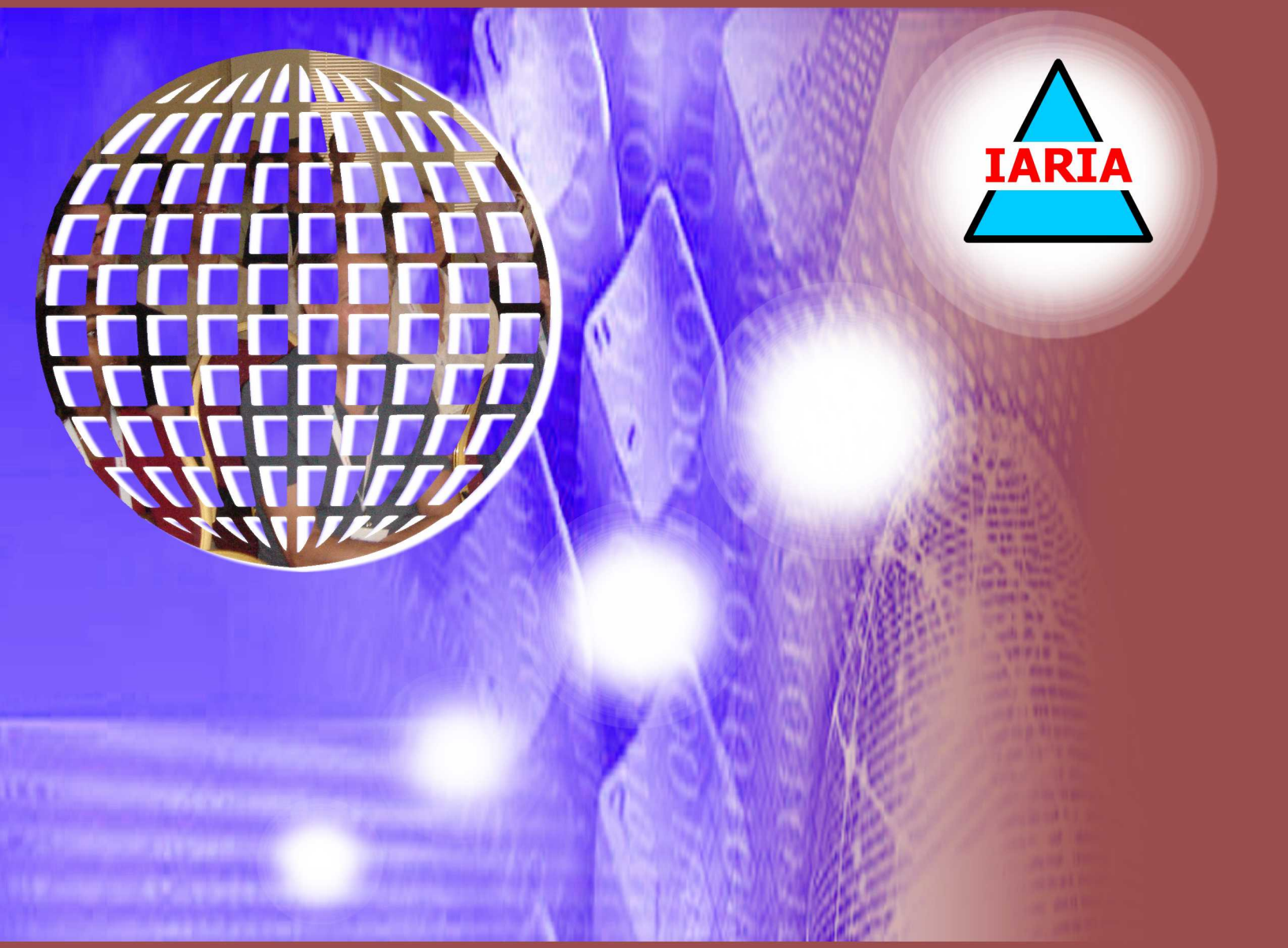
International Journal on Advances in Networks and Services



2024 vol. 17 nr. 3&4

The *International Journal on Advances in Networks and Services* is published by IARIA.
ISSN: 1942-2644
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Networks and Services
Volume 17, Number 3 & 4, 2024

International Journal on Advances in Networks and Services
Volume 17, Numbers 3 & 4, 2024

CONTENTS

Yusheng Ji, Information Systems Architecture Research Division, National Institute of Informatics, Japan
Tokumi Yokohira, Faculty of Interdisciplinary Science and Engineering in Health Systems, Okayama University, Japan
Tutomu Murase, Graduate School of Information Science, Nagoya University, Japan

Simulation Based Energy Efficiency Analysis and Evaluation of DUDe 5G Networks

Chrysostomos-Athanasios Katsigiannis
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
email: up1072490@upnet.gr

Apostolos Gkamas
Department of Chemistry
University of Ioannina
Ioannina, Greece
email: gkamas@uoi.gr

Konstantinos Tsachrelias
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
email: up1096511@upatras.gr

Christos Bouras
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
email: bouras@upatras.gr

Vasileios Kokkinos
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
email: kokkinos@upatras.gr

Philippos Pouyioutas
Computer Science Department
University of Nicosia
Nicosia, Cyprus
email: pouyioutas.p@unic.ac.cy

***Abstract* - Meeting the escalating demands for data traffic in fifth-generation networks and beyond requires efficient solutions like Heterogeneous Networks, which enhance spectral and energy efficiency by deploying small cells close to users. Traditional Downlink and Uplink Coupling often limits uplink efficiency due to power imbalances across base stations. Downlink and Uplink Decoupling addresses this by allowing separate access points for uplink and downlink, optimizing user association and energy use. This research expands upon previous conference work by introducing a new scenario that evaluates Downlink and Uplink Decoupling's performance at a 25 decibel milliwatts user equipment power setting, along with an additional experiment for 1500 user devices in the 20 and 30 decibel milliwatts scenarios. The extended analysis offers deeper insights into the energy efficiency and resource allocation of Downlink and Uplink Decoupling under various network conditions, confirming its suitability for scalable, efficient fifth-generation networks.**

Keywords - Downlink and uplink decoupling; Downlink and uplink coupling; Energy efficiency; Heterogeneous networks; Resource allocation evaluation.

I. INTRODUCTION

Modern 5G Networks offer great benefits compared to the 4G Long-Term Evolution (LTE) technology, with some of them being high speed, low latency and increased bandwidth. However, the volume of mobile traffic and the number of connected devices is predicted to increase significantly in the 5G era, which will lead to inevitable implications regarding the resource allocation and the total throughput of the networks. An important issue of modern 5G Networks is the energy efficiency evaluation [1]. The 4G technologies had already achieved extreme densification by utilizing Small Base Stations (BSs) throughout the network, leading to the modern Heterogeneous Networks (HetNets) [2]-[5].

In 4G HetNets the User Equipment (UE) devices were associated with the same BS for both Downlink (DL) and Uplink (UL) signals, resulting in the method known as Downlink/Uplink Coupling (DUCo) (see Figure 1). This access scheme, though, has a major drawback. The existence of major inequalities between the transmit power among high powered Macro BS and low powered Small BSs, results in suboptimal BS association and thus in performance degradation, affecting the UL. A fine solution to this problem is the decoupling of DL and UL signals in the current HetNets, commonly referred to as Downlink/Uplink Decoupling (DUDe), where the UE is connected to the optimal Macro BS for the DL and the optimal Small BS (Micro-Pico-Femto) for the UL. The UL and DL are treated as separate network entities and a UE can connect to different serving nodes in the UL and DL, resulting in improved user/BS association and improved resource allocation.

BS association in cellular networks has traditionally been based only on the received signal strength, despite the fact that transmit power and interference levels vary significantly. This approach was adequate in homogeneous networks with Macro BSs that all have similar transmit power levels. However, with the development of HetNet, there is a significant difference between the transmission power of different types of BSs, as stated above, making this approach extremely inefficient.

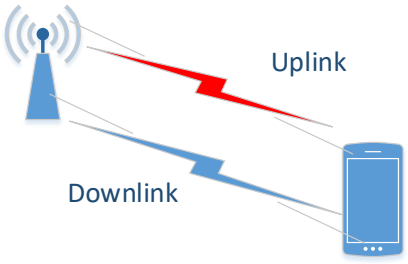


Figure 1. DUCo Example.

The motivation of this work is the improvement of energy efficiency in 5G networks. Energy efficiency is crucial for the success of 5G networks, as these networks will require a significant increase in energy consumption compared to their

predecessors. With the proliferation of 5G-enabled UE and the explosion of data traffic, the demand for energy-intensive infrastructure components, such as BSs and data centers, will rise dramatically. By optimizing network architecture, using low-power components, and implementing intelligent power management strategies, operators can significantly reduce energy usage without sacrificing performance.

DUDe has the potential to significantly improve the energy efficiency of 5G networks. By separating the uplink and downlink channels, operators can dynamically allocate network resources to match the requirements of different applications and services. This results in a more efficient use of resources and reduces the energy consumption of network components, such as BSs and routers. Additionally, decoupling can enable intelligent power management strategies, such as sleep mode for low-traffic devices, further reducing energy consumption.

The main objective of this paper is to validate the findings of previous research by investigating the performance of the system in terms of the number of users and considering different decibel (dBm) values. The paper aims to fill the research gap by conducting a comprehensive analysis that incorporates various factors and parameters. By doing so, the paper intends to provide a deeper understanding of how the system performs in real-world conditions and assess its suitability for different deployment scenarios.

The rest of this paper is organized as follows. Section II presents similar research work. Section III presents the DUDe technology and its key features. Section IV presents the mathematical model used in our simulation environment. Section V presents the analysis of the algorithms compared in our scenarios. Section VI presents the simulation environment used for the implementation of our experiments. Section VII presents the results of the simulation and provides a detailed analysis of the findings. Finally, Section VIII concludes the paper and provides suggestions for future research.

II. State of the Art

DUDe has been researched by various studies. In one of these studies, researchers consider the resource allocation problem in LTE-U networks using DUDe, formulating the problem as a game theoretic model incorporating UE association, spectrum allocation and load balancing, for which they propose a decentralized expected Q-learning algorithm to solve [6]. Another paper proposes an UL and DL Supplementary UL (SUL) decoupling technology and an UL enhancement technology to coordinate New Radio Time Division Duplexing (NR TDD) and New Radio Frequency Division Duplexing (NR FDD. Lastly, several researchers study the concept of DUDe where DL BS association is based on received power DL, while UL is based on path loss [7].

However, another paper proposed a DUDe model where Macro-BS selection for DL is based on received power (as usual), but Micro, Pico and femto-BSs selection for UL is not based solely on path loss (link quality), but on a combination of parameters such as: link quality, BS load and BS backhaul capacity [8]. Authors in [9] focus on how to use DUDe technology improves the distribution of network resources based on UE distribution. The study found that by considering the capacity limitations of each type of BS, the DUDe technology results in a more even distribution of UEs within the network.

Paper [10] highlights the limitations of traditional resource allocation techniques in efficiently managing bandwidth within 5G networks. DUDe technology offers a dynamic approach by adjusting resource allocation based on UE demand and network conditions, thereby optimizing bandwidth distribution. Experimental results from this study have shown that DUDe effectively balances UE equipment distribution across BS, reducing the bandwidth usage of Macro BSs and consequently enhancing the Quality of Service (QoS) for UEs. These findings underscore the potential of DUDe in Macro BS offloading, providing valuable insights for network operators and researchers aiming to develop advanced resource allocation strategies in 5G networks.

Paper [11] provides an in-depth analysis of how DUDe enhances resource allocation by introducing an additional lower frequency signal on the uplink, complementing the existing signal. This approach effectively rebalances the uplink/downlink disparity at the BS edge, improving coverage and network capacity. Through extensive literature review and industry trend analysis, the study examines the benefits and challenges of DUDe, focusing on its impact on network performance, UE experience, and future advancements. Utilizing a simulation-based methodology, the research evaluates DUDe's effectiveness in terms of coverage, capacity, latency, and energy efficiency. The findings demonstrate that DUDe significantly enhances network performance, particularly in environments with high data transmission demands, and reduces outage rates in networks with high minimal throughput requirements. These insights are crucial for researchers and network operators aiming to implement efficient resource allocation strategies to optimize 5G network performance.

Authors in [12] address the high energy consumption associated with mmWave Small Cell Base Stations (SCBSs), which are integral to 5G networks. Dynamic TDD is employed to improve SCBS throughput by allowing flexible TDD time fractions. Given the coexistence of mmWave SCBSs with microwave Macro Base Stations (MBSs), DUDe is proposed to mitigate transmit power imbalances. This research formulates the joint optimization of energy efficiency and resource allocation for dynamic TDD with DUDe, analyzing the trade-off between throughput and energy efficiency using a generalized a-fair scheduler. The findings indicate a significant throughput gain of 28.4% with minimal impact on energy efficiency for dynamic TDD systems with DUDe compared to static TDD systems. These results demonstrate that dynamic TDD with DUDe improves throughput (52.45%) with only a marginal decrease (2.3%) in energy efficiency compared to static TDD without DUDe. These insights are crucial for developing efficient resource allocation strategies that balance throughput and energy efficiency in 5G networks.

In paper [13], the authors address the challenges of uplink power control in HetNets using the DUDE mode. They identify that traditional BS association rules, which are based on maximum downlink received power, are inadequate for current heterogeneous cellular networks. To mitigate co-channel interference from Small UE (SUE) and DUDE UE to Macro UE (MUE), the authors extend three existing power control schemes from homogeneous networks to HetNets. They compare the convergence and optimality of these schemes through

theoretical analysis. Additionally, the authors propose an Improved Distributed Power Control (IDPC) scheme. Simulation results demonstrate that the IDPC scheme significantly enhances system performance, particularly in scenarios with a high number of UEs experiencing severe mutual interference. The findings confirm that IDPC is more suitable for uplink power control in DUDE mode HetNets, improving QoS and average signal-to-interference-plus-noise ratio (SINR).

Authors in paper [14] explore the optimization of joint uplink and downlink scheduling and resource allocation in a millimeter-Wave (mmWave)-based cellular network using dynamic Time Division Duplexing (TDD). Recognizing the potential of mmWave SCBSs to enhance data rates and network capacity in 5G networks, the study also addresses the transmit power imbalance between Macro BS and SCBSs through DUDe. The authors formulate the scheduling and resource allocation problem within a dynamic TDD system as an optimization problem, employing a generalized α-fair scheduler. They derive the dynamic TDD and UE scheduling time fractions for a relaxed version of the problem, showing that the results are a function of the system load. Simulation results validate the derived dynamic TDD results, demonstrating a 17% throughput gain in certain scenarios. The findings indicate that the proposed approach outperforms existing schemes, offering a robust solution for dynamic resource allocation in mmWave-based 5G networks.

III. DUDE ENERGY EFFICIENCY OVERVIEW

DUDe is a complex technique that requires meticulous planning and coordination to be implemented effectively. This method involves assigning separate frequency bands and resources to DL and uplink UL channels, necessitating close collaboration between network operators and device manufacturers to ensure seamless integration and maximize benefits.

One of the most compelling advantages of DUDe is its potential to significantly reduce energy consumption in wireless communication systems. These systems typically consume substantial amounts of energy, which not only increases operational costs but also has adverse environmental effects, contributing to global warming. By decoupling the DL and UL channels, DUDe optimizes energy utilization within the network infrastructure. This process, illustrated in Figure 2, minimizes the energy required to operate the network by reducing interference and improving signal quality. Consequently, the system operates more efficiently, leading to considerable energy savings over time. These savings not only reduce operational costs for network operators but also support environmental sustainability efforts by lowering the carbon footprint associated with wireless communications [15]- [18].

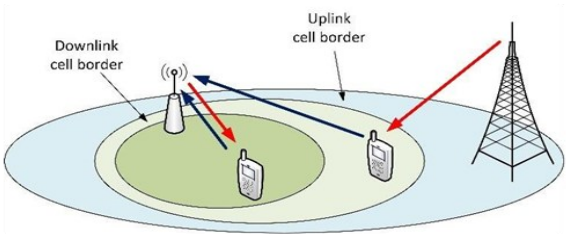


Figure 2. DUDe Example.

In addition to energy efficiency, DUDe offers significant improvements in network performance and reliability. By eliminating interference through the decoupling of DL and UL channels, DUDe enhances overall signal quality and network stability, resulting in a more reliable and robust network. This reliability translates into a superior UE experience, as the reduction in interference improves the quality of service and extends the lifespan of network components by reducing strain on the infrastructure.

Furthermore, DUDe facilitates more flexible and efficient resource allocation. With separate frequency bands and resources for DL and UL channels, network operators can allocate resources dynamically based on current demand and usage patterns. This flexibility helps mitigate the risk of congestion, ensuring that the network can handle high traffic volumes without compromising service quality. During periods of high network demand, efficient resource allocation ensures that UEs receive the expected quality of service. This dynamic resource management is particularly beneficial in urban areas and during peak usage times when network congestion can be a significant issue. By improving resource allocation, DUDe helps maintain high UE satisfaction and operational efficiency.

In conclusion, DUDe is a promising radio resource management technique that offers substantial benefits for both network operators and end- UEs. By decoupling the DL and UL channels, DUDe reduces energy consumption, enhances network performance and reliability, and enables more efficient resource allocation. The implementation of DUDe represents a significant advancement in the evolution of wireless communication systems, aligning operational efficiency with environmental sustainability.

IV. MATHEMATICAL MODEL ANALYSIS

To determine the minimum distance between UE and BS antennas, a mathematical model defined in TR 38.901 Section 7.4.1 is used [19]. The paper does not delve into a detailed analysis of this particular model, so this paper do not extensively scrutinize its equations from (1) to (3) as a result. The model calculates the Path Loss (PL) in different scenarios, such as Line-Of-Sight (LOS) and Non-Line-Of-Sight (NLOS) conditions.

$$PL_{\mathrm{RMa-LOS}} = \begin{cases} PL_1 & 10m \le d_{\mathrm{2D}} \le d_{\mathrm{BP}} \\ PL_2 & d_{\mathrm{BP}} \le d_{\mathrm{2D}} \le 10\mathrm{km} \end{cases} \quad (1)$$

$$PL_1 = 20\,log_{10}(40\pi d_{3D} f_c/3) + min(0.03h^{1.72}, 10)\,log_{10}(d_{3D}) - min(0.044h^{1.72}, 14.77) + 0.002\,log_{10}(h)d_{3D} \quad (2)$$

$$PL_2 = PL_1(d_{BP}) + 40\,log_{10}(d_{3D}/d_{BP}) \quad (3)$$

$$\text{SNR} = \text{Psignal}/\text{Pnoise} \quad (4)$$

The path loss is calculated using equations (1), (2), and (3). Equation (1) defines the path loss in LOS conditions and NLOS conditions based on the distance between the UE and the BS antennas. Equation (2) calculates the path loss based on the three-dimensional distance, carrier frequency, user height, and other parameters. Equation (3) modifies the path loss based on the breakpoint distance and the three-dimensional distance.

Once the minimum distance for each user from different types of BS is determined, the next step is to compute the Signal-to-Noise Ratio (SNR) to find the BS type that provides the best connection. Equation (4) represents the SNR as the ratio of the signal power to the noise power [20].

V. Energy Efficiency Algorithm

Our HetNet includes Macro BS (MB) Small BS (Pico and Micro) and UEs. Consider a set of MBs (M = 1, 2, 3, 4.....| M|), a set of Small BSs (Pico => p = 1, 2, 3, 4.... |P|, Micro => m= 1, 2, 3, 4 |m|) and a set of UEs (U = 1, 2, 3, 4.....|U|). The MBs are placed at high levels to provide continuous uninterrupted coverage to large BSs. In addition, the BSs with the least sensitivity are placed at lower levels within an area, and as a result, the coverage of the NLOS locations is as wide as possible in the entire area, even in the most remote/obstructed points to efficiently serve static users or users who are constantly in motion within the area.

Algorithm 1 Algorithm for 20dBm or 30dBm Transmit Power

```
//Initialize variables
Macro_BSs
Micro_BSs
Pico_BSs
N = total number of UEs
occurrences_for_scenarios
SNR_matrix = zeros(N, occurrences_for_scenarios)
// calculate best SNR for each UE for occurrences_for_scenarios
for i in range(N):
    for j in range(occurrences_for_scenarios):
        // calculate SNR for current occurrences_for_scenarios
        SNR = calculate_SNR(UE_i, occurrences_for_scenarios _j)
        SNR_matrix[i][j] = SNR
    end
end
// calculate standard SNR value for each UE for each
//occurrences_for_scenarios
standard_SNR = zeros(N, occurrences_for_scenarios)
for i in range(N):
    for j in range(occurrences_for_scenarios):
        // calculate standard SNR for current snapshot
        standard_SNR[i][j]=sum(SNR_matrix[i][j])/occurrences_for_sce
        narios
        // calculate transmit power for each UE for each
        //occurrences_for_scenarios
        transmit_power = zeros(N, snapshots)
    end
end
for i in range(N):
    for j in range(occurrences_for_scenarios):
        // calculate transmit power for current snapshot
        transmit_power[i][j] = calculate_transmit_power(UE_i,
        standard_SNR[i][j])
        // build coupled scenario and distribute UEs in the network
        coupled_power = zeros(N)
    end
end
for i in range(N):
    // if transmit power is less than 20 or 25 or 30 dBm, keep value
    if transmit_power[i][-1] < 20  if transmit_power[i][-1] < 25 if
    transmit_power[i][-1] <30:
        coupled_power[i] = transmit_power[i][-1]
        // if transmit power is above 20 dBm, change value to 20 dBm

        // if transmit power is above 25 dBm, change value to 25 dBm

        // if transmit power is above 30 dBm, change value to 30 dBm
    else:
        coupled_power[i] = 20 or 25 or30
        // build decoupled scenario and distribute UEs in the network
        decoupled_power = zeros(N)
end
for i in range(N):
    // calculate transmit power using decoupling technology
    decoupled_power[i] = calculate_decoupled_transmit_power(UE_i,
    standard_SNR)
    // compare energy efficiency between coupled and decoupled
    //scenarios
end
if sum(coupled_power) > sum(decoupled_power):
    output("Decoupling technology is more energy efficient.")
else:
    output("Coupling technology is more energy efficient.")
```

Figure 3. Algorithm 1.

