Understanding and Managing Quality-of-Service in Grid Communications

May 1, 2024

Prepared by:
U.S. DEPARTMENT OF ENERGY,
OFFICE OF ELECTRICITY

Part of a series of white papers on
Secure Pathways for Resilient Communications.

SPaRC
Secure Pathways for Resilient Communications

U.S. DEPARTMENT OF ENERGY | OFFICE OF ELECTRICITY

Executive Summary

In the evolving landscape of technology-driven energy solutions, the shift towards a carbon-free grid necessitates the seamless coordination of distributed energy resources (DERs), including renewable generation (solar, wind), energy storage systems (batteries, electric vehicles (EVs)), and demand response. This transition marks a profound departure from the conventional paradigm of electrical energy generation on the grid. It not only involves a fundamental shift in the underlying physics, transitioning from large rotating thermal generators to inverter-based resources (IBRs), but also introduces new ownership models outside the realm of traditional utilities as well as changes in plant numbers and distribution on the grid. With an increasing number of these DERs integrated into the grid, effective coordination becomes paramount for ensuring a resilient and reliable energy infrastructure, reliant on data communications with specific performance requirements.

The orchestration of these DERs hinges on robust communications infrastructure. The establishment of communication systems capable of meeting the stringent performance requirements for efficient coordination is of utmost importance. Traditionally, electric utilities have taken on the responsibility of constructing, maintaining, and owning their communication networks to meet the needs of their operational processes, such as metering, substation monitoring, protection systems, and generation dispatch, each with unique communication system demands. It's worth noting that each of these operational processes may have distinct performance expectations from the underlying communication system. New grid services will likely challenge traditional communication methods and require different performance of the communication channels. This whitepaper explores a series of attributes and characteristics of a network or communications system that together describe the overall performance of that network or system, called Quality-of-Service (QoS), which will be critical to the reliability and resiliency of the future electric grid.

Introduction

Welcome to the sixth paper in a series of white papers authored by the Secure Pathways for Resilient Communications (SPaRC) project. The first four white papers provided a high-level description of the project and discussed the challenges facing the evolving grid, including the communication sector, grid collaboration, and grid device interoperability. The fifth white paper began a series of deep-dive papers with a look at latency and its impact on grid communications.

This whitepaper, the second deep-dive discussion, explores a series of attributes and characteristics of a network or communications system that together describe the overall performance of that network or system, called Quality-of-Service (QoS). As the grid evolves and information on the grid components, their characteristics, and the ability to dispatch a variety of energy resources become increasingly important, the performance of the communications systems in delivering data for orchestrating system operation is more critical than ever. These QoS attributes are key measures in ensuring seamless and timely balancing of sources and loads.

What is QoS?

QoS in the context of communications refers to the set of attributes and characteristics of a network or communication system that define its overall performance. The performance of the network is relative to both the technology of the communication platform as well as the operations, design, and implementation of the communication system.

> The QoS of a communication system is the set of attributes and characteristics that define its overall performance.

Operators often utilize different techniques and services to manage the network, prioritize certain types of data, ensure efficient use of resources, and improve the overall performance of the network. It is important to note that as technology changes so can the performance characteristics; new characteristics may be determined and/or other characteristics may change in their application or usage. One example could be the application of jitter in a packet-switched network versus a Time Division Multiplexing (TDM) network. Jitter is a performance characteristic for both types of technologies, but it translates differently relative to each technology.

In addition to advancement in technologies altering QoS characteristics, how the technology is integrated to a service can directly affect QoS (i.e., not all services of the same technology have equivalent QoS). This is readily apparent as packet-switched networks have been displacing TDM technology, just as TDM displaced analog and Frequency Division Multiplexing (FDM) systems. TDM was built to transition to digital systems, lowering costs and improving bandwidth, with a focus on voice in the late 1980-1990's. TDM's performance characteristics focused upon this transition and associated services, while packet-switch networks evolved to handle data, growth of the internet, and e-commerce while also reducing cost.

QoS attributes and characteristics are relative to both the technology and the services provided. Now traditional TDM-supported applications, which include real-time services, need to be supported over packet-switched networks, thus necessitating the appropriate quality-of-service considerations for the architectures. For example, the transition from voice over TDM to voice over IP (VoIP) required prioritization of the voice traffic across a provider's IP-based network.

The displacement of TDM by IP and Ethernet marks a significant evolution in network technologies, driven by the demand for higher bandwidth, greater flexibility, and the integration of diverse communication services onto a unified structure.

In the previous whitepaper, we identified seven common communications system characteristics: *latency, jitter, bandwidth*, *throughput, packet loss, availability*, and *security*. The requirements for these characteristics will vary, both with the deployed technology and the operational service offered with that technology. To better represent these variations, we will discuss the performance characteristics in the context of TDM and packet-switched networks.

Latency

From the previous whitepaper, *latency* refers to the delay in the transmission of data from the sender (source) to the receiver (destination) over a network or communication path. Latency components include Propagation Delay, Transmission Delay, Queueing Delay, and Processing Delay. The key difference in latency between TDM and packet-switched networks lies in their operational principles. TDM networks offer more predictable latency due to their fixed allocated bandwidth channels and inband synchronized nature but are less efficient in handling varying traffic loads and types.

Packet-switched networks offer greater flexibility and can be more efficient with bandwidth usage, but they introduce more variable latency due to the dynamic routing of packets and potential congestion, often reflected in queuing delays.

This variability in latency between technologies and services can be a critical factor in applications requiring real-time communication, such as voice over IP (VoIP), video, and rapid control actions, where packet-switched networks need mechanisms like QoS to prioritize traffic and manage latency effectively.

Bandwidth and Throughput

Bandwidth and throughput are both key performance metrics in communications, each representing different aspects of network capacity and efficiency. The differences in these metrics, within the context of TDM networks versus packet-switched networks, highlight the fundamental differences in how these two types of

technologies and services operate and manage data transmission, ultimately illustrating how QoS is dependent upon technology and service.

