Latency Implications for Grid Communications

January 31, 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

Our Nation's electric system is evolving rapidly: an increasing variety of new energy resources is being integrated throughout the system while new sensing, computing, and control technologies promise to facilitate more efficient, flexible system operation. The shift towards a renewable, carbon-free grid necessitates the seamless coordination of a multitude of distributed energy resources (DERs), including renewable generation (solar, wind), energy storage systems (batteries, electric vehicles), and demand response. These changes mark a profound departure from the conventional paradigm of grid operation to one that will rely more heavily on reliable, secure two-way communication to deliver timely, accurate data throughout the system. The evolved grid will feature ubiquitous sensors collecting data and distributed computing at utility control centers and consumer premises to process it. Secure communications with consistent, well-defined latency and adequate bandwidth will enable sharing this data to facilitate effective coordination between DER and grid operators, ensuring a resilient and reliable energy infrastructure.

A secure communications system protects the end-to-end physical pathway that transports data from origin to destination. That pathway may involve different transmission methods, such as optical fiber, copper wire, and wireless technologies; transport diverse data, including grid state information and control messaging; and use a variety of analog and digital formats. Securing this end-to-end communications pathway—which is essential for reliable grid operations—involves preventing unauthorized access and monitoring traffic to identify anomalous activity without compromising the confidentiality, integrity, or availability of the data. Communications security methods complement cybersecurity approaches used to protect data at origin and destination.

This series of papers, part of the Secure Pathways for Resilient Communications (SPaRC) program, follows an initial series that covered high-level descriptions of communications challenges facing the evolving grid. This series of whitepapers delves into various technical dimensions of grid transformation as they relate to communications: creating secure communication pathways, understanding, and managing data communication requirements, and the critical roles of latency, bandwidth, and throughput in grid communications. The series begins with latency and its impacts on grid communications. Three additional papers will follow. One will explore communications quality of service (QoS) parameters and associated characteristics such as throughput and bandwidth. Another will discuss the implications of using the Internet for transport of grid communications and whether existing communications products fit the requirements. The final white paper will explore the applications and implementation of data communications requirements on the necessary network architecture.

Introduction

In an era marked by the rapid transformation of the electrical grid, understanding the role of data communications, especially latency, has become paramount. As the grid shifts towards a carbon-free model with increased integration of distributed energy resources (DERs), including renewables, energy storage, and demand response, effective coordination becomes essential for grid stability and efficiency. Traditional, utility-owned communication networks are giving way to more distributed and diverse systems and the demands on communications systems are evolving. This paper emphasizes the need for low-latency communication systems to manage this complexity, ensuring resilient and reliable energy infrastructure. The discussion extends to the challenges of coordinating DERs and the performance requirements of future communication systems in grid services and sets the stage for a comprehensive discussion on the need for flexible, low-latency, secure communication networks in managing future grid operations.

Data Communication Characteristics

Data communications systems evolved rapidly in the 1990s, enabling the growth of the Internet and, later, the Internet of Things (IoT). Industry has shifted from synchronized time-division multiplexing (TDM) communications built to support voice (1960-1990s) to asynchronous packet switched communication systems built to support data (1990s-2020s). As communication systems have evolved, so have the characteristics used to describe them. Previously, in a synchronized TDM voice world, we considered characteristics such as bandwidth, bit-error rate, and timing synchronization. In today's asynchronous data communication world, we often characterize data communications in terms of *bandwidth*, *throughput, packet loss, availability, security*, *latency,* and *jitter*.

Bandwidth

Bandwidth is the maximum amount of data that can be transmitted over a link where *throughput* refers to the actual amount of data transmitted over time between the sender and receiver. It is often the primary parameter discussed when ordering a communication circuit. Adequate bandwidth is essential for efficiently transmitting large volumes of data from multiple sources, such as smart meters and sensors. Insufficient bandwidth can lead to delays in data transmission, affecting real-time decision-making in grid operations. Bandwidth is typically measured in megabits per second (Mbps).