The Energy Efficiency Algorithm (Algorithm 1), depicted in Figure 3, explores two different scenarios for distributing UEs in a network: DUCo and DUDe. The algorithm initializes key variables, including the number of Macro, Micro, and Pico BSs, the total number of UEs, and the number of snapshots used in the simulation.

Initially, for each UE and each snapshot, the algorithm calculates the SNR of each UE in the network, storing these SNR values in a matrix. Subsequently, it computes the standard SNR value for every UE in each snapshot by averaging the SNR values across all snapshots, and these values are stored in a standard SNR matrix. Following this, the algorithm calculates the transmit power for each UE in each snapshot and stores these values in a transmit power matrix. The algorithm then constructs the DUCo and DUDe scenarios by distributing the UEs within the network. For the DUCo scenario, it sets the transmit power of each UE to the last value in the transmit power matrix for that UE unless the value exceeds a threshold (20 or 25 or 30dBm in the simulations). If the value surpasses this threshold, it is capped at the pre-selected limit. This ensures that the power levels remain within acceptable bounds, optimizing energy efficiency and maintaining regulatory compliance. In the DUDe

scenario, the algorithm calculates the transmit power for each UE based on the standard SNR values. It runs the scenarios over numerous snapshots (1000 in our simulations) to mitigate the impact of random variations or uncertainties in the simulation outcomes. This approach ensures a robust comparison between DUCo and DUDe scenarios.

In summary, Algorithm 1 provides a comprehensive evaluation of energy efficiency in HetNet configurations, comparing DUCo and DUDe scenarios. The results validate that DUDe is a more sustainable and efficient method for managing radio resources in 5G networks. Future research could focus on optimizing bandwidth allocation using DUDe, further enhancing energy efficiency, minimizing interference, and maximizing throughput. Integrating advanced technologies like machine learning could also improve dynamic resource management, solidifying DUDe's role in future wireless communication systems.

VI. Simulation Environment

Specifically, a 5G DUDe network is considered, which consists of 2 Macro BSs, 4 Micro BSs, and 8 Pico BSs, each equipped with specific transmit power in dBm. Furthermore, it should be noted that the capacity of the Macro BSs is 2000 users, the capacity of the Micro BSs is 200 users, and the capacity of the Pico BSs is 46 users. This information is crucial in determining the optimal number of users that can be allocated to each type of cell. A total of N number of users are distributed within the network, each with their own transmit power in dBm. The gain from all BS antennas, including bandwidth and noise in the network, is also considered. For the implementation of our model and scenarios, MATLAB [21] was used, due to the fact that the application provides appropriate libraries and, consequently, functions, which make it easy and reliable to create a demanding algorithm like the one above. In addition, Figure 4 depicts the layout of our network, where the two Macro BS are located at the center, surrounded by Small BSs that are distributed around them.

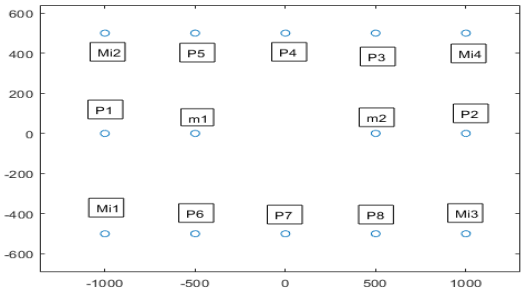


Figure 4. Topology of our network. (m) for Macro (mi) for Micro and (p) for Pico.

It is important to mention that users are randomly located between 1 and 2 meters apart from each other. The connection of the users is done in such a way that, for the DL processes the user will be connected to the Macro BS. During UL processes, the user will connect to the Small BS, which can either be Micro or Pico BS. The selection of the appropriate Small BS is based on the lowest path loss value, in addition to the transmission power.

Once the path loss has been calculated, the SNR is calculated using the variables mentioned in Table I. Utilizing the highest SNR, each user is connected to the best BS choice from the three categories, namely Macro BS, Micro BS, and Pico BS.

The direct result of this is that our model guarantees the non-interruption of the connection and less power consumption since the BS does not consume resources to serve users with great losses.

TABLE I. Simulation Parameters

Parameter	Value
Amount of BSs	Macro BS = 2 Micro BS = 4 Pico BS = 8
Transmit power(dBm)	UE = 20, 30 Macro BS = 45 Micro BS = 33 Pico BS = 24
BS height (m)	Macro height = 30 Micro height = 10 Pico height = 5
Antenna gain (dBi)	Macro BS = 21 Micro BS = 10 Pico BS = 5
Bandwidth (MHz)	20
Environmental parameters	UEs = 200,500,1000,1500,2000 Position = random
Power Noise	Pnoise = -74+10log(Bandwidth(hz))

The purpose of the evaluation is to demonstrate the superior energy efficiency of the DUDe technique compared to the DUCo technique in a 5G network. This goal is achieved by calculating a common SNR value for each type of BS (Macro, Micro, Pico) using the mathematical formula presented in Section III.

Next, the transmission power is calculated for different scenarios involving 200, 500, 1000, 1500 and 2000 UEs for each BS instance. The findings reveal that for the same SNR value, the power consumption of the DUCo technique is significantly higher than that of the DUDe technique.

VII. Results Evaluation and Analysis

Two scenarios, DUDe and DUCo, are implemented with transmit power of 20dBm, 25dBm and 30dBm. As the performed evaluation shows, the DUDe scenario requires less transmission power compared to the DUCo scenario, making a network that uses the DUDe technique in a more energy-efficient and environmentally friendly way.

The limit for UE transmission power (20dBm, 25dBm 30dBm) in our scenarios and in general in mobile telecommunications is set by the Mobile Broadband Standard Partnership Project (3GPP) [22].

Also, in the context of the diagrams provided, initially the average SNR values was determined for each UE across 1000 snapshots. These average SNR values are our target for subsequent calculations. When calculating the required transmission power for each UE to achieve these target SNR values, a threshold was taken into consideration: if the calculated

power exceeds 20dBm or 30dBm, the transmission power was set to those respective limits, regardless of the calculation outcome.

A. Evaluation and analysis of 1st scenario

In this evaluation scenario, DUDe and DUCo was compared in terms of energy efficiency by setting the UE transmission power at 20dBm. The results of the evaluations are displayed in Figures 5 to 9. In the diagrams presented in these figures, the x axis is the average transmission power of the UEs in dBm and the Cumulative Distribution Function (CDF) of the performance metric F(x) axis is the possibility of successful DUDe or DUCo scenario. The implementation of the scenario was successful in meeting the goal of achieving the same result with less transmission power in the DUDe method. This means that our scenario was able to accomplish the desired outcome while using less transmission power, which is a significant achievement in the field of mobile telecommunications.

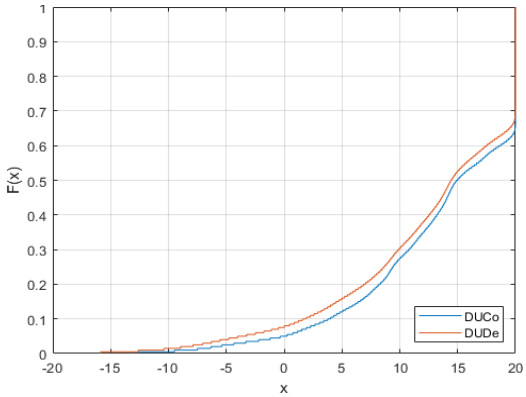


Figure 5. DUDe/DUCo comparison with 20dBm UE limit for N=200.

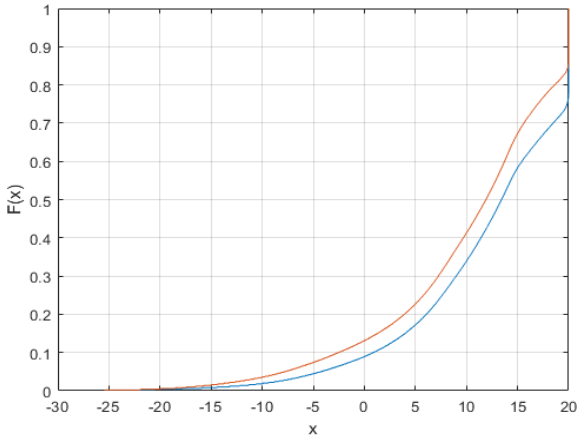


Figure 6. DUDe/DUCo comparison with 20dBm UE limit for N=500.

Based on the provided diagrams, it is evident that the DUDe method exhibits a higher likelihood of establishing successful connections compared to the DUCo method. For instance, at a UE transmit power of 10 dBm, the DUDe method demonstrates a 50% chance of establishing a successful connection, whereas the DUCo method achieves a success rate of less than 40% across all three simulations involving (200, 500, 1000, 1500, and 2000) UEs. These results suggest that DUDe technology consistently outperforms DUCo technology, offering at least 20% more successful connections. This higher success rate implies that DUDe technology provides better reliability and improved connectivity for UEs in wireless communication systems.

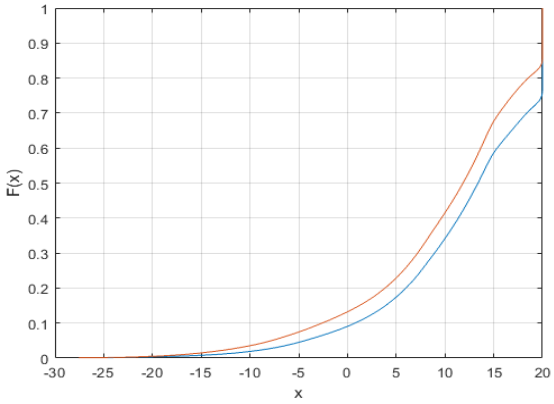


Figure 7. DUDe comparison with 20dBm UE limit for N=1000.

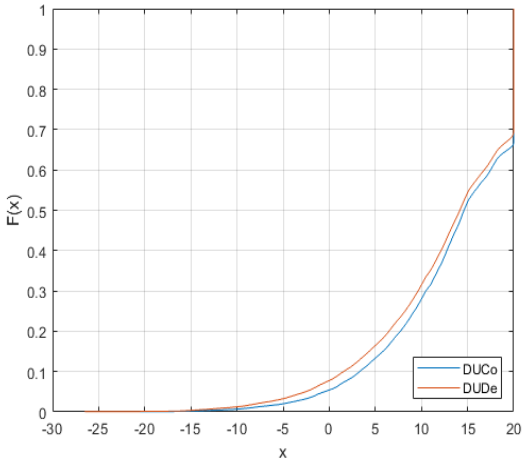


Figure 8. DUDe comparison with 20dBm UE limit for N=1500.

The research includes probability diagrams that illustrate the likelihood of successful connections for both DUDe and DUCo technologies. These diagrams visually support the claim of superior performance by showing the higher probability of successful connections with DUDe technology. As more successful connections result in fewer retransmissions and less energy-intensive signaling processes, DUDe technology can contribute to reducing overall energy consumption compared to DUCo technology. This combination of higher success rates and lower energy consumption underscores the efficiency and preference for DUDe technology.

Similarly, when considering an SNR of 15 dB, the diagrams illustrate that the probability of achieving low consumption with the DUDe method is 80%, whereas the DUCo method reaches only 62%. The difference between the two technologies, approximately 29%, further substantiates the

hypothesis that DUDe technology achieves lower power consumption for the same performance level. These findings provide robust evidence that DUDe is more energy-efficient than DUCo. The significant advantage of DUDe in terms of reduced power consumption highlights its potential for practical implementation, aligning with the objective of developing sustainable and environmentally friendly 5G networks. The research demonstrates that DUDe not only enhances connectivity and reliability but also promotes energy efficiency, making it a preferable choice for modern wireless communication systems.

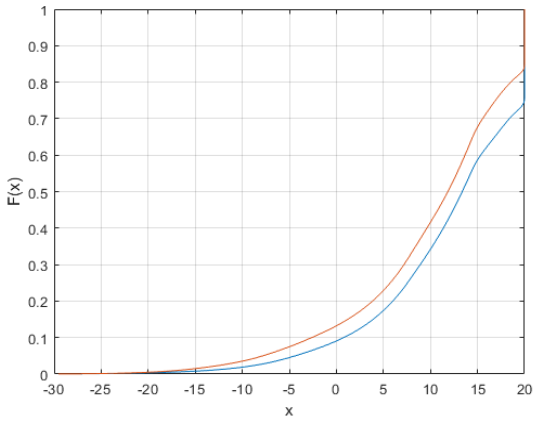


Figure 9. DUDe comparison with 20dBm UE limit for N=2000.

The extended analysis of the diagrams and simulation results clearly indicates that DUDe technology surpasses DUCo technology in both connection success rates and energy efficiency. By minimizing retransmissions and optimizing signaling processes, DUDe reduces overall energy consumption, thereby contributing to more sustainable network operations. The superior performance of DUDe, as evidenced by the probability diagrams and research findings, reinforces its suitability for implementation in 5G networks aimed at achieving higher efficiency and environmental sustainability. Continued research and development in this area will further solidify the advantages of DUDe and support its broader adoption in future wireless communication infrastructures.

B. Evaluation and analysis of 2nd scenario.

In the conducted evaluation, in the second scenario, DUDe and DUCo were meticulously compared to assessing their energy efficiency, with UE transmission power set at 30dBm. The outcomes of these evaluations are vividly illustrated in Figures 10 to 14. In these diagrams, the x-axis represents the average transmission power of the UEs in decibels (dBm), while the F(x) axis denotes the probability of a successful DUDe or DUCo scenario. This second scenario also adheres to a 30dBm limit imposed by 3GPP, which caps the maximum transmission power allowable for signal transmission in this context.

Despite this power limitation, the findings remain consistent with those observed in the first scenario. It was determined that the DUDe method demonstrates greater environmental friendliness compared to the DUCo method when implementing a heterogeneous 5G network. This conclusion is valid even within the constraints of the 30dBm limit. It is noteworthy, however, that the difference in energy consumption between the DUDe and DUCo methods is less pronounced in the second scenario compared to the first. Nonetheless, this does not alter the overall conclusion that the DUDe method is more energy-efficient and exerts a more positive environmental impact.

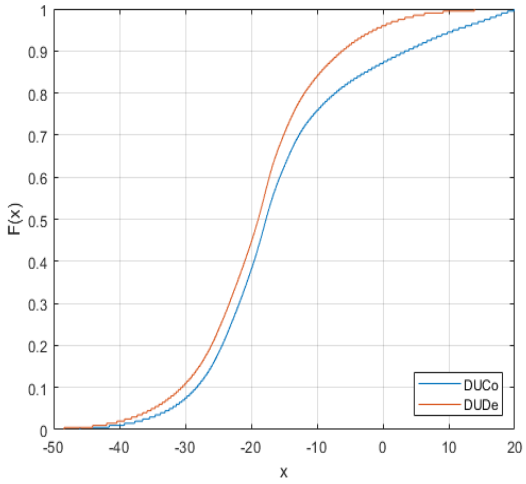


Figure 10. DUDe comparison with 30dBm UE limit for N=200.

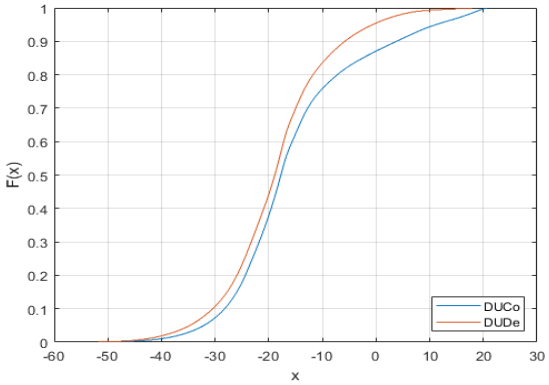


Figure 11. DUDe comparison with 30dBm UE limit for N=500.

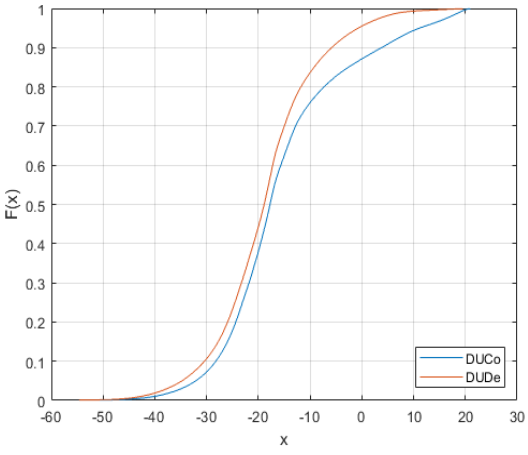


Figure 12. DUDe comparison with 30dBm UE limit for N=1000.

Figures 10 to 14 clearly indicate that increasing the transmission power of a UE significantly enhances the probability of establishing a DUDe association. Specifically, when the transmission power exceeds 2dBm, there is a 70% or greater likelihood of forming a DUDe connection. This data suggests that the DUDe communication algorithm is more efficient, requiring less transmission power to achieve similar results compared to the DUCo scenario.

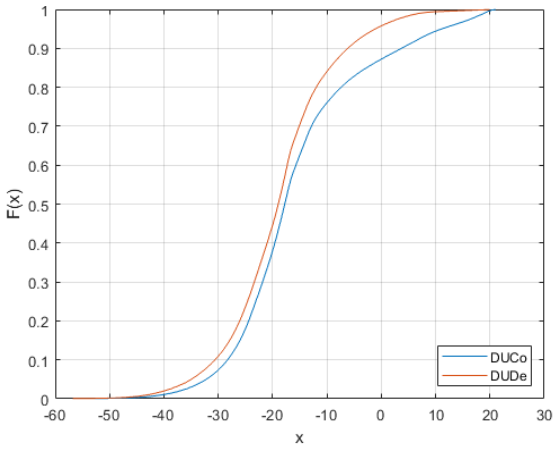


Figure 13. DUDe comparison with 30dBm UE limit for N=1500.

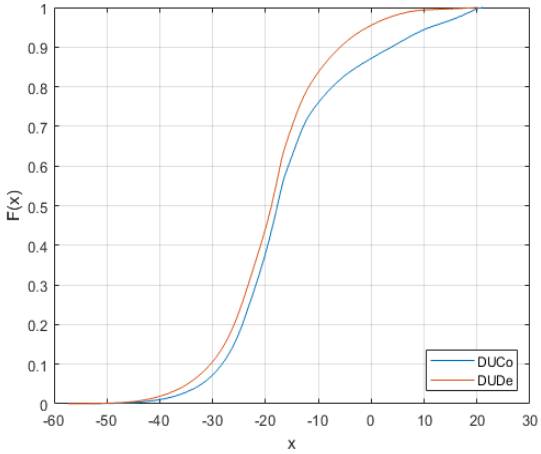


Figure 14. DUDe comparison with 30dBm UE limit for N=2000.

Moreover, the diagrams demonstrate that the DUDe method consistently exhibits a higher probability of creating a successful connection than the DUCo method. For example, at a 10 dBm transmission power level, the DUDe method achieves a 100% probability of establishing a successful connection. In contrast, the DUCo method requires a significantly higher transmission power—double that of DUDe—to attain the same success rate across various simulations (200, 500, 1000, 1500, and 2000 UEs).

The analysis conclusively shows that while the 30dBm limit is an important consideration in the implementation of a 5G network, it does not detract from the conclusion that the DUDe method is the superior choice for achieving a more energy-efficient and environmentally friendly network. The consistent results from the data and diagrams underscore the significant advantages of DUDe in terms of reduced power consumption, improved efficiency, and enhanced UE satisfaction.

Furthermore, the inherent flexibility and dynamic resource allocation capabilities of the DUDe method contribute to its superior performance. By allowing for the decoupling of downlink and uplink channels, DUDe optimizes the usage of available resources, leading to a more efficient network operation. This optimization not only conserves energy but also ensures that the network can handle high traffic volumes without compromising service quality.

C. Evaluation and analysis of 3nd scenario.

Figures 15 through 19 present a comparative performance analysis between the DUDe and DUCo approaches under a 25dBm power constraint is illustrated. These experiments were conducted for scenarios with 200, 500, 1000, 1500, and 2000 UEs, respectively.

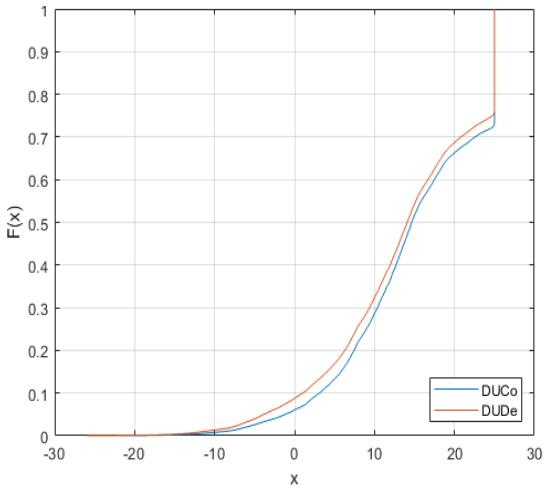


Figure 15. DUDe comparison with 25dBm UE limit for N=200.

