Trond Vatten

Enhancing the Resilience of Modern Mobile Networks

June 2023

NTNU
Norwegian University of
Science and Technology

Master's thesis

2023

Trond Vatten

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Information Security and Communication
Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Enhancing the Resilience of Modern Mobile Networks

Trond Vatten

Communication Technology and Digital Security

Submission date:	June 2023
Supervisor:	Poul Einar Heegaard
Co-supervisor:	Yuming Jiang

Norwegian University of Science and Technology
Department of Information Security and Communication Technology

Title: Enhancing the Resilience of Modern Mobile Networks
Student: Vatten, Trond

Problem description:

As the world becomes increasingly connected and reliant on mobile networks, it is crucial that these networks support both critical and non-critical services. Despite advancements in technology, current mobile networks are not robust enough to fully support these services and ensure they function at a minimum level at all times, even in the face of larger undesired events like natural disasters and network failures. Ensuring service guarantees is a vital aspect of future mobile networks, but a significant challenge that must be addressed. One technology that is key to achieving this goal in 5G networks is network slicing, which allows operators to tailor each logical network to the specific needs of a particular service, without compromising the performance of other services.

While network slicing in 5G networks has great potential, there are still significant challenges to be addressed. One of them is determining the optimal way to allocate resources throughout the network, also called the resource allocation problem. The high level of virtualization and software-based architecture of modern mobile networks makes the optimization of network resources particularly difficult. To address this, a promising algorithm called ClusPR has been developed. ClusPR uses a heuristic approach, and provides a near-optimal solution to the resource allocation problem.

In this project, I will investigate the survivability performance of a network that implements ClusPR. Even though ClusPR has already shown a near-optimal solution to the problem, it is yet to be tested in adverse circumstances where large parts of the network fails. To address this gap, I will develop a disaster model that simulates larger network failures, and use it to assess the survivability of the network.

The results of this research will provide insights into how network slices should be defined to ensure robust and resilient networks for various critical and non-critical services. Additionally, the research will identify possible shortcomings and improvements that need to be addressed in ClusPR and similar algorithms.

To accomplish these goals, I will start by modeling a realistic ISP topology of nodes and links and deploy the ClusPR algorithm on it. Thereafter, I will create the disaster model and apply it to the network to simulate the adverse conditions. I will then make performance assessments of the disrupted network, deploy ClusPR again, and assess its performance on the degraded topology. By evaluating the performance of ClusPR in these scenarios, this project will contribute to the advancement of 5G technologies and improve our understanding of how to build robust and resilient networks for critical services.

Approved on:	2023-02-19
Main supervisor:	Heegaard, Poul Einar, NTNU
Co-supervisor:	Jiang, Yuming, NTNU

Abstract

As the world becomes increasingly dependent on the internet, more and more services demand internet connectivity - from self-driving cars and remote surgical operations to the integration of augmented and virtual reality in our daily lives. These services require extreme performance, such as a maximum latency of 1 millisecond and uptime of more than 99.9999%. Meeting these requirements necessitates a revolution in network design, wherein Network Function Virtualization (NFV) plays an integral role. Modern mobile networks have transitioned from reliance on dedicated, inflexible hardware to extensive softwarization and virtualization. This thesis addresses the essential need for these networks to maintain extreme resilience, even in the face of network failures or targeted attacks. We employ a comprehensive framework that quantifies and evaluates network survivability and use the findings to devise strategies that NFV networks can implement to enhance their resilience. These strategies center around guiding NFV network operators in prioritizing node repair after network failures and leveraging softwarization to automatically re-optimize the virtual layer of the network post-failure. Our findings underscore the significant influence of network structure on resilience and introduce a novel metric called *flow centrality* that outperforms traditional metrics such as betweenness and closeness centrality in identifying critical nodes in an NFV network. Furthermore, we demonstrate the benefits of adopting a re-optimization strategy post-network failures and discuss the challenges of a one-size-fits-all recovery strategy for NFV networks, thereby highlighting the complex and context-dependent nature of network survivability. This thesis contributes to enhancing the resilience of NFV networks, investigates recovery strategies, proposes a novel metric for node importance, and suggests directions for future research, ultimately contributing to a more robust and secure digital infrastructure.

Sammendrag

Ettersom verden blir stadig mer avhengig av internett, krever flere og flere tjenester internettforbindelse - alt fra selvkjørende biler og fjernkirurgiske operasjoner til integrasjon av utvidet og virtuell virkelighet i vårt daglige liv. Disse tjenestene krever ekstreme ytelse, som for eksempel maksimal latenstid på 1 millisekund og oppetid på mer enn 99,9999%. For å møte disse kravene kreves det en revolusjon innen nettverksdesign, der virtualisering av nettverksfunksjoner NFV spiller en sentral rolle. Moderne mobilnettverk har gått fra å være avhengige av dedikert og ufleksibel maskinvare til omfattende softwarisering og virtualisering. Denne avhandlingen adresserer det essensielle behovet for at disse nettverkene opprettholder ekstrem resiliens, selv i møte med nettverksfeil eller målrettede angrep. Vi bruker et omfattende rammeverk som kvantifiserer og evaluerer overlevelsevnene til et nettverk, og bruker funnene til å utarbeide strategier som NFV-nettverk kan implementere for å styrke sin resiliens. Disse strategiene fokuserer på å veilede NFV-nettverksoperatører i å prioritere rekkefølge av nodereparasjoner etter nettverksfeil, og å utnytte softwarisering for å automatisk re-optimere det virtuelle laget av nettverket etter feil. Våre funn understreker den betydelige påvirkningen av nettverksstruktur for resiliens, og introduserer en ny metrikk kalt *flow centrality*, som overgår tradisjonelle metrikker som mellomplasseringssentralitet (betweenness centrality) og nærhetssentralitet (closeness centrality) i å identifisere kritiske noder i et NFV-nettverk. Videre demonstrerer vi fordelene med å ta i bruk en re-optimaliseringsstrategi etter nettverksfeil, og diskuterer utfordringene med en "one-size-fits-allstrategi for reparering av NFV-nettverk, og fremhever dermed den komplekse, kontekstavhengige karakteristikken av overlevelsevnene til et nettverk. Denne avhandlingen bidrar til å styrke NFV-nettverks resiliens, undersøker ulike repareringsstrategier, foreslår en ny metrikk for nodesentralitet, og foreslår retninger for fremtidig forskning som til slutt vil bidra til en mer robust og sikker digital infrastruktur.

Preface

This thesis represents a significant milestone for me: the beginning of my academic journey! To mark this occasion, I wish to express my gratitude to some important people.

Firstly, my gratitude goes to my supervisor and co-supervisor, Poul and Yuming. Our weekly meetings have been a perfect arena for me to discuss and get challenged on my ideas. You have pushed and motivated me while keeping me in check when I get too ambitious; as Poul reminded me: "We haven’t saved the world, yet".

I must also acknowledge my family. To my mother, Marit, and my father, Torgeir, thank you for your tireless support, no matter what challenge I begin. To my brother, Lars, thank you for your compassion and friendship, you are my closest. Finally, a heartfelt thanks to my girlfriend, Ingrid. Your daily support and enduring understanding will always be a source of inspiration to me.

Contents

List of Figures

List of Tables

List of Acronyms

3GPP 3rd Generation Partnership Project.

CI Critical Infrastructure.

eMBB Enhanced Mobile Broadband.

ETSI European Telecommunications Standards Institute.

mMTC Massive Machine Type Communication.

NFV Network Function Virtualization.

NFV-RA Network Function Virtualization Resource Allocation.

NORCICS Norwegian Centre for Cybersecurity in Critical Sectors.

NTNU Norwegian University of Science and Technology.

SDN Software Defined Networking.

SFC Service Function Chain.

SFI Centre for Research-based Innovation.

URLLC Ultra Reliable Low Latency Communication.

VNF Virtualized Network Function.

Chapter 1 Introduction

1.1 Background and Motivation

The evolution of modern mobile networks has facilitated the creation of a wide range of innovative services and use cases. These use cases, set to be enabled by a fully operational 5G network, range from daily-life integration of augmented reality to remote surgical procedures and autonomous transport systems. Each emerging service necessitates unique performance parameters, such as low delay or latency, as well as dependability measures such as reliability and availability. However, the full realization of these services has yet to be realized due to the partial implementation of critical technologies intrinsic to the 5G infrastructure, such as network slicing.

Network slicing is perceived as a fundamental enabler in satisfying the stringent and widely diverse requirements of services within the 5G domain [3GP18]. Given that different use cases necessitate such a heterogeneous and varied set of performance criteria, often in conflict with each other, providing all services on a single network is not viable. Network slicing addresses this challenge by partitioning a physical network into smaller, logical entities known as network slices. From an operator or client perspective, each network slice is perceived as a fully functional standalone network, complete with all the management capabilities typically offered by a standalone network. This mechanism allows operators to tailor their services precisely to the requirements of a particular service without compromising the performance of other services since each operates within its independent network slice.

Despite network slicing coming to light as far back as 2016, network operators and research communities have yet to establish a universally accepted technical definition. Research and development efforts continue in this arena, with numerous methodologies still under proposal. Central to all this research is the transformative paradigm of virtualization and softwarization of network infrastructure, offered by NFV.

NFV is an emerging technology in network architecture that transforms rigid, physical network infrastructure into flexible, software-defined entities. This shift towards software-focused solutions aims to reduce the cost of network operations, enhance service delivery, and make networks more flexible and easy to scale. However, the complexity of NFV networks introduced by their dynamic and virtualized characteristics presents unique challenges in maintaining network resilience and survivability.

A critical aspect of NFV networks is their dependence on VNFs hosted on network nodes. The failure of these nodes can have a profound impact on network performance, making them a potential target for attackers. While network failures are inevitable, swift and effective recovery is vital in maintaining a network's operational continuity. This is particularly significant in scenarios involving random node failures or targeted node (cyber) attacks, necessitating effective strategies to enhance network resilience.

However, designing these strategies and determining their efficacy is not straightforward. They rely heavily on multiple factors such as network structure, types of network failures, and how to prioritize node repair order, among others. The effectiveness of a strategy might vary significantly depending on the combination of these factors, creating a complex landscape that is challenging to navigate.

1.2 Research Objectives

The overall aim of this thesis is to investigate the impact of (random and targeted) node failures and their recovery on network service degradation in an NFV network context. The specific objectives are:

1. Assess the performance of NFV networks during and after an undesired event, aiming to develop recovery strategies and heuristics to enhance network resilience.
2. Assess the impact of these strategies on network survivability.
3. Identify key factors that influence the resilience of NFV networks, informing more resilient network design.

1.3 Thesis Structure

Following this introduction, the thesis is structured as follows:

Chapter 2 presents various use cases that serve as the motivation for this study, as well as thoroughly investigating the technical solutions that exist, and are missing, to meet the requirements of these use cases.

Chapter 3 describes the methodology of the research, detailing the experimental setup, performance measures used, and the developed recovery strategies.

Chapter 4 presents and discusses the results of the experiments.

Chapter 5 discusses the implications of the findings, drawing key conclusions from the research, exploring practical implications, and suggesting potential directions for future work.

Finally, Chapter 6 concludes the thesis, summarizing the main findings, the contributions made to the field, and the potential impact of this research.

1.4 Note on Related Publication

This master thesis is part of a PhD project conducted under the umbrella of the Norwegian Centre for Cybersecurity in Critical Sectors (NORCICS), a Centre for Research-based Innovation (SFI) dedicated to promoting cybersecurity and resilient digital solutions.

Simultaneously with this thesis, a paper incorporating similar methods and results was prepared and submitted to the 13th International Workshop on Resilient Networks Design and Modeling (RNDM 2023) [RNDM23]. The initial version of this paper, as submitted for the first deadline of RNDM 2023, is included in Appendix A.

While both the thesis and the paper share objectives and findings, they differ in their level of detail and format. This thesis provides a more comprehensive and in-depth exploration of the research topic.

Chapter 2 Background and Literature Review

Core parts of this chapter were conducted during the Norwegian University of Science and Technology (NTNU) pre-project course TTM4502 and supplemented by a literature survey in the IMT4203 course at NTNU in January 2023. Therefore, substantial elements of this chapter are also included in the literature survey, which is available in its entirety in Appendix B.

2.1 5G for Critical Infrastructure

One of the domains expected to witness substantial innovation and novel services owing to the advancements in modern mobile networks is CI. Previous generations of mobile networks, such as 3G and 4G, have undergone rapid developments in terms of data rates and capacity. However, these performance metrics alone are insufficient for CI. The potential consequences of network failures in CI are too significant, underscoring the necessity of including resilience as a performance requirement in communication systems.

Rising demands from CI necessitate advanced, specialized communication systems. Historically, dedicated physical infrastructure has been used to meet these requirements, but this strategy often results in high costs, slow deployment, and complex management as the network expands. To tackle these challenges, the 3rd Generation Partnership Project (3GPP) formulated the 5G standard, segmenting it into three primary markets: Enhanced Mobile Broadband (eMBB), Massive Machine Type Communication (mMTC), and Ultra Reliable Low Latency Communication (URLLC) [3GP18]. The eMBB segment delivers higher data rates and enhances broadband access over a broader geographical range. In contrast, the mMTC segment facilitates large-scale communication between devices, essential for the Internet of Things (IoT). The URLLC segment addresses services requiring high reliability and fast data transmission. These segments each answer to unique needs, enhancing the flexibility of 5G in supporting a wide array of services and applications.

Evolving Cyber Threats in an Increasingly Interconnected World

Today's world is highly interconnected, with few infrastructures remaining entirely isolated from the internet [MT19]. This ongoing trend will likely expand the potential attack surfaces on CI [All15]. As CI increasingly incorporate data systems, their management must recognize the inherent cyber-physical system attributes to enhance resilience against cyber threats [BMK+21]. Addressing the security of CI calls for international cooperation, as a single country's CI breach could have global repercussions. Within this scenario, 5G, and specifically URLLC, play a pivotal role in strengthening CI and facilitating cross-border security requirements alignment [3GP18].

The RESISTO project [ercpfciO18], funded by the EU, exemplifies efforts to increase the resilience of telecommunications for CI in light of expanding attack surfaces. RESISTO aims to develop comprehensive and innovative solutions for mitigating emerging security challenges stemming from the rapid deployment and adoption of 5G networks.

Cascading Effects: Interdependence of Critical Infrastructures

CIs are interdependent and rely on each other for proper operation. These interconnections extend beyond physical to logical and cyber dependencies, underscoring the significance of system resilience. A disruption in one area can trigger cascading effects such as error propagation or failure escalation, highlighting the importance for 5G networks to maintain high resilience levels to prevent such incidents [BMF+19].

Service Isolation

Service isolation is crucial in modern communication networks, as a compromise in one service could impact others sharing the same channels [3GP16]. This issue becomes particularly severe when services have varied security and performance requirements. In the context of CI, a disruption in one service could result in a ripple effect on others. For example, the disruption of surgical operations due to nearby excessive network traffic. Thus, service isolation is paramount for smooth CI operations, ensuring attacks and failures are confined to a single service.

Scaling

CI confront the challenge of scaling to meet fluctuating demand. Traditional strategies often involve each vertical deploying separate infrastructure, a process which is both costly and time-consuming [BBKW19]. Communication networks in CI must maintain high levels of availability and reliability, even in the face of changing demand. This flexibility is one of the main reasons behind the slower-than-anticipated deployment of future automation solutions, such as self-driving cars.

With CI moving towards more interconnected and interdependent systems, the necessity for service isolation and scalable solutions arises. This is where 5G comes into play, a technology designed with the versatility to meet these emerging demands. In the following section, we will further explore the capabilities and potential of 5G in addressing the complex needs and challenges of critical infrastructures.

2.2 5G

As discussed earlier, 3GPP aimed to address some of the key problems in traditional networks by defining use cases that will be enabled once 5G is fully implemented. One of these problems lies in the fact that the use cases have very diverse performance requirements, often conflicting with each other. While some services demand performance features such as low delay and high throughput, others need to compromise on these to ensure high dependability. In traditional networks, where these use cases share the same "one-size-fits-all" communication channels, satisfying such varied, stringent, and heterogeneous requirements is not feasible. However, 5G introduces a key concept that addresses this problem, called network slicing.

2.3 Network Slicing

The implementation of network slicing has been identified as a key enabling technology in the pursuit of meeting the stringent and varied requirements of 5G use cases. Network slicing allows for the creation of virtual networks, referred to as "network slices", on top of existing physical infrastructure. Each network slice can be tailored to meet the specific requirements of a tenant, facilitating serving both low latency slices and high dependability slices on the same physical infrastructure. This virtualization of network infrastructure enables multiple tenants to optimize their connectivity needs without incurring the cost and maintenance of owning their physical infrastructure.

In the context of CI, the introduction of network slicing allows for the specification of highly specific network requirements without compromising the demands of other services. Before the adoption of network slicing, it was often necessary to make trade-offs in the performance of one sector to meet the security requirements of another [FPEM17]. With slicing, each sector can customize a dedicated network slice to meet its own unique needs. Additionally, the logical isolation and separation provided by network slicing enable CI to maintain strict security requirements even when sharing physical infrastructure with less secure networks.

Although network slicing was introduced in the full set of 5G standards by 3GPP in 2018 [3GP18] as a means of dividing a physical network into smaller, logically isolated virtual networks, research communities, and network operators have yet to arrive at a consistent technical definition for network slices. Various methods have been proposed, and research in this area remains ongoing. One common understanding of network slices originates from the one introduced in 3GPP's 5G release. Here, three predefined slices were detailed with specific characteristics, each serving the requirements of the three use cases: eMBB, mMTC, and URLLC. Although there is some overlap in the requirements for these slices, they have distinct performance characteristics. The URLLC slice has much higher requirements for latency and reliability such as sub $1ms$ end-to-end latency and 99.999% reliability [3GP21], while the mMTC slice must support a large number of smaller devices transmitting control information. All three slices are important for the proper functioning of the 5G network, but URLLC and mMTC is particularly critical for the operation of CI. In most cases, slices operating CI will need higher grades of isolation, whether it is incorporated with an URLLC slice or an mMTC slice. This is due to the stringent requirements for service isolation and availability that CI must adhere to, owing to the vital societal functions they deliver.

As an example of the potential benefits of network slicing, Kurtz *et al.* [KBDW18] discusses the use of communication in driving automation. Using traditional networks, the communication required for tasks such as tracking vehicle location, speed, and direction would not be fast or reliable enough, potentially leading to catastrophic accidents. However, the work also shows that if the specifications [3GP18] for URLLC slices are met, the data could be used to improve real-time road analysis, optimize traffic flow, or even increase vehicle automation. All of these improvements are possible while also including slices for passenger entertainment (eMBB) and smart metering (mMTC), as long as proper slice prioritization is followed to ensure hard service guarantees for the most critical slices. Such technological advancements could help reduce the number of fatal accidents on the roads.

Network slicing also has the potential to address the issue of service isolation in communication networks, as it allows for the creation of completely isolated slices within a single network. This can prevent the interference of one slice from affecting the operation of others, as described in Gonzalez *et al.* [GOH+20]. While there are challenges to be addressed in the implementation of this concept, the ability to achieve successful slice isolation is a major step to meeting the stringent security requirements of CI.

One of the key advantages of 5G network slicing is its ability to quickly scale and dynamically adapt to changes in the network, as noted in Foukas *et al.* [FPEM17], Li *et al.* [LSC+17], and Zhang [Zha19]. By providing a shared physical infrastructure with a virtual layer on top that can be dynamically updated in an automated manner, network slicing can replace the need for multiple dedicated networks. This programmable and adaptable nature is crucial for ensuring the continuous operation of CI, even in the event of compromise or failure. The flexible nature of softwarization and virtualization is a key enabler for network slicing and is referred to as one of the biggest paradigms in networking, called SDN and NFV.

Several remaining technical challenges to fully implement network slicing will be addressed within the domains of SDN and NFV, as these are critical elements to enable network slicing. Therefore, in this project, we will transition to exploring the existing research gaps in this area. However, the underlying motivation of the project remains the same; Increase knowledge and insight into using 5G and beyond technologies to build secure, resilient, and survivable critical infrastructures to provide critical services [NORCICS23].