Throughput

Throughput refers to the rate of the message delivery over a communication channel. In an IP-based network, which is dynamically routed, the throughput and bandwidth may not be equal depending the packet stream characteristics (packet size, packet rate, and bandwidth available) and the architecture of the network. In contrast, a synchronized TDM network is channel-based, with static capacity per channel and by carrier, so throughput and bandwidth are equal.

Packet Loss

Packet loss refers to the loss or non-delivery of entire data packets during transmission over a network. A packet is a unit of information, typically a small segment of a larger message, that is packaged up for transmission across a network. Packet loss occurs when one or more of these packets do not reach their intended destination, which can result in errors in the larger message. An example of this might be an intermittent voice call over a cellular network. Packet loss can result from various issues in network operation, including network congestion, hardware failures, buffer overflows, or errors in routing decisions. Packet loss can also be due to physical factors including noise, interference, signal attenuation, and distortion in the communication channel. Loss of data packets during transmission can result in incomplete or delayed information, affecting the accuracy of grid management decisions.

Availability

Availability, in the context of data 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. High availability of communication networks ensures continuous monitoring and control of grid assets. Unavailability can lead to gaps in data, potentially causing operators to miss critical changes in grid conditions or be unable to send control commands when needed.

Security

Security, in the context of communications can refer 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 data communication systems. Secure communication channels are crucial to protecting the grid. Secure channels in data communications can rely upon the inherent information provided by network devices, such as switches, routers, and data transmission equipment. Breaches can compromise the control of grid assets and lead to operational disruptions.

Latency

Latency refers to the delay in the transmission of data from the sender (source) to the receiver (destination) over a network or communication channel. It is the time it takes for data to travel from one point to another and is typically measured in units of time. Latency is a critical factor in various types of networked applications, and its impact depends on the specific use cases. Reducing latency is a key goal in network design and optimization to enhance the performance of the application and user experience of various communication and data applications. Latency can be considered one of the parameters of QoS. Latency can also be impacted by the insertion of security, reliability, or other QoS measures. Low latency is critical for real-time control and coordination of grid assets, especially for quick response needs like load balancing and frequency regulation. High latency can delay responses to grid fluctuations, risking stability. Latency's relationship to the changing electric grid and its operation is the focus of this whitepaper.

Jitter

Jitter is the variability in latency over time. In other words, jitter is the fluctuation or inconsistency in the delay of 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) [1]. In grid operations, consistent latency is crucial for synchronizing actions across the network. If there is significant jitter, the variability in communication delays can lead to problems in coordinating actions, such as the timing of control signals for grid stability. High jitter in a network can undermine the benefits of low latency by making the system less predictable and reliable. Jitter can result in poor quality voice services. or in relay mis operations

Latency Explained

Latency can be described as how much time it takes for data to travel from one point to another. In digital applications with data communications, total end-to-end latency is the sum of each of the individual latency components. Most of us experience latency in our day-to-day online activities: when viewing a webpage, making a purchase, watching a video, or playing a game, we sometimes see a spinning icon or must wait for an application to complete an action before proceeding to the next step. This delay, which is an example of round-trip delay (Figure 1), includes end-to-end latency in ***both*** directions as well as the time it takes for host processing on the opposite node. Sometimes end-to-end latency also includes retransmission of data due to some error. When trying to reduce latency, it is important to recognize contributions from both the host and the data networking system.