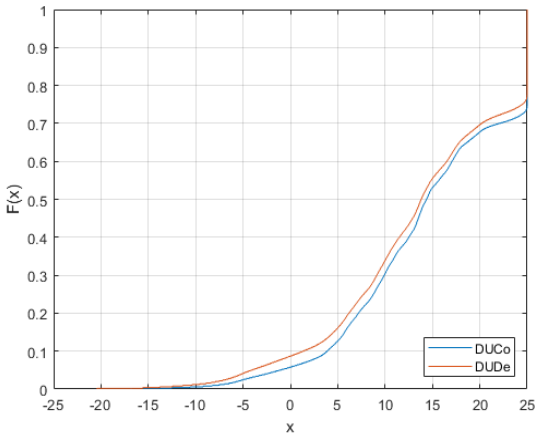


Figure 16. DUDe comparison with 25dBm UE limit for N=500.

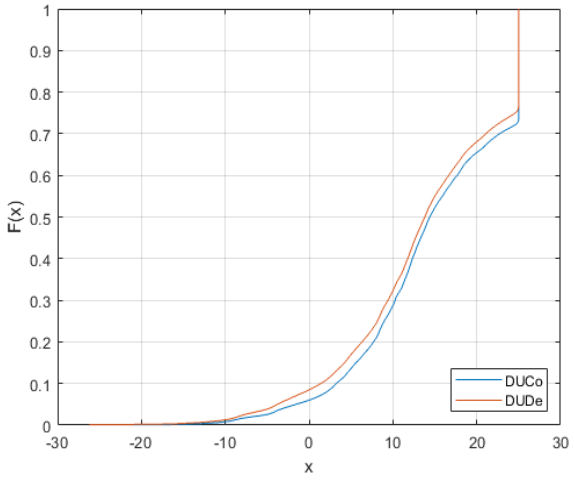


Figure 17. DUDe comparison with 25dBm UE limit for N=1000.

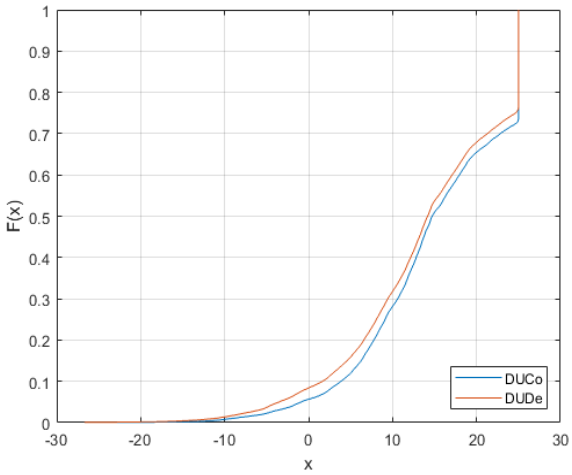


Figure 18. DUDe comparison with 25dBm UE limit for N=1500.

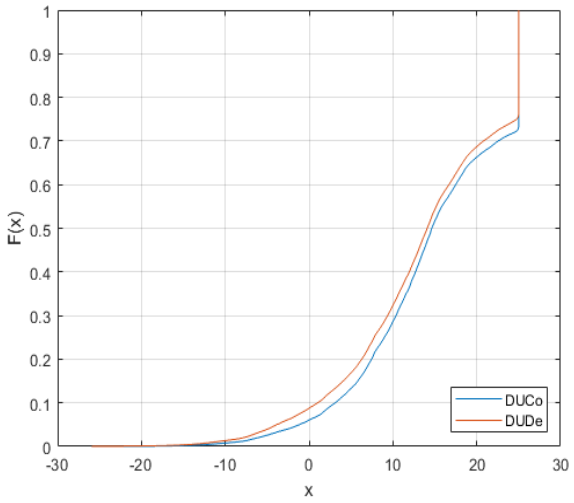


Figure 19. DUDe comparison with 25dBm UE limit for N=1500.

Figure 15 illustrates the scenario for 200 UEs. The CDF of the performance metric F(x) reveals that the DUDe approach (in red) provides better performance than the DUCo approach (in blue). For instance, at F(x)=0.5 the value of x is approximately 14dBm for DUDe, compared to about 13dBm for DUCo. This indicates a moderate performance improvement with the decoupled approach even with lower network loads.

Figure 16, depicting the scenario for 500 UEs, shows a clearer distinction between the two approaches. At F(x)=0.5, the value of is around 14.5dBm for DUDe and approximately 13.5dBm for DUCo. This improvement demonstrates the effectiveness of the DUDe approach in managing increased network loads.

Figure 17 represents the scenario for 1000 UEs. Here, the performance gained with DUDe becomes more pronounced. The DUDe curve remains to the right of the DUCo curve throughout the distribution. At F(x)=0.6, the value of x is about 18dBm for DUDe, while it is around 16dBm for DUCo. This significant difference underscores the advantage of the decoupled approach in more congested network conditions.

Furthermore, Figure 18, showing the scenario for 1500 UEs, further confirms the superior performance of the DUDe approach. The CDFs reveal that at F(x)=0.7, the value of x is approximately 21dBm for DUDe, compared to around 18dBm for DUCo. This greater separation between the curves indicates that DUDe continues to perform better as the network load increases.

Figure 19 illustrates the scenario for 2000 UEs, where the performance gain with DUDe is the most substantial. At $F(x)$=0.5, the value of is significantly higher for DUDe, reaching around 18dBm, compared to about 15dBm for DUCo. This result highlights DUDe's effectiveness in extremely high-density network environments.

These results suggest that the DUDe approach consistently offers better performance, particularly as the network load increases. The performance gain is evident across all scenarios, with more significant improvements observed in higher UE densities. For example, the value of x at F(x)=0.7 is approximately 20dBm for DUDe and 17dBm for DUCo in the 2000 UEs scenario, illustrating a clear advantage. Similarly, at F(x)=0.9, DUDe achieves around 23dBm, while DUCo reaches about 21dBm, reinforcing the trend.

This consistent performance advantage of the DUDe approach is attributed to its superior management of interference and more efficient resource allocation. As network density increases, these factors become increasingly critical, making DUDe a preferable choice for future 5G networks with high UE density.

In the set of Figures 5 to 19, in both scenarios, 'x' represents the average transmission power of the UEs Upon analyzing these figures, the following conclusion can draw: Increasing the transmission power of a UE results in a greater than 60% likelihood of establishing a DUDe association. Especially when the transmission power exceeds 10 dBm, the likelihood of DUDe correlation notably increases to over 50%. Also, with a stronger signal, which means a higher SNR value, a steady increase was observed in the DUDe correlation probability. Based on the insights gained in Figures 5 to 19 and the data

presented, it can be asserted that the DUDe scenario requires less transmission power to achieve similar outcomes compared to the coupled scenario. This observation has several benefits: it implies reduced BS power consumption, more efficient user service, and a higher overall level of user satisfaction compared to the coupled scenario.

VIII. Conclusion and Future Work

In conclusion, this study has conducted a thorough comparison of the energy efficiency between DUCo and DUDe within a 5G HetNet. The findings clearly demonstrate that DUDe is a more energy-efficient method for achieving comparable network performance, as it requires less energy consumption. The evidence from our analysis indicates that DUDe holds a significant promise for reducing energy use in 5G networks. By decoupling the downlink and uplink transmissions, DUDe optimizes resource utilization and reduces energy consumption while maintaining high-quality network performance.

This research underscores the potential of DUDe to contribute to more sustainable and efficient network operations. The advantages of separating DL and UL transmissions are evident in the enhanced energy savings and improved network reliability observed in our evaluations. These benefits not only support cost-effective network management but also align with broader environmental goals by lowering the overall carbon footprint of wireless communication systems.

Looking ahead, future research could explore further optimization of bandwidth allocation using the DUDe approach. By strategically allocating bandwidth to each cell, it is possible to enhance energy efficiency even further, minimize interference, and maximize throughput. This would involve developing advanced algorithms for dynamic bandwidth management that can adapt to varying network conditions and demands. Additionally, examining the scalability of DUDe in different network configurations and deployment scenarios would provide valuable insights for its broader implementation.

Further investigation into the integration of DUDe with emerging technologies, such as machine learning and artificial intelligence, could also yield significant improvements in network efficiency and performance. These technologies can enable more intelligent and adaptive resource management, further enhancing the benefits of the DUDe method.

In summary, DUDe presents a compelling case for adoption in future 5G network deployments due to its superior energy efficiency and potential for resource optimization. Continued research and development in this area will be crucial for realizing the full potential of DUDe and achieving more sustainable and high-performing wireless communication networks.

Acknowledgment

The research project was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "2nd Call for H.F.R.I. Research Projects to support Faculty Members & Researchers" (Project Number: 02440)

References

[1] Ch. A. Katsigiannis et al., “Simulation Based Energy Efficiency Analysis of DUDe 5G Networks”, The Twentieth Advanced International Conference on Telecommunications (AICT 2024), April 14 - 18, 2024, Venice, Italy.

[2] S. Ruiza, et al., “5G and beyond networks”, ScienceDirect, Inclusive Radio Communications for 5G and Beyond, pp. 141-186, 2014.

[3] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, “Downlink and Uplink Decoupling: A disruptive architectural design for 5G networks”, IEEE Global Communications Conference, Austin, TX, USA, pp. 1798-1803, 2014.

[4] H. Khan, M. Ali, I. Rashid, A. Ghafoor, and M. Naeem, “Cell Association for Energy Efficient Resource Allocation in Decoupled 5G Heterogeneous Networks”, 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, pp. 1-5, 2020.

[5] N. Garg, “Downlink and Uplink UE Association in Dense Next-Generation Wireless Networks”, Ph.D dissertation, University of Texas at Austin 2015 [Online]. Retrieved on 15/11/2024: http://hdl.handle.net/2152/30074.

[6] Y. Hu, R. MacKenzie, and M. Hao, “Expected Q-learning for Self-Organizing Resource Allocation in LTE-U with DL-UL Decoupling”, European Wireless; 23th European Wireless Conference, pp. 1-6, 2017.

[7] J. Guo,Y. Zhang, B. Guo, Z. Fan, and H. Song, “5G UL Coverage Enhancement Based on Coordinating NR TDD and NR FDD”, Signal and Information Processing, Networking and Computers. Lecture Notes in Electrical Engineering, vol 917. 2023, Springer, Singapore. https://doi.org/10.1007/978-981-19-3387-5_168.

[8] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, “Load & backhaul aware DUDe DL/UL access in 5G systems”, IEEE International Conference on Communications (ICC), pp. 5380-5385, 2015, doi: 10.1109/ICC.2015.7249179.

[9] T. Konstantinos et al., “On the Optimization of UE Allocation in Heterogeneous 5G Networks Using DUDe Techniques”, The 14th International Conference on Ubiquitous and Future Networks, July 4 - 7 2023, ECE – Ecole d’ingénieurs, Paris, France & Virtual Conference.

[10] T. Konstantinos et al., “Bandwidth Optimization Techniques in Heterogeneous 5G Networks Using DUDe”, 6th International Conference on Advanced Communication Technologies and Networking (CommNet), Rabat, Morocco, pp. 1-6, 2023.

[11] C. Bouras et al., “Optimizing Network Performance in 5G Systems with Downlink and Uplink Decoupling”, 6th International Conference on Advanced Communication Technologies and Networking (CommNet), Rabat, Morocco, pp. 1-6, 2023.

[12] Y. Ramamoorthi and A. Kumar, “Energy Efficiency in Millimeter Wave based Cellular Networks with DUDe and Dynamic TDD”, International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, pp. 670-673, 2020.

[13] K. Sun, J. -m. Wu, X. -y. Sun, and W. Huang, “Uplink Power Control of the DUDE Mode in HetNets”, IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, pp. 653-657, 2018.

[14] Y. Ramamoorthi and A. Kumar, “Dynamic Time Division Duplexing for Downlink/Uplink Decoupled Millimeter Wave-Based Cellular Networks”, in IEEE Communications Letters, vol. 23, no. 8, pp. 1441-1445, Aug. 2019.

[15] A. J. Muhammed, Z. Ma, Z. Zhang, P. Fan, and E. G. Larsson, “Energy-Efficient Resource Allocation for NOMA Based Small Cell Networks With Wireless Backhauls”, in IEEE Transactions on Communications, vol. 68, no. 6, pp. 3766-3781, June 2020.

[16] Y. Shi, E. Alsusa and M. W. Baidas, “A survey on downlink–uplink decoupled access: Advances, challenges, and open problems.” In ScienceDirect, Computer Networks, vol. 213 2022: 109040.

[17] H. Z. Khan et al., “Resource allocation and throughput maximization in decoupled 5G.", IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2020.

[18] F. A. Niasar and A. R. Momen, “A novel approach to fairness-aware energy efficiency in green heterogeneous cellular networks,” Wireless Networks, vol. 28, pp. 2651–2667, 2022.

[19] 3GPP, TR 38.901 Section 7.4.1"Study on channel model for frequencies from 0.5 to 100 GHz." (Release 17), March 2022.

[20] B. Błaszczyszyn and H. P. Keeler, “Studying the SINR Process of the Typical UE in Poisson Networks Using Its Factorial Moment Measures”, in IEEE Transactions on Information Theory, vol. 61, no. 12, pp. 6774-6794, Dec. 2015.

[21] MathWorks. (2023). 5G Toolbox [Software]. Retrieved on 15/11/2024 from https://www.mathworks.com/products/5g.html.

[22] 3rd Generation Partnership Project (3GPP). (n.d.). 3GPP References Search. Retrieved on 15/11/2024 from 3GPP website: “https://portal.3gpp.org/3gppreferences/SearchReferences.aspx”.

Byte Consistency Verification Method with Dynamic Threshold Adjustment for Each Node in Software-Defined Networking

Naoya Kitagawa
Research and Development Center for Academic Networks
National Institute of Informatics
Tokyo, Japan
kitagawa@nii.ac.jp

Jumpei Sato
Graduate School of Marine Science and Technology
Tokyo University of Marine Science and Technology
Tokyo, Japan
m234020@edu.kaiyodai.ac.jp

Kohta Ohshima
Marine Electromechanical Engineering Division
Tokyo University of Marine Science and Technology
Tokyo, Japan
kxoh@kaiyodai.ac.jp

***Abstract*—Software-defined networking (SDN), which enables flexible routing control based on communication content, has been widely studied as a countermeasure against possible attacks on the data plane by compromised SDN switches and hosts. We previously proposed a byte consistency verification method that uses information such as transfer volume collected from SDN switches to detect anomalous communications, even when the communications are encrypted. In addition, we improved the anomaly detection performance of this method by implementing high-precision time synchronization and an SDN switch function for each host. In this study, we extend the scope of information collection to each host (in addition to SDN switches) and propose a data plane anomaly detection method that monitors the communication volume of each process at each host. We also propose a method that automatically adjusts the threshold, which can be set individually for each node, used for detection. Furthermore, we implement and evaluate the proposed method on a network testbed. The results confirm that it can be used to improve anomaly detection accuracy.**

Index Terms*—Software-defined networking; Data plane verification; Byte consistency verification; Anomaly detection.

I. Introduction

This paper is an extended version of our study presented at the Twentieth International Conference on Networking and Services (ICNS 2024) [1]. Network equipment that uses software-defined networking (SDN) and network functions virtualization (NFV) technology has recently been introduced into carrier and data center networks; further widespread use is expected [2]. Unlike conventional router devices, which have fixed settings, SDN enables flexible routing control using various types of information, such as the content of transmitted data, information on sending and receiving terminals, and networks passed through. The SDN switches that make up an SDN network cooperate according to control information from the SDN controller, enabling fine control of communication on a flow-by-flow basis.

In the operation of SDN, it is important to ensure compatibility with security-related technologies. Encrypted communication is becoming a mainstream measure against information leaks, with Google reporting that 95% of its total communication traffic was encrypted as of November 2023 [3]. While encrypted communication can ensure end-to-end security, it makes it difficult for network operators to use intrusion detection and prevention systems, which provide security by checking the payload of exchanged packets. In addition, network administrators must also be aware of countermeasures against SDN switches and silent failures. SDN networks are often realized using software switches, making them possibly more vulnerable than networks consisting of conventional hardware switches [4] [5] [6] [7]. Specifically, SDN controllers may not be able to detect SDN switches that are compromised or defective. To solve these security issues, byte integrity verification has been proposed, where anomalies are detected by collecting and processing communication status data from a group of SDN switches.

We previously proposed a method for increasing the granularity of anomalies that can be detected in SDN communications by using high-precision time synchronization via IEEE1588 PTPv2 [8] to ensure the time resolution of collected communication status data and by handling transfer volume information in units of flows [9]. Furthermore, to solve the problem of conventional byte consistency verification, where the accuracy of information collected from a terminal SDN switch cannot be verified, we developed a method for expanding the range of devices that can detect anomalies by incorporating a reporting function similar to that of SDN switches in the host connected to a terminal SDN switch [10]. The results of our previous research indicate that the quality and variety of data that can be collected from SDN switches and hosts are useful for improving the anomaly detection

performance of SDN.

In this paper, we confirm the applicability of byte consistency verification for anomaly detection in SDN networks by collecting communication status data for each host. In our approach, statistical data on the communication status, which can be obtained using commands provided by the host operating system (Linux), are formatted to be compatible with SDN networks. They can be used by SDN controllers and nodes that perform byte integrity verification. We implement this method on a network testbed to obtain per-process communication volume information measured at each host and confirm that it is applicable to anomaly detection in SDN networks.

In addition, a conventional anomaly detection method (see Section II) depends on the knowledge and experience of the administrator because the threshold values need to be set manually. Since conventional methods use the same threshold value for the entire network, they cannot detect minute anomalies or identify the location of anomalies. To address these issues, we propose and implement a method for detecting anomalous switches that automatically sets the threshold value for each node individually. The improvement in anomaly detection accuracy is determined through experiments using a testbed.

The rest of this paper organized as follows. Section II describes related techniques and existing research. Section III explains the proposed network verification scheme, which deals with the process-level communication volume of hosts. Section IV describes the dynamic adjustment of thresholds. Section V describes an experiment in which the proposed method was implemented on a testbed. Section VI discusses considerations based on the results of evaluation experiments. Finally, Section VII presents the concluding remarks.

II. Related Work

In this section, we review related technology and existing research.

A. Software-Defined Networking

SDN allows network devices to be centrally controlled through software. In a conventional network, shown in Figure 1, the network administrator configures each router for routing control. The router forwards packets according to its configuration. In contrast, in an SDN network, forwarding control instructions can be issued to all SDN switches by configuring the SDN controller. The SDN switches perform packet forwarding based on these instructions. Therefore, SDN allows dynamic control based on the operation status of each SDN switch. Flexible control in SDN is achieved by separating the data plane, which handles data forwarding functions, and the control plane, which handles control functions [11].

OpenFlow [12] is widely used for implementing SDN. There are several OpenFlow controller implementations, such as Floodlight [13]. Although SDN allows for flexible control of the network, several security issues have been reported [4] [14]. For example, there are known attacks in which malicious switches attack the data plane or mislead the SDN controller about the network topology. Methods have been developed to solve these problems [5] [6].

B. Data Plane Security in SDN

If an SDN network is compromised, unintended packets may be discarded or generated and routes may be changed. To prevent such problems, verification techniques can be used to protect the data plane. SPHINX verifies compromised switches using byte consistency verification [5]. This method detects anomalies by having each switch collect and compare transfer volume information.

Figure 2 shows the operation of byte consistency verification by SPHINX. A report of the forwarding volume information from each switch is received. Based on the received information, the method calculates a moving average ($\sum$) of the transfer volume information for each SDN switch and a value ($\sum_{avg}$) obtained by averaging the moving average for each SDN switch over all SDN switches. Then, the method checks whether the average value deviates from the moving average value of each SDN switch by the inequality in Equation 1 using a predetermined value of threshold τ.

$$\frac{1}{\tau} < \frac{\sum}{\sum_{avg}} < \tau \tag{1}$$

If threshold τ is excessively small, false positives (FPs) are likely to occur; if it is excessively large, false negatives (FNs) are likely to occur. The appropriate value of τ depends on the network configuration and type of switches used. It is thus necessary to set an appropriate value for each network. Since SPHINX performs byte integrity verification using network switch forwarding volume information, it cannot verify whether an edge switch is malicious and it does not support flow aggregation. Various other SDN data plane security measures have been proposed [14].

C. WhiteRabbit

As described in Section II-B, for SPHINX, detection accuracy is affected by variations in the timing of obtaining statistic information from switches. To address this issue, WhiteRabbit reduces the deterioration of verification accuracy due to acquisition timing deviations by using IEEE1588 PTPv2 for high-precision time synchronization and scheduling the timing of the acquisition of transfer volume information [9]. However, WhiteRabbit, like SPHINX, does not verify edge switches and does not support flow aggregation.

