wise. And while two-step digital signatures result in a factor of 7 increase in total execution time over conventional CRT-based RSA signatures, that overhead is generally not significant in the context of software image signing. Note that in the total execution time, roughly half of the computation is normally performed by the TCSS. Between the TCSS and clients used by authorized users, only the TCSS is a centralized resource. Therefore, for the TCSS the performance overhead is only about half of what is suggested by the seven-fold increased execution time.

5. Security Considerations

1) As mentioned in section 4, RSA optimization based on the Chinese remainder theorem cannot be used in this two-step digital signature approach. Doing so would defeat the goal of having two

[1] The measured execution time includes both the computation performed by the TCSS as well as that performed by the authorized user.

parties involved in the code signing process, each having access to only one share of the private key.

2) Key splitting requires a cryptographically strong random number generator to ensure unpredictability and that all valid choices for the combination of private key shares are nearly equally likely. The same random number generator which qualifies for key pair generation can also be utilized for the purpose of key splitting.
3) There are known attacks against RSA based on “small” private exponents, such as the ones in [6] and [14]. We do not believe they are applicable to the key splitting methods we described. We note that when a private exponent is split into two shares, they are not always valid RSA exponents. Even when they are, no corresponding public keys are calculated or published for the shares. Also, in the methods we described, a pair of private key shares d_1 and d_2 is chosen uniformly at random from among all valid choices. Only a negligible fraction of such choices have one or both of d_1 and d_2 is less than, say, half of the bit length of the modulus.

6. Conclusions

We described an architecture for applying split-key digital signature to create a code signing service with enhanced security and explained the advantages it offers. Two approaches of splitting RSA signing keys are presented as examples. We anticipated that signature generation using split keys would incur performance overhead because a commonly employed optimization technique becomes unavailable when split keys are used. We performed experiments using a software implementation to gauge the computational overhead. Measurements from the experiments confirmed our expectation that the overhead is very acceptable in typical code signing usage scenarios.

7. Abbreviations and Definitions

7.1. Abbreviations

CRT	Chinese remainder theorem
HSM	hardware security module
RSA	Rivest-Shamir-Adleman public-key cryptosystem
TCSS	trusted code signing service
TTP	trusted third party

7.2. Definitions

authorized signer	a person authorized by a subscriber (organization) to have possession of a share of a split signing key belonging to the subscriber and to request signatures from a trusted code signing service
partly-computed signature	an intermediate result computed using one share of a split private signing key. It is used later in combination with a second share of the signing key to compute a complete digital signature.
share	one of the outputs from splitting a private key. Knowledge of all the shares of a private key is equivalent to knowledge of the private key. Knowledge of only one share does not make it feasible to generate a valid signature.

subscriber	a party that uses a TCSS for two-step code signing. A subscriber may be an organization but may alternatively be an individual.

8. Bibliography and References

[1] "Application Signing", *Android Open Source Project*, undated. [Online]. Available: https://source.android.com/security/apksigning. [Accessed: November 17, 2021].

[2] Microsoft, "Windows Authenticode Portable Executable Signature Format," March 21, 2008. [Online]. Available: http://download.microsoft.com/download/9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184a/authenticode_pe.docx.

[3] CableLabs, "Data-Over-Cable Service Interface Specifications DOCSIS® 3.1 Security Specification," *CableLabs*, CM-SP-SECv3.1-I04-150910, September 10, 2015. [Online]. Available: https://community.cablelabs.com/wiki/plugins/servlet/cablelabs/alfresco/download?id=00d39889-0af8-4722-b8a2-0063eeaa460a.

[4] B. Lynn, "The Chinese Remainder Theorem", undated. [Online]. Available: https://crypto.stanford.edu/pbc/notes/numbertheory/crt.html.

[5] Ç. K. Koç, "Montgomery Arithmetic", in *Encyclopedia of Cryptography and Security*, 2011 ed., Boston, MA: Springer, 2011.

[6] M. Wiener, "Cryptanalysis of short RSA secret exponents," *IEEE Trans. Inform. Theory*, vol. 36, pp. 553–558, May 1990.

[7] K. Moriarty, B. Kaliski, J. Jonsson, and A. Rusch, "PKCS #1: RSA Cryptography Specifications Version 2.2", *Internet Engineering Task Force*, RFC 8017, November 2016. [Online]. Available: https://www.rfc-editor.org/rfc/rfc8017.txt.

[8] P. Nohe, "Code Signing Compromise Installs Backdoors on Thousands of ASUS Computers," *thesslstore.com*, April 1, 2019. [Online]. Available: https://www.thesslstore.com/blog/code-signing-compromise-installs-backdoors-on-thousands-of-asus-computers/.

[9] T. Anderson, "HashiCorp reveals exposure of private code-signing key after Codecov compromise," *The Register*, April 26, 2021. [Online]. Available: https://www.theregister.com/2021/04/26/hashicorp_reveals_exposure_of_private/.

[10] D. Goodin, "Crooks steal security firm's crypto key, use it to sign malware," *Ars Technica*, February 8, 2013. [Online]. Available: https://arstechnica.com/information-technology/2013/02/cooks-steal-security-firms-crypto-key-use-it-to-sign-malware/.

[11] Center for Internet Security, "The SolarWinds Cyber-Attack: What You Need to Know," *Center for Internet Security*, March 15, 2021. [Online]. Available: https://www.cisecurity.org/solarwinds/.

[12] C. Cimpanu, "Nightmare week for security vendors: Now a Trend Micro bug is being exploited in the wild," *The Record*, April 22, 2021. [Online]. Available: https://therecord.media/nightmare-week-for-security-vendors-now-a-trend-micro-bug-is-being-exploited-in-the-wild/.

[13] T. Seals, "Zoho ManageEngine Password Manager Zero-Day Gets a Fix, Amid Attacks," *Threatpost*, September 9, 2021. [Online]. Available: https://threatpost.com/zoho-password-manager-zero-day-attack/169303/.

[14] D, Boneh and G. Durfee, "Cryptanalysis of RSA with private key *d* less than $N^{0.292}$," *IEEE Trans. Inform. Theory*, vol. 46. pp. 1339–1349, 2000.

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Multilayered LoRaWAN Networks For Smart Cities

A Technical Paper prepared for SCTE by

Mohamed Daoud, Principal Engineer, Charter Communications
6360 S. Fiddlers Green Circle
Greenwood Village, CO 80111
Mohamed.Daoud@Charter.com
720-699-5077

Muhammad Khan, Principal Engineer, Charter Communications
6360 S. Fiddlers Green Circle
Greenwood Village, CO 80111
Muhammad.J.Khan@Charter.com
720-536-1578

Hossam Hmimy, Senior Director, Charter Communications
6360 S. Fiddlers Green Circle
Greenwood Village, CO 80111
Hossam.Hmimy@Charter.com
720-536-9396

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

1. Introduction

1.1. What's LoRaWAN™

LoRaWAN stands for Long Range Wide Area Network. It's a low power wide area network (LPWAN) that utilizes open-source technology and transmits over unlicensed frequency bands. It uses the industrial scientific and medical (ISM) band which is 902MHz-928MHz in the United States. Designed for the Internet of Things (IoT), LoRaWAN technology provides longer range connectivity than Wi-Fi or Bluetooth, works well indoors and outdoors, and is especially valuable for applications in remote areas where cellular networks have poor coverage.

LoRaWAN end devices communicate with a LoRaWAN gateway which relays the messages to a LoRaWAN network server and finally the messages and data are visualized in a human readable format on an application server.

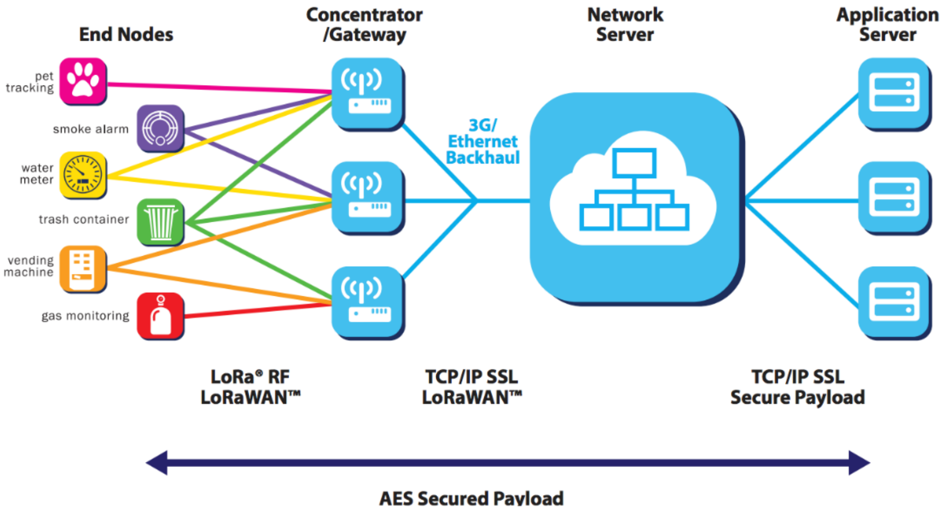


Figure 1 - LoRaWAN End-to-End Architecture

1.2. LoRa vs. LoRaWAN

LoRa and LoRaWAN are always used interchangeably; however, they aren't the same thing.

LoRa (Long Range) is an LPWAN protocol that defines the physical layer of a network. It's a proprietary technology owned by Semtech (chip manufacturer) that uses chirp spread spectrum to convert radio frequencies (RF) into bits so they can be transported through a network. LoRa is one of the technologies that makes LoRaWAN possible, but it's not limited to LoRaWAN, and can be used for different technologies.

LoRaWAN (long range wide area network) is an upper layer protocol that defines the network's communication and architecture. LoRaWAN provides the medium access control (MAC) layer protocol with some network layer components. It leverages LoRa, but it specifically refers to the network and how data transmissions travel through it.

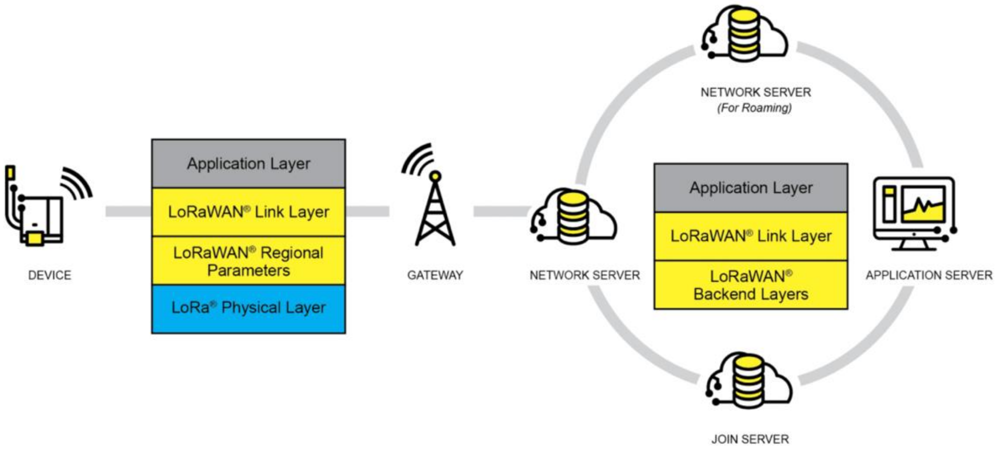
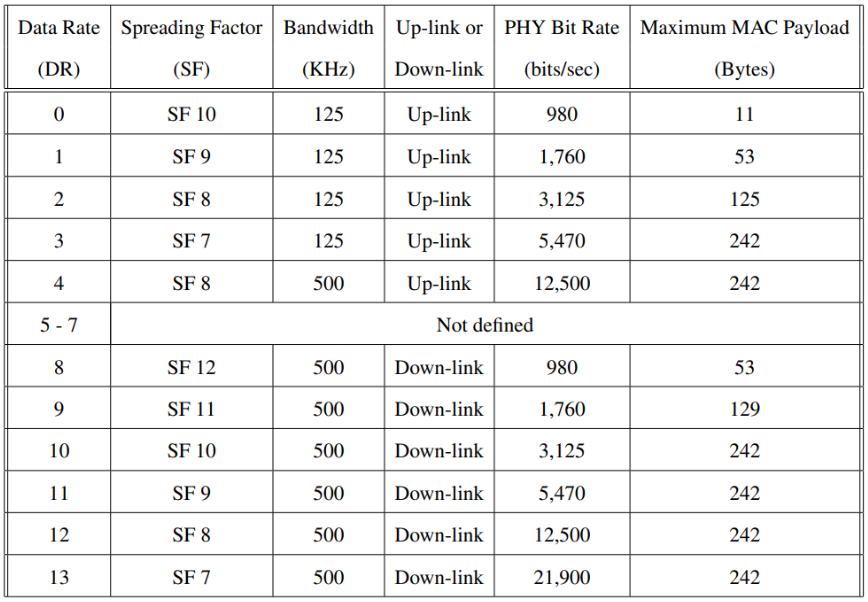


Figure 2 - LoRa and LoRaWAN Protocol Layers

1.3. Advantages and Problems of LoRaWAN

The main advantage of LoRaWAN is being able to utilize unlicensed band for communications. This is a very important aspect since spectrum is a scarce resource that is managed by the Federal Communications Commission (FCC). The unlicensed spectrum enables more communication providers to participate and it also means a bigger ecosystem of vendors offering devices, gateways, network servers, and application server solutions. The sub 1 GHz frequency for LoRaWAN means wide coverage outdoors and good indoor penetration, along with low power consumption of the device level which makes the device last for many years.

One of the limitations currently with LoRaWAN is the low adoption rate among businesses since it's a fairly new technology. However, more and more members are joining the LoRa Alliance and contributing to the standard which will help increase the LoRaWAN adoption. Another item to keep in mind is that LoRaWAN has limited data message capabilities given the maximum payload is 256 bytes and bandwidth of 125KHz/500KHz on the uplink and 500KHz on the downlink. However, this limitation is by design in order to carry small messages over long ranges.



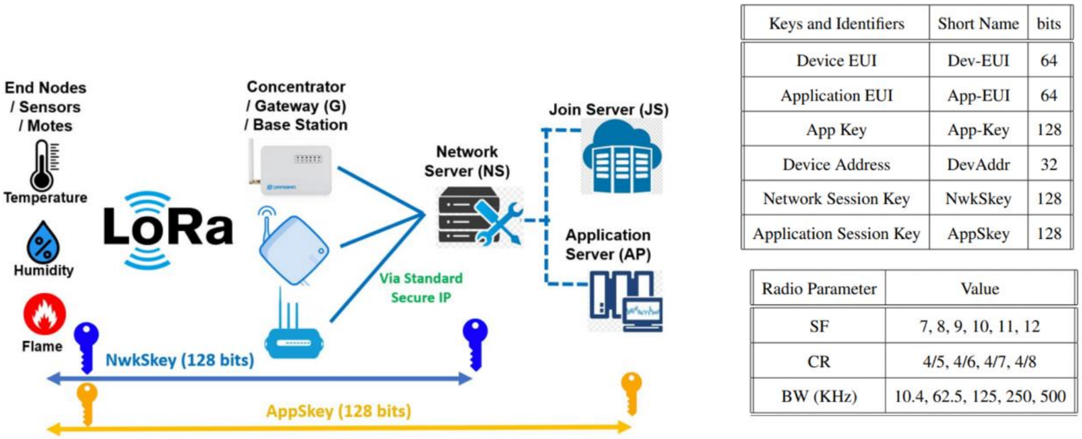
Data Rate (DR)	Spreading Factor (SF)	Bandwidth (KHz)	Up-link or Down-link	PHY Bit Rate (bits/sec)	Maximum MAC Payload (Bytes)
0	SF 10	125	Up-link	980	11
1	SF 9	125	Up-link	1,760	53
2	SF 8	125	Up-link	3,125	125
3	SF 7	125	Up-link	5,470	242
4	SF 8	500	Up-link	12,500	242
5 - 7	Not defined				
8	SF 12	500	Down-link	980	53
9	SF 11	500	Down-link	1,760	129
10	SF 10	500	Down-link	3,125	242
11	SF 9	500	Down-link	5,470	242
12	SF 8	500	Down-link	12,500	242
13	SF 7	500	Down-link	21,900	242

Figure 3 - LoRaWAN Spreading Factor and Data Rates

2. LoRaWAN Technology

2.1. LoRaWAN Architecture

LoRaWAN networks typically are laid out in a star-of-stars topology in which gateways relay messages between end devices and a central network server. The network server routes the packets from each device of the network to the associated application server. To secure radio transmissions, the LoRaWAN protocol relies on symmetric cryptography using session keys derived from the device's root keys. In the backend, the storage of the device's root keys and the associated key derivation operations are provided by a join server.



Keys and Identifiers	Short Name	bits
Device EUI	Dev-EUI	64
Application EUI	App-EUI	64
App Key	App-Key	128
Device Address	DevAddr	32
Network Session Key	NwkSkey	128
Application Session Key	AppSkey	128

Radio Parameter	Value
SF	7, 8, 9, 10, 11, 12
CR	4/5, 4/6, 4/7, 4/8
BW (KHz)	10.4, 62.5, 125, 250, 500

Figure 4 - LoRaWAN Security Keys and Parameters

Gateways are connected to the network server via secured standard IP connections while end devices use single-hop LoRa communication to one or many gateways. All communication is generally bidirectional, although uplink communication from an end device to the network server is expected to be the predominant traffic. Communication between end devices and gateways is spread out on different frequency channels and data rates. The selection of the data rate is a trade-off between communication range and message duration. Communications with different data rates do not interfere with one another. To maximize both battery life of the end devices and overall network capacity, the LoRa network infrastructure can manage the data rate and RF output for each end device individually by means of an adaptive data rate (ADR) scheme.

Some countries impose a duty cycle restriction which is the maximum amount of time a device can spend communicating. In Europe, the European Telecommunications Standards Institute (ETSI) sets the maximum duty cycle for the EU863-870 frequency at 1%. However, the FCC has not imposed any duty cycle regulation in the United States.

2.2. Chirp Spread Spectrum

LoRa is based on chirp spread spectrum (CSS) technology, where chirps are the data carriers.

Spreading factor (SF) controls the chirp rate, and thus controls the speed of data transmission. Lower SFs mean faster chirps, and, therefore, a higher data transmission rate. For every increase in SF, the chirp sweep rate is halved, and so the data transmission rate is halved.

Lower SFs reduce the range of LoRa transmissions, because they reduce the processing gain and increase the bit rate. Changing the SF allows the network to increase or decrease the data rate for each end device at the cost of range.

LoRa modulation has a total of six spreading factors from SF7 to SF12. SFs influence data rate, time-on-air, battery life and receiver sensitivity.

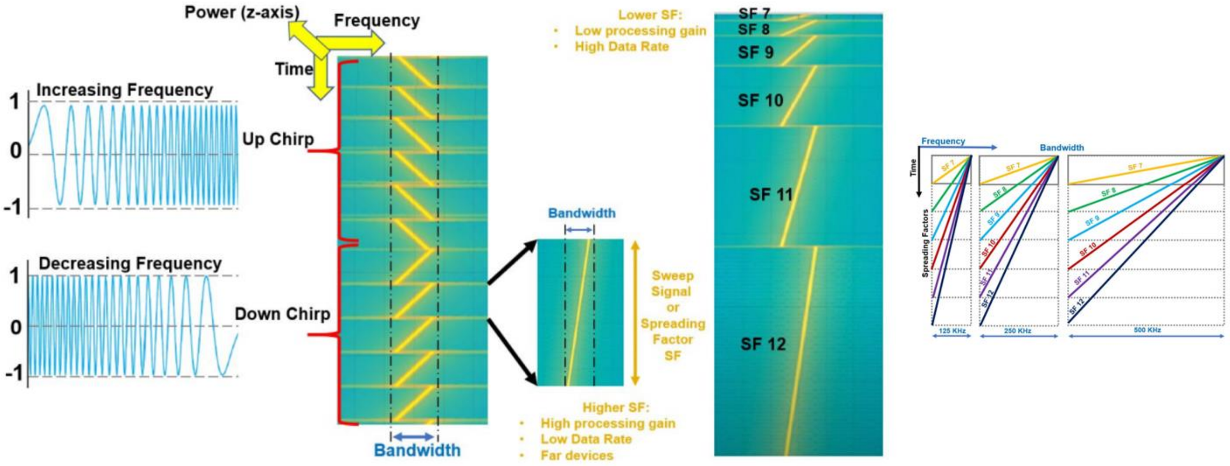


Figure 5 - Different Chirps Levels

It’s always a trade-off with the SF because a big SF means long range, but low bit rate while a small SF means short range, but high bit rate. A high SF means a large processing gain. A signal modulated with a high SF can be received with less errors compared to a signal with a low SF, and, therefore, travel a longer distance. For instance, a signal modulated with SF12 will travel a longer distance than a signal modulated with SF7. In addition, sending a fixed amount of data with a high SF and a fixed bandwidth needs longer time-on-air. Thus, higher SFs result in longer active times for the radio transceivers and shorter battery life.

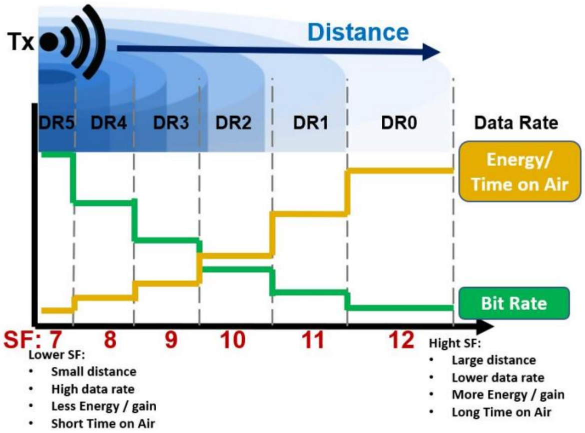


Figure 6 - SF and Data Rate Tradeoff

2.3. End Devices

End devices are the LoRaWAN sensors that connect to LoRaWAN gateways. Every end device must be registered with a network before sending and receiving messages. This procedure is known as activation. There are two activation methods available:

- The first is over-the-air activation (OTAA). This is the most secure and recommended activation method for end devices. Devices perform a join procedure with the network, during which a dynamic device address is assigned and security keys are negotiated with the device.
- The second is activation by personalization (ABP). This requires hardcoding the device address as well as the security keys in the device. ABP is less secure than OTAA, and also has the downside that devices cannot switch network providers without manually changing keys in the device.

There are three classes of LoRaWAN devices: Class A, B, and C.

Class A devices support bidirectional communication between a device and a gateway. Uplink messages (from the device to the server) can be sent at any time. The device then opens two receive windows at specified times (RX1 Delay and RX2 Delay) after an uplink transmission. If the server does not respond in either of these receive windows, the next opportunity will be after the next uplink transmission from the device.