Bandwidth refers to the maximum rate at which data can be transmitted over a network connection, typically measured in bits per second (bps). It is a measure of network capacity and is a theoretical maximum that doesn't account for real-world conditions that can affect actual data transmission rates.

In TDM networks, bandwidth is allocated in fixed segments to different channels via time slots. Each channel receives a portion of the total bandwidth of the medium, dedicated to it for the duration of its time slot. This allocation is static, meaning each channel's bandwidth is constant and guaranteed, regardless of whether there is data to send or not. The total bandwidth of the system is divided among all channels, limiting the maximum bandwidth available to each.

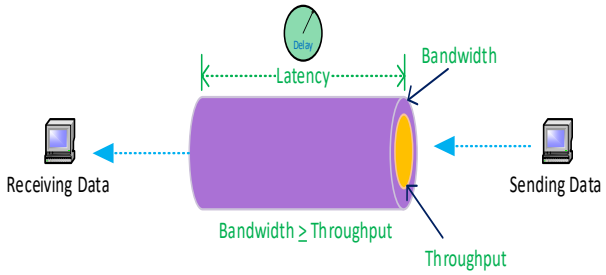

Figure 1: Representing Bandwidth versus Throughput.

In packet-switched networks, bandwidth is dynamic and shared among all users and applications. There is no inherent fixed allocation of bandwidth to specific timeslots, channels, or data streams. Instead, data packets are transmitted as needed, potentially allowing the network to utilize the available bandwidth more efficiently. Some services have operational features that may cap maximum rate, burst capability, or similar traffic shaping features. The maximum bandwidth available to an individual data stream can fluctuate based on overall network load and the current demand from other users.

Throughput refers to the actual rate at which data is successfully transmitted over a network, measured over a specific time period. It's influenced by various factors including the network's bandwidth and also by other elements such as network congestion, data packet size, and transmission errors.

In TDM networks, throughput can closely match the allocated bandwidth for each channel, assuming no technical issues with the transmission. Since each channel's bandwidth is fixed, the throughput is typically stable and predictable, making TDM suitable for applications requiring guaranteed bandwidth and throughput. However, the static allocation can lead to inefficient use of the total network bandwidth if some channels have no data to transmit during their allocated slots.

In packet-switched networks, throughput is more variable than in TDM networks. It can be significantly affected by factors such as network congestion, the efficiency of routing protocols, and the nature of the traffic. While the maximum bandwidth might be high, actual throughput can vary widely depending on current network conditions. Packet-switched networks are designed to maximize the efficient use of available bandwidth, but this can lead to fluctuations in throughput for individual data streams.

Overall, efficiency and flexibility are key strengths of packet-switched networks. Packet-switched networks are better suited to handle variable traffic loads and types, adapting in real time to changes in demand.

Jitter and Bit Error Rate

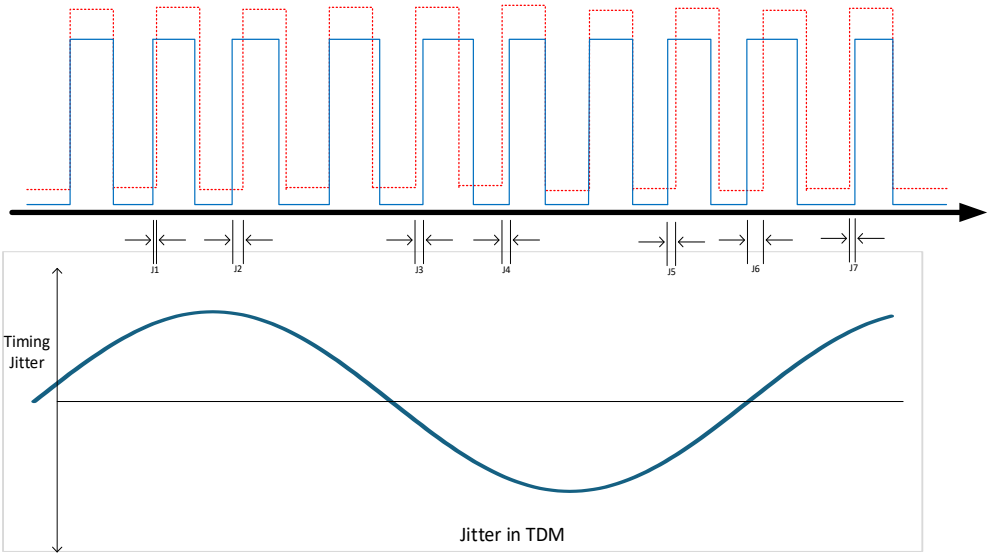

Figure 2: Example of Jitter in Synchronous Networks.

In the broadest sense, *jitter* is the deviation of a sequence of reference instants from their ideal values [1]. While this sounds straight-forward, jitter is not a singular metric. There are many standards that contain jitter definitions and calculation methods for both time and frequency domains. The IEEE Standard for Jitter and Phase Noise [1] (IEEE 2414-2020) pulls many of the jitter standards together in one document, providing definitions and calculation methods for several types of jitters, including Random Jitter, Timing Jitter, Periodic Jitter, and Deterministic Jitter. Real-time applications have timing dependencies, and, historically, real-time systems and equipment received timing support from synchronous protocols like TDM systems. TDM is based upon fixed timing and provides the synchronization between the transmitter and receiver, which can provide a higher level of QoS for real-time applications including lower latency and jitter. Packet-switched networks are inherently asynchronous but have developed several mechanisms to handle synchronization and support real-time applications effectively, despite the inherent challenges posed by their asynchronous nature. These mechanisms are crucial for delivering voice, video, and other time-sensitive services with the quality and reliability users expect.

One specific type of jitter, timing jitter, is especially important for protection systems and equipment associated with the electric grid. As shown in Figure 2, timing jitter is the “deviation of the actual reference instants associated with a timing waveform with respect to their ideal values” [1]. The ideal value of the timing signal is obtained from a highly accurate reference clock. In a network, the reference clock might be an atomic clock used as a primary reference source, or it could be represented by a high-quality oscilloscope or spectrum analyzer in a test environment.