D. Edge Switch Validation with In-host Switches

As mentioned in Sections II-B and II-C, byte consistency verification using only SDN switch information cannot verify edge switches. To solve this problem, we previously proposed a method for obtaining the communication volume of each host [10]. This method builds a switch inside the host to obtain the host's communication volume and behaves like any other SDN switch, allowing byte integrity verification between the edge switch and the host. However, the method requires the threshold τ in Equation 1 to be larger than that for the

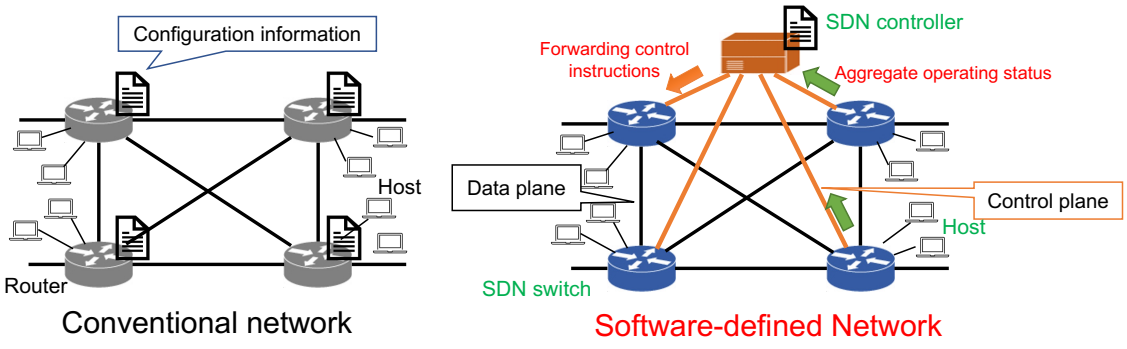


Fig. 1. Comparison of conventional network and software-defined network.

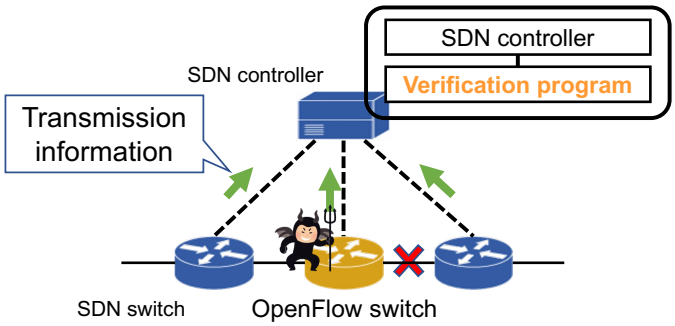


Fig. 2. Byte consistency verification by SPHINX.

conventional method, which may make it miss minute network anomalies or attacks that take place in a very small amount of time.

III. Proposed Method

To overcome the issues described in Section II, in this section, we describe a network verification scheme that deals with the process-level communication volume of hosts. Figure 3 shows an overview of the proposed method. As shown, host information is collected by implementing an in-host information collection function on hosts in a conventional SDN network. In addition, we deploy a host information collection server to compare the SDN controller's collection of each SDN switch's forwarding volume information. This allows the verification system to perform host-information-aware verification. This system improves the accuracy of detecting abnormal networks by classifying communication volume using detailed host information, which cannot be obtained using the conventional method.

This system requires the implementation of the following two functions.

1) A function for each host to send its collected data (process-level traffic information) to the host information collection server.
2) A function for the SDN controller to send the traffic information of each switch to the host information collection server.

In addition, the host information collection server needs to know which host sent the data and compare the data with the transfer volume information of each switch. Furthermore, each host needs to implement a function to collect its own process-level traffic and send the collected information to the server.

A. Host Information Collection Server

The host information collection server monitors the traffic information of all hosts that have executed the intra-host information transmission agent and alerts the user according to the conditions based on the statistics of the traffic information. The host information collection server collects per-process communication volume information from each host, compares it with the transfer volume information of each switch collected by the SDN controller, and sends an alert to the network administrator if any abnormality is found.

B. Host Information Collection Agent

The host information collection agent, which is implemented on each host, executes the *ss* (socket statistics) command provided by the Linux operating system as an external command to obtain the cumulative number of received packets as statistical information for each process. Then, the agent sends the acquired information to the host information collection server. By repeating these processes periodically, the host information collection agent collects transfer volume information for each host.

IV. Automatic Adjustment of Threshold

In this section, we describe a method that automatically sets threshold τ for byte consistency verification and allows the threshold to be fine-tuned for each node, thereby improving the granularity of anomaly detection. As described in Section III, in our scheme, the SDN controller collects transfer volume information from the host switches and from each SDN switch. The anomaly detection system uses this information to detect anomalies. Conventional anomaly detection methods such as

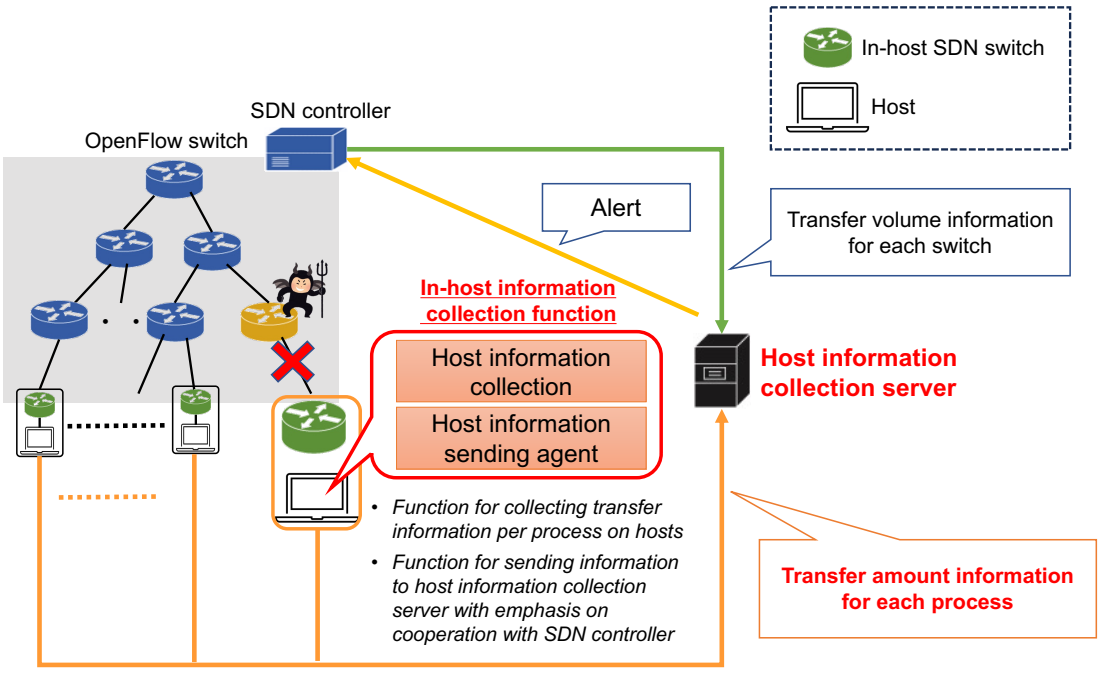


Fig. 3. Overview of proposed method.

SPHINX require manual setting of threshold values used for detection at the discretion of the administrator. In contrast, we automate the setting of threshold values used for detection so that they can be adjusted without relying on the knowledge and experience of the network administrator. This also allows the setting of individual thresholds for nodes, which is not possible with the conventional method. Setting an appropriate threshold for each node enables the detection of minute anomalies and the identification of switches with anomalies that are missed with the conventional method.

A. Calculation Method

The first step in the thresholding calculation is to determine the reference threshold value, as shown in Equation 2. Figure 4 shows an example of switch placement. As shown, when traffic flows from left to right, the switch closest to the origin is defined as $FormerSwitch$ (FSW) and the switch closest to the end is denoted as $LatterSwitch$ (LSW) .

$$Reference\ Threshold\ = \frac{FSW's\ transfer\ volume}{LSW's\ transfer\ volume} \quad (2)$$

Then, as shown in Figure 5, the upper and lower threshold limits (tolerance rate) are set and the reference threshold is given as the range $\pm\ x\ \%$. By giving the threshold as a range, the sensitivity of anomaly detection can be adjusted and FPs can be prevented.

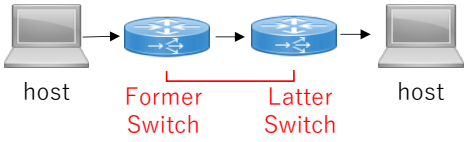


Fig. 4. Example of switch placement.

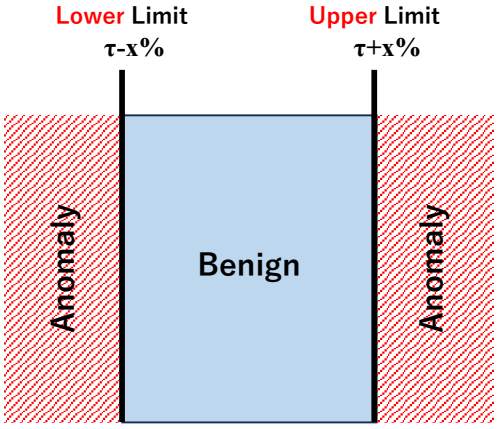


Fig. 5. Threshold tolerance.

B. Individual Threshold Setting for Nodes

Figure 6 shows the method used to set individual threshold values for nodes. Using the calculation method described in Section 3.2, threshold values $\tau 1$–$\tau 4$ are set between nodes SW1 and SW2, SW2 and SW3, SW3 and SW4, and SW4 and SW5, respectively. By setting individual threshold values in this way, it is possible to identify which threshold value was used to detect an anomaly and thus the switch to which

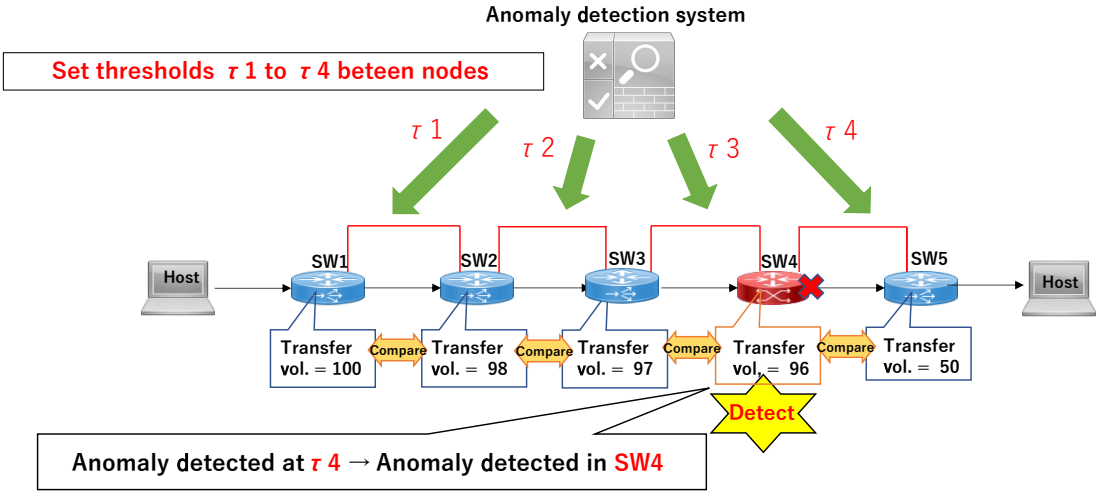


Fig. 6. Individual threshold setting and anomaly detection for nodes.

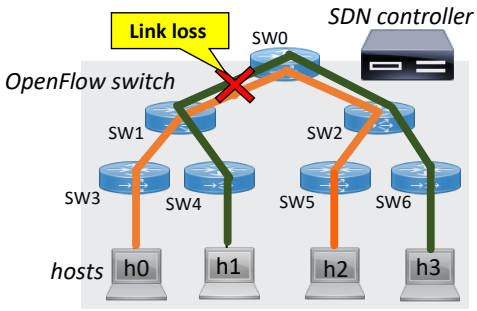


Fig. 7. Network configuration used in evaluation experiment.

TABLE I
SPECIFICATIONS OF MICROCLOUD NODES IN DETERLAB USED IN EXPERIMENT.

Type	Specifications
CPU	Intel(R) Xeon(R) E3-1260L Quad-Core Processor Running at 2.4 GHz
Memory	16 GB
Storage	250 GB SATA Western Digital RE4 Disk Drive
OS	Ubuntu 16.04 LST

that threshold value was assigned can be identified as the anomalous switch. In the example shown in Figure 6, the anomaly is detected at threshold value $\tau 4$, which means that SW4 is anomalous.

V. EXPERIMENT

A. Environment

To verify and evaluate the operation of the host information collection function based on this method, we implemented an experimental network on DeterLab, a network testbed operated by the University of Southern California Information Sciences Institute and the University of Utah [15].

We used a total of 12 nodes on DeterLab, each with an SDN controller, a verification component, SDN switches (7 nodes), and hosts (4 nodes). As shown in Figure 7, the network for this experiment had a tree network topology with $Depth = 2$ and $Fanout = 2$, where $Depth$ indicates the depth of the hierarchy from the root node and $Fanout$ indicates how many nodes are connected in one branch.

Table I shows the specifications of the MicroCloud on DeterLab used in this experiment. All 12 nodes in this experiment used equipment with the same specifications. We used Floodlight v1.2 [18] as an SDN controller. We also implemented an OpenFlow proxy, stopcock, between the group of switches and the controller as the verification component for route verification. ofsoftswitch13_EXT340 [19] was used as the SDN switches. Since this experimental network consisted of actual equipment rather than simulators or emulators, the evaluation environment was close to that in actual operation.

B. Evaluation of Detection Accuracy

In this section, we describe an experiment conducted to determine the anomaly detection accuracy for the proposed method and the conventional method SPHINX. In this evaluation experiment, we investigated the effectiveness of the thresholds automatically set by the proposed method and the effect of partial threshold setting on anomaly detection accuracy.

We used the FN rate for malicious traffic and the FP rate for benign traffic as evaluation metrics. We measured TCP communications over a five-hop path using iperf and compared the anomaly detection accuracy of the proposed method with that of SPHINX based on the amount of forwarded data sent from each switch to the SDN controller.

1) False Negative Rate: We set up a link with a loss rate between SW0 and SW1 and discarded packets on this link, as

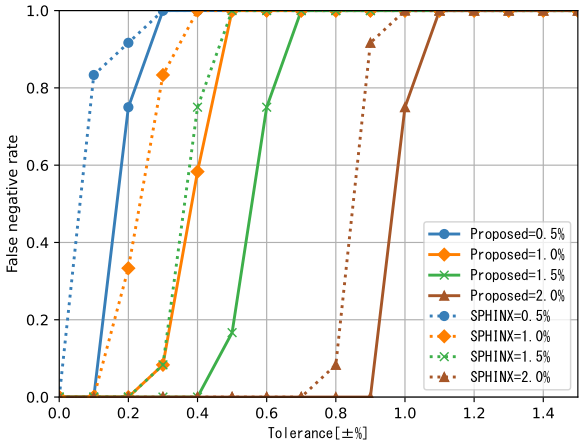


Fig. 8. FN rate comparison between SPHINX and proposed method.

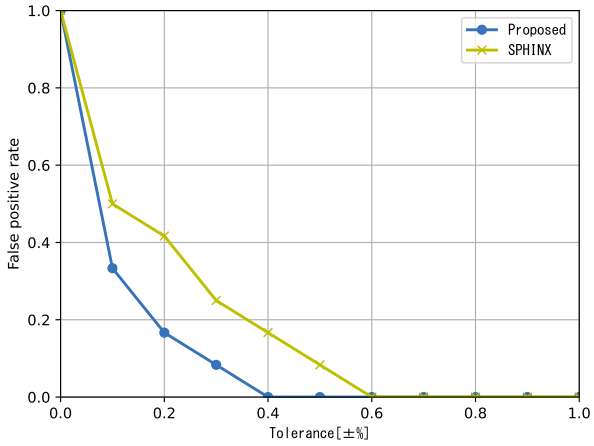


Fig. 9. FP rate comparison between SPHINX and proposed method.

shown in Figure 7. The link loss rates in this experiment were 0.5%, 1.0%, 1.5%, and 2.0%.

Figure 8 shows a comparison of the FN rate between the proposed method and SPHINX. The vertical axis indicates the FN rate and the horizontal axis indicates the tolerance rate x %, which is a $\pm$ x % variation of the reference threshold obtained in Equation 2 in Section IV-A. An FN rate that is sufficiently low in the range where the tolerance is greater than 0% indicates that there is no detection failure (i.e., that anomalies were detected). For SPHINX, the FN rate increased rapidly when the link loss rate was 0.5%, with the tolerance rate increasing from 0%. This indicates that SPHINX was unable to detect an anomaly when the link loss rate was 0.5%. In contrast, for the proposed method, when the link loss rate was 0.5%, the FN rate remained at 0 up to a tolerance rate of $\pm$ 0.1%, confirming that our method could detect anomalies. The results for a link loss rate of 2.0% indicate that the tolerance rate x at which the FN rates for SPHINX and the proposed method begin to increase is 0.7% and 0.9%, respectively.

2) False Positive Rate: We measured the FP rate after generating traffic flows, as done in the evaluation of the FN rate. Figure 9 compares the FP rates for SPHINX and the proposed method. As shown, the tolerance rate at which the FP rate becomes zero is 0.6% for SPHINX and 0.4% for the proposed method.

C. Effect of Threshold Calculation Method

To evaluate the effectiveness of the threshold calculation method presented in Section IV-A, we evaluated the FN and FP rates using another threshold calculation method that automatically sets the threshold. The calculation method is shown in Equation 3.

$$Reference\,Threshold = \frac{Ingress\ SW's\ transfer\ volume}{Egress\ SW's\ transfer\ volume} \quad (3)$$

1) False Negative Rate: As done above, the link loss rate was set to 0.5%, 1.0%, 1.5%, and 2.0%. Figure 10 shows a comparison of the FN rate between the proposed threshold calculation method, shown in Equation 2, and the calculation method in Equation 3. For a link loss rate of 0.5%, minute anomalies were detected only when the proposed method was used. It can also be seen that the proposed method has a larger tolerance rate x at which the FN rate begins to increase when the link loss rate is 2.0%.

2) False Positive Rate: The FP rate was measured after traffic flows were generated for the evaluation experiment. Figure 11 shows a comparison of the FP rate between the proposed threshold calculation method, shown in Equation 2, and the calculation method in Equation 3. The acceptable rate at which the FP rate becomes zero is 0.4% for the proposed method and 0.8% for the calculation method in Equation 3.

VI. Considerations and Discussion

A. Comparison of SPHINX and Proposed Method

When there is a link with a loss rate, the FN rate increases rapidly for SPHINX, with the tolerance rate increasing from 0% when the link loss rate is 0.5%. This result indicates that SPHINX is unable to detect an anomaly when the link loss rate is 0.5%. In contrast, for the proposed method, when the link loss rate is 0.5%, the FN rate remains at 0 up to a tolerance rate of $\pm$ 0.1%. This indicates that the proposed method can detect minute anomalies that SPHINX cannot.

In the experiment with a link loss rate of 2.0%, the tolerance rate x at which the FN rates for SPHINX and the proposed method begin to increase was 0.7% and 0.9%, respectively. This confirms that the proposed threshold setting method is superior to that of SPHINX.

For benign traffic, the tolerance rate at which the FP rate becomes 0 is 0.6% for SPHINX and 0.4% for the proposed

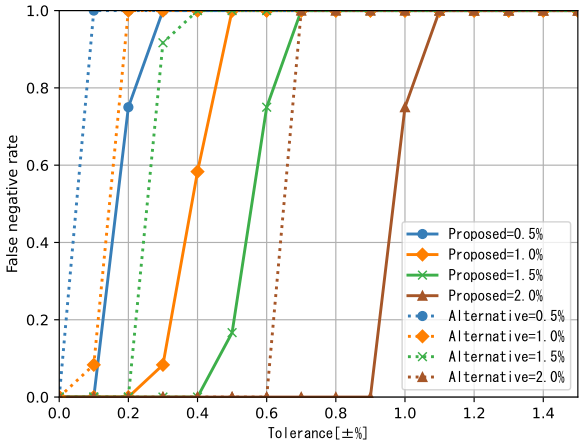


Fig. 10. FN rate comparison between proposed threshold calculation method and alternative method in Equation 3.

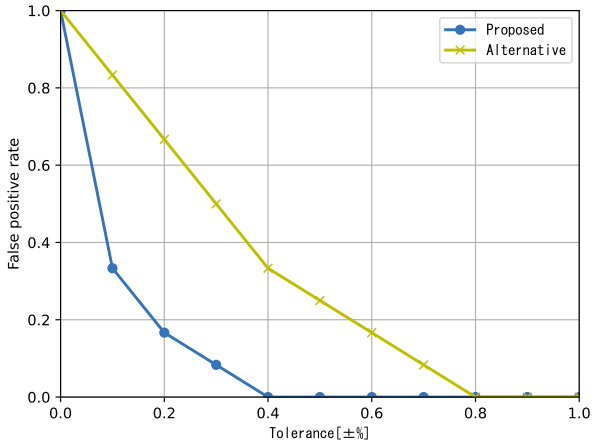


Fig. 11. FP rate comparison between proposed threshold calculation method and alternative method in Equation 3.

method. The FP rate is thus not considered to be significantly different between SPHINX and the proposed method. As can be seen in Figures 8 and 9, the acceptable rate of no false detection and no missed detection for the proposed method is 0.4% to 0.9%. Link loss rates of 1.5% and 2.0% within this range can be correctly detected without false detection or missed detection.

These results show that there is a trade-off between the FN rate and the FP rate. The improved FN rate for the proposed method despite similar FP rates between SPHINX and the proposed method can be attributed to the improvement in anomaly detection accuracy by the automatic setting of individual thresholds.

B. Discussion of Threshold Calculation Methods

As described in Section V-C, to evaluate the validity of the threshold calculation scheme of the proposed method, the FN and FP rates obtained for an alternative calculation method were compared. Regarding the FN rate, it was confirmed that the alternative calculation method was unable to detect small link loss rates (i.e., small anomalies). In addition, the threshold tolerance X was larger for the proposed method for all link loss rates. Regarding the FP rate, the proposed method had an FP rate of 0.4% and the alternative method had an FP rate of 0.8%. These results confirm the validity of the proposed threshold calculation method.