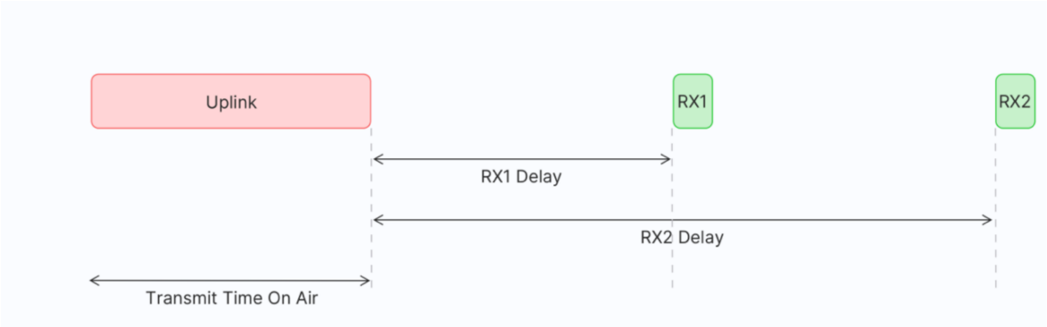


Figure 7 - Class A Device Operation

Class B devices add scheduled receive windows for downlink messages from the server. Using time-synchronized beacons transmitted by the gateway, the devices periodically open receive windows.

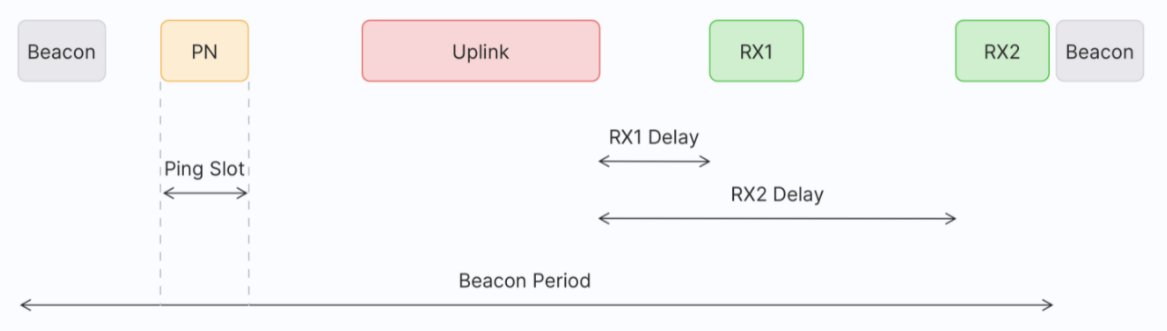


Figure 8 - Class B Device Operation

Class C devices keep the receive windows open unless they are transmitting.

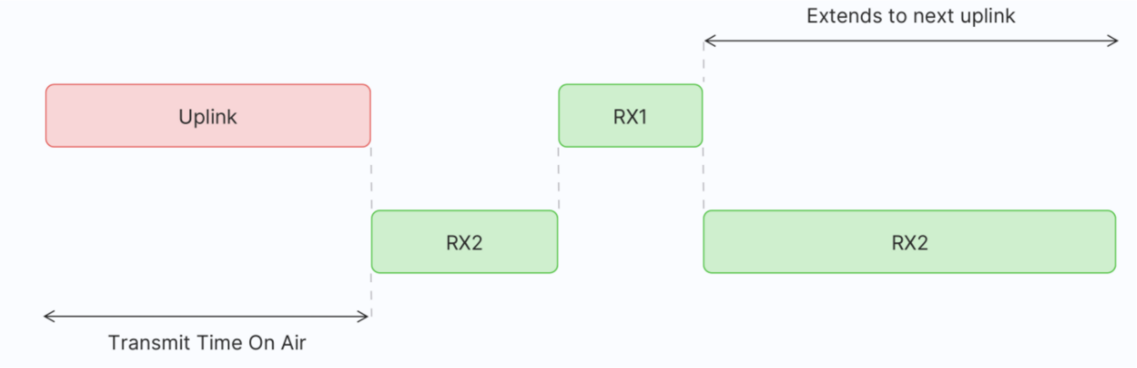


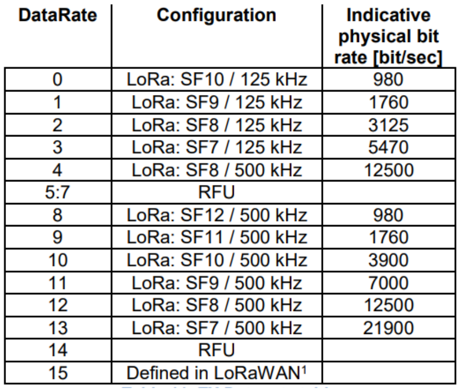
Figure 9 - Class C Device Operation

With regard to uses, class A is for battery operated sensors in the field where one of the goals is to reduce battery consumption, while class C is for latency sensitive sensors like luminaires on light poles or smart meters for example. Class B is in between both classes. The variety of classes gives the end user choices for their applications.

2.4. Network Planning and RF Design

When doing network planning for a LoRaWAN network, three factors must be considered:

- **Desired coverage area**: As with any other wireless RF design, one must model the environment as accurately as possible. For indoor scenarios, this means modeling the correct wall type, material loss (dBm/feet) and dimensions. For outdoor scenarios, the correct terrain and clutter classes, along with associated properties, must be modeled. Furthermore, the designer must establish the minimum desired signal level and the percentage of area to be covered.
- **Desired capacity**: The required data rate (DR) and SF is an important consideration in the design. In the U.S. frequency plan, end devices are allowed to communicate between SF10 (DR 0) and SF7 (DR3) in the uplink. A network designed for SF10 will have the largest coverage, and is good for use cases in which there are small, infrequent transmissions of uplink data and redundant coverage is not important. When the device sends a larger payload or when there is little room for errors, SF8 or SF7 is better. The SF vs. DR mapping is shown in Figure(10) from the LoRa alliance regional parameters.



DataRate	Configuration	Indicative physical bit rate [bit/sec]
0	LoRa: SF10 / 125 kHz	980
1	LoRa: SF9 / 125 kHz	1760
2	LoRa: SF8 / 125 kHz	3125
3	LoRa: SF7 / 125 kHz	5470
4	LoRa: SF8 / 500 kHz	12500
5:7	RFU	
8	LoRa: SF12 / 500 kHz	980
9	LoRa: SF11 / 500 kHz	1760
10	LoRa: SF10 / 500 kHz	3900
11	LoRa: SF9 / 500 kHz	7000
12	LoRa: SF8 / 500 kHz	12500
13	LoRa: SF7 / 500 kHz	21900
14	RFU	
15	Defined in LoRaWAN[1]	

Figure 10 - Different SF Values and Bandwidth

- **Overlap area**: Unlike traditional cellular networks, the end devices in LoRaWAN communicate to multiple gateways in the uplink. Thus, it is better to have overlapping coverage and design for redundancy in case one of the gateways goes offline. The desired number of gateways seen by the end device can depend on the use case. For example, in a critical scenario like emergency buttons, there should be at least two to three gateways covering any given area since the packet error rate (PER) must be extremely low. In other less critical cases like environmental monitoring, it is acceptable even if a few packets are lost due to gateways being disconnected. There is always a trade-off between redundancy and cost.

Charter has deployed both outdoor and indoor gateways in customer pilots and lab environments. The figures below show the outputs of these designs.

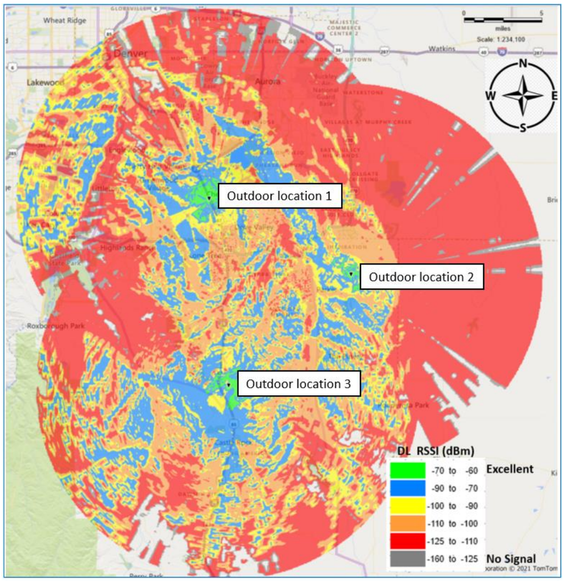


Figure 11 - Coverage Plot for Outdoor Deployment

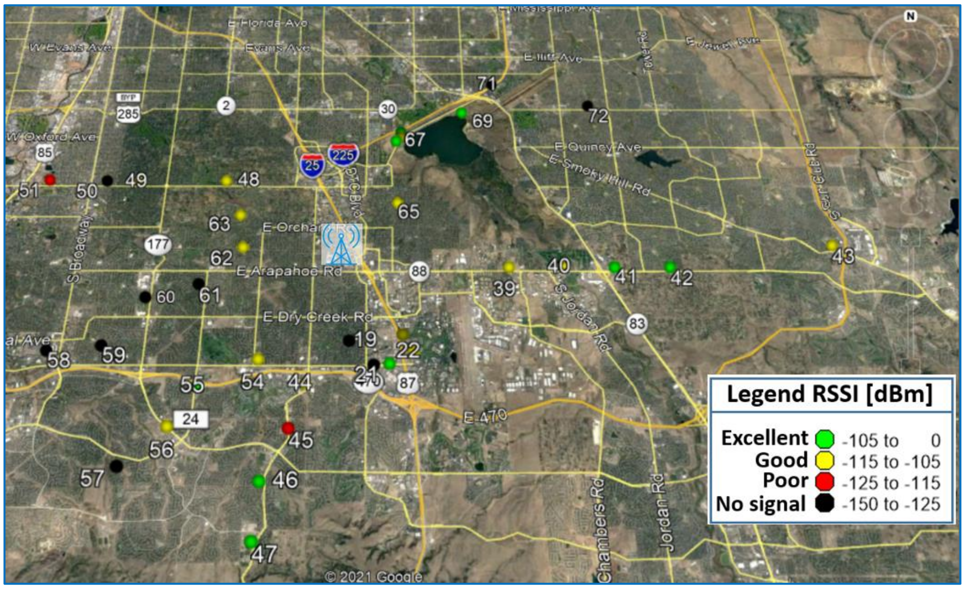


Figure 12 - Outdoor Deployment Network Field Test Result

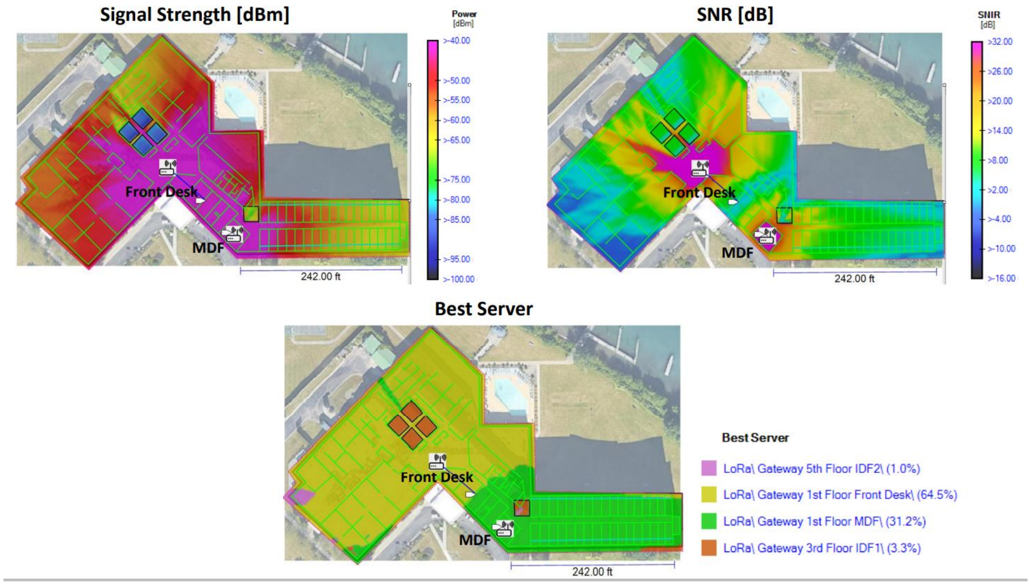


Figure 13 - Indoor LoRaWAN Design

3. LoRaWAN Use Cases

3.1. Indoor Localization

An important use case is indoor localization and tracking. Consider, for example, the need for hospitals and factories to locate key equipment inside their facilities where GPS doesn’t work. While there are many indoor location technologies based on Bluetooth beacons, their range is limited and also requires adding beacon Bluetooth devices around the building.

An innovative indoor localization solution is combining LoRaWAN, Wi-Fi, and artificial intelligence (AI). The WI-FI equipment building is scanned to determine the Wi-Fi signal strength at different locations in the building. The result of the scan is a Wi-Fi heatmap that’s uploaded to a prediction server. LoRaWAN gateways are installed throughout the building to ensure adequate coverage, and equipment to be tracked is equipped with LoRaWAN sensors. The LoRaWAN sensor is triggered periodically or on-demand. With each trigger, the sensor scans the WI-FI in the surrounding area for signal strength and reports the signal strength over LoRaWAN to the LoRaWAN network server. The LoRaWAN network server decodes the packets and forwards them to the prediction server which, in turn, uses AI to predict the location based on the preloaded Wi-Fi heat map. The diagram below shows a high-level architecture for the solution.

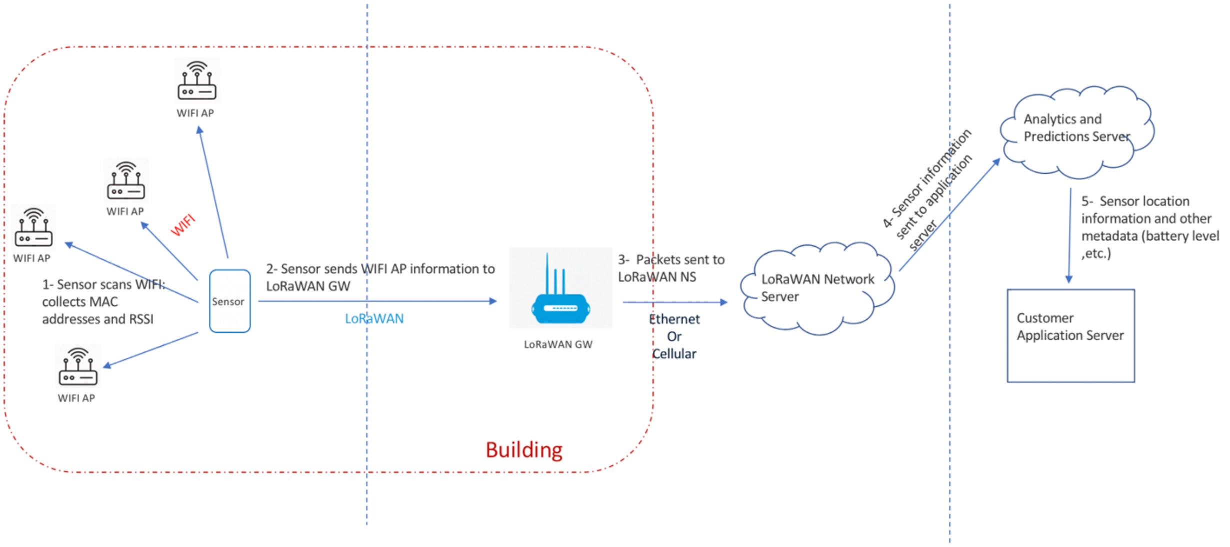

WIFI AP
WIFI AP
WIFI AP
WIFI AP
WIFI
1- Sensor scans WIFI: collects MAC addresses and RSSI
Sensor
2- Sensor sends WIFI AP information to LoRaWAN GW
LoRaWAN
LoRaWAN GW
3- Packets sent to LoRaWAN NS
Ethernet Or Cellular
Building
LoRaWAN Network Server
4- Sensor information sent to application server
Analytics and Predictions Server
5- Sensor location information and other metadata (battery level ,etc.)
Customer Application Server

Figure 14 - Indoor Localization and Tracking Solution

3.2. Outdoor Tracking

Another use case for LoRaWAN is outdoor tracking. For example, this is a common use case for many verticals like fleet management. The solution requires a LoRaWAN outdoor gateway mounted on a building rooftop or tower for a large coverage area. The vehicles are equipped so that GPS is connected to LoRaWAN sensors. The LoRaWAN sensors read the GPS data periodically and send the location over LoRaWAN. The benefit of this solution is the wide area coverage of LoRaWAN and relatively inexpensive infrastructure.

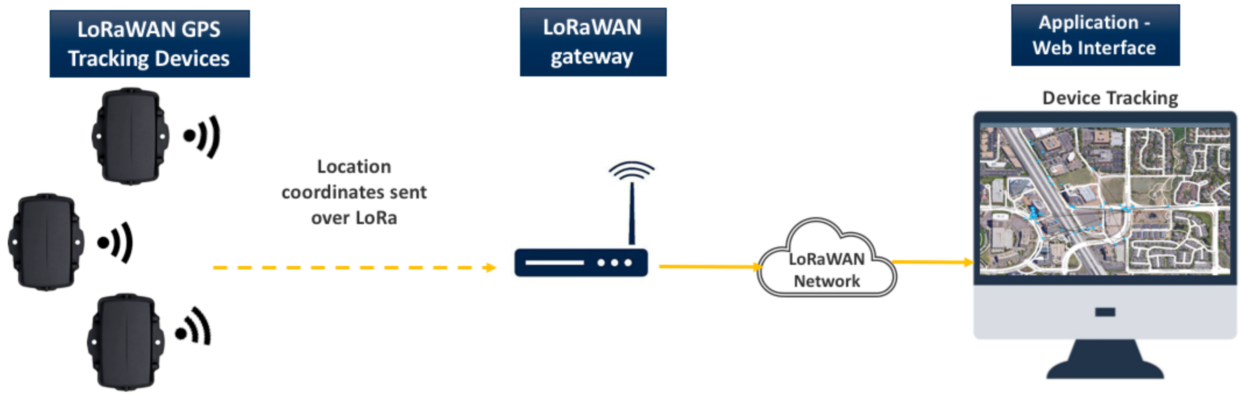

LoRaWAN GPS Tracking Devices
LoRaWAN gateway
Application - Web Interface
Device Tracking
Location coordinates sent over LoRa
LoRaWAN Network

Figure 15 - Outdoor LoRaWAN Tracking Solution

3.3. LoRaWAN for Smart Buildings

Smart buildings are another use case where LoRaWAN can be leveraged to improve occupant experience and building management efficiency. At Charter Communications, we piloted smart buildings by installing LoRaWAN sensors to monitor things like room occupancy, air quality, temperature, humidity, motion, gas/leak detection, and door status.

In office buildings, occupancy sensors can help building management derive insights on the actual occupancy and better utilize meeting rooms and desks.

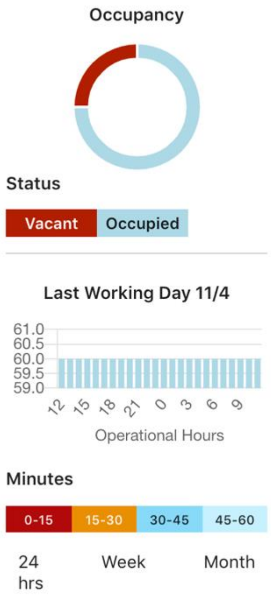


Figure 16 - Desk Occupancy

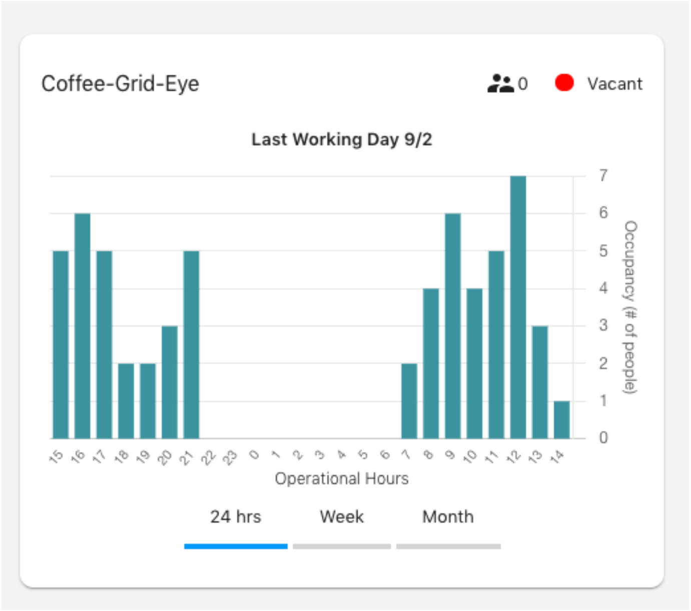
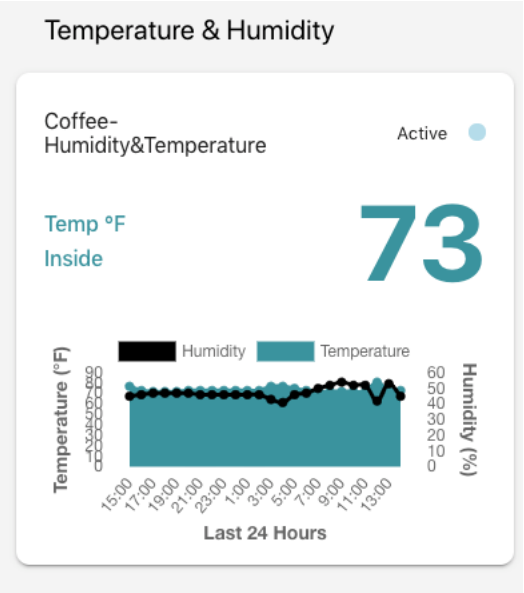


Figure 17 - Room Occupancy

Temperature and humidity sensors enable buildings to get real-time data to a central dashboard. A sudden change in temperature could indicate a fault. Thanks to LoRaWAN sensors, this sudden change can be reported immediately to the central dashboard.



Temperature & Humidity

Coffee-Humidity&Temperature

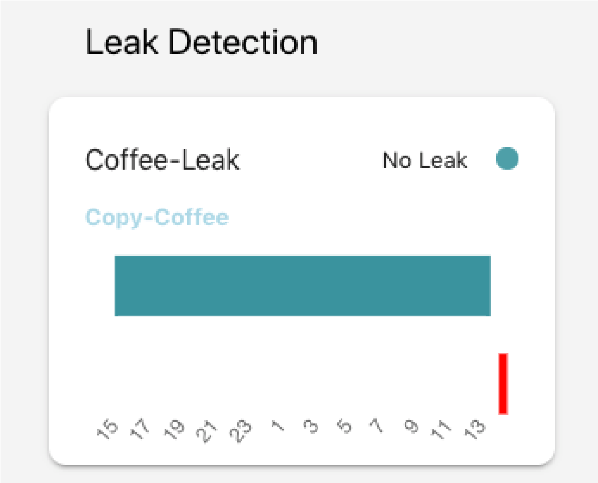
Active

Temp °F Inside

73

Figure 18 - Temperature and humidity readings

The leak detection allows building management to discover a leak as soon as it happens and react quickly.