If the jitter is a larger amplitude and the phase slowly varies below 10 Hz, it is a subset of jitter called “wander”. (If the variation is very slow, below 1 Hz, it is called “drift.) Wander does not normally cause bit errors since the changes are slow enough to be followed by the clock. It can, however, cause frame slips in some technologies.

The overall jitter of a system is the sum of all jitter contributions present in the system, regardless of type and source. Often total jitter is expressed as Peak-to-Peak Jitter or RMS (root-mean-squared) jitter. If the total jitter is high enough, it will cause bit errors. The probability that any given bit will be received in error at its destination is the *bit error rate* (BER). In general practice, the BER is tested by sending bit patterns through a system and measuring the ratio of the resulting bit errors to the total number of bits received. This is normally performed using a specialized type of test equipment J– a Bit Error Rate Test (BERT) set. This device generates a

test signal that stresses the system but still mimics some subset of expected data and then records the resulting number of bit errors. The BER is a common specification for communications and data equipment and typically ranges from 10^{-10} to 10^{-16} for TDM and Ethernet Systems, [2] however, for some wireless systems, 10^{-6} is a typical range.

In TDM networks, jitter, especially timing jitter, is a frequent concern. The effect of timing jitter is serious because it impacts the basic control for sending and receiving bits of data. All data movement is dependent upon time slots, so if the clock signal is not accurate, this can cause significant errors and frame slips in the system. This can be challenging when a Primary Reference Clock (PRC) signal is distributed across a large system with both asynchronous and synchronous transport technologies.

In packet-switched networks, the fundamental method for handling data from source to destination changed from a synchronized periodic transport protocol to a more flexible, addressed series of network paths or devices. Understandably, the definitions of many QoS characteristics and metrics changed with it. Jitter is one example. There is no longer a reference to waveforms, but a description of impacts to data packets moving through the network. As such, the previous paper explained jitter as "the variability in latency over time". This is a specific example of a jitter definition that applies to IP packet delivery. The Internet Engineering Task Force (IETF) defines jitter as "the difference between the one-way-delay of the selected packets" in a stream of packets and can also be called IP Packet Delay Variation (IPDV) [3].

Figure 3 shows the variation in delay between packets. The top row of data packets is a dataflow with little to no IPDV. The packets arrive at regular intervals. The bottom row of data shows a dataflow with significant IPDV, as evidenced by the large variation of packet arrival time within the dataflow. IPDV is a useful performance metric whether a system is owned by the user or furnished by a third party. In ITU-T Y.1540 [4], mean, minimum, and median packet transfer delays are defined for a series of packets from source to destination in a network. Normally the minimum packet delay is used as a reference value for the IPDV as in ITU-T Y.1541 [5].

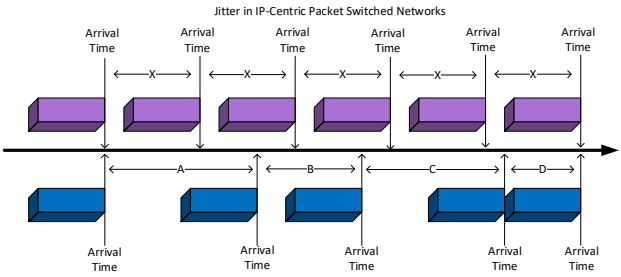

Figure 3: Jitter in packet switched networks.

Multiple tools can be used to check latency and its variation, including Traceroute and Ping. Traceroute has the added advantage of showing layer 3 routers and minimum, maximum, and average latency for that hop, which allows an IPDV calculation per hop. With data traveling over many systems not owned by the user, as well as changing network conditions with routing changes, the individual sources of the jitter are not easily determined. However, some common causes are congestion, jitter buffers that are too small, and lack of traffic prioritization for jitter-sensitive traffic. With the proper tools and network monitoring, IPDV becomes a standardized, easily measured overall performance metric of the network in use against which application requirements can be verified.

Packet Loss

Packet loss refers to the loss or non-delivery of entire data packets during transmission over a network. Packet loss is applicable to IP packet-switched networks and not to TDM systems. Data on a TDM system can have bit errors or be sampled incorrectly, but there is no equivalent concept of packet loss until the whole signal is lost.

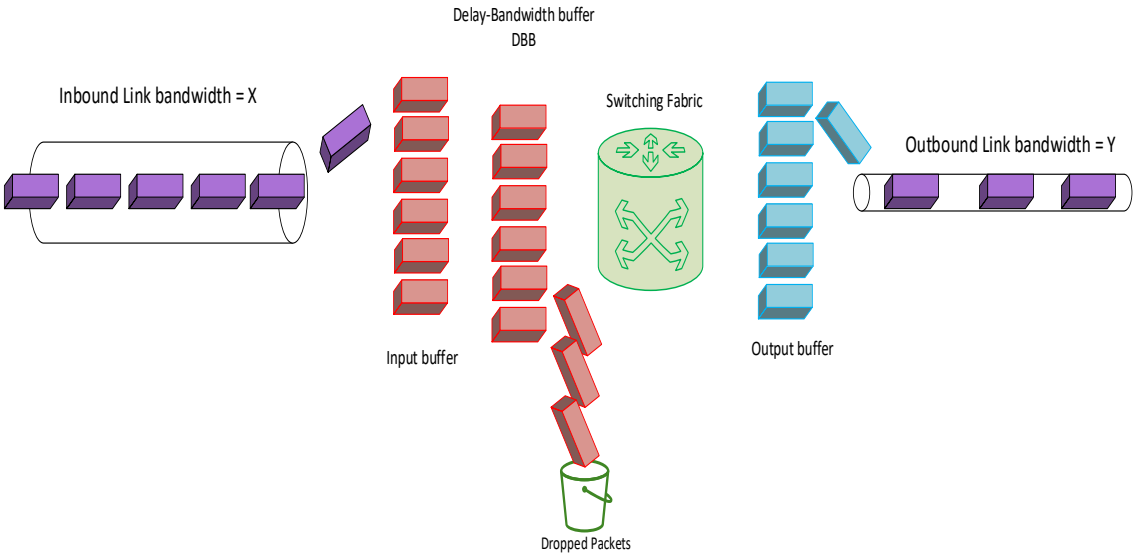

Figure 4: Example of packet loss.

