

UC Davis

UC Davis Electronic Theses and Dissertations

Title

A Simulation-based Study for Achieving Deterministic and Reliable Large-Scale Data Transfers in a Scientific Network

Permalink

<https://escholarship.org/uc/item/3816n8wz>

ISBN

9798189226885

Author

Kumbarahally Lakshminarayana, Vijeth

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

A Simulation-based Study for Achieving Deterministic and Reliable Large-Scale Data Transfers in
a Scientific Network

By

VIJETH KUMBARAHALLY LAKSHMINARAYANA
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Computer Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Dipak Ghosal, Chair

Alexander Gamero-Garrido, Co-Chair

Biswanath Mukherjee

Committee in Charge

2026

To my friends and family

Abstract

Modern scientific workflows are data-intensive, geographically distributed, and time-sensitive, demanding reliable, low-latency data transfers across high-performance networks. While current solutions, such as Energy Sciences Network ([ESnet](#))’s On-Demand Secure Circuits and Advance Reservation System ([OSCARS](#)), offer static bandwidth reservation, they lack the fine-grained, deterministic guarantees essential for emerging real-time applications. To address this gap, we explore the applicability of Cyclic Queuing and Forwarding (CQF), a deterministic scheduling mechanism traditionally designed for Local Area Networks (LANs), to the unique challenges of Wide Area Networks (WANs). We propose a novel architecture that integrates Cyclic Queuing and Forwarding ([CQF](#)) with multi-path routing to enhance reliability and evaluate it through a comparative simulation against [OSCARS](#) on a realistic [ESnet](#) topology, considering both time-triggered and event-triggered flows. Our results demonstrate that with proper parameter tuning, [CQF](#) can achieve a performance comparable to that of [OSCARS](#) in terms of latency, acceptance rate, and network utilization. As one of the first efforts to adapt [CQF](#) for WAN-scale scientific workflows, this study highlights its potential to enable deterministic communication for high-volume, long-distance scientific networks.

Acknowledgments

I express my deepest gratitude to my advisor, Professor Dipak Ghosal. His patient guidance and insightful mentorship have been crucial to my journey. Through every challenge and discovery, his unwavering support and encouragement not only helped me overcome obstacles, but also inspired me to think more critically and push the boundaries of this research. This thesis would not have been possible without him.

My sincere thanks go to my thesis committee members, Prof. Alexander Gamero-Garrido, and Prof. Biswanath Mukherjee. I am deeply grateful for the time and expertise they dedicated to my work. Their invaluable feedback and thought-provoking questions have immeasurably strengthened this thesis.

I am also immensely grateful to my research collaborator and friend, Carolina Minami Oguchi. Our countless brainstorming sessions were a constant source of inspiration and energy. Her sharp insights and collaborative spirit made the entire research process not just productive but genuinely enjoyable.

I would like to express my sincere gratitude to my research collaborator and friend, Nishant Acharya, for his invaluable partnership on the Internet Exchange Point (IXP) indispensability project. His sharp technical insights and collaborative spirit were essential in conducting the active measurement campaigns. The many hours we spent brainstorming network resilience immeasurably strengthened this thesis.

To my colleagues and friends in the research group, thank you for creating such a supportive and stimulating environment. The shared knowledge, discussions, and camaraderie were an invaluable part of my experience.

I am also grateful for the administrative staff of the Computer Science department, especially Tamara Riggs, Jessica Stroller, and Alyssa Bates, whose guidance helped me seamlessly navigate the complexities of academic requirements.

Finally, to my friends, who have been my pillars of support throughout the ups and downs of my life, thank you! Your encouragement, advice and much-needed moments of laughter kept me grounded and reminded me of the world beyond my research. Your belief in me has meant more than words can say.

Contents

Abstract	iii
Acknowledgments	iv
List of Figures	vii
List of Tables	viii
List of Acronyms	ix
Chapter 1. Introduction	1
1.1. Motivation: The Data Deluge in Modern Science	1
1.2. The Challenge: Limitations of Current Network Architectures	1
1.3. Proposed Solution and Contributions	2
1.4. Thesis Outline	3
Chapter 2. Background	4
2.1. Scientific Networks and Their Demands	4
2.2. On-Demand Secure Circuits and Advance Reservation System (OSCARS)	5
2.3. A Deterministic Network Architecture	6
Chapter 3. Related Work	10
Chapter 4. Motivation	12
4.1. OSCARS	12
4.2. Need for Reliable and Deterministic Networking	12
4.3. Need for Multi-path Flow Control	13
Chapter 5. Methodology	14
5.1. Traffic Patterns	14
5.2. Network Load	15

5.3. Bandwidth Allocation Algorithm	16
5.4. Performance Metrics	17
Chapter 6. Results and Discussion	18
6.1. Cycle Time vs Latency and Acceptance Rate	18
6.2. Comparison of Different Schemes	19
6.3. Link Utilization vs Betweenness Centrality	20
6.4. CQF with Multi-path Routing	21
Chapter 7. Conclusion and Future Work	24
7.1. Conclusion	24
7.2. Future Work	24
Chapter 8. Analysis of Critical Internet Infrastructure	26
8.1. Introduction and Motivation	26
8.2. Related Work	27
8.3. Data Sources	27
8.4. Methodology	28
8.5. Results	33
8.6. Limitations	37
8.7. Future Work	38
Bibliography	39

List of Figures

2.1 Core concept of Cyclic Queuing and Forwarding (CQF)	6
2.2 Illustration of minimum cycle time in Cyclic Queuing and Forwarding (CQF)	8
2.3 Illustration of queue states in Cyclic Queuing and Forwarding (CQF)	9
6.1 Flow latency and acceptance rate vs. cycle time T	18
6.2 Workload vs Flow Latency and Acceptance Rate	19
6.3 Workload vs median link utilization	19
6.4 Betweenness centrality vs link utilization	20
6.5 Flow Acceptance rates vs workload for different numbers of paths used	22
6.6 Flow Latency vs workload for different numbers of paths used	22
6.7 FRER Success rates and Overall Flow acceptance rates and Latency vs workload	23
8.1 Pipeline to find the group of indispensable Internet Exchange Points (IXP)s	29
8.2 IXP Reachability Example	30
8.3 Observing the IXPs with a set of low-impact IXPs	34
8.4 Observing the IXPs with a set of high-impact IXPs.	34
8.5 Observing sparse IXPs with a single high-impact IXP.	34

List of Tables

5.1 Parameters used in our experiments	15
8.1 Verification Table For Italy	36
8.2 Verification Table for Australia	37
8.3 Verification Table for South Africa	37
8.4 Correctness Scores	38
8.5 Bangladesh Traceroute Hits	38

List of Acronyms

WAN: Wide Area Network	2
LAN: Local Area Network	2
ESnet: Energy Sciences Network	iii
CQF: Cyclic Queuing and Forwarding	iii
TSN: Time Sensitive Networking	2
DetNet: Deterministic Networking	10
OSCARS: On-Demand Secure Circuits and Advance Reservation System	iii
ET Flow: Event Triggered Flow	15
TT Flow: Time Triggered Flow	15
FRER: Frame Replication and Elimination for Reliability	22
IXP: Internet Exchange Points	vii
AS: Autonomous System	26
CAIDA: Center for Applied Internet Data Analysis	28
RIPE: Réseaux IP Européens Network	29

CHAPTER 1

Introduction

1.1. Motivation: The Data Deluge in Modern Science

Modern science is undergoing a transformation. Scientific workflows are becoming increasingly complex, data-intensive, and geographically distributed. Across the globe, high-resolution instruments and large sensor networks generate massive amounts of data that must be transferred, stored, and analyzed in real time. These workflows frequently connect different types of facilities, such as experimental stations, data centers, and high-performance computing platforms. Because of this, high-throughput and low-latency communication has become a critical tool for enabling new scientific discoveries [1].

The need for time-sensitive communication is growing across many different scientific fields. In Nuclear Physics, for example, workflows analyze ever-growing datasets, making it more practical to use remote high-performance computing resources instead of relying only on local infrastructure. In High Energy Physics, fast and reliable communication helps scientists detect rare physical events by allowing massive data volumes to be captured and processed with very little delay. The Fusion Energy Sciences domain often relies on a small number of experimental facilities, which creates a need for long-distance, high-throughput data transfers to remote computing centers. Similarly, in Basic Energy Sciences, autonomous experimentation systems driven by artificial intelligence require tight coordination between data acquisition and processing systems. These examples underscore the importance of transmitting and processing data with minimal latency for scientific progress to continue.

1.2. The Challenge: Limitations of Current Network Architectures

Traditional Wide Area Network Communication protocols often rely on a best effort model, which does not provide any delivery and latency guarantees. This has pushed the scientific community toward deterministic networking solutions that can offer predictable and bounded communication behavior.

Current approaches often depend on reservation-based systems, such as the On-Demand Secure Circuits and Advance Reservation System, or [OSCARS](#), used in the Energy Sciences Network (ESnet) [2,3]. [OSCARS](#) allows researchers to reserve bandwidth and provision dedicated circuits for their data transfers [3]. While this is a step in the right direction, its configuration is often coarse-grained and static, which limits the responsiveness and efficiency of the network. For instance, [OSCARS](#) supports reservations that can last for multiple years, which shows it operates at a broad control level rather than managing traffic in real time. Once a reservation is made, its configuration typically remains fixed, which can lead to the inefficient use of network capacity.

1.3. Proposed Solution and Contributions

To address these limitations, we turn to solutions from the domain of Local Area Networks (LANs), specifically Time Sensitive Networking ([TSN](#)) architectures [4]. We propose applying Cyclic Queuing and Forwarding ([CQF](#)) [5], a [TSN](#) mechanism that provides fine-grained, per-flow control orchestrated at millisecond timescales. The precise, cycle-based scheduling offered by [CQF](#) directly counters the static, coarse-grained nature of [OSCARS](#), making it a compelling candidate for high-performance scientific workflows.

However, adapting a Local Area Network ([LAN](#))-centric protocol like [CQF](#) to a scientific Wide Area Network ([WAN](#)) presents significant challenges. Unlike [LANs](#), [WANs](#) are characterized by long-distance topologies with high propagation delays, creating a large bandwidth-delay product. In this environment, the cost of packet loss recovery is extremely high for large data transfers, as retransmissions can easily lead to congestion and packet re-ordering. This work is one of the first comprehensive efforts to adapt and evaluate [CQF](#) for a [WAN](#)-scale scientific network, considering these real-world propagation delays and link bandwidths. We further investigate integrating multi-path routing with [CQF](#) to enhance data transfer reliability across geographically dispersed nodes. The key contributions of this work are as follows:

- We provide a detailed description of Cyclic Queuing and Forwarding and evaluate its suitability for use in wide-area scientific networks. We determine how sensitive the network’s performance is to the [CQF](#) cycle time, a critical parameter for its operation.

- We demonstrate that for the [ESnet](#) topology, a cycle time in the range of a few hundred milliseconds achieves good performance in terms of latency and the acceptance rate of time-sensitive data flows.
- We present a simulation-based comparative analysis of [CQF](#) and [OSCARS](#) using a realistic network topology and traffic patterns from [ESnet](#), showing that with tuned parameters, [CQF](#) achieves comparable performance.
- We propose the integration of multi-path routing with [CQF](#) to enhance the reliability of data transfers in a [WAN](#) setting.

1.4. Thesis Outline

The remainder of this thesis is organized as follows. Chapter 2 will provide a detailed background on the technologies central to this work, including the [ESnet](#) architecture, deterministic networking principles. Chapter 3 provides a review of related literature. Chapter 4 outlines the motivation behind the research. Chapter 5 will describe the proposed system architecture and the simulation methodology used to evaluate its performance. Chapter 6 will present and analyze the results of the simulation experiments. Chapter 7 will conclude the thesis with a summary of the findings and a discussion of potential directions for future research. Finally, Chapter 8 presents a study on identifying critical components in public networks by quantifying the indispensability of Internet Exchange Points ([IXPs](#)).

CHAPTER 2

Background

This chapter provides the necessary background on the foundational concepts and technologies relevant to this thesis. We begin by introducing the Energy Sciences Network (ESnet) as a representative large-scale scientific network and discuss the unique requirements of its time-sensitive workflows. We then describe its current bandwidth reservation system, [OSCARS](#), and its limitations. Subsequently, we delve into the principles of deterministic networking, with a focus on Cyclic Queuing and Forwarding ([CQF](#)), the core mechanism of our proposed architecture. Finally, we situate our research within the context of existing literature, highlighting the gaps this work aims to address.