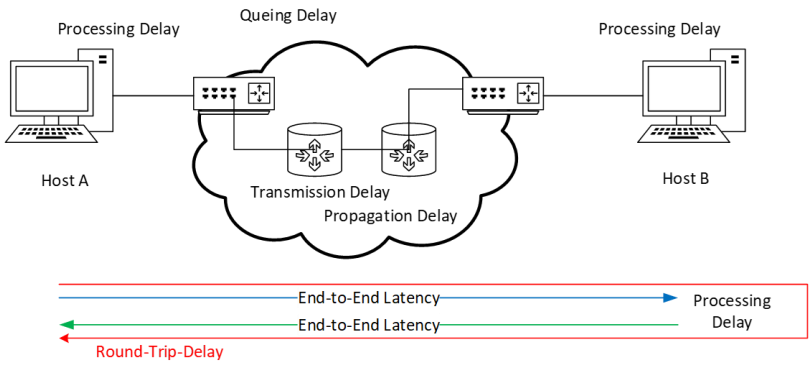


Figure 1: Round-Trip Delay or Round-Trip Time (RTT)

End-to-end latency is commonly broken down into four components: *propagation delay, transmission delay, queueing delay,* and the portion of *processing delay* not attributable to the host.

Propagation delay is the amount of time a bit on the link needs to travel from the source to the destination, where the speed is dependent on the communications medium. Propagation delay is difficult to alter since it depends on the underlying physics and distance.

Transmission delay is the time from when the first bit of a file reaches a link to when the last bit reaches the link. The transmission delay is calculated as the size of the file divided by the data rate of the link. Transmission

delay is a relatively small factor in overall latency.

Queueing delay occurs when packets are held in a buffer on a network device and is dependent on factors such as the number of packets arriving in a time interval, transmission capacity, and the size of the queue. In networked systems, data packets may be placed in queues or buffers at various points along the network path, waiting their turn for transmission. Queueing delays can be large especially in cases of network congestion.

Processing delay includes the time it takes for routers, switches, or other network devices to process and forward data packets. It can also include the host processing of the data and packetization. It also encompasses any data processing or protocol-related delays incurred by the sender and receiver as they prepare, parse, or interpret data.

Several factors affect these latency delay components: the communication protocol; physical media and media access protocols; network size, design, and architecture; network traffic volume; and communication equipment performance and configuration. Optimizing or minimizing latency within a data communication network requires a robust design that considers each of these factors in meeting traffic requirements, and even then, nondeterministic behavior of networks in operation result in latency variation (jitter). Latency and jitter in data communications can have a significant impact in applications where near-real-time communication and decision-making are critical—such as electric grid operation. The level of acceptable latency varies depending on the specific application, but reducing latency is often a priority to ensure efficient and effective communication and control.

Using ICMP to measure latency characteristics.

Observing IP networks as traffic levels and routing paths change demonstrates variation in latency. The Internet Control Message Protocol (ICMP) can provide an estimated measure of latency of an IP network. ICMP messages can also be used for diagnostic or control purposes or generated in response to errors in IP operations. The ICMP Echo request (ping) measures round-trip delay for the network at a particular time. Figure 3 shows results of a ping command where the average round-trip time was 13 ms, with the range of 12 ms to 14 ms. As high-capacity and high-bandwidth capability devices have been deployed, additional capabilities like Simple Network Management Protocol (SNMP) exist on networking devices that can also help determine latency and overall health of the network.