2.4 Virtualized Networking

For network slicing to be effectively realized, NFV and SDN are regarded as key enabling technologies [OAL+17; SBT+17; YBSS17]. NFV involves the virtualization of traditional network functions such as firewalls and routers, which were previously implemented on specialized hardware. By adopting NFV, a network function is instantiated as software running on generalized hardware, known as a VNF. These VNFs are flexible and can be operated independently from any location. In 2012, the European Telecommunications Standards Institute (ETSI) released a standard for NFV, which has been further developed through hundreds of publications [ETS12]. VNFs serve as fundamental building blocks for constructing network slices by combining them in a Service Function Chain (SFC). To equip each slice with the necessary network functions, virtualized network function chaining is crucial. Such a chain is programmable and can be dynamically modified and allocated throughout the network using SDN.

A simplified example of the connection between CI, 5G slicing, VNF, SDN is depicted in Figure 2.1. Nodes in the figure represent virtualized 5G network components, which are mapped to physical nodes in the infrastructure (omitted from the figure for readability). In the core network, these nodes consist of generalized core routers where VNFs are deployed. In the Radio Access Network of 5G, these components consist of logical units such as the Central Unit and the Distributed Unit, as well as radio units, connecting equipment to the system [BPD+20]. VNFs are deployed on the virtual nodes and can easily be migrated to other nodes based on changing network demands. All VNF deployment and migration of VNFs, together with other network management, happens in the centralized SDN controller in the separated control plane. Lastly, we see a network slice consisting of an SFC of various VNFs in a specific order. Communication in this slice is dedicated to a certain URLLC service, ensuring service guarantees can be upheld even though other network traffic is present in the same network.

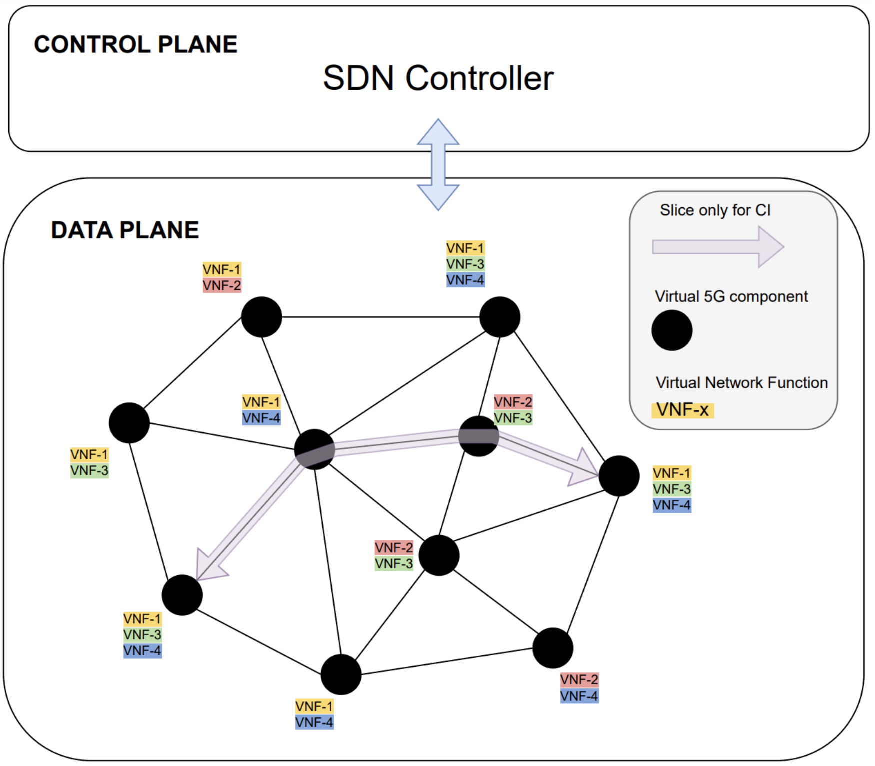


Figure 2.1: Overview of the connection between CI, 5G slicing, VNF and SDN. The example network slice is dedicated to a CI use case, meeting its stringent requirements.

SDN is a novel approach to networking that segregates the data plane and the control plane [BAMH20]. The data plane is accountable for physically forwarding traffic, while the control plane determines how traffic is routed. In SDN, the control plane centralizes routing decisions and maintains a comprehensive view of the entire network, simplifying the identification and response to changing demands and loads. Combining SDN with NFV facilitates traffic routing through a network in specific ways to satisfy diverse requirements such as bandwidth, end-to-end latency, security, and other factors.

The benefits of NFV and SDN for CI are considerable. The programmable nature of 5G networks empowered by NFV and SDN provide greater flexibility and adaptability in addressing changing demands and risks. This can help guarantee the continuous operation of CI, even in the face of disruptions to other infrastructures.

Moreover, NFV and SDN can enable more efficient use of resources and cost savings as they support the dynamic allocation of network resources and the capability to scale up or down as required. This is particularly important for CI, where the cost of downtime can be substantial.

In Kurtz *et al.* [KBDW18], a novel solution was proposed that builds upon ETSI's NFV standard and offers network slicing functionality. The solution was evaluated on a small testbed and validated through physical testing using real-world data from a CI communication scenario. The study demonstrated scalability and provided evidence that network guarantees can be maintained even under conditions of partial network overload. Hence, it is crucial to continue developing and enhancing such solutions and testing them with real-world data from CI scenarios before deploying them in actual networks.

Further, the ability to program and customize network slices for specific needs can enhance the security and reliability of CI by facilitating the implementation of more personalized security measures. For instance, in disaster recovery, traditional methods for establishing emergency networks can be slow. By utilizing SDN and NFV, network traffic can be swiftly rerouted, or a dedicated slice can be allocated for communication. Research presented in Gajić, Furdek, and Heegaard [GFH20] has explored the use of network modeling after an undesired event to potentially cover blind spots and identify the need for redundancy in networks.

In summary, the incorporation of NFV and SDN into 5G networks are instrumental in the successful implementation of network slicing, yielding numerous benefits for both CI and non-critical infrastructure. Nonetheless, several research questions remain unaddressed, particularly concerning technical solutions to fulfill the stringent requirements of CI services. Among these issues is the complex problem of efficiently allocating resources in this newly virtualized network layer.

2.5 Network Function Virtualization Resource Allocation

With the advent of NFV, a new level of flexibility has been introduced, allowing more adaptive responses to rapidly changing network environments. The transition from reliance on dedicated hardware to a more flexible virtual layer allows for efficient repositioning of network functions. However, while this adaptability offers greater scalability and efficiency, it also introduces increased complexity. The Network Function Virtualization Resource Allocation (NFV-RA) problem, which concerns the optimal deployment of VNFs to form SFCs, has been recognized as one of the major challenges in NFV [WMRJ18].

The flexibility facilitated by network virtualization and softwarization introduces a multitude of potential optimization objectives when designing a network. For example, deploying VNFs along the shortest path between a source and a destination optimizes the delay for a given flow or slice. Applying this strategy for all source-destination pairs or slices results in minimum delay across the network. However, this approach requires deploying a large number of VNFs, leading to high redundancy and low network utilization, which can hinder efficiency and scalability. By contrast, sharing VNFs between slices could enhance network utilization, but would also increase delay for some slices. Additionally, the design must consider objectives such as slice isolation, link and node throughput, and execution time of network functions. Thus, due to its complexity, the NFV-RA problem has been proven to be NP-hard. Numerous solutions have been proposed, each of which optimizes different objectives using heuristic or meta-heuristic algorithms [GB16].

2.5.1 Examples of NFV-RA

Wang *et al.* [WLW+16] introduced a solution named JoraNFV, which offers a heuristic-based approach to balance network costs with high-quality service performance. Simulations demonstrated near-optimal results, and the system proved to be efficient and practical.

A comprehensive survey by Gil Herrera and Botero [GB16] revealed that none of the existing NFV-RA frameworks adequately addressed network failure scenarios by incorporating resilience into their optimization objectives. A more recent survey [YLT+20] noted some attempts to fill this gap. Several studies proposed methods for ensuring backup nodes or links in the event of failures [BBS16; FYG+15; DYL17], while some incorporated resilience constraints such as availability in designing algorithms to route and place VNFs as SFCs [BBS16; QASK17; QKA18]. However, while these studies propose strategies based on standard metrics such as node and link availability, they do not adequately conduct comprehensive resilience or survivability analyses on NFV networks to develop strategies specific to NFV networks' characteristics.

2.5.2 ClusPR

In our experiments, introduced in Chapter 3, we will use ClusPR when applying an NFV-RA framework Woldeyohannes *et al.* [WMRJ18]. ClusPR balances several objectives: minimizing path stretch, evenly distributing load among NF instances, and maximizing total network utilization. The framework has demonstrated its capability of achieving near-optimal results for large-sized networks in reasonable time frames.

More specifically, the main problem the framework solves is the balance between minimizing delay in the network and maximizing network utilization. By placing VNFs in the shortest paths for all flows in the network, all traffic would experience as low a delay as possible. However, this also yields low network utilization, leaving minimal capacity for scaling the network with more services. Instead, ClusPR groups the flows with similar shortest paths and shares VNF instances between these flows. As a result, flows will get close to their shortest paths and receive their required VNFs while deploying close to as few VNFs in the network as possible.

2.6 Survivability Quantification in Networks

Survivability and resilience are key concepts in systems that need to perform. Survivability refers to a system's ability to fulfill its mission without interruption, even in the presence of threats or failures [EFL+99; KS00; Wes04]. It often implies the system's capacity to withstand degradation or failures, maintain essential operational capabilities, and recover functionality as fast as possible. On the other hand, resilience is a broader concept that encompasses survivability. Resilience refers to the ability of a system to anticipate, adapt to, and rapidly recover from a potentially disruptive event. This includes handling unexpected disruptions, adapting to new demands, and even learning from previous disruptions to improve future performance. While survivability focuses more on withstanding and continuing through undesired events, resilience underscores the importance of adaptability and growth in the face of adversity.

Survivability was first quantified in computer networks in Liu and Trivedi [LT06], and later in Heegaard and Trivedi [HT09], which provided a framework to quantify network survivability. This framework demonstrated broad applicability for various network sizes and disaster scenarios, including different parameters in assessing network performance. Xie, Heegaard, and Jiang [XHJ13] extended this framework to model a propagating failure scenario, highlighting the cascading effects disasters could have on a network over time. Gajić, Furdek, and Heegaard [GFH20] further considered both spatial and temporal evaluations of network disaster recovery in a content delivery network example. While these studies provide frameworks for quantifying the survivability of different networks in different scenarios, they do not use NFV-specific measures in their survivability quantifications or conduct survivability assessments in NFV networks.

2.7 Network Failures and Attacks

To conduct survivability assessments, network failures or attacks need to be simulated. In virtualized networks, the susceptibility to failures and attacks increases, as new vulnerabilities and attack angles are introduced by adding a virtual layer to the infrastructure. Network failures can occur due to numerous factors such as hardware faults, software bugs, natural disasters, human errors, and cyber-attacks. Failures can be categorized into two main types: link failures and node failures. Link failures refer to the loss of connectivity between two nodes due to damage to the physical or logical link, while node failures refer to a node in the network becoming unavailable or nonfunctional due to a variety of reasons, including physical damage, software malfunction, or cyber-attack.

Cyber-attacks represent a growing threat to networked systems, particularly for CI. Such attacks can originate externally and internally, intending to compromise the integrity, availability, or confidentiality of data and services. Given the nature of virtualized networks and the increasing dependence on software, these systems are vulnerable to a variety of attacks.

2.7.1 Node Centrality

The concept of node centrality plays a crucial role in understanding the vulnerability of a network to failures and attacks. Node centrality refers to a measure of the relative importance of a node within the network. Highly central nodes often represent critical points for the flow of traffic within the network, and their failure or compromise can lead to significant disruptions to the network's performance.

Betweenness Centrality

Betweenness centrality is a measure of the number of shortest paths that pass through a particular node. Nodes with high betweenness centrality often act as key connectors within the network, and their failure can significantly disrupt the network's connectivity and performance. These nodes can also be attractive targets for attackers aiming to cause maximum disruption to the network.

Closeness Centrality

Closeness centrality measures the average shortest path between a node and all other nodes in the network. Nodes with high closeness centrality are typically those that can reach others in the network with the least amount of hops. These nodes, due to their accessibility, can be significant for maintaining network performance, especially for keeping network delay down, and may be considered as strategic points for network protection and failure recovery measures, but also for attackers to degrade network performance.

Understanding node criticality in a virtualized network has implications for resilient network design. In traditional networks, betweenness and closeness centrality are crucial to identify the most critical nodes. However, introducing a virtualized layer may change the criteria for determining node importance for network survivability.

In the next section, we discuss the research gap in existing literature concerning network survivability in virtualized networks.

2.8 Research Gap

The current body of research has not sufficiently addressed the integration of survivability quantification into the design and operation of NFV networks. There have been attempts to incorporate resilience in the NFV-RA problem, but a comprehensive survivability analysis incorporating NFV-specific measures is missing. Existing solutions primarily focus on traditional network measures such as betweenness and closeness centrality, thereby failing to fully capture the unique features and potential vulnerabilities of NFV networks. This gap in research necessitates further investigations into how survivability quantification can be incorporated into the NFV-RA problem, to design more resilient and robust networks that can proactively handle failures while ensuring optimal performance.

Chapter 3 Methodology

This chapter outlines the framework of our approach and describes the steps taken in our experimental study.

This chapter outlines the methodology employed to fulfill the objectives of the study. We detail the experimental framework, where we quantify survivability to evaluate the performance of an NFV network during and after undesirable events. Subsequently, we describe the developed recovery strategies and how their efficiency is evaluated.

3.1 Experimental Setup

Our study investigates survivability and recovery strategies in an NFV network under various failure scenarios. Figure 3.1 illustrates our experimental setup, consisting of five primary steps:

1. Initialization of the network topology
2. Generation of network traffic flows
3. Execution of an NFV-RA algorithm to optimize the deployment of service components
4. Simulation of node failures
5. Assessment of network survivability

Two loops in the setup are introduced to account for the randomness inherent in generating network traffic flows and network failures in certain scenarios. We implement all components of this experiment in Python, with the setup designed for modularity, facilitating the easy transition between different topologies, methods of flow creation, NFV-RA algorithms, failure scenarios, or modifications of the survivability assessment method.

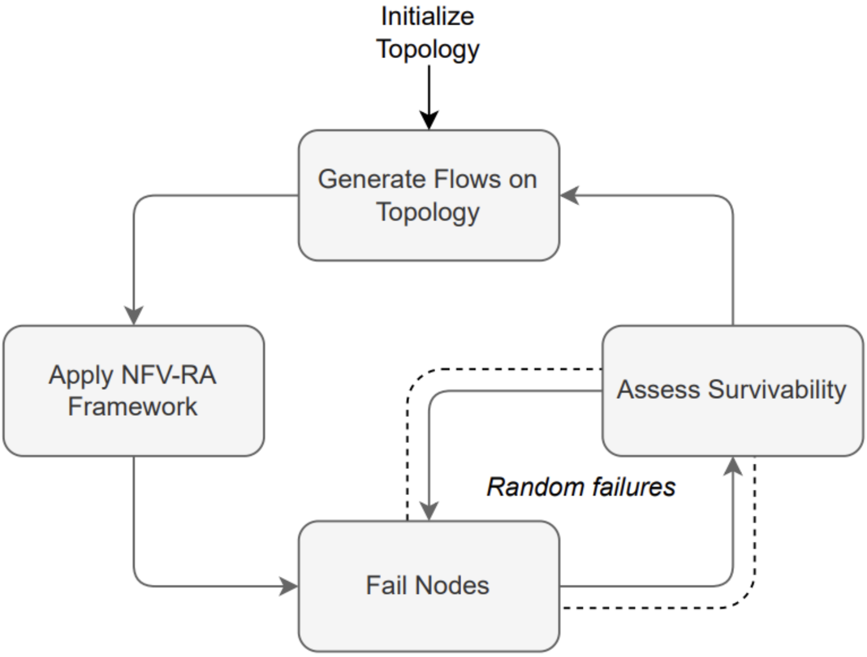


Figure 3.1: Overview of the experimental setup

3.2 Topology and Flow Generation

We created a Python program that extracts topologies from the Rocketfuel ISP topology mapping engine to generate the topologies [Rocketfuel02]. Rocketfuel employs real-world IP data to construct realistic ISP topologies stored in open-source datasets. Then we used the NetworkX package to represent this data in Python, enabling the creation and analysis of network structures with customized characteristics [NetworkX]. We visualized a model of these realistic networks, consisting of access nodes and core nodes, interconnected through edge nodes, all featuring realistically defined node capacities and link delays. To ensure the validity of our results, we conducted all experiments on two distinct topologies.

We generated network traffic by randomly creating a set of flows, each originating from an access node, passing through edge and core nodes, and terminating at another access node. Each flow required a designated set of network functions in a specific order, forming an SFC. The network functions are deployed at edge and core nodes. Flows and SFC, alongside their delay requirements simulating real-world latency constraints, constitute a service or a network slice.

Figure 3.2 shows an example of the flow represented by the following values in the network:

source: 26

destination: 16

required_nfs: 2, 1

shortest_path: 29

delay_requirement: 60.9

flow_path: 111, 112

actual_delay: 35

All irrelevant node labels and network functions are omitted in the figure for readability. The flow originates at node 26, terminates at node 16, and requires network function "2" and network function "1" to form its SFC. The theoretical shortest path between the source and destination is 29, and the maximum tolerated delay is 60.9. The routing phase of ClusPR has set node 111 and node 112 as nodes this flow must pass to receive its network functions. Doing this, the actual delay of the flow will be 35. Section 4 provides the specific values and characteristics of the various topologies, flows, and requirements we use in our experiments.

3.3 Application of NFV-RA

In an NFV network, a wide range of optimization strategies exists for the placement of VNFs forming the SFCs for flows. Our study uses the ClusPR framework, described in detail in Chapter 2, which optimizes the NFV-RA problem based on two metrics: delay and network utilization [WMRJ18]. These factors are highly relevant for 5G use cases, as delay is a crucial client-side QoS metric, and network utilization is vital for cost reduction and scalability on the operator side. As the ClusPR framework only implicitly handles service dependability, it is ideal for identifying the potential benefits of incorporating recovery strategies to improve system dependability.

3.4 Monte Carlo Simulation of Network Failures

Our study models random and targeted node failures, to capture and compare the diverse effects such failures can have on a network. Random failures can affect multiple nodes due to events such as natural disasters or software bugs when network components are upgraded. Targeted attacks aim at nodes with specific characteristics,

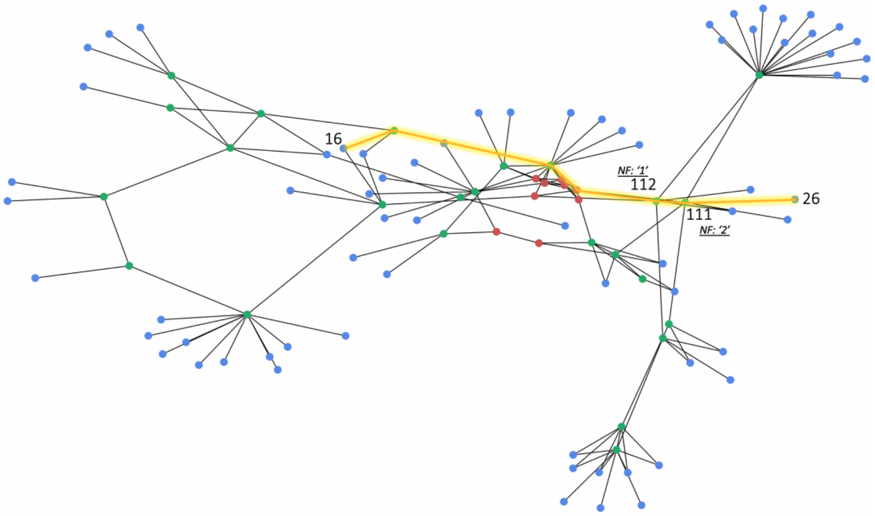


Figure 3.2: Example of generated flow in a network. Originating at access node 26, passing through edge nodes 111 and 112 which host the flows required network functions (NF 1 and NF 2), and terminating at access node 16

which could stem from cyberattacks based on an attacker's knowledge of the network structure.