2.1. Scientific Networks and Their Demands

2.1.1. The Energy Sciences Network (ESnet). The Energy Sciences Network (ESnet) is a high-performance Wide Area Network (WAN) funded by the U.S. Department of Energy’s Office of Science and managed by Lawrence Berkeley National Laboratory (LBNL) [2]. It serves as the primary network infrastructure for the U.S. national laboratory system, connecting numerous research facilities across the United States and extending to select locations in Europe, including several Research and Educational (R&E) networks. [ESnet](#) is engineered to support the massive data requirements of modern science, providing high-capacity links capable of transferring approximately 20 petabytes of data each month [6]. The network underpins a wide array of scientific endeavors by linking experimental sites, data centers, and supercomputing facilities.

2.1.2. Time-Sensitive Scientific Workflows. A report on [ESnet](#) usage patterns identifies time-sensitivity, data integration, and long-term data access as common data transfer patterns [1]. This need arises from the necessity to transmit data from experimental facilities to computational resources in real-time, enabling low-latency analysis and timely decision-making in dynamic workflows. The importance of time-sensitive communication is evident across several key scientific domains:

- **Nuclear Physics:** Workflows involve incremental analysis over growing datasets, making it more feasible to leverage remote high-performance computing resources rather than relying solely on local infrastructure.
- **High Energy Physics:** Time-sensitive data transfer is crucial for detecting rare or anomalous physical events, which requires that massive data volumes be captured and processed with minimal delay.
- **Fusion Energy Sciences:** The limited number of experimental facilities necessitates long-distance, high-throughput data transfers to remote computing centers, for which time-sensitive communication is essential.
- **Basic Energy Sciences:** The integration of autonomous, AI/ML-driven experimentation systems depends on tight coordination between data acquisition and processing systems, facilitated by low-latency networking.
- **Biological and Environmental Research:** Progress in this domain benefits from the ability to share raw data from geographically distributed sensors and experiments across multiple facilities in near real-time.

These examples illustrate the growing importance of time-sensitive communication across diverse scientific domains. As experimental workflows become increasingly data-intensive, distributed, and reliant on real-time feedback, the ability to transmit and process data with minimal latency is essential for scientific progress. As a high-performance research network, [ESnet](#) plays a central role in enabling these capabilities by bridging experimental facilities and computational resources. Continued investment in low-latency high-throughput networking infrastructure will support the next generation of data-driven science, accelerating discovery through responsive, integrated workflows across the research environment.

2.2. On-Demand Secure Circuits and Advance Reservation System ([OSCARS](#))

To manage large-scale scientific data flows, [ESnet](#) utilizes the On-Demand Secure Circuits and Advance Reservation System ([OSCARS](#)) [3]. [OSCARS](#) is an open-source software that automates the provisioning of dedicated network paths, a task that was traditionally manual and time-consuming. It provides end-to-end circuit provisioning, allowing users to request and configure network paths

with specific attributes such as guaranteed bandwidth, latency, and routing preferences. Furthermore, [OSCARS](#) supports dynamic circuit management, enabling connections to be modified, renewed, or terminated in minutes to meet evolving research demands.

2.3. A Deterministic Network Architecture

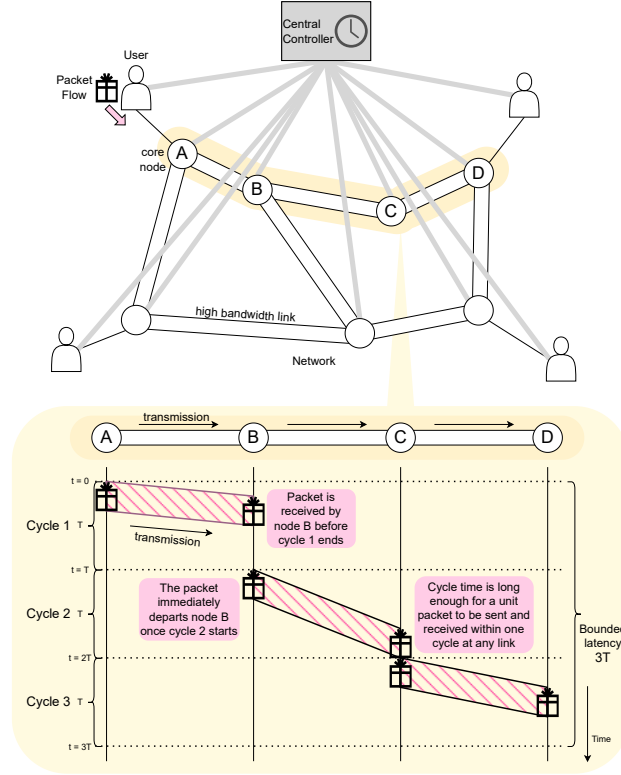


FIGURE 2.1. Core concept of Cyclic Queuing and Forwarding ([CQF](#)). A flow is transmitted from source node A to destination node D via nodes B and C under the control of a centralized controller. All nodes are time-synchronized, and the network uses high-bandwidth links. Each hop occurs in one cycle of duration T , resulting in a bounded end-to-end delay of $3T$ for the three-hop path.

2.3.1. Overview. To meet the growing demand for deterministic and reliable data transfer in scientific workflows, we propose an architecture for routing and scheduling data across large-scale scientific networks. The key approach in our work is the integration of [CQF](#) into a [WAN](#). As illustrated in Figure 2.1, [CQF](#) ensures bounded latency regardless of variations in delay between nodes. There is a central controller that has global visibility of the network and communicates with nodes over a control plane where signaling delays are assumed to be negligible relative to the

data plane cycle times. It synchronizes the cycles across all nodes in the network and manages all the flows in the network. Flows can be sent from one to another using high-bandwidth links in the network, which operates in cycles. The latency for a flow to hop from one node to another is always equal to the cycle time, making the transfer deterministic. The central controller rejects flows before transmission if it is impossible to maintain determinism.

2.3.2. Cyclic Queuing and Forwarding (CQF). Cyclic Queuing and Forwarding (CQF) [5] is a core algorithm used in Time-Sensitive Networks (TSN) to achieve deterministic and low-latency packet delivery with guaranteed timing constraints. It operates on a cycle-based transmission principle, where all network nodes are tightly time-synchronized and follow a shared global cycle clock, and each node has a pair (or pairs) of queues.

A central controller controls all flows in the network. It is responsible for computing paths for incoming flows, ensuring that the strict latency guarantees defined by TSN requirements can be met. Once this is successful, the traffic flow is admitted and the controller reserves the necessary bandwidth on all the links in the path. Then, the flow moves deterministically from node to node, one hop per cycle. The total end-to-end delay for any given packet is easily calculated as the product of the fixed cycle time and the number of hops in the selected path. This mathematical simplicity is one of the key advantages of the CQF, as it enables precise delay estimation and supports guaranteed latency and jitter requirements.

Cycle time T , a key parameter in the CQF algorithm. The whole network operates based on this cycle time, especially when flipping the states of the queues in each node. It is typically set within a range of milliseconds to hundreds of milliseconds to align with the latency requirements of the TSN standards. The cycle time must be long enough to allow adjacent nodes to send and receive a certain amount of flow within a single cycle. To be specific, it must be longer than any minimum possible cycle time $T_{i,j}^{min}$, where $e_{i,j} \in E$ in the network (V, E) . $e_{i,j}$ represents a link from node i to node j . V is a set of all nodes in the network, and E is a set of all links in the network.

$$(2.1) \quad T \geq \max_{(i,j)}(T_{i,j}^{min})$$

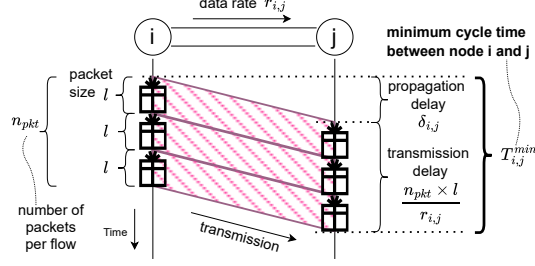


FIGURE 2.2. Illustration of minimum cycle time in Cyclic Queuing and Forwarding (CQF). Cycle time must be longer than any minimum cycle time $T_{i,j}^{min}$ where link $e_{i,j} \in E$ of the network (V, E) . V is a set of all nodes in the network, and E is a set of all links in the network. $T \geq \max_{(i,j)}(T_{i,j}^{min})$.

Here, minimum possible cycle time of the link $e_{i,j}$ is denoted as $T_{i,j}^{min}$, and it is calculated as follows

$$(2.2) \quad T_{i,j}^{min} = \frac{n_{pkt} \times l}{r_{i,j}} + \delta_{i,j}$$

where n_{pkt} is the number of packets per flow, l is the fixed size of the packet, $\delta_{i,j}$ is the propagation delay of link $e_{i,j}$, and $r_{i,j}$ is the data rate. Substitute n_{pkt} with the largest number of packets per flow among all flows. This is visualized in Figure 2.2.

Within each node, there is a pair of queues per egress port. One is the receive queue, and the other is the send queue. The roles alternate exactly at the end of each cycle. When a queue is in the receive state, it receives the flows from adjacent nodes. A node can send no more flows than it can fit in the queues and be sent and received within a single cycle. The central controller pre-determines routing and scheduling to meet this requirement. When a queue is in the send state, flows are sent to the adjacent node corresponding to the queue. All the flows received in the previous cycle must be sent and received by the next node within the next cycle. Figure 2.3 describes how a flow is sent from a node to the next node in the CQF mechanism.

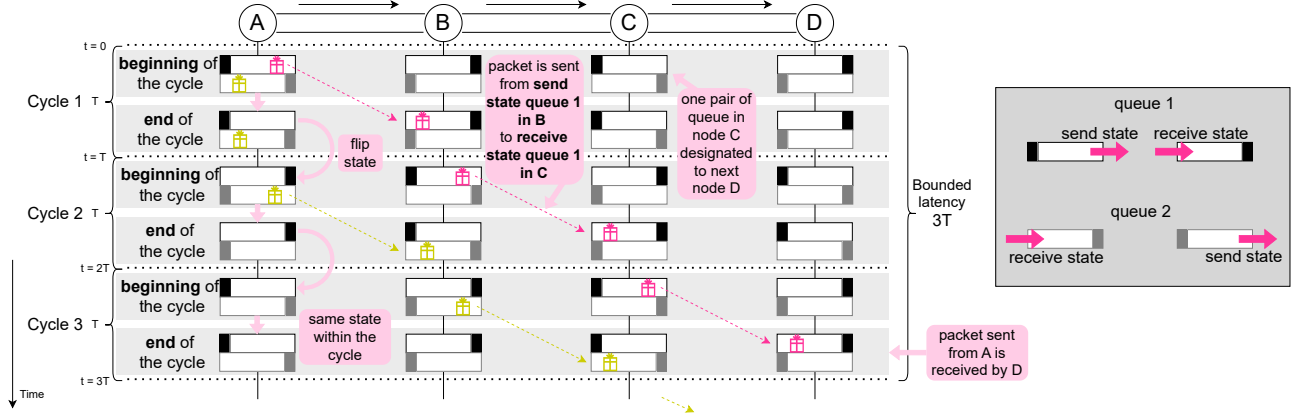


FIGURE 2.3. Illustration of queue states in the Cyclic Queuing and Forwarding (CQF) algorithm. Each node has two queues per egress port that alternate between send and receive states at each cycle boundary. The flow is sent from queue 1 in node A at the beginning of Cycle 1 and received by queue 1 in node B by the end of Cycle 1. At the next cycle, Cycle 2, the flow is sent from queue 1 in node B at the beginning of Cycle 2 and received by queue 1 in node C by the end of Cycle 2. At the next cycle, Cycle 3, the flow is sent from queue 1 in node C at the beginning of Cycle 3 and received by queue 1 in node D by the end of Cycle 3.

2.3.3. Multi-path Flow Control. Effective network control requires precise management of the path selection for each flow and the timing of its transmission. There are two primary routing strategies: single-path routing, where a flow is transmitted along a single end-to-end path from source to destination, and multi-path routing, where a flow may be duplicated or distributed across multiple disjoint or partially joint paths. Multi-path routing may offer advantages in load balancing, fault tolerance, and reduced end-to-end latency.

CHAPTER 3

Related Work

CQF [5] enables predictable, bounded-latency communication by enforcing strict timing control over packet transmission at each node. CQF has traditionally been applied in LANs, such as industrial automation and automotive networks [7], where time synchronization between nodes is achievable and low propagation delays are typical.