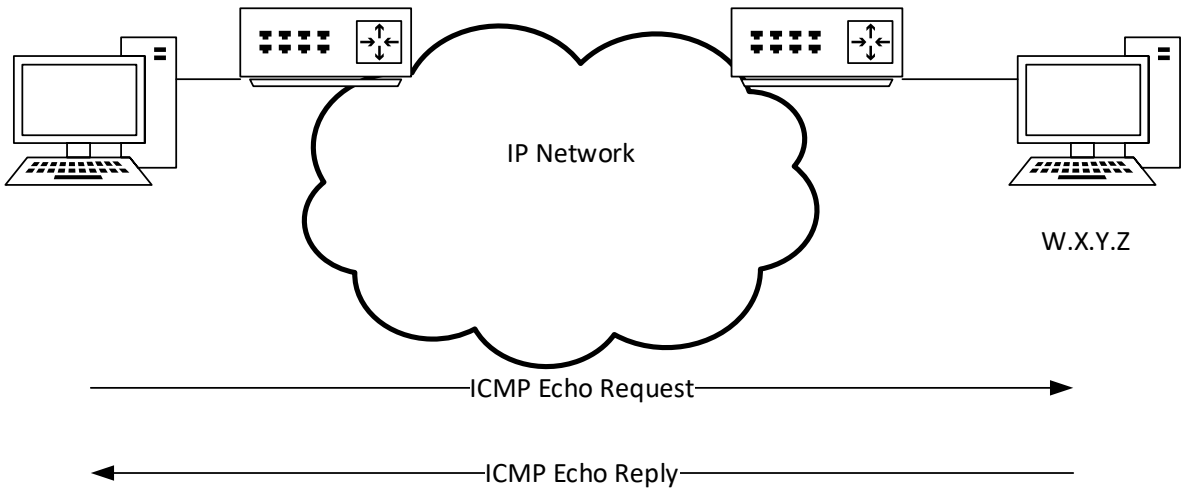


Figure 2: ICMP Ping

It is important to note that the latency (13 ms) measured represents latency at the time of the measurement and that, while ICMP requires minimal processing, it does contribute host processing delay in the round-trip time measurement. Because multiple factors can impact latency such as traffic levels and congestion, network design, equipment health, and network outages, the measured latency will vary over time.

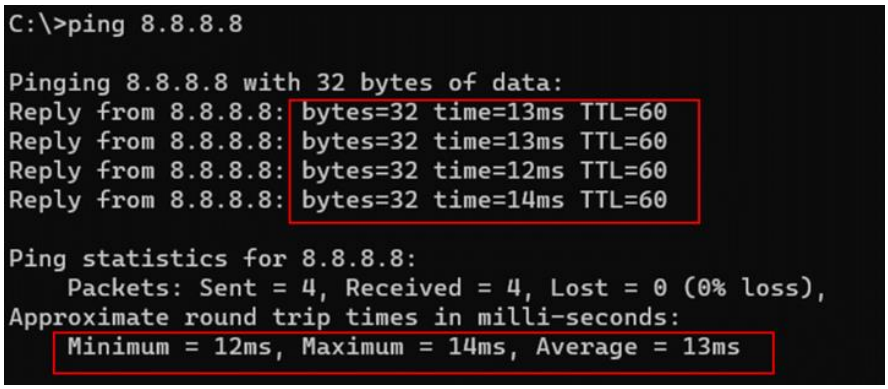


Figure 3: RTT for ICMP Ping

Impacts of Internet Protocols on Data Latency

As portions of grid data communications networks move to commercial networks, it is important to consider how commercially available tools and components treat latency. Today's predominant data communication protocols include Ethernet and Internet Protocol (IP), which make up the internet core and most industrial, commercial, and residential networks. These protocols have become the de facto components of computers and embedded systems on devices. Ethernet (IEEE 802.3 [2]) and TCP/IP (IETF RFC 9293 [3] and 791 [4]) have provided a robust basis to support many types of services on the internet.

In the early development of data link protocols, significant proliferation and cost reductions advanced data rates rapidly from 10Mbps to 10Gbps (a factor of 1000). Deployment of 10Gbps Ethernet followed rapidly by 100Gbps networks displaced SONET OC-192 synchronous TDM networks by rapidly integrating with local area networks and reducing cost and complexity. The result is a network with endpoints capable of 10/100/1000Mbps, but absent the deterministic latency that was provided by TDM.

The transition from TDM networks to IP networks signified a shift from predictable, fixed latency to unpredictable, variable latency.