3.4.1 Random Failures

The impact of random failures heavily depends on the nodes that fail. Some nodes do not significantly affect the network as they do not host any VNFs, while others can severely disrupt network performance if they fail (for instance, by separating large portions of the network). We capture the expected effects of such failures in our simulation experiments by repeating and averaging the random failure stage multiple times and assessing the network's survivability for each iteration, as depicted in the smaller loop in Figure 3.1.

3.4.2 Targeted Failures

Targeted failures are based on metrics identifying node importance. Thus, for a given set of flows on a specific topology, the same nodes will be targeted in each attack, requiring only one iteration per set of flows. We assume the attacker possesses knowledge of the topology structure and can infer node centrality measures from this knowledge. Our study assesses two types of attacks: one target nodes with high *betweenness centrality*, and another targets nodes with high *closeness centrality*. These measures are well-established metrics for identifying important nodes within a network and also play a crucial role in the ClusPR framework.

3.5 Assessment and Quantification of Survivability

Our primary goal is to accurately assess the performance of the NFV network during and after network failures. This allows us to evaluate how well NFV networks perform under various conditions and to devise strategies for improving recovery times.

We quantify network survivability over time following a failure, ignoring the rate of failure occurrence and instead focusing on the network's performance during the recovery phase post-failure. The performance of the NFV network is measured based on the number of admitted flows in the network, defined as those flows that receive their SFC, implying that they obtain all required VNFs in the correct sequence and thus can deliver their intended services. To be considered admitted, a flow must also meet its delay requirement; if not it is classified as failed. Immediately after a failure, the percentage of admitted flows is at its lowest, but as nodes are repaired, the number of admitted flows returns to its initial value.

Our assessment considers *n+1* states of the network, where n is the number of concurrent node failures, in line with the Markov model illustrated in Figure 3.3. Each state represents the number of failed nodes, and at *t = 0*, the system transitions from zero to n failed nodes. We assume nodes are repaired independently following an exponential distribution with rate (μ), until the network reaches a fully operational state in the absorbing state, 0.

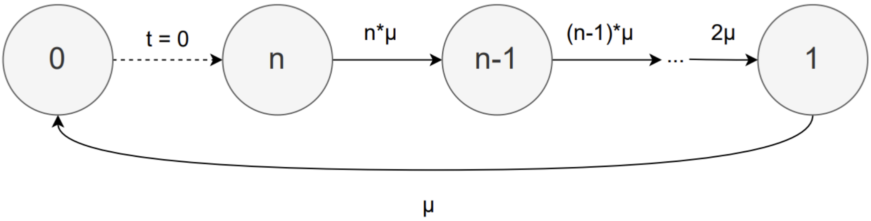


Figure 3.3: State transition diagram for network recovery following n node failures

To calculate network survivability, we first compute state probabilities as a function of time, given the repair intensity. This is computed with the cumulative distribution function of an exponential distribution, according to the function in Equation (3.1).

$$F(t, \mu) = 1 - e^{-\mu t} \tag{3.1}$$

With this, we can find the probability of being in each state at any given time according to Equation (3.2).

$$p_i(t) = \binom{n}{i} \cdot [1 - F(t, \mu)]^i \cdot F(t, \mu)^{n-i} \quad \text{for } i = 0, 1, \ldots, n \tag{3.2}$$

In our experiments, presented in Chapter 4, we use $n = 3$ node failures, giving the state probabilities in Equations (3.3), (3.4), (3.5), (3.6).

$$p_0(t) = F(t, \mu)^3 \tag{3.3}$$

$$p_1(t) = 3 \cdot [1 - F(t, \mu)] \cdot F(t, \mu)^2 \tag{3.4}$$

$$p_2(t) = 3 \cdot [1 - F(t, \mu)]^2 \cdot F(t, \mu) \tag{3.5}$$

$$p_3(t) = [1 - F(t, \mu)]^3 \tag{3.6}$$

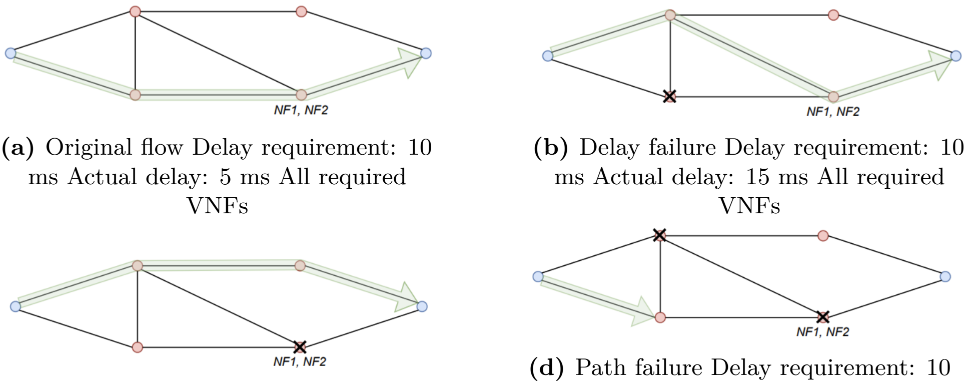
Then, we determine the network's performance (i.e., the number of admitted flows) for each state over time, given a specific set of flows and failed nodes. We then multiply the state probabilities at each time point with the network performance at the corresponding time point. The sum of these products over all time points provides network survivability. This process repeats for each set of flows and failed nodes.

Because the system's survivability depends on the repair order of failed nodes, we assess survivability for all possible permutations of node repair orders. From this, we extract three key statistics: the expected performance (average of all permutations of all iterations), the best performance (node repair order leading to the quickest recovery), and the worst performance (node repair order resulting in the slowest recovery). We note that for three node failures, which we will consider in this study, the number of permutations is six. As this number rapidly increases with the number of failures, a more efficient approach than brute force might be necessary for larger numbers of failures.

Complementing the survivability assessment, we also examine the flow failure causes. There are three potential causes for flows failing to meet their requirements, illustrated in Figure 3.4.

1. A flow may not meet its delay requirement along its path (Figure 3.4b);

2. Network fragmentation may block a feasible path between the source and destination node of a flow (Figure 3.4d);
3. In a case unique to NFV networks, a node hosting a VNF required by a flow may fail, preventing the flow from obtaining its necessary SFC (Figure 3.4c).



(a) Original flow Delay requirement: 10 ms Actual delay: 5 ms All required VNFs

(b) Delay failure Delay requirement: 10 ms Actual delay: 15 ms All required VNFs

(c) VNF failure Delay requirement: 10 ms Actual delay: 5 ms No required VNFs

(d) Path failure Delay requirement: 10 ms Actual delay: inf ms No required VNFs

Figure 3.4: All three possible causes of a flow not meeting its requirements

3.6 Validation of Results

To validate our findings, we employ two primary strategies: first, we conduct all experiments on two different topologies to ensure that our results are not specific to one scenario. Second, we simulate the experiment to compare the outcomes with the analytical approach. This simulation provides an additional level of validation for our study.

3.7 Recovery Strategies

In this section, we evaluate the network's survivability under four distinct recovery strategies for all topologies and failure types. We assume that failed nodes are repaired in either a random or a predetermined order, and we take advantage of a significant attribute of NFV networks: the ability to migrate VNFs from one node to another. These factors form the basis of the recovery strategies we outline in the following subsections.

3.7.1 Baseline Strategy

Initially, we do not adopt any specific node repair order or implement any VNF migration. This approach implies a random node repair order, and it represents the most computationally intensive method for survivability assessments, given the evaluation of all permutations of the node repair order. This method serves as a reference for the comparison with other strategies, as the average performance across all permutations depicts the expected performance of the network. We also extract the best and worst-performing node repair orders to establish benchmarks for subsequent strategies. This approach demonstrates the potential variability in the performance of different node repair orders and offers best and worst-case scenarios.

3.7.2 Node Repair Order Selection

The second strategy suggests heuristics to select a specific order of repair following failures based on three different rules. In contrast to the baseline strategy, this strategy considers only one node repair order, thereby reducing variability in survivability assessments. We compare this node repair order to the best, worst, and expected performances of the no strategy method to determine if some node repair orders consistently improve the network's recovery time across all network configurations and attack types.

The first heuristic involves repairing nodes with the highest betweenness centrality, potentially focusing on nodes that connect large parts of the network. The second rule considers closeness centrality, potentially prioritizing nodes that are essential for minimizing delays for flows. Both metrics are conventional measures used to characterize node importance in a network.

Lastly, we introduce a novel metric for node importance specific to NFV networks: **flow centrality**. Flow centrality is similar to betweenness centrality, which counts the shortest paths that pass through each node. Flow centrality counts the number of flows dependent on a VNF that a node hosts. Nodes are ranked accordingly, with nodes hosting a VNF that most flows depend on as most important. Therefore, the third rule repairs nodes based on their flow centrality.

3.7.3 Re-optimizing VNF Deployment

The third developed strategy involves re-optimizing VNF deployment in response to network changes. Applying ClusPR to the initial topology results in an optimized network based on the specific objectives of ClusPR. However, this optimization is conditioned on the given flows and the network structure of the topology. Any change to the network structure will change the conditions, and the applied optimization is no longer guaranteed optimal. In this strategy, we re-optimize VNF deployment by migrating VNFs in response to structural changes. As summarized in Figure 3.5, we optimize the initial topology with ClusPR, induce network failures, and then automatically re-optimize the network based on the degraded network structure. This re-optimization occurs on each node repair until all nodes are repaired, and the network returns to its initial state. This strategy leverages a key functionality of NFV networks: centrally controlled network management through virtualization.

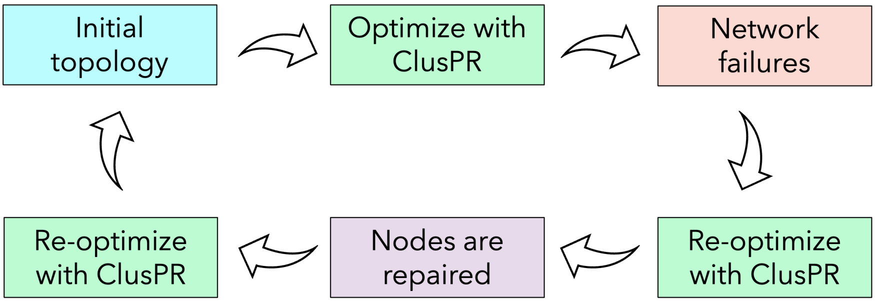


Figure 3.5: Re-optimization strategy

3.7.4 Combined Re-optimization and Specific Node Repair Order Strategy

The final strategy integrates the two previous strategies into one unified approach: it re-optimizes the VNF deployment every time the network structure changes while also prioritizing node repair order. This combined approach sheds light on how the two strategies interact, whether they complement or possibly duplicate each other, which could lead to redundant computation.

Chapter 4
Results

In this chapter, we present the results. After consecutively presenting each set of results, we also briefly discuss each one, as the main takeaways from some results lay the foundation for the experiments coming after. This gives more context to the choices made when applying recovery strategies. In Chapter 5 we conduct a more in-depth discussion that summarizes all results and discussions and also looks at them in their entirety.

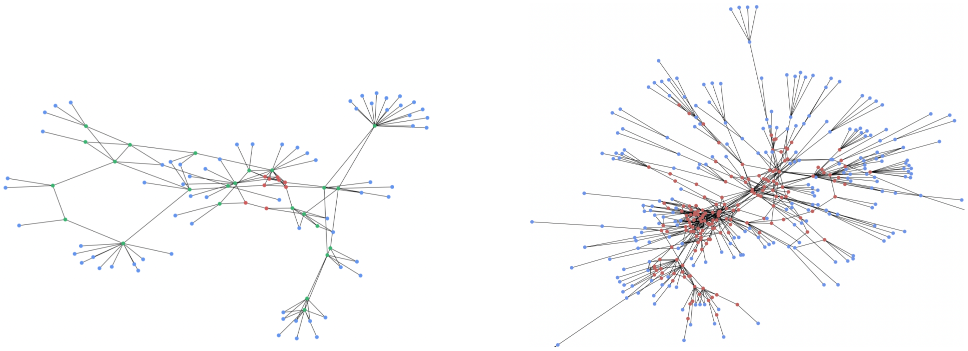
4.1 Experimental Setups

This section details the specific values for the experiments. Subsequent sections present the results.

4.1.1 Network Topologies and Flows

Experiments were conducted on two distinct network topologies, illustrated in Figure 4.1. The first topology, visualized in Figure 4.1a, comprises 102 nodes and 141 links, characterized by a high degree of clustering with a global clustering coefficient of 0.16. The second topology, shown in Figure 4.1b, contains 229 nodes and 471 links, exhibiting minimal clustering with a global clustering coefficient of 0.03. For simplicity, we will refer to them as *small topology* and *large topology* in the remainder of the thesis.

For each topology, we assigned a link delay of 3 ms from access to edge, 10 ms from edge to core, and 40 ms between core nodes. Each link has a capacity of 1 Gbps. Network traffic was simulated by randomly generating 720 flows per iteration for each topology, originating and terminating at access nodes and passing through edge and core nodes. Each flow was assigned a maximum tolerated delay, randomly chosen to be 1 to 2.5 times the delay of the flow's shortest path. Further, each flow required a unique sequence of network functions to form a SFC.



(a) Small topology: 102 nodes and 141 links, high clustering coefficient

(b) Large topology: 229 nodes and 471 links, low clustering coefficient

Figure 4.1: Test Topologies

4.1.2 Node Failures

We considered two main types of network failures: random node failures and targeted node attacks. Random node failures were simulated by disabling nodes at random for each simulation. In contrast, targeted failures involved selecting nodes based on node importance rankings, determined by either betweenness centrality or closeness centrality.

4.1.3 Node Repairs

We assumed nodes would be repaired and functional over time after a failure. The repair time of a node is assumed to follow an exponential distribution with a mean repair time of 5. We also assumed nodes to recover independently; the recovery of one node does not influence the recovery of others.

In all cases, we consider a scenario where three nodes fail concurrently, resulting in the Markov model depicted in Figure 4.2

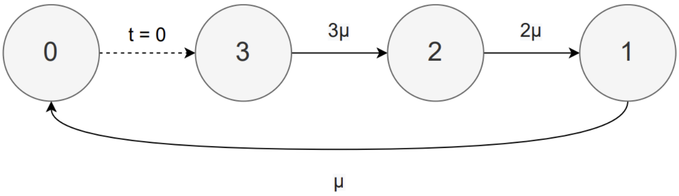


Figure 4.2: Markov model showing the recovery stages of a network where three nodes fail concurrently at time $t = 0$

4.2 Interpretation of Results

This section provides brief explanations of the two types of results obtained from the experiments.

4.2.1 Survivability Assessments

A survivability assessment is based on a given topology, a specific set of generated flows, and a particular set of failed nodes. The survivability assessment curve provides an evaluation of network performance following node failures. We define performance as the percentage of admitted flows at a specific time post-failure. The survivability curves show the recovery period for a network following network failures. Figure 4.3 illustrates example survivability curves from the initial experiments, showcasing the lowest performance immediately after the failure and gradual recovery over time. Each curve in the plot represents the average performance of all iterations for each case.

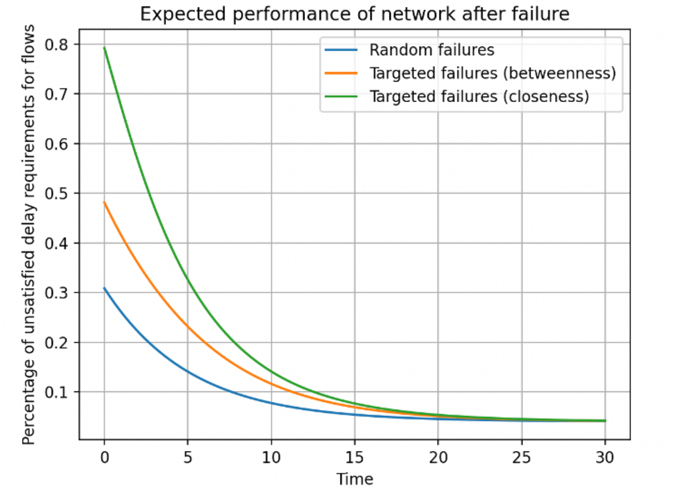


Figure 4.3: Example survivability curve showing the expected performance in the recovery phase of a network in three different cases: random failure, targeted failure (betweenness), and targeted failure (closeness)

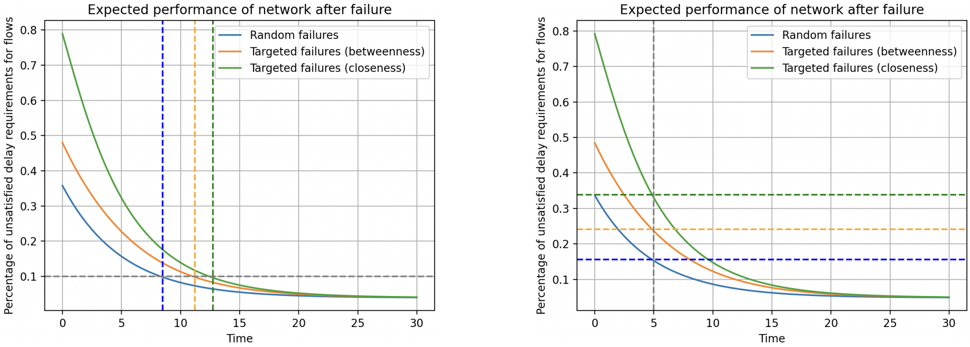
We analyze the survivability curves in two dimensions, temporal and spatial, depicted in Figure 4.4.

Temporal

Temporal evaluation of the network following failures means *how long before the network recovers to an acceptable performance level.* Figure 4.4a illustrates an example where we define 10 % of failed flows as acceptable performance and the time it takes for the network to recover to this level for each failure scenario (vertical lines).

Spatial

Spatial evaluation of the network following failures means *how much of the network recovers after a certain amount of time.* In Figure 4.4b we define time $t = 5$ as the time we want to evaluate network recovery, with horizontal lines depicting the performance (proportion of failed nodes) of the networks after time $t = 5$.



(a) Temporal evaluation

(b) Spatial evaluation

Figure 4.4: Temporal and spatial evaluation of the recovery phase of a network following three different failure scenarios

Both analysis methods are relevant to assess network recovery and often show similar trends. The temporal evaluation may be more important when a network needs a specific minimum accepted level of performance to function, while the spatial evaluation may be important for networks with a minimum amount of allowed downtime. In our experiments, both evaluations showed similar trends. While the horizontal and vertical lines used in the temporal and spatial evaluations are vital for our initial understanding of the network recovery, they are omitted in subsequent plots for clarity and readability. Despite their absence in these figures, it's important to remember that they lay the foundation for our analysis of the recovery strategies.

4.2.2 Causes of Failure

To add context to the results, we also computed the average cause of flow failure in each case. Figure 4.5 shows an example of this analysis, presenting the percentage of each failure cause for each number of failed nodes. The orange bar shows flows that receive all VNFs and reaches their destination but breaches their maximum tolerated delay. The blue bar displays that the source and destination of the flow are unreachable due to a separated topology, and the green bar shows that a flow does not receive its required VNF, as the node with the VNF has failed.