VII. Conclusion

This study proposed a method for collecting forwarding volume information for each host in an SDN network to improve network verification accuracy.

The threshold values used for detection are automatically set so that they can be adjusted without relying on the knowledge and experience of the network administrator. This allows the setting of individual thresholds for nodes, which is not possible with the conventional method. Setting an appropriate threshold for each node enables the detection of minute anomalies and the identification of switches with anomalies that are missed with the conventional method.

We evaluated the effectiveness of the automatically set thresholds and the impact of the threshold range. The results confirm that the proposed method improves the FN rate and maintains the FP rate compared to those for SPHINX. The improvement in the FN rate and maintenance of the FP rate confirm the effectiveness of the proposed threshold calculation method and demonstrate that applying individual thresholds improves anomaly detection accuracy.

Acknowledgement

We would like to thank the team at the University of South California Information Sciences Institute and the University of Utah, which operated the DeterLab project, for providing the network testbed for this research.

A part of this work was supported by JSPS KAKENHI Grant Number JP19K20252 and JP24K14925.

References

[1] N. Kitagawa, N. Moriyama, and K. Ohshima, “Anomaly Detection by Monitoring Communication Volume at the Process Level of Each Host in SDN,” Proc. of The Twentieth International Conference on Networking and Services (ICNS 2024), pp. 1-6, 2024.

[2] R. Souza, K. Dias and S. Fernandes, “NFV Data Centers: A Systematic Review,” IEEE Access, vol. 8, pp. 51713-51735, 2020.

[3] “HTTPS encryption on the web,” https://transparencyreport.google.com/https/overview?lang=en&hl=en (Accessed: Nov. 10, 2024).

[4] D. Kreutz, F. M. V. Ramos, and P. Verissimo, “Towards secure and dependable software-defined networks,” Proc. of the 2013 ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking, pp. 55-60, 2013.

[5] M. Dhawan, R. Poddar, K. Mahajan, and V. Mann, “SPHINX: Detecting Security Attacks in Software- Defined Networks,” 2015, doi: 10.14722/ndss.2015.23064.

[6] A. Shaghaghi, M. A. Kaafar, and S. Jha, “WedgeTail: An intrusion prevention system for the data plane of software defined networks,” Proc. of the 2017 ACM Asia Conference on Computer and Communications Security, pp. 849-861, 2017.

[7] A. Feldmann, P. Heyder, M. Kreutzer, S. Schmid, J. Seifert, H. Shulman, et al., "NetCo: Reliable Routing with Unreliable Routers," Proc. of 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, pp. 128-135, 2016.
[8] IEEE, 1588-2008, "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", pp. 1-300, 2008.
[9] T. Shimizu, N. Kitagawa, K. Ohshima and N. Yamai, "WhiteRabbit: Scalable Software-Defined Network Data-Plane Verification Method Through Time Scheduling," IEEE Access, vol. 7, pp. 97296-97306, 2019.
[10] T. Amano, T. Shimizu, N. Kitagawa, and K. Ohshima, "SDN Data-Plane Verification Method using End-to-End Traffic Statistics," IEICE Tech. Rep., vol. 120, no. 413, NS2020-163, pp. 238-243, 2021 (in Japanese).
[11] K. Benzekki, A. Fergougui, and A. Elalaoui, "Software- defined networking (SDN): a survey," Security and communication networks 9.18, pp. 5803-5833, 2016.
[12] N. McKeown, T. Anderson, H.i Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, et al., "OpenFlow: enabling innovation in campus networks," ACM SIGCOMM computer communication review 38.2, pp. 69-74, 2008.
[13] S. Scott-Hayward, S. Natarajan and S. Sezer, "A Survey of Security in Software Defined Networks," IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 623-654, 2016.
[14] C. Black and S. Scott-Hayward, "A Survey on the Verification of Adversarial Data Planes in Software-Defined Networks," Proc. of the 2021 ACM International Workshop on Software Defined Networks & Network Function Virtualization Security, pp. 3-10, 2021.
[15] "DeterLab," https://www.isi.deterlab.net/ (Accessed: Nov. 10, 2024).
[16] "Prometheus," https://prometheus.io (Accessed: Nov. 10, 2024).
[17] "GitHub - prometheus/Prometheus: The Prometheus monitoring system and time series database," https://github.com/prometheus/prometheus (Accessed Nov. 10, 2024).
[18] "Floodlight Controller," https://floodlight.atlassian.net/wiki/spaces/floodlightcontroller/overview (Accessed: Nov. 10, 2024).
[19] "ofsoftswitch13_EXT-340," https://github.com/oronanschel/ofsoftswitch13_EXT-340 (Accessed: Nov. 10, 2024).

Lessons Learned from Building Sustainable Municipal LoRaWan Infrastructure

André Nitze
Department of Business
Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
andre.nitze@th-brandenburg.de

Josephine Jahn
[Sustainability –Transformation - Transfer] Research Center
Eberswalde University for Sustainable Development
Eberswalde, Germany
josephine.jahn@hnee.de

Tingting Wang
Department of Business
Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
tingting.wang@th-brandenburg.de

Sabah Ali
Department of Business
Brandenburg University of Applied Sciences
Brandenburg an der Havel, Germany
sabah.ali@th-brandenburg.de

Timon Miesner
[Sustainability –Transformation - Transfer] Research Center
Eberswalde University for Sustainable Development
Eberswalde, Germany
timon.miesner@hnee.de

***Abstract*—This paper introduces the LoRaWAN Collaboration Framework (LCF), a strategic blueprint for deploying and managing LoRaWAN infrastructures in smart cities with an emphasis on rural and small municipalities. LoRaWAN technology distinguishes itself by its capability to support long-range, low-power IoT applications, making it ideal for extensive and sparsely populated areas. The LCF aims to address common challenges in these settings, such as limited technical expertise, financial constraints, and the need for cross-municipal cooperation. It outlines roles and responsibilities across various stakeholders including municipal authorities, IT service providers, application developers, and end-users. The framework emphasizes the balance of technological, economic, social and ecological sustainability in line with the United Nations' Sustainable Development Goals. In this paper, we describe the experiences from several LoRaWAN projects in small towns and municipalities in Germany and give some insights to these use cases, the derived collaboration framework, and other arguments to consider before implementing LoRaWAN infrastructures.**

Keywords-component; Service delivery; collaboration; sustainability; smart city; infrastructure; LoRaWAN Collaboration Framework (LCF)

I. INTRODUCTION

Sustainability [1] and efficient resource utilization is an indispensable reality of today's world. Sustainability is the primary element to execute all development goals that are a part of Sustainable Development Goals 2030 [26] and a sustainable earth for all stages of life [2, 3]. It is assumed to be the ultimate target of not only the United Nation, but of many developed and developing countries [4,5] and for example the World Economic Forum. Lastly, it is expected that sustainability brings long lasting socio-economic benefits to the general masses and the environment [6].

To achieve this, a significant variation in control and monitoring conditions is required, which underlines the importance and high demand for digitalization in the future. The Internet of Things (IoT) is helping to minimize the difference between the digital world and the physical world. Cities may tackle this demand with the help of Internet of Things technologies and develop smarter cities. Long-range wide-area-networks (LoRaWANs) are a big contribution in this development, as their range and costs are well-suited for providing a network infrastructure for smart city applications.

The biggest advantage of LoRaWan lies in the high range and its efficient energy consumption. A single gateway can cover distances up to 5 kilometers in urban settings and up to 15 km in rural areas. The range is dependent on various variables, for example, if there is no impediment between the transmitter and the receiver, the range can also increase, whereas the presence of an hindrance (buildings, trees, heavy rain, snow) can affect the signal quality negatively and reduce the range drastically. That means in order to cover a larger area, multiple gateways or a stronger LoRaWAN gateway antenna are necessary. Although the range depends on the environment and other obstacles, it still has the most advanced and reliable power transfer balance when compared to other communication technologies.

LoRaWAN provides wireless data transmission that is comparable to Wi-Fi and Bluetooth but has its own distinct properties. LoRaWAN is a low-power wide-area network

(LPWAN) technology that facilitates communication of connected devices covering long distances while consuming low energy [7]. This makes LoRaWAN particularly appropriate for rural areas. The LoRa Alliance developed LoRaWAN specification, whose basic modules as open-source software are available [8].

LoRaWAN is an enabler technology [9] that not only helps achieve the Sustainable Development Goals (SDGs) by measuring climate impacts, soil moisture, sealing, modal split, water levels et cetera but it also allows for automating and initiating counter measures, for example to save CO2-binding trees, moorland and so on.

Despite the growing need, only technical experts can perform the required construction work, data integration, processing, and visualization required for an end-to-end LoRaWAN use case, and small towns and municipalities cannot maneuver this alone. This group lacks basic technical knowledge and has insufficient human resources – in particular IT staff. Moreover, there are only limited financial resources available for digitalization. Hence, many projects concerning IoT or "smart city" are conducted using third party funding (public or private) instead of household budgets, or not at all.

Despite the urgent need for climate adaptation at all levels, the willingness of small towns and communities to contribute to climate goals, and the relatively low cost of LoRaWAN, in many cases this is not enough to build technically and organizational sustainable LoRaWAN infrastructures in these regions. This raises the question of a common operating model for LoRaWAN in rural areas, for example across several small towns and municipalities. This would demonstrate efficiency and cost benefit (cost savings, volume benefits, and production efficiencies) to be attained that would make a business economically viable. That can only be achieved if all the required actors have clear areas of responsibility and associate the LoRaWAN infrastructure with a benefit for themselves.

There is a clear need for a framework that:

- defines responsibilities for different stakeholders in LoRaWAN projects,
- balances ecological, economic, technical, and social objectives, and
- enables local authorities and municipalities to sustainably operate LoRaWANs.

Therefore, in this paper we propose the LoRaWAN Collaboration Framework (LCF), which addresses and tries to solve these issues.

The rest of this paper is organized as follows: Section II describes the related work in the areas of sustainability and LoRaWAN service management. Section III describes how we gained our findings and gives some insights into our use cases. Section IV describes the framework including the responsibilities of the various stakeholders and the organizational interfaces and other arguments to consider before implementing LoRaWAN infrastructures from a rural municipalities or rural town's point of view. The conclusions close the article.

II. RELATED WORK

Due to its impact on the economic and social sphere, digitalization is no longer seen as an isolated technical phenomenon. Digitalization as an encompassing process is related to massive changes in the economic production, in communication patterns and in other social aspects of everyday life and therefore has an influence on the society.

LoRaWAN technology helps in attaining Maslow's hierarchmuof needs along with sustainability. It's the hierarchy level that distributes the term "need" into five components, i.e., physiological needs (air, food, water), security needs (health, security, financial performance), social needs (relationships and belonging), respect needs and self-actualization needs. LoRaWAN supports the first two base levels of Maslow's hierarchy of needs (physiological needs and security needs). As they are foundation levels, the other three levels cannot be achieved until these are attained.

Nölting and Dembski found that digitalization, as a process, has no normative direction itself, but, in contrast, is governed by individual and organizational entities in accordance with their own goals [10]. Digitalization technologies such as LoRaWAN provide many opportunities, which can or cannot be used in terms of sustainability in its diverse dimensions. Due to this, some authors underline the necessity to regard the big societal trends, forward digitalization and sustainability, together as a twin transformation. In this setting, digitalization is aimed at normative common goals, and functions as an enabler for sustainability [11, 12, 13].

According to Farsi, Hosseinian Far, Daneshkhah, and Sedighi [14] sustainability is important to maintain the basis for sustainability assessment. In 1987, the United Nations Brundtland Commission defined sustainable as "meeting the needs of the present without compromising the ability of future generations to meet their own needs" [15], and hence highlighted the aspect of inter- and intra-generational equity. While sustainability is generally accepted as an important goal in our time, it is important to clarify the meaning and the different dimensions of the term sustainability as used here. In accordance with the current understanding in the scientific community, we consider four different dimensions of sustainability: technological, economic, social and ecological sustainability.

A. Technological sustainability

Sustainability of technology means that the IT infrastructure can be used and maintained in the long term and does not require any extensive adjustments in its foreseeable life cycle. No matter who is developing it, a sustainable combination of software technologies that are used together (i.e., "tech stack") should therefore be beneficial in the long term.

This also includes a reference data model for a clear database that is compatible with data from other municipalities. Such a data model would have to define which data is recorded by which sensor in which configuration. Even with seemingly straightforward devices such as soil moisture sensors, various critical factors—such

as the depth of installation, soil type, measurement intervals, and calibration—significantly affect the data obtained. Moreover, it delves into the data's transformation processes throughout its lifecycle, including storage practices (data lineage), the mechanisms of data provision, and its semantic description. Finally, the integration of the data into broader metadata portals is essential for maximizing utility and accessibility. By capturing comprehensive, structured metadata, municipalities can ensure data quality, interoperability, scalability, and long-term utility. LoRaWAN is an emerging technology that is helping in technological sustainability across cities, utilities and buildings. It helps in regulating traffic to find a park space in an overcrowded urban area. It increases productivity and efficiency not only for the provider but also for the consumer, for example, to adaptability of power grids in extreme weather conditions. In other words, LoRaWAN has the right blend of market position, technology with the aim of sustainability that gives LoRaWAN an engrossing redeeming feature and appealing aspects.

B. Economic sustainability

Ikerd defines economic sustainability as scarcity, efficiency, and sovereignty [16]. Technology is nowadays an important factor of the economy, businesses and companies to achieve economic sustainability alongside monitoring the operations and increasing efficiency, productivity and to meet needs of the society.

To increase the economic sustainability, private and public companies, universities, and colleges should contribute their knowledge and services. It is important to process the division of labor. Economic efficiency and thrift often meet each other, for example, when reduced consumption of resources and energy correspond with lower financial costs. Sustainability in terms of economy also refers to long-term usability of investments. In the case of building a LoRaWAN infrastructure, a municipality must be convinced that the benefit will exceed the costs of implementation (economic viability). This can be achieved, for example, by reducing personnel costs for manual reading of measured values and other routine activities that can be automated using actuators and sensors.

C. Social sustainability

This dimension of sustainability considers that social equity and cohesion continue to be indispensable for sustainable development. Social sustainability refers to equal opportunities for “good living” and participation in the society for every individual [12]. With LoRaWAN technology in municipalities, the participation of citizens can be strengthened by promoting citizen science projects to utilize the collected data for their own needs, or to improve the local provision of public services.

A socially sustainable LoRaWAN model prioritizes accessibility, benefits all segments of society, and fosters positive community impacts. For example, low-cost connectivity and easy deployment foster the development of local solutions that address specific community challenges, such as agriculture, healthcare, education, and environmental monitoring. Robust data protection measures must be in place to safeguard user privacy and enhance trust in LoRaWAN services. To tailor LoRaWAN installations to the needs and priorities of the community, collaboration with local authorities and community groups is required.

The three dimensions of social, ecological, and economic sustainability are all part of the sustainable development approach. Consequently, they are all represented in the 17 sustainable development goals of the United Nations [26]. The challenge for practitioners is to find integrated solutions to achieve these goals in a holistic way.

D. Ecological sustainability

Ecological sustainability refers to the reduction of consumption and pollution of natural resources and energy. The main goal is to preserve the biosphere [12]. The question here is how technology such as LoRaWAN can be helpful in the protection of natural resources. Examples are the monitoring of environmental data in combination with sensor technology, which can be implemented in the countryside, in parks, and in nature related industries such as agriculture. In the field of agriculture, the monitoring data can be used to reduce herbicides, fertilizers, irrigation and allows precision farming [17]. Other use cases in that field appear when looking at sensors that measure air pollution or traffic dynamics, in order to take measures for improvements. In addition, IoT solutions can be also used for the prediction of extreme weather conditions including heat, flood, storm and wildfire. It is possible to gather data, which has not been measured before or to provide data in a better quality and frequency than before. The data puts the local and regional entities in a better position to take measures of precaution rather than only on repairing damages.

The three dimensions of economic, social, and ecological sustainability are all part of the sustainable development approach established by the United Nations. Some institutions have already shifted from the equality of the three dimensions to a perspective where the SDGs are organized in a hierarchical structure. Following the dependencies of social and economic life on an intact biosphere (see, e.g., [12]), the ecological dimension is prioritized due to building the ground for the social and economic sphere [18]. In line with this, an increasing number of institutions recommend municipal services of general interest on sustainability [12, 19]. The challenge for practitioners is to find integrated solutions for meeting the requirements of these goals in a holistic way.

The scientific discourse on the role of digital technologies in sustainability is ambiguous. On the one hand, digital technologies, including IoT technologies, support progress in sustainability, and on the other hand, they can be part of the problem [11, 17, 19].

Nonetheless, the awareness on the relationship of IoT technology and sustainable development is still limited. An analysis of the World Economic Forum came to the conclusion that, in 2018, 84% of IoT deployments (including the private sector) had the potential to address the Sustainable Development Goals that have already been

addressed. Another conclusion they found was that aligning IoT with development goals did not reduce their commercial viability [20].

E. LoRaWAN Service Management

Information Technology Service Management (ITSM) is a process-focused discipline that is concerned with the efficient and structured delivery and support of IT services [21]. While the concept and its popular implementation, the IT Infrastructure Library (ITIL), have long been known and practiced, no work could be found examining the application of service management principles to the LoRaWAN realm.

LoRaWAN sensors and smart city applications have been thoroughly compiled by Bonilla, Campo Verde and Yoo [22]. The contributions edited by Song, Srinivasan, Sookoor, and Jeschke [23] as well demonstrate the breadth of smart city and IoT applications, but also the need for sustainable operating models and service management.

Zanella et al. [24] found that smart city projects can be complex due to heterogeneous technology (wireless transmission standards, sensors, software architecture), the multitude of different use cases, and the integration of data sources and sinks in order to facilitate diverse digital services.

However, most of these findings are based on case studies in larger cities. While extensive research exists on LoRaWAN infrastructure in rural areas, showcasing applications such as agriculture, smart meters, and fire monitoring. No specific studies could be identified that is concerned with LoRaWAN infrastructure for small towns and rural areas with low population density that still want to leverage the technology.

III. METHOD

We draw our experience from several projects in Germany in which we established LoRaWAN infrastructures, deployed sensors of different types, and created data visualizations. We synthesized our findings from projects in the municipalities Michendorf, Rüdersdorf, and Wiesenburg, and the town Brandenburg an der Havel.

The town of Brandenburg an der Havel is the home of the university where a part of the research team is located. The rural municipalities are nearby, and have contacted us following initial reports of successful deployments. We therefore assume that these are municipalities that are consciously seeking to drive digitalization forward. In some cases, we have already come across acquired funding or ongoing smart city efforts.

Table I illustrates the details of the covered sites. All projects deal with technology transfer in the sense that the university research team applies their knowledge and skills to practical problems of local companies and municipal administrations. Some of the projects are still ongoing, so findings might not be conclusive. Areas of application of the different projects include soil moisture, water temperature, water level, parking spots, presence detection, people counting, and traffic counting.

After clarifying project goals and scope, the project team selected and configured appropriate gateways, sensors, and data visualization platforms. The town or municipality then usually installed the configured sensors themselves. However, a large part of the project duration was spent coordinating with various project participants, organizing site visits, and waiting for service providers, key supporters from within the administration, or decision makers to clarify responsibilities. We had to explain to authorities and network operators that the technology is safe and will not interfere with other radio equipment. Site visits usually took a lot of planning and alignment due to the various ownership structures of buildings, in particular, towers and other tall structures that are already used for other purposes, e. g., sirens and webcams of fire departments, air traffic beacons, and other radio cell systems.

TABLE I. OVERVIEW OF PROJECT SITES

Site	Population	Pop. density (people per sq. km)	Time frame	Deployment	Types of Applications
Brandenburg an der Havel	72,100	320	2022-2024	12 gateways, 25 sensors	weather stations; people counter; water level
Rüdersdorf	15,500	228	2024	2 gateways, ~10 sensors	garbage can levels; water level; indoor temperature, humidity and movements
Michendorf	11,600	202	2023	2 gateways, 38 sensors	traffic counter; parking counter
Wiesenburg	4,900	19	2022-2024	4 gateways, 30 sensors	soil moisture; water level

Funding has been and is a crucial part of every project. Limited short-term funding is usually available, particularly for the procurement of sensors. Due to the way public budgets are planned, there is rarely a permanent funding option for LoRaWAN projects. These are often seen as one-off digitization or technology evaluation projects.

As part of these endeavors, we had frequent talks with all involved stakeholders, such as the municipal administrations, regional utilities, private network suppliers and end-users from other areas such as citizen science and climate initiatives. By accompanying and promoting these processes we learned not only that LoRaWAN projects tend to face similar difficulties in different places, but also that there are many similarities in the needs and capabilities of the stakeholders involved.

Analyzing and conflating these learnings led to the creation of an operational template for LoRaWAN infrastructure and services. This framework names the individual actors as well as their tasks, or responsibilities, in the process of establishing said infrastructure to meet all requirements of the participating stakeholders and especially the end-users.

Before laying out the collaboration framework, we provide a closer look at two of the projects from different perspectives, in order to illustrate our experiences:

A. Project in Wiesenburg

In Wiesenburg, we had the chance to accompany the establishment of LoRaWAN-Infrastructure and many of the strategic processes surrounding it. The initial deployment was driven by a citizen science project aimed at establishing a network of soil moisture measurement stations. Soon, the municipal administration and the office of the Smart City project which Wiesenburg shares with its neighboring municipality of Bad Belzig, saw the potential for cooperation. The municipal administration was interested in utilizing LoRaWAN-Sensors for counting visitors to the local castle and for monitoring water levels in fire ponds and wells. As we accompanied the strategic process, we learned about the challenges of elevating a community-driven network to the purpose of fulfilling statutory duties of the municipal administration, such as different requirements in terms of reliability and data storage. On the other hand, joining forces with the administration provided substantial benefits for the other parties, for example when looking for gateway locations and continuous infrastructure support. The project is still ongoing: The sensor network continues to expand, and different solutions for data storage and visualization are under development as of mid-2024.