The low cost and capability of Ethernet and IP have allowed us to rapidly deploy networks to support common applications we use daily including near-real-time applications such as video, voice and video calls, financial transactions, remote control, and online gaming. Rapid growth and expansion of devices will continue to put upward pressure on the bandwidth and throughput capabilities of existing networks, impacting latency and performance of the network. Over the last two decades we have seen advancement in other network devices such as firewalls and software network devices that are inspecting packets based upon application and protocol. These devices are frequently deployed as a cybersecurity measure to improve defense in depth strategy. This type of inspection does not happen within the communication stack and comes at a cost of processing latency at intermediate nodes. Thus, for latency-critical applications, we need to also consider the impact of cybersecurity measures on data communication characteristics such as latency and throughput. One of many goals is an architecture that supports internet connectivity to any location over multiple technologies, with delay and bandwidth requirements dependent on the mission the network is supporting. The usefulness of asynchronous networks like the internet for electric grid data communications depends on their ability to achieve consistent latency.

Utility Operational Processes

Within an electric energy delivery system or electric grid, a core set of operational processes are applied to ensure electricity is delivered to end customers. Figure 4 represents typical electric utility operational processes with performance requirements relating to data communication latency characteristics and tolerances.

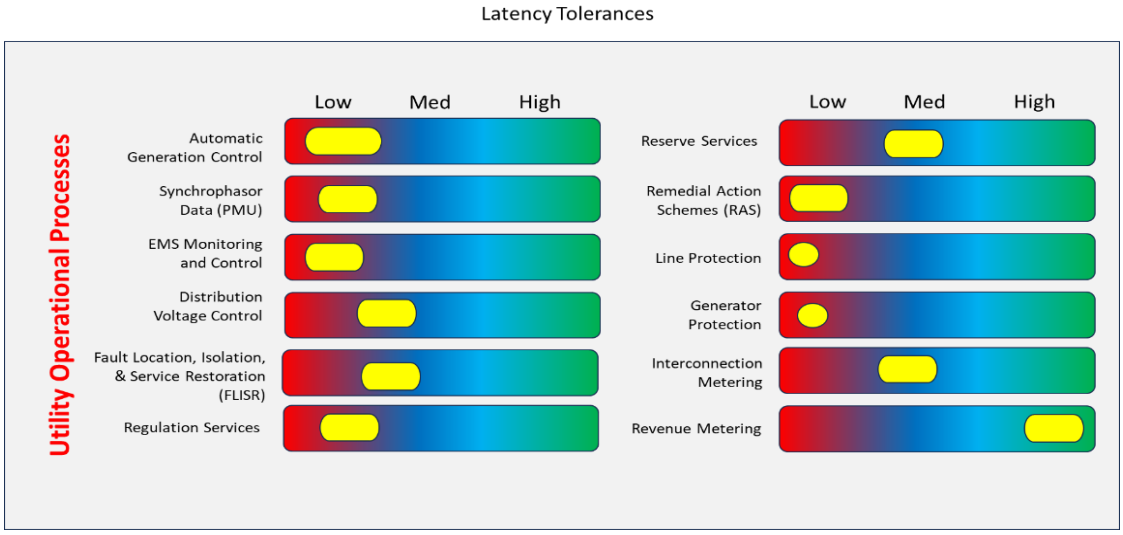


Figure 4: Relative Operational Process Latency Tolerances

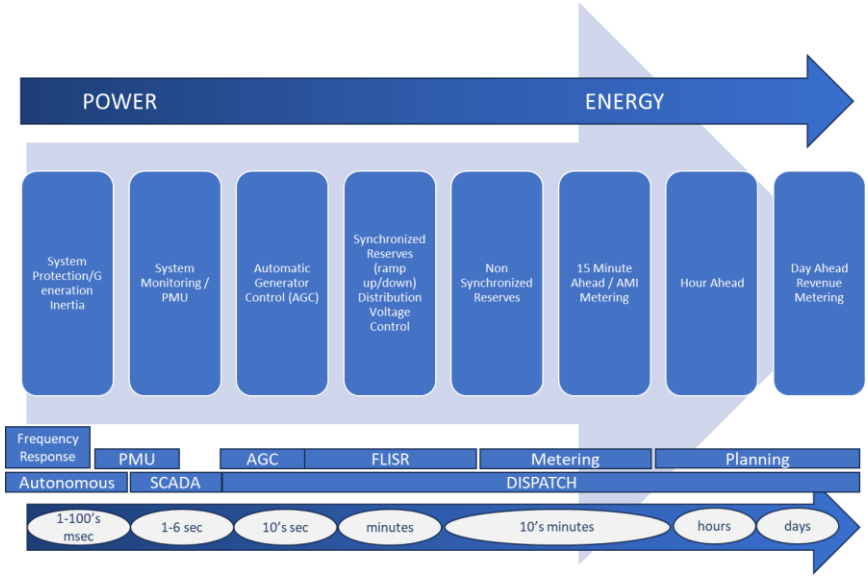
As evident in Figure 4, above, latency is a key parameter, especially for protective relay systems. These systems automatically open circuit breakers to de-energize transmission and distribution lines during an abnormal event, such as a tree in the line or a power pole being hit by a car. For these protection systems, a low-latency system is critical for safety, wildfire prevention, and the protection of multi-million-dollar substation equipment. The longer the transmission or distribution line carries power, the more damage is done.

> Latency is a key performance communication parameter for coordinating DER assets to ensure resilient and reliable energy infrastructure.