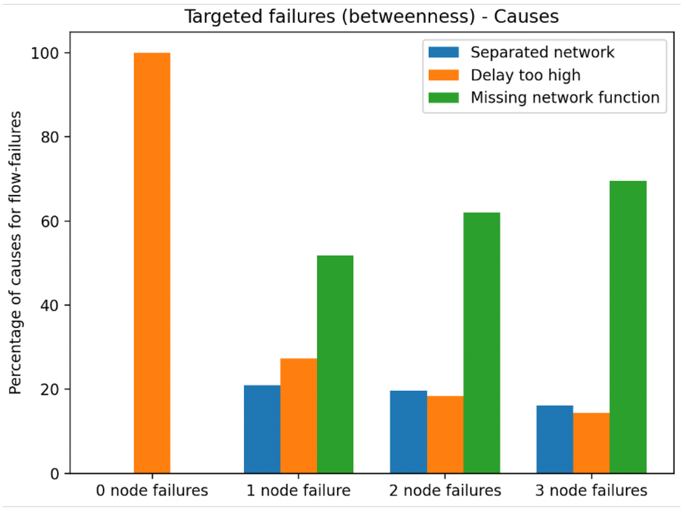


Figure 4.5: Example of flow failure causes, showing the most prevalent causes of flow failures for each number of failed nodes in the targeted failure (betweenness) case

4.3 Validation of Results

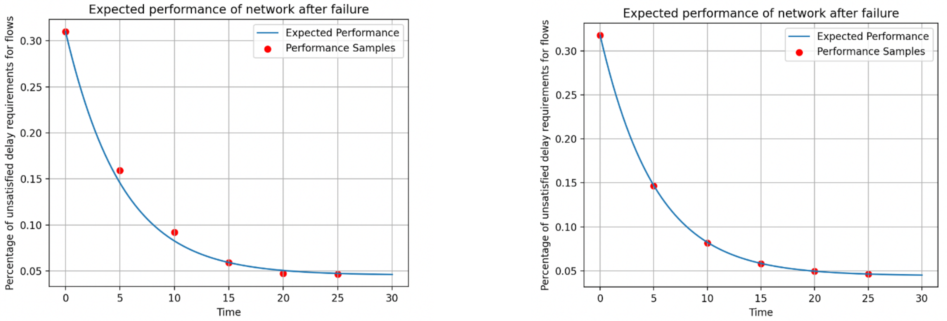
This section presents the methods we used to validate our findings, including conducting all experiments on two distinct topologies and simulating the same experiments to validate our analytical findings.

4.3.1 Analytical vs. Simulation

In addition to computing survivability curves using state probabilities and performing computations in each state in a Monte Carlo simulation, we also simulated the same experiments with random repair times and repair orders. Matching results in the two cases imply consistent results, and that we use enough iterations in the analytical approach to capture the variance of the inherent randomness. Figure 4.6 illustrates the difference between increasing the number of iterations conducted in the process versus 50 sets of generated flows and 50 sets of failed nodes per flow.

4.3.2 Two Topologies

Experiments were conducted on two distinct topologies to ensure the results were not specific to a given scenario, and to highlight possible differences between different topology structures.



(a) Example where simulation samples and analytical approach do not match, indicating variance caused by too few iterations in the analytical study (10 sets of generated flows and 10 sets of failed nodes per flow)

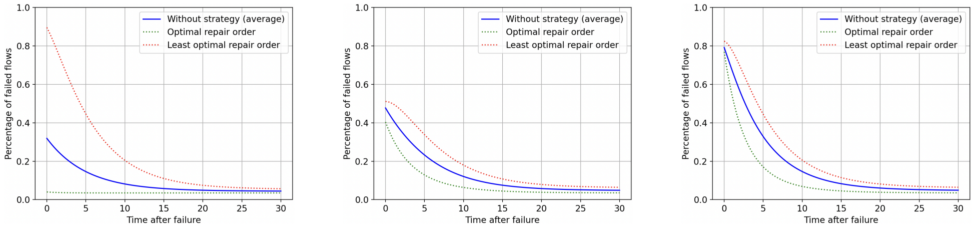
(b) Example where simulation samples and analytical approach match, validating that results from the analytical approach are consistent (50 sets of generated flows and 50 sets of failed nodes per flow)

Figure 4.6: Comparison between simulation of experiments and analytical approach

4.4 Baseline Strategy

Our baseline results, where we employ no specific strategy to recover the network after a failure, provide a reference for comparing our recovery strategies. These experiments involve scenarios with no pre-determined repair order or VNF migration. Figure 4.7 presents the average, best, and worst-case outcomes of three different node failure scenarios. The random failures have the least impact on the network performance and show the quickest recovery, but also exhibit the greatest variance, as illustrated by the best case having minimal impact on performance, whereas the worst case eliminates nearly 90% of the network. Betweenness and closeness failures degrade the network more, with closeness failure inflicting the most damage in the small topology (however, betweenness inflicted the most damage in the large topology). The observed variance in these scenarios stems from our node repair strategy, which examines all possible sequences of node repair orders. This allows us to contrast the worst and best repair order in both attack cases, demonstrating the significance of effective node repair orders.

In Figure 4.8, we observe the causes of failure in each case when no recovery strategy is in place. Across all scenarios, flow failures are solely due to delay requirement breaches when there are no node failures. Additionally, the primary cause of flow failure in all cases is node failure for nodes hosting the flow's VNF. We note that the severe impact of a closeness attack on the small topology is due to significant network segmentation. This was not the case with the large topology.

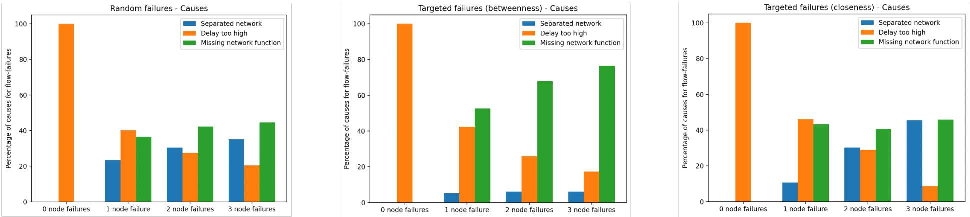


(a) Performance after random failures, without recovery strategy.

(b) Performance after betweenness attack, without recovery strategy.

(c) Performance after closeness attack, without recovery strategy.

Figure 4.7: Three different failure cases: random and targeted (betweenness and closeness), with average (blue), best (green), and worst (red) case scenarios



(a) Random failures, average causes of flow failure

(b) Betweenness attacks, average causes of flow failure

(c) Closeness attacks, average causes of flow failure

Figure 4.8: Average percentage of flow failure causes for all three cases (random failures, betweenness attacks, and closeness attacks)

4.4.1 Discussion of Baseline Strategy

First, our results underline how node failures negatively impact a network's ability to deliver its services. Some nodes are more critical than others, as exemplified by the significant variance between the best and worst-case outcomes in random failure scenarios (Figure 4.7a). This outcome underscores the presence of critical nodes whose failure can drastically hinder a VNF network's performance (as seen in the worst-case curve). Conversely, some nodes have minimal effect on performance upon failure, emphasizing their relative insignificance in the network.

Targeted attacks, on average, yield more detrimental effects on network performance than random failures. Notably, the impact of attack types appears to be contingent on the network topology. While the small topology suffers more from closeness attacks, the large one is more affected by betweenness attacks. This outcome suggests that enhancing network resilience may require a tailored approach instead of a one-size-fits-all strategy, given its dependence on the topology. The same principle applies to attackers aiming to inflict maximum damage.

We had expected the nature of betweenness centrality and closeness centrality to yield different causes of flow failures when applied in attacks. Betweenness centrality, which measures how well a node "bridges" large network parts, should result in more network partitioning. In contrast, closeness centrality, indicating the number of shortest paths associated with a node, should lead to increased delay-induced failures. This held for the larger topology but reversed for the smaller one, as depicted in Figure 4.7c, where the closeness attack heavily partitioned the network. We attribute this unexpected outcome to the smaller network's size, where each node failure can have a disproportionately large impact. Regardless, it illustrates that an attack may not always yield the anticipated effects.

Most importantly, we found that the majority of flow failures stem from node failures hosting VNFs. This observation is significant because, unlike betweenness or closeness attacks, these nodes operate on the virtual layer rather than the physical layer. This could imply that the extent of damage an attacker can cause depends on their knowledge of these different layers. An attacker equipped with information on the virtual layer could exploit this knowledge to inflict even more harm. This finding laid the groundwork for the recovery strategy discussed in Section 4.5.2, where we introduce the concept of *flow centrality* to identify nodes hosting VNFs for numerous flows.

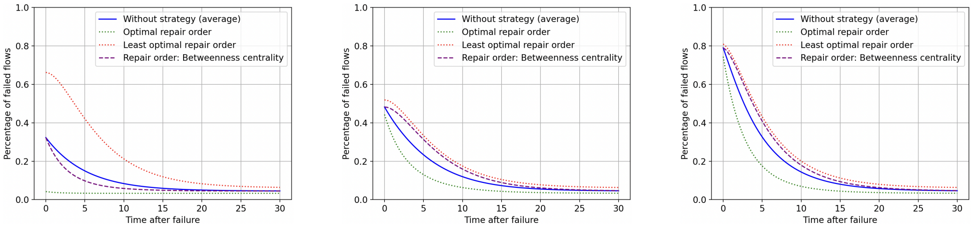
To evaluate the effects of implementing dedicated recovery strategies during failure simulations, we first examine each strategy separately and then consider a combination of both.

4.5 Node Repair Order Selection

Experiments utilizing pre-defined node repair order were conducted three times, each time prioritizing nodes based on different measures: betweenness centrality, closeness centrality, and flow centrality (as explained in Section 3.7). Given the similar trends observed in the results of betweenness and closeness repairs, we illustrate only the betweenness centrality results for conciseness.

4.5.1 Betweenness and Closeness Centrality as Node Repair Order

The results of using either betweenness or closeness centrality as the basis for node repair order varied depending on the network structure. For the small topology, these repair orders performed worse than random node repair orders, as shown in Figure 4.9. In this figure, the blue, green, and red lines represent the same failures as before but without employing any recovery strategy, while the purple dashed line illustrates the failure rate when applying the strategy. The pre-defined repair strategy improves the network's resilience against random failures but performs worse in the event of betweenness and closeness attacks. Notably, having a defined repair order removes the variance in the results, leading to more predictable outcomes. For the large topology, we saw a slight improvement in all cases.



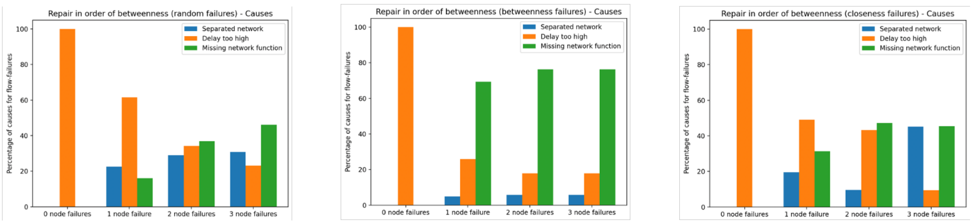
(a) Random failures, repair order: betweenness (purple) compared to random repair order

(b) Betweenness attacks, repair order: betweenness (purple) compared to random repair order

(c) Closeness attacks, repair order: betweenness (purple) compared to random repair order

Figure 4.9: Effect of applying betweenness centrality as a priority when repairing nodes compared to a random repair order

A deeper understanding of this strategy's performance can be obtained by examining the causes of flow failures in Figure 4.10. For random failures, the two reasons—"separated network" and "missing network function"—recover more quickly than with a random repair order. The opposite is observed in the case of betweenness and closeness attacks. Alternatively, this can be interpreted as an increased rate of "too high delay" as the reason for flow failures in the case of betweenness and closeness attacks. These results indicate that pre-defined node repair order is more beneficial in random failure scenarios rather than targeted attacks.



(a) Random failures, betweenness repair - average causes of flow failure

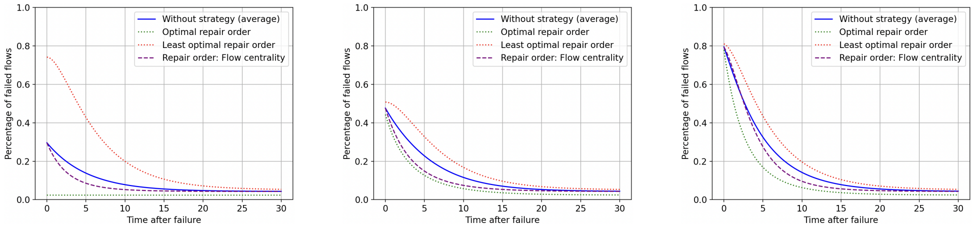
(b) Betweenness attacks, betweenness repair - average causes of flow failure

(c) Closeness attacks, betweenness repair - average causes of flow failure

Figure 4.10: Average percentage of flow failure causes for all three failure cases, with betweenness repairs as recovery strategy

4.5.2 Flow Centrality as Node Repair Order

Turning our attention to the application of flow centrality as a recovery strategy, it can be seen that flow centrality outperforms betweenness centrality as node repair order in the tested network topology, as shown in Figure 4.11. Compared to a random repair order, flow centrality as a repair order increases the network's resilience to random, betweenness, and closeness attacks, maintaining the network's performance above 40% in all cases. However, the rate of recovery does not improve significantly. This suggests that flow centrality as a repair order helps to mitigate the initial impact of the failure but does not expedite the recovery process.



(a) Random failures, repair order: flow (purple) compared to random repair order

(b) Betweenness attacks, repair order: flow (purple) compared to random repair order

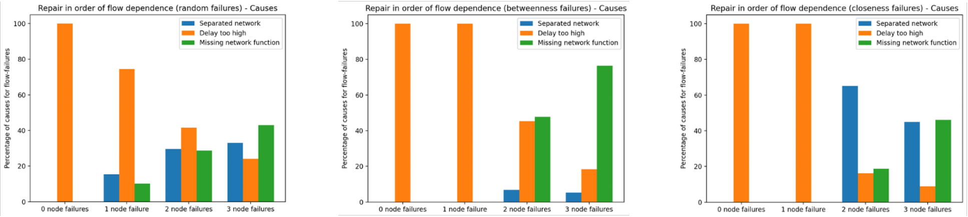
(c) Closeness attacks, repair order: flow (purple) compared to random repair order

Figure 4.11: Effect of applying flow centrality as a priority when repairing nodes compared to a random repair order

A detailed analysis of the causes of flow failures in the case of flow centrality as a repair order, presented in Figure 4.12, reveals that the flow centrality strategy's advantage comes from its ability to preserve network functions and maintain connectivity. However, it cannot reduce the number of failures due to high delay. This indicates a limitation of the flow centrality strategy—it is effective in maintaining a functional network but is unable to ensure that the network's performance meets all requirements.

4.5.3 Discussion of Node Repair Order Selection Strategy

Key insights gleaned from implementing a pre-defined order of node repair highlight its potential to both expedite and delay node recovery. The strategy accelerated recovery time for the large topology, yet performed nearly as poorly as the worst-case repair order the small one. This finding illustrates the challenges associated with developing a one-size-fits-all rule for node prioritization. Surprisingly, even when nodes with high betweenness centrality were targeted, pre-defining their repair led to poorer performance.



(a) Random failures, flow centrality repair - average causes of flow failure

(b) Betweenness attacks, flow centrality repair - avg. causes of flow failure

(c) Closeness attacks, flow centrality repair - average causes of flow failure

Figure 4.12: Average percentage of flow failure causes for all three failure cases, with flow centrality repairs as recovery strategy

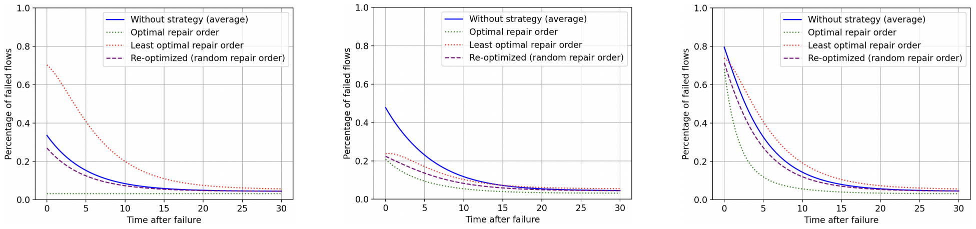
A review of the failure causes when implementing this strategy, as demonstrated in Figure 4.10, shows a high number of "missing VNF" causes, persisting even after both 1 and 2-node repairs. This suggests that the strategy fails to adequately address the most crucial nodes in these NFV networks, specifically those hosting VNFs for many flows.

Motivated by these findings, we developed a novel metric called flow centrality to identify critical nodes in these networks. The results from prioritizing these nodes for repair were promising in all cases, with accelerated recovery times observed across both topologies and failure types. An examination of the flow failure causes revealed a faster removal rate of the targeted failure cause as nodes were repaired. Therefore, pre-defined node repairs based on flow centrality appear to be a sound strategy for accelerating network recovery post-failures.

4.6 Re-optimization Strategy

In this subsection, we examine re-optimization as a recovery strategy. Re-optimization refers to the automatic migration of VNFs from failing nodes to other operational nodes upon the occurrence of failures. The repair order of the nodes is random, meaning we do not employ any specific node repair order strategy. Figure 4.13 depicts the improvements in all scenarios across all topologies when we apply this strategy. Importantly, it also shows the variance in node repair order while utilizing the re-optimization strategy, revealing that the performance of this strategy is not solely dependent on the repair order.

Significantly, even the worst-case order during the re-optimization strategy outperforms the expected value without re-optimization for betweenness attacks, as shown in Figure 4.13b. For closeness attacks, while the performance improvement is not as significant as that for betweenness attacks, it is still an improvement.



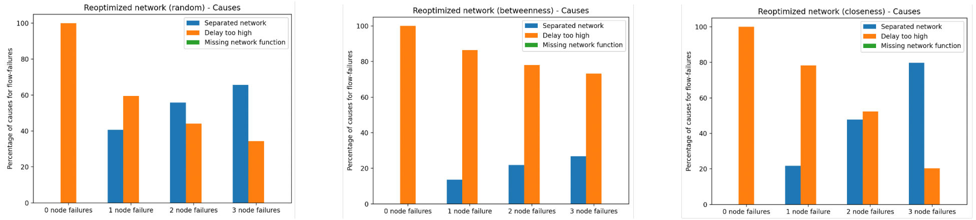
(a) Random failures, with re-optimization (purple) compared to no re-optimization

(b) Betweenness attacks, with re-optimization (purple) compared to no re-optimization

(c) Closeness attacks, with re-optimization (purple) compared to no re-optimization

Figure 4.13: Effect of applying the re-optimization strategy compared to a no re-optimization approach

The causes of flow failures when using the re-optimization strategy are presented in Figure 4.14. Interestingly, the primary cause of flow failure from previous experiments—the missing VNF—is eradicated when the re-optimization strategy is applied. However, this strategy does not speed up the resolution of network separation in the case of random failures and closeness attacks, as shown in Figures 4.14a and 4.14c. This explains the lesser performance gain from this strategy in these scenarios compared to betweenness attacks. Nevertheless, the worst-case repair order in closeness attacks performs worse than the expected performance using the baseline strategy, while the best-case repair order coupled with re-optimization hints at substantial performance gains.



(a) Random failures, with re-optimization - average causes of flow failure

(b) Betweenness attacks, with re-optimization - avg. causes of flow failure

(c) Closeness attacks, with re-optimization - average causes of flow failure

Figure 4.14: Average percentage of flow failure causes for all three failure cases, with re-optimizations as recovery strategy

4.6.1 Discussion of Re-optimization as Strategy

The application of re-optimization as a strategy has consistently expedited network recovery post-failure across all topologies and failure instances. This strategy, much like the flow centrality repair order, directly addresses the principal cause of flow failures. It eradicates the issue, as VNFs are redeployed across the network. However, as illustrated in Figure 4.13c, if the network is severely fragmented due to failures, this does not necessarily improve the situation as source and destination nodes within flows become disconnected. From the worst-case scenario depicted in the same plot, we infer that despite the application of re-optimization and the subsequent removal of the root cause of failure, a poor repair order can result in an outcome that is inferior to not implementing any strategy at all, highlighting the challenge in devising a universal strategy.

A significant constraint to consider is the cost and time consumed by the re-optimization process each time the network is re-optimized. If VNFs are redistributed across the network, flows previously unaffected by network failures may now be impacted. Additionally, the re-optimization process could require substantial computational resources, which may be a crucial consideration for network operators. Although we've demonstrated that a re-optimization strategy can significantly aid in network recovery, exploring more efficient ways to re-optimize a network following failures is an intriguing direction for future research, which we intend to pursue.