For packet-switched systems, packet loss occurs when one or more packets do not reach their intended destination, which can result in errors in the larger message. In addition, lost packets can decrease throughput since they are never received at the destination. A common cause of packet loss is network congestion, depicted in Figure 4. Network congestion happens when traffic flowing through the network exceeds the maximum capacity of the network. There are many reasons for congestion. The example in Figure 4 represents one possible condition where the packets arriving at a network device far exceed the ability of the network device to forward based upon the outbound link bandwidth. Hence, as delay-bandwidth buffers fill up, the network device is forced to drop packets. There might be intentional oversubscription in the network for system cost reduction or network outages that result in periods of time where traffic exceeds the capacity when users utilize more bandwidth than normally expected. An example of this might be an intermittent voice call over a cellular network or slow data transfers.

Packet loss can also result from buffer overflows or errors in routing decisions. In a wireless network, a low signal-to-noise ratio caused by interference, signal attenuation, or distortion in the communication channel may also contribute to packet loss. As with jitter, both Ping and TraceRoute tools are helpful in finding the magnitude of the packet loss issues and what network segments are contributing to the overall impact.

Availability

Availability, in the context of communications, refers to the accessibility and usability of services when needed by users. Availability ensures that data and systems are consistently accessible and operational, without experiencing excessive downtime or disruptions. In a secure system, this is modified by adding a qualifier that the availability must be only for authorized users. See Security, below, for additional details.

A simple measure of availability is to divide the uptime for a system by the sum of its uptime and downtime for the same time period and express it as a percentage. Availability requirements are typically determined by assessing the tolerance of the downtime for a system. For example, certain processes may only be allowed to be inaccessible for a matter of minutes, while other less critical elements can be unavailable for a week or more with no negative impacts. Ensuring high availability is an active process that can occur at many levels of a system. Redundancy at multiple levels, distributed storage, physical separation, media diversity, manufacturer diversity, and 3rd party provider diversity can all be methods to increase the availability of a communications network. The trade-off with increasing availability is increasing cost for each level of improvement. Loss of

system availability imposes other costs, which could come in the form of customer dissatisfaction, penalties for compliance violations, or curtailing the grid due to lack of visibility, to name a few. At the end, it becomes an economic decision regarding what measures to use and where to use them to attain the highest availability improvement with the best rate of return.

Security

Security, in the context of communications, refers to the protection of data, information, and communication channels from unauthorized access, disclosure, alteration, and disruption. It encompasses a wide range of practices, technologies, and protocols aimed at safeguarding the confidentiality, integrity, and availability of data as it is transmitted and received across communication systems. This paper is focused on the security of the communications pathways as they carry data from source to destination. The equipment that makes up the communication pathway has capabilities that can add to defense-in-depth for the cyber posture of any organization.

The security aspects of a communication system can add to a cybersecurity program's posture of defense-in-depth.

Defense-in-depth is a layered approach to security that employs multiple mechanisms to protect the integrity, confidentiality, and availability of information systems. This paper assumes that there is defense-in-depth in use and that other security measures are being taken to secure the data before it enters the communications pathways and after it exits the communications pathways.

A Network Management System (NMS) using Simple Network Management Protocol (SNMP) [6] can contribute to a defense-in-depth strategy for communications in several ways for both IP and TDM networks. In TDM networks, some inherently limiting qualities of the technology act as security measures. For example, specific primary and backup routes are deterministic. The routes are provisioned to build explicit pathways from device to device through a TDM system before traffic is added, with each device expecting to receive or send information to or from the specific adjacent device. On-ramps and off-ramps for the information being transported are built in advance as well.

SNMP is one tool available in most communication equipment and is an example of adding to defense-in-depth for a cybersecurity program.

In contrast, Ethernet and IP bring the flexibility of having multiple routes from source to destination providing dynamic routing via different paths during link failures or a congested network. The TDM provisioning makes it very noticeable if traffic is re-directed to different pathways or an unexpected device is inserted in the middle of a pathway, however the additional pathways, equipment, and channels can add significantly to costs. For additional visibility of a TDM system, an NMS can be used to monitor and secure the control plane of the TDM transport equipment via SNMP traps configured to alarm upon performance changes, including signal attenuation and loss of signal. While SNMP itself is primarily designed for network management (monitoring and configuring network devices), when used wisely within an NMS, it can enhance the following cybersecurity efforts:

1. **Network Visibility and Monitoring**

   Device Inventory: SNMP can be used to automatically discover network devices, creating a comprehensive inventory. Knowing exactly what devices are on the network is a foundational cybersecurity principle.

   Performance Monitoring: By monitoring device performance, SNMP can help identify anomalies that might indicate a cybersecurity threat, such as unusual traffic patterns, changes in signal strength, or unexpected

changes in device behavior.

2. **Configuration Management and Control**
   Standardized Configuration: SNMP can be used to enforce standardized configurations across network devices, reducing the risk of vulnerabilities due to misconfiguration and multiple software versions.

   Change Management: It allows for tracking and managing changes to device configurations, helping to ensure that unauthorized changes are detected and addressed.
3. **Security Alerts and Notifications**
   Real-time Alerts: SNMP traps can be configured to send real-time alerts to the NMS in response to specific security events, such as repeated login failures, enabling rapid response to potential threats.

   Event Logging: SNMP can be used to collect logs from network devices, which are crucial for detecting, investigating, and responding to security incidents.
4. **Vulnerability Management**
   Patch Management: An NMS using SNMP can help identify devices that are running outdated firmware or software versions, facilitating vulnerability management through timely patching.

   Compliance Reporting: SNMP can assist in generating reports that demonstrate compliance with security policies and standards, which often include requirements for patching and configuration management.