B. Project in Michendorf

The Michendorf Project, launched in 2023, aimed to leverage LoRaWAN technology to enhance mobility and optimize public services. The project involved the deployment of 2 gateways and 38 sensors, specifically focusing on park sensors and traffic counting sensors. This section, more detailed than the last one, provides a comprehensive overview of the project's objectives, implementation strategies, costs and economic sustainability, the outcomes achieved and challenges encountered. It also discusses the sustainability implications of the project, including its economic sustainability and its contributions to long-term ecological and social benefits.

1) Objectives

The primary goal is to enhance mobility by improving local solutions through real-time monitoring and data analysis. This includes deploying traffic-counting sensors to collect data on various roadways, as well as implementing parking sensors to monitor the usage of the "Mitnahmebank", a local initiative aimed at optimizing public services. The “Mitnahmebank” service is intended to make it easier for people without a driver's license or their own car to reach the next district outside of bus times. Nineteen colorful benches have been placed near bus stops in the Michendorf municipal area and are clearly marked as “Mitnahmebank”. Anyone who needs a ride can select the desired district from the directional sign on the bench and then sit down. Passing cars traveling in that direction can stop, pick up the waiting person, and drop them off at a mutually agreed destination. Additionally, intuitive data visualization tools are employed to interpret and present the collected data. These tools ensure that the information is accessible and actionable for various stakeholders, including municipal authorities and citizens, facilitating informed and transparent decision-making and effective communication.

2) Implementation

Michendorf Project was executed in several phases, each focusing on a key aspect of implementation. Initially, detailed planning was conducted to identify optimal locations for gateway and sensor deployment, with coordination from local authorities to ensure necessary permissions and support were in place. The project team selected and configured the appropriate hardware, including gateways and sensors, which the municipality installed under the guidance of the university's technical team. To promote ecological sustainability, the devices were installed using existing structures wherever possible, minimizing environmental damage and preventing ecological disturbances. Once installed, the sensors began collecting data. This data was integrated into a central database for structured analysis. Data visualization tools were employed to transform the collected data into accessible and insightful visual representations. These tools facilitated informed decision-making for municipal authorities and ensured that citizens remained well-informed and engaged with the project's outcomes. The project's successful implementation provided enhanced decision-making capabilities to municipal authorities, improved infrastructure efficiency, and fostered community engagement by providing transparent access to the project's insights. By making data-driven decisions, this approach also promotes social sustainability by ensuring public resources are used efficiently and equitably.

3) Cost

The Implementing an IoT network using LoRaWAN technology involves several key cost components. The initial hardware costs include purchasing gateways and sensors, which are essential for data collection and transmission within the network. For this project, a total of 2 gateways and 38 sensors have been acquired. In addition to hardware, managing already connected IoT devices through platforms like The Things Network (TTN) incurs costs related to subscription or usage fees, particularly as the network grows to include more devices. Hosting a LoRaWAN Network Server (LNS) also adds financial considerations, encompassing server infrastructure expenses, such as hosting fees for cloud or physical servers, along with ongoing operational costs like maintenance, security, and technical support. Michendorf is now actively looking for an affordable hosted LoRaWAN Network Server (LNS).

LoRaWAN devices are designed to be energy-efficient, which can significantly reduce operational costs, especially when deployed in remote or hard-to-reach areas. The technology's long-range capabilities minimize the need for multiple gateways, further lowering infrastructure expenses. LoRaWAN networks are also highly scalable, allowing new devices to be added with minimal additional investment. This scalability ensures that costs remain aligned with network growth, supporting economic sustainability. Using

an Internet of Things platform like The Things Network provides a set of open tools and resources for low-cost device management. This platform enables efficient network administration and reduces software expenses. If Michendorf can effectively leverage the low-cost, long-range, and scalable nature of LoRaWAN technology, while also securing affordable hosted LoRaWAN Network Server solutions and optimizing operational efficiencies, the implementation can achieve economic sustainability.

4) Results

In Michendorf project traffic data are collected and analyzed to gain a comprehensive understanding of the traffic in the street network. Traffic counting sensors monitor and record the number of pedestrians, two wheelers, cars, and heavy vehicles traveling on the street. These sensors are installed across various locations, including school zones, main thoroughfares, and bicycle areas. The traffic data can additionally be analyzed in relation to factors like school holidays, weather conditions, social events, and more. For example, Figure 1 illustrates the daily traffic data overview for street "Stückener Dorfstraße". The graph indicates that traffic counts were notably high on November 4th, due to the open house event at the fire station on this street, which attracted a large number of guests.

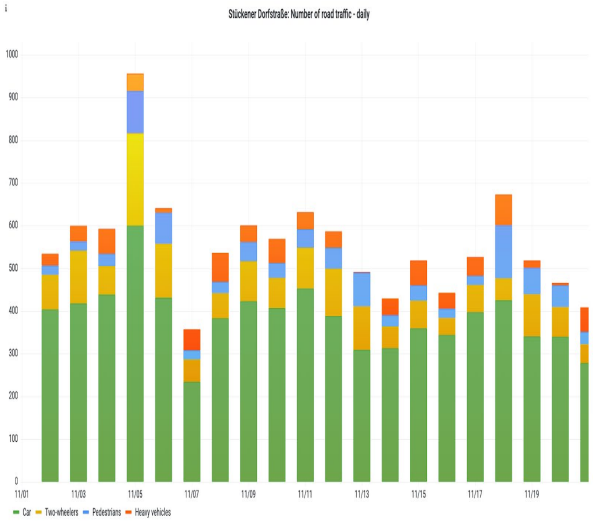


Figure 1. Daily traffic data for street: Stückener Dorfstraße

The use of park sensors helps to assess the efficiency of the "Mitnahmebank", enabling adjustments to enhance public services for the benefit of both citizens and local businesses. Figure 2 illustrates the frequency of use of the "Mitnahmebank".

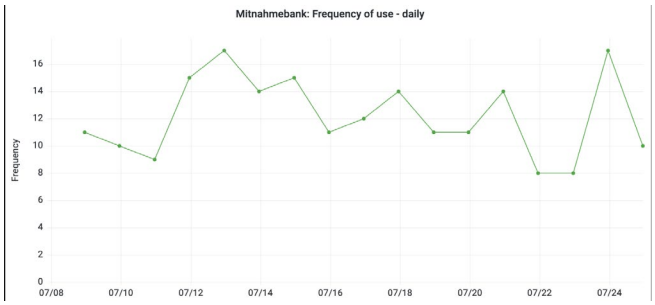


Figure 2. Mitnahmebank: Frequency of use - daily

5) Challenges

Challenges in the project included ensuring clear and continuous communication among all stakeholders, which was crucial for keeping the project on track: This required regular updates and meetings. Proper calibration of sensors was another significant challenge, as it was essential for obtaining accurate data. This necessitated meticulous attention during installation and regular maintenance checks. Additionally, managing large volumes of data and ensuring its accuracy and reliability was a significant challenge, requiring robust data management strategies.

IV. LoRaWAN Collaboration Framework (LCF)

The "LoRaWAN Collaboration Framework" (LCF) serves as a blueprint for effective stakeholder collaboration in the deployment and maintenance of sustainable LoRaWAN infrastructures for communities. It delineates the roles, responsibilities, and necessary capabilities for each participant, ensuring that all involved parties understand what is required of them and what they need from others. This clarity facilitates not only the identification of existing and potential contributors to the LoRaWAN ecosystem but also the establishment of seamless interfaces and partnerships. By highlighting specific needs and capabilities across stakeholders, the LCF aims to streamline operations, foster innovation, and enhance service delivery and citizen engagement within the realm of LoRaWAN-enabled services. Figure 2 illustrates the structure of the LCF and provides details on each role.

	Infrastructure		Application		
	Network operators and utility companies	Hosting and IT service providers	Start-ups, universities, and specialized IT companies	Administrations	End-users
Wants & Needs what stakeholders want and need	• Market expansion • New revenue streams and business models	• Stable contracts and partnerships • Maintainenance-friendly software components • Opportunities to showcase and deploy new technologies • Support by hardware and software vendors	• Evaluating cutting-edge technology • Funding for R&D or transfer projects • Partnerships • Visibility and recognition	• Education on technology options • Fast and efficient delivery of services to citizens • Enhanced citizen engagement and satisfaction • Limited Total Cost of Ownership (TCO) • User feedback on provided services	• Convenient access to municipal services • Intuitive interfaces and good user experience • Integration with other IT systems • Security features to protect user data and privacy • Regular updates and improvements
Capabilities what stakeholders are able to contribute to the LoRaWAN value chain	• Provide sites and connectivity for gateways • Operate the LoRaWAN Network Server (LNS) • On-site servicing of sensors (batteries, cleansing, replacement)	• Provide computing resources (hardware, software, virtual machines, and networks) • Install and configure standard software packages • Configure sensors • Backups & security • Customer support and training for municipal staff and end-users	• Build innovative prototypes • Find or build the right sensors for use cases • Develop custom software • Integrate custom data sources and platforms • Create machine learning models • Support and train end-users	• Problems / use cases • Provide sites for gateways • Deploy sensors • Budget or funding for public IT projects • Collaboration with private entities and other government agencies	• Problems / use cases • Give feedback on services

Figure 3. LoRaWAN Collaboration Framework (LCF) with stakeholders and their responsibilities within LoRaWAN projects.

A. Framework structure: Roles, needs and capabilities

The stakeholders are grouped by "infrastructure" and "application". The infrastructure group is mainly concerned with deployment, operation, and maintenance of the hardware, software, and networking infrastructure. The "application" group, which also includes the end-users, represents the rest of the value chain, including implementation and customization of software and data platforms, deployment of sensors and the provision of the actual use cases or problems to be solved. Each stakeholder group takes on a role in the LoRaWAN ecosystem. The interaction of all roles and the correct distribution of tasks is critical to the success of LoRaWAN projects.

To define each role, the framework is split into "wants and needs" (top row) and the "capabilities" (bottom row) of stakeholders. The wants and needs of one stakeholder group should correspond to the capabilities of another stakeholder group. This creates a balanced system of responsibilities that should make sure projects can be executed successfully. The roles are assigned to columns, which are ordered on a spectrum from technological necessities (left) to an end-user application (right).

By clearly defining the roles and interfaces, the LCF not only ensures that each stakeholder's capabilities are effectively utilized but also fosters a cooperative and dynamic ecosystem. The framework supports the strategic alignment of wants and needs with available capabilities, facilitating a balanced and effective collaboration. The stakeholder roles are briefly described in the following.

1) Network operators and utility companies: Network operators and utility companies are driven by goals of market expansion and the development of new revenue streams. These stakeholders are adept at providing sites and connectivity for gateways, managing LoRaWAN Network Servers (LNS), and conducting essential on-site maintenance such as battery replacement and cleansing of sensors.

2) Hosting and IT service providers: On the technological service front, hosting and IT service providers aim for stable, long-term contracts and opportunities to deploy emerging technologies for their customers. They provide vital capabilities such as the provisioning of computing resources, software installation, sensor configuration, and rigorous data security measures. Moreover, they can provide support and training for municipal staff and end-users, ensuring smooth operation and adoption of technologies.

3) Startups, universities, and specialized IT companies: In the innovation and research sector, startups, universities, and specialized IT companies are focused on evaluating and implementing cutting-edge technologies. These entities are key in building innovative prototypes, selecting or creating appropriate sensors for specific use cases, and developing customized software solutions. They can also handle the integration of custom data platforms and are instrumental in developing advanced machine learning models to support complex data analysis and decision-making processes based on the collected sensor data, e.g., for predictive maintenance applications.

4) Administrations: Administrative bodies concentrate on delivering efficient and enhanced services to citizens while maintaining cost-effectiveness. Their role includes defining clear use cases and overseeing sensor deployment. They might also secure locations for gateway installations. Their collaboration with private sectors and other government agencies is crucial for securing the necessary budget and support for public IT projects, which is essential for sustained technological advancement and community service enhancement.

5) End users): T End-users, crucial to the success of the whole LoRaWAN value chain, require easy access to municipal services via intuitive and potentially mobile-friendly user interfaces, seamless integration with existing IT systems, and robust security features to protect their data and privacy. Their ongoing feedback is instrumental in driving the continuous refinement and user-centered optimization of services.

B. Sharing and swapping responsibilities

Some capabilities in the LoRaWAN value chain can be provided by several stakeholders.

For example, providing sites for gateway installations might be a task a communal administration might want to contribute to in a project. However, ensuring long-term connectivity at the site via wired or wireless connections, having trained maintenance staff on standby in case of breakdowns, and having constant access to necessary equipment (lift trucks, spare parts, etc.) are things which network operators or utility companies have established processes for, resulting in lower cost and higher quality of service.

Another common example is the installation of affordable sensors by citizens configured on municipal or public LNS. The existence of tech-savvy communities and individuals can be considered a substantial benefit for any municipality or town. But it might jeopardize the long-term support of these sensors and by that, data quality. If administrations want to rely on the collected data, there must be some kind of alignment and trust between the two groups.

A last example illustrates another problem of voluntary work. When volunteers create custom data integration layers using Python or JavaScript, this "glue code" might not be documented as required, complicating maintenance and future extensions (e.g., a "temperature" attribute is added to the next sensor model). This also includes simple things like patch management and storage of access credentials.

The same goes for configuring dashboards, providing end-user training, integrating data sources, and creating machine learning models. All these tasks can be handled by

different stakeholders with varying degrees of quality, cost, and availability. Initial interest in certain stages of the value chain by any one party does not guarantee that all the required tasks are fulfilled by the role. So, for a stable and sustainable operation, we recommend the responsibilities as componentsd in the LCF, or at most, one "column" away from the original stakeholder group.

In summary, although some capabilities might be taken over by another party than designated in the framework, generally, this hurts sustainability.

C. Business models

A solid business model is needed to sustain a LoRaWAN infrastructure and application ecosystem. The identified "wants and needs" indicate a demand in the market, while the "capabilities" are potential services which satisfy needs in one of the following elements of the value-chain.

1) Supply-side business models: One solution is renting out the network on a per-sensor and per-time basis, thus creating a very low barrier to market entry for customers and allowing for rapid adoption of the service. However, like free Wi-Fi, this might eventually become a commodity and network operators need to find other ways to generate revenue. Netwrok operations and suppliers can capitalize on existing infrastructures, such as data centers and network backbones, to gain a competitive edge. Additionally, revenue generation extends beyond network access fees to include value-added services like sensor commissioning. R&D stakeholders contribute by offering scientific and technical support, optimizing gateway placements, selecting appropriate sensors, and providing custom data integrations, visualizations, and project management, adding significant value to the LoRaWAN ecosystem. Value, of course, is understood differently by different stakeholders. When compared to commercial projects, public projects are rather focused on social and ecological sustainability. This might mean making it possible for citizens to participate in local decision-making like defining speed limits, deciding on the desired quality of air and water, increasing comfort with digital services, or improving public health and safety. In conclusion, the business model of the supply-side of municipal LoRaWAN projects today relies on forward-thinking administrations which actively seek to contribute to achieving the SDGs by using advanced technology like LoRaWAN. In the future, the collection of such data could become a legal requirement. Only then are more local authorities likely to look for joint operating models.

2) Demand-side business models: A shared operations model would be beneficial for small-scale deployments like the ones we described above. For example, a properly set-up LoRaWAN network server (LNS) can easily process data packages from several hundred gateways. Each gateway is technically capable of supporting thousands of LoRaWAN nodes, i.e., sensors and actuators. Sharing the infrastructure costs would therefore be an obvious way to achieve sustainable funding. The problem with this approach in a municipal setting is twofold. First, administrations need to align their demands and timing, and agree on a fair share of the (still) required funding. So, there is a cost for coordinating interested parties. Second, the infrastructure needs to be installed and administered in the partnering regions by the same operator. When these challenges can be overcome, e.g., by applying systematic project and stakeholder management and finding a way to align the diverse interests, there is potential for a low-cost infrastructure that benefits all the stakeholders and thereby provides a holistic societal value.

D. Decision making

The LoRaWAN Collaboration Framework gives an overview on the different stakeholders and roles required to operate a LoRaWAN infrastructure in rural municipalities or rural towns efficiently. In addition, some other questions arise in municipalities, for example, considering potential costs and benefits from a local LoRaWAN infrastructure before deciding to implement such an infrastructure. The following questions and arguments mainly refer to the presented different dimensions of sustainability in section II.

1) Is there a good balance on costs and benefits?: This question addresses economic sustainability: There are numerous different use cases on LoRAWAN technology already practiced, and a lot more may arrive in the future. Use cases can be found for example in the fields of smart cities & smart regions, energy and resource monitoring, disaster control and environmental data collections [17, 25]. To argue for an investment in LoRaWAN infrastructure with public funds in the logic of economic sustainability, one needs to take into account the number of use cases relevant for a region. A larger number of use cases can justify investment costs, in particular at the beginning. Due to that, many stakeholders like municipal utilities (municipal or district level), companies (e.g., waste collection companies, agricultural farms), the civil society (as fire departments, nature conservation groups) and similar stakeholders should be asked for their use cases. The more parties that will share the infrastructure and benefits, the easier a return of the investment can be reached.

2) Can the technology address relevant community problems?: This question relates to ecological and social sustainability: In addition to considerations on the balance of financial costs and benefits, another relevant focus is on benefits for the common good provided by the LoRaWAN infrastructure, which cannot be measured in financial terms. For example, if the technology helps to protect local buildings and citizen's lives due to improved disaster control or improved prediction and monitoring of extreme

weather events, the financial costs for implementing it may be high.

3) Is the technology itself harmful for the environment?*:* Although LoRaWAN and sensor infrastructure are characterized by low energy consumption, it should not be overlooked that resources are required to build and operate them. Considerations on where the hardware is manufactured – and under what kind of working conditions – arise. It should also be noted that servers consume a lot of power, which should be taken into account as an invisible factor, although it is hard to meter [25]. These are arguments not to implement the technology if the expected normative value cannot compensate for these negative effects. In order to reduce the energy consumption, there are several viable steps that might be taken: For the case of the measuring sensors it is possible to operate the sensor kit with a solar or photovoltaic panel instead of a battery, and for the case of the servers it is possible to use server farms operation with power from renewable energies.

V. CONCLUSION

The proposed LoRaWAN Collaboration Framework provides a solid foundation for municipalities to successfully set up LoRaWAN projects. By following the framework, municipalities can search for and align with partners, knowing which responsibilities need to be covered by those partners. They now have proper criteria for selecting partners like hardware vendors, utilities, citizen science projects, and innovation leaders from higher education and start-up ecosystems.

A key finding from our projects was that collaboration is vital to successful LoRaWAN deployments in rural areas and small municipalities. Not only allows collaboration access to good practices, but it will also provide an opportunity to pool financial resources and use cases for an efficient acquisition of infrastructure and partners. While we were in the role of technical project management for all the projects, we learned that a more holistic approach is necessary to create a sustainable LoRaWAN deployment.

Further research needs to be done on the economic viability of the described business models. In particular, the shared operation models and its organizational and coordinative prerequisites as well as the proper involvement of citizen initiatives and individual volunteers.

According to the different dimensions of sustainability discussed in this paper, further questions arise for rural municipalities and towns to decide for or against the implementation of LoRaWAN infrastructures. Apart from measurable cost and benefit scenarios, there are also arguments that do not fit the financial cost perspective and some other arguments refer to partially not visible and measurable factors, which should be taken into account in a holistic view on sustainability.

A comprhensive analysis on the overall project outcomes is needed to validate that the ecological, economic, technical, and social objectives are in the desired balance. Because some sustainability dimensions are hard to measure, it must always be carefully weighed up which use cases can realistically make improvements and which are just greenwashed fig leaves. Further research on metering sustainability costs of IoT infrastructures are recommended to lower the remaining uncertainties.

ACKNOWLEDGMENT

This work was supported in part by the German Federal Ministry of Education and Research (grant code 13IHS230A) as part of the “Innovative Hochschule” project.

REFERENCES

[1] A. Nitze, T. Wang, J. Jahn, and S. Ali, "Beyond Connectivity: A Sustainable Approach to Municipal LoRaWAN Infrastructure and Services," ICDS 2024: The Eighteenth International Conference on Digital Society.

[2] E. B. Weiss, “In fairness to future generations and sustainable development”. American University International Law vol. 8(1), pp. 19-26, 1992.

[3] K. A. Emina “Sustainable development and the future generations”, Social Sciences, Humanities and Education Journal (SHE Journal), vol. 2(1), pp. 57-71, Jan. 2021, doi: 10.25273/she.v2i1.8611.

[4] B. O. Linnér and H. Selin, “The United Nations Conference on Sustainable Development: forty years in the making” Environment and Planning C: Government and Policy, vol. 31(6), pp. 971-987, 2013, doi: 10.1068/c12287.

[5] M. Bexell and K. Jönsson, “Responsibility and the United Nations’ sustainable development goals”. Forum for development studies, vol. 44(1), pp. 13-29, Jan. 2017, doi: 10.1080/08039410.2016.1252424.

[6] A, Szymańska,. “Reducing Socioeconomic Inequalities in the European Union in the Context of the 2030 Agenda for Sustainable Development”. Sustainability, 2021, vol. 13(13), 7409, doi: 10.3390/su13137409.

[7] R. Rana, Y. Singh, and P. K. Singh, "A systematic survey on the internet of things: energy efficiency and interoperability perspective," Trans. Emerg. Telecommun. Technol., vol. 32(8), pp. 1-41, 2021, doi:10.1002/ett.4166.