4.7 Combined Strategy

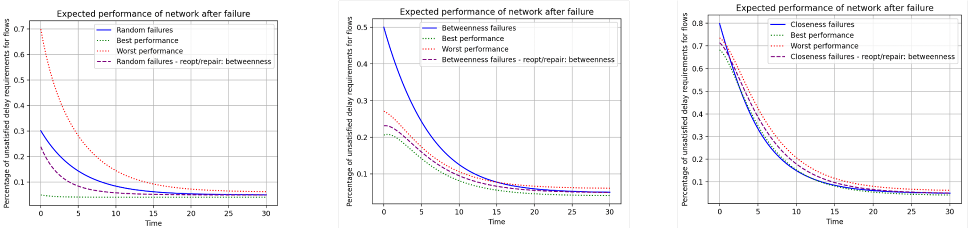
Finally, we study the impact of combining two strategies—re-optimization and pre-defined node repair. While we re-optimize the network upon any structural change, we also determine a specific order for node repairs. Figure 4.15 illustrates the impact of this combined strategy. Notably, the performance variance is further reduced due to the defined repair order, and the remaining variance is attributable to different sets of flows for each simulation iteration.

Recall that re-optimization alone improved performance compared to the baseline strategy, while the strategy of pre-defined node repair based on betweenness resulted in worse performance in this specific case. Interestingly, combining these strategies results in worse performance than using re-optimization alone due to the negative effect of betweenness centrality on recovery time.

When flow centrality is used for repair orders (which improved performance as a stand-alone strategy), the combined strategy led to improved performance across all aspects. This trend was consistent across both topologies.

4.7.1 Discussion of Combined Strategy

For the combined strategy we in all cases saw the summation of the two strategies. By summation we mean that whenever both strategies showed increased performance individually, the combined strategy performed better than both, and while one of



(a) Random failures, with combined re-optimization and pre-defined node repair order based on betweenness

(b) Betweenness attacks, combined re-optimization and pre-defined node repair order based on betweenness

(c) Closeness attacks, with combined re-optimization and pre-defined node repair order based on betweenness

Figure 4.15: Effect of applying the combined re-optimization and node repair order (betweenness) strategy, compared to a no re-optimization approach

the strategies was better and one was worse, the combined strategy performed worse than the best one, showing that a bad strategy will negatively impact a good strategy if combined.

One of the most interesting findings is the one in Figure 4.15c, where the combined strategy performs worse than the expected performance of applying the baseline strategy. This shows the disadvantage of employing a bad repair order, even though it is combined with a good strategy. It should however be noted that the variance is much higher when not applying any strategy, as you would not migrate any VNFs and could potentially have a very bad node repair order.

4.8 Summary of Results

Given the multitude of plots across multiple topologies, we have summarized the results in Table 4.1. The comments column notes any particularities about the results. The table applies only to the small topology, as the strategies improved results in all cases for the large topology.

In summary, the re-optimization strategy generally provides improved performance. When combined with pre-defined node repair based on betweenness, however, performance worsens in some cases. On the other hand, the combined strategy of re-optimization and pre-defined node repair based on flow centrality improves performance across all aspects.

Table 4.1: Performance of Various Strategies Compared to The Average Performance of the baseline strategy

Strategy	**Failure Type**	**Performance**	**Comment**
Betweenness repair	Random failures	Improved	-
	Targeted failures	Worse	Close to worst case
Closeness repair	Random failures	Improved	-
	Targeted failures	Worse	Close to worst case
VNF repair	Random failures	Improved	-
	Targeted failures	Improved	Close to best case
Re-optimization	Random failures	Improved	-
	Targeted failures	Improved	Low variance
Combined (Betweenness repair)	Random failures	Improved	-
	Targeted failures	Improved	Worse than re-optimization, lower variance
Combined (Closeness repair)	Random failures	Improved	-
	Targeted failures	Worse	Even best case has worse performance
Combined (VNF repair)	Random failures	Improved	-
	Targeted failures	Improved	-

Chapter 5 Discussion

5.1 Implication of the Findings

Our research compared various strategies for shortening the recovery time of NFV networks during and after network failures. These failures refer to multiple concurrent random node failures and targeted node (cyber) attacks. We investigated the performance of different strategies in terms of network flow survivability, which included prioritizing node repair order based on three metrics (betweenness centrality, closeness centrality, and flow centrality). The second strategy was re-optimizing the network after changes in network structure (after failure and repairs), while the last strategy was a combination of these strategies. Our findings, which display diverse outcomes depending on the selected strategy and the type of network failures, provide valuable insights for network operators and researchers.

Our study relied on a quantifiable measure of survivability, where we evaluated how network performance evolves post-failure. This framework enabled us to compare different strategies, observing the speed at which each strategy returns the network to its standard operational state. Within this framework, we looked at both the temporal and the spatial dimension. The temporal dimension evaluates the time taken to restore the network to its normal performance, while the spatial dimension considers what size of the network is restored after a specified duration. The choice of comparing the spatial or temporal dimension depends on the specific assessment objectives. A comparison of the speed of full network recovery versus the extent of network capacity recovered after a certain duration could yield different outcomes. Nonetheless, given our findings, both cases yielded similar conclusions.

5.1.1 Primary Conclusions

Here, we list the most noteworthy observations drawn from our study.

Topology Structure Affects Resilience

Network topology structure profoundly impacts network resilience. As a result, the topology structure determines the vulnerability of a network to different types of attacks and the effectiveness of the recovery strategies implemented. Some topologies are degraded significantly more by attacks than others. The same also holds for recovery strategies. Attackers could utilize this insight in physical sabotage or virtual cyber attacks, depending on their knowledge of the network.

Flow Centrality is an Effective Repair Order Indicator

In the domain of NFV networks, flow centrality outperforms betweenness and closeness centrality as a heuristic for determining the node repair order. In all cases, repairing nodes based on flow centrality recovered the network faster than traditional metrics. The leading cause of performance degradation in these networks is the failure of nodes hosting VNFs. Consequently, these nodes become the prime targets for attackers and network operators alike, particularly in network design.

Re-optimization Proves Beneficial

Re-optimizing VNF deployment (i.e., migrating VNF) after structural network changes consistently enhanced network performance, regardless of the topology structure or attack type. However, it's worth noting that VNF re-deployment has consequences, such as migration time and computational costs, that need to be accounted for and further researched.

Optimal Strategies May Yield Suboptimal Results

Figure 4.15c illustrates that the best-case scenario using a combined strategy may result in lower performance than a scenario in which no specific recovery strategy is applied (baseline scenario). Hence, optimal performance in a strategy doesn't necessarily yield superior results compared to a baseline approach with no specific strategy. This highlights the complex and context-dependent nature of network survivability.

5.1.2 Practical Implications

Network Design and Recovery Strategies

Introducing a virtual layer in VNF networks creates a new vulnerability, i.e., the nodes hosting VNFs. If attackers gain insight into the virtual topology of a network, they can exploit this knowledge and target nodes with high flow centrality to inflict maximum damage. These nodes, therefore, become critical for network operators when designing their networks. They demand priority in VNF deployment, network design, recovery strategies, and redundancy planning.

Services with Stringent Delay and Resilience Requirements

Certain services have stringent delay requirements, some demand high resilience, and some require both. In NFV-RA frameworks, priorities must be set when determining VNF deployment. If the delay is prioritized, survivability objectives are downplayed, and vice versa. Our study offers insights into this trade-off by considering strategies and survivability assessments that do not affect the objective functions of NFVRA. This enables the optimization of the delay in NFVRA while still ensuring survivable service delivery.

The Absence of a Universal Strategy

Our study emphasizes that there is no universal strategy that ensures survivability. Our experiments show vastly diverse outcomes across different network structures, attack types, and recovery strategies. However, the study provides valuable insights that can inform more resilient network designs, including automatic strategies to respond to network outages. Using flow centrality as a heuristic when prioritizing node repair order can enhance network recovery strategies in NFV networks, as it outperforms traditional metrics such as betweenness and closeness centrality.

5.1.3 Limitations and Future Directions

Can We Provide Performance Guarantees for Network Slicing?

Network slicing is a process that divides networks into multiple virtual networks that can run different applications and services. Each slice can be prioritized, with some demanding performance guarantees in delay and availability, especially in 5G networks. The insights from our study can inform the design of resilient VNF deployment, which in turn can be used in the implementation of network slicing. An interesting extension to our research could be to incorporate flow priority. This could enable network operators to guarantee capacity within the network, even in the face of failures, and use re-optimization to ensure the operation of critical (top priority) flows.

Trade-offs of Re-optimization

While our study highlights the potential benefits of re-optimization, we also acknowledge that the strategy overlooks certain practicalities associated with VNF migration. There are costs involved, such as migration time and computing power, which we did not account for. Future work could involve studies on effective VNF migration that consider these trade-offs, or the development of more cost-effective high-level strategies that only re-optimize the affected parts of the network.

Different Attack Types and Virtual Vulnerabilities

The knowledge an attacker has about the network is a crucial factor. Whether the attacker knows the physical or the virtual infrastructure can drastically impact the damage they can inflict. Therefore, future work could include more realistic attack scenarios and investigation of security vulnerabilities related to knowledge of the virtual layer, thereby informing a more robust and resilient virtual network design.

Chapter 6 Conclusion

This thesis explored the enhancement of resilience in NFV networks. Our research revealed insights into the significance of the network structure, the importance of certain nodes, and recovery strategies of NFV networks. We studied multiple failure types, including random failures (natural disasters or software bugs) and targeted (cyber) attacks. We introduced a novel metric for node importance in NFV networks called *flow centrality* and identified it as a reliable metric for pre-defining node repair order post failure. Additionally, we demonstrated the benefits of leveraging softwarization for post-failure network re-optimization. However, we also found that seemingly optimal strategies do not necessarily guarantee the best outcomes, highlighting the complexity and context-dependent nature of network survivability.

Our findings carry practical implications, providing network operators with insights into network design, particularly concerning the nodes hosting VNFs. Moreover, our research emphasizes the importance of striking a balance between the objectives in NFV-RA, and how recovery strategies can provide solutions to trade-offs when optimizing for conflicting objectives.

Future research could explore multiple avenues suggested by this study. Among them the work on performance guarantees for network slicing in 5G networks, the costs and practicalities associated with re-optimization post-failure, and the impact of an attacker's knowledge of the network structure on security vulnerabilities introduced by virtualized networks.

In conclusion, while our research has advanced the understanding of NFV network resilience, it also underscores the field's complexity. As the digital infrastructure continues to grow, the findings of this research contribute towards building more robust and resilient networks better equipped to withstand future challenges. This work sets the stage for more nuanced investigations into network design and resilience, which should focus on developing more sophisticated and context-aware repair and recovery strategies to optimize network survivability in NFV environments.

References

[3GP16] 3GPP, «3gpp tr 22.891», 3rd Generation Partnership Project (3GPP), Technical Report (TR) 22.891, version 14.2.0, Sep. 2016. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2897.

[3GP18] 3GPP, «Release 15 description; summary of rel-15 work items», 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 21.915, version 15.0.0, Oct. 2018. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3389.

[3GP21] 3GPP, «3gpp ts 22.261», 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.261, version 16.16.0, Dec. 2021. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3107.

[All15] N. Alliance, «5g white paper», *Next generation mobile networks, white paper*, vol. 1, no. 2015, 2015.

[BAMH20] A. A. Barakabitze, A. Ahmad, *et al.*, «5g network slicing using sdn and nfv: A survey of taxonomy, architectures and future challenges», *Computer Networks*, vol. 167, p. 106 984, 2020.

[BBKW19] C. Bektas, S. Bocker, *et al.*, «Reliable software-defined ran network slicing for mission-critical 5g communication networks», in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[BBS16] M. T. Beck, J. F. Botero, and K. Samelin, «Resilient allocation of service function chains», in *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, IEEE, 2016, pp. 128–133.

[BMF+19] M. Belesioti, R. Makri, *et al.*, «A new security approach in telecom infrastructures: The resisto concept», in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2019, pp. 212–218.

[BMK+21] M. Belesioti, R. Makri, *et al.*, «Security and resilience in critical infrastructures», in *Technology Development for Security Practitioners*, Springer, 2021, pp. 317–333.

[BPD+20] L. Bonati, M. Polese, *et al.*, «Open, programmable, and virtualized 5g networks: State-of-the-art and the road ahead», *Computer Networks*, vol. 182, p. 107 516, 2020.

[DYL17] W. Ding, H. Yu, and S. Luo, «Enhancing the reliability of services in nfv with the cost-efficient redundancy scheme», in *2017 IEEE international conference on communications (ICC)*, IEEE, 2017, pp. 1–6.

[EFL+99] R. Ellison, D. Fisher, *et al.*, «Survivability: Protecting your critical systems», *IEEE Internet Computing*, vol. 3, no. 6, pp. 55–63, 1999.

[ercpfciO18] R. enhancement and risk control platform for communication infraSTructure Operators (RESISTO). «Resisto». (2018), [Online]. Available: https://www.resistoproject.eu/ (last visited: Jan. 5, 2023).

[ETS12] E. T. S. I. (ETSI). «Network functions virtualisation (nfv)». (2012), [Online]. Available: https://www.etsi.org/technologies/nfv (last visited: Jan. 5, 2023).

[FPEM17] X. Foukas, G. Patounas, *et al.*, «Network slicing in 5g: Survey and challenges», *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94–100, 2017.

[FYG+15] J. Fan, Z. Ye, *et al.*, «Grep: Guaranteeing reliability with enhanced protection in nfv», in *Proceedings of the 2015 ACM SIGCOMM Workshop on Hot Topics in Middleboxes and Network Function Virtualization*, 2015, pp. 13–18.

[GB16] J. Gil Herrera and J. F. Botero, «Resource allocation in NFV: A comprehensive survey», *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, 2016.

[GFH20] M. Gajić, M. Furdek, and P. Heegaard, «A framework for spatial and temporal evaluation of network disaster recovery», in *2020 32nd International Teletraffic Congress (ITC 32)*, IEEE, 2020, pp. 37–45.

[GOH+20] A. J. Gonzalez, J. Ordonez-Lucena, *et al.*, «The isolation concept in the 5g network slicing», in *2020 European Conference on Networks and Communications (EuCNC)*, 2020, pp. 12–16.

[HT09] P. E. Heegaard and K. S. Trivedi, «Network Survivability Modeling», *Computer Networks*, vol. 53, no. 8, pp. 1215–1234, Jun. 2009.

[KBDW18] F. Kurtz, C. Bektas, *et al.*, «Network slicing for critical communications in shared 5g infrastructures-an empirical evaluation», in *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*, IEEE, 2018, pp. 393–399.

[KS00] J. C. Knight and K. J. Sullivan, «On the definition of survivability», *University of Virginia, Department of Computer Science, Technical Report CS-TR-33-00*, 2000.

[LSC+17] X. Li, M. Samaka, *et al.*, «Network slicing for 5g: Challenges and opportunities», *IEEE Internet Computing*, vol. 21, no. 5, pp. 20–27, 2017.

[LT06] Y. Liu and K. S. Trivedi, «Survivability quantification: The analytical modeling approach», *International Journal of Performability Engineering*, vol. 2, no. 1, pp. 29–44, 2006. [Online]. Available: http://www.ee.duke.edu/~kst/surv/IoJP.pdf.

[MT19] I. Mikhalevich and V. Trapeznikov, «Critical infrastructure security: Alignment of views», in *2019 Systems of Signals Generating and Processing in the Field of on Board Communications*, IEEE, 2019, pp. 1–5.

[NetworkX] The NetworkX Website. [Online]. Available: https://networkx.org/ (last visited: Jun. 5, 2023).

[NORCICS23] NORCICS, 2023. [Online]. Available: https://www.ntnu.edu/norcics/ (last visited: May 31, 2023).

[OAL+17] J. Ordonez-Lucena, P. Ameigeiras, *et al.*, «Network slicing for 5g with sdn/nfv: Concepts, architectures, and challenges», *IEEE Communications Magazine*, vol. 55, no. 5, pp. 80–87, 2017.

[QASK17] L. Qu, C. Assi, *et al.*, «A reliability-aware network service chain provisioning with delay guarantees in nfv-enabled enterprise datacenter networks», *IEEE Transactions on Network and Service Management*, vol. 14, no. 3, pp. 554–568, 2017.

[QKA18] L. Qu, M. Khabbaz, and C. Assi, «Reliability-aware service chaining in carrier-grade softwarized networks», *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 3, pp. 558–573, 2018.

[RNDM23] RNDM 2023, 2023. [Online]. Available: http://www.rndm.pl/2023/ (last visited: May 31, 2023).

[Rocketfuel02] Rocketfuel: An ISP Topology Mapping Engine, 2002. [Online]. Available: https://research.cs.washington.edu/networking/rocketfuel/ (last visited: Mar. 23, 2023).

[SBT+17] T. Soenen, R. Banerjee, *et al.*, «Demystifying network slicing: From theory to practice», in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, 2017, pp. 1115–1120.

[Wes04] V. R. Westmark, «A definition for information system survivability», in *37th annual hawaii international conference on system sciences, 2004. proceedings of the*, IEEE, 2004, 10–pp.

[WLW+16] L. Wang, Z. Lu, *et al.*, «Joint optimization of service function chaining and resource allocation in network function virtualization», *IEEE Access*, vol. 4, pp. 8084–8094, 2016.

[WMRJ18] Y. T. Woldeyohannes, A. Mohammadkhan, *et al.*, «ClusPR: Balancing multiple objectives at scale for NFV resource allocation», *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1307–1321, 2018.

[XHJ13] L. Xie, P. E. Heegaard, and Y. Jiang, «Network survivability under disaster propagation: Modeling and analysis», in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 4730–4735.

[YBSS17] F. Z. Yousaf, M. Bredel, *et al.*, «Nfv and sdn—key technology enablers for 5g networks», *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2468–2478, 2017.

[YLT+20] S. Yang, F. Li, *et al.*, «Recent advances of resource allocation in network function virtualization», *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 295–314, 2020.

[Zha19] S. Zhang, «An overview of network slicing for 5g», *IEEE Wireless Communications*, vol. 26, no. 3, pp. 111–117, 2019.

Appendix A
Paper Under Review

Assessing NFV Network Recovery Strategies after Random Failures and Targeted Attacks

Trond Vatten, Yuming Jiang, Poul E. Heegaard
Department of Information Security and Communication Technology
NTNU - Norwegian University of Science and Technology
{trond.vatten|yuming.jiang|poul.heegaard}@ntnu.no

***Abstract*—This paper investigates the critical role of network survivability under random failures or targeted attacks in Network Function Virtualization (NFV) environments and assesses the survivability of NFV networks under different recovery strategies. Through a comprehensive application of the ClusPR framework, we simulated various failure events, which include both random and targeted failures. Our survivability assessment focuses on the temporal dimension and compares different repair sequences in network recovery and explores potential mitigation strategies. These strategies involve re-optimization of network function placement upon structural network changes and prioritization of node repairs according to different centrality measures. Our findings revealed that the resilience of the network was greatly dependent on its specific topology and that the efficacy of the recovery strategies varied accordingly. In some cases, the use of sub-optimal strategies appeared to decrease the network survivability, demonstrating the importance of the recovery strategy. The results provide valuable insights into improving the survivability of NFV networks in the face of failures and can have implications for network management and policy.**

***Index Terms*—Survivability, 5G, network slicing, critical infrastructure, NFV, random failures, targeted attacks**

I. Introduction

5G networks have defined use cases that will be enabled once 5G is fully implemented. Examples of these use cases include autonomous vehicles, remote surgical operations, and the integration of augmented reality into daily lives, each representing a unique traffic class. These applications present diverse performance needs, including low delay or latency, alongside dependability factors such as reliability and availability, which are often in conflict with each other. The full-scale deployment of these services has yet to be realized due to the incomplete implementation of essential technologies, such as network slicing.

Network slicing is considered a crucial enabler for meeting the stringent requirements of services in the 5G domain [1]. Even though network slicing was first introduced in 5G as early as 2016 as a means of dividing a physical network into smaller, logically isolated virtual networks, research communities, and network operators have yet to arrive at a consistent technical definition for network slices. A variety of methods have been proposed and research in this area remains ongoing. Without the flexibility offered by virtualization and softwarization, network slicing cannot be efficiently realized.

Network Function Virtualization (NFV) represents the transition from dedicated hardware components for each network function to running Virtualized Network Functions (VNFs) as software on generalized hardware. This offers a level of flexibility that is more scalable and effective in rapidly changing network environments, a necessary condition for creating network slices that meet the stringent requirements of modern network services [2].

NFV Resource Allocation (NFV-RA) refers to the optimal deployment of VNFs that form service chains, one of the main challenges in NFV [3]. Despite the scalability and efficiency provided by the added flexibility, it also introduces an increased complexity. The NFV-RA problem is proven to be NP-hard, and numerous solutions have been proposed, each optimizing different objectives using heuristic or meta-heuristic algorithms [4].