Recent studies have begun to explore the challenges of extending Time-Sensitive Networks' principles to WANs or Large-Scale Deterministic Networks (LDNs). While most existing work focuses on smaller or less geographically distributed networks, Simon et al. [8] introduced deterministic principles in a Metropolitan Area Network (MAN), which is more geographically spread than a LAN but less than a WAN. Peng et al. [9] introduced deterministic principles in a WAN context, but their study is limited to a relatively small network of nine core nodes and employed Cyclic Specified Queuing and Forwarding (CSQF), which does not require nodes to be synchronized at each cycle level through precise distributed clock synchronization as in CQF. A document by the Internet Engineering Task Force (IETF) [10] discusses how TSN at Layer 2 and Deterministic Networking (DetNet) at Layer 3 interoperate, extending deterministic principles beyond local TSN domains into larger DetNet-enabled networks. To the best of our knowledge, there are no existing studies on the applicability of CQF or CQF-like algorithms to a WAN, especially for scientific networks such as ESnet that support complex time-sensitive scientific workflows with very large data transfers.

Furthermore, most of the related studies focus on Time-Triggered traffic [11, 12, 13, 14]. In contrast, Event-Triggered flows, which occur spontaneously and without an explicit pattern of arrival time, have received less attention despite their relevance to real-world scientific applications. Additionally, prior work on large-scale data transfers has tackled challenges from various angles, including mathematical analysis of scheduling algorithms [15], joint optimization of routing and scheduling [16], control mechanisms [17, 18], and online scheduling strategies [19]. Some of these studies have considered CSQF rather than the CQF.

To address these gaps, our work applies CQF to a geographically distributed scientific WAN based on ESnet’s architecture. We incorporate event-triggered flows and examine CQF’s potential in a realistic, large-scale environment. We aim to enhance deterministic and reliable high-volume data transfers across long-distance scientific networks. Our work builds on these foundations by adapting CQF to ESnet, which is a WAN that carries a huge amount of scientific data that requires time-sensitive communication. To the best of our knowledge, this is one of the first comprehensive efforts to simulate and benchmark CQF and OSCARS under WAN-scale scientific workloads using realistic traffic types and topologies.

CHAPTER 4

Motivation

4.1. OSCARS

[OSCARS](#) in [ESnet](#) supports dynamic provisioning and long-term reservations of bandwidth-aware network circuits, with reservation durations ranging up to multiple years. Allowing year-scale reservations indicates that [OSCARS](#) operates at a coarse-grained control level, reserving a certain amount of bandwidth on paths without specifying the exact time that flows are transmitted. Once a reservation is established, its configuration is often static, leading to inefficient capacity utilization. In contrast, [CQF](#) provides fine-grained, per-flow control, orchestrated at millisecond time scales, allowing precise scheduling and assignment of links to flows. This enables tight latency bounds, deterministic forwarding, and real-time resource allocation, leading to more efficient routing and a greater possibility of better utilization of the network. [CQF](#) addresses the static and coarse-grained limitations of [OSCARS](#) and is well-suited for high-performance scientific workflows requiring predictable, low-latency data transfer.

4.2. Need for Reliable and Deterministic Networking

For [WANs](#) such as [ESnet](#), there are multiple reasons why a reliable and deterministic networking paradigm is imperative. First, scientific workflows are complex and require distributed processing. They involve data collected at remote locations that need to be transported to high-performance computing sites, where it needs to be processed and saved in data storage at various locations. Reserving computational and storage resources is crucial for processing pipelines that require deterministic and reliable data transfer. Additionally, as mentioned previously, [ESnet](#) is a [WAN](#) with very high bandwidth links and large propagation delays, and the bandwidth-delay product is very high. This places large and unpredictable buffering requirements on the routers. Furthermore, any packet loss due to buffer overflow can significantly impact the flow’s throughput and latency.

While there have been several advancements in traffic engineering strategies to optimize routing and scheduling, our work does not aim to develop novel traffic engineering solutions. Instead, we

focus on applying CQF to a scientific network. Our approach leverages established techniques for flow assignment and path selection, but the contribution lies in demonstrating how CQF can be adapted, rather than in proposing new traffic engineering methods.

4.3. Need for Multi-path Flow Control

Introducing multi-path routing to the CQF mechanism in a WAN is critical for enhancing network utilization, reducing end-to-end latency, and potentially increasing redundancy. By distributing traffic across multiple paths, it can leverage alternative routes to shorten transmission delays, particularly in WANs where long physical distances increase latency. Furthermore, it applies to integrating CQF with mechanisms such as Frame Replication and Elimination for Reliability (FRER). It can increase network redundancy, ensuring that time-sensitive traffic meets strict latency and loss requirements even in network faults or heavy-load traffic.

CHAPTER 5

Methodology

To assess the feasibility of integrating [CQF](#), we perform a comparative performance analysis that simulates and contrasts three approaches to route and schedule flows:

- (1) [OSCARS](#) (Static Resource Allocation) (Baseline): This works similarly to the circuit reservation scheme used in [OSCARS](#). A fixed bandwidth is assigned to each flow at the beginning, and the flows transmit at this bandwidth regardless of the availability of additional bandwidth. Flows are rejected if there is no available bandwidth.
- (2) [OSCARS](#) (Dynamic Resource Allocation): Assigned bandwidth depends on the available bandwidth and flow deadlines. Flows use more bandwidth than initially required when available and can use less when running ahead of their schedule.
- (3) Cyclic Queuing and forwarding (Proposed system)

We simulated the [ESnet](#) topology using NetworkX [20] and ran discrete event simulations for routing and scheduling using SimPy [21]. The network comprises 68 geographically distributed nodes connected by 197 links. The low bandwidth links used for administrative purposes were excluded along with some nodes, leaving only the nodes with high incoming and outgoing bandwidth (100 and 400 Gbps). Parameter settings are listed in Table 5.1.

5.1. Traffic Patterns

We model two types of flows:

- **Event-triggered concurrent flows (ET Flows):** Concurrent data transfers representing multiple data streams are initiated in response to specific events or experiments. An [ET Flow](#) request consists of a flow size and a deadline, t_{fc}^{deadline} . These two features determine the minimum bandwidth, b_{fc}^{required} needed to meet the deadline.
- **Time-triggered sensor flows (TT Flows):** Persistent constant rate flows representing periodic sensor-generated data that follow a predefined schedule. [TT Flow](#) requests consist of the required bandwidth and duration.

Parameter	Description	Value
Number of episodes	Total number of episodes in an experiment	10
Number of iterations	The number of Event Triggered Flow (ET Flow) iterations in each episode	50
Maximum number of ET Flows	Maximum number of ET Flows that could be transmitting concurrently in the network	Depends on the load factor
Number of Time Triggered Flow (TT Flow)	Number of TT Flow in each episode	20
Required bandwidth of an ET Flow	Minimum bandwidth required by an incoming ET Flow determined by the flow size and deadline	Random between 7000 and 9000 Mbps
Required bandwidth of a TT Flow	Bandwidth requested by an incoming TT Flow	Random between 3000 and 5000 Mbps
Packet size	Size of a packet (aligned with the size of a jumbo frame)	9000 bytes
Deadline	Deadline of a flow	Larger flows have a higher deadline
Latency	Time taken to fully transmit a flow	Should not exceed the deadline
Cycle time	Duration of cycles in CQF	Between 6 milliseconds to 10 seconds
Scheduling window	Time window for which the path bandwidths are reallocated	1 second

TABLE 5.1. Parameters used in our experiments (**TT Flow**: Time-triggered flow, **ET Flow**: Event-triggered flow)

Traffic is organized into episodes. In each episode, n_{TT} **TT Flows** run for the entire duration. **ET Flows** launch in randomized batches of n_{ET} flows every t_{interval} seconds. The sources and destinations are chosen randomly while ensuring they are reachable. The flow size is scaled according to the load factor, which is outlined below.

5.2. Network Load

We define a metric called load factor to quantify the network load. It is based on the link betweenness centrality, which is a measure that quantifies the importance of a link in terms of how many shortest paths pass through it. Betweenness centrality $B(e_{i,j})$ of link $e_{i,j} \in E$ is defined as follows [22].

$$(5.1) \quad B(e_{i,j}) = \sum_{s,t \in V} \frac{n_{\text{shortest path}}(s,t|e_{i,j})}{n_{\text{shortest path}}(s,t)}$$

Where $n_{\text{shortest path}}(s, t)$ denotes the number of paths from node s to node t , and $n_{\text{shortest path}}(s, t|e_{i,j})$ denotes the number of paths that pass through a link $e_{i,j}$.

We define load factor, $L(e_{i,j})$ as the ratio of the average utilized bandwidth $\bar{b}_{f_c}^{\text{utilized}}$ to the total available bandwidth $b_{e_{i,j}}^{\text{available}}$ on link $e_{i,j}$.

$$(5.2) \quad L(e_{i,j}) = \frac{B(e_{i,j}) \times \bar{b}_{f_c}^{\text{utilized}} \times n_{\text{concurrent flow}}^{\text{max}}}{b_{e_{i,j}}^{\text{available}}}$$

$$(5.3) \quad n_{\text{concurrent flow}}^{\text{max}} = n_{\text{ET}} \times \left(\frac{\max_{f_c \in \text{ET}}(t_{f_c}^{\text{deadline}})}{t_{\text{interval}}} \right)$$

Intuitively, $n_{\text{concurrent flow}}^{\text{max}}$ expresses the maximum possible number of flows that can occupy a link. In our experiments, we scale traffic volume based on the proportion of bandwidth it could consume on the median betweenness centrality link.

5.3. Bandwidth Allocation Algorithm

We define scheduling interval to denote the time between scheduling epochs. At each scheduling epoch, we follow Algorithm 1 to calculate the available bandwidth in a path. When a new flow is scheduled, it is routed through the path with the highest available bandwidth, provided the available bandwidth exceeds the required bandwidth.

In the static version of [OSCARS](#), the bandwidth allocation is static, and all flows constantly transmit at a fixed bandwidth. In [CQF](#) and dynamic [OSCARS](#), the bandwidth assigned to each accepted [ET Flow](#) f at time t on a path p is defined as follows, with a flow size of $|f|$ and a deadline of t_f^{deadline} :

$$(5.4) \quad b_{f,t,p}^{\text{assigned}} = \frac{|f|}{t_f^{\text{deadline}}} + \frac{b_p^{\text{available}}}{n_{\text{ET}}}$$

Intuitively, the minimum bandwidth assigned to each accepted flow is the average bandwidth needed to transmit all the packets before the deadline, which is the flow size divided by the remaining time.

In addition, any residual bandwidth of the path is distributed equally among all [ET Flows](#).

Algorithm 1 Calculate Available Bandwidth Along a Path

Require: $path, network, f_s \in F_{TT}, f_c \in F_{ET}, packet_size$

```
1:  $path\_bw \leftarrow +\infty$ 
2: for all  $link \in path$  do
3:    $b_{F_{TT}}^{required} \leftarrow 0$  ▷ bandwidth required by TT Flows on the link
4:   for all  $f_s$  in  $link$  do
5:      $b_{F_{TT}}^{required} \leftarrow b_{F_{TT}}^{required} + b_{f_s}^{required}$ 
6:    $b_{F_{ET}}^{required} \leftarrow 0$  ▷ bandwidth required by ET Flows on the link
7:   for all  $f_c$  in  $link$  do
8:      $b_{F_{ET}}^{required} \leftarrow b_{F_{ET}}^{required} + \frac{|f_c|}{t_{deadline_{f_c}}}$ 
9:    $link\_bw \leftarrow network.bandwidth[link]$ 
10:   $avail\_bw \leftarrow link\_bw - (b_{F_{TT}}^{required} + b_{F_{ET}}^{required})$ 
11:  if  $avail\_bw < path\_bw$  then
12:     $path\_bw \leftarrow avail\_bw$ 
13: return  $path\_bw$ 
```

5.4. Performance Metrics

5.4.1. Acceptance Rate. Acceptance rate is the ratio of the number of accepted flows to the total number of requested flows. In all three network architectures, **TT Flows** are only accepted if the controller can find paths to distribute the flows, while ensuring that they get the requested bandwidth. **ET Flows** are accepted if the data can be transmitted without exceeding the deadline.

5.4.2. Latency. Latency is the total amount of time spent (in seconds) transmitting a flow from the source to the destination, calculated from the moment the flow is accepted. The lower the value, the better the performance.