Latency plays a pivotal role in the coordination of generation assets within an electric grid. The process of Automatic Generation Control (AGC) involves the precise and timely dispatch of generation through autonomous generator actions (like inertia and governor response) and directives from utility control centers. AGC's primary function is to fine-tune generator output, aligning it with fluctuating electrical demands to maintain area control error near zero. This process is integral to critical grid operations such as frequency regulation, load following, and droop control. The significance of low latency cannot be overstated, as it is crucial for the swift rebalancing of the grid to prevent equipment damage and mitigate the risk of area separation, thereby ensuring reliable grid operation.

In both regulated and deregulated markets, and regardless of whether the electric grid relies on fossil fuels or renewable sources, the principles of coordinating and dispatching generation to meet load demands remain constant. Traditionally, utilities have built and managed their own data communication networks to ensure reliable grid operation. However, as we shift towards a carbon-free grid, integrating a growing number of diverse renewable energy assets becomes increasingly complex. These assets, often under varied ownership and spread over large areas, necessitate near-real-time operational processes for efficient coordination. This is especially challenging with variable renewable generation, which differs significantly from dispatchable, fuel-based generation methods.

Figure 5 provides another view of grid operations and planning processes at multiple time scales. Electric utilities have modernized these processes with digital technologies supported by data communications.



Metering, for example, was originally completely manual. As new technologies have been adopted to include the automation of billing and advanced meter systems, they also became highly dependent upon data communications technologies. Most operational processes use a combination of centralized control and localized autonomous decision making.

Figure 5: Electric Utility Operational Processes on Timescale

Each process has different information requirements that translate to different performance characteristics of the underlying data communication network. For example, gathering metering data from customers is not as time sensitive as system protection, but the bandwidth required to collect all customer data is often larger than the data required for communication among protective relays. Overall response time requirements (latency) for synchronized reserves and AGC, however, may be similar.

Latency, and its consistency, in data communications is crucial in this context. Effective large-scale coordination hinges on understanding and maintaining specific latency requirements in the network, encompassing communications, hosts, and cybersecurity. Neglecting these requirements can significantly impact operational processes. Therefore, as the grid evolves, continuously evaluating the performance characteristics—particularly latency—of data communication networks against data requirements is fundamental to maintaining the grid's reliability and resilience.