In network slicing, proper network function chaining is crucial to equip each slice with the necessary network functions. Different 5G use cases have a mix of high requirements for capacity, delay, dependability, and security. These requirements vary across use cases, all of which need to be delivered on the same physical network, with virtualized networks serving as a potential solution. At the same time, the Internet Service Providers (ISPs) focus on network scaling and cost reduction, and therefore network utilization must also be considered.

Despite the importance, dependability-focused NFV-RA algorithms and methodologies for assessing the survivability of NFV networks have received limited attention. To meet the diverse service guarantees required across all use cases, it is crucial to thoroughly evaluate an NFV network's performance during and after failures. The development of automated, on-demand strategies for managing network failures is equally important [5].

The main objective of this paper is to investigate the impact of (random or targeted) node failures and their recovery on network service degradation in an NFV network context. Specifically, the focus is on assessing the performance during and after an undesired event (e.g. an outage with multiple node failures), with the aim to compare different recovery strategies and heuristics to limit the consequences of random failures and/or targeted cyber attacks.

The remainder of this paper unfolds as follows: Section II provides a review of the relevant literature. Section III introduces our experimental setup and methodology, designed to evaluate the survivability of an NFV network effectively. In Section IV, we outline the strategies we have developed to mitigate the impact of node failures in diverse contexts. Subsequently, Section V showcases our results. In Section VI, we discuss our findings and their implications. Finally, Section VII concludes the paper, synthesizing the key points and proposing potential avenues for future research.

II. Related Work

A framework for quantification of system *survivability* was introduced in [6]. The framework aims to assess spatial and temporal effects on system performance when a major undesired event (such as a natural disaster) has occurred and during the system's recovery. The framework was applied to study three different events and their consequences on a network with virtual connections. This framework has also been applied to other systems and events, such as modeling disaster propagation [7] and combining temporal and spatial survivability evaluations [8]. These studies attest to the framework's adaptability across different network sizes and failure types. In this paper, we continue to expand the application of this framework, investigating its use in the assessment of virtualized networks employing Network Function Virtualization Resource Allocation (NFV-RA) algorithms.

Although survivability has been addressed within the context of virtualized networks, previous studies have primarily focused on the Virtualized Network Embeddings (VNE) problem [9], which shares similarities with the NFV-RA problem. Both problems concern resource allocation within virtualized networks. However, they differ distinctly in their nature: the NFV-RA problem addresses more dynamic network demands, while the VNE problem is inherently static.

A comprehensive survey on NFV-RA is conducted in [4]. Although the survey reviewed numerous frameworks, none adequately addressed survivability within the NFV-RA context. As of our knowledge, only a single recent study has attempted to address this gap [10], proposing a novel solution to the NFV-RA problem that integrates resilience constraints.

The advent of the 5G standard has introduced a wide array of traffic classes, each bearing stringent and often conflicting requirements that must be met within a shared physical infrastructure [1]. These demands range from high data rates to ultra-reliable low-latency communication. To accommodate these diverse needs, many Network Function Virtualization Resource Allocation (NFV-RA) algorithms prioritize performance parameters such as end-to-end latency.

However, focusing predominantly on performance may risk undermining the survivability of the network, especially under failure conditions. Instead of integrating survivability directly into the NFV-RA algorithms, which may potentially compromise performance, we propose and assess separate strategies that the network can employ to react to network failures. These strategies aim to enhance network survivability without significantly impacting the other objectives of the NFV-RA algorithms.

III. Experimental Setup and Methodology

To assess the survivability performance under different failure recovery strategies for NFV networks, an extensive experimental study has been conducted. Our experimental approach consists of several steps as depicted in Fig. 1: (1) Initialize network topology; (2) Generate flows; (3) Run NFV-RA algorithm to optimize service component deployment; (4) Simulate node failures; (5) Evaluate the network's survivability. The loop of simulating node failures and assessing survivability is iteratively executed to capture the variance in impacts of different node failures. Ultimately, we return to the flow generation stage and repeat the entire process to account for variability in the generated flows. The approach has been implemented in Python, emphasizing modularity to ease the replacement for different setups, e.g. networks and node failures.

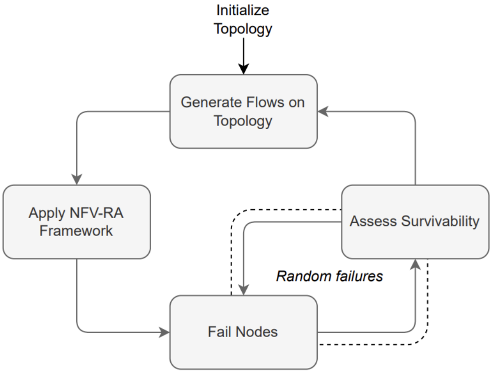


Fig. 1: Overview of the experimental setup

A. Initializing Topology and Generating Flows

Our experimental NFV networks consist of access nodes and core nodes, connected through edge nodes. The link delays and capacities are defined realistically to replicate practical network conditions. These networks employ realistic topologies, modeled using the Rocketfuel topology mapping engine based on real-world IP data [11]. We investigate distinct topologies to offer a broader perspective on the results, each having different characteristics regarding node degree distribution and clustering levels.

To generate network traffic, we create a set of flows that originate from one access node and terminate at another. These flows are designed to require a sequence of network functions, effectively forming a network function chain. Furthermore, each flow is subject to a maximum tolerable delay, intended to simulate real-world latency constraints.

The specifics about the topologies, the number of flows, their characteristics, and tolerable delay ranges, will be presented in Section V, for each experimental scenario.

B. Application of NFV-RA

To optimize the NFV service components deployment, we use the method implemented in the ClusPR framework [3]. It is designed to optimize NFV-RA considering both delay and network utilization. This dual-objective approach is suitable for 5G use cases, ensuring client performance (meeting delay requirements) and operator scalability and cost-efficiency (optimized network utilization). ClusPR only implicitly considers service dependability. This setup will enable us to pinpoint mechanisms that mitigate the network's survivability during failures.

C. Simulation of Network Failures

Our study encompasses both random node failures, such as multiple failures caused by e.g., a natural disaster or bugs introduced during software updates, and targeted node attacks, such as a cyber attack utilizing structural knowledge about the system/network.

For random failures, the impact significantly depends on the specific nodes that fail. Simulation of the effect of such random failures requires multiple iterations with different sets of node failures to capture this variance. For targeted failures, our attacks are based on node importance under some centrality measures, which means that the same set of nodes would be attacked for a specific set of flows and topology. Therefore, only a single iteration of node failures and survivability assessment is conducted. We assume that the attacker has knowledge of the topology and can infer centrality measures from this information. Two types of attacks are assessed: those targeted nodes with the highest *betweenness centrality* and those targeted nodes with the highest *closeness centrality*. The attacks are described and discussed in more detail in Section VI.

D. Assessment and Quantification of Survivability

The performance of the network in every state (a specific flow configuration and a different number of failed nodes) is evaluated. This is used as the basis for a reward in a Markov reward model representing the stages in the recovery of the network after the node failures. Specifically, the reward of each state is the network performance, defined as the percentage of admitted flows under this state. A flow is considered admitted if it visits all required network functions (service components) in the correct order while still complying with its delay requirement. Immediately after a failure, the number of admitted flows is at its lowest but gradually increases as nodes are repaired.

This process can be visualized in Fig. 2, which illustrates a state transition diagram of the network recovery process. The diagram captures the states as the number of failed nodes, from three initially down to zero, which is an absorbing state representing a fully functioning network. The transitions between states are dictated by the repair intensity μ, which depicts the rate at which nodes are repaired.

For each node repair or re-optimization, the network performance is re-evaluated. Ultimately, the network returns to its initial pre-failure state. The survivability is the sum of rewards, weighted with the probabilities of the states in the Markov model where the initial state ($t = 0$) is node failures, and the absorbing state is when all nodes have been repaired and the network functions have been re-optimized.

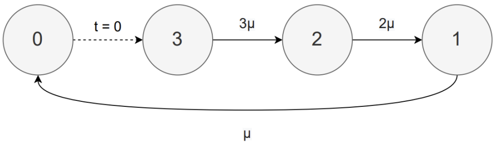


Fig. 2: State transition diagram for network recovery following 3 node failures

The survivability depends on the sequence of node repairs. The optimal sequence is generally not known, so to find the best sequence we take a brute force approach where we apply all possible permutations of node repair order, and select the one with the shortest time until recovery. This is used as a baseline to compare alternative failure recovery heuristics that we will consider. We remark that while for 3 node failures, there are only 6 permutations, the number of permutations increases quickly (more than exponentially) with the number of failures, and hence an alternative approach to the brute force approach may be needed.

As an enhancement to the survivability assessment, we also investigate the causes for flows failing to meet their requirements at each state, which are:

- A flow may not meet its delay requirement along its path;
- Network fragmentation may prevent a feasible path between the source and destination node of a flow;
- The last cause, unique to NFV networks, arises when the node hosting a VNF required by a flow fails, preventing the flow from obtaining its necessary network function chain.

IV. Recovery Strategies

When recovering from a major failure, we assume that the failed nodes can be repaired (in random or specific order). In addition, we assume that the appealing property of NFV will be exploited, i.e., the VNFs can be migrated (e.g. by re-optimizing the deployment). They form the basis for our recovery strategies as introduced in the following.

A. Prioritizing Node Repair Order

An initial experimental study indicated that some sequences for node repair allowed the network to return to its normal performance levels more quickly than others. Additionally, given the distinct structures and characteristics of different network topologies, a repair order that proves beneficial in one scenario may not necessarily yield the same effectiveness in another. We investigate three different rules for prioritizing node repair in this paper.

One is to repair nodes based on their betweenness centrality levels: the node with the highest betweenness centrality is repaired first. Another is to repair nodes based on their closeness centrality levels: the node with the highest closeness centrality is repaired first. Both betweenness and closeness are classical centrality measures that are used to characterize the importance level of a node to the network.

In this work, we introduce a novel metric for node importance that is specific to NFV networks. Unlike betweenness centrality, which counts the number of shortest paths in the physical or logical network structure, we consider the number of affected flows dependent on the VNFs hosted by a node if it fails. Accordingly, under this new centrality measure, nodes are ranked based on the number of flows assigned to a VNF on that specific node. For simplicity, we will refer to this metric as **flow centrality** in the remainder of the paper. The third rule is to repair nodes based on their flow centrality levels.

B. Re-optimizing VNF Deployment

Migrating VNFs to other nodes implies re-optimization of the NFV-RA problem. In other words, any failure and recovery in the network structure will change the system on which the VNF deployment was optimized. To this aim, we propose to re-apply the ClusPR framework to re-optimize resource allocation in the remaining NFV network. Accordingly, the experimental setup then becomes to initialize network topology and flows, apply ClusPR, inject node failures, apply ClusPR, first node repair, apply ClusPR, and repeat until all nodes have been repaired and the network returns to normal operation. This strategy targets the primary cause of node failures - flows not receiving their required network functions.

V. Survivability Assessment Results

A. Experiment Setups

1) Topologies and Flows: We conducted experiments on two distinct topologies, each demonstrating different characteristics. The first topology, with 229 nodes and 471 links, showcased a low degree distribution and a low level of clustering, with a clustering coefficient of 0.03. In contrast, the second topology, with 102 nodes and 141 links, exhibited a high degree distribution and a high level of clustering, characterized by a clustering coefficient of 0.16. The last topology is illustrated in Fig. 3. Both topologies consisted of access nodes and core nodes, connected through edge nodes. We used the following link delays: 3ms from access to edge, 10ms from edge to core, and 40ms for core to core connections. Each link possessed a capacity of 1 Gbps.

To simulate network traffic, we randomly generated 720 flows across these topologies. Each flow originated from an access node and terminated at another. We assigned a unique set of five network functions to each flow, which had to be executed in a specific order to form a network function chain. Furthermore, we assigned each flow a maximum tolerable delay, randomly chosen between 1 - 2.5 times the delay of the flow's shortest path.

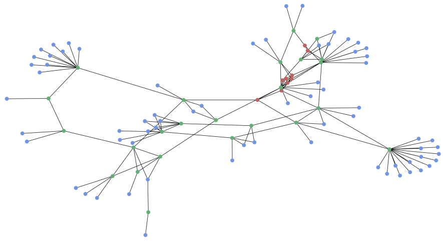


Fig. 3: One of the experimental topologies

2) Random Node Failures and Attack Strategies: We considered two types of network disruptions: random node failures and targeted node attacks. The aim was to understand how different types of disruptions can impact a network, forming the basis for our subsequent assessments of network recovery strategies.

Random node failures were simulated to represent scenarios such as natural disasters or software update errors. In these cases, nodes affected by the failure were selected at random, without regard to their role or position within the network.

Targeted node attacks were modeled to mimic deliberate cyber attacks, where the attacker has knowledge of the network structure and specifically targets nodes based on their centrality. We considered two attack strategies: one focused on nodes with the highest betweenness centrality and the other on nodes with the highest closeness centrality.

The recovery time of a failed node is assumed to follow an exponential distribution with a mean repair time of 5. We consider nodes to recover independently, meaning the recovery of one node does not affect the recovery time of other nodes.

B. Survivability for No Strategy

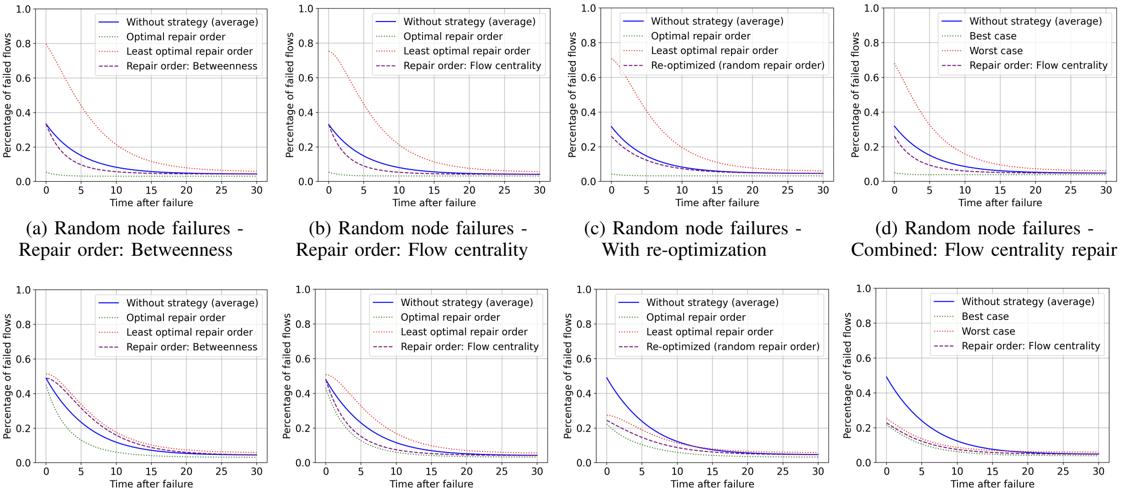
In this set of experiments, we did not implement any recovery strategy: Upon node failures, no VNFs were migrated, and node repairs were performed in random order.

We found that targeted attacks, which focused on nodes with high betweenness centrality or closeness centrality, had a more significant impact on network performance compared to random failures. This effect was consistent across both topologies.

We also observed that the two topologies were differently impacted by the type of node failure. One topology was particularly affected by attacks targeting nodes with high closeness centrality, while the other was most affected by attacks targeting nodes with high betweenness centrality.

Upon analyzing the causes of network service degradation, we identified a primary factor that significantly influenced the outcomes. The primary determinant was the failure of nodes hosting VNFs. In all scenarios, irrespective of the type of node failure or the topology configuration, the majority of flows failed to meet their service requirements when nodes hosting VNFs experienced failure.

Before discussing recovery strategies, we will describe the plots in Fig. 4, showcasing network performance post-failure



(a) Random node failures - Repair order: Betweenness

(b) Random node failures - Repair order: Flow centrality

(c) Random node failures - With re-optimization

(d) Random node failures - Combined: Flow centrality repair

(e) Targeted node failures - Repair order: Betweenness

(f) Targeted node failures - Repair order: Flow centrality

(g) Targeted node failures - With re-optimization

(h) Targeted node failures - Combined: Flow centrality repair

Fig. 4: Survivability of network under different node failures and recovery strategies.

for the different strategies. The top row (Fig. 4a - 4d) shows random failures; the bottom row (Fig. 4e - 4h) shows targeted failures based on betweenness centrality. The x-axis denotes elapsed time after failure, while the y-axis represents the percentage of flows not meeting their requirements.

Each plot contains four lines. The blue line shows average performance without a recovery strategy (random repair order). The green dotted line represents the best-case recovery speed for that failure type (best repair order). The red line indicates the slowest recovery, while the purple line portrays the average performance for the corresponding failure type when applying a specific strategy.

C. Survivability for Specific Node Repair Orders

The survivability of the network, under the influence of varying node repair orders as described previously, is depicted in Fig. 4. Fig. 4a portrays the survivability curves corresponding to random node failures, where the purple dashed line highlights the performance improvements realized when nodes are repaired in an order dictated by betweenness centrality.

On the other hand, Fig. 4e indicates the performance trajectory for targeted failures when nodes are repaired based on betweenness centrality, showcasing worse performance than random repair order. The impact of repair order based on flow centrality on network survivability is illustrated in Fig. 4b and Fig. 4f. Across both random and targeted attacks, flow centrality consistently improved performance.

D. Survivability with Re-optimization

For the second approach to enhancing network survivability, we employed a dynamic strategy involving the migration of VNFs in response to changes in network structure. These changes encompassed both instances of node failures and node repairs. The effectiveness of this strategy can be seen in Fig. 4c and Fig. 4g. The purple curve in these plots, representing cases of failure with immediate network re-optimization, shows a clear trend of improved expected performance across both failure types. This was the case for all failure types in both topologies.

E. Survivability with Combined Strategy

Fig. 2d and 2h display the results of combining both strategies: re-optimization upon network change and specifying the repair order. This combined strategy resulted in improved performance across most scenarios. When the repair order was detrimental as an individual strategy (e.g., Fig. 4e), it also under-performed when paired with re-optimization. When flow centrality, which proved beneficial in all individual cases, was used for repair order, the combined strategy further improved performance in all scenarios.

Fig. 5 reveals an important case: In one of the topologies, employing the combined strategy with closeness centrality for node repair order performed worse than applying no strategy at all (random repair order), even in the best case. Yet, it's worth noting that the combined strategy led to a reduced variance in the estimated performance compared to not specifying the repair order.

In one of the topologies, all implemented strategies, regardless of the node failure type, led to performance improvements, with varying degrees of enhancement ranging from minor to significant. For the second topology, the results displayed a broad spectrum of performance, with some strategies enhancing survivability, while others escalated the impact of the failure.

Tab. I: Performance of Various Strategies Compared to No-strategy Method

Strategy	**Failure Type**	**Performance**	**Comparison to No-strategy**
Betweenness repair	Random failures Targeted failures	Improved Worse	- Close to worst case
Closeness repair	Random failures Targeted failures	Improved Worse	- Close to worst case
VNF repair	Random failures Targeted failures	Improved Improved	- Close to best case
Re-optimization	Random failures Targeted failures	Improved Improved	- Lower variance
Combined (Betweenness repair)	Random failures Targeted failures	Improved Improved	- Worse than re-optimization, but lower variance
Combined (Closeness repair)	Random failures Targeted failures	Improved Worse	- Even best case has worse performance
Combined (VNF repair)	Random failures Targeted failures	Improved Improved	- -

Tab. I provides a summarized overview of the performance outcomes for each combination of strategy and node failure type, as compared to the no-strategy method (i.e., random repair order with no re-optimization). The comments column provides additional details on the performance.

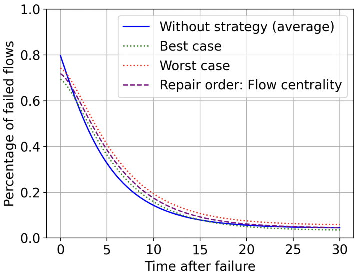


Fig. 5: Example of worse performance

VI. Discussion

Several strategies have been investigated to enhance the performance of NFV networks during and following network failures, specifically (multiple concurrent) random node failures and targeted node (cyber) attacks. We evaluated the efficacy of these strategies concerning the survivability of the network flows, including prioritizing node repair order based on different metrics (betweenness centrality, closeness centrality, and flow centrality), re-optimizing the network after failures and repairs, and combining these approaches. Our results underscore the diverse outcomes achievable depending on the selected strategy and the nature of network failures, offering valuable insights for network operators and researchers alike.