5.4.3. Network utilization. We define network utilization as the ratio of the amount of bandwidth utilized in a link to its total capacity. In **CQF** and dynamic **OSCARS**, the utilized bandwidth is defined as the sum of the bandwidth assigned to **TT Flows** and the minimum bandwidth required to transmit all the **ET Flows** by their deadline.

CHAPTER 6

Results and Discussion

6.1. Cycle Time vs Latency and Acceptance Rate

Figure 6.1 shows the latencies and acceptance rates as a function of the CQF cycle time T at a load factor of 0.8. The latencies are slightly higher for the lower cycle times because most of the cycle time is wasted to account for the propagation delay. For longer cycle times, latency increases because a flow may finish transmitting early in a cycle, but the network resources allocated to it remain locked until the cycle completes. This inefficient use of bandwidth can delay the scheduling of subsequent flows. This effect is even more pronounced in the acceptance rates. We can observe the importance of carefully choosing the cycle time for a higher acceptance rate and lower latency.

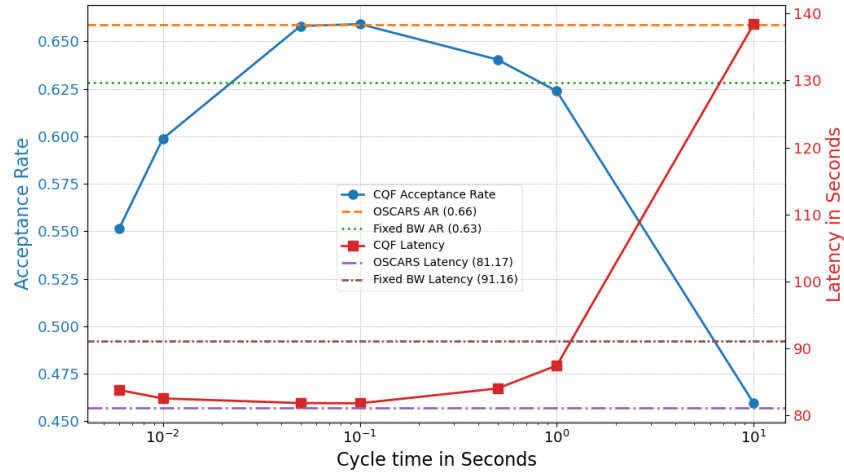


FIGURE 6.1. Flow latency and acceptance rate vs. cycle time T . The flow latencies and acceptance rates for different values of T at a load factor of 0.8. The results show an optimal performance range for the cycle time between 10 ms and 100 ms, with performance degrading outside this window. At very short or long cycle times, latency and acceptance rate degrade due to propagation delay and bandwidth underutilization, respectively. This highlights the significance of cycle time in CQF.

6.2. Comparison of Different Schemes

We evaluated the performance of the different schemes for varying workloads. Figure 6.2 shows the average flow latency and acceptance rates for different workloads.

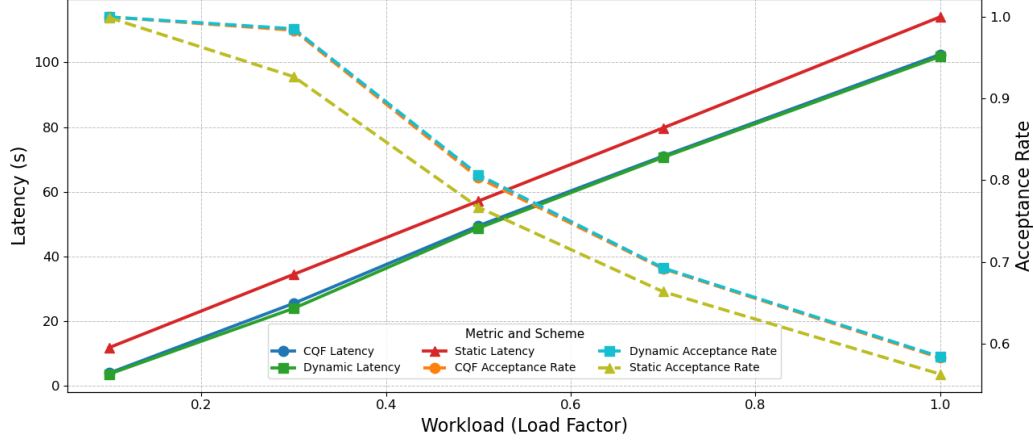


FIGURE 6.2. Workload vs Flow Latency and Acceptance Rate. CQF and the dynamic reservation scheme achieve lower latency and higher acceptance rates than static reservation. This is enabled by more flexible resource scheduling, allowing flows to utilize excess bandwidth when available and reduce usage when ahead of schedule.

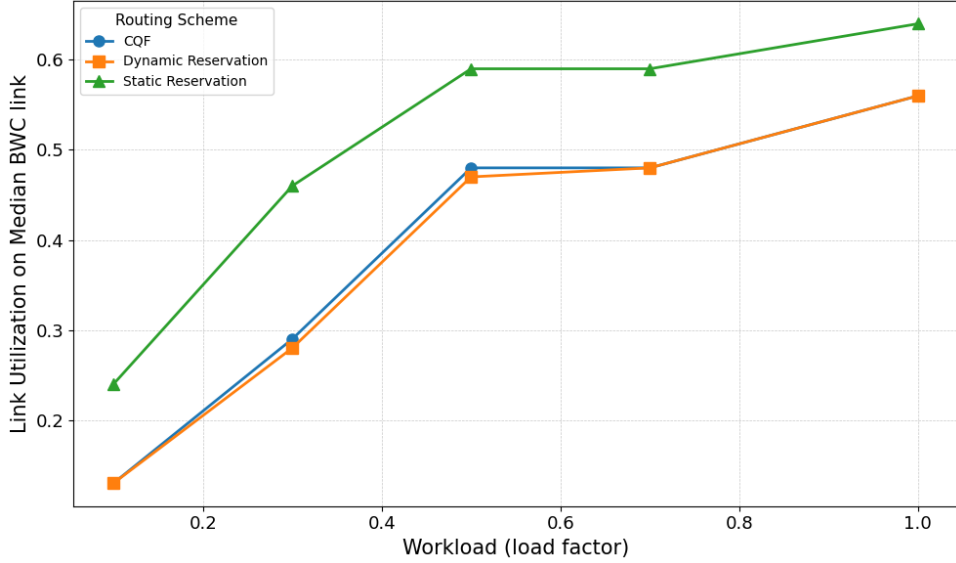


FIGURE 6.3. Workload vs median link utilization. CQF and dynamic reservation exhibit lower utilization than static reservation, demonstrating improved load balancing.

CQF and the dynamic reservation scheme achieve lower latency and higher acceptance rates than static reservation by flexibly allocating bandwidth. Rather than strictly reserving the initial request, they adapt allocations to workload demands, sometimes below the requested rate if deadlines are still met. This allows more flows to be admitted while keeping latency low. A similar trend is observed in the bandwidth utilization on the median betweenness centrality link as seen in Figure 6.3. These results show that both CQF and the dynamic reservation scheme support communication of scientific workflows.

6.3. Link Utilization vs Betweenness Centrality

We observed a positive correlation between the bandwidth utilization of links and their betweenness centrality. We also noticed that the lower bandwidth links have high utilization even on lower centrality. This trend can be observed in the regression lines in Figure 6.4.

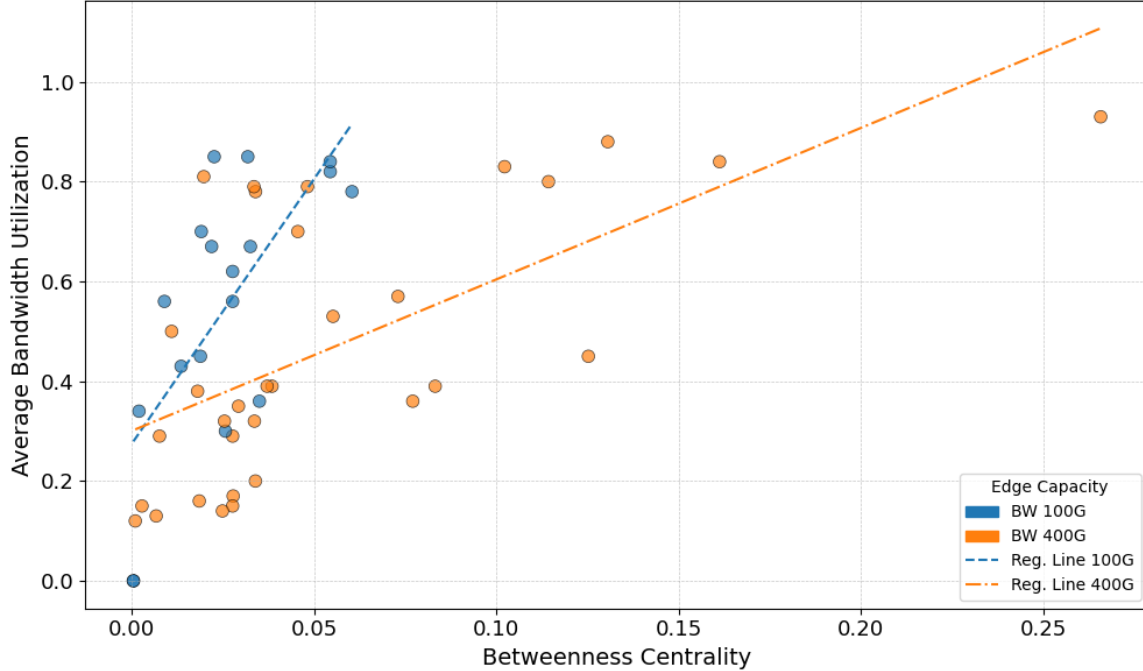


FIGURE 6.4. Betweenness centrality vs link utilization. The positive correlation indicates that links traversed by more paths experience higher utilization. Links with a smaller capacity tend to experience higher utilization.

6.4. CQF with Multi-path Routing

6.4.1. Multi-path vs single-path CQF. An investigation into multi-path routing was conducted to determine if distributing traffic could improve performance. In this experiment, for a given flow, traffic and bandwidth reservations were divided evenly across a set of k paths, where k was varied from one to three. The results, shown in Figure 6.5 and Figure 6.6, indicate that this multi-path approach did not yield a significant performance improvement over the single-path ($k=1$) baseline in either acceptance rate or latency.

This outcome can be attributed to the naive traffic engineering strategy employed. Simply splitting traffic evenly is insufficient in a heterogeneous Wide Area Network like [ESnet](#) for several reasons. The paths between any two nodes are rarely symmetrical; they often differ in hop count, total propagation delay, and, most importantly, bottleneck bandwidth. When a flow is split evenly, its overall completion time is dictated by the slowest and most constrained path. A single path with high congestion or a longer hop count can nullify the potential benefits of other, more efficient paths, resulting in negligible overall performance gains.

A more sophisticated traffic engineering approach would be necessary to unlock the true benefits of multi-path routing in a [CQF](#) environment. Such an approach would require the central controller to be path-aware, making intelligent decisions based on the characteristics of each available route. For example, instead of an even split, the controller could distribute traffic proportionally based on the bottleneck bandwidth of each path, assigning a larger portion of the flow to higher-capacity routes. This transforms the task from a simple division into a complex optimization problem that must be solved in real-time. Developing these advanced, path-aware scheduling algorithms for multi-path [CQF](#) remains a promising direction for future work.