5. **Access Control**
   SNMP Version 3 (SNMPv3): SNMPv3 includes security features for authentication and encryption, ensuring that SNMP traffic cannot be easily intercepted or manipulated and that only authorized roles and individuals can access the system. Using SNMPv3 can help protect management data and control commands from eavesdropping and tampering.

   By integrating SNMP-based monitoring and management into a comprehensive communications strategy, organizations can enhance their ability to detect, respond to, and prevent security incidents, adding an important layer to their defense-in-depth approach.

Why is QoS important in a network?

The ubiquity of high-speed communication technologies like Ethernet and IP has revolutionized the way we connect and communicate. Many new utility engineers have grown up only knowing the 1 Gbps interface on computers. While these technologies provide tremendous bandwidth and connectivity, they have also introduced challenges in managing QoS for different applications and business processes. High-speed communications technologies like Ethernet and IP can mask QoS in several ways:

1. **Uniform Treatment of Traffic:** Ethernet and IP networks often treat all types of traffic uniformly, using a best-effort approach to packet delivery. This means that both critical and non-critical traffic are handled in the same manner, without regard to their specific QoS requirements. As a result, QoS-sensitive applications may not receive the prioritization they require, unless specific guarantees are purchased from the provider. Individual providers have mechanisms for managing competing traffic on their own networks (for example voice vs video) which may also be available in provider service contracts for grid utilities.
2. **Bandwidth Abundance:** The high-bandwidth capabilities of Ethernet and IP networks can create an illusion of abundant resources, leading to the assumption that QoS is not a concern. However, even in networks with ample bandwidth, contention for resources (throughput) can still occur, particularly

during peak usage periods. Without proper QoS mechanisms in place, critical applications may experience degradation in performance due to competing traffic.

3. **Overprovisioning:** In some cases, network administrators may resort to overprovisioning bandwidth to mitigate potential QoS issues. While this approach can temporarily alleviate congestion, it is often costly and inefficient in the long run. Overprovisioning may also mask underlying QoS issues rather than addressing them directly, leading to suboptimal network performance and resource utilization.
4. **Assumption of Homogeneous Traffic:** Ethernet and IP networks are designed to accommodate a wide range of traffic types and applications. However, this diversity in traffic characteristics can pose challenges for QoS management. Without proper classification and prioritization mechanisms, network operators may struggle to differentiate between different types of traffic, and as a result, apply inappropriate QoS policies.
5. **Lack of Visibility:** In complex network environments, gaining visibility into traffic patterns and performance metrics can be challenging. This difficulty in gaining visibility is exacerbated by the potentially transient and intermittent nature of the possible QoS issues. Without comprehensive monitoring and analytics tools, network administrators may have limited insight into how different applications and services are impacted by QoS issues. This lack of visibility can hinder the timely identification and resolution of QoS-related problems.

While high-speed communication technologies offer significant advantages in terms of bandwidth and connectivity, they can also mask QoS needs by promoting a perception of unlimited resources and uniform treatment of traffic. To address these challenges, network administrators must implement robust QoS mechanisms and monitoring solutions to ensure that critical applications receive the necessary performance and reliability levels.

These networks were originally designed to efficiently transmit data packets from source to destination without prioritizing one type of traffic over another. While this approach works well for many applications, it falls short when it comes to meeting the more demanding QoS requirements of certain applications and business processes.

The increasing convergence of voice, video, and data traffic on a single IP network exacerbates the QoS dilemma. Without proper QoS mechanisms in place, bandwidth-hungry applications can monopolize network resources, leading to degraded performance and poor user experience for critical applications.

QoS policies identify the mechanisms designed to ensure performance and reliability of a communication network.

A Quality-of-Service policy for a communication system refers to a set of rules or mechanisms designed to manage network resources efficiently and to ensure the performance, reliability, and priority of various types of data traffic on the network. The goal of QoS is to provide a better experience to users by prioritizing certain types of traffic, minimizing network congestion, and managing latency, jitter, and packet loss. This is especially important in networks that carry a diverse mix of traffic types, including real-time and near real-time applications alongside traditional data traffic.

Key components and strategies often included in a QoS policy for a communication system include:

Traffic Classification and Marking: Traffic is identified and classified into different categories based on criteria such as application type, source, destination, and service level agreements (SLAs). After classification, traffic can be marked to reflect its priority level.

Traffic Shaping and Policing: Traffic shaping involves adjusting the flow of traffic to meet desired

performance metrics, such as bandwidth limits or delay characteristics, while traffic policing monitors the traffic flow and can drop packets or downgrade their priority if they exceed predefined limits.
Priority Queuing: Packets are placed in different queues based on their priority level. Higher-priority traffic is processed before lower-priority traffic, reducing latency and ensuring timely delivery for critical applications.
Bandwidth Allocation: Bandwidth is allocated to different types of traffic to ensure that high-priority services receive the necessary bandwidth to function optimally, even during times of congestion.
Congestion Management: Mechanisms are put in place to manage and mitigate congestion on the network, ensuring that even during high traffic periods, performance levels for priority services are maintained.
Reliability and Redundancy: Strategies to enhance network reliability and implement redundancy, ensuring continuous service availability and performance even in the case of network failures or disruptions.

Implementing a QoS policy involves both hardware and software solutions, including configuring network devices such as routers, switches, and firewalls to support QoS mechanisms. It's a critical aspect of network management for service providers and businesses, ensuring that network resources are used efficiently and that user experiences are optimized, especially for applications requiring high levels of performance and reliability.

For enterprise networks and provider-specific traffic (such as voice carried by a cable company), network engineers and administrators often deploy some mix of the above QoS mechanisms. However, these QoS mechanisms do not typically persist across carrier boundaries, so service guarantees would need to be negotiated as part of the various provider contracts. Not all services provide QoS standards or guarantees hence it is imperative to investigate which carriers and services may provide some level of QoS agreements. These mechanisms prioritize traffic based on predefined criteria, ensuring that critical applications receive the necessary bandwidth, latency, and reliability while preventing nonessential traffic from overwhelming the network.