Our results are grounded in a quantified analysis of survivability, studying how network performance, in terms of survivability, changes over time after a failure. This temporal dimension of the analysis is key to comparing different recovery strategies, as it allows us to measure how fast each strategy can bring the network back to its normal functioning state. At the same time, our study also facilitates the possibility of using a spatial dimension within the same survivability quantification framework. This alternative approach would depend on the specific goals of the assessment. The spatial perspective would let us compare recovery strategies by evaluating the amount of network recovery they can achieve within a given time frame. Given our findings, we believe such an approach would lead to similar conclusions.

A. Summary of Findings

Our experiments yield several key conclusions. First, network topology significantly influences the network's resilience to failures and the design of node repair strategies. This insight may also be utilized by an attacker (sabotage or cyber attack). Second, the proposed *flow centrality* is a more reliable heuristic for determining the order of node repair than the betweenness and closeness centrality measures for an NFV network. We found that the primary cause of flows failing to meet their requirements was the failure of nodes hosting these VNFs. These nodes should be prioritized in network design and repair strategies, and even considered to be taken into account in VNF deployment. On the other hand, betweenness and closeness yielded variable results, dependent on the failure and attack type, and on the network topology. Third, re-optimizing in response to structural changes in the network consistently enhanced performance, irrespective of attack type or network topology. Finally, as Fig. 5 demonstrates, there can be instances where seemingly optimal strategies perform worse than doing nothing, underlining the complex and highly context-dependent nature of network survivability.

B. Practical Implications

This study presents novel perspectives on the survivability of NFV networks under different node failure conditions, which can be utilized to increase network survivability. While previous research, such as the NFV-RA algorithm study [10], has made strides towards survivability-focused solutions, our work explores the impact of novel survivability enhancement strategies on a delay-focused NFV-RA algorithm.

We discovered that nodes hosting VNFs play a pivotal role in network performance, with their failure identified as a primary cause of service degradation. These findings have important practical implications for network design, repair strategies, and VNF deployment. In particular, these nodes should be prioritized for protection, given their critical role. However, this also suggests a potential vulnerability: if attackers gain insights into the topology of the network, they can target these critical nodes to inflict maximum damage. Therefore, the defense strategies for these nodes should be designed with this risk in mind.

Our study underscores that no one-size-fits-all strategy exists for survivability assurance, given the diverse outcomes observed across different strategies and failure scenarios. Instead, it offers key insights that can inform more survivable and adaptive network designs. We found that flow centrality proved to be a more reliable heuristic for node repair order determination, outperforming traditional metrics such as betweenness and closeness centrality. Utilizing this insight can enhance recovery strategy effectiveness.

One noteworthy insight concerns networks aiming to deliver high-performance services. While prioritizing speed and delay in NFV-RA algorithms, these networks may inadvertently compromise their survivability as a trade-off. Our study offers strategies to address this trade-off, emphasizing the value of incorporating survivability assessments in the design and operation of high-performance networks. This approach ensures a survivable service delivery even in the face of undesired events and failures.

C. Limitations and Directions for Future Work

Our study offers valuable insights into service degradation caused in NFV networks during node outages and also outlines potential future research directions. We effectively identified patterns by examining various failure types and network topologies. However, further analysis of attack types could enhance the realism of future studies, an area we plan to explore.

The knowledge an attacker possesses about the network emerged as an important aspect of our study. With information on network topology, attackers can infer strategies using betweenness and closeness centrality. However, knowing how the topology is considered in NFV-RA allows an attacker to exploit flow centrality, causing further network damage. Investigating this security vulnerability could inform a more resilient NFV network design.

Regarding the re-optimization strategy, our study acknowledged its benefits but overlooked practical VNF migration considerations. As VNF migration involves costs, incorporating these into future studies can improve their realism. Combining existing research on seamless VNF migration with our findings could yield beneficial insights into managing NFV networks under various failure scenarios.

VII. Concluding Remarks

The investigations in this paper demonstrate the importance of network topology, flow assignments, and VNF deployments on network survivability after both random and targeted failures. The effect of node failures on survivability varied significantly across the two topologies considered. The impact of a failure and the efficiency of a recovery strategy is (significantly) dependent on the network's specific structural characteristics. The results also show potential benefits of network re-optimization and repair order strategies, particularly those that leverage an NFV network specific *flow centrality* metric. While all strategies led to performance improvements for one topology, the results for the other topology were more variable. Some strategies led to minor or significant enhancements, but others appeared to even amplify the performance decline following a failure, performing even worse than a scenario with no defined strategy (i.e., random priority for node repair). These findings highlight the need for careful selection of repair and recovery strategies, emphasizing that a sub-optimal strategy could inadvertently decrease network resilience rather than improve it. Future research should focus on developing more sophisticated and context-aware repair and recovery strategies to optimize network survivability in NFV environments.

References

[1] 3GPP, "Release 15 description; summary of Rel-15 work items," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 21.915, 10 2019. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/21_series/21.915/

[2] M. He, A. M. Alba, A. Basta, A. Blenk, and W. Kellerer, "Flexibility in softwarized networks: Classifications and research challenges," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2600–2636, 2019.

[3] Y. T. Woldeyohannes, A. Mohammadkhan, K. K. Ramakrishnan, and Y. Jiang, "ClusPR: Balancing multiple objectives at scale for NFV resource allocation," *IEEE Transactions on Network and Service Management*, vol. 15, no. 4, pp. 1307–1321, 2018.

[4] J. Gil Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Transactions on Network and Service Management*, vol. 13, no. 3, pp. 518–532, 2016.

[5] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Network*, vol. 28, no. 6, pp. 18–26, 2014.

[6] P. E. Heegaard and K. S. Trivedi, "Network Survivability Modeling," *Computer Networks*, vol. 53, no. 8, pp. 1215–1234, Jun. 2009.

[7] L. Xie, P. E. Heegaard, and Y. Jiang, "Network survivability under disaster propagation: Modeling and analysis," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 4730–4735.

[8] M. Gajić, M. Furdek, and P. Heegaard, "A framework for spatial and temporal evaluation of network disaster recovery," in *2020 32nd International Teletraffic Congress (ITC 32)*. IEEE, 2020, pp. 37–45.

[9] M. R. Rahman and R. Boutaba, "SVNE: Survivable virtual network embedding algorithms for network virtualization," *IEEE Transactions on Network and Service Management*, vol. 10, no. 2, pp. 105–118, 2013.

[10] M. T. Beck, J. F. Botero, and K. Samelin, "Resilient allocation of service function chains," in *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, 2016, pp. 128–133.

[11] "Rocketfuel: An ISP Topology Mapping Engine," 2002. [Online]. Available: https://research.cs.washington.edu/networking/rocketfuel/

Appendix B
Literature Survey - IMT4203

A Survey of 5G Network Slicing As An Element Of Critical Infrastructures

1st Trond Vatten
Department of Information Security and Communication Technology, NTNU, Trondheim, Norway
trond.vatten@ntnu.no

***Abstract*—In the increasingly interconnected world of future mobile networks (5G and beyond), critical infrastructures (CIs) face a growing demand for increased security. At the same time, these networks offer numerous opportunities for improving such infrastructures through technologies such as network slicing. The extreme performance guarantees provided by 5G and beyond networks enable improvements such as increased automation in transportation, smart energy grids, and ultra-reliable emergency networks. Network slicing is viewed as a key enabling technology in these networks, enabling efficient, secure, and reliable connectivity for multiple verticals on a single network. This survey aims to identify the service requirements of critical infrastructures and explore how network slicing can fulfill these requirements. It will then examine the state of the art for the underlying technologies, before examining real-world use cases and identifying remaining challenges and future research directions.**

***Index Terms*—5G, network slicing, critical infrastructure, critical services, URLLC**

I. Introduction

The increasing demands of critical infrastructures (CIs) have led to the need for more diverse and specialized communication systems. In the past, dedicated infrastructure has been used to meet these demands, but these systems can be expensive, slow to deploy, and complex to manage as the network expands. To address these issues, 3GPP defined the 5G standard with a focus on dividing different market segments, including one with specifically strict requirements for latency and availability, called Ultra Reliable Low Latency Communication (URRLC) [3].

The implementation of network slicing has been identified as a key enabling technology in the pursuit of improving the security and reliability of critical infrastructures in future mobile networks, such as 5G and beyond. Network slicing allows for the creation of virtual networks, referred to as ”network slices,” on top of existing physical networks. These network slices can be tailored to meet the specific requirements of various tenants, including those with demands for high levels of latency, reliability, and the ability to handle large volumes of incoming traffic. This virtualization of network infrastructure enables multiple tenants to optimize their connectivity needs without incurring the cost and maintenance of owning their own physical infrastructure.

In the context of CIs, the introduction of network slicing allows for the specification of highly specific network requirements without compromising the demands of other infrastructures. Prior to the adoption of network slicing, it was often necessary to make trade-offs in the performance of one sector in order to meet the security requirements of another. With slicing, each sector is able to customize a dedicated network slice to meet its own unique needs. Additionally, the logical isolation and separation provided by network slicing enables CIs to maintain strict security requirements even when sharing physical infrastructure with less secure networks.

The use of network slicing in CIs has the potential to significantly enhance service functions in sectors such as transportation, healthcare, and energy grids. The highly customizable and automated nature of network slices, which offer performance guarantees, enable advancements in areas such as automated transportation, remote healthcare services, and real-time coordination between machines and control systems in industrial settings. Network slicing offers a promising solution for optimizing the connectivity and performance of CIs in the interconnected world of future mobile networks.

The purpose of this literature survey is to explore the use of network slicing in critical infrastructures (CIs) as outlined in the 3GPP’s 5G specification [3]. The survey aims to identify the motivations for implementing network slicing in CIs, with a focus on the specific service requirements of these infrastructures in the context of communication systems, and to examine how network slicing concepts can be utilized to meet these requirements. It will also provide a comprehensive overview of the current state of the technology and examine real-world use cases in which network slicing has been applied, with a particular emphasis on its application in CIs. Additionally, the survey will address future challenges and identify potential areas for further research in this field.

II. Methodology

This literature review aims to provide a comprehensive overview of the application of network slicing in critical infrastructures (CIs), including its motivations, proposed use cases, real-world examples, and future challenges. To identify relevant literature, advanced searches were conducted on IEEE Xplore, ScienceDirect, and Google Scholar using the keywords ”network slicing” and ”(critical infrastructures OR critical services OR URLLC)”. The search was limited to articles published from 2016 to the present, when 3GPP began work on the first release defining 5G networks. To ensure the thoroughness of the findings, the search was also extended to include related concepts such as ”slicing AND scalability”, ”slice isolation AND critical infrastructure”, ”SDN”, and ”NFV”. The resulting articles were carefully evaluated and curated to select the most relevant papers for inclusion in the review.

To guide the review, a set of research questions is formulated:

- What are the service requirements for critical infrastructures?
- How can network slicing be used to fulfill these requirements?
- What technological capabilities are needed to deploy these network slicing functionalities?
- How is network slicing used for critical infrastructures today?
- What are the remaining challenges to be solved in the realm of network slicing for critical infrastructures?

III. Critical infrastructure requirements

In this section, we will analyze the demands and threats facing critical infrastructures (CIs) in an interconnected world and explore how network slicing can enable the fulfillment of these demands while mitigating potential threats and vulnerabilities. While we will provide an overview of the concept of network slicing and its relevance to CIs, the specifics of the various concepts, current use cases, and challenges associated with the implementation of network slicing will be addressed in Section IV. The focus of this section will be on linking the broad concept of network slicing with the specific demands and requirements of CIs.

A. Requirements and challenges

1) Interconnected world: In today's interconnected world, few infrastructures are completely disconnected from the internet [24]. This trend is expected to continue in the future [6], which means that the number of attack surfaces on CIs is likely to increase. As CIs increasingly incorporate data systems [10], they must be viewed as cyber-physical systems in order to be resilient in the face of cyber threats. To address these challenges, it is necessary for nations to work together to secure CIs, as the compromise of a single country's CIs can have far-reaching consequences. In 2018, the 5G specifications introduced ultra-reliable low-latency communication (URLLC) as a distinct market segment, setting strict performance standards [3]. These specifications represent an important step in ensuring the security of CIs within individual countries, as well as promoting harmonization of security requirements across borders.

The RESISTO project [14], funded by the European Union (EU), is an example of an initiative addressing the issue of increased attack surfaces on telecommunications for CIs. The project aims to develop innovative, comprehensive solutions to enhance the resilience of communication networks in the face of emerging security challenges posed by the rapid deployment and increased utilization of 5G networks, both in critical and non-critical infrastructures. The project emphasizes the need for holistic, scalable, and flexible approaches in order to effectively address the security threats that will arise.

2) Interdependence: Critical infrastructures (CIs) are interconnected and rely on each other for proper functioning. In the realm of telecommunications, these dependencies extend beyond geography and physical connections to also include logical and cyber connections. For instance, a disruption in the electricity sector could have a cascading effect on the efficiency of the banking system, highlighting the importance of extremely high resilience in these systems. To be utilized as a communication network for these services, 5G networks must operate at very high levels of resilience to prevent such cascading effects [9].

3) Isolation: Service isolation is a critical issue in traditional communication networks, as the compromise of one service can potentially affect others that are sharing the same communication channels. This is particularly problematic when different services have diverse security and performance requirements, such as when a banking service prioritizes security over speed. Ensuring that unrelated services and operators are not affected by the compromise of each other is essential for the smooth operation of these networks. This is especially important in the context of CIs, where the disruption of one service can have cascading effects on others. For instance, it would not be desirable for the functioning of a surgical operation to be impaired due to a nearby sports stadium consuming all available network traffic. Ensuring service isolation is therefore critical for the proper operation of CIs.

4) Scaling: Another challenge faced by CIs in communication networks is the ability to adapt to fluctuating demand. As the demand for various infrastructures, both critical and non-critical, can vary, so too does the load on their communication networks. Traditional approaches to meeting these varying demands involve the deployment of separate infrastructure by each vertical, as described in Bektas et al. [8]. However, this is problematic due to long roll-out times and the high cost of dedicated hardware. Communication networks in CIs must maintain high levels of availability and reliability, even in the face of changing demand. This is one of the reasons why the deployment of "automation solutions of the future" such as self-driving cars has been slower than expected. These systems must consistently operate at a reasonable level, regardless of fluctuating network load. Using traditional methods, it is not feasible to achieve the necessary level of flexibility.

B. Advantages of network slicing for critical infrastructures

One of the key features that allow 5G networks to meet their performance guarantees is network slicing, which enables the division of physical networks into smaller, logically separated networks. This is a departure from traditional approaches, which often required the deployment of dedicated hardware to meet specific requirements.

According to 3GPP's earliest 5G release [3], network slices can be either predefined or defined by an operator. The three predefined slices are:

1) *Enhanced mobile broadband (eMBB) slice* for high data rates in large areas.
2) *URLLC slice* for mission critical communication.
3) *Mobile Internet of Things (mIoT) slice* for IoT devices such as sensors.

Although there is some overlap in the requirements for these slices, they have distinct performance characteristics.

The URLLC slice has much higher requirements for latency and reliability, while the mIoT slice must support a large number of smaller devices transmitting control information. All three slices are important for the proper functioning of the 5G network, but the latter two are particularly critical for the operation of CIs. In most cases, slices operating CI will need higher grades of isolation, whether it is incorporated with an URLLC slice or an mIoT slice. This is due to the stringent requirements for service isolation and availability that CIs must adhere to, owing to the vital societal functions they deliver.

As an example of the potential benefits of network slicing, Kurtz et al. [21] discusses the use of communication in driving automation. Using traditional networks, the communication required for tasks such as tracking vehicle location, speed, and direction would not be fast or reliable enough, potentially leading to catastrophic accidents. However, the work also shows that if the specifications for URLLC slices are met, the data could be used to improve real-time road analysis, optimize traffic flow, or even increase vehicle automation. All of these improvements are possible while also including slices for passenger entertainment (eMBB) and smart metering (mIoT), as long as proper slice prioritization is followed to ensure hard service guarantees for the most critical slices. Such technological advancements could help reduce the number of fatal accidents on the roads.

Network slicing also has the potential to address the issue of service isolation in communication networks, as it allows for the creation of completely isolated slices within a single network. This can prevent the interference of one slice from affecting the operation of others, as described in Gonzalez et al. [17]. While there are challenges to be addressed in the implementation of this concept, discussed further in IV, the ability to achieve successful slice isolation is a major step to meeting the stringent security requirements of CIs.

One of the key advantages of 5G network slicing is its ability to quickly scale and dynamically adapt to changes in the network, as noted in Foukas et al. [15], Li et al. [23], and Zhang [34]. By providing a shared physical infrastructure with a virtual layer on top that can be dynamically updated in an automated manner, network slicing can replace the need for multiple dedicated networks. This programmable and adaptable nature is crucial for ensuring the continuous operation of critical infrastructures (CIs), even in the event of compromise or failure.

In summary, the demands and threats faced by critical infrastructures (CIs) in an interconnected world pose significant challenges for communication networks. The interdependence of CIs and the increasing number of attack surfaces due to increased connectivity, as well as the need for high resilience and performance, highlight the importance of ensuring the security and reliability of these networks. Network slicing offers a promising solution to these challenges, enabling the creation of separate, independent slices within a single network and providing a programmable, adaptable infrastructure that can meet the varying demands and requirements of CIs. While there are challenges to be addressed in the implementation of this concept, the ability to achieve the stringent security and performance requirements of CIs through successful network slicing is a major advantage.

IV. State of the art network slicing

In the preceding section, we examined the various requirements and threats faced by critical infrastructures (CIs) and discussed how the concepts of network slicing can help address these challenges. In this section, we delve into the technical aspects and state of the art of these concepts, examining the current state of technological advancement and identifying remaining challenges. We also consider real-world examples and proof-of-concepts to provide a practical context for our analysis.

A. Programmable networks

In section III, we highlighted the interdependence of different infrastructures as a potential risk to the continuous operation of critical services, and mentioned the programmable nature of 5G networks as a potential solution to mitigate these risks. The enabling technologies that make this capability possible are Network Function Virtualization (NFV) and Software Defined Networking (SDN) [25, 30, 33].

NFV involves the virtualization of traditional network functions such as firewalls and routers, which were previously implemented on specialized hardware. With NFV, network functions are implemented as software running on generalized hardware, known as Virtual Network Functions (VNFs). These VNFs are flexible and can be run autonomously from any location. In 2012, the European Telecommunications Standards Institute (ETSI) released a standard for NFVs, which has been further developed through hundreds of publications [2]. VNFs serve as basic building blocks for creating network slices by combining them in a service function chain. Such a chain is programmable and can be changed and dynamically allocated throughout the network by leveraging Software Defined Networking (SDN).

SDN is a new approach to networking that separates the data plane and the control plane [7]. The data plane is responsible for physically forwarding traffic, while the control plane determines how traffic is routed. In SDN, the control plane centralizes routing decisions and has a global view of the entire network, making it easier to identify and respond to changing demands and loads. By combining SDN with NFV, it is possible to route traffic through a network in specific ways to meet varying requirements for bandwidth, end-to-end latency, security, and other factors.

The benefits of NFV and SDN for critical infrastructure can be significant. The programmable nature of 5G networks enabled by NFV and SDN allows for greater flexibility and adaptability in responding to changing demands and risks. This can help to ensure the continuous operation of critical services, even in the face of disruptions to other infrastructures.

Additionally, NFV and SDN can allow for more efficient use of resources and cost savings, as they allow for the dynamic allocation of network resources and the ability to scale up or down as needed. This can be especially important for critical infrastructure, where the cost of downtime can be high.