Leak Detection

Coffee-Leak

No Leak

Copy-Coffee

Figure 19 - Leak Detection Sensor

LoRaWAN will help accelerate the deployment of smart building use cases due to the technology and ecosystem maturity.

3.4. Private Networks and LoRaWAN

The architecture for LoRaWAN allows for both cloud-hosted and private on-premises networks. There are several advantages of a private network. Enterprise and end customers may have special requirements that the data cannot leave the premises. This could be accomplished with a private network on which both the network and application are hosted on either servers or mini edge compute on premises. Second, there are use cases in which low latency is important because an automated action needs to be taken based on sensor data. For example, if the temperature of a cold storage room goes above a certain threshold, then immediate action could be taken. Similarly, if carbon dioxide levels rise above the recommended value in a factory or other venue, then automated actions can be set up to alleviate the problem. Third, private networks allow for the end customer to set their own quality of service (QoS) and adaptive data rate (ADR) parameters to customize their network based on their specific use cases.

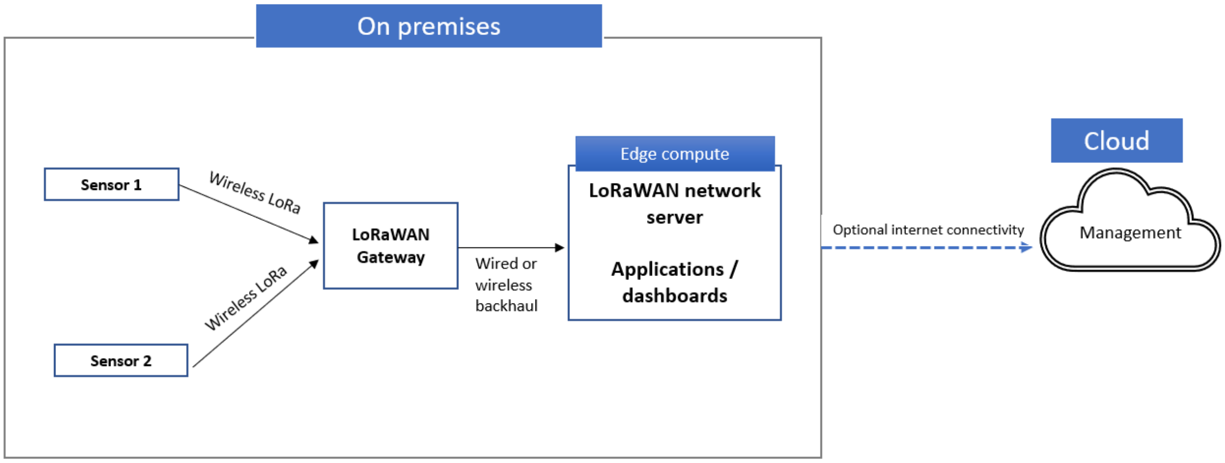


Figure 20 - Private LoRaWAN Network Server

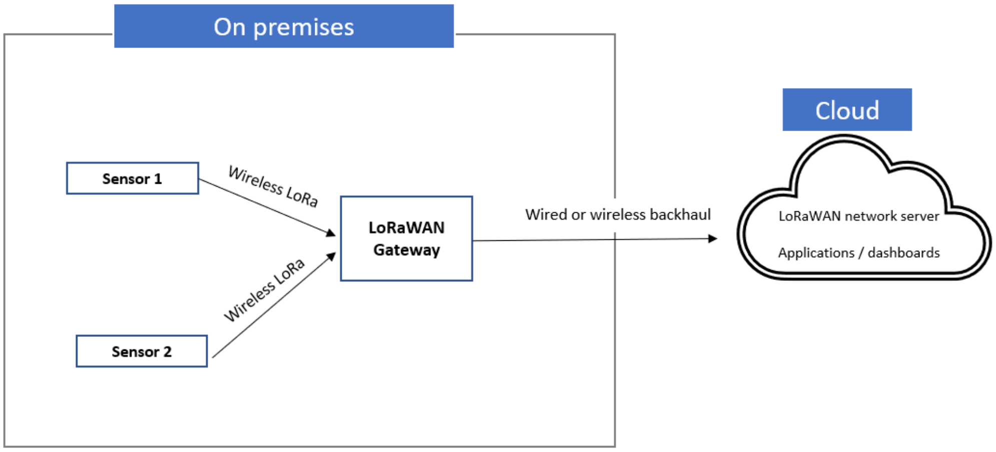


Figure 21 - Cloud Hosted LoRaWAN Network Server

LoRaWAN can also be used along with other wireless technologies to take advantage of the strength that each communication protocol brings. In a factory scenario, the indoors can be covered with higher bandwidth technologies like LTE/5G or Wi-Fi. If the outdoor use case consists primarily of low data rate sensors, then LoRaWAN is ideal to cover the large parking areas and loading docks, eliminating the need to deploy a high-capacity network. This LoRaWAN gateway can then be backhauled as part of the same private network. This way, LoRaWAN can complement existing deployments rather than acting as a separate network.

3.5. LoRa in a Box

Charter Communications has developed a LoRa in a box proof of concept (POC) and reference design based on lightweight version of the network server on edge, mini gateways and application servers. It is a plug and play prototype network, which can be used to set up LoRaWAN trials anywhere. For example, in remote sites where access to backhaul is challenging, this box can be powered on to provide long range connectivity for various sensors and devices. This setup is also ideal for showing the power of LoRaWAN with some use cases in conferences and to potential customers.

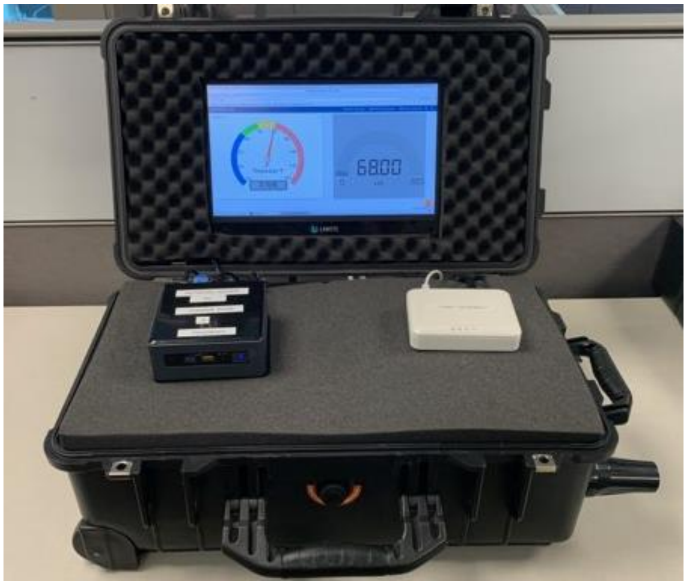


Figure 22 - LoRaWAN In a Box

4. Conclusion

LoRaWAN IoT is a disrupter technology and can be easily leveraged for operational efficiency and solving everyday problems. LoRaWAN is promising because it uses ISM unlicensed spectrum, covers wide areas using low frequency, is less susceptible to interference as it uses CSS modulation, and has built-in security measures.

5. Abbreviations

ABP	activation by personalization
ADR	adaptive data rate
AI	artificial intelligence
CSS	chirp spread spectrum
DR	data rate
ETSI	European Telecommunications Standards Institute
FCC	Federal Communications Commission
IOT	Internet of Things
ISM	industrial scientific and medical
LPWAN	low power wide area network
MAC	medium access control
OTAA	over-the-air-activation
PER	packet error rate
QoS	quality of service
SF	spreading factor

6. Bibliography and References

https://lora-alliance.org/

https://www.thethingsnetwork.org/

https://lora-alliance.org/wp-content/uploads/2021/05/RP002-1.0.3-FINAL-1.pdf

https://www.semtech.com/

https://aws.amazon.com/iot-core/lorawan/

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

Rethinking DDoS Security in the Era of Volumetric Attacks

New DDoS Security Requirements for the Era of the 5G, Cloud and IoT

A Technical Paper prepared for SCTE by

Dr. Craig Labovitz, CTO, Nokia Deepfield
Nokia USA
Ann Arbor, MI
craig.labovitz@nokia.com
+1 734 276 4194

Stefan Meinders, Product Manager, Nokia Deepfield
Nokia Germany
Stuttgart
stefan.meinders@nokia.com
+49 17 0559 5862

Alex Pavlovic, Director of Product Marketing, Nokia Deepfield
Nokia Canada
Ottawa, ON
alex.pavlovic@nokia.com
+1 343 417 0930

Table of Contents

Title **Page Number**

List of Figures

Title **Page Number**

1. Introduction

DdoS threats and attacks are becoming more frequent and impactful in the era of the cloud, 5G and the Internet of Things (IoT). The growth of all-IP networks has extended the security perimeter and expanded the threat and attack surface. Attacks now come from both outside and within service provider networks and are aimed at internet hosts and servers, customers and users – as well as network infrastructure. A more intelligent and agile approach – using big data analytics, advanced IP routing, software-defined DdoS security and automation is required to thwart and minimize the security risks associated with a new generation of DdoS threats and attacks.

In the era of the cloud, 5G and the Internet of Things (IoT), networks matter more than ever. They are critical for all businesses, from manufacturing and supply chains to logistics. They are also critical for the functioning of society, from energy and resources to transportation and the public sector. Networks bring us together — at work, for remote learning, and at play.

The COVID-19 pandemic changed consumer online behavior and internet traffic patterns. Among the traffic types that indicated the highest growth was the traffic that corresponds to distributed denial of service (DDoS) attacks; these attacks have grown in number, frequency, intensity and sophistication.

2. Types of DDoS Attacks

There are many types of DDoS – some inflict harm at the network level, others at the application layer. One type of DDoS – volumetric DDoS – stands out as it poses the greatest danger because of its immense impact on the network, services and users, often also causing large collateral damage. Volumetric DDoS attacks comprise more than 95 percent of all DDoS traffic.

DDoS has been exploiting IP protocol and system vulnerabilities for more than two decades. Some protocols, such as Domain Name System (DNS), have gained additional security features, but many protocols still rely on open principles set by the internet community a long time ago.

Volumetric attacks can appear as high-bandwidth attacks, described by their total bandwidth, expressed in bits per second (b/s). These attacks aim to exhaust transmission capacity by the sheer volume of traffic.

Alternatively, volumetric attacks can appear as high packet-rate attacks, described by their packet intensity, expressed in packets per second (pps). These attacks aim to exhaust the processing capacity of network hosts and other network elements such as firewalls.

The most common volumetric attacks are amplification DDoS attacks. Amplification here denotes the ability of the attacker to send a small packet (e.g., 40 bytes request) and leverage all of the misconfigured servers (there are about 40 distinct types, including DNS servers, time servers, Connection-less Lightweight Directory Access Protocol - CLDAP servers) that will respond to a tiny packet with a response size of tens of MB of data (in some cases hundreds of MB) to the victim. As a result, a victim system gets a very large volume of traffic originating from many points (amplifiers) on the Internet.

Volumetric DDoS attacks, especially amplification and reflection attacks, generally require IP address spoofing (or IP header modification - IPHM). Spoofing the source IP address(es) hides the originator's IP address(es). More importantly, using the ability to spoof forge their source IP address, the attacker can send packets pretending their address is the IP address of their targets or victims. For example, the attacker might send traffic to regular internet servers like servers from Google, Akamai or Microsoft. These servers, fooled into thinking that they are responding to the original requester, will respond (e.g., in TCP language, with SYN-ACK packets), sending millions of packets to the victim and target systems.

Amplification and reflection of responses to spoofed IP traffic can lead to a high volume of traffic going to victims' systems.

Volumetric attacks also include protocol flooding attacks such as TCP, UDP, ICMP and other spoofed-packet floods. Here, DDoS attacks come in irregular protocol message exchanges aimed at confusing and saturating servers. Examples are synchronization packet (SYN) floods and fragmented packet attacks. The extraordinary amount of this traffic type leads to the exhaustion of state information on (stateful) firewalls on load balancers or on different parts of the infrastructure, again leading to end users experiencing serious delays and eventually being disconnected from the service.

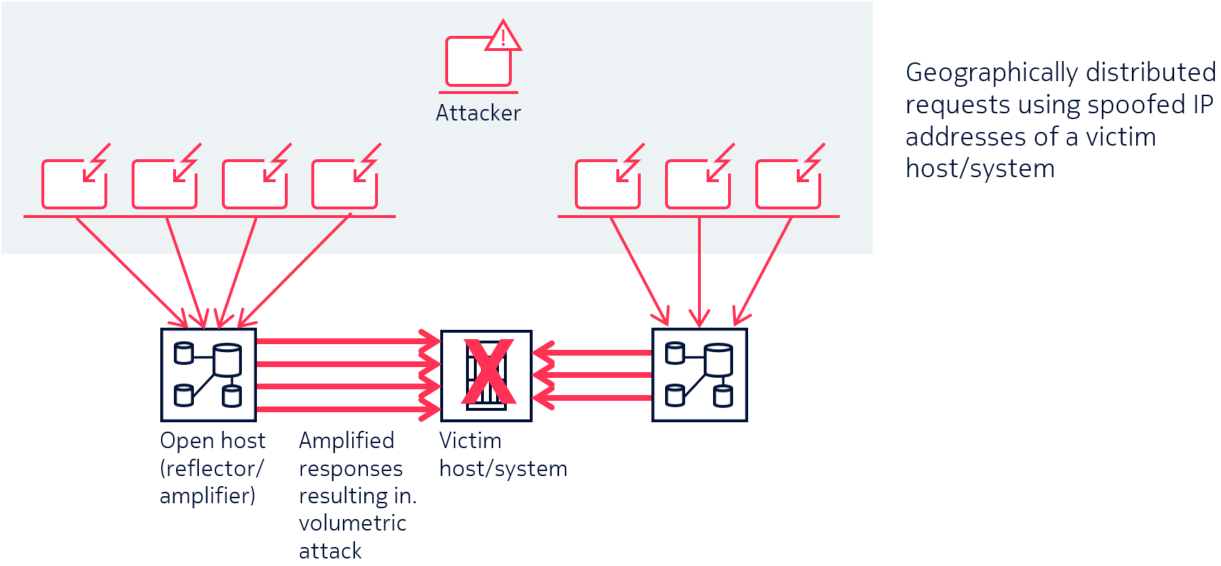

Attacker
Geographically distributed requests using spoofed IP addresses of a victim host/system
Open host (reflector/ amplifier)
Amplified responses resulting in. volumetric attack
Victim host/system

Figure 1 - Volumetric DDoS

A part of overall DDoS traffic comes from botnets – sets of compromised devices or systems that have been exploited and can be remotely controlled by a bot-master.

Before 2021, botnet-sourced attacks did not represent a significant part of DDoS. However, we have seen significant growth of botnet DDoS traffic in 2021, at times used as ransom DDoS, asking the targets to pay ransom money in cryptocurrency so the attacks would stop or not escalate to terabit levels.

Application-level attacks are aimed at state exhaustion of the target system. Attack types include low-and-slow attacks, GET/POST floods and other forms of attacks that target specific servers/applications or hosts.

Recent new DDoS techniques such as carpet bombing use an extended range of IP addresses as targets (instead of an IP address of a single host) or hide the "real" attack in a range of simultaneous attacks (e.g., a bits-and-pieces attack).

3. Motivation for DDoS Attacks

The motivation for DDoS attacks varies widely; while some attacks are just a nuisance, others are tools to achieve various goals.

Online gamers do it to win a round of a game as well as get an adrenaline rush.

Hacker activists called "hacktivists" are motivated by ideology and have a political or social agenda.

Extortion is common, with perpetrators using DDoS attacks — or the threat of attacks — to demand ransom from individuals or corporations (ransom DDoS). In some cases, DDoS attacks are combined with other malware attacks and are used to obfuscate or hide the real attack.

DDoS attacks are made easier with the advent of DDoS-for-hire services and the wider use of cryptocurrency. They have gone from being an annoyance to causing major business and service disruptions. As attack ROI and incentives increase, so do attackers' skill sets.

4. Impact and Damage from Volumetric DDoS Attacks

DDoS attacks spare no one. Targets range from individual users to networks belonging to service providers, cloud builders, and large digital enterprises.

While most DDoS attacks are a nuisance (e.g., to individual gamers), the bandwidth represented by high-bandwidth and high-packet intensity volumetric attacks is cause for concern. With volumetric amplification DDoS types, the attackers need a little bit of bandwidth and connectivity. They can then launch attacks leveraging millions of servers and IoT devices across the Internet to overwhelm interfaces, routers, load balancers, firewalls and network hosts. As a result, the performance is reduced, and services are downgraded or stopped.

These attacks can damage connectivity and service availability and result in losses costing thousands or even millions of dollars in production and operational losses. Additionally, there may also be legal costs, not to mention reputational damage.

Some segments, such as banking, insurance, and healthcare, can also be subject to high regulatory fines. In August of 2020, an attack on the New Zealand Stock Exchange left the exchange out of service for four days and incurred significant monetary loss plus a warning from the country's financial regulator.

In May of 2021, a Belgian network provider supplying connectivity services to the Belgian government, including remote learning and COVID-19 vaccines registration, was hit by a DDoS attack originating from 257,000 IP addresses in 29 countries — leaving many customers without vital connectivity.

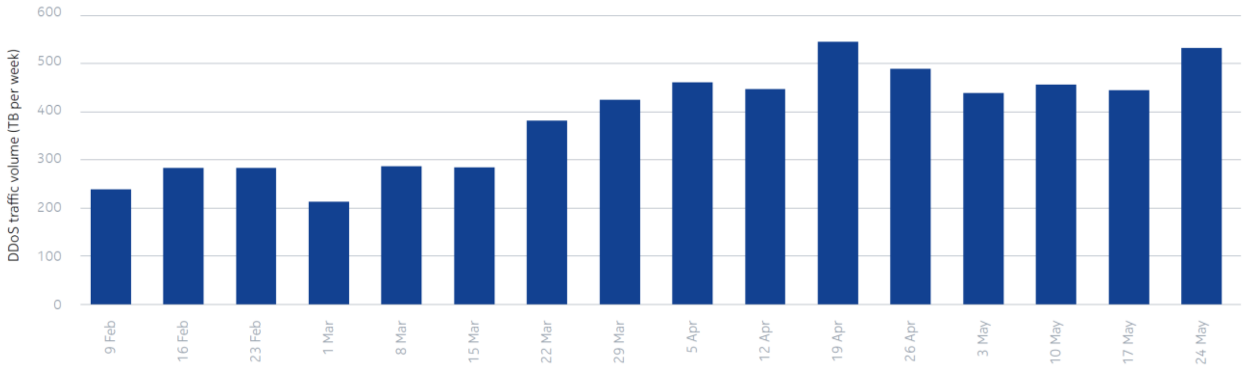
Volumetric DDoS is particularly dangerous to network routers and infrastructure. These attacks can affect tens of thousands of enterprises and millions of consumers.

While some big attacks get the headlines, many attacks go unreported because service providers do not want to expose details about their security capabilities or vulnerabilities. Even worse, many attacks go undetected or are reported by users on social media.

5. Quantifying the DDoS Danger

Volumetric DDoS attacks have increased dramatically in recent years. Around 2016, we entered the era of the first terabit-level DDoS attacks. Today, DDoS attacks are a reality for most networks. Aggregate DDoS volume levels show the constant threat to service providers. DDoS traffic rose significantly in 2020. As shown in Figure 2, in the short period from early February to late May, aggregate DDoS volume levels in the United States rose by more than 40 percent.

Note: The data in Figure 2 were aggregated across multiple US service providers.

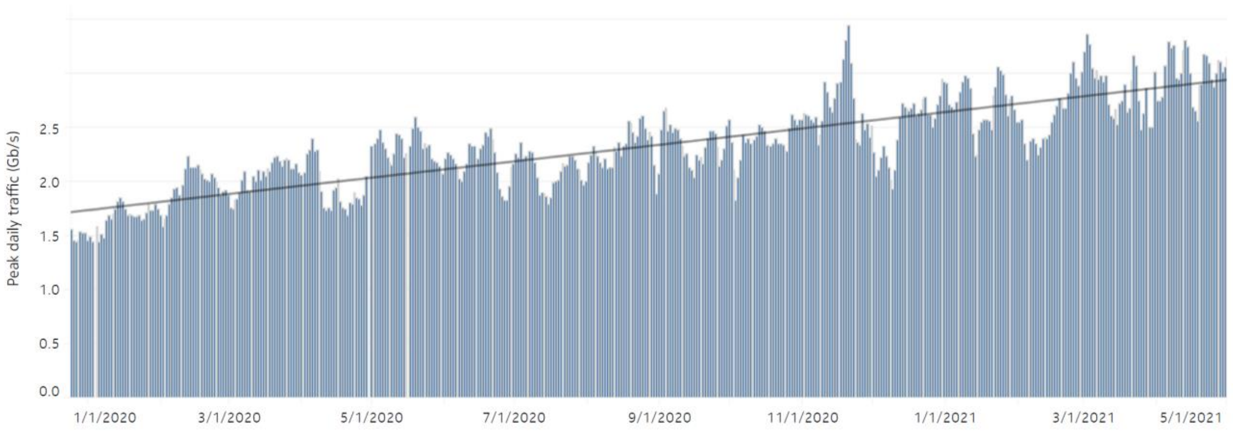


Source: Nokia Deepfield Network Intelligence Report, 2020, Page 43.

Figure 2 - Weekly DDoS traffic February - May 2020

Some might argue that this percentage is skewed because it covered the first wave of the COVID-19 pandemic and lockdowns, when remote work increased significantly, schools and universities went online, and people connected via video chats (when they were not gaming or streaming videos).

However, when compared over a longer-term and globally, the situation appears even worse. As shown in Figure 3, daily DDoS peaks have more than doubled in a little over a year – from 2020 to 2021. In January 2020, average daily 5-minute peaks were at 1.5 Tb/s. By May 2021, the average daily peaks were exceeding 3.0 Tb/s.



Source: Nokia Deepfield, May 2021.

Figure 3 - Peak daily DDoS traffic January 2020 - May 2021 across select service providers

It is also worth stating that in 2021 we have observed an increased incidence of botnet DDoS attacks using a variety of IoT devices – from insecure (and hijacked) VoIP terminal adapters to always-on high-bandwidth security cameras as well as CPE devices. With a growing number of IoT devices and increasing network bandwidth available to them, this category of DDoS may become even more threatening.