[8] LoRa-Alliance, "LoRa-Alliance," [Online]. Available: https://lora-alliance.org. [Retrieved April 3, 2024].

[9] BusinessDictionary.com, "Enabling Technology," September 2020 [Online]. Available: https://web.archive.org/web/20200926020742/http://www.businessdictionary.com/definition/enabling-technology.html. [Retrieved April 3, 2024].

[10] B. Nölting and N. Dembski, "Digitalization for sustainable management and corporate sustainability management," in Corporate Sustainability Management, A. Baumast and J. Pape, Eds. pp. 405-422, 2022.

[11] Kompetenzzentrum Öffentliche IT (Competence center public IT) "Wertebasierte Digitalisierung für nachhaltige Entwicklung im öffentlichen Sektor, 7" („Value-based digiatlization for sustainable development in the public sector“), 2023 [Online]. Available: https://www.oeffentliche-it.de/documents/10181/14412/Wertebasierte+Digitalisierung+f%C3%BCr+nachhaltige+Entwicklung+im+%C3%B6ffentlichen+Sektor [Retrieved August 12, 2024].

[12] S. Beer, N. Wulf, and M. Großklaus, "Sustainable-digital local supply. Development perspectives and recommendations for action for a reorientation of the municipal supply mandate,", Sep. 2022. [Online] Available: https://codina-transformation.de/wp-content/uploads/CODINA_Forschungslinienbericht_Daseinsvorsorge.pdf. [Retrieved April 5, 2024].

[13] J. Quaing, J. Fink, B. Bilfinger, F. Vorländer. „Doppelte Transformation gestalten – Praxisleitfaden Nachhaltigkeit und Digitalisierung“ („Shaping double transformation – practical manual sustainability and digitalization“), oekom verlag, 2023, doi: 10.14512/9783962389444.

[14] M. Farsi, A. Hosseinian-Far, A. Daneshkhah, and T. Sedighi, "Mathematical and Computational Modelling Frameworks for Integrated Sustainability Assessment (ISA)," in Strategic Engineering for Cloud Computing and Big Data Analytics, A. Hosseinian-Far, M. Ramachandran, and D. Sarwar, Eds. Springer, pp. 3-27, 2017.

[15] United Nations, "Report of the World Commission on Environment and Development. Our Common Future", 1987. [Online] Available: https://digitallibrary.un.org/record/139811?v=pdf. [Retrieved April 5, 2024].

[16] J. Ikerd, The Essentials of Economic Sustainability, Kumerian Press, Sterling, VA, 2012.

[17] Technopolis and IÖW (Eds.), „Metastudie „Nachhaltigkeitseffekte der Digitalisierung“. Eine Auswertung aktueller Studien zur (quantitativen) Bemessung der Umwelteffekte durch die Digitalisierung“ („Meta-study „sustainability effects of digitalization“, an evaluation of contemporary studies fort he quantification of enviromnental effects of digitalization“), Berlin, 2024 Available: https://www.ioew.de/fileadmin/user_upload/BILDER_und_Downloaddateien/Publikationen/2024/Technopolis-IOEW_2024-Metastudie_Nachhaltigkeitseffekte-der-Digitalisierung.pdf [Retrieved August 12, 2024].

[18] Stockholm Resilience Centre, “Stockholm Resilience Centre’s contribution to the 2016 Swedish 2023 Agenda HLPF report” 2017, [Online] Available: https://www.stockholmresilience.org/download/18.2561f5bf15a1a341a523695/1488272270868/SRCs%202016%20Swedish%202030%20Agenda%20HLPF%20report%20Final.pdf [Retrieved August 12, 2024].

[19] German Advisory Council on Global Change, "Main report. Our shared digital future," p. 2. 2019, Available: https://www.wbgu.de/fileadmin/user_upload/wbgu/publikationen/hauptgutachten/hg2019/pdf/wbgu_hg2019.pdf. [Retrieved April 5, 2024].

[20] World Economic Forum, "Future of Digital Economy and Society Systems Initiative. Internet of Things. Guidelines for Sustainability", 2018, [Online] available: https://www3.weforum.org/docs/IoTGuidelinesforSustainability.pdf [Retrieved August 12, 2024].

[21] S. D. Galup, R. Dattero, J. J. Quan, and S. Conger, "An overview of IT service management," Communications of the ACM, vol. 52 (5), pp. 124-127, 2009.

[22] V. Bonilla, B. Campoverde, S. G. Yoo, "A Systematic Literature Review of LoRaWAN: Sensors and Applications," Sensors, vol. 23, pp. 8440, 2023, doi: 10.3390/s23208440.

[23] H. Song, R. Srinivasan, T. Sookoor, and S. Jeschke, Eds., Smart Cities: Foundations, Principles, and Applications. John Wiley & Sons, 2017.

[24] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," in IEEE Internet of Things Journal, vol. 1(1), pp. 22-32, Feb. 2014, doi: 10.1109/JIOT.2014.230632.

[25] Umweltbundesamt (Federal environment agency), „Digitalisierung nachhaltig gestalten. Ein Impulspapier des Umweltbundesamts“ (“Shaping digitalization sustainably: A discussion paper of the federal environment agency”), 2019, [Online] Available: https://www.umweltbundesamt.de/sites/default/files/medien/376/publikationen/uba_fachbroschuere_digitalisierung_nachhaltig_gestalten_0.pdf [Retrieved August 12, 2024].

[26] United Nations, "The 17 Sustainable Development Goals" [Online]. Available: https://sdgs.un.org/goals. [Retrieved April 3, 2024].

[27] A. H. Maslow, Motivation and personality. New York: Harper and Row, 1954.

Relations Between Entity Sizes and Error-Correction Coding Codewords and Effective Data Loss

Ilias Iliadis
IBM Research Europe – Zurich
8803 Rüschlikon, Switzerland
email: ili@zurich.ibm.com

***Abstract*—Erasure-coding redundancy schemes are employed in storage systems to cope with device and component failures. Data durability is assessed by the Mean Time to Data Loss** (MTTDL) **and the Expected Annual Fraction of Entity Loss** (EAFEL) **reliability metrics. In particular, the** EAFEL **metric assesses losses at an entity, say file, object, or block level. This metric is affected by the number of codewords that entities span. The distribution of this number is obtained analytically as a function of the size of the entities and the frequency of their occurrence. The deterministic and the random entity placement cases are investigated. It is established that for certain deterministic placements of variable-size entities, the distribution of the number of codewords that entities span also depends on the actual entity placement. To evaluate the durability of storage systems in the case of variable-size entities, we introduce the Expected Annual Fraction of Effective Data Loss** (EAFEDL) **reliability metric, which assesses the fraction of stored user data that is lost by the system annually at the entity level. The** MTTDL**,** EAFEL**, and** EAFEDL **metrics are assessed analytically for erasure-coding redundancy schemes and for the clustered, declustered, and symmetric data placement schemes. These metrics are derived in closed-form for the case of lazy rebuilds and in the presence of correlated latent symbol errors. It is demonstrated that an increased variability of entity sizes results in improved** EAFEL**, but degraded** EAFEDL**. It is established that both reliability metrics are adversely affected by the size of the erasure-coding symbols. The** EAFEL **and** EAFEDL **reliability metrics are evaluated for some real-world erasure coding schemes employed by enterprises. The analytical reliability expressions derived can identify efficient erasure coding schemes and can be used to dimension and provision storage systems to provide desired levels of durability.**

Keywords–Storage; Reliability analysis; MTTDL; EAFDL; EAFEL; EAFEDL; MDS codes; Unrecoverable or latent symbol errors; Deferred recovery or repair; stochastic modeling.

I. INTRODUCTION

The durability of data storage systems and cloud offerings is affected by device and component failures [1]. Desired reliability levels are ensured by employing erasure-coding redundancy schemes for recovering lost data [2-5].

The frequency of data loss events is assessed by the Mean Time to Data Loss (MTTDL) metric that has been widely used to assess the reliability of storage systems [4][5]. Also, the amount of data loss is obtained by the Expected Annual Fraction of Data Loss (EAFDL) metric that was introduced in [6]. This metric was recently complemented by the Expected Annual Fraction of Entity Loss (EAFEL) metric [7]. The EAFEL metric assesses data losses at an *entity*, say file, object, or block level, whereas the EAFDL metric assesses data losses at a lower data processing unit level.

The smallest accessed unit of a storage device is a *sector* in Hard-Disk Drives (HDDs), a *page* in flash-based Solid-State Drives (SSDs), and a *data set* in Linear Tape-Open (LTO is the trademark of HP, IBM, and Quantum in the Unites States and other countries) tape systems [8]. A sector has a typical size of 512 bytes or 4 KB, a page has a size that ranges from 4 KB to 16 KB, and a data set currently has a size of 5 MB or more. Erasure-coding redundancy schemes are implemented by treating the units that contain user data as symbols and complementing them with parity symbols (units) to form codewords. In the case of HDDs and SSDs, one or more units are allocated to an entity and the last unit may be partially filled. Depending on the file system employed, the remaining space of a partially-filled unit may or may not be used to store the contents of another entity. Therefore, user data may or may not be stored in an aligned fashion with units (symbols), which in turn implies that entities may or may not be aligned with codewords. The case where entities are aligned with codewords was considered by the reliability model presented in [7]. By contrast, in the case of tape, user data is written sequentially such that a unit may contain data of multiple entities. Therefore, user data and entities are not aligned with symbols and codewords, respectively. Moreover, the reliability model presented in [7] assumed that entities have a fixed size, whereas in practice they have variable sizes. It turns out that the MTTDL metric does not depend on the placement and size of the entities, but the EAFEL metric does. More specifically, EAFEL depends on the number of codewords that stored entities span. Furthermore, the EAFEL metric reflects the fraction of lost user data only when entities have a fixed size. To evaluate system durability in the case of variable-size entities, in this article we introduce the Expected Annual Fraction of Effective Data Loss (EAFEDL) reliability metric, that is, the fraction of stored user data that is expected to be lost by the system annually at the entity level.

The key contributions of this article are the following. The reliability model presented in [7] for the assessment of the EAFEL metric is enhanced in two ways. First, entities are considered to be stored such that they are not aligned with codeword boundaries. Second, the size of entities is considered to be variable. The objective of this article is to assess system reliability by deriving the distribution of the number of codewords that entities span. We address the following question. Does this distribution only depend on the statistics of the entities stored, that is, on their size and frequency of occurrence, or does it also depend on their placement? In the present work, we shed light on this issue by investigating the cases of deterministic and of random entity placement. The

distribution of the number of codewords that entities span is obtained analytically as a function of the size of the entities and the frequency of their occurrence. We also establish that for certain deterministic placements of variable-size entities, this distribution also depends on the actual entity placement.

The general non-Markovian methodology that was applied in prior work to assess the EAFDL and EAFEL metrics for erasure-coding redundancy schemes and for the clustered, declustered, and symmetric data placement schemes, was extended to derive analytically the EAFEL and the new EAFEDL reliability metrics for the case of variable-size entities [1]. It was demonstrated how the erasure-coding capability as well as the entity and symbol sizes affect system reliability in the entire range of bit error rates. In this article, we extend our previous work by deriving MTTDL for the case of lazy rebuilds and in the presence of correlated latent symbol errors. We also evaluate the EAFEL and EAFEDL reliability metrics for some real-world erasure coding schemes employed by enterprises. The model developed provides useful insights into the benefits of the erasure coding schemes and yields results for the entire parameter space, which allows a better understanding of the design tradeoffs.

The remainder of the article is organized as follows. Section II reviews prior relevant work and analytical models presented in the literature for assessing the effect of latent errors on the reliability of erasure-coded systems. Section III describes the storage system model and the corresponding parameters considered. In Section IV, the distribution of the number of codewords that entities span is derived analytically as a function of the entity size distribution when entities are not aligned with symbols and when entity sizes are either fixed or variable. In Section V, the MTTDL metric is derived analytically for the case of lazy rebuilds and correlated latent symbol errors. Also, the EAFEL and EAFEDL metrics are derived analytically for the case of random placement of variable-size entities. Section VI presents numerical results demonstrating the effect of the erasure-coding capability and of the entity sizes on system reliability, as well as the adverse effect of an increased symbol size. The reliability of real-world erasure coding schemes employed by enterprises to protect their stored data is assessed in Section VII. Finally, we conclude in Section VIII.

II. Related Work

Analytical reliability expressions for MTTDL that take into account the effect of latent errors have been obtained predominately using Markovian models, which assume that component failure and rebuild times are independent and exponentially distributed [9][10][11][12]. The effect of latent errors on MTTDL and EAFDL of erasure-coded storage systems for the realistic case of non-exponential failure and rebuild time distributions was assessed in [4][5].

Disk scrubbing has been used to mitigate the adverse effect of latent errors on system reliability [9][13][14][15]. The scrubbing process identifies latent errors at an early stage and attempts to correct them before disk failures occur. This in effect reduces the probability of encountering a latent error during the rebuild process. The resulting latent-error probability was derived in [9] as a function of the scrubbing and workload parameters. Subsequently, it was shown that the reliability level achieved when scrubbing is used can be obtained from the reliability level of a system that does not use scrubbing by adjusting the probability of encountering a latent error accordingly. The methodology presented in [9] for deriving the adjusted latent error probability when scrubbing is employed is also applicable for assessing the efficiency of other scrubbing schemes, such as the adaptive scrubbing schemes proposed in [14][15]. Moreover, this methodology can also be applied in conjunction with the reliability results presented in this article to assess the reliability of erasure-coded systems when scrubbing is used.

The efficiency of applying erasure coding in storage systems that employ solid state disks (SSDs) was studied in [16]. It was demonstrated that the reliability improvement achieved by erasure coding is in general greater than the reliability degradation induced. Also, the reliability of SSD arrays using a real-system implementation of conventional and emerging erasure codes was investigated in [17] using realistic storage traces.

A simulation analysis of reliability aspects of erasure-coded data centers was presented in [18]. Various configurations were considered and it was shown that erasure codes and redundancy placement affect system reliability. In [19] it was recognized that it is hard to get statistically meaningful experimental reliability results using prototypes, because this would require a large number of machines to run for years. This underscores the usefulness of the analytical reliability results derived in this article.

III. Storage System Model

The reliability of erasure-coded storage systems was assessed in [7] based on a model that considers codeword rebuilds for reconstructing lost symbols and assess system reliability when entities (files, objects, blocks) are lost. Maximum Distance Separable (MDS) erasure codes (m, l) that map l user-data symbols to codewords of m symbols are employed. They have the property that any subset containing l of the m codeword symbols can be used to reconstruct (recover) a codeword. The MTTDL and EAFEL reliability metrics were derived analytically for systems that employ a lazy rebuild scheme.

The corresponding storage efficiency s_{eff} and amount U of user data stored in the system is

$$s_{\text{eff}} = l/m \quad \text{and} \quad U = s_{\text{eff}}\, n\, c \;=\; l\, n\, c/m\;, \qquad (1)$$

where n is the number of storage devices in the system and c is the amount of data stored on each device. The storage space of devices is partitioned into units (symbols) of a fixed size s, such that the number C of symbols stored in a device is

$$C = c/s\;. \qquad (2)$$

Our notation is summarized in Table I. The parameters are divided according to whether they are independent or derived and are listed in the upper and lower part of the table, respectively.

To minimize the risk of permanent data loss, the m symbols of each of codeword are spread and stored in m devices. This

TABLE I. NOTATION OF SYSTEM PARAMETERS

Parameter	Definition
n	number of storage devices
c	amount of data stored on each device
l	number of user-data symbols per codeword ($l \geq 1$)
m	total number of symbols per codeword ($m > l$)
(m, l)	MDS-code structure
e_s	entity size
s	symbol (sector or data set) size
k	spread factor of the data placement scheme, or group size (number of devices in a group) ($m \leq k \leq n$)
b	average reserved rebuild bandwidth per device
B_{max}	upper limitation of the average network rebuild bandwidth
X	time required to read (or write) an amount c of data at an average rate b from (or to) a device
$F_X(.)$	cumulative distribution function of X
$F_\lambda(.)$	cumulative distribution function of device lifetimes
P_{b}	probability of an unrecoverable bit error
s_{eff}	storage efficiency of redundancy scheme ($s_{\text{eff}} = l/m$)
U	amount of user data stored in the system ($U = s_{\text{eff}}\, n\, c$)
$\tilde{r}$	MDS-code distance: minimum number of codeword symbols lost that lead to permanent data loss ($\tilde{r} = m - l + 1$ and $2 \leq \tilde{r} \leq m$)
C	number of symbols stored in a device ($C = c/s$)
μ^{-1}	mean time to read (or write) an amount c of data at an average rate b from (or to) a device ($\mu^{-1} = E(X) = c/b$)
λ^{-1}	mean time to failure of a storage device ($\lambda^{-1} = \int_0^\infty [1 - F_\lambda(t)]dt$)
P_s	probability of an unrecoverable sector (symbol) error
s_s	shard size ($s_s = e_s/l$)
J	shard size measured in symbol-size units ($J = s_s/s = e_s/(l\, s)$)
Y	number of lost entities during rebuild
$\breve{Q}$	amount of lost user data during rebuild

way, the system can tolerate any $\tilde{r} - 1$ device failures, but $\tilde{r}$ device failures may lead to data loss, with

$$\tilde{r} = m - l + 1\,, \quad 1 \leq l < m \quad \text{and} \quad 2 \leq \tilde{r} \leq m\,. \tag{3}$$

Examples of MDS erasure codes are the following:

Replication: A replication-based system with a replication factor r can tolerate any loss of up to $r - 1$ copies of some data, such that $l = 1$, $m = r$ and $\tilde{r} = r$. Also, its storage efficiency is equal to $s_{\text{eff}}^{\text{replication}} = 1/r$. The mirroring scheme is the special case where $r = 2$. The corresponding storage efficiency of only 50% can be improved by employing erasure codes.

RAID-5: A RAID-5 array comprised of N devices uses an $(N, N-1)$ MDS code, such that $l = N - 1$, $m = N$ and $\tilde{r} = 2$. It can therefore tolerate the loss of up to one device, and its storage efficiency is equal to $s_{\text{eff}}^{\text{RAID-5}} = (N-1)/N$.

RAID-6: A RAID-6 array comprised of N devices uses an $(N, N-2)$ MDS code, such that $l = N - 2$, $m = N$ and $\tilde{r} = 3$. It can therefore tolerate a loss of up to two devices, and its storage efficiency is equal to $s_{\text{eff}}^{\text{RAID-6}} = (N-2)/N$.

In terms of encoding operations, MDS erasure codes are either bitwise exclusive-OR (XOR) or non-XOR. The computation complexity of the non-XOR-based codes, such as Reed–Solomon, is much higher than that of the XOR-based ones. Also, in the context of storage, Reed-Solomon codes are preferable to Turbo codes owing to their simpler implementation and the fact that they are more suitable in environments where bit error rates are low, and errors occur in bursts.

Two different ways (A and B) for storing user data on devices were shown in Figure 1 of [7]. According to way A, user data contained in entities is divided into chunks with the contents of a chunk stored on different devices, whereas according to way B, user data contained in entities is divided into *shards* with the contents of a shard stored on the same device. More specifically, according to way B, user data contained in entities is divided into l shards with each one being stored on a different device, as shown in Figure 1(a). Entities were assumed to have a fixed size e_s with the corresponding shard size s_s then obtained by $s_s = e_s/l$.

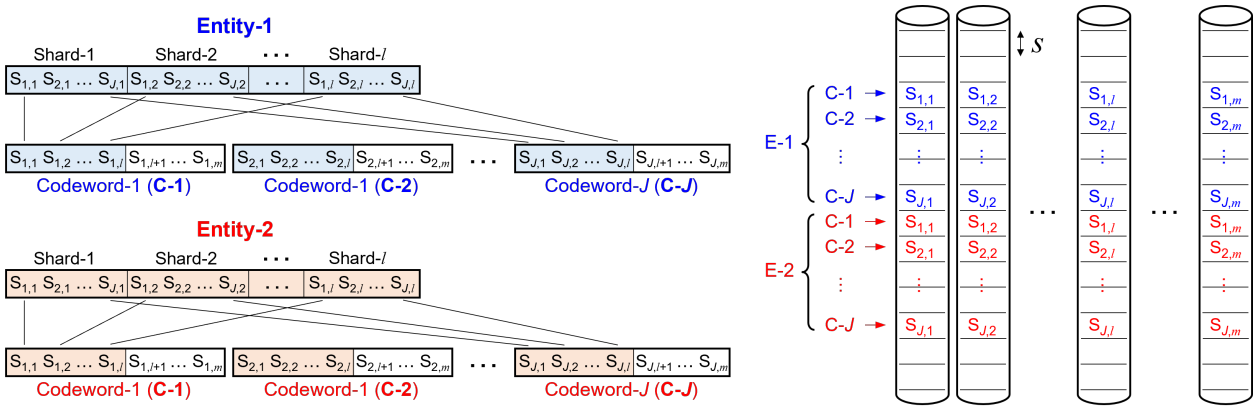
The storage space of devices is partitioned into units (symbols) of a fixed size s and complemented with parity symbols to form codewords. Each shard was assumed to be stored in an integer number of J symbols that is determined by

$$J = \frac{s_s}{s} = \frac{e_s}{l\, s}\,. \tag{4}$$

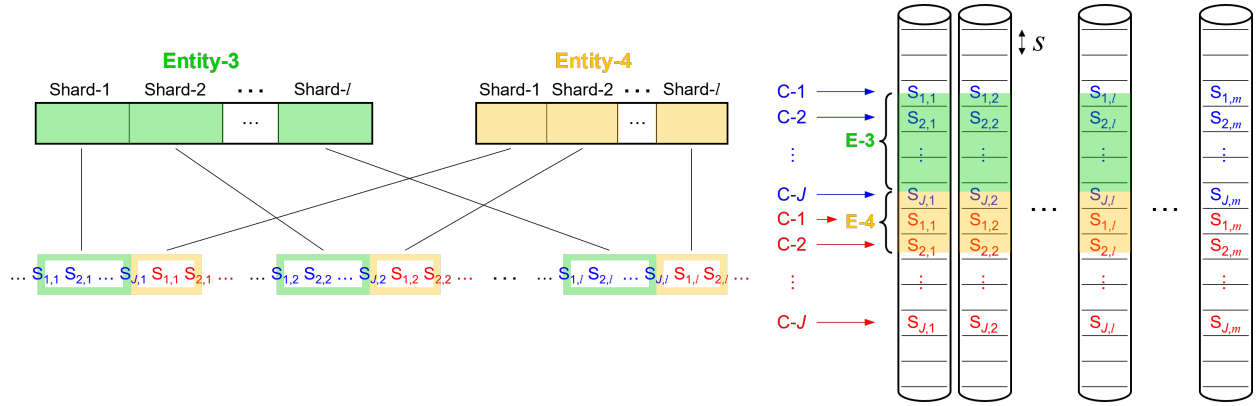
Consequently, the contents of each entity, such as Entity-1 and Entity-2, are stored in $J\, l$ user-data symbols with these symbols being stored in an integer number of J codewords. These codewords also contain $J\,(m - l)$ parity symbols for a total number of $J\, m$ symbols per entity, as shown in Figure 1(a). Note that $S_{j,i}$ denotes the ith symbol of the jth codeword. Thus, $S_{1,2}$, which is the second symbol of codeword C-1, is the first symbol of the second shard. Successive symbols of a shard are stored on the same device. To minimize the risk of permanent data loss, the m symbols of each of the J codewords are spread and stored successively in a set of m devices.