Latency Work in Standards and Potential Upcoming Technologies

The impact of latency on robust communications systems has been recognized by the standards community and various potential improvements are being discussed. Today, several efforts continue work in the standards and development space to help improve issues with latency in IP/Ethernet networks but are at different stages of the technology readiness scale. Standards and working groups continue to examine potential solutions to improve the latency issues including reducing and improving deterministic delay. Some examples include:

Time Sensitive Networks (TSN) is a developing set of standards under IEEE 802.1 that focuses on making Ethernet deterministic. It will provide "guaranteed packet transport with bounded latency, low packet delay variation and low packet loss." TSN sits on layer 2 of the OSI/ISO Model and adds definitions to guarantee determinism and throughput in Ethernet networks. This work is being accomplished by the TSN Task Group under the IEEE 802.1 Working Group [5].

Deterministic Networking (DetNet) – The DetNet Working Group, under the Internet Engineering Task Force (IETF), focuses on "deterministic data paths that operate over Layer 2 bridged and Layer 3 routed segments, where such paths can provide bounds on latency, loss, packet delay variation (jitter), and high reliability." This group is currently concentrating on both wired and wireless systems, but for privately administered LANs and WANS as opposed to the internet. The Working Group collaborates with the IEEE802.1 Time-Sensitive

Networking (TSN) Working Group as well as other related IETF Working Groups [6].

Low Latency, Low Loss, Scalable Throughput (L4S) Internet Service Architecture - This architecture "enables Internet applications to achieve low queuing latency, low loss, and scalable throughput control." The L4S architecture primarily concerns incremental deployment mechanisms that support a new class of L4S congestion controls which coexist with 'Classic' congestion controls in a shared network. These mechanisms aim to ensure that the latency and throughput performance using an L4S-compliant congestion controller is usually much better (and rarely worse) than performance would have been using a 'Classic' congestion controller, and that competing flows continuing to use 'Classic' controllers are typically not impacted by the presence of L4S. This is especially interesting because it can be deployed selectively to bottlenecks on the network without deploying on the entire network[7][8].

Conclusion

As the electric grid evolves, the shift from centralized, utility-owned telecommunications systems to a more distributed model is becoming increasingly evident. This change, crucial for integrating a diverse array of generation and load resources, demands telecommunications systems that are not only less centralized but also agile and capable of maintaining consistent latency. This evolution in communication technology is key to effectively managing the grid's complexity and ensuring its reliability and resilience. The forthcoming papers in this series will delve deeper into the potential configurations of these future systems and their implications for grid management.

References

[1] "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)," IETF Datatracker RFC 3393 – 2002. Accessed: Dec. 19, 2023. [Online.] Available: https://datatracker.ietf.org/doc/html/rfc3393

[2] "IEEE Standard for Ethernet," IEEE Std 802.3-2022. Accessed: Dec. 19, 2023. [Online.] Available: https://standards.ieee.org/ieee/802.3/10422/

[3] "Transmission Control Protocol (TCP)," IETF Datatracker RFC 9293-2022. [Online.] Available: https://datatracker.ietf.org/doc/html/rfc9293

[4] "Internet Protocol," IETF Datatracker RFC 791-1981. Accessed: Dec. 19, 2023. [Online.] Available: https://datatracker.ietf.org/doc/html/rfc791

[5] "Time-Sensitive Networking (TSN) Task Group," IEEE 802.1. Accessed: Dec. 19, 2023. [Online.] Available: https://1.ieee802.org/tsn/

[6] "Deterministic Networking (detnet)," IETF Datatracker. Accessed: Dec. 19, 2023. [Online.] Available: https://datatracker.ietf.org/wg/detnet/about/

[7] "RFC 9330: Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture", Association for Computing Machinery. Accessed: Dec. 19, 2023. [Online.] Available: https://dl.acm.org/doi/10.17487/RFC9330

[8] "Low Latency, Low Loss, and Scalable Throughput (L4S) Internet Service: Architecture," IETF Datatracker RFC 9330. Accessed: Dec. 19, 2023. [Online.] Available: https://datatracker.ietf.org/doc/rfc9330/