Despite these efforts, achieving optimal QoS remains a complex and ongoing challenge in modern networking environments. As the demand for bandwidth-intensive applications continues to grow, network operators must continuously adapt and refine their QoS strategies to meet the evolving needs of their users and applications while maintaining a balance between performance, cost, and scalability. There can be serious consequences to the grid if these challenges are not addressed.

Why is QoS important for the future of the electrical grid?

Communications have become an integrated part of everyday activities that are necessary to support everything from business processes to critical infrastructure operations. QoS provides the performance metrics to make informed decisions to choose and deploy the best communication pathway to meet the operational needs. Understanding and matching operations and process requirements to a communication system with the appropriate QoS characteristics is paramount for successful operations.

QoS is a set of performance metrics that are defined not only by the underlying technology, but by the design and operations of the network offering services. From a QoS perspective, similar services may have significantly different performance with the same technology—for example, if one service has guarantees for failover performance during stressed or transient conditions, and another has failover without guarantees. In this example, outages and maintenance may drastically affect asymmetrical routing, causing variations in jitter, delay, availability, and throughput. QoS needs to be examined in both steady state conditions and degraded conditions, where they may be significantly different. In the utility world, architecting the communications

network utilizing QoS parameters is like adding load shedding plans in the grid.

Today, multiple electric markets exist in the U.S., from vertically integrated, regulated monopolies to cooperatives to deregulated markets. No matter the market, the primary function of these systems is to deliver electrical energy reliably at a low cost. To deliver electrical energy, these systems must work to balance instantaneous power with load every second of every day. This balancing of power and load is choreographed between load, resources, and energy schedules, and depends heavily upon the flow of information within and between components of the grid to navigate the changing conditions. Communication systems provide the underlying connections throughout the grid at the required performance levels to ensure that key processes have the data they need to execute their functions.

Historically, meeting load has been accomplished via scheduling and dispatching a small number of large-scale fossil fuel plants. As a large number of variable renewable generators and "dispatchable loads" continue to enter the grid, the number of systems participating in balancing power and load, the complexity of orchestration, the flow of information and controls, and the dependency on grid communications all increase. As this scaling continues, we need to ensure that internet service providers supporting grid communications implement the QoS levels needed for reliable operation.

QoS should be considered for all critical infrastructure processes focused on real-time operations.

Although QoS characteristics and associated policies and mechanisms are important, they are not all important for every operational process or objective. The key to ensuring that all processes operate as intended is to understand the interaction of the processes with the performance capabilities of the communications system in use. The following several paragraphs will assist the reader in understanding which QoS characteristics are important for each operational process to meet its objectives. It can be considered as a starting point from which to determine which communications products and services are needed and what guarantees and Service Level Agreements (SLAs) might be necessary from the provider.

Automatic Generation Control

The process of Automatic Generation Control (AGC) involves the precise and timely dispatch of generation to fine-tune generator output, aligning it with fluctuating electrical demands to ensure that the interchange power on a control area tie-line remains at the scheduled value. This process is integral to critical grid operations such as frequency regulation, load following, and droop control. Because swift action by the AGC system is sometimes needed to re-balance the electrical grid after a load is dropped or a generation source is lost, the AGC system has a low latency tolerance. For the same reason, it has a lower tolerance to jitter/IPDV and to packet loss since they can cause increased delay.

Synchrophasor Data

QoS considerations for collecting and storing synchrophasor data (IEEE C37.118) include low latency requirements, if the data is being used for real-time situational awareness, to more relaxed latency requirements if the data is being stored for disturbance recording and planning model validation. Synchrophasor data is obtained by Phasor Measurement Units (PMUs). The QoS requirements for PMU data vary based upon application and use. Many PMU applications are used for post analysis where lower QoS is acceptable, however more applications are emerging that are near real-time and require more attention to QoS to support the application. PMU data increases based upon the number of devices, parameters being collected, and sampling rates. Multiple architectures exist to aggregate and distribute data via Phase Data Concentrators (PDCs) utilizing User Datagram Protocol (UDP) and Transmission Control Protocol (TCP) over IP

along with multicast IP. Generally, the PMU data packets are relatively small, however as the number of devices and sampling rates increase so do the requirements on the network for throughput and bandwidth. For applications where PMU data is being used for real-time decision making by grid operators, both the level of security and the availability must be high, such as identical backup circuits that are on different facilities and routes.

Energy Management System (EMS) Monitoring and Control

EMS monitoring and control processes collect status information from substations on breaker status, line status, current, voltage, reactive power, and many other inputs, often in the hundreds of points per location through a Supervisory, Control, and Data Acquisition (SCADA) system. These SCADA systems can also initiate control actions in the substations. Because of these two attributes, monitoring and control, the SCADA system is critical for real-time situational awareness, equipment alarm notifications, and control actions. Besides the SCADA functions, the EMS is also key for marketing and trading functions, such as participating in the Energy Imbalance Market or the Day Ahead Market, due to the critical role it plays in optimizing energy trading decisions. All of these functions drive the security and availability requirements to a high level. The data gathered and stored by these systems is significant and drives the bandwidth and throughput requirements from relatively low at substations to high both at control centers and between control centers and Independent System Operators (ISOs).

Distribution Voltage Control

Distribution voltage control has been handled in different manners by utilities but generally includes strategies in compensating reactive power to improve power factor as power extends radially from the substation. These strategies include employing substation voltage regulators, feeder capacitor banks, load tap changers, conservation voltage reduction, and standards for customer usage. Generally, voltage control is needed as the load changes over the day or seasons as feeders become loaded during peak usage. To this end, many voltage control strategies are self-adjusting over a range or can be deployed over communication architectures such as LTE service, fiber with IP service on the feeder, modems using dial-up, or even lower power wide-area networks for metering. Historically, voltage control has not had sensitive QoS requirements; however as we continue to shift our generation resources to inverter-based resources (IBRs), these requirements will change. Current methodologies for voltage regulation will fall short as IBRs' energy is consumed locally, since the predominant grid following inverters tend to raise voltage levels by improving power factors through reducing load on the feeder. Additionally, as IBRs continue to deploy, their increased level of penetration can shift the flow of power back to the substation, defeating original radial distribution power flow assumptions. To address these changes, communications to both IBR units as well as utility owned equipment on the feeder may become more dependent upon high QoS requirements in latency and bandwidth. Adding to the complexity of voltage control is coordinating feeder re-sectionalization under restoration conditions, below, in FLISR.