The model in [7] considered shards that have a fixed size of J symbols and are stored aligned with the symbol boundaries, which are indicated by the horizontal black lines in Figure 1(a). However, in practice user entities, and in turn shards, do not have a fixed size and, in the case of tape, are not necessarily aligned with symbols, because, as discussed in Section I, entity data is stored in a way that is agnostic to symbol boundaries. This is demonstrated in Figure 1(b) that shows two entities of two different sizes, Entity-3 and Entity-4, and the way they are stored on l devices of the system. For instance, Shard 1 of Entity-3 spans J symbols, i.e., the blue symbols $S_{1,1}$, $S_{2,1}$, $\cdots$, $S_{J,1}$, with its data partially occupying the first and last symbol, $S_{1,1}$ and $S_{J,1}$, respectively. Subsequently, Shard 1 of Entity-4 spans three symbols, namely, the blue symbol $S_{J,1}$ and the two red symbols $S_{1,1}$ and $S_{2,1}$, with its data partially occupying the first and the last symbol, that is, the blue $S_{J,1}$ and the red $S_{2,1}$ symbol. Thus, symbol $S_{J,1}$ contains data from both these entities. More generally, depending on the entity and symbol sizes, a symbol may contain data from multiple entities. Clearly, shard and entity sizes do not necessarily correspond to an integer number of symbols, which implies that the size J of a shard, expressed in number of symbols by (4), is in general a real number, which is less than 1 when the shard size is less than the symbol size. Codewords are formed by combining symbols containing user-data to generate and store parity symbols, as shown in Figure 1(b), regardless of the entities involved.

As pointed out in [7], the MTTDL metric does not depend on the entity size. This is due to the fact that the degree to which permanent data losses occur depends on the capability of the erasure-coding redundancy scheme employed and the resulting codeword formation, which in turn is agnostic to the entity placement and size characteristics. Note that an entity is lost if any of the codewords that it spans is permanently



(a) Symbol-aligned shards of integer size



(b) Non-symbol-aligned shards of arbitrary size

Figure 1. Data placement of entities and formation of codewords.

lost. Consequently, the EAFEL and EAFEDL metrics, which consider data loss at the entity level, depend on the number of codewords that entities span. The corresponding derivation is performed in Section IV.

The reliability of storage systems degrades by the presence of unrecoverable or latent errors. According to the specifications of enterprise quality HDDs, the unrecoverable bit-error probability P_{b} is equal to 10^{-15}. In practice, however, P_{b} can be orders of magnitude higher, reaching $P_{\mathrm{b}} \approx 10^{-12}$ [5]. On the other hand, according to Figure 13 in [20], tapes are more reliable than HDDs with a Bit Error Rate (BER) in the range of 10^{-22} to 10^{-19}. Assuming that bit errors occur independently over successive bits, the unrecoverable symbol error probability P_s is determined by

$$P_s = 1 - (1 - P_{\mathrm{b}})^s \,, \tag{5}$$

with the symbol size s expressed in bits. For a symbol size of 512 bytes, the equivalent unrecoverable sector error probability is $P_s \approx P_{\mathrm{b}} \times 512 \times 8$, which is 4.096×10^{-12} and 4.096×10^{-9} for $P_{\mathrm{b}} \approx 10^{-15}$ and 10^{-12}, respectively. Moreover, latent errors are found to exhibit spatial locality and they occur in bursts of B contiguous symbol errors. The degree to which symbol errors are correlated is captured by the factor f_{cor} whose value is determined by [5, Eq. (29)]

$$f_{\mathrm{cor}} = \begin{cases} 1 \,, & \text{for independent symbol errors} \\ \frac{1}{\bar{B}} \,, & \text{for correlated symbol errors} \,, \end{cases} \tag{6}$$

where $\bar{B}$ denotes the average length (in number of symbols) of bursts of latent symbol errors. Thus, $f_{\mathrm{cor}} \geq 1$.

IV. Codewords Spanned by Entities

Here, we obtain the distribution of the number of codewords, K, that entities span, which also represents the number of symbols that shards span. We proceed by considering the cases of fixed- and variable-size entities (shards).

A. Fixed-Size Entities

Let us consider fixed-size entities, which in turn result in fixed-size shards, such that J is fixed. Owing to periodicity, it suffices to study the process within a window of $S = J \times 10^k$ symbols, where k represents the number of decimal digits of J. This window corresponds in a symbol interval $[\epsilon, S + \epsilon]$ where ϵ is the starting position of the first shard within the first symbol, such that $0 < \epsilon < 1$. This interval contains S symbol boundaries and stores 10^k shards. For example, for $J = 4.287$, we have $k = 3$, and it suffices to consider the process in a window of $S = 4.287 \times 10^3 = 4,287$ symbols that store 1000 shards.

Let us now consider the example shown in Figure 2 whereby the shard size is 2.3. In this case, it holds that $k = 1$ and therefore it suffices to consider the process within a window of $S = 2.3 \times 10^1 = 23$ symbols that store 10 shards depicted between the black circles with the symbol boundaries

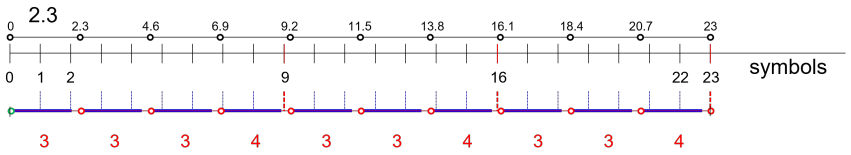


Figure 2. Number of symbols that shards span. Fixed-size shards of size $J = 2.3$ symbols.

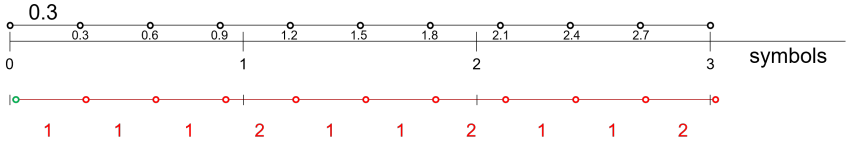


Figure 3. Number of symbols that shards span. Fixed-size shards of size $J = 0.3$ symbols.

indicated by the black vertical lines and with the first shard aligned with the first symbol. However, given that in practice shards are not aligned with symbols, their actual placement is indicated between the red circles, with the first shard starting at position ϵ, as indicated by the green circle. Figure 2 shows the case where $\epsilon = 0^+$.

Owing to periodicity, it suffices to study the process in the symbol interval $[\epsilon, 23 + \epsilon]$. The red integers indicate the number of symbols spanned by the successive shards. We note that 7 shards span 3 symbol and the remaining 3 shards span 4 symbols. Note that this holds for any $\epsilon \in (0, 1)$. Therefore, the probability density function (pdf) $\{p_j\}$ of the number of symbols K that an arbitrary shard spans is

$$P(K = i) = p_i = \begin{cases} 0.7\,, & \text{for } \; i = 3 \\ 0.3\,, & \text{for } \; i = 4\,. \end{cases} \tag{7}$$

Returning to the general case, we note that each shard can be decomposed into two components. The size of the first components, as indicated by the horizontal blue lines shown in Figure 2, corresponds to the number of symbols determined by the integer part of the shard size J, which is $\lfloor J \rfloor$ symbols. In the example considered, the integer part is 2. The size of the second components, as indicated by the horizontal red lines shown in Figure 2, corresponds to the fractional part, which is $J - \lfloor J \rfloor$ symbols. In the example considered, the fractional part is 0.3. Clearly, to each of the first (blue) components correspond $\lfloor J \rfloor$ symbol boundaries, which implies that each shard spans at least $\lfloor J \rfloor + 1$ symbols. In the example considered, to each of the first (blue) components correspond 2 symbol boundaries, as indicated by the blue vertical dotted lines, and, consequently, each shard spans at least 3 symbols.

As there are 10^k first components, one for each shard, the number of the corresponding symbol boundaries is $\lfloor J \rfloor \times 10^k$, which, in the example considered, is $2 \times 10^1 = 20$, as indicated by the blue vertical dotted lines. Consequently, there are $S - \lfloor J \rfloor \times 10^k = (J - \lfloor J \rfloor) \times 10^k$ additional symbol boundaries that correspond to $(J - \lfloor J \rfloor) \times 10^k$ out of the 10^k second components. In the example considered, there are $23 - 20 = 3$ additional symbol boundaries, as indicated by the red vertical dotted lines at positions 9, 16, and 23, that correspond to 3 out of the 10 red components. Consequently, these 3 components are associated with 3 shards, each of which spans one additional symbol for a total of 4 symbols. In general, each of the corresponding $(J - \lfloor J \rfloor) \times 10^k$ shards spans one additional symbol for a total of $\lfloor J \rfloor + 2$ symbols. Therefore, the percent of shards that span $\lfloor J \rfloor + 2$ symbols is $(J - \lfloor J \rfloor) \times 10^k / 10^k$ which is equal to $J - \lfloor J \rfloor$, that is, the fractional part of J denoted by $fr(J)$. Consequently, for any ϵ $(0 < \epsilon < 1)$, it holds that

$$P(K = i) = p_i = \begin{cases} 1 - fr(J)\,, & \text{for } \; i = \lfloor J \rfloor + 1 \\ fr(J)\,, & \text{for } \; i = \lfloor J \rfloor + 2 \\ 0\,, & \text{otherwise}\,, \end{cases} \tag{8}$$

where $fr(x)$ denotes the fractional part of the real number x,

$$fr(x) \triangleq x - \lfloor x \rfloor\,, \quad \forall\, x \in \mathcal{R}\,. \tag{9}$$

Let us also consider the case where $J < 1$ and the example shown in Figure 3 whereby the shard size is 0.3. Let us consider the first 10 shards indicated between the black circles with the first shard aligned with the first symbol. However, given that in practice shards are not aligned with symbols, their actual placement is indicated between the red circles, with the first shard starting at position ϵ, as indicated by the green circle. Owing to periodicity, it suffices to study the process in the symbol interval $[\epsilon, 3 + \epsilon]$. The red integers indicate the number of symbols spanned by the successive shards. We note that 7 shards span 1 symbol and the remaining 3 shards span 2 symbols and this holds for any $\epsilon \in (0, 1)$. Therefore, the pdf $\{p_j\}$ of the number of codewords (symbols) K that an arbitrary entity (shard) spans is

$$P(K = i) = p_i = \begin{cases} 0.7\,, & \text{for } \; i = 1 \\ 0.3\,, & \text{for } \; i = 2\,, \end{cases} \tag{10}$$

which is also the result determined by (8).

Next, we consider the case where the shard size is 2.7 symbols, as shown in Figure 4. Owing to periodicity, it suffices to study the process in the symbol interval $[\epsilon, 27 + \epsilon]$. The red integers indicate the number of symbols spanned by the successive shards. We note that 7 shards span 4 symbol and the remaining 3 shards span 3 symbols. According to (8), the pdf $\{p_j\}$ of the number of codewords (symbols) K that an arbitrary entity (shard) spans is

$$P(K = i) = p_i = \begin{cases} 0.3\,, & \text{for } \; i = 3 \\ 0.7\,, & \text{for } \; i = 4\,, \end{cases} \tag{11}$$

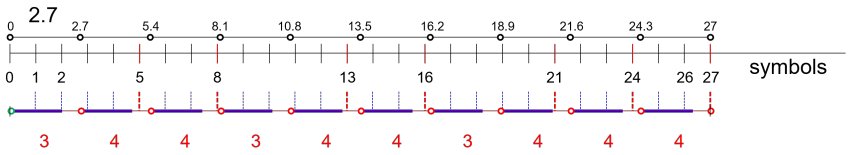


Figure 4. Number of symbols that shards span. Fixed-size shards of size $J = 2.7$ symbols.

which is also the result determined by (8).

B. Variable-Size Entities

We proceed to relax the assumption that all entities have the same size, by considering entities of E_s different sizes, $e_{s,1}, e_{s,2}, \cdots, e_{s,E_s}$. Without loss of generality, we assume that $e_{s,1} < e_{s,2} < \cdots < e_{s,E_s}$. Subsequently, let $\{v_j\}$ denote the corresponding pdf of the entity size, that is,

$$v_j \triangleq P(e_s = e_{s,j}) , \quad \text{for } j = 1, 2, \ldots, E_s , \tag{12}$$

such that the average entity size $E(e_s)$ is determined by

$$E(e_s) = \sum_{j=1}^{E_s} e_{s,j}\, v_j . \tag{13}$$

From (4), it follows that the shard size J_j corresponding to entity $e_{s,j}$ is determined by

$$J_j = \frac{e_{s,j}}{l\, s} \quad \text{for } j = 1, 2, \ldots, E_s . \tag{14}$$

Consequently, the pdf of the shard size J is determined by

$$P(J = J_j) = v_j , \quad \text{for } j = 1, 2, \ldots, E_s , \tag{15}$$

such that the average shard size $E(J)$ is determined by

$$E(J) = \sum_{j=1}^{E_s} J_j\, v_j \overset{(13)(14)}{=} \frac{E(e_s)}{l\, s} , \tag{16}$$

where the notation $\overset{(x)(y)}{=}$ implies that the final expression is derived using Equations (x) and (y).

The preceding discussion begs the following questions. Can the probability density function $\{p_j\}$ that was theoretically obtained in (8) for the case of a single fixed shard size be extended for the case of variable-size entities? Does it depend on the sequence according to which the variable-size entities are stored? Next, we address these critical questions. We shed light on these issues by considering the following cases regarding the placement and the way according to which the various shards are stored.

1) Segregated Shard Placement: According to this placement, shards of any given size are stored successively. One particular realization is to first store the shards of size J_1, followed by the shards of size J_2, and so on. For a large number of shards stored, from (8) and (15) we deduce that

$$P(K = i) = p_i = \begin{cases} [1 - fr(J_j)]\, v_j , & \text{for } i = \lfloor J_j \rfloor + 1 \\ fr(J_j)\, v_j , & \text{for } i = \lfloor J_j \rfloor + 2 \\ 0 , & \text{otherwise,} \end{cases}$$
$$\text{for } j = 1, 2, \ldots, E_s . \tag{17}$$

Let us consider the special case of a discrete bimodal distribution for the shard size, that is, $E_s = 2$, and let us assume that half of the shards have a size of 0.3 symbols and the remaining half of the shards have a size of 2.7 symbols. In this case we have $J_1 = 0.3$, $J_2 = 2.7$, and $v_1 = v_2 = 0.5$. For the particular realization where first the shards of size 0.3 are stored followed by the shards of size 2.7, (17) yields

$$P(K = i) = p_i = \begin{cases} 0.7 \times 0.5 = 0.35 , & \text{for } i = 1 \\ 0.3 \times 0.5 = 0.15 , & \text{for } i = 2 \\ 0.3 \times 0.5 = 0.15 , & \text{for } i = 3 \\ 0.7 \times 0.5 = 0.35 , & \text{for } i = 4 \\ 0 , & \text{otherwise.} \end{cases} \tag{18}$$

2) Alternating Shard Placement: According to this placement, shards of various sizes are stored interleaved by also considering the v_j values. One particular realization in the case where $v_j = 1/E_s$, for $j = 1, 2, \ldots, E_s$, is to first store a shard of size J_1, followed by a shard of size J_2, and so on. The first cycle is completed by storing a shard of size J_{E_s} and is followed by a second cycle that begins by storing a shard of size J_1.

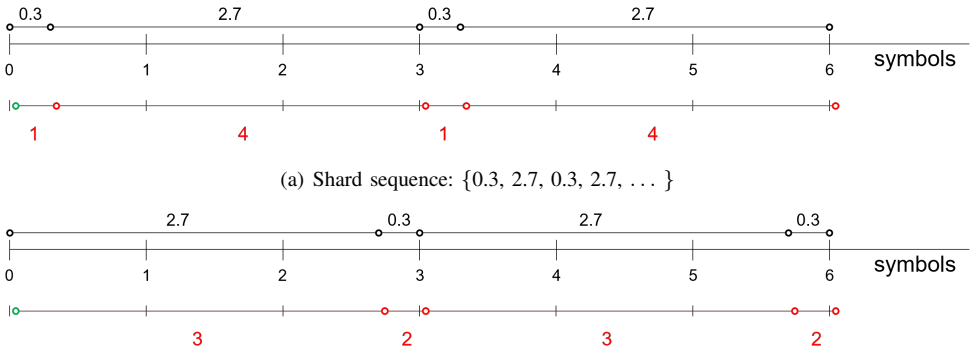
We proceed by investigating the special case considered in Section IV-B1 for the discrete bimodal distribution of the shard size, with the sizes of 0.3 and 2.7 symbols. The alternating placement of the shards corresponding to these two sizes lead to two possible sequence realizations, as shown in Figure 5.

The realization for the alternating sequence $\{0.3, 2.7, 0.3, 2.7, \ldots\}$ is depicted in Figure 5(a). Owing to periodicity, it suffices to study the process in the symbol interval $[\epsilon, 3+\epsilon]$. Figure 5(a) shows the case where $\epsilon = 0^+$. The red integers indicate the number of symbols spanned by the successive shards. We note that half of the shards span 1 symbol and the remaining half of the shards span 4 symbols and this holds for any $\epsilon \in (0, 0.7)$. Consequently, the pdf $\{p_j\}$ of the number of symbols K that an arbitrary shard spans is

$$P(K = i) = p_i = \begin{cases} 0.5 , & \text{for } i = 1 \\ 0.5 , & \text{for } i = 4 . \end{cases} \tag{19}$$

On the other hand, the realization for the alternating sequence $\{2.7, 0.3, 2.7, 0.3, \ldots\}$ is depicted in Figure 5(b). Owing to periodicity, it suffices to study the process in the symbol interval $[\delta, 3+\delta]$. Figure 5(b) shows the case where $\delta = 0^+$. In this case, half of the shards span 3 symbols and the remaining half of the shards span 2 symbols and this holds for any $\delta \in (0, 0.3)$. Consequently, the pdf $\{p_j\}$ of the number of symbols K that an arbitrary shard spans is

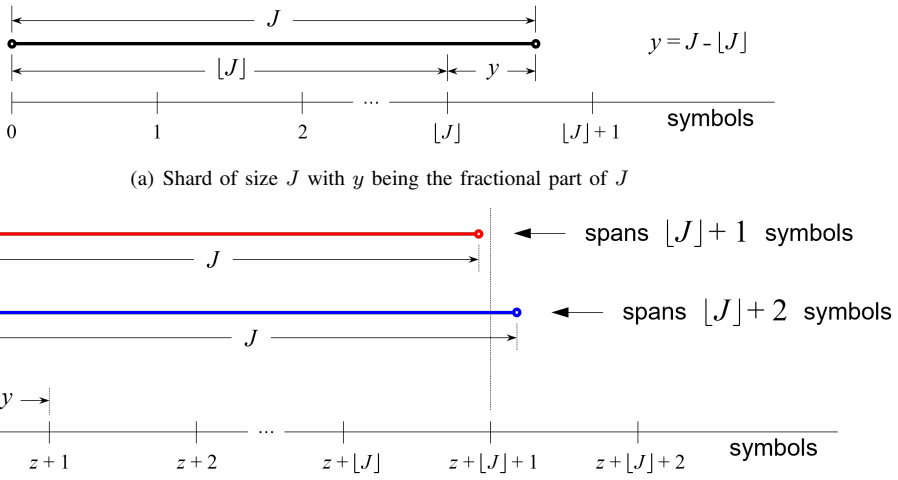
$$P(K = i) = p_i = \begin{cases} 0.5 , & \text{for } i = 2 \\ 0.5 , & \text{for } i = 3 . \end{cases} \tag{20}$$



(a) Shard sequence: {0.3, 2.7, 0.3, 2.7, . . . }

(b