6. June 2021 DDoS Study

In June 2021, we completed multi-year research - looking at the Internet, a significant number of Tier-1 service provider networks, and encompassing thousands of network routers - and announced key findings.

Most DDoS analyses stop at amplifiers because this is where the DDoS attack traffic seems to originate. Tracing spoofed traffic beyond amplifiers and reflectors typically requires a lot of manual tracing and investigation. Our study used fingerprinting techniques to identify DDoS traffic. The use of other techniques and tools, such as tracking time-to-live (TTL) in addition to fingerprinting, allowed us to trace DDoS back from victims to the closest internet entry point.

Using these steps, we were also able to identify the specific hosting providers from which this DDoS traffic originated. This constituted about 40% to 50% of the cases.

For the remaining 50% to 60% of the traffic, additional steps were performed.

The first step was to obtain accounts and details on basically every DDoS-for-hire service known to the researchers. This included information obtained from darknet lists and even telegram messages used to negotiate paid access to some of these services), and resulted in a list that captures the majority of DDoS-for hire services offered from so-called "Booter" and "Stresser" websites. Hosting providers used to supply these websites were also identified.

Using packet captures and tracking down amplifiers, DNS payloads, NTP servers, Internet Control Message Protocol (ICMP) and other relevant metrics of the attacks, “fingerprints” that were taken that resulted in a broad range of measurements of both the hosting and the booter/stresser illegal DDoS services.

Leveraging techniques and algorithms or queries that were shared at NANOG82, such as analyzing TTL-values - to correlate these signatures to DDoS attacks happening across the Internet, the researchers were able to greatly narrow down the originating domains that previously had largely been hidden behind reflectors or amplifiers.

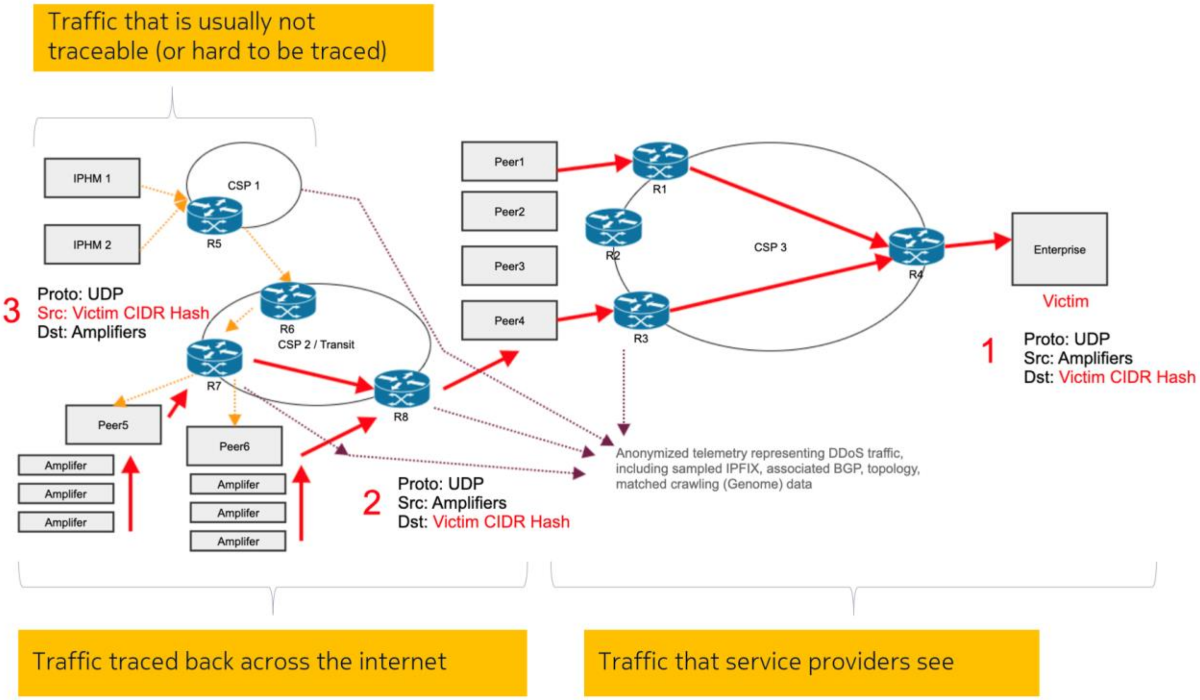


Figure 4 - Tracing DDoS back to its origins

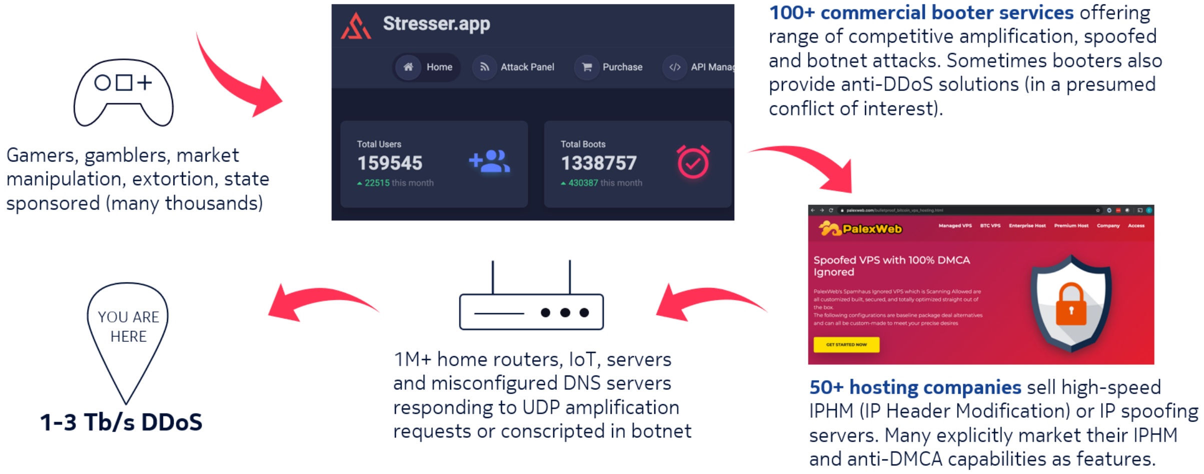
6.1. Booter and Stresser Websites: DDoS-for-Hire is a Major Source of DDoS

Contrary to the conventional wisdom that "DDoS comes from everywhere" or is unstoppable, the analysis showed that most DDoS comes from a tiny number of hosting providers that are providing a safe haven for these services.

There are approximately 100 booter websites on the Internet that will (for small fees, starting from about $50 a month) provide a menu of options to attack your "favorite" enterprise for extortion, your "favorite" gaming server, or anyone in a service provider network - a particular subscriber, a banking institution, or an online gambling company, for example. These websites advertise their services for “penetration testing,” keep no logs, and accept payments in cryptocurrencies. They have thousands of customers and rake in millions of dollars per month.

Interestingly, all booter sites typically look pretty similar. They offer a ranges of plans from $50 to $1,000 a month depending on the attacks' sophistication, frequency, and intensity. They allow IP address spoofing; a malicious party can specify that they want the attack to appear that it originates from Google, Microsoft, or Cloudflare, for example. Typically, these booter sites will offer a range of services; some of them are so bold to advertise the capability to launch an attack over 2 terabits per second (Tbps).

So, most DDoS is not coming from individual hackers; instead, the vast majority of DDoS is coming from a small number of around 100 websites, which in turn, are hosted by 50 or so hosting providers.



Spoofed VPS with 100% DMCA Ignored

PalexWeb's Spamhaus Ignored VPS which is Scanning Allowed are all customized built, secured, and totally optimized straight out of the box.
The following configurations are baseline package deal alternatives and can all be custom-made to meet your precise desires

Gamers, gamblers, market manipulation, extortion, state sponsored (many thousands)

100+ commercial booter services offering range of competitive amplification, spoofed and botnet attacks. Sometimes booters also provide anti-DDoS solutions (in a presumed conflict of interest).

50+ hosting companies sell high-speed IPHM (IP Header Modification) or IP spoofing servers. Many explicitly market their IPHM and anti-DMCA capabilities as features.

1M+ home routers, IoT, servers and misconfigured DNS servers responding to UDP amplification requests or conscripted in botnet

1-3 Tb/s DDoS

Figure 5 - DDoS-for-hire ecosystem

6.2. DDoS Hosting Providers

The research found approximately 50 globally distributed providers that allow DDoS attacks to originate from the web servers they are hosting. Several of them are in the US and in Asia, with a large number of them in Eastern Europe. They allow hosting DDoS booter services used by people who want to launch DDoS attacks. The geographical choice largely reflects jurisdictions where DDoS attacks are not a criminal offense.

These providers are not your familiar, large commercial web hosting providers. Most of the well-known and reputable hosting providers check and block spoofed IP traffic.

Providers allowing hosted websites to offer DDoS-for-hire services and found that they fall into three categories:

- The first category involves **typical content piracy servers**. They advertise their offerings on shady websites, and they are quite open that they support IP spoofing. Typically, they also offer to host "private content," mostly referring to copyrighted content, adult content, or any other type of illegal content. About one-half of the DDoS hosting providers are "out in the open" and are very explicit about their services. Typically, they operate in "gray area" jurisdictions (e.g., some countries in Eastern Europe), where they may be out of reach of legal enforcement.
- The second category is **hosting companies that try to hide.** Typically, this hiding is done in layers. Most hosting on the Internet today is resold. Often, at the bottom layer are reputable businesses. On top of these, they have two or three layers of increasingly less reputable businesses all reselling the same infrastructure and hardware. So, all of these seemingly disparate businesses belong to the same company. They are simply different brands within the same ownership structure. Websites hosted in these hosting companies do not advertise as openly. For example, they may not advertise DDoS or spoofing capabilities directly on their web pages.

Instead, they have these terms embedded in the source code (to allow internet search engines to point to them).

- A major surprise to us was the third type of hosting companies. These are **DDoS mitigation companies** – or companies that advertise they will provide DDoS mitigation services. It turned out that three of the top DDoS mitigation services are also three of the top sources of DDoS attacks on the Internet, in what we can only presume is a conflict of interest (imagine criminals and robbers offering home protection services).

6.3. DDoS Has a Threat Potential of Over 10 Tbps

The largest reported attacks so far have ranged between two to three Tbps.

It's important to keep in mind that not every attack is reported, but generally, what at least is publicly reported falls within this range of two to three terabits per second. This represents the "high watermark" for attacks today and certainly is more than enough to interrupt service to many corporate and residential customers.

The recent series of attacks against VoIP providers (September 2021), leaving many customers without voice services for days, is a clear example of the damage that can be inflicted.

What is concerning is that significant terabit-level attacks were quite infrequent until a couple of years ago. Today, we see aggregate DDoS traffic levels of that capacity every day, indicating volumetric DDoS attacks happening on the Internet almost constantly. In other words, DDoS has become a significant "traffic noise" in networks today.

One of our study's major findings was the measure of the overall aggregate capacity for DDoS attack traffic originating from these DDoS hosting companies – a combined DDoS threat potential for attacks that are yet to be launched. When we consider the global internet infrastructure that can be leveraged for DDoS (including insecure, misconfigured servers and hosts, hijacked IoT devices etc.), we see a potential for attack volumes of more than **10 Tbps**.So, there is (fortunately untapped for now) a potential to launch DDoS attacks three to 10 times larger than any attack we have seen until today.

The situation with DDoS is not getting better. These numbers have doubled from 2020 and will likely double again throughout this year or next year. Considering DDoS as an application or a traffic type, DDoS is far outpacing any other application or any other type of traffic on the Internet, really posing an existential threat to service provider networks and the Internet – if this traffic continues untamed.

The DDoS protection issue is also about economics. What is the cost to mitigate DDoS attacks (OPEX and CAPEX)? If it costs only $50 per month to launch a very damaging DDoS attack, and if such an attack is used to ask for ransom amounts of $10,000 or $20,000, this becomes an arbitrage opportunity for malicious players who may be launching tens of these attacks per month.

This results in a large and growing pronounced economic disparity between the attacker and the defender while the Internet continues to expand and the criticality of services continues to increase.

7. Drivers Behind Further DDoS Growth

Historically, most DDoS attacks have been received from the Internet across peering and transit links. These are called inbound or ingress DDoS attacks on the peering boundary.However, a growing number of attacks originate from within service providers' networks and are aimed at targets within the network or outside of it. These are called internal, east-west or egress attacks, depending on where the attack traffic enters and leaves the service provider network. A growing number of attacks are also launched from cloud infrastructures as insecure loads hosted in data centers.

Several factors are driving the increased severity and damage caused by DDoS attacks:

- Expanded threat and attack surface;
- Increased bandwidth;
- More malware and IoT bots;
- Availability of DDoS toolkits and DDoS for hire services; and
- Sophistication of new DDoS techniques.

7.1. Expanded Threat and Attack Surface

DDoS threats and attacks are becoming common and more damaging in the era of the cloud, 5G and the IoT.

The growth of all-IP networks has extended the security perimeter and expanded the threat and attack surface. The number of IP addresses has grown exponentially because of new technologies and services. Examples of these new technologies include:

- Localized content delivery networks that deliver streaming services;
- Connecting cloud points of presence (PoPs) in new metro architectures; and
- 5G adoption of edge cloud architectures such as Multi-access Edge Computing (MEC).

This expanded threat and attack surface allows threats that have been known for more than a decade, such as combined amplification and reflection attacks (e.g., DNS/NTP/TCP reflection attacks), to aim at a wider range of targets.

7.2. Increased Bandwidth

Users today have broadband connections with speeds ranging from tens of megabits to gigabits. This increased bandwidth is available to all users and devices, including malicious users and DDoS-capable devices (bots) that launch orchestrated DDoS attacks.

Combining individual high-bandwidth connections with well-known reflection, amplification and botnet techniques can result in terabit-level and high packet-intensity attacks that employ millions of packets per second.

The sheer volume of attack traffic can take a victim's system or an entire service provider network out of service or heavily degrade the target’s performance. Volumetric attacks can simultaneously affect thousands or even millions of users.

The rollout of technologies such as 5G, fiber-to-the-home (FTTH) and DOCSIS 4.0 with continued and accelerated delivery of gigabit access speeds will provide additional power for DDoS attacks.

7.3. More Malware and IoT Bots

The huge proliferation of IoT devices also includes many devices with substandard or default security that can easily be compromised.

Nokia Threat Intelligence Center's research shows that it often takes just a few minutes for an insecure IoT device with a public IP address to be compromised and to potentially be used as a remotely controlled device (bot) that can be exploited in a DDoS attack.

With high projected IoT growth, these bots (and botnets created out of them) represent a significant threat potential for increased frequency and impact of DDoS attacks.

7.4. Availability of DDoS Toolkits and DDoS for Hire

Throughout the darknet (today commonly referred to as areas of the Internet that are accessed using specific software and often act as criminal marketplaces for various illegal services), there is a growing number of websites where DDoS toolkits can be downloaded or where DDoS services can be ordered.

These toolkits and the availability of DDoS-as-a-service put the destructive power into the hands of a broader set of malicious actors, which are now able to launch DDoS attacks easily.

7.5. The Sophistication of New DDoS Techniques

Attackers use combinations of attack techniques and vectors to "shape-shift" their attacks, changing the mix and intensity of DDoS attacks over time and across different parts of the network. In addition, botnet-generated attacks are more similar to valid traffic profiles. Bots use non-spoofed source IPs, run a full tcp-stack and bypass transport-layer authentication mechanisms while establishing sessions to attack servers at the application layer.

8. Current Approaches to DDoS Security

There are a few main approaches to DDoS security used today.

One approach is based on "security in the cloud" by moving services into content delivery networks (CDNs). This approach is beneficial for a lot of content that is hosted and lives in the cloud. However, the issue is that not all content and services can be moved into the cloud.

Residential subscribers and enterprises need to connect to the cloud through links to their internet service providers (ISPs). For the content that is not cloud-based and the hundreds of thousands of subscribers and enterprises under attack that aren't in the cloud, this leaves the ISPs with two options:

One option is to buy a secure service from the upstream provider. Typically, secure services are more expensive because they require custom hardware to be deployed. Also, in many service provider environments, DDoS security is not fully automated. When operational costs of security teams that monitor the network and provide secure connectivity are added, this becomes a very expensive approach.

The **other option**, taken by many service providers, is to "own your security" and implement DDoS protection within their networks.

The great variety of DDoS techniques and continued efforts to combine/evolve them and change attack dynamics continue to pose DDoS detection and mitigation challenges. Legacy DDoS detection and mitigation approaches that used to be effective no longer work.

8.1. DDoS Detection

Historically, DDoS detection was done using tools and technologies that provide additional insight into traffic and services: dedicated network probes; inline traffic processing; and deep packet inspection (DPI) technology. Detection techniques focused on recognizing known DDoS traffic patterns or monitoring traffic volumes for irregularities.

This approach evolved to the distributed gathering of DDoS intelligence by obtaining insights from distributed data plane probes (hardware or software). This information was passed to a centralized location for further processing.

When a specific threat or attack was detected, the knowledge about that threat/attack was then disseminated to all network-wide data collection points, added to their localized knowledge bases, and used for localized DDoS mitigation.

DDoS detection techniques and approaches have included:

- DDoS signature analysis;
- Heuristic and behavioral analysis;
- Traffic anomalies monitoring; and
- Analysis of traffic packet samples.

8.2. DDoS Mitigation

Because DDoS traffic manifests itself at the network level, network-level protection using IP addresses has been the main protection mechanism. DDoS mitigation techniques has included DNS-based blocking, but DNS-based protection can be circumvented easily. More effective approaches have focused on IP address-based filtering.

Filtering techniques have included:

- Inline mitigation;
- Remotely triggered black hole (RTBH routing);
- BGP Flowspec; and
- Traffic scrubbing.

8.3. Challenges with Current DDoS Approaches

There are several challenges with the current approaches to DDoS protection:

- Provisioned protection for a select number of customers or systems;
- Inability to scale;
- Performance degradation and added latency; and
- Exorbitant cost to scale.

8.3.1. Protection for a Select Number of Customers or Systems

Traditionally, DDoS protection has been offered to only the most valuable and demanding customers or the most critical systems in the network. The concept of creating "monitored objects" dictated this strategy. The result was selective DDoS protection capabilities.

This approach left many small and medium enterprises as well as residential broadband customers without comprehensive DDoS protection. Now that "always-on" broadband connectivity is becoming a must for everyone, DDoS protection needs to extend to the whole network and all services, systems and customers.

8.3.2. Inability to Scale

The proliferation of new and distributed network architectures such as content delivery networks (CDNs and edge cloud) and the introduction of internet exchange points (IXP) have expanded the number of systems and interfaces and greatly increased connectivity. This, in turn, has raised the number of interfaces and endpoints that can be affected by DDoS.

The number of IP flows to monitor for anomalies and malicious activity has also increased exponentially.

The challenge of monitoring the wider and more dynamic network environment for a much larger and ever-evolving threat landscape makes current approaches to DDoS detection and mitigation inadequate for the cloud, 5G and IoT era.

8.3.3. Performance Degradation

BGP Flowspec provided a framework for DDoS mitigation through improved filtering and policing capabilities across BGP peering routers. However, BGP Flowspec has not been widely adopted and trusted for DDoS mitigation for two main reasons:

To use Flowspec-based mitigation, a service provider needs BGP routers supporting Flowspec and with the Flowspec-capability activated on the BGP session. Many providers have had concerns about performance degradation on routers when enabling Flowspec in addition to their core routing functions.

In addition, Flowspec announcements need to be carefully programmed on routers. Misconfiguration or the incorrect order of BGP Flowspec announcements can impact services or even other service providers.

Since the introduction of BGP Flowspec, routers have advanced enough to have minimal performance degradation when Flowspec filtering is used, and service providers have become comfortable enabling it.

Meanwhile, DDoS has evolved to the degree that only a very dynamic application of large sets of access control list (ACL) filters are now adequate for comprehensive DDoS protection.

As a result, Flowspec remains a viable DDoS mitigation approach only if DDoS detection is agile enough to create or activate large sets of ACL filters and if the routers engaged in DDoS mitigation can install and remove those ACLs filters in real time. In addition, NETCONF has become another option that can facilitate the agile activation of large ACL sets on routers.

8.3.4. Exorbitant Cost to Scale

One approach that has become a de-facto standard for mitigating DDoS attacks is the use of scrubbing centers. This approach has been severely challenged by the exponential rise in network traffic, including volumetric DDoS threats. This rise requires at least a proportional scaling of scrubbing centers. The continual addition of required capacity in the scrubbing centers is becoming cost-prohibitive.

Scrubbing also introduces additional latency because all traffic sent to scrubbing centers needs to be diverted, processed, cleaned and reinjected. Finally, scrubbing introduces more additional costs related to backhauling network traffic to scrubbers and back.

Cost efficiency is a paramount concern for service providers and cloud builders. They work to reduce the overall costs of DDoS protection, especially as they aim to drive down a key metric: the cost per protected customer or protected network infrastructure object.

9. New DDoS Protection Requirements for the Cloud, 5G and IoT Era

A new, forward-looking approach to DDoS protection is a vital aspect of overall network security.

The DDoS defense must be context-aware to protect from the new generation of threats. It needs to provide cloud-era visibility beyond IP addresses to extend the visibility into services, CDNs, websites and IoT devices. Finally, an effective defense needs to be flexible and capable of detecting new and emerging threats as they develop and evolve.

Hybrid network architectures which combine physical and virtualized network domains are proliferating. They establish more distributed network boundaries that require to be monitored for both ingress and egress DDoS.

With the increased number of endpoints that need to be protected (including customers, end devices and systems, plus network infrastructure), DDoS security must deliver improved performance, along with scalability and automation.

The DDoS threats we see today and envisage for tomorrow demand a new way of thinking about DDoS protection.

9.1. Protection for Everything and Everyone

The new global networking environment, with its ever-evolving technologies, requires a new type of protection that will encompass all customers, services and the network infrastructure. This is a major paradigm shift from the legacy approach in which DDoS protection was reserved for the most valuable and demanding customers and the most critical network entities.

In addition to protecting the hosts and servers, next-generation DDoS protection must include the ability to monitor network infrastructure within the entire network perimeter, from peering edge to centralized and distributed data centers to service edge.

DDoS protection also needs to encompass aggregation networks that serve both wireline and wireless access.

Most important, DDoS protection must extend to cover most — or all — customers and protect them from attacks coming from any direction.

Attacks aimed at service provider infrastructure are also rising, so providers need to be able to protect their entire IP infrastructure.

All systems that provide or enable connectivity services and wherever systems are located:

- In the access/edge/backbone network;
- In cloud data centers, where servers host CDNs and other services;
- In a software-defined network; and
- in hybrid network domains.