Fault Location, Isolation, and Service Restoration (FLISR)

FLISR has evolved with technology in the distribution systems to improve restoration times and recovery of the system. The focus of distribution systems that implement some form of FLISR is to quickly identify the fault location on the feeder and restore, if possible, based upon fault type and ability of the distribution to re-configure around the fault. This type of action requires coordination of software, planning, communication, and sensors. It also involves deploying switching equipment at predetermined and planned locations to be able to effectively "pick-up" load from faulted feeders and transfer it to other feeders. Multiple communication strategies have been developed to integrate these actions depending upon the level of integration of FLISR devices deployed. From a communication QoS perspective, the requirements can vary depending upon the integration strategies and what is being accomplished. The recovery of the feeder load, where possible, is a priority but can range in timeframe from minutes to hours.

Regulation Services and Reserve Services

Operational processes such as regulation services and reserve services make up a significant part of the balancing of generation resources and load. Reserve services reserve a specified capacity to produce or consume power within a specific time frame (e.g., 10 to 30 minutes) and duration (e.g., 1 or 4 hours). Regulation services increase or decrease real power generation or demand against a predefined real power basepoint following a service requestor's signal (e.g., every 2 or 4 seconds). These operational processes drive separate physical functions of balancing supply and demand over different time intervals. Both are important for a stable reliable grid, however the response period for regulation services is much shorter than reserve services. From a QoS perspective, these services may not require high bandwidth and throughput, but availability, packet loss, and latency are important to support the physical objectives targeted by these services. Regulation services require lower latency than reserve services, which may be more tolerant to delays due to the time period over which they operate.

Remedial Action Schemes (RAS)

RAS, also known as Special Protection Systems, are systems that take rapid action to re-balance the electric grid after a disturbance. Pre-programmed algorithms in RAS controllers, normally in the utility's control center, monitor the grid for key combinations of outages or other events, such as loss of certain transmission lines, that together initiate specific balancing actions, such as generator dropping or suspending AGC, to counteract that specific disturbance scenario. Many of these schemes execute within 100-200 milliseconds, with some schemes being faster. As expected, low, deterministic latency is key, as is high availability. Much like line protection, if these systems are delayed or fail to operate, the disruptions triggering the scheme are exacerbated and can cause islanding and separation. Traffic volume is low, but due to the high availability requirements and large ramifications if the control signals do not get through, class-of-service (CoS) marking or priority queuing should be used to prevent packet loss.

These schemes are mainly used in transmission systems, but they could be triggered by large virtual power plants (VPPs) connected to the distribution or transmission system. In addition, automated actions from a RAS scheme could drop renewable generation as part of interconnection agreements with the transmission provider.

Line Protection

Line protection, whether on a transmission line or a distribution line, is normally composed of smaller sets of data packets but can have the most critical time constraints of any grid traffic. On a distribution system, this requirement often relaxes as the line voltage drops, allowing more latency than is allowable for transmission line protection. Even so, when the protection scheme needs to operate due to a fault on the line or some other disturbance, it is critical that the control signals are received quickly and accurately at the location they are needed to re-balance the electrical grid. The availability must be high for any line protection to prevent physical damage to the grid devices and more importantly, to protect the public and workers from hazards to life and safety. Bandwidth and throughput for these signals must be preserved when business functions, video, and other large packet streams are sharing the same data paths. CoS marking, priority queuing, or traffic shaping can be used to ensure that the line protection commands get through the first time, without needing any re-transmission of packets.

Generation Protection

Generation protection does not have the same communication requirements for high-speed clearing of transmission lines. In fact, generation protection is locally monitored and acted upon by the relay and associated generator breaker. The data communication for generation protection falls to communication back to the generator operator and perhaps a central authority or utility EMS for notification of the relay and breaker status, so QoS is not an issue.

Interconnection Metering

Interconnection metering (also known as Interchange Metering) is generally used to compensate neighboring utilities for energy and capacity and typically has the same monthly billing cycle as traditional customers. These meters are used to measure the in or out flow of energy based upon agreements and network conditions. These values are typically separate from EMS signals sent from relays or PMU units that may be examining power, voltage, and current levels at a second-to-second interval to ensure Area Control Error (ACE) is maintained by the regions' reliability coordinators. Hence, communication requirements for interconnection metering have been focused on reliability and availability to ensure the data is retrieved regularly.

Revenue Metering

Revenue meters measure total energy used (kWh) and the rate of power usage (kW), on a regular cadence, at a customer connection point. The meters are highly accurate and must conform to both ANSI C12.20 requirements and those of the utility providing the power. These meters are part of a system that stores these values for a set time period and then retrieves the data as part of a billing and settlement cycle. Additional data may be collected from the same meters and sent locally to the SCADA remote unit, then the SCADA database for incorporation into the EMS system. (See discussion on EMS for this data stream.)

Often revenue meters are polled every few minutes for some uses and once an hour for other uses; however, the data is only recovered from the on-site storage hourly or daily, so latency is not a significant issue. Similarly, bandwidth and throughput are not of high importance since there is ample time to retransmit any dropped or delayed packets. Security is a high priority, since the integrity of the information is important.

What standards and upcoming technologies are associated with QoS?

As noted previously, TDM technologies were meant to be a transition to digital technology, supporting voice traffic, which represented the largest revenue for telecommunication companies. In the process of transitioning to digital voice optimization, TDM provided electric utilities with an opportunity for a communications system that offered dedicated, synchronized timeslots, dedicated bandwidth and most importantly, low and predictable latency, high reliability, and high availability. In short, solving communication QoS requirements to implement their control and protection schemes.

With TDM use declining and IP technologies increasing, new technologies and standards variations will be needed to deliver deterministic QoS capabilities.