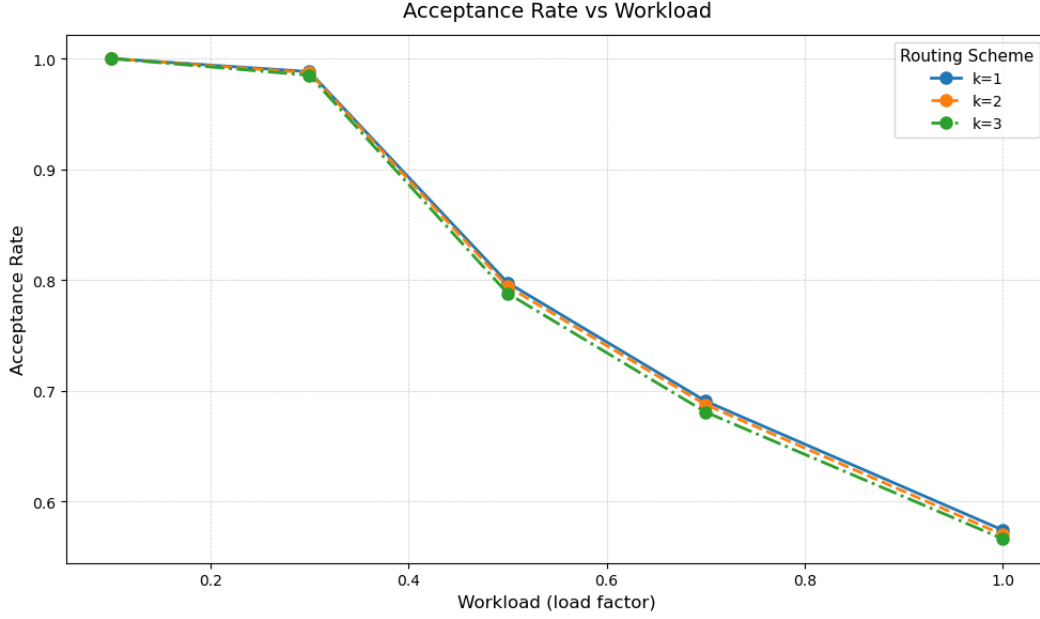


FIGURE 6.5. Flow Acceptance rates vs workload for different numbers of paths used. Results display very little variation across different numbers of paths.

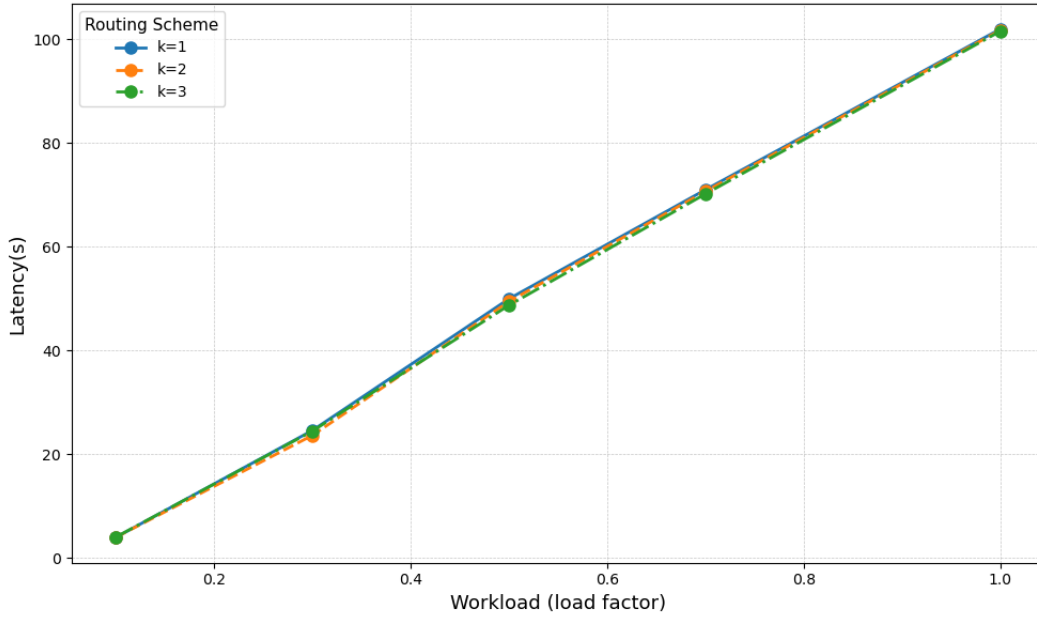


FIGURE 6.6. Flow Latency vs workload for different numbers of paths used. Results display very little variation across different numbers of paths.

6.4.2. Frame Replication and Elimination for Reliability). Figure 6.7 shows the impact of increasing network workload and varying proportions of flows requesting Frame Replication and Elimination for Reliability (FRER) on three key metrics: overall flow acceptance rate, latency

(top plot), and the success rate of multi-path allocation for requesting flows (bottom plot). As the workload (network load factor) increases, the acceptance rate declines due to congestion and limited resources. Although acceptance rates decline as the proportion of FRER flows increases, the degradation relative to single-path routing is not severe. Latency remains stable across all FRER levels, demonstrating that CQF can support multi-path routing with FRER without imposing substantial performance penalties.

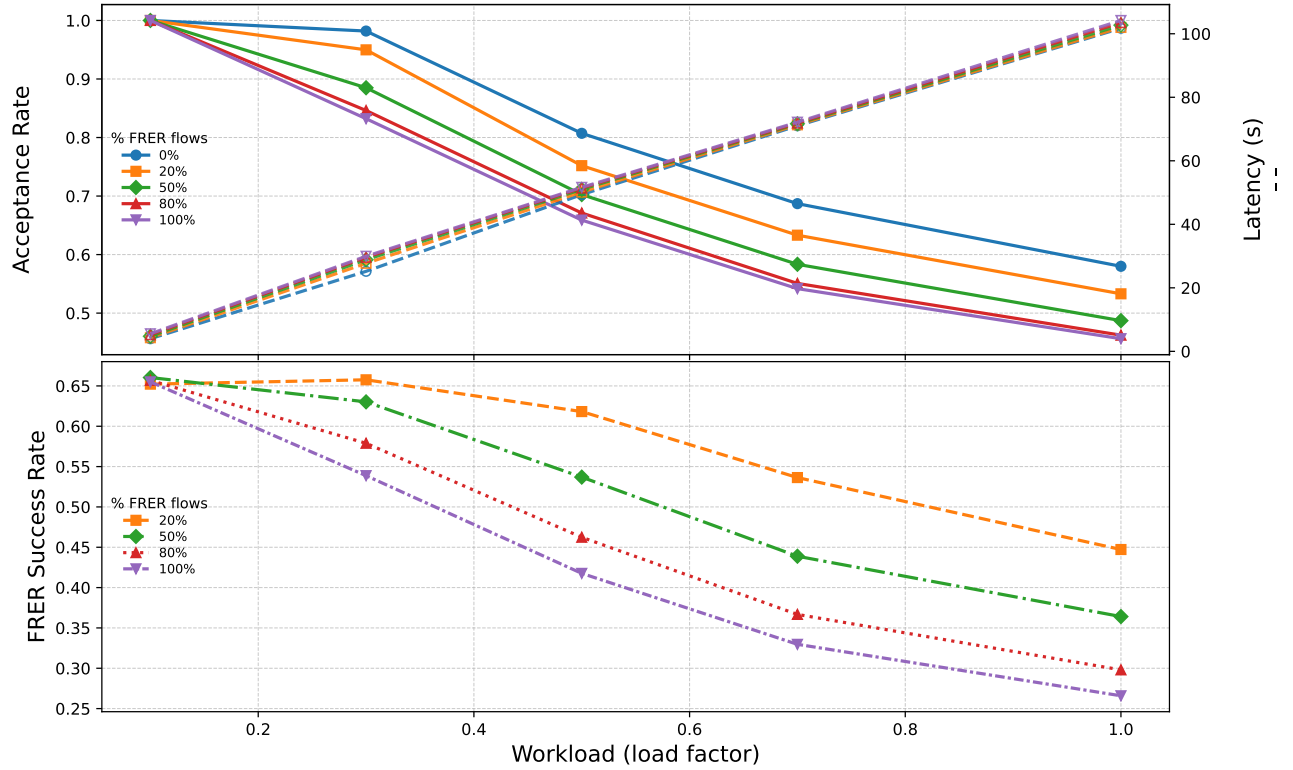


FIGURE 6.7. FRER Success rates and Overall Flow acceptance rates and Latency vs workload. Results display only a small degradation in acceptance rate and latency while enabling frame replicated flows.

By transmitting flows across multiple paths, the likelihood of total data loss due to a single-path failure is significantly reduced, thereby enhancing overall reliability even in the presence of unexpected failures. It also reduces the overhead associated with packet loss recovery. For large data transfers across long distances, retransmissions can be costly, leading to congestion and packet reordering.

Overall, the results demonstrate that multi-path routing can improve reliability while not causing a significant reduction in performance.

CHAPTER 7

Conclusion and Future Work

7.1. Conclusion

This thesis explored the applicability of Cyclic Queuing and Forwarding (CQF), a deterministic networking technology, in the demanding environment of wide-area scientific networks. We investigated its potential to provide low-latency communication for dynamic workloads composed of both time-triggered and event-triggered data flows. The core of this work demonstrated that by integrating CQF into a realistic simulation of the ESnet topology, it is possible to achieve performance comparable to existing reservation systems while offering finer-grained control.

Our simulation results confirm several key findings. First, carefully tuning the CQF cycle time is critical to balancing efficiency and performance. Second, leveraging multi-path routing alongside CQF can significantly improve network performance and reliability while maintaining the predictability required for sensitive scientific workflows. This work represents one of the first comprehensive efforts to apply CQF at the WAN scale, illustrating its potential to support the increasing demands of data-intensive, geographically distributed scientific applications. Ultimately, our findings suggest that CQF, though originally designed for LANs, is a strong candidate for addressing the future high-performance, deterministic networking needs in WANs like ESnet.

7.2. Future Work

While the simulation results are promising, several avenues for future research can build upon this work to further validate and enhance the proposed approach.

First, to gain more practical insights beyond a simulated environment, experiments could be conducted on real-world scientific networks such as the wide area Fabric testbed [23] to deploy and evaluate the performance of CQF with physical hardware and network links. This will allow us to observe its behavior under real-world conditions and uncover challenges not present in a simulation. Furthermore, a critical aspect of implementing CQF in a WAN is the requirement for precise time synchronization across geographically distant nodes. Therefore, we plan to investigate the required

precision of time synchronization for CQF to operate effectively and reliably. This investigation opens avenues for exploring emerging and highly precise technologies like quantum time synchronization [24], which could one day make ultra-precise, WAN-scale clock alignment a practical reality.

We would also like to research novel traffic engineering methods, including reinforcement learning methods, to improve the performance, especially in multipath routing.

Analysis of Critical Internet Infrastructure

This chapter presents a study on identifying critical components in public Wide Area Networks (WANs), specifically focusing on **Internet Exchange Points (IXPs)**. This work addresses the broader theme of network resilience and reliability, which is a foundational requirement for the stable operation of high-performance scientific networks. The material presented here was co-authored by the author and is based on the research titled “Quantifying IXP Indispensability.”

8.1. Introduction and Motivation

Networks across the globe have been affected by numerous synthetic (Censorship, Observation, and others) and natural (Hurricanes, Earthquakes, Floods, and others) outages. The shutdowns in Bangladesh are a natural example [25]. In such cases, having redundancy in network pathways between IPs and Autonomous System (AS)es can alleviate the impact of such outages. IXPs are prominent institutions where this redundancy can help maintain connectivity during outages, due to multiple ASes peering together. Not all IXPs in a country’s domestic network have the same peering profile, and it is imperative to pinpoint IXPs with the least amount of redundancy in their peering profile as potential critical points. To that end, we aim to quantify the impact of an IXP on a country’s domestic network, based on their AS profile.

Multiple avenues have been used to address this issue, ranging from measuring the reachability of the IXP to understanding the centrality of an IXP in a country’s network topology. One missing aspect is the impact of the removal of an IXP on the connection between consumer IPs and AS peers. We would call this the **indispensability** of the IXP. Indispensable IXPs are often in a unique position to provide internet services to an under-served portion of a country’s domestic network. The impact of their removal on a country’s network, especially for IXPs that have a high number of unique ASes with no alternative peering-based IP-AS connections in the country’s domestic network, is significant.

In this study, we aim to analyze a set of countries’ domestic connectivity-based [IXP](#) profiles and identify and verify the group of [IXPs](#) that are indispensable. To address the above questions, we build a country’s [IXP](#) profile, which contains information per [IXP](#) about its peers, members’ neighbors, IP coverage, and active traceroute data. We then use this profile to find the set of [IXPs](#) that are indispensable to a country, based on a scoring function, and verify it using empirical data. Our main contribution through this work is:

- We develop a scoring function to identify the indispensable set of [IXPs](#).
- We verify the correlation between the product of the number of unique [ASes](#) and the percentage of IPs related to an [IXP](#) and the [IXP](#)’s indispensability.