9.2. Real-Time Detection with Better Accuracy

Next-generation DDoS security must detect DDoS threats and attacks in real time with improved accuracy, resulting in a lower percentage of false positives and false negatives.

9.3. Cost-Effective, Agile, Terabit-Level Mitigation

Protection against the most damaging DDoS attacks needs to minimize the impact of DDoS traffic on target and victim systems and users.

DDoS protection also needs to limit the side effects on the network caused by network bandwidth congestion by wasteful and malicious traffic.

In addition, DDoS protection needs to scale across the whole network cost-effectively so that costs do not increase in line with bandwidth.

9.4. Automation

A forward-looking DDoS solution for millions of users, services and network infrastructure entities must be able to scale operationally. This will only be possible with a solution that allows automated mitigation of complex security policies to drive real-time surgical removal of DDoS threats and attacks.

10. Leveraging Big Data Analytics and Advanced IP Network Infrastructure to Fight DDoS

Much better and more accurate DDoS detection can be achieved by using big data network analytics with the added internet security context. Using telemetry from the network, enhanced with a detailed internet security context, a dramatic increase in DDoS detection accuracy can be achieved, resulting in much lower rates of false positives and false negatives.

DDoS mitigation is as important as detection. With major technology innovation and progress in network routing equipment, which enables accelerated evolution of capabilities and performance, the latest generations of production routers can be used as security enforcement points to instantiate granular countermeasure filters quickly and fully automated.

Leveraging the capabilities of the latest generation of advanced IP routers for security can radically change the economics of DDoS protection. This study shows cost savings of over 65% over legacy approaches.

Using existing network infrastructure (with embedded security capabilities) brings significantly lower CAPEX/OPEX than performing security enforcement with custom, inline, multi-terabit capacity. This is a major technology leap from using dedicated and expensive appliances or custom-built hardware.

11. Conclusion

We believe that a combination of advanced big data network security analytics and the latest generations of commercial routers can effectively block the most damaging, volumetric DDoS attacks.

Automation, combined with next-generation DDoS detection and mitigation, enables the network to defend itself with a high degree of performance, scale, efficacy and economic efficiency - to quickly block DDoS attacks and ensure valid traffic continues to pass through without affecting the network, its services and users.

At the same time, the internet community needs to take a firmer stance, not just with technology but with policy and cooperation of all parties – service providers, hyperscale cloud builders and probably end-users – to block DDoS and minimize DDoS effects on internet services and applications.

12. Abbreviations

ACL	access control list
BGP	Border Gateway Protocol
CDN	content delivery network
CLDAP	Connectionless Lightweight Directory Access Protocol
DDoS	distributed denial of service
DNS	Domain Name System
Flowspec	flow specification
ICMP	Internet Control Message Protocol
IoT	Internet of things
ntp	Network Time Protocol
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
VoIP	voice over IP

13. Bibliography and References

How a Dated Cyber-Attack Brought a Stock Exchange to its Knees (Bloomberg Business News, February 4, 2021)

Regulator Blasts NZ's Stock Exchange Over DDoS Meltdown (Bank Info Security, January 29, 2021)

Massive DDoS Attack Disrupts Belgium Parliament (May 5, 2021)

Nokia Deepfield research, June 2021.

Nokia Deepfield NANOG82 presentation (June 2021)

Nokia Threat Intelligence Center

Nokia Deepfield Network Intelligence Report 2020 (November 2020)

Nokia Deepfield global analysis shows most DDoS attacks originate from fewer than 50 hosting companies (June 14, 2021)

The business case for a future mode of DDoS attack mitigation

SCTE STANDARDS
JOURNAL

SCTE
a subsidiary of CableLabs®

Federated Learning for the Cable Industry

A Position Paper for Contributions in Federated Learning

Letter to the Editor prepared for SCTE by

Thomas Sandholm, Principal Architect, CableLabs, SCTE Member
3960 Freedom Circle
Santa Clara, CA 95054
t.sandholm@cablelabs.com
669-777-9042

Sayandev Mukherjee, Principal Architect, CableLabs, SCTE Member
3960 Freedom Circle
Santa Clara, CA 95054
s.mukherjee@cablelabs.com
669-777-9038

1. Introduction

Federated learning (FL) is a machine learning (ML) setting where many clients (e.g., mobile devices or whole organizations) collaboratively train a model under the orchestration of a central server (e.g., service provider), while keeping the training data decentralized (Kairouz, et al., 2021). During training the central server acts as a coordinator that receives model parameter updates from individual clients, also called learners (based on local data), and then updates a global model and broadcasts the parameter updates back to all the learners. Thus, the entire collection of clients collectively behaves like a single learner. Each learner incorporates the global model parameter updates into its local model.

There are several reasons why the federated learning model is popular:

- Only parameter updates are exchanged between the coordinator and the learner – training data remains where it is (i.e., at the local learners) and is never exchanged or shared
- The learning process is distributed and effectively parallelized across all learners, thus allowing the learners to collectively train a global model faster and with more data
- Each learner is autonomous and may decide on weights given to local vs global model parameters, thereby achieving customization to the specific conditions at the learner while still enjoying the robustness and stability advantages of a global model trained on a large dataset

Federated algorithms exist for common training optimization and data fitting algorithms such as gradient descent, allowing the same type of ML problems to be solved as with a single learner.

2. Why Should the Cable Industry Care About Federated Learning?

For a cable operator there are many domain-specific benefits to employing federated learning beyond the advantages mentioned above. Below we highlight some of them.

2.1. Network Operations Architecture

A network is, of course, distributed by its very nature, but its management is commonly centralized around a network operations center (NOC) where telemetry from all the different network equipment and end-user devices may be collected, monitored, and analyzed. The NOC would be a natural place to run an FL coordinator, and learners may be distributed across the network on routers, switches, gateways, nodes, access points or customer premises equipment (CPE) to (a) limit the volume of data that is sent to the coordinator and (b) do preprocessing and local learning close to where data is generated. The benefit of processing close to data is that decisions can be made with lower latency and in a more informed way by taking more data into account, including contextual parameters that are only known locally. Furthermore, most of the critical telemetry data used to improve the customer experience originate from the deep edge such as CPE and user equipment (UE), and federated coordination across endpoints is crucial in terms of spectrum planning and load management.

2.2. Single Purpose Devices

Network and IoT equipment are designed to perform a single task cost-effectively. In contrast to a multi-purpose personal computing device or server, processing resources such as central processing unit (CPU), memory and disk are significantly more constrained. Most of these devices can thus not deploy a full-fledged ML platform locally but would need to rely on upstream federated processing power.

2.3. Privacy and Customer Trust

All the customers' data traffic for various applications and services must flow through the operator's network. Thus, even if the operator does not do deep packet inspection, it is clear that it knows a lot about a customer just from the existence and timing of flows.

In today's environment there is increasing sensitivity on the part of operators, regulators, and customers to not only preserving the privacy of customer data but also its collection, retention, and governance. A central store of customer data in the operator's network or cloud for training ML models presents an attractive target for hackers. While it may be an option not to send customer data at all over the network, this is not always practical, i.e., if some remote control is required. Sending only partial data, or better still just ML model parameter updates aggregated across clients to a federated processor would be the next best thing from a privacy perspective, as even a breach of the coordinator would not yield usable data.

2.4. Data and Model Sharing without Privacy Risks

One common use case for (vertical) FL is coopetition, in which multiple organizations cooperate to improve their individual decision making by sharing data and models securely with privacy guarantees. In banking this is common for fraud detection. Network providers in general and cable operators in particular also often fit into this mix between competition and cooperation across industry partnerships. For example, FL allows for the existence of a marketplace for models, through which companies may share models (not data) after updating the parameters based on private training data. Partnership organizations such as CableLabs and SCTE can play key roles in enabling this ecosystem among cable operators. This would also increase the scope and value of model marketplaces like Acumos.

2.5. Low Latency Edge

Both the Wi-Fi and LTE/5G protocols make use of various timers for acknowledgements to drive the protocol, so delays in responding to these acknowledgments will not only lead to lagging behavior and sluggish user experience but would break the protocol itself, effectively blocking any form of communication. In such a scenario it is critical to make decisions fast, which means local evaluation is needed. Out-of-band communication could be utilized for coordinating with remote endpoints asynchronously, again in a federated architecture.

Many enterprises are struggling with the decision between on-prem and cloud offload. Network operators are no exception. There are many issues to consider, such as latency, privacy, cost and bandwidth capacity overhead of remote telemetry communication. This complexity leads many to opt for hybrid models in which some services reside in the cloud, others are on-prem, and yet others are deployed in edge devices to mediate between the two. This three-way split is a natural environment for deploying FL models, allowing operators to make informed and data driven tradeoffs between what gets processed locally, in the edge, or in the cloud.

2.6. Smart Pipes

Cable operators trying to fight off competition from fixed wireless access or fiber are eager to increase the value of their services beyond just being utility-providing connectivity solutions. One way to add value is by providing a superior customer experience (CX), essentially providing a "smart" pipe. An autonomously managed proactive network that predicts problems and fixes them before they occur is one example of a value-added service that earns customer loyalty and leads to revenue growth. This kind of closed loop diagnostics and corrective control requires smartness or learning behavior deep in the network as well as coordinated central consolidation. FL was designed to meet those objectives.

3. State-of-the-Art Tooling

There are a number of open-source initiatives actively developing tools to facilitate building of FL systems. The following list is not intended to be exhaustive, but informative of the current state of the art.

3.1. FATE

FATE (Federated AI Technology Enabler) implements multi-party computation (MPC) and homomorphic encryption algorithms to secure existing ML algorithms, such as logistic regression and deep learning. The original developers come from the financial sector and are thus interested in use cases such as vertical federated learning for fraud detection across banking institutions. The toolkit was accompanied by one of the first survey books on federated learning (Yang, et al., 2019).

3.2. TensorFlow Federated

TensorFlow, introduced by Google in the early days of deep learning, has become an enormously popular library for developing advanced ML algorithms with good, dedicated hardware support. TensorFlow Federated (TFF) allows for TensorFlow models to make use of federated training and evaluation through an application programming interface (API), and also allows custom federated learning algorithms to be added. One such model is the famous mobile keyboard prediction model (Yang, et al., 2018).

TFF also allows multi-machine simulations when developing new federated learning algorithms.

3.3. AWS SageMaker Neo

Although AWS as of this writing does not offer a specific product or open source tool targeted directly at federated learning, there are many building blocks that can be used to implement FL, and Amazon has also been active in improving some core FL algorithms such as federated averaging (Nandury, et al., 2021). The tool that comes closest to achieving FL is arguably AWS SageMaker Neo, which allows for models to be trained in the cloud and then executed on a wide array of platforms locally.

3.4. IBM Federated Learning

IBM offers an open-source federated learning library that implements the latest research algorithms in the field in an easy-to-use Python framework. Features include various state-of-the art fusion and fairness algorithms as well as neural network, decision tree and reinforcement learning model support.

3.5. FLoC

The new proposal from Google on how to enable interest-based advertising in Web browsers without cookies (FLoC federated learning of cohorts) has received a fair amount of pushback for solving some privacy issues with current cookies, but introducing new ones, such as sharing some behavior across sites. Nevertheless, the proposed architecture is intriguing from a wider perspective and may find other use cases. The basic idea is to train local models in each browser to capture user behavior and then map that behavior to a fixed set of cohorts, where each cohort would have at least thousands of users. When visiting a web page that has opted into FLoC, the site developers will have access to the cohort to which the visiting user belongs in order to customize the experience.

4. Gaps

Given the state-of-the-art tooling, what are some of the gaps that need further attention for a cable operator to realize federated learning effectively?

A popular mode to date has been to send all your data to a cloud provider, let them crunch it with some AutoML toolkit to come up with the best model fits for a prediction task, then train it on a distributed set of cloud resources, before deploying and evaluating the models locally.

Although this model is certainly attractive to a cable operator for a number of reasons, e.g., speed to market and being able to tap into external data science talent without having to re-train the internal workforce, it comes with some drawbacks.

With massive amounts of data collected for telemetry on data flows it quickly becomes impractical to move all the data into the cloud. Even when overcoming the mechanics of efficient data transfers, processing the data will not be free. And even though most clouds will be able to scale elastically to meet the demand, it doesn't mean your budget will! Launching hundreds of VMs to train deep neural networks is not uncommon, and the cost is exacerbated if you leave it to the cloud provider to pick the best model for you (with AutoML).

Let's say your data is not that massive and the cost of training does not outweigh the other benefits of cloud processing. The final obstacle is privacy and trust. Can sensitive data be sent to cloud providers, or does it need to be anonymized first? Do current regulations such as General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act of 1996 (HIPAA) even allow you to share PII data?

Depending on how one looks at it, FL can be both the solution to privacy protection and the villain. It is clear that FL researchers are deeply concerned about privacy, and most toolkits cited above provide some means for securing the federation exchanges. Algorithms to secure the core FL aggregation step are almost as old as FL itself (Bonawitz, et al., 2017).

However, not all the technical and engineering challenges have been completely solved. Below we will discuss some gaps we believe deserve further research.

4.1. Constrained Device Support

Many network equipment devices have limited resources due to their single-purpose design. ML toolkits typically focus on algorithm and feature support as opposed to running efficiently on constrained devices. Most cloud providers tend to take the approach of just supporting model execution on constrained devices, but if you can also train on local constrained devices many of the privacy and scalability challenges to which we alluded above may be mitigated. With the current trends in single board computer (SBC) and industrial Internet of Things (IoT) we believe it is feasible to push training as well as execution down to the deep edge of the network.

4.2. Scalability

Closely related to the issue of running on constrained devices in a resource efficient way is the issue of scalability. Many FL algorithms assume you run the local learners on a powerful edge gateway and the controllers in an infinitely elastic cloud. But what if you run the learners on constrained devices and the controllers on limited edge devices? Both FL algorithms in general and secure aggregation algorithms in particular have tended to focus on meeting privacy and ML model guarantees rather than scalability guarantees. For example, the most famous secure aggregation algorithm proposed scales quadratically in terms of communication required with the number of learners (Bonawitz, et al., 2017).

4.3. Testing

As with all infrastructure innovations, the ability to thoroughly test the new mechanisms tends to lag behind their implementation. What if you have a smart reasoning autonomous network designed to proactively increase the customer experience, and you need to test it before going into production, but you don't have the budget or time to recruit test users, or would like to do validation before letting any real users on the system? How could test tools evolve in parallel with infrastructure tools to also get smarter?

4.4. Network Operations Application

It is easy to see the benefits of various advances in ML and FL to the domain of network operations in theory, but the devil is in the details and we do not know how this plethora of algorithms and tools performs until we deploy them in a specific use case. Hence, to truly show the benefits to network operations and advance the field, practitioners need to exchange war stories from real applications and deployments in the field. Just mapping a known problem to a set of custom ML models is non-trivial, even with brute force algorithms such as AutoML and deep neural networks training themselves. This mapping is thus often done through a painstakingly slow collaboration between business stakeholders, data scientists and data engineers. At some point we need to support these deployed systems and when things go wrong, we need to have a clear idea of how to fix them. The good news is that once this process has run its course in at least one loop, the next iteration will be less painful. Therefore, we again believe that sharing use cases of successful application of ML and FL in network operations is the ultimate way of advancing the field.

5. Conclusions

At CableLabs we have started exploring some of these challenges. In (Sandholm, et al., 2021) we address the current complexity and scalability of secure aggregation algorithms to incentivize cross-organizational data sharing. We propose a novel algorithm that is targeted at constrained devices and designed to scale linearly with the number of learners while preserving strong privacy guarantees using standard public key infrastructure (PKI) encryption and a logical ring topology. In (Sandholm & Mukherjee, 2021) we explore a new way of generating user traces realistically in Wi-Fi and cellular wireless testbeds and simulations. Our proposed tool is targeted at testing smart network services by using context-aware generative adversarial networks to simulate user traffic that adheres to statistical properties of a real trace.

We believe there are great opportunities ahead both in applying tools like these and in exploring new tools to ensure successful adoption of FL and ML within an organization as well as across organizations in the cable industry.

References

Bonawitz, K. et al., 2017. *Practical secure aggregation for privacy-preserving machine learning.* s.l., s.n., p. 1175–1191.

Kairouz, P. et al., 2021. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977.*

Nandury, K., Mohan, A. & Weber, F., 2021. *Cross-Silo Federated Training in the Cloud with Diversity Scaling and Semi-Supervised Learning.* s.l., s.n., p. 3085–3089.

Sandholm, T., Huberman, B., Hamzeh, B. & Clearwater, S., 2019. Learning to wait: Wi-fi contention control using load-based predictions. *arXiv preprint arXiv:1912.06747.*

Sandholm, T. & Mukherjee, S., 2021. MASS: Mobile Autonomous Station Simulation. *arXiv preprint arXiv:2111.09161.*

Sandholm, T., Mukherjee, S. & Huberman, B. A., 2021. SAFE: Secure Aggregation with Failover and Encryption. *arXiv preprint arXiv:2108.05475.*

Yang, Q. et al., 2019. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning,* Volume 13, p. 1–207.

Yang, T. et al., 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903.*

Resources

Acumos	https://marketplace.acumos.org
AutoML	https://www.automl.org/automl
AWS SageMaker Neo	https://aws.amazon.com/sagemaker/neo
FATE	https://fate.fedai.org
FLoC	https://github.com/WICG/floc
FLoC Critique	https://www.eff.org/deeplinks/2021/03/googles-floc-terrible-idea
IBM Federated Learning	https://github.com/IBM/federated-learning-lib
TensorFlow Federated	https://www.tensorflow.org/federated

Extended Upstream

A Pragmatic Approach

A Technical Paper prepared for SCTE by

Steve Condra, Senior Director, Product Management, Teleste Intercept, SCTE Member
440 Forsgate Drive
Cranbury, NJ 08512
steven.condra@telesteintercept.com
+16786418099

Kari Mäki, R&D Manager, Teleste Corporation, SCTE Member
Telestenkatu 1
Kaarina, 20660
kari.maki@teleste.com
+358405540506

Niko Suo-Heikki, Product Manager, Teleste Corporation
Telestenkatu 1
Kaarina, 20660
niko.suo-heikki@teleste.com
+358405930193

Arttu Purmonen, Vice President, System Marketing, Teleste Corporation
Telestenkatu 1
Kaarina, 20660
arttu.purmonen@teleste.com
+358405562942

Table of Contents

List of Figures

List of Tables

1. Introduction

The cable industry relies on Data Over Cable Service Interface Specification (DOCSIS) 4.0 technology to serve the needs of future applications requiring multi-gigabit symmetric services over Hybrid Fiber-Coax (HFC) networks. The capacity increase of downstream due to extended frequencies has been studied both in theory and practice. However, the increase of upstream capacity on the grounds of higher split frequencies has been analyzed theoretically, but practical considerations are conspicuously absent. The cable industry is keenly aware of the importance of ultrahigh upstream speeds due to the competitive offers from fiber to the home (FTTH) providers, motivating cable to invest in next generation DOCSIS.

Although fiber deep architecture (N+0, Node plus Zero Amplifiers) exist in cable networks, the construction costs can make them impractical in many cases. Therefore, we need to understand how amplifier cascades behave when frequencies are extended. In this paper, we focus on the usefulness of a principle called “return follows forward” (RFF), that has been developed widely deployed with European operators who have deployed amplifiers supporting 204 MHz return path. While DOCSIS 3.1 supports a 204 MHz return path, DOCSIS 4.0 recognized that more upstream bandwidth will be required in the future.   In this paper, we demonstrate how a 396 MHz return path (one of the DOCSIS 4.0 standardized frequency return paths) is even more sensitive than 204 MHz return path. We will show the relative variation that can be expected and show one method to compensate for that RF signal variation. The attenuation of coaxial cables changes as a function of temperature and as a function of frequency. Thus, it is beneficial or even mandatory to manage these multifaceted changes in order to reach maximum upstream throughput over aerial coaxial cables in low and high temperatures and cope with the fluctuation in temperatures.

Our technical paper is arranged in the following way. First, we introduce the RFF principle, often known as continuous return path automatic level and slope control (ALSC). Second, we calculate how a cascade consisting of three amplifiers and aerial coaxial cables behaves in low (-4°F), reference (77°F), and high (140°F) temperatures at frequencies below 204 MHz and at frequencies below 396 MHz Third, we report real return path measurement results of the cascade comprising three cascaded amplifiers and coaxial cables that we modeled theoretically in the earlier section. Using these measurement results, we then demonstrate what happens when the return path is extended to 396 MHz Our results illustrate how amplifier cascades supporting ultrahigh split frequencies can achieve remarkable upstream throughput when the continuous return path ALSC supports interplay of DOCSIS remote physical device (RPD) and cable modem (CM). As our results describe how the physical layer works, they are valid when the RPD is replaced by a remote mac device (RMD) and can be applied even in traditional cable networks where a centralized cable modem termination system (CMTS) communicates with cable modems (CMs). Last, we report limitations in our approach and discuss practical concerns that industry experts must consider in their journey toward multi-gigabit symmetric services over HFC networks.

2. From Theory to Practice

Rather than unknowingly replicating and renaming history, we acknowledge the patent application US5724344 filed in 1996 called “Amplifier using a single forward pilot signal to control forward and return automatic slope circuits therein” (Beck, 1996). Although cable television technology has developed dramatically in the past 25 years, the idea to adjust the return path automatically is old. Despite the age of the idea, the need for modern return path ALSC today is more topical than ever as return path frequencies are extended.

2.1. Return Follows Forward Principle

Figure 1 shows three cascaded amplifiers and the attenuation of coaxial cabling at 396 MHz During or even before installation, the amplifiers have been configured to know what downstream signal levels give the best performance at the output port. They can automatically regulate electrically adjustable pads and equalizers to provide the right gain and slope settings to compensate for unideal signal levels at the input. These settings can be used to calculate ideal upstream gain and slope settings to compensate for the attenuation and slope of coaxial cables between the amplifier and the device that feeds the amplifier in the downstream. In Figure 1, the feeding device is the network analyzer feeding the first amplifier, the first amplifier feeding the second, and the second feeding the third.