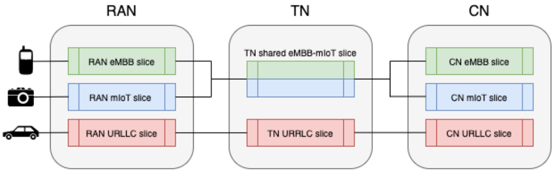


Fig. 1. Example of E2E slice implementation

In Kurtz et al. [21], a novel solution that builds on ETSI's NFV standard and provides network slicing functionality was proposed. The solution was evaluated on a small testbed and validated through physical testing with real-world data from a CI communication scenario. The study demonstrated scalability and provided evidence that network guarantees can be upheld even in the event of partial network overload. As such, it is important to continue building on and improving these types of solutions, evaluating and validating them using real-world data for CIs before deployment in actual networks.

Furthermore, the ability to program and customize network slices for specific needs can improve the security and reliability of critical infrastructure by allowing for the implementation of more tailored security measures and the creation of redundant systems. As an example, in the field of disaster recovery, traditional methods for establishing emergency networks can be slow. By utilizing SDN and NFV, network traffic can be quickly rerouted or even allocated a dedicated slice for communication. Research presented in Gajić, Furdek, and Heegaard [16] has explored the use of network modeling after an undesired event in order to potentially cover blind spots and identify the need for redundancy in networks.

Overall, the use of NFV and SDN in 5G networks provides numerous benefits for critical infrastructure, helping to ensure the continuous operation of critical services and improving the efficiency, cost-effectiveness, and security of these systems.

B. *Slice isolation*

In the literature, there is a range of approaches to addressing slice isolation in 5G networks. These approaches range from considering only the isolation of the Radio Access Network (RAN), to a holistic approach that takes into account the entire end-to-end (E2E) architecture of the network, to simply considering slice isolation as a concept [17]. It is crucial to address slice isolation at the E2E level in order to ensure the proper functioning of CIs. According to 3GPP [4], the E2E architecture of a 5G network consists of the RAN, Core Network (CN), and Transport Network (TN). The RAN is the wireless component responsible for communication with devices such as mobile phones, while the CN is responsible for routing data across the network. The transport network includes all data transmission components, such as cables, switches, and routers. A simplified example of a slice implementation is illustrated in Figure 1. In order to effectively manage slice isolation, a comprehensive approach that considers the entire E2E architecture is necessary.

According to Elayoubi et al. [13], there is a trade-off between resource sharing efficiency and slice isolation when it comes to addressing slice isolation in the Radio Access Network (RAN) of 5G networks. A completely stand-alone RAN slice with its own spectrum and infrastructure would have a high grade of isolation, but also require dedicated infrastructure. On the other hand, if the RAN is completely unaware of slices and slice isolation is only handled by the CN, there is more flexibility in terms of resource sharing, but a lower grade of slice isolation. This latter approach may not be suitable for CIs, which have higher requirements for not being affected by events in other slices.

In Sattar and Matrawy [28], a combination of mathematical modeling and simulation was used to test various levels of slice isolation in the context of a distributed denial of service (DDoS) attack on the 5G core network. The study found that architectures with a high level of slice isolation performed significantly better than a non-isolated slicing architecture in terms of mitigating the effects of the attack. However, the study also noted that complete inter-slice isolation can reduce efficiency in resource utilization. This trade-off is particularly relevant for critical infrastructures, which must protect against the effects of attacks while also efficiently utilizing resources. The study highlighted challenges with large-scale 5G testing capabilities, noting that the experiments were currently too small to be certain.

Contreras and Ordonez-Lucena [11] surveys methods for providing slice isolation in the transport network of 5G networks. The study evaluates four different approaches and discusses the advantages and disadvantages of each, stating that the requirements of the slice customer should be taken into consideration when deciding on an approach. The study also notes that providing indicators for different slicing approaches can give the slice user more flexibility and facilitate compliance with Service Level Agreements (SLAs).

Artificial intelligence (AI) has also been explored as a means of optimizing slice isolation in 5G networks. Vittal and A [32] presents a novel solution for self-optimizing slices based on demands, which utilizes deep learning-based Long Short-Term Memory (LSTM) methods to adapt to varying demands on the network. The framework has demonstrated improvements of up to 35% in serving users with uninterrupted connectivity for ultra-reliable low-latency communication (URLLC) traffic compared to standard slice deployment strategies. Self-learning and flexible optimization techniques such as these could prove particularly useful for critical infrastructures.

C. *Bridging the gap: 5G, AI and Blockchain*

Network slicing is considered a revolutionary technology in the domain of mobile networks. Similarly, AI and blockchain technology have been recognized as disruptive technologies in recent years. These fields provide novel and innovative techniques and strategies, and recent research has started to investigate the intersection of these technologies to address the challenges in network slicing.

In a recent survey, Ssengonzi, Kogeda, and Olwal [31] further addresses the topic of deep reinforcement learning in 5G and beyond networks, with a focus on the application of deep learning in 5G network slicing. The study aims to bridge the gap between these two areas and examine relevant research problems and directions for future research, many of which are relevant to CIs. The survey highlights the increasing complexity introduced by high data rates and diverse network use cases, which can make the task of managing and monitoring data fall under the category of NP-hard problems in the context of optimizing multiple parameters [22]. As such, the potential role of AI techniques in addressing these issues is emphasized, and may even be necessary to meet the requirements of 5G networks and CIs.

Singh et al. [29] propose a framework for improving spectrum sharing efficiency in the radio access network using a combination of blockchain and reinforcement learning. The study emphasizes the dynamic and transparent nature of the framework and uses OMNET++ simulations to demonstrate its potential for energy efficiency and improved transmission success rate. While the work lacks detail on the implementation of blockchain, using emerging technologies to address known problems is an innovative approach that may hold promise for further research.

In Hu et al. [19], blockchain is used to eliminate the need for trust in a multi-tenant slicing architecture. The authors suggest that blockchain can be used for fairness and machine learning-accelerated optimization to address resource allocation problems, resulting in a more decentralized system that may reduce attack vectors from centralized management units.

Similarly, in a study by Dai et al. [12], blockchain is proposed as a promising technology for secure and decentralized resource sharing environments, with AI (specifically, deep reinforcement learning) used to enhance effectiveness. The authors discuss the different types of blockchains (public, private, and consortium) and their potential applications in mobile network architectures. They propose a blockchain and AI-based scheme for wireless networks, based on visions for sixth-generation networks (6G) presented at the Mobile World Congress Americas (MWCA), and validate the framework numerically, demonstrating high system utility.

While the focus on incorporating AI and blockchain in mobile networks is primarily aimed at future generations (6G), the literature suggests that there are benefits to using these technologies. They can create a more distributed and secure network through encryption in blockchain and enhance resource utilization, efficiency, and flexibility through AI. If these early results hold, future mobile networks may be well-suited to meet the requirements of CIs.

V. Case studies

In this section, we will examine various cases in which 5G slicing is used in the context of CIs. Given the critical nature of CIs, it is not feasible to test and validate solutions from different verticals in real-world deployments. As such, there is a need for methods to trial such solutions in a controlled environment.

A. RESPOND-A

The RESPOND-A project [27], funded by the EU, aims to provide modern network capabilities to first responders at emergency sites. The project focuses on user-friendly solutions using 5G technology, augmented reality, autonomous robots, and coordinated unmanned aerial vehicles (UAVs). In a recent evaluation of the RESPOND-A platform [20], the methods used within the RESPOND-A platform were examined within the context of 5G network slicing. The evaluation presents RESPOND-A as a portable communication platform for first responders that utilizes network slicing to meet the needs and requirements of various emergency units. Furthermore, it mapped the different emergency unit's requirements onto different types of network slices. For example, a URLLC slice is required for UAV control, while an mMTC slice is necessary for sensor deployment. The prototype platform was tested on a real testbed to determine optimal performance, and holds promising potential for creating easily accessible applications for first responders in emergencies.

B. 5G PPP, 5Growth and 5G-VINNI

1) 5G PPP: The 5G Infrastructure Public Private Partnership (5G PPP) [26] is a collaboration between the European Union and European industry that was launched in 2018 to develop and innovate future mobile network infrastructures. The project aims to enable Europe to compete in the global technology market by dealing with the biggest challenges of 5G infrastructure.

2) 5Growth: The 5Growth project [5], which is funded by over €14 million from the EU's Horizon 2020 Research and Innovation Programme, is a part of the 5G PPP infrastructure. Its vision is centered on empowering vertical industries to leverage their 5G-enabled technologies, by collaborating with telecommunications industries to create large-scale 5G testing infrastructures.

3) 5G-VINNI: The 5G Verticals Innovation Infrastructure (5G-VINNI) [1] is one of the platforms under the 5Growth umbrella. It provides a comprehensive end-to-end 5G facility that allows vertical industries to test and develop their use cases on a realistic 5G infrastructure. The 5G-VINNI facility consists of eight interconnected sites across Europe and has validated over 30 different vertical use cases.

In a study conducted by the Norwegian defense and Telenor [18], the use of 5G network slicing to meet the demanding service requirements of the military was explored. These requirements included a high degree of isolation, the removal of attack vectors, end-to-end encryption, 99.999% availability, prioritized Quality of Service (QoS), assured throughput, and in some cases, ultra-reliable low-latency communication (URLLC). The slice implementation to meet these requirements was carried out using two slices: one *military slice* based on the URLLC slice type and one *commercial slice* based on the mMTC and eMBB slice types. To ensure complete isolation for the military slice VNFs had to be dedicated, while some VNFs could be shared for the commercial slice. The use of a large-scale facility like 5G-VINNI made it possible to conduct these kinds of trials, which would have otherwise been extremely challenging for an infrastructure such as the military.

The services provided by 5G-VINNI make it easier to develop and validate 5G-enabled technologies, lowering the barriers to entry and promoting technological progress. This is especially important for CIs, which have a critical nature and cannot afford errors in real-world use. Testing slices dedicated to CIs in these kinds of testing sites help to optimize the technology and mitigate potential vulnerabilities.

Projects such as 5G-VINNI demonstrate the value of pre-competition collaboration between tenants in the advancement of 5G technologies. This type of collaboration, which allows for innovation and experimentation across industries, has the potential to accelerate technological progress and ultimately contribute to the development of a more robust critical infrastructure.

VI. Conclusion and future research

In summary, the increased connectivity afforded by 5G and beyond network technologies has introduced a range of new requirements and threats, but also opportunities in CIs. These threats are largely due to the interconnectedness and interdependence of CIs, which creates a larger attack surface when all components are connected to the internet. Additionally, CIs have stringent requirements for service isolation and availability, which needs extra consideration in the face of high network loads or fluctuating network demands.

Subsequently, we examined the role of network slicing in addressing the challenges and realizing the opportunities presented by the incorporation of 5G and beyond technologies in CIs. Through the use of softwarization and virtualization, network slicing enables dynamic reconfiguration of network structures to accommodate changes in demand and requirements. Additionally, the ability to create slices with diverse specifications, including complete isolation from one another, is crucial for the operation of CIs, as it ensures the integrity and robustness of the network by preventing interference from other sectors.

Furthermore, we analyzed the state of the art for the underlying technologies of network slicing, namely SDN and NFV, which enable the creation of highly flexible networks. These technologies facilitate the development of customized slices for various verticals, allowing for the optimization of network resources to meet specific requirements. However, we also addressed the difficulties in achieving slice isolation, particularly in situations where trade-offs for resource sharing efficiency may not be desirable. Various approaches have been proposed to address this optimization issue, including the development of novel algorithms and the use of artificial intelligence frameworks.

Additionally, we examined several studies that utilize emerging technologies, including blockchain and artificial intelligence techniques, to address the challenges in network slicing. While these approaches have shown promise and the literature suggests that they hold significant potential, it appears that they are primarily being researched for future generations of mobile networks such as 6G and beyond. Nonetheless, the incorporation of these technologies into network slicing for CIs warrants further investigation due to their potential to enhance the transparency, security, and efficiency of the network.

Finally, we examined various case studies that utilized network slicing in their solutions, with a particular focus on its application in critical infrastructures. RESPOND-A demonstrated promising progress in the development of an easy-to-use platform for first responders in emergencies, though it is currently in the pilot stage. We also examined the 5G PPP and 5G-VINNI initiatives, which are heavily funded by the European Union and have established large-scale trial facilities for 5G networks across Europe. Their work has proven to be valuable for trialing and validating 5G use cases for over 30 different verticals, and will likely be instrumental in the innovation, testing, and validation of new solutions for CIs.

A. Future research

There are several areas in which further research is necessary to fully realize the potential of network slicing in meeting the requirements of CIs. These include improving resource allocation efficiency with trade-offs, ensuring proper slice isolation, and enabling the dynamic reconfiguration of networks in changing environments. While the use of slicing-based networks for the most critical infrastructures may not yet be advisable due to the complexity of the attack and failure surface, the technology is mature enough to support the development of novel solutions for improving smaller but still critical services. One example of such a solution is RESPOND-A for emergency responders. These efforts can be pursued concurrently with ongoing research thanks to the modularity provided by SDN and NFV. Ultimately, the advancement of network slicing in CIs will require a combination of technological innovation and careful consideration of the unique requirements and challenges of these environments.

References

[1] 5G Verticals Innovation Infrastructure (5G-VINNI). *5G-VINNI*. 2020. URL: https://www.5g-vinni.eu/ (visited on 01/05/2023).

[2] European Telecommunications Standards Institute (ETSI). *Network Functions Virtualisation (NFV)*. 2012. URL: https://www.etsi.org/technologies/nfv (visited on 01/05/2023).

[3] 3GPP. *Release 15 Description; Summary of Rel-15 Work Items*. Technical Specification (TS) 21.915. Version 15.0.0. 3rd Generation Partnership Project (3GPP), Oct. 2019. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3389.

[4] 3GPP. *System architecture for the 5G System (5GS)*. Technical Specification (TS) 23.501. Version 15.0.0. 3rd Generation Partnership Project (3GPP), Dec. 2022. URL: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144.

[5] 5Growth. *Opening 5G to European vertical industries*. 2020. URL: https://5growth.eu/ (visited on 01/05/2023).

[6] NGMN Alliance. "5G white paper". In: *Next generation mobile networks, white paper* 1.2015 (2015).

[7] Alcardo Alex Barakabitze et al. “5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges”. In: *Computer Networks* 167 (2020), p. 106984.

[8] Caner Bektas et al. “Reliable Software-Defined RAN Network Slicing for Mission-Critical 5G Communication Networks”. In: *2019 IEEE Globecom Workshops (GC Wkshps)*. 2019, pp. 1–6. DOI: 10.1109/GCWkshps45667.2019.9024677.

[9] Maria Belesioti et al. “A New Security Approach in Telecom Infrastructures: The RESISTO Concept”. In: *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*. 2019, pp. 212–218. DOI: 10.1109/DCOSS.2019.00056.

[10] Maria Belesioti et al. “Security and Resilience in Critical Infrastructures”. In: *Technology Development for Security Practitioners*. Springer, 2021, pp. 317–333.

[11] Luis M. Contreras and Jose Ordonez-Lucena. “On Slice Isolation Options in the Transport Network and Associated Feasibility Indicators”. In: *2021 IEEE 7th International Conference on Network Softwarization (NetSoft)*. 2021, pp. 201–205. DOI: 10.1109/NetSoft51509.2021.9492546.

[12] Yueyue Dai et al. “Blockchain and Deep Reinforcement Learning Empowered Intelligent 5G Beyond”. In: *IEEE Network* 33.3 (2019), pp. 10–17. DOI: 10.1109/MNET.2019.1800376.

[13] Salah Eddine Elayoubi et al. “5G RAN Slicing for Verticals: Enablers and Challenges”. In: *IEEE Communications Magazine* 57.1 (2019), pp. 28–34. DOI: 10.1109/MCOM.2018.1701319.

[14] RESIlience enhancement and risk control platform for communication infraSTructure Operators (RESISTO). *RESISTO*. 2018. URL: https://www.resistoproject.eu/ (visited on 01/05/2023).

[15] Xenofon Foukas et al. “Network Slicing in 5G: Survey and Challenges”. In: *IEEE Communications Magazine* 55.5 (2017), pp. 94–100. DOI: 10.1109/MCOM.2017.1600951.

[16] Marija Gajić, Marija Furdek, and Poul Heegaard. “A Framework for Spatial and Temporal Evaluation of Network Disaster Recovery”. In: *2020 32nd International Teletraffic Congress (ITC 32)*. 2020, pp. 37–45. DOI: 10.1109/ITC3249928.2020.00013.

[17] Andres J. Gonzalez et al. “The Isolation Concept in the 5G Network Slicing”. In: *2020 European Conference on Networks and Communications (EuCNC)*. 2020, pp. 12–16. DOI: 10.1109/EuCNC48522.2020.9200939.

[18] Pål Grønsund et al. “5G Service and Slice Implementation for a Military Use Case”. In: *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*. 2020, pp. 1–6. DOI: 10.1109/ICCWorkshops49005.2020.9145236.

[19] Qiwei Hu et al. “Blockchain Enabled Federated Slicing for 5G Networks with AI Accelerated Optimization”. In: *IEEE Network* 34.6 (2020), pp. 46–52. DOI: 10.1109/MNET.021.1900653.

[20] Michail-Alexandros Kourtis et al. “5G Slicing for Emergency Communications”. In: *2021 Eighth International Conference on Software Defined Systems (SDS)*. 2021, pp. 1–6. DOI: 10.1109/SDS54264.2021.9732142.

[21] Fabian Kurtz et al. “Network slicing for critical communications in shared 5G infrastructures-an empirical evaluation”. In: *2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft)*. IEEE. 2018, pp. 393–399.

[22] Mathieu Leconte et al. “A Resource Allocation Framework for Network Slicing”. In: *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 2018, pp. 2177–2185. DOI: 10.1109/INFOCOM.2018.8486303.

[23] Xin Li et al. “Network Slicing for 5G: Challenges and Opportunities”. In: *IEEE Internet Computing* 21.5 (2017), pp. 20–27. DOI: 10.1109/MIC.2017.3481355.

[24] IF Mikhalevich and VA Trapeznikov. “Critical infrastructure security: alignment of views”. In: *2019 Systems of Signals Generating and Processing in the Field of on Board Communications*. IEEE. 2019, pp. 1–5.

[25] Jose Ordonez-Lucena et al. “Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges”. In: *IEEE Communications Magazine* 55.5 (2017), pp. 80–87. DOI: 10.1109/MCOM.2017.1600935.

[26] 5G Infrastructure Public Private Partnership (5G PPP). *5G PPP*. 2018. URL: https://5g-ppp.eu/ (visited on 01/05/2023).

[27] RESPOND-A. *RESPOND-A Project*. 2020. URL: https://respond-a-project.eu/ (visited on 01/05/2023).

[28] Danish Sattar and Ashraf Matrawy. “Towards Secure Slicing: Using Slice Isolation to Mitigate DDoS Attacks on 5G Core Network Slices”. In: *2019 IEEE Conference on Communications and Network Security (CNS)*. 2019, pp. 82–90. DOI: 10.1109/CNS.2019.8802852.

[29] Saurabh Singh et al. “BENS- B5G: Blockchain-Enabled Network Slicing in 5G and Beyond-5G (B5G) Networks”. In: *Sensors* 22.16 (2022), p. 6068.

[30] Thomas Soenen et al. “Demystifying network slicing: From theory to practice”. In: *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. 2017, pp. 1115–1120. DOI: 10.23919/INM.2017.7987450.

[31] Charles Ssengonzi, Okuthe P Kogeda, and Thomas O Olwal. “A survey of deep reinforcement learning application in 5G and beyond network slicing and virtualization”. In: *Array* (2022), p. 100142.

[32] Shwetha Vittal and Antony Franklin A. “Self Optimizing Network Slicing in 5G for Slice Isolation and High Availability”. In: *2021 17th International Conference on Network and Service Management (CNSM)*. 2021, pp. 125–131. DOI: 10.23919/CNSM52442.2021.9615546.

[33] Faqir Zarrar Yousaf et al. “NFV and SDN—Key Technology Enablers for 5G Networks”. In: *IEEE Journal on Selected Areas in Communications* 35.11 (2017), pp. 2468–2478. DOI: 10.1109/JSAC.2017.2760418.

[34] Shunliang Zhang. “An Overview of Network Slicing for 5G”. In: *IEEE Wireless Communications* 26.3 (2019), pp. 111–117. DOI: 10.1109/MWC.2019.1800234.