8.2. Related Work

Several studies map the internet from different aspects and identify critical points. Roberts et al. [26] introduced a method to identify critical [ASes](#) in the national network. Leyba et al. [27] introduced NCP, an [AS](#)-level metric to assess a country’s network for critical points. Prehn et al. [28] analyzed the reachability of prefixes from bilateral and multi-lateral peering locations. Gamero-Garrido et al. [29] developed CTI, to measure critical points for transit-dominant countries. Gregori et al. [30] added [IXPs](#) to [AS](#) level topology and analyzed connection strengths, Fan et al. added [IXP](#) information in [AS](#) level topology, by creating a two-layered network model [31], Lodhi et al. analyze PeeringDB [32], a public database that has information about participating network’s peering information and the traffic related to it, for measuring [IXP](#) specific properties. Trusin et al. [33] analyzed the effects of the Ukraine-Russia War from the perspective of critical [IXPs](#) and found the effects of connection loss at the [IXPs](#) for both sides.

The most similar to our work is by Tashiro et al. [34], where they find the dependency of networks on [IXPs](#) topologically using active measurement. Our work differs, as we attempt to identify the [IXPs](#) in the country that have the highest potential to majorly affect connections between the peering [ASes](#) and the connected IPs, in the case of a shutdown, given [AS](#) uniqueness and IP connectivity.

8.3. Data Sources

In the span of this study, we used these datasets:

- **CAIDA IXP Dataset [35]:** A collection of IXPs and their subnet prefix, geolocation, AS peers, and other metadata. We used the country code to filter the IXPs of each country and extracted the peering db IDs of all the IXPs.
- **Peering DB Dumps [32]:** Peering DB includes information related to ASes and their peering locations. We used the Peering DB IDs from the Center for Applied Internet Data Analysis (CAIDA) IXP dataset [35] to filter out the AS IPs and their bandwidth details.
- **Active Traceroutes:** We generated a longitudinal dataset through our traceroute campaign, over 3 months, collected at random times of the day, to the IP addresses of the selected ASes in each IXP.
- **RIPE Atlas probes archive [36]:** Provides information about their probes, their location, their ASes, and others via this dataset.
- **IP2Location [37]:** A commercial dataset that includes the geolocations of prefixes and provides a country-level distinction between these prefixes.

8.4. Methodology

In this study, we analyze the indispensability of an IXP. To do so, we first establish a varied set of candidate countries to analyze. Countries are chosen to get a variance across topologically and physically different countries. Once the countries are selected, to constrain the analysis, top ASes based on their assigned bandwidth are selected per IXP in a country. Our analysis includes active measurement of the IXP via traceroutes, to enable that, we select candidate probes per day of the campaign. The overall synopsis of the ranking methodology can be seen in Figure 8.1.

8.4.1. Getting Candidates for the study. For our study, processing data using all the countries in the world is not feasible, we need a representative set of countries to analyze. To enable this we identify candidate countries such that, per IXP, we select candidate ASes based on the bandwidth assigned to them in the IXP’s subnet. Furthermore, we also select candidate probes for our active measurement, per country. We expand on our selection process in the following sections.

8.4.1.1. Candidate Country Selection. We aim to identify candidate countries based on factors like the number of domestic IXPs, country size, number of IPs, and others. To enable this we use the following criteria:

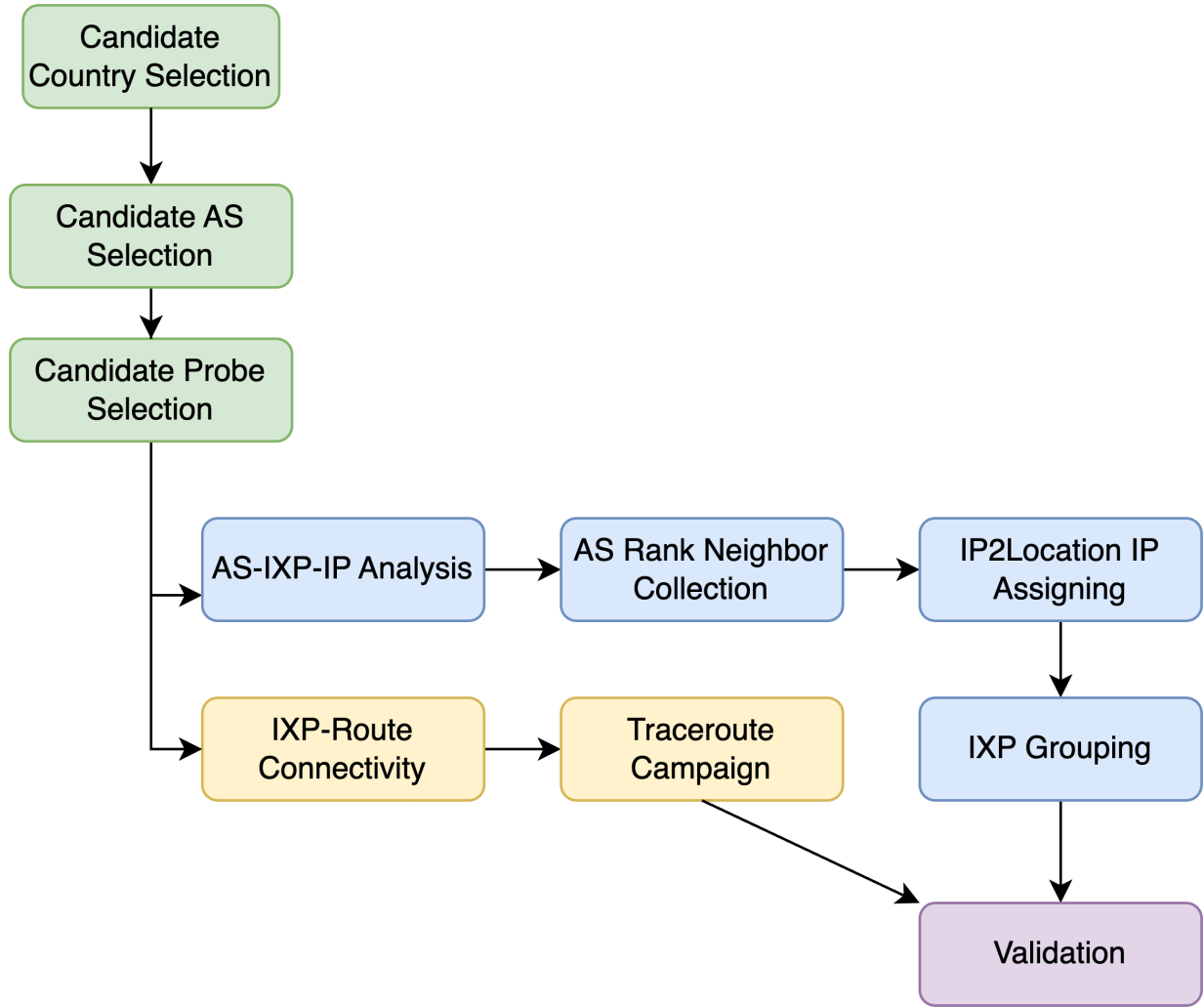


FIGURE 8.1. Pipeline to find the group of indispensable [IXPs](#)

- (1) Have at least 1 country from each continent
- (2) Have at least 1 Réseaux IP Européens Network ([RIPE](#)) Atlas probe to ping the [IXP](#) from within the country.
- (3) Variance in physical size.
- (4) Variance in number of [IXPs](#).
- (5) Variance in number of IPs.

Using the above criteria, we select a set of 10 countries from all over the world. The candidate countries include: *Argentina, Australia, Bangladesh, Brazil, Ireland, India, Italy, Saudi Arabia,*

Turkey, and South Africa. We explicitly avoid choosing countries from North America due to the abundance of studies done regarding countries that region. [38, 39, 40].

8.4.1.2. *Candidate Peer Selection per IXP.* In a candidate country, we choose all the available IXPs in the country that have active peers and not just router servers.

Each IXP has multiple ASes peering at it and not all ASes have the same assigned bandwidth within the IXP. PeeringDB [32] provides information about the bandwidth for the AS peering at an IXP. For each IXP, we select candidate ASes such that they are in the top 10 ASes in terms of assigned bandwidth. In doing so we aim to identify the important AS peers of the IXP while minimizing the number of ASes to be processed for our analysis.

8.4.1.3. *Candidate Probe Selection per country.* We select all available probes from RIPE Atlas [41] for that country per campaign day. A probe is available if its status is shown as "Connected". Probe data updates may be made before the active measurement starts, in which case the most recent probe information is used. Furthermore, we always check the probe activity before and during an active measurement.

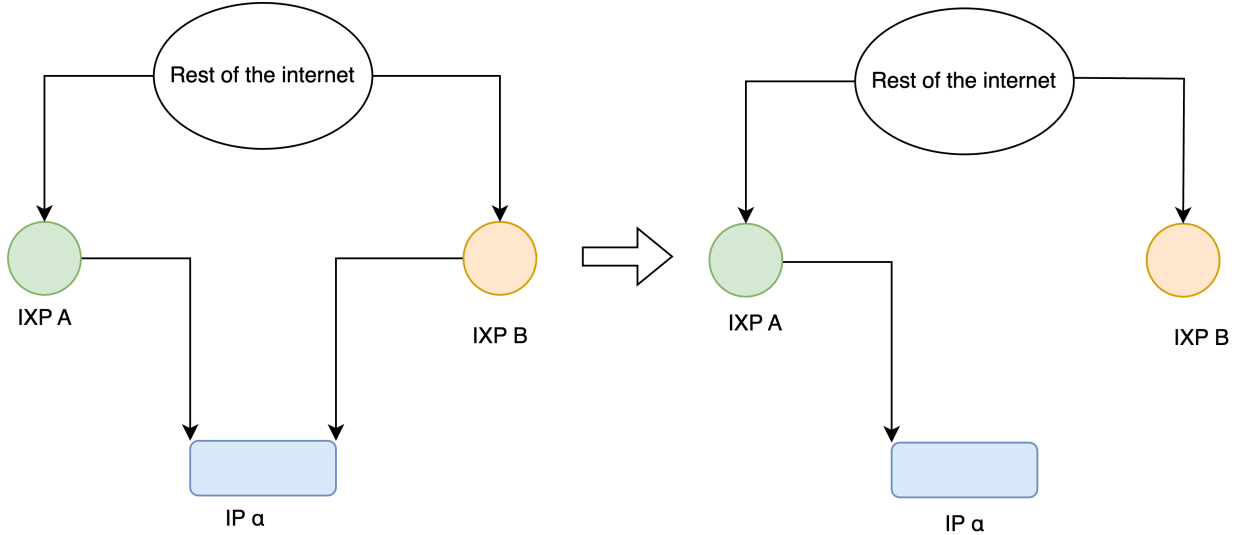


FIGURE 8.2. IXP Reachability Example

8.4.2. AS-IXP-IP Analysis. We start our analysis by using the candidate countries and the IXPs in them. In this step, we aim to identify the indispensability of the IXPs based on the reachability of their ASes. We define **reachability** as the ability of an AS to connect its customer IPs to the rest of the internet via an IXP. For example, as seen in Figure 8.2, a single AS is peering

at both IXP A and IXP B, and IP α is connected to the rest of the internet at IXP A or IXP B, via that AS. In the case that IXP B is disconnected from the network, the impact on IP α will be low, in regards to reaching the rest of the internet, as α has a way through IXP A. An AS with a high reachability reduces the impact of an IXP shutdown on the connected IPs. We quantify this, by designing a function to find the uniqueness score of an IXP. We elaborate on this scoring function in the following section.

We also observe the impact of IXP from another angle, by looking at the percentage of IPs associated with the IXP. Intuitively, if an IXP directly serves a low number of IPs then the impact on its removal is low. For example, an IXP with mostly unique ASes and fewer IPs connected with it is more dispensable than an IXP with some unique ASes and a large number of IPs connected with it.

8.4.2.1. *IXP Grouping.* We utilize the information from the subsections 8.4.2.4 & 8.4.2.3 to group the IXPs that can potentially be indispensable ones in a country. The grouping is done by selecting IXPs with the comparatively higher Normalized Weighted AS Score and Percentage of IPs. We don't define a ranking between the IXPs in the grouping.

To quantify and better visualize this grouping process, we define a scoring function,

$$(8.1) \quad \textbf{Scoring Function} = \textit{NormalizedWeightedASScore} \times \textit{PercentageofIPs}$$

Each IXP, per country, is scored and then all scores, per country, are normalized for comparison. We will elaborate on the components of this function in the following subsections.