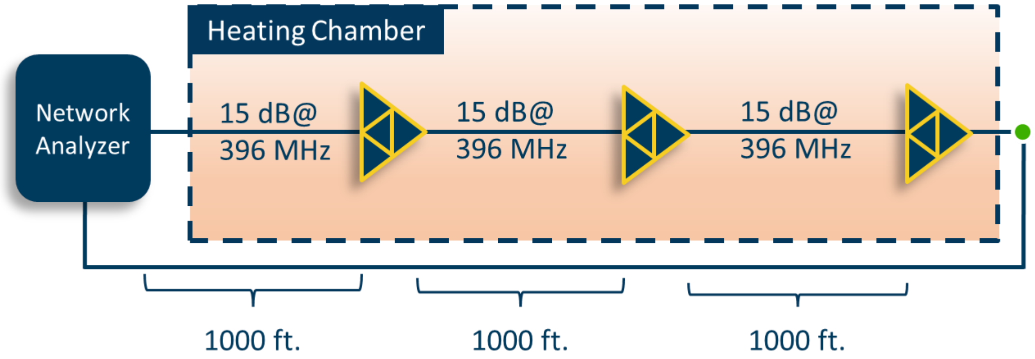


Figure 1 - Three Cascaded Amplifiers and Cables in a Heating Chamber

2.2. Theoretical Modelling

In our calculations we used a typical trunk cable with 15 dB attenuation at 396 MHz and temperature drift of 0.11%/°F/dB (0.2%/°C/dB). The attenuation of the 1,000-ft cable is shown in Figure 2, where the green line shows attenuation at 77°F (25°C), blue at -4°F (-20°C), and red at 140°F (60°C).

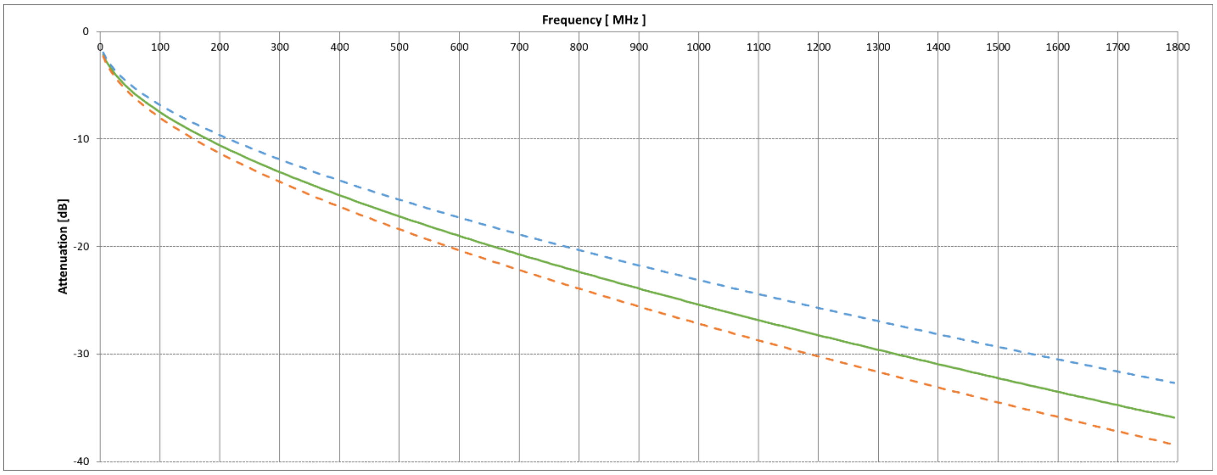


Figure 2 - The Attenuation of Coaxial Cable in Different Temperatures

The difference between the attenuations in five different frequency points at low (-4°F) and high (140°F) temperatures is shown in Table 1. In Figure 1, we have three amplifiers. To understand how the frequency response of their cascade differs in low and high temperatures, we multiplied the results in Table 1 by three. This is possible because the coaxial cables have equal length. Results are shown in Table 2.

Table 1 - Attenuation Difference in the 1,000-ft Cable at Different Frequencies

Frequency	50 MHz	100 MHz	150 MHz	204 MHz	396 MHz
Loss at -4°F	4.92 dB	7.02 dB	8.45 dB	9.69 dB	13.75 dB
Loss at 140°F	5.79 dB	8.25 dB	9.93 dB	11.39 dB	16.17 dB
Difference	0.87 dB	1.23 dB	1.48 dB	1.70 dB	2.42 dB

Table 2 - Upstream Frequency Response Difference in the Cascade

50 MHz	100 MHz	150 MHz	204 MHz	396 MHz
2.61 dB	3.69 dB	4.44 dB	5.10 dB	7.26 dB

The values in Table 2 are interpreted in the following way. In theory, the upstream frequency response of three cascaded amplifiers and the cabling between them fluctuates as a function of temperature if these amplifiers do not contain upstream ALSC. This assumption is valid unless some of the amplifiers receive signal levels so high that they become overloaded, causing distortions that decrease modulation error ratio (MER). Similarly, if the amplifiers receive very low signal levels, the upstream noise decreases MER. This worse MER is a more serious consequence than the fluctuating signal level. DOCSIS CMs can change their upstream output level, but can they compensate for these fluctuations so that modems are not forced to use a less advanced but a more robust modulation method? Unfortunately, this is not always the case; many times, the signal fluctuations exceed the adjustment capability of the cable modem and result in signal levels at the amplifier that can create excessive noise figure or excessive RF level thus creating distortions. In the next chapter, we move from theory to practice and report real measurement results of the cascade that we analyzed theoretically in this chapter.

2.3. Measurements

We built the cascade shown in Figure 1. In Figure 3, the used output and input levels are explained.

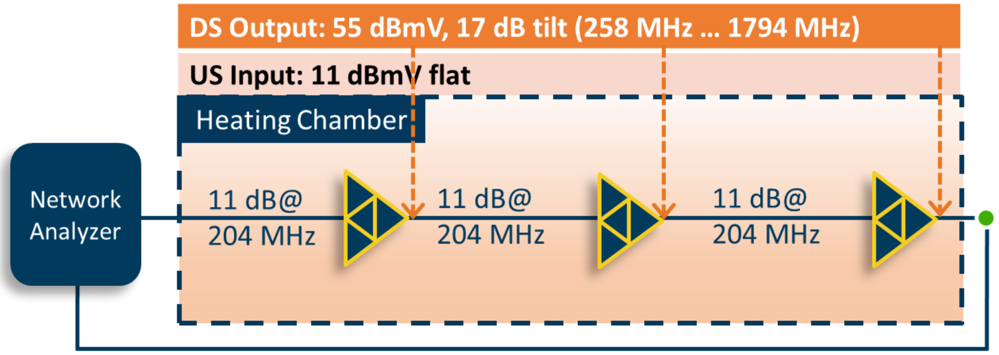


Figure 3 - The Real Test Setup and Upstream Frequency Response

Figure 3 above depicts the attenuation of the 1,000-ft coaxial cables at 204 MHz used in our test setup.. The first test was run without upstream ALSC and resulted in the frequency response shown in Figure 4. Although upstream ALSC was switched off, the amplifiers did precise internal temperature compensation for amplifier gain versus temperature variation, as we wanted only to see the effect of pure cabling. The arrows and corresponding values in dB illustrate theoretical calculations. While the theoretical values and measured values are not equal, the trend of the theoretical calculations are undeniable. The difference between theoretical and actual at frequencies near the band edge of the diplexer being utilized is considerable, but that is an artifact of the 204/258 MHz diplexers in the amplifiers. When, using 396/492 MHz diplexers the theoretical and actual results at frequencies close to 200 MHz will be very similar, however those will show similar differences closer to their cut-off frequency around 400 MHz, again this is an artifact of being close to the band edge of the diplex filter.

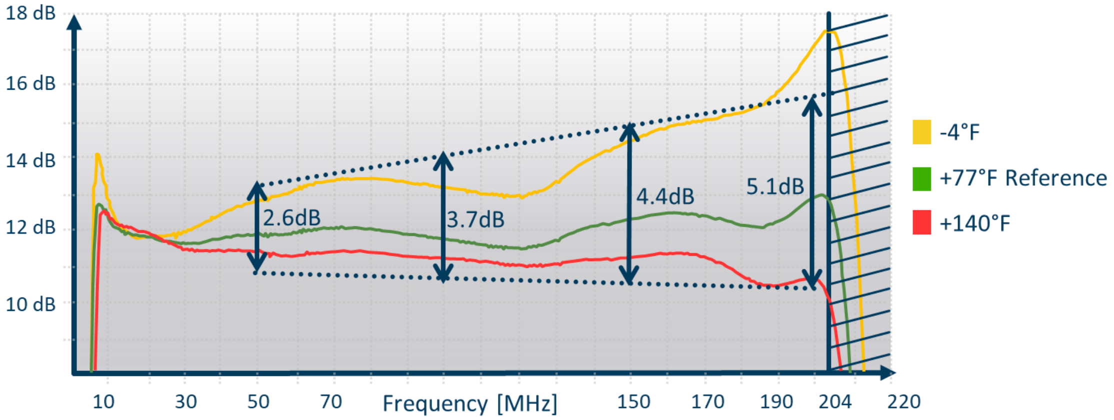


Figure 4 - Frequency Response of the Cascade with Upstream ALSC off

In the second test, the upstream ALSC was switched on, and there was the precise internal temperature compensation of the amplifiers. The results are shown in Figure 5. The graphs demonstrate how the amplifiers remove all frequency response fluctuations caused by low and high temperatures. The diplexers again impact the frequency response close to the cut-off frequency.

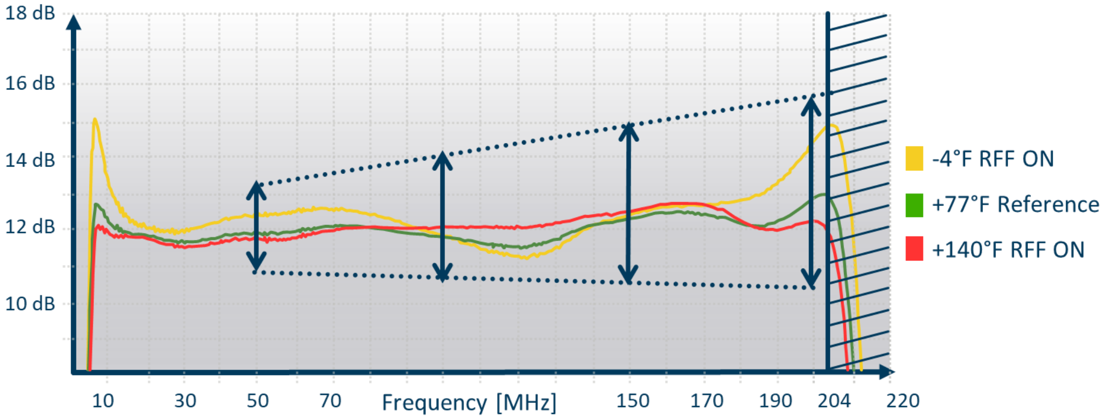


Figure 5 - Frequency Response of the Cascade with Return Path ALSC on (RFF on)

3. Return path extension to 396 MHz

3.1. Considerations

We showed how the theoretical calculations can be used to estimate the impact of outdoor temperature changes to the frequency response of the return path comprising amplifiers and the cabling between them. For the three cascaded cable spans we considered, each with a cable loss of 15 dB at 400 MHz, the variation of level at 400 MHz due to a moderate temperature change from -4 degrees F to +140 degrees F was 7.3 dB. In practice, most networks have longer cables, so the impact would be greater signal lever variation. Our calculations also disregarded the effect of temperature change on the last drop cables and the hardline coaxial cable between the last amplifier and the tap that feeds the drop cables, although in practice, they increase the impact. The 7.3 dB RF signal variation is equivalent to 490-ft of coaxial cable.

Outdoor temperature is not the only reason cable modems must change their upstream signal level; the 490-ft length difference between cold and warm ambient temperatures already causes major issues. Every cable network in practice includes cable modems that are at or near their maximum output level, in addition to amplifiers that do not tolerate extremely high return path signal levels. So, in practice some amplifiers would see a signal level that overloads them causing nonlinear behavior that decreases MER. Alternatively, in some networks, the amplifiers would see very low signal levels such that noise figure in the amplifier causes the MER to decrease. The last active device in the network—the amplifier that feeds cable modems through drop cables—is forced to tolerate signal level fluctuations caused by the network above and below it. Accordingly, it makes sense to compensate for signal level fluctuations in every amplifier and leave cable modem upstream signal level margins (if they exist) to cope with changes caused by the last drop itself.

3.2. Limitations

In our calculations, we assumed that outdoor cables are predisposed to 144°F temperature changes in cold and warm seasons. This is not true in every case, as -4°F is extremely rare in some locations. However, in these same locations, a black coaxial cable exposed to direct sunlight can easily reach 144°F, and hot temperature can cause higher attenuation requiring higher signal levels that cable modems are not necessarily capable of transmitting. We also assumed three cascaded amplifiers having a 1,000-ft distance from each other. While some operators may claim that future amplifier cascades are shorter, we argue that virtually every cable operator has longer cascaded amplifier in parts of their network that they may not like to admit. It is also rational to ask: Does automatically adjusting amplifiers equipped with upstream ALSC interfere with automatically adjusting DOCSIS cable modems? Can these two control mechanisms oscillate? The answer is no. The return path ALSC is adjusted according to the forward path attenuation and the corresponding forward path signal levels.

4. Conclusions

While existing cable networks may work without the upstream ALSC, the gradual move toward high-split and ultra-high-split networks urges cable network operators to reconsider traditional approaches. In our view, the 204 MHz return path can benefit from the upstream ALSC, whereas the 396 MHz return path demands it. As our results show, for operators to offer high upstream capacity and stay competitive, every decibel counts. The decibels that the upstream ALSC provides are neither difficult nor expensive to achieve.

5. Abbreviations and Definitions

5.1. Abbreviations

ALSC	automatic level and slope control
CM	cable modem
CMTS	cable modem termination system
dB	decibel
dBmV	decibel millivolt
DOCSIS	data over cable service interface specification
DS	downstream
FTTH	fiber to the home
HFC	hybrid fiber-coax
Hz	hertz
MAC	media access control
MER	modulation error ratio
MHz	Megahertz
N+0	node plus zero
RFF	return follows forward
RMD	remote mac device
RPD	remote physical device
SCTE	Society of Cable Telecommunications Engineers
US	upstream

5.2. Definitions

Cable modem	A modulator-demodulator at the subscriber premises intended for use in conveying data communications on a cable television system.
Cable modem termination system	A device located at the cable television system headend or distribution hub, which provides complementary functionality to the cable modems to enable data connectivity to a wide-area network.
Decibel	Ratio of two power levels expressed mathematically as $dB = 10\log_{10}(P1/P2)$.
Decibel millivolt	Unit of RF power expressed in terms of voltage, defined as decibels relative to 1 millivolt, where 1 millivolt equals 13.33 nanowatts in a 75 ohm impedance. Mathematically, dBmV = 20log10(value in mV/1 mV).
DOCSIS	Data-Over-Cable Service Interface Specifications. A group of specifications that defines interoperability between cable modem termination systems and cable modems.
Downstream	The direction of RF signal transmission from headend or hub site to subscriber. Also called forward.
Drop	The coaxial cable and related hardware that connects a residence or other service location to a tap or to an amplifier. Also called drop cable or subscriber drop.
Forward	See downstream

Frequency response	A complex quantity describing the flatness of a channel or specified frequency range, and which has two components: amplitude (magnitude)-versus-frequency, and phase-versus-frequency.
Hertz	A unit of frequency equivalent to one cycle per second.
Hybrid fiber-coax	A broadband bidirectional shared-media transmission system or network architecture using optical fibers between the headend and fiber nodes, and coaxial cable distribution from the fiber nodes to the subscriber locations.
Layer	One of seven subdivisions of the Open System Interconnection reference model.
Media access control	A sublayer of the Open Systems Interconnection model's data link layer (Layer 2), which manages access to shared media such as the Open Systems Interconnection model's physical layer (Layer 1).
Megahertz	One million hertz
Modulation error ratio	The ratio of average signal constellation power to average constellation error power – that is, digital complex baseband signal-to-noise ratio – expressed in decibels.
Millivolt	One thousandth of a volt
Nonlinear distortion	A class of distortions caused by a combination of small signal nonlinearities in active devices and by signal compression that occurs as RF output levels reach the active device's saturation point.
Return	See upstream
Subscriber	End user or customer connected to a cable network.
Upstream	The direction of RF signal transmission from subscriber to headend or hub site. Also called return.

6. Bibliography and References

(1) Beck William Federick. (1996). Amplifier using a single forward pilot signal to control forward and return automatic slope circuits therein. Retrieved November 20, 2021, from https://patents.google.com/patent/US5724344

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

AI/ML Industry Best Practices Series

AI Powered Customer Care

A Technical Paper prepared for SCTE by

Utpal Mangla, Vice President, IBM

Luca Marchi, Lead Architect, IBM

SCTE STANDARDS JOURNAL

SCTE a subsidiary of CableLabs®

Table of Contents

List of Figures

1. AI Powered Customer Care – Cogntive Care

1.1. Industry Challanges

The cable TV industry in North America is facing intense television service competition from a number of sources, especially over-the-top (OTT) providers. OTT providers have leveraged advanced technologies, in particular artificial intelligence, to disrupt the status quo and provide a new customer experience, based on digital, omnichannel and personalized engagement. In this paper we will cover how the same technologies can be leveraged by incumbent cable companied to provide a differentiating customer experience and defend their growth.

Defending and driving growth is an industry imperative for every player, in order to maximize viewership, subscriptions and advertising revenues.

Growth can be defended and secured through the following strategies:

- Exploding video and digital content – enabling secure storage and delivery, managing the rights, and providing personalized advertising and *aligned customer care;*
- Fending off threats to customer base and revenue from streaming services, social media giants, tech-platform companies, and post-video services; and
- Shifting to distribution and revenue models supported by an IT infrastructure and network that is resilient, scalable and can adapt quickly.

Growth should not be only defended but also optimized in two ways:

1. Leveraging data and advanced analytics to *Grow share* of the $129B (and growing) US Digital Advertising market through personalized advertising, particularly around OTT.
2. Managing video customer churn by *providing an outstanding customer experience at every touchpoint*

In the following paragraphs we will demonstrate how a key cable industry challenge like customer experience can be addressed with artificial intelligence and machine learning. In particular, we will explore the best practices to make the experience of the customer effortless and to effectively infuse artificial intelligence into the key business processes and client journeys.

1.2. The Issue of Customer Experience

As customer experience becomes one of the key success factors of any business, cable, satellite and internet service providers consistently rank lowest in Net Promoter Score (NPS), with scores ranging from -10 to +10. Providing the ultimate customer experience and consequently growing the NPS is now on the agenda of every cable provider. In fact, in most sectors, industry leaders in NPS outgrow competitors on average by a factor greater than two times. In the following section we will discuss how to deliver this exceptional experience, while keeping in mind that the strategic and tactical choices to be made are numerous, and the playing field is crowded with tech giants and niche, high growth specialty companies. An enhanced customer experience that meets the higher digital expectations needs to be consistent across all devices, channels and interactions.

1.3. Achievable Improvements in Cable TV Business

Cable TV providers that reinvented and redesigned the customer experience leveraging new technologies, in particular artificial intelligence, have been able to achieve outstanding results:

- 15-20 point increase in NPS.
- 20-40% reduction of contact center OpEx.
- Up to 10% increase in revenue.

These results were obtained by placing the customer at the center of the design process and following the best practices:

- Delight your customers with personalized care across channels.
- Deliver transformed experiences.
- Leverage AI and data for richer personalization.
- Predict the next best action to be recommended to the customer.
- Integrate and scale artificial intelligence across the operator
- Co-create and co-execute in a “virtual garage.”
- Sustain continuous improvement with a center of excellence (COE).

The impact of deploying artificial intelligence to additional call center KPI’s can be seen as:

- Reducing operating cost: ~ 50%
- Reducing customer effort: ~60%
- Reducing average handle Time: ~30%
- Increasing customer satisfaction (CSAT): ~35%
- Increasing NPS: ~20 points
- Increasing Revenue: ~15%
- Reducing inconsistent responses: ~50%

1.4. Cognitive Care

Cognitive care defines the application of artificial intelligence to an improved and comprehensive customer experience. Cognitive care is more than just communication marketing and chatbots – it’s about how consumers interact with your organization every day. It’s about how they experience companies’ services; how companies communicate and interact; and how technology and humans collaborate to reduce customer effort and improve customer experience. It’s about how companies support and ***connect*** everything together– all enhanced by data, analytics, personalization, trust, and consistency across every touchpoint.