Over the past few decades, TDM has increasingly been displaced by IP (Internet Protocol) and Ethernet technologies due to several compelling reasons including cost-effectiveness, scalability and flexibility, improved data rates, and increased vendor support. New technologies and standards variations will be needed to deliver these QoS characteristics and fill this void, especially when using third-party providers. The following examples describe standards and architecture efforts underway to address QoS requirements for IP-based networks.

Time Sensitive Networking (TSN), under the IEEE 802.1 working group, focuses on making Ethernet deterministic. TSN sits on layer 2 of the OSI/ISO Model and adds definitions to guarantee determinism and throughput in Ethernet networks. While this emerging standard has been used in vehicles and industrial processes, it is still being investigated for use in electrical power systems. Unlike standard Ethernet, TSN places an upper limit on latency and controls jitter/IPDV. One example of this is a recent paper [7], where using TSN for differential protection demonstrated low latency and lack of both packet loss and IPDV.

Deterministic Networking (detnet) – The detnet working group, in collaboration with the IEEE802.1 TSN task group, focuses on deterministic data paths that operate over Layer 2 bridged and Layer 3 routed segments.

Detnet is compatible with both TSN and Multiprotocol Label Switching (MPLS) systems [8]. Functionality is provided by encapsulating DetNet flows and applying MPLS labeling. This is particularly significant since many utilities use MPLS with traffic engineering to manage latency for time-critical traffic. Like TSN, this combination puts an upper bound on both latency and jitter/IPDV and leverages the characteristics of MPLS -TE systems. As of January 2024, the DetNet Operations, Administrative, and Maintenance (OAM) functions are still in draft form [9].

Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service Architecture – The L4S architecture enables internet applications to achieve low queuing latency, low congestion loss, and scalable throughput control [10]. The architecture focuses on a new class of L4S congestion controls, using Explicit Control Notification to signal the host to adjust the queue before the queuing delay gets large and packets are dropped. While these L4S packets are identified and queued separately, they can coexist with 'Classic' congestion controls in a shared network without impacting non-L4S data flows. This allows for incremental adoption of this architecture. As of December 2023, wide deployment of L4S has not occurred, perhaps due to a limited number of devices on the market supporting this architecture, but trial deployments are underway [11].

Impact of Physical Layer and Technology on QoS

As indicated previously, the technology and operations of communication system determine the QoS of the overall system. Included in the technology discussion is the physical media used by the system, such as fiber, copper cables, and wireless. As renewables are increasingly deployed on the distribution system, the communications paths traveled from the utility side of the customer meter to the distribution and transmission utilities will be quite varied. The protocols used will be many, and will likely be carried over internet, wireless, fiber optic, and point-to-point radio systems. The QoS performance characteristics that can be attained with these protocols and technologies will vary as well. Fiber optic systems, for example, can provide higher bandwidth than point-to-point radios, but at a potentially higher cost. For more information on these and other technology options, watch for the next white paper in this series, which will examine communications components and technologies.

Conclusion and Upcoming Opportunities

The ubiquity of high-speed communication technologies like Ethernet and IP has revolutionized the way we connect and communicate. Offering unprecedented bandwidth and connectivity, these technologies also pose challenges in ensuring QoS for diverse applications and business operations. As we adapt to new communication technologies, QoS evolves, highlighting its significance not only as a technological function but also as a crucial aspect of network design and operations.

Amidst the ongoing grid transition, the core operational goals of the grid remain unchanged, yet the adoption of new energy resources and technologies necessitates a transformation in the underlying operational processes to maintain grid reliability, resilience, security, and affordability. The effectiveness of this transformation critically depends on the performance of communication systems, which must align with the requirements of these new technologies and processes.

This paper defined seven QoS parameters and discussed the QoS requirements for many common grid processes. The upcoming whitepaper will explore the current landscape of technologies and services, shedding light on their capability to meet these evolving requirements and sustain grid efficiency and reliability in an era of significant change.

References

[1] IEEE, "IEEE Standard for Jitter and Phase Noise," *IEEE Std 2414-2020,* pp. 1-42, 26 Feb. 2021.

[2] J. Redd, "Explaining those BER testing mysteries," *Lightwave magazine,* 31 Aug 2004.

[3] Internet Engineering Task Force (IETF), "IETF RFC 3393 IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)," 17 June 2022. [Online]. Available: https://datatracker.ietf.org/doc/rfc3393. [Accessed 19 Dec. 2023].

[4] International Telecommunications Union, "Recommendation ITU-T Y.1540 Internet protocol data communication service-IP packet transfer and availability performance parameters," Geneva, 2019.

[5] International Telecommunications Union, "Recommendation ITU-T Y.1541 Network performance objectives for IP-based services," Geneva, 2011.

[6] Internet Engineering Task Force (IETF), "IETF RFC 1157 A Simple Network Management Protocol (SNMP)," 1990.

[7] F. Tullenburg and J. L. Dui, "The Effect of Time-Sensitive Networking Onto Performance and Robustness of Power Grid Protection," in *IEEE 11th International Conference on Advanced Computer Information Technologies*, Deggendorf, 2021.

[8] Internet Engineering Task Force (IETF), "IETF RFC 8964 Deterministic Networking (DetNet) Data Plane: MPLS," January 2021. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc8964. [Accessed 7 March 2024].

[9] Internet Engineering Task Force (IETF), "Framework of Operations, Administration and Maintenance (OAM) for Deterministic Networking (DetNet)," 08 January 2024. [Online]. Available: https://datatracker.ietf.org/doc/draft-ietf-detnet-oam-framework/11/. [Accessed 08 March 2024].

[10] Internet Engineering Task Force (IETF), "IETF RFC 9330 Low Latency, Low Loss, and Scalable Throughput (L4S) internet Service: Architecture," January 2023. [Online]. Available: https://datatracker.ietf.org/doc/rfc9330/. [Accessed 08 March 2024].

[11] M. Clark, "The quiet plan to make the internet feel faster," 09 December 2023. [Online]. Available: https://www.theverge.com/23655762/l4s-internet-apple-comcast-latency-speed-bandwidth. [Accessed 08 March 2024].