8.4.2.2. *AS Rank Neighbor Collection.* To quantify IXP reachability, we first need to find the number of ways a candidate AS can be reached by an IP in the country. To do so, we utilize AS Rank [42] to find all the neighboring (Providers, Customers, and Peers) ASes to the candidate AS, per IXP, in a country. Since we iterate over the IXP's peering ASes, we skip any AS that peers at IXP, to avoid duplicate data. The result of this process is a set of candidate IXPs, in a country and a mapping to their neighboring ASes. In the following subsection, we define how we quantified this mapping to get a scoring function.

8.4.2.3. *Percentage of IPs.* The Normalized Weighted AS score provides information about the uniqueness of the AS profile of an IXP, but it doesn't show the impact this uniqueness has on the

country’s network. To understand that, we correlate the [AS](#) score from the section 8.4.2.4, with the percentage of IPs associated with the [IXP](#)’s [AS](#) profile.

Percentage of IPs = $\frac{P}{Q}$, where P is the number of IPs directly connected to the [IXP](#) and Q is the total number of IPs geolocated in the country

We first get all the IPs that are geolocated in a candidate country. We use a commercial IP geolocation database,

IP2Location [37] filtering the IP ranges associated with the candidate country. We then use [RIPE](#) Stat [36] to get the most specific prefix, to address the issue of multiple [ASes](#) announcing a prefix. We then use regional registries like AFRINIC, RIPE NCC, and others, along with WHOIS [43] to find the prefix ownership information. Using the [AS](#) neighbor mapping from the section 8.4.2.2 and prefix information from IP2Location, we find the percentage of total IPs in a candidate country associated with an [IXP](#).

8.4.2.4. *Normalized Weighted AS score.* To quantify the [AS](#) to [IXP](#) mapping, we come up with a scoring function that scores how unique an [IXP](#)’s [AS](#) profile is, based on reachability.

Weighted AS Score: $\sum_{i=1}^N \frac{1}{n_i}$, where n_i is the number of [IXPs](#), where any of the candidate [ASes](#) peer at, for the i th candidate peer [AS](#). N is the total number of candidate peers at the [IXP](#). Intuitively, we are trying to find the number of unique [ASes](#) an [IXP](#) has, given all the [IXPs](#) in a country. To make this score comparable, we normalize it over all the [IXPs](#) in a candidate country.

8.4.3. IXP-Route Analysis. To verify the passive data-based scores, we utilize active measurements. Our score is dependent on the percentage of IPs and number of [ASes](#), related to an [IXP](#). We use the number of successful traceroutes as our verification source, due to it having a positive correlation with both the number of IPs and the number of [ASes](#), which are the components of our scoring function (Number of unique [ASes](#) and number of [ASes](#) are positively correlated).

Additionally, BGP paths often don’t reflect [IXP](#) subnets, due to them not advertising the subnet IPs to the border routers. During our study, we attempted to find the [IXP](#) subnets in BGP dumps collected from Route Views monitors [44] located in Sao Paulo, Amsterdam, Oregon, and Singapore. We checked the occurrence of [IXP](#) subnets on these dumps for 2 [IXPs](#), one in Brazil and one in Argentina. We did not observe the [IXP](#) subnets in those dumps, hence we decided to center our verification process around traceroutes.

To verify the set of indispensable IXPs, obtained from the scoring function, we find the set of IXPs that have the number of daily % hits and compare the two sets, an IXP is considered verified if it is a member of both these sets.

- The campaign was held over 3 months, and IXPs were pinged at random intervals during this time. All available candidate probes were used for these traces.
- Traceroutes were sent from the selected probes, in section 8.4.1.3, to candidate ASes in the IXP. We avoided sending traceroutes to IPs that are connected to the IXP due to not being able to identify which IXP is showing up on the traceroute, using existing tools [45].
- We used information from the traceroute (destination reply status and Last Hop ASN) and used AS Rank to identify peers to check the status of a traceroute reaching the IXP, as seen in previous approaches [45].
- We found a day-by-day percentage of the number of traceroutes that were successful using the previous information.
- Additionally, we also used RIPE Stat [46] to do an IP to AS matching to understand the negative results for the traceroutes.

Using the information collected, we attempt to verify the results from the previous section, which can be seen in section 8.5.

Verification Process: We first select the set of IXPs that are considered indispensable according to the scoring function. We then select the IXPs that have a high number of daily peer hit % of successful traceroutes to the IXP (Verification Set). We then do an intersection of the sets and check for overlap. Any IXP that is in the indispensable set and verification set is considered verified, any IXP that is in the indispensable set but not in the verification set is considered a false positive, and any IXP in the verification set but not in the indispensable set is considered a false negative. The statistics on this can be seen in Table 8.4. We further elaborate on this in the following section.

8.5. Results

In this section, we summarize our findings from the study. We utilized the information gathered from existing databases like PeeringDB, ASRank, CAIDA, RIPE [32, 42, 47, 48] to identify the group of IXPs that are indispensable to each of the candidate countries. We found that the set of indispensable IXPs in a country is much smaller than the non-indispensable IXP set, for

the candidate set, we further elaborate on this finding in section 8.5.1. We then used active measurements in the form of traceroutes to verify the IXPs in this set using the process defined in 8.4.3. On verification, we observe that some IXPs are not in the verified set of indispensable IXPs. We discuss this while explaining our outliers in section 8.5.2.

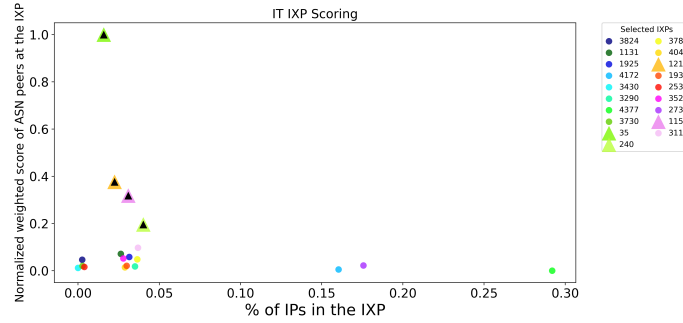


FIGURE 8.3. Observing the IXPs with a set of low-impact IXPs. Outliers have too low of a Normalized Weighted AS Score or % of IPs

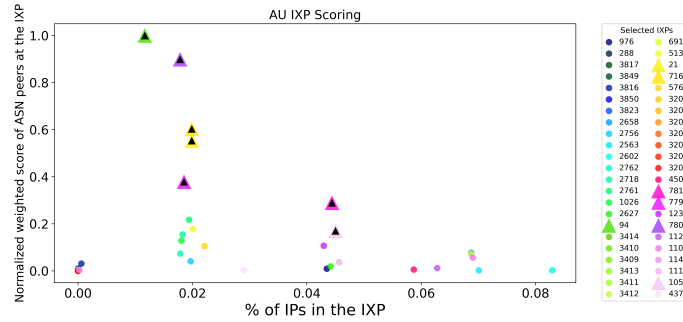


FIGURE 8.4. Observing the IXPs with a set of high-impact IXPs.

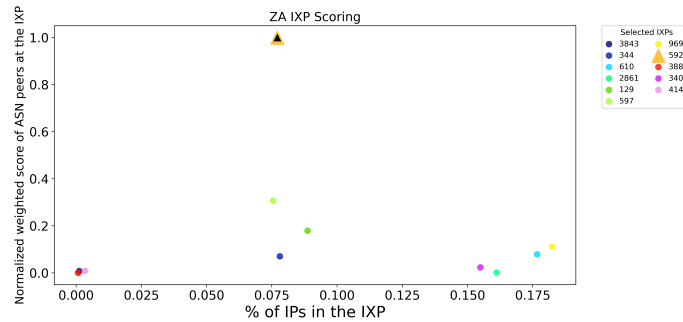


FIGURE 8.5. Observing sparse IXPs with a single high-impact IXP.

8.5.1. Identifying Indispensable IXPs. In this section, we visualize the indispensable set of IXPs. As mentioned in section 8.4.2.1, the scoring function is used to identify the set of indispensable IXPs and Figures 8.3, 8.4, 8.5 we use the components of the scoring function from section 8.4.2.4 as the axis, i.e. Normalized Weighted AS Score as the Y axis and Percentage of IPs as the X axis. The intersection of these points is used to compare the score of an IXP in the country. The top right represents the max value of the scoring function, i.e. 1.0 which is represented by (1,1) while the bottom left represents the lowest value the scoring function can have i.e. (0,0).

Indispensable IXPs, as defined in section 8.1, are the IXPs that have comparatively higher number of unique ASes (higher Normalized Weighted AS Score) and comparatively higher number of IPs associated with them (higher percentage of IPs). In the graphs, triangles represent the IXPs in the indispensable set.

In Figure 8.3, we observe that most IXPs are grouped with a few IXPs that are by themselves. These IXPs are in the indispensable set except IXPs that have either a comparatively low number of IPs associated with the IXP or the Normalized weighted AS score is low, i.e. more non-unique ASes peer at this IXP. This trend is observed in all the candidate countries.

Additionally, we also make observations about the country's IXPs, based on our scoring function visualization. A country with a majority of its IXPs with a low Normal Weighted AS Score or low percentage of IPs means that the impact of IXP shutdown on the country's network is low compared to the country where IXPs don't show this behavior. As we can see from Figure 8.3, Italy's IXPs have low Normalized Weighted AS Score and low percentage of IPs as compared to Australia (Figure 8.4) and South Africa (Figure 8.5).

8.5.2. Verification using Traceroutes. To verify the set of indispensable IXPs and check if our scoring function works, we use Average Daily Peer Hit % as described in section 8.4.3. We use Average Daily Peer Hit % as it is directly correlated with our metric, due to being dependent on similar components, number of ASes, and number of IPs. This can be observed in Tables 8.1, 8.2, 8.3 such that the IXPs above the partition in the table are IXPs that were confirmed on verification, the IXPs with a * are the IXPs in the initial indispensable set and the bold IXPs are the ones in the verification set (found using Average Daily Peer Hit %). Table 8.4 shows the correctness of our scoring function per country, excluding Bangladesh.

IXPs	Average Daily Peer Hit %	Scores
* 35	0.12885500	0.01584590
* 115	0.14122900	0.00984620
* 121	0.09664110	0.00847165
* 240	0.07736390	0.00787317
273	0.00908953	0.00387411
311	0.04148940	0.00358506
1131	0.04682790	0.00187600
1925	0.04051270	0.00183156
1933	0.00535018	0.00061980
2537	0.00295567	0.00006204
3290	0.01048060	0.00063409
3528	0	0.00144799
3730	0	0.00005513
3782	0.03846150	0.00175606
4046	0.02244900	0.00040879
4172	0	0.00082726
4377	0	0

TABLE 8.1. Verification Table For Italy. All the [IXPs](#) are verified to be indispensable

The average correctness score over 9 countries was 0.8299. The selection of False Positives and False Negatives is described in section 8.4.3. Our scoring function does not consider peering policy, but it affects our verification set, as restrictive or selective peering policies for the top bandwidth [ASes](#) can reduce the number of traceroute hits. We believe considering the peering policy during candidate [AS](#) selection in the future would help in increasing the correctness of the scoring function.

8.5.3. Bangladesh Outages. During our traceroute campaign for Bangladesh, we encountered outages in the region. Bangladesh experienced a country-wide internet shutdown between July 18th - July 28th, with social media being down till July 31st [25]. This shutdown affected the domestic Ripe Atlas probes [41], such there were a total of 4 probes available during the campaign, out of which 1 probe was not available in August, 1 probe was not available in October, 1 probe was only available in September and 1 probe was available throughout the campaign. We observed the country during its recovery period to analyze the behavior of indispensable [IXPs](#). We observed no traceroute hits till August 17th. We observe that three [IXPs](#) with the highest scores, BDIXMain, ISPABNIX, and SUMMITNIX (0.2651,0.0492,0.0241) received the highest number of traceroute hits, and follow the trend we observed in the previous section.

IXPs	Average Daily Peer Hit %	Scores
* 21	0.112003	0.01100880
* 94	0.0823308	0.01167470
* 716	0.105201	0.01199740
* 105	0.0560853	0.00766892
110	0.0319046	0.00386142
111	0.0318841	0.00168151
112	0	0.00073330
114	0	0.00000097
288	0	0.00000034
513	0.0776603	0.00356832
576	0.0376303	0.00232731
691	0.120116	0.00490527
* 779	0.0438699	0.00702338
* 780	0.0405229	0.0160428
* 781	0.0160428	0.0128512
976	0	0.00040207
1026	0.0377385	0.00231736
1235	0.0035338	0.00458436
2563	0.0389206	0.00015468
2602	0.0465443	0.00020589
2627	0.0554197	0.00085105
2658	0.0439302	0.00524436
2718	0.0950784	0.00283091
2756	0.0394886	0.00080179
2761	0.0232675	0.00421586
2762	0	0.00130629
4375	0	0.00007787

TABLE 8.2. Verification Table for Australia. There are 4 false positives and 3 false negatives. The main reason for this behavior is due to the peering policy differences in the candidate [ASes](#).