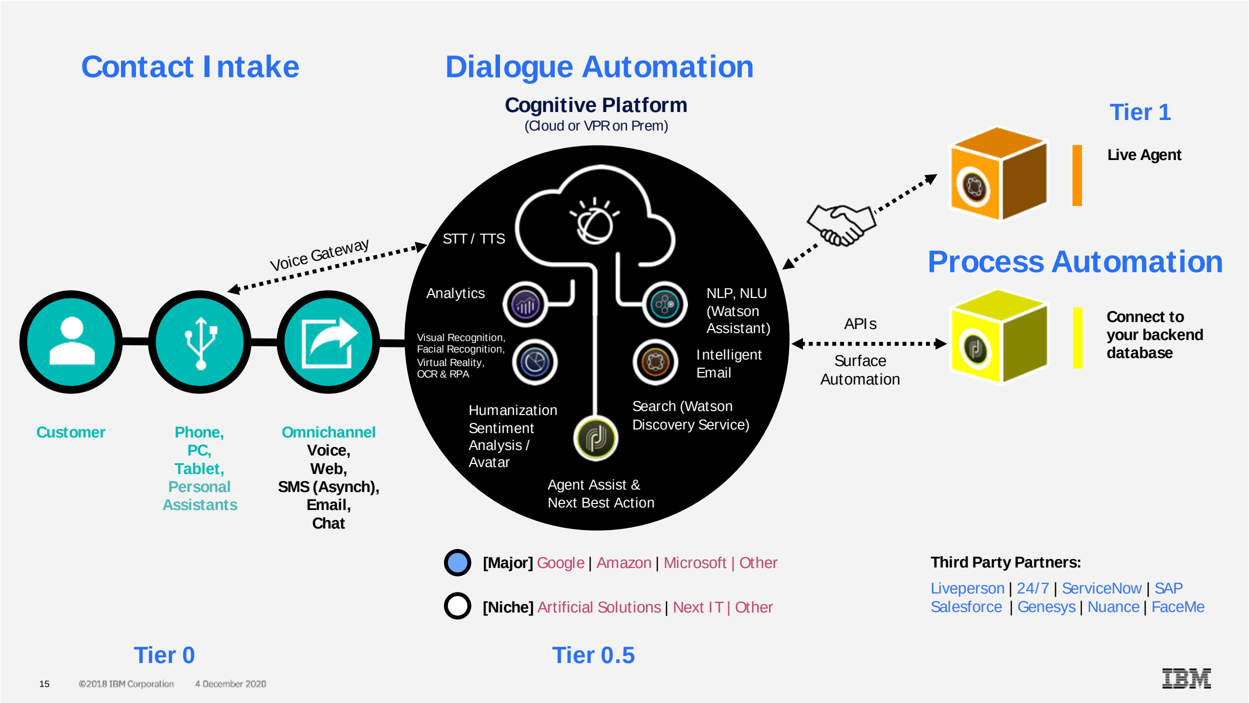

Contact Intake
Dialogue Automation
Cognitive Platform
(Cloud or VPR on Prem)
Tier 1
Live Agent
Process Automation
Connect to your backend database
Voice Gateway
STT / TTS
Analytics
Visual Recognition, Facial Recognition, Virtual Reality, OCR & RPA
NLP, NLU (Watson Assistant)
Intelligent Email
APIs
Surface Automation
Humanization Sentiment Analysis / Avatar
Search (Watson Discovery Service)
Agent Assist & Next Best Action
Customer
Phone, PC, Tablet, Personal Assistants
Omnichannel Voice, Web, SMS (Asynch), Email, Chat
[Major] Google | Amazon | Microsoft | Other
[Niche] Artificial Solutions | Next IT | Other
Third Party Partners:
Liveperson | 24/7 | ServiceNow | SAP
Salesforce | Genesys | Nuance | FaceMe
Tier 0
Tier 0.5
15
©2018 IBM Corporation
4 December 2020
IBM

Figure 1 - Cognitive Care Framework

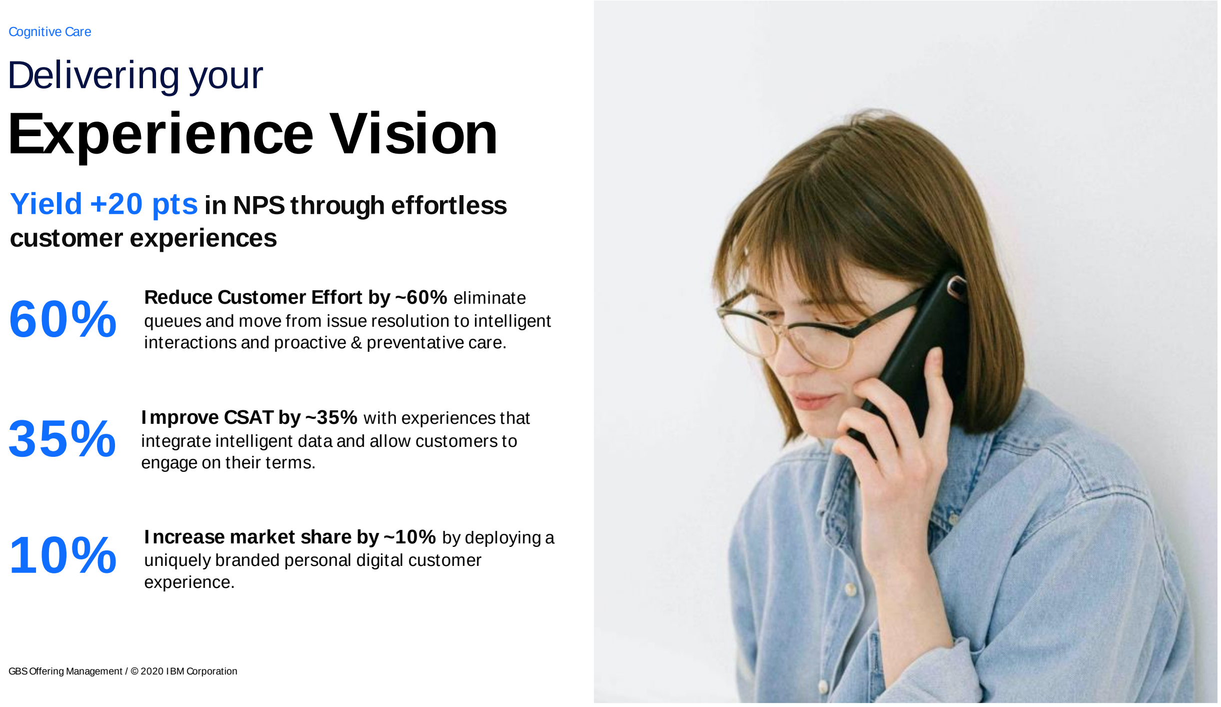
Figure 1 shows a high-level representation of a cognitive care system. The starting point is the customer who interacts with the corporation in an omnichannel fashion through different devices. At the core of the system is a cognitive platform. The cognitive platform owns AI and analytics capabilities that enable interaction with the customer in natural language (text or voice), the contextualization of the interaction, the use of all data regarding the client and the issue, the communication of a resolution or next best action, and finally the support to human agents.

Because the cognitive platform alone cannot resolve all issues, integration with systems such as BSS (billing, CRM etc.) and OSS (network, field service management etc.) is fundamental in the automation of the issue resolution. The goal of cognitive care is to provide the best experience possible according to customers' expectations and company values. The experience can be zero-touch and fully automated, as when customers interact only with software, or the experience can be hybrid, with human agents supporting customers with the help of AI-enabled tools. Finally, cognitive platforms can be deployed on a hybrid cloud, mixing private and on-prem components with public cloud services, in order to maximize security, privacy, and scalability.

1.5. Best Practice #1: Enable effortless customer experiences

The first step for the delivery of an excellent customer experience through cognitive care is to design and build **an intelligent experience.** An intelligent experience implies interacting in natural language and proactively with the customer, leveraging context information about the customer herself and the issue she is facing.

The capabilities of a cognitive platform that support such interaction are the tools that enable the design of intelligent dialogue and workflows, as well as the machine learning models that provide customer care agents or the customers directly with real time recommendations and next best actions. The role of artificial intelligence is to support a set of capabilities that automate dialogue and the processes. Figure 2 shows the impact of an intelligent experience on customer care results.



Cognitive Care

Delivering your

Experience Vision

Yield +20 pts in NPS through effortless customer experiences

60% **Reduce Customer Effort by ~60%** eliminate queues and move from issue resolution to intelligent interactions and proactive & preventative care.

35% **Improve CSAT by ~35%** with experiences that integrate intelligent data and allow customers to engage on their terms.

10% **Increase market share by ~10%** by deploying a uniquely branded personal digital customer experience.

GBS Offering Management / © 2020 IBM Corporation

Figure 2 - Enable Effortless Customer Experience

A few examples of the features that a cognitive care platform needs to have to meet the customer's expectations are:

- Sentiment analysis – Does your virtual agent sense customer frustration and generate a proactive call from a human to the customer when the virtual agent does not resolve the issue?
- Responsiveness – Does your virtual agent take seconds, rather than minutes, to reply to a customer?
- Customer Monitoring – Are you tracking if your customer has contacted multiple company centers and if the situation was resolved?
- Cross-functional workflows – Is your customer's chatbot log being passed to the sales center or tech support so the customer is not repeating themself?
- Personal touch – Do you follow-up with a "how did we do" email?
- Experience management – Do you follow-up with customers who give you a poor service review?

1.6. Best Practice #2: Infuse Intelligence

As mentioned in the previous section, the cognitive care experience is driven by AI and automation. The two technologies have the task of solving business processes and improving process workflows, and have an impact on outcomes when they keep the customer at the center of the strategy.

Infusing intelligence means interacting with a customer with deep context, as explained before, and also adding industry-specific knowledge to a company customer care center. In fact, artificial intelligence has the ability to harvest knowledge from subject matter experts and leverage it to solve clients' problems. This is done using training tools that allow experts to train machine learning models and transfer their own knowledge. For example, a customer care agent can provide feedback and correct the answers of a virtual agent or a field technician can tell from a visual recognition app whether or not a certain piece of network equipment is damaged.

To make sure that the right level of industry knowledge is infused in a platform, companies need to involve their experts in the platform design and training and to leverage pre-integrated, industry-specific solutions and reusable assets.

The advantages of this approach include:

- **Scaling rapidly across the enterprise** by leveraging prebuilt assets to jump start the transformation journey.
- **Reducing Time to Value** by deploying in weeks rather than months with proven ~6mo ROI.
- **Leveraging Intelligent Data and AI Capabilities** to create personalized intelligent experiences.
- **Futuring-proof technology decisions** by using a vendor agnostic platform.

Examples of pre-integrated and pre-developed industry assets are:

- Lists of industry-specific use cases to accelerate the training of the virtual agents
- Industry-specific language models to accelerate the analysis and insights extraction from user generated data.
- Use-case accelerators with pre-built workflows and interfaces that orchestrate specific use cases (B2B or B2B services order management; churn prevention; troubleshooting; human agent assist; proactive care etc.)

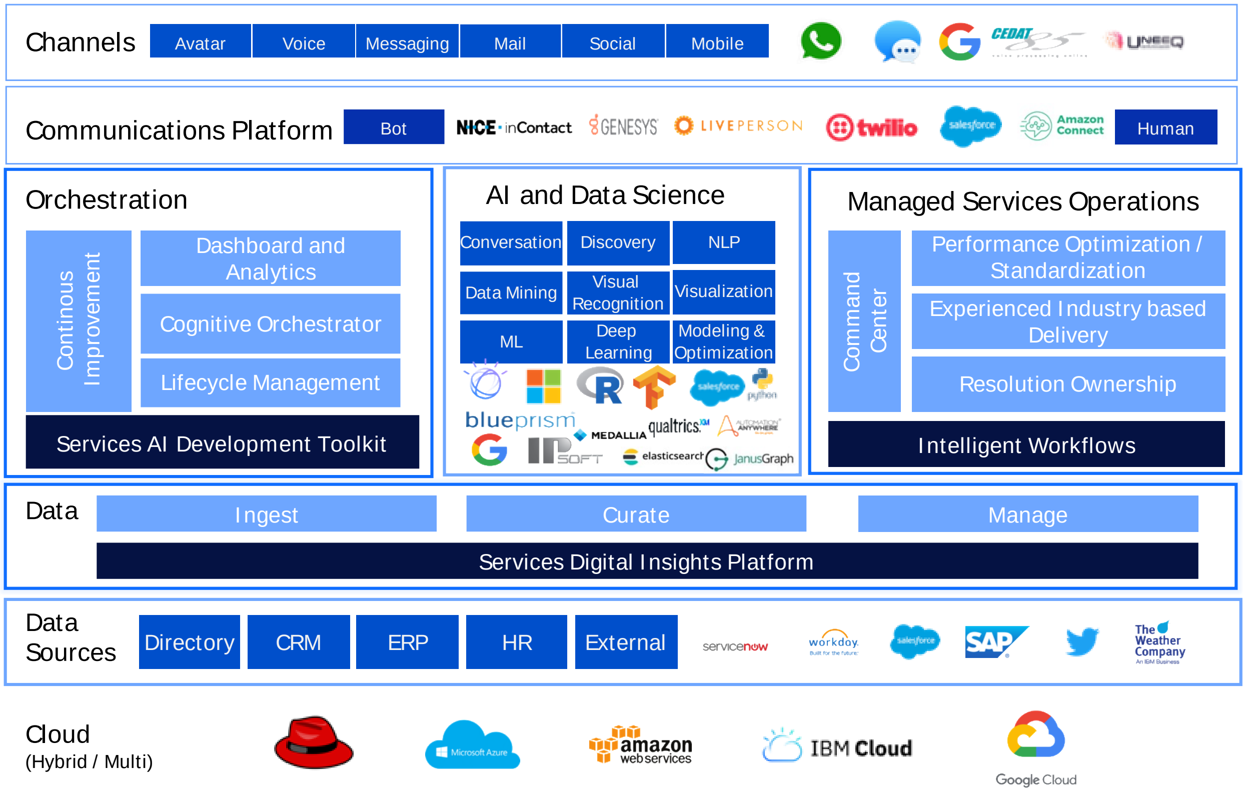


Figure 3 - Cognitive Care Architecture

Cognitive care is a multichannel approach: customers can reach the contact center through voice or text, phone or PC, messaging or social media. The optimal channel mix depends on the communication strategy of the provider. Cognitive care channels integrate with established communication/customer care platforms like Liveperson, Genesys etc. Connectors and integrators to these platforms are the industry-specific assets mentioned above that accelerate the deployment of a cognitive care platform. The core of the architecture leverages AI, data science and an orchestration layer. AI and data science are the services or APIs that bring machine learning, natural language processing, and deep learning capabilities to the platform. In the market there are many different providers, spanning from point solutions to large cloud vendors.

AI and data services can be grouped in major categories. The most used services in the context of cognitive customer care are the natural language processing ones (NLP). A good example of NPL services are conversation services: those are APIs able to understand written human language and tools to build a dialog or a conversation with a human. Usually these services focus on “short tail” interactions: high volume, predictable customer requests, that are similar but not limited to frequently asked questions (FAQ). Discovery services, on the other hand, focus on the “long tail” of requests: less frequent and more complex interaction. To answer these requests, an AI system needs to read and understand large portions of text, like a product manual or contract terms and feed the answer back into the dialog. Finally, natural language understanding services (NLU) provide real time analytics of human generated language, giving insights on tone, sentiment and text structure (entities, keywords, relations etc.). NLU services provide both out of the box models and the ability to customize your own models, for example creating dictionaries and ontologies of your company knowledge base. Another set of services related to natural language are speech services: speech to text (turning human language in audio format into text) and text to speech (vice versa, turning text into a human voice). Again, these services are available both as out of the box or with customization capabilities, for instance to teach a text to speech model how to read the labels of a company's products. The last services leveraging unstructured data are the visual recognition services that analyze pictures and videos. An example of an application in the cognitive care context is the ability to read screenshots or pictures that subscribers send to a virtual agent. The agent is then able to recognize invoices, contracts and other company communications and start the proper support workflow.

Natural language processing, speech and visual recognition services are the most used in cognitive care since they analyze and interact with human generated content, supporting customer engagement and experience. There are other more analytical services that use structured and unstructured data like machine learning, deep learning, modeling, and optimization to support the customer engagement. For example, companies leverage client history data, product catalogs, and customer care logs to develop next best action machine learning models and provide the best possible offer to their subscribers.

The services are essential but not sufficient. They need to be coordinated by an orchestration layer that manages the rules of the system: providing insights and analytics; triggering different services; switching between channels; handing off to humans; and facilitating the continuous improvement of the platform.

As with any solution based on AI, the main ingredient of cognitive care is data. The AI services feed on data; this may be internal or external, coming from the network or the BSS or generated by the client. The cognitive care workflows are only as intelligent as the quality of the underlying data allows. Companies deploying an intelligent customer care approach need to map the data currently available against the target data architecture required to deliver an excellent customer experience; they then need to work to fill the gap between the current and the desired architecture.

Finally, cognitive care plugs into existing data sources such as ERPs, HR systems, CRM, billing platform et al based on the needs of each use case. For example, a cable provider using cognitive care to support their customers may connect the CRM, billing system, and order management system to manage activation of service. It could also add the field service management platform to coordinate the intervention of field agents for installation.

Cognitive care does not necessarily support only external customers; it also can be deployed to support employees via an HR management system.

1.7. Best Practice #3: Achieve Sustained Excellence

Cognitive care is not just about technology: the customer experience needs to be properly designed and implemented, continuously improved and measured, and supported by the right culture.

The first key success factor to the implementation of cognitive car is the involvement from the beginning of cross-functional stakeholders and subject matter experts, to make sure that the right level of company, process and industry knowledge is infused in the design.

The deployment process consists of three major steps:

1. **Co-create**
    - Design the Experience
        - Build the target customer experience
        - Design the prioritized customer journeys

        - Elaborate the target use cases
    - Understand the enabling capabilities
        - Map required capabilities
        - Understand and design target architecture
        - Iterate on the customer experience vision
    - Build a backlog
        - Establish and agree on the business case
        - Finalize the vision
        - Draw out a value-based roadmap
        - Build a prioritized, implementation ready backlog of user stories
2. **Co-execute**
    - Iterate to minimum viable product (MVP)
        - Implement first instance of MVP
        - Incorporate feedback from agents and customers
    - Market Launch
        - Train agents
        - Go-live preparation & execution
        - Stand up center of excellence (COE)
3. **Co-operate**: Continually iterate to incorporate learning and scale along roadmap

Once the cognitive care platform is in place, companies need to establish a framework and governance to sustain an environment of continuous improvement. They also need to track and measure rigorously the KPIs that matter the most.

Establishing a center of excellence (COE) is a recommended tool to:

- Govern and monitor the development and performance of the platform across multiple channels and dimensions.
- Assess and prioritize new experience use cases across three dimensions: value, time and business objectives.
- Deploy COE platform enhancement squads with specialized skills to maximize efficiency and quality for new and existing use cases.
- Leverage intelligent data and AI skills to achieve superpersonalized customer interactions

Finally, a successful transformation requires a **cultural shift** that can be achieved through the following tactics:

- **Learn by doing** – Cross-functional stakeholders adopt new agile and design-thinking practices as they design the future state.
- **For Users, By Users** – New process, tools, and interfaces are rooted in user-centric research and co-created with the users themselves to ensure successful future adoption.
- **Clear communication** from leaders and change agents
- **Cross-functional stakeholders champion** the new tools, processes, and ways of working they are creating with a cadence of playbacks to socialize changes with their teams

2. Conclusion

Customer experience is the key for success and growth in the cable TV business. Growing customer expectations and external competition have pushed the boundaries of what is considered exceptional. Cognitive care, with the application of AI, analytics, data science and automation is the enabler that lets providers deliver a consistent, multichannel, digital outstanding experience to each one of their customers.

3. Abbreviations

AI	artificial intelligence
API	application programming Interface
B2B	business to business
B2C	business to consumer
BSS	business support systems
COE	center of excellence
CSAT	customer satisfaction
ERP	enterprise resource planning
FAQ	frequently asked questions
HR	human resources
KPI	key performance indicator
MVP	minimum viable product
NLP	natural language processing
NPS	Net Promoter Score
OSS	operations support systems
OTT	over the top
ROI	return on investment
STT	speech to text
TTS	text to speech

4. Bibliography and References

Total Economic Impact, Forrester, 2020

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

MapiFi: Using Wi-Fi Signals to Map Home Devices

A Technical Paper prepared for SCTE by

Yonatan Vaizman, Senior Researcher Machine Learning, Comcast, SCTE Member
1800 Arch Street
Philadelphia, PA 19103
Yonatan_Vaizman@comcast.com
202-524-5077

Hongcheng Wang, Distinguished Engineer, Comcast, SCTE Member
1800 Arch Street
Philadelphia, PA 19103
Hongcheng_Wang@cable.comcast.com
332-301-5055

Table of Contents

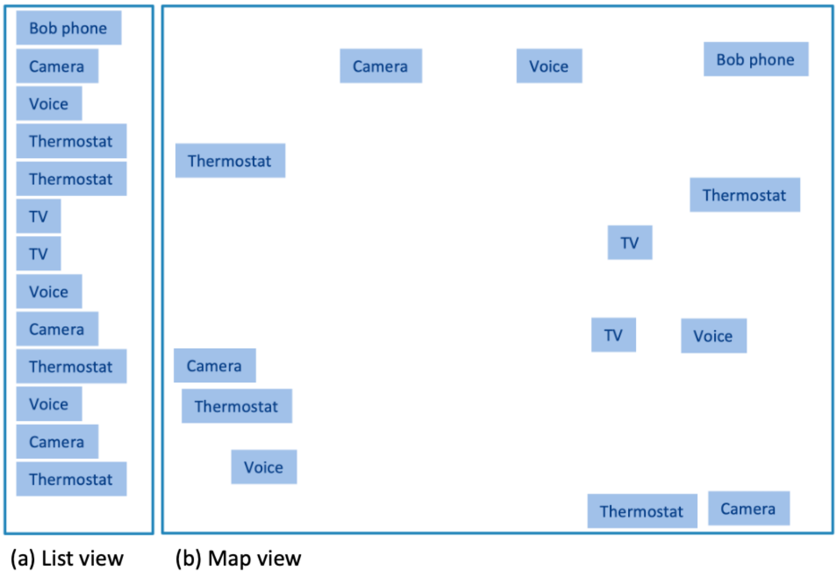
List of Figures

1. Introduction

Imagine a map of your home with all of your connected devices (computers, TVs, voice control devices, printers, security cameras, etc.) in their locations. You could then easily group devices into user-profiles, monitor Wi-Fi quality and activity in different areas of your home, and even locate a lost tablet in your home. MapiFi is a method to generate that map of the devices in a home. The first part of MapiFi involves the user (either a technician or the customer) walking around the home with a mobile device that listens to Wi-Fi radio channels. The mobile device detects Wi-Fi packets that come from all of the home's devices that connect to your gateway and measures their signal strengths (ignoring the content of the packets). The second part is an algorithm that uses all the signal-strength measurements to estimate the locations of all the devices in the home. Then, MapiFi visualizes the home's space as a coordinate system with devices marked as points in this space. A patent has been filed based on this technology [3] .

Today's Wi-Fi access points (home-internet gateways) often come with applications to manage the home network and the devices connected to it (phones, computers, voice control devices, cameras, etc.). Such applications may include features like creating user-profiles (assigning devices to people in the home), pausing internet access to certain devices (e.g., for parental control over a child's device), and Wi-Fi usage tracking (how much download/upload does each device take). These apps may also monitor Wi-Fi quality (e.g., channel interference, weak signals) and facilitate troubleshooting Wi-Fi problems for specific devices or for the overall home network.

A traditional way for a Wi-Fi management application to present the home devices is in a list view (Figure 1 (a)). Such a view can be inconvenient for the user: it is hard to identify which device is which (for example if your home has three voice control devices). We propose to present the home devices in a map view, showing *where* each device is in the home (Figure 1 (b)).



(a) List view (b) Map view

Figure 1 – Managing home devices: list view (a) vs. map view (b)

A map view of the devices in the home will provide a convenient user interface to manage and control devices in the home. Before we explain *how* to create the map (the MapiFi method), we cover several features for customer experience improvement and troubleshooting, showing *why* such a map is useful:

1.1. Device Identification and User Profiling

A home may have multiple web-cameras, multiple voice-control devices, multiple smart TVs, etc. When looking at a list view of the devices (Figure 1 (a)), it can be hard to identify which device is which. A map view (Figure 1 (b)) can resolve this ambiguity – the user can clearly distinguish between the TV in the living room and the TV in the bedroom (see illustration in Figure 2). Similarly, viewing all the devices on a map can help group devices into user-profiles based on functional space (e.g., grouping the camera and thermostat located in Alice’s room into a user-profile for Alice (Figure 2)). Such user profiles can then be useful for tracking internet usage by user, or for control features like parental control (e.g., easily pausing internet access at night for all the child’s devices).

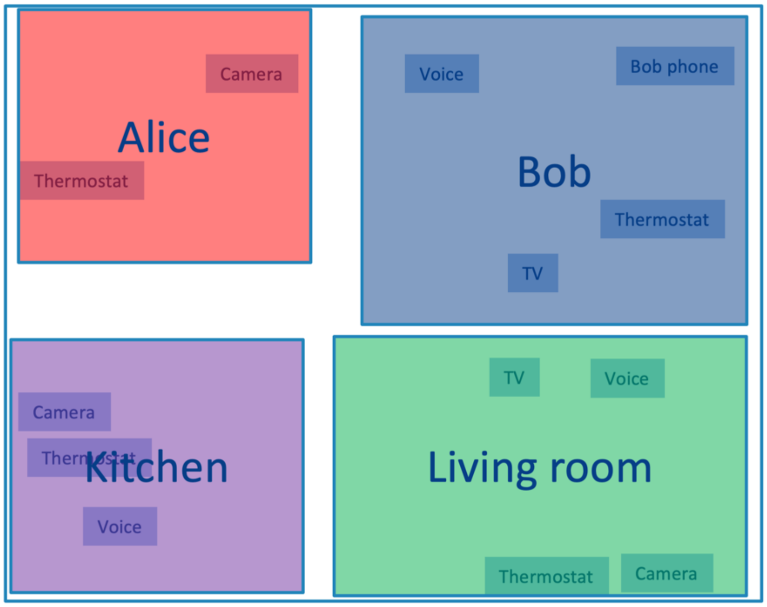


Figure 2 – User profiling with map view

1.2. Monitoring Wi-Fi Activity and Quality

Users are often interested in tracking how much internet they are using (e.g., to make sure they have the appropriate internet service plan, to track how much time they spend online gaming, etc.). With a map view, the app can present internet activity (e.g., how much download) as a heat map – this can help the user see which areas of the house (possibly associated with a certain person or family activity) use much traffic (Figure 3). It can also help detect unusual internet activity. The user can also visualize Wi-Fi quality measures (like signal strength or the rate of corrupted packets) as a heat map on the map of the home – this can help detect areas in the house that consistently get poor Wi-Fi conditions.

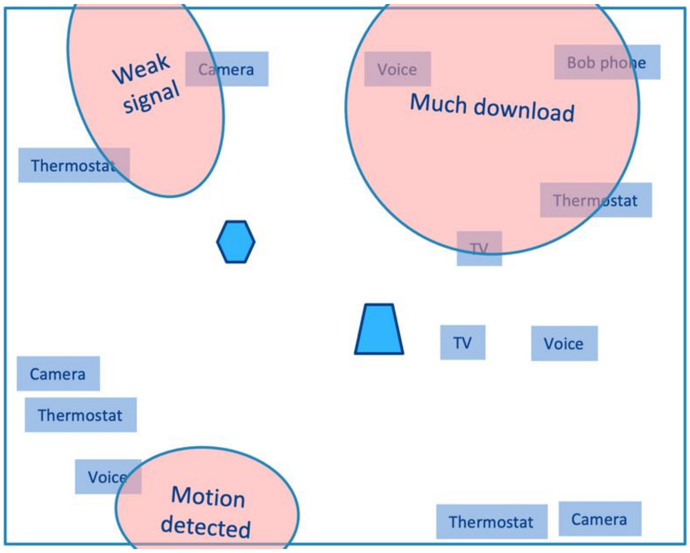


Figure 3 – Wi-Fi monitoring with map view. The trapezoid represents an access point and the hexagon represents a Wi-Fi extender.

1.3. Recommending Placement of Gateway and Wi-Fi Extenders

Based on the Wi-Fi signal coverage around the house, mapping the devices can help decide on the best place to put the main access point (the home's internet gateway, illustrated by the trapezoid in Figure 3). This can either be done by the user looking at the map view with the visualized signal strength and deciding on their own, or by some automated algorithm that considers devices' locations, their signal strengths, amount of traffic, and availability of internet wall-outlets to optimize gateway placement. Similarly, such algorithms can recommend the user to add Wi-Fi extenders (illustrated by the hexagon in Figure 3), and where to put them, typically in locations leading to a place in the home that consistently gets weak signal.

This analysis can be done on the day of installing a new Wi-Fi network, or after a few weeks or months of the residents using the Wi-Fi network. Visualizing signal strength on the map can be a helpful tool to explain to the user why they experience Wi-Fi trouble in some areas of the house or to convince them to change the gateway position or install Wi-Fi extenders.

1.4. Localizing Motion

There's growing research area about using Wi-Fi radio signals to detect motion in the home ([4] [1]). Methods usually rely on tracking the RSSI (received signal strength indicator) of signals coming from devices, or more detailed measurements like CSI (channel state information). When there is no motion in the home, typically the RSSI or CSI from devices should be stable. On the other hand, when someone is moving in a room, their body affects the obstruction and reflection of the radio signals in the room, causing fluctuations in the measurements. When we know the location of all the home devices, an algorithm can analyze the signal changes measured from all the devices and infer the location of the motion in the home. For example (Figure 3) if the gateway (the trapezoid in the center of the house) senses co-occurring fluctuations in signals from the three devices in the kitchen (lower left quadrant of the map), but stable signals from the other devices in the house, the algorithm can infer that there is motion in the kitchen.

1.5. Locating a Lost Device

The features mentioned so far are mainly practical for a map of the stationary devices in the home – devices that typically stay in their fixed location (TV, voice control, thermostat, printer, etc.). However, MapiFi can also help with mobile devices, like tablets – it can help locate a device that you lost in your home (if the device is still connected to the home network). Whenever you lose a device, you can use the MapiFi method (including a new "walk around" to take measurements, and then running the localization algorithm) to produce an ad-hoc map of your home. Then you'd see all the devices in the home – some of them will be the stationary devices (which you may have already mapped before), and one of them will be the lost device – the map will show you where the device is.

In the next section, we describe the MapiFi method.

2. Method

The MapiFi method has two parts: Taking Measurements (which should take a few minutes) and a Localization Algorithm (taking a few seconds). A typical time to use MapiFi is after connecting all the home devices to a new Wi-Fi network – this will produce a map with the locations of all the stationary devices in the home. The user can decide to ignore/delete locations of mobile devices, like phones, if they assume that these devices will keep moving around the house. The same map can be used for a long time, possibly months or years. The user can use MapiFi again whenever they want if they add new devices to the home or if they re-arrang the locations of devices in the home.

2.1. Taking Measurements

The first part of MapiFi requires participation from the user. This can be the end user of the Wi-Fi network (the resident of the home) or a technician coming to help troubleshoot the home network. The user carries a mobile device ("the measuring device") and walks around with it in the home, pausing for a few seconds in multiple points along the path – we refer to these as "anchor" points. In each anchor point, the user runs a script that listens to the home network's Wi-Fi channels and captures Wi-Fi packets coming in from all the home devices. The measuring device ignores the content of the packets (which is typically encrypted anyway), and only registers the MAC (media access control) address of the sending device and the RSSI of each packet.

It doesn't matter which channel the home network is using. It is fine if the network changes channels during the walk (as long as the measuring device knows which channel to listen to). The key thing is to catch incoming packets from the devices and measure the signal strength (as a proxy for distance). It is also possible to "ping" the home devices (e.g., send some http request to the smart TV, either from the measuring device or from another collaborating device) in order to "wake them up" and make them reply and send something over Wi-Fi – the point is to induce the devices to transmit anything over the air, so that the measuring device can measure signal strength.

The user can make the walking path longer, more winding, exploring all rooms, with more anchor points, and can hold the measuring device in different heights. All these factors can help gather more diverse evidence (measurements), which will help the localization algorithm estimate device locations more accurately.

2.2. Localization Algorithm

2.2.1. From Signal-Strength to Range

MapiFi relies on the fact that radio signals lose power as they travel through the air: generally, we will sense signals coming from near devices with higher power than signals coming from far devices. In practice, the level of signal attenuation depends on the frequency band, obstacles along the way, reflection paths, etc. We start simple by assuming a constant decay factor, meaning that the power of the signal reduces (from the transmit power Tx to the measured, received power Rx) by a constant power γ of the distance (or "range" r) that it travelled (similar to previous models of signal decay through space, e.g., [2]).

$$r^{\gamma} = \frac{Tx}{Rx}$$

Or, if the powers are in decibel scale:

$$r^{\gamma} = 10^{\frac{TxdB-RxdB}{10}}$$

An important part of MapiFi is the realization that every client-device may use different (but typically constant) transmit power. To simplify the method, we calculate the range of a signal assuming a fixed arbitrary transmit power ($TxdB = -50dB$) and we model these inter-device differences later – by a distance-gain variable. We assumed a constant decay factor of $\gamma = 2.5$. So, for every measurement, we use the RSSI $RxdB$ to calculate the range of the signal:

$$r = 10^{\frac{-50-RxdB}{10\gamma}}$$

We then treat the range r as an uncalibrated measure of the distance between the measuring (anchor) point and the transmitting device. We model the variability among devices with a multiplicative gain, g, and we assume that every device j has a fixed gain g_j: so, the distance between device j and anchor i is modeled as $g_j r_{i,j}$.

This simplified model assumes that the device stays in its place during the few minutes of walk-around and that it is trying to communicate with the access point, which also stays in its place. Remember, even if the measuring device keeps moving, it doesn't matter, because the packets are not targeted to it, they are always targeted to the access point. It is possible that during the walk there is suddenly more noise, and the home device increases its transmitting power. Also, if during the walk the home device switches Wi-Fi frequency band from 2.4GHz to 5GHz than its signals will attenuate more quickly with distance. In such cases, the fixed-distance-gain assumption may fail. Future methods can mitigate these issues by adding variables to the model, or by helping the measurement walk be quicker and more efficient.

This model describes the distance that the signal travels from the transmitting device to the measuring device, which can be a non-straight path. In this preliminary MapiFi version, we further simplify assumptions and treat it as an approximation of the direct distance in 3d space between the transmitting and measuring devices. Next, we describe how the localization algorithm uses all these distances to estimate the location of every device (and every anchor point) in the 3d space of the home.

2.2.2. Finding a Device's Location and Gain

In the basic problem, we assume we know the N anchor locations $a_i \in \mathbb{R}^3$ for $i \in \{0, \dots, N-1\}$, and we have the signal strength measurements (which we convert to ranges $r_i \in \mathbb{R}_+$ between the device and all the N anchors). We need to find out the device's location $d \in \mathbb{R}^3$ and its gain $g \in \mathbb{R}_+$. The evidence (measurements) gives us a set of N quadratic equations in d and g:

$$\forall i \in \{0, \dots, N-1\}:$$

$$\|d - a_i\|_2^2 = (gr_i)^2$$

or:

$$\forall i \in \{0, \dots, N-1\}:$$

$$\|d\|_2^2 + \|a_i\|_2^2 \, - \, 2a_i^T d = r_i^2 g^2$$

By subtracting equations $i \in \{1, \dots, N-1\}$ from equation 0, we get a system of $N-1$ linear equations (in d and g^2):

$$\forall i \in \{1, \dots, N-1\}:$$

$$\|a_0\|_2^2 - \|a_i\|_2^2 = 2(a_0 - a_i)^T d + \left(r_0^2 - r_i^2\right) g^2$$

With enough equations (enough anchor points, in general position), this system can theoretically be solved for the device location d and its gain g (total of 4 unknown variables) – it would require 4 linearly independent equations, which would require 5 quadratic equations. This means that in perfect conditions, 5 anchor points will suffice to fully recover the device's location and gain. However, because the measurements are noisy, these equalities can be approximated by solving a least squares problem, possibly with constraints (g needs to be positive, and the coordinates of d can be constrained by the dimensions of the house if these are readily available). Having more than 5 anchor points would provide more evidence (the more, the better) and help better infer the device's location. If the anchor points (the locations where the user stops to take measurements) are on a straight line, it may result in redundant (linearly dependent) equations – that is why it is best to take a winding path through the home, as well as to hold the measuring device at different heights.

2.2.3. Finding an Anchor's Location

In the complementary problem, we assume that we know the locations $d_j \in \mathbb{R}^3$ and gains $g_j \in \mathbb{R}_+$ of all the M home devices ($j \in \{0, \dots, M-1\}$). From the measurements, we have the ranges $r_j \in \mathbb{R}_+$ between all the M devices and a specific anchor, and we need to find out this anchor's location $a \in \mathbb{R}^3$. The evidence gives us a set of M quadratic equations in a:

$$\forall j \in \{0, \dots, M-1\}:$$

$$\left\|a - d_j\right\|_2^2 = r_j^2 g_j^2$$

or:

$$\forall j \in \{0, \dots, M-1\}:$$

$$\|a\|_2^2 + \left\|d_j\right\|_2^2 - 2a^T d_j = r_j^2 g_j^2$$

As in the previous problem, we can turn this to a system of $M - 1$ linear equations in a, and treat this as a least squares problem, with constraints on a's coordinates to be inside the house (if the house's coordinate boundaries are available).

2.2.4. Full Algorithm

In some cases, the locations of the anchors (the measurement points) can be known in advance – either if the measurements are taken from multiple stationary network devices (like Wi-Fi extenders), or if the user carefully recorded the coordinates of every point along the path where they took measurement. In such rare cases, you can simply solve the first problem for each home-device separately to get its location.

However, in the general case, the anchor locations are not known because the user walks freely in a winding path through the house. So, in the general case, we have N unknown anchor locations a_i ($i \in \{0, \dots, N-1\}$), M unknown device locations d_j and gains g_j ($j \in \{0, \dots, M-1\}$). The observed variables are the measured ranges $r_{i,j}, i \in \{0, \dots, N-1\}, j \in \{0, \dots, M-1\}$ (where for some of the anchor-device pairs we may be missing a measurement).

The solution is an alternating algorithm: starting with a random initialization (guess) of the anchor locations a_i, and then iterating over the steps:

1. estimating the device locations d_j and gains g_j given the currently assumed anchor locations a_i (2.2.2),
2. estimating the anchor locations a_i given the currently assumed device locations d_j and gains g_j (2.2.3),
3. (optional) normalizing the estimated anchor locations, making sure the points don't explode in space or decay to the origin (or make sure they remain within a desired boundary box).

3. Initial Experiment

We tested the MapiFi method in one initial experiment, in an actual single-floor apartment with one bedroom and one den (see Figure 4). The internet gateway and five devices were placed in known locations. We used a laptop as the measuring device. The user walked with the laptop in a path through the rooms (marked in blue curve in the figure), stopped in 40 points along the path (the "anchor" points) and took measurements (signal strengths of Wi-Fi packets from the devices in the home). This walk took approximately five minutes.

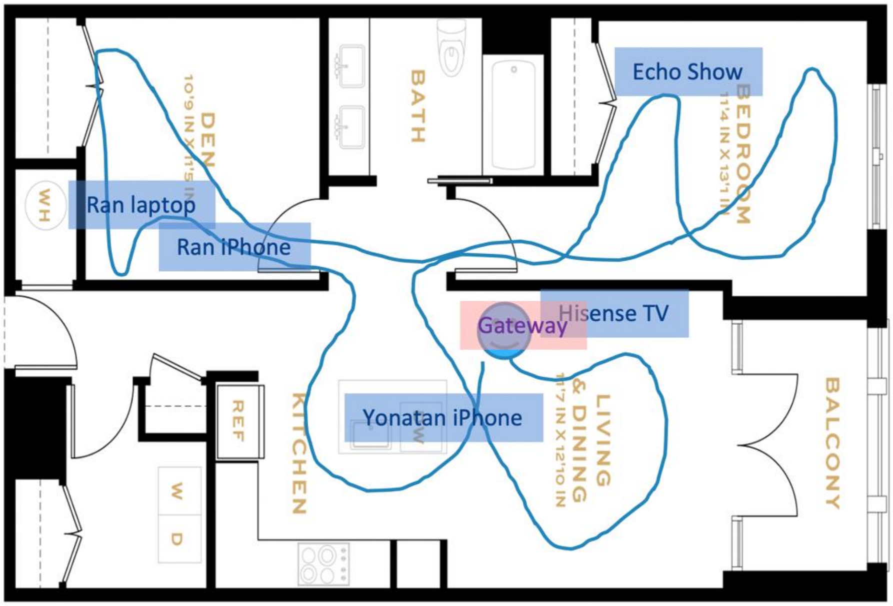


Figure 4 – Taking measurements in a real home

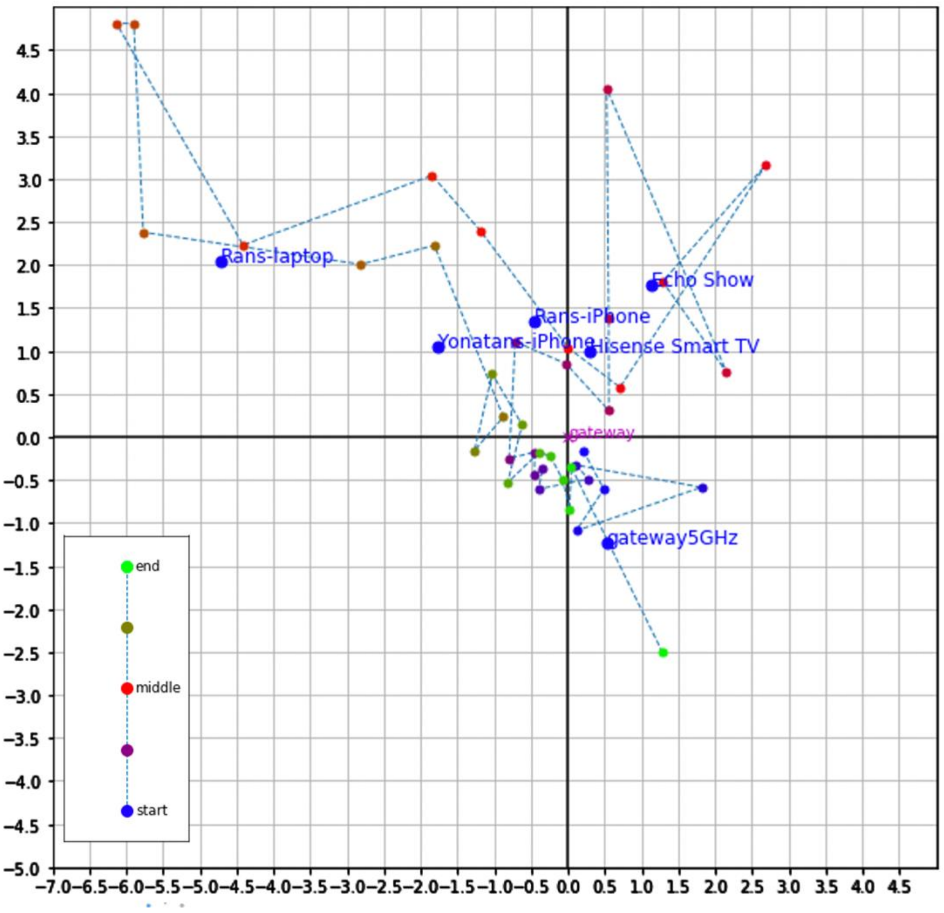


Figure 5 – The resulting map with MapiFi

Figure 5 shows the resulting map after six iterations of the MapiFi localization algorithm. The map is projected to two dimensions (ignoring the height dimension). The estimated measurement path is marked with dashed line with anchors as small dots with colors ranging from blue at the start of the path to green at the end of the path. The devices are marked with bold blue dots at their estimated locations. The localization algorithm managed to reconstruct an estimated path that resembled the actual path that the user took: starting close to the gateway, going through the bedroom (close to the “Echo”), passing through the den (close to “Ran’s iPhone” and “Ran’s laptop”), then passing by “Yonatan’s iPhone”. The estimated locations of the devices approximate the relative locations of the actual devices in the home.

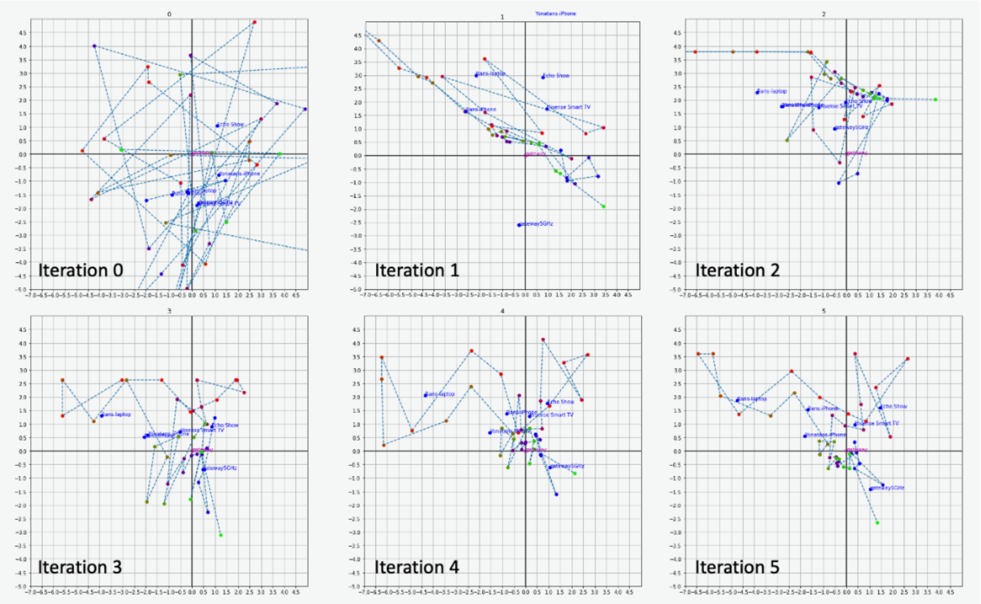


Figure 6 – Iterations of the MapiFi localization algorithm

Figure 6 shows what happens in the algorithm: it starts with random initialization (iteration 0). As the iterations progress, the estimated path becomes more coherent and resembles the actual path that the user took in the home. The devices' estimated locations change with every iteration, following the estimated locations of the anchors.

4. Conclusions

We describe a method for mapping Wi-Fi devices in a home's space. The generated map can help the resident manage their home devices and monitor their Wi-Fi quality and usage. We report promising results from an initial experiment in a real home environment. The method itself can be further improved with various adjustments, like combining known device-locations as constraints, giving stronger weight to measurements of closer-devices (where distance estimation is more reliable), modeling obstacles (walls, furniture, etc.), and combining with a structural model of rooms and walls from computer-vision based methods.

5. Abbreviations and Definitions

5.1. Abbreviations

RSSI	received signal strength indicator
CSI	channel state information
MAC	media access control

6. Bibliography and References

[1] Gong, L. Y. (2015). WiFi-based real-time calibration-free passive human motion detection. Sensors, 15(12), 32213-32229.

[2] Shang, F. S. (2014). A Location Estimation Algorithm Based on RSSI Vector Similarity Degree. International Journal of Distributed Sensor Networks., https://doi.org/10.1155/2014/371350.

[3] Vaizman, Y. W. (2021). Locating Devices Within a Premises (patent pending). U.S. Patent and Trademark Office.

[4] Youssef, M. M. (2007). Challenges: device-free passive localization for wireless environments. Proceedings of the 13th annual ACM international conference on Mobile computing and networking (pp. 222-229). ACM.

SCTE STANDARDS JOURNAL

SCTE
a subsidiary of CableLabs®

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