TABLE 8.3. Verification Table for South Africa. Example of when the verification set and indispensability set don't match. There are 2 false negatives and 1 false positive.

8.6. Limitations

Our study relies on passive measurement datasets that are updated periodically. This can lead to inconsistencies in our passive analysis. Our traceroute campaign had an element of variability based on candidate probes in a country since the number of probes could change between days. This could lead to inconsistencies in data collection.

Country	False Positives	False Negatives	Correctness
IT	0	0	1.0
AR	0	0	1.0
TR	0	0	1.0
BR	3	3	0.85
IN	2	4	0.823
SA	1	0	0.75
AU	4	3	0.747
ZA	1	2	0.7
IE	1	1	0.6

TABLE 8.4. Correctness Scores: $\frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$

Date	BDIXMain	BTCLIX	ISPABNIX	SUMMITNIX
Aug-17-2024	0	0	8	0
Sep-02-2024	14	4	11	8
Sep-21-2024	7	3	5	12
Sep-22-2024	4	3	15	8
Sep-23-2024	8	2	15	8
Sep-24-2024	9	4	8	13
Sep-26-2024	7	1	4	7
Sep-30-2024	10	4	10	3
Oct-02-2024	8	1	14	8
Oct-03-2024	7	4	8	6
Grand Total	74	26	98	73

TABLE 8.5. Bangladesh Traceroute Hits

We define our scoring function based on peering connections between an [IXP](#) and IP, which might not be enough in some cases where backbone [AS](#) connections are prominent.

8.7. Future Work

In this study, we devised an approach to find the indispensable [IXPs](#) in a country. Based on our results, we found that most candidate countries had a small set of indispensable [IXPs](#). Having a small set of indispensable [IXPs](#) can have negative implications on the country’s connectivity during outages due to reduced redundancy in [AS](#) peering.

In the future, we aim to scale our study by adding more candidate countries and collecting data for a longer period. We also plan on incorporating the peering policy during our candidate [AS](#) selection to further improve our scoring correctness.

Bibliography

- [1] E. Dart, J. Zurawski, C. Hawk, B. Brown, and I. Monga, “ESnet requirements review program through the IRI lens: A meta-analysis of workflow patterns across DOE Office of science programs,” 2023.
- [2] “Esnet homepage,” <https://www.es.net/>, 2024, accessed: 2025-05-03. [Online]. Available: <https://www.es.net/>
- [3] C. Guok and D. Robertson, “ESnet on-demand secure circuits and advance reservation system (OSCARS),” presented at Internet2 Joint, 2006, additional information available at <https://www.es.net/engineering-services/oscars/>. [Online]. Available: <https://www.es.net/assets/ESnet-Research/OSCARS/oscars-gridnets-20061001.pdf>
- [4] Time-sensitive networking (TSN) task group. <https://1.ieee802.org/tsn/>. Last accessed: 2024-7-29.
- [5] “P802.1Qch – cyclic queuing and forwarding,” <https://1.ieee802.org/tsn/802-1qch/>, accessed: 2024-9-27.
- [6] “Network facts and stats,” <https://www.es.net/engineering-services/the-network/network-facts-and-stats/>, 2025, accessed: 2025-05-03. [Online]. Available: <https://www.es.net/engineering-services/the-network/network-facts-and-stats/>
- [7] Y. Seol, D. Hyeon, J. Min, M. Kim, and J. Paek, “Timely survey of time-sensitive networking: Past and future directions,” *IEEE Access*, vol. 9, pp. 142 506–142 527, 2021.
- [8] S. Tschöke, F. Lynker, H. Buhr, F. Schreiner, A. Willner, A. Vick, and M. Chemnitz, “Time-sensitive networking over metropolitan area networks for remote industrial control,” in *2021 IEEE/ACM 25th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. IEEE, 2021, pp. 1–4.
- [9] G. Peng, S. Wang, Y. Huang, R. Huo, T. Huang, and Y. Liu, “Traffic shaping at the edge: Enabling bounded latency for large-scale deterministic networks,” in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–6.
- [10] B. Varga, Ed, J. Farkas, A. Malis, and S. Bryant, “Deterministic Networking (DetNet) Data Plane: IP over IEEE 802.1 Time-Sensitive Networking (TSN),” Jun 2021.
- [11] S. Hou, W. Yang, Y. Cheng, and L. Feng, “Reliable routing of time-triggered traffic in time sensitive networks,” in *2023 5th International Conference on Natural Language Processing (ICNLP)*. IEEE, 2023, pp. 469–474.
- [12] K. Huang, X. Wan, K. Wang, X. Jiang, J. Chen, Q. Deng, W. Xu, Y. Peng, and Z. Liu, “Reliability-aware multipath routing of time-triggered traffic in time-sensitive networks,” *Electronics*, vol. 10, no. 2, p. 125, 2021.
- [13] Z. Chen, Y. Lu, H. Wang, J. Qin, M. Wang, and W. Pan, “Dynamic stream partitioning for time-triggered traffic in time-sensitive networking,” *Computer Networks*, vol. 248, p. 110492, 2024.

- [14] G. Liu, R. Lin, S. Luo, and X. Wang, “Rescheduling for dynamic traffic in time-sensitive networks,” in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2024, pp. 239–244.
- [15] X. Jin, C. Xia, N. Guan, C. Xu, D. Li, Y. Yin, and P. Zeng, “Real-time scheduling of massive data in time sensitive networks with a limited number of schedule entries,” *IEEE access*, vol. 8, pp. 6751–6767, 2020.
- [16] A. Banerjee, W.-c. Feng, D. Ghosal, and B. Mukherjee, “Algorithms for integrated routing and scheduling for aggregating data from distributed resources on a lambda grid,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 1, pp. 24–34, 2008.
- [17] X. Huang, J. Chen, K. Guo, and D. Gao, “Cqcf: A cyclic queuing cluster forwarding based transmission mechanism for large-scale deterministic network,” in *International Conference on Emerging Networking Architecture and Technologies*. Springer, 2022, pp. 557–569.
- [18] M. Wang, S. Zhu, L. Wang, W. Li, and Y. Zhang, “Heuristic fast routing in large-scale deterministic network,” in *2023 IEEE International Performance, Computing, and Communications Conference (IPCCC)*. IEEE, 2023, pp. 9–17.
- [19] M. Wang, S. Zhu, and Y. Zhang, “A heuristic online algorithm for routing in large-scale deterministic networks,” in *2023 24th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, 2023, pp. 189–194.
- [20] NetworkX — NetworkX documentation. <https://networkx.org/>. Accessed: 2024-8-17.
- [21] Overview — SimPy 4.1.1 documentation. <https://simpy.readthedocs.io/en/latest/>. Accessed: 2024-8-17.
- [22] NetworkX Developers. (2024) Betweenness centrality - networkx 3.4.2 documentation. Accessed: 2025-05-03. [Online]. Available: <https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms centrality.betweenness centrality.html>
- [23] Fabric - adaptive programmable research infrastructure for computer science and science applications. <https://portal.fabric-testbed.net/about/about-fabric/>. Accessed: 2025-8-28.
- [24] E. Yu, A. Butko, D. Ghosal, M. Kiran, and N. Rao, “A simulation study of quantum clock synchronization using teleportation,” in *2024 International Conference on Quantum Communications, Networking, and Computing (QCNC)*. IEEE, 2024, pp. 265–270.
- [25] S. Abdullah, “Shutdown watch: Building a community against internet shutdowns in bangladesh,” *Internet Society Pulse*, 2024. [Online]. Available: <https://pulse.internetsociety.org/blog/shutdown-watch-building-a-community-against-internet-shutdowns-in-bangladesh>
- [26] H. Roberts, D. Larochelle, R. Faris, and J. Palfrey, “Mapping local internet control,” in *Computer Communications Workshop (Hyannis, CA, 2011)*, IEEE, 2011.
- [27] K. G. Leyba, B. Edwards, C. Freeman, J. R. Crandall, and S. Forrest, “Borders and gateways: measuring and analyzing national as chokepoints,” in *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*, 2019, pp. 184–194.

- [28] L. Prehn, F. Lichtblau, C. Dietzel, and A. Feldmann, “Peering only? analyzing the reachability benefits of joining large ixps today,” in *International Conference on Passive and Active Network Measurement*. Springer, 2022, pp. 338–366.
- [29] A. Gamero-Garrido, E. Carisimo, S. Hao, B. Huffaker, A. C. Snoeren, and A. Dainotti, “Quantifying nations’ exposure to traffic observation and selective tampering,” in *International Conference on Passive and Active Network Measurement*. Springer, 2022, pp. 645–674.
- [30] E. Gregori, A. Improta, L. Lenzini, and C. Orsini, “The impact of ixps on the as-level topology structure of the internet,” *Computer Communications*, vol. 34, no. 1, pp. 68–82, 2011.
- [31] Z. Fan and W. K. Tang, “Unraveling the impacts of ixp in internet ecosystem using bi-layered network,” *Physica A: Statistical Mechanics and its Applications*, vol. 456, pp. 327–339, 2016.
- [32] PeeringDB, “Peeringdb: The interconnection database,” <https://www.peeringdb.com/>, 2024, accessed: 2024-05-21.
- [33] C. Trusin, L. Bertholdo, and J. J. Santanna, “The effect of the russian-ukrainian conflict from the perspective of internet exchanges,” 2022. [Online]. Available: <https://arxiv.org/abs/2211.06123>
- [34] M. Tashiro, R. Fontugne, and K. Fukuda, “Following the data trail: An analysis of ixp dependencies,” in *International Conference on Passive and Active Network Measurement*. Springer, 2024, pp. 199–227.
- [35] “Caida ucsd ixps dataset,” <https://www.caida.org/catalog/datasets/ixps/>, 2024, accessed: 2024-02-21.
- [36] 2024, rIPE Routing Information Service (RIS), <https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris/>. Accessed: 2024-04-15.
- [37] “Ip2location,” 2024, accessed: 2024-03-12. [Online]. Available: <https://lite.ip2location.com/database-download>
- [38] C. A. Rose, “A measurement procedure for queueing network models of computer systems,” *ACM Computing Surveys (CSUR)*, vol. 10, no. 3, pp. 263–280, 1978.
- [39] U. Goel, M. P. Wittie, K. C. Claffy, and A. Le, “Survey of end-to-end mobile network measurement testbeds, tools, and services,” *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 105–123, 2015.
- [40] A. Yassine, H. Rahimi, and S. Shirmohammadi, “Software defined network traffic measurement: Current trends and challenges,” *IEEE Instrumentation & Measurement Magazine*, vol. 18, no. 2, pp. 42–50, 2015.
- [41] R. Atlas, “Ripe atlas probe dataset,” <https://ftp.ripe.net/ripe/atlas/probes/archive/>, 2024, accessed: 2024-05-21.
- [42] A. Rank, “A ranking of the largest autonomous systems (as) in the internet,” 2021.
- [43] L. Daigle, “Whois protocol specification,” RFC, Tech. Rep., 2004.
- [44] Routeviews, “RouteViews Archive Project,” <https://archive.routeviews.org/>, 2023, accessed: 2024-03-11.
- [45] G. Nomikos and X. Dimitropoulos, “traixroute: Detecting ixps in traceroute paths,” in *Passive and Active Measurement: 17th International Conference, PAM 2016, Heraklion, Greece, March 31-April 1, 2016. Proceedings 17*. Springer, 2016, pp. 346–358.
- [46] “Ripe stat application,” <https://stat.ripe.net/>, 2024, accessed: 2024-03-11.

- [47] “Ipv4 based ranking,” 2024, accessed: 2024-03-11. [Online]. Available: <https://www-public.imtbs-tsp.eu/~maigron/rir-stats/ripe-allocations/ipv4/ripence-ipv4-by-number.html>
- [48] “Caida ucsd routeviews prefix to as mappings dataset (pfx2as) for ipv4 and ipv6 dataset, <https://www.caida.org/catalog/datasets/routeviews-prefix2as/>,” 2024, accessed: 2024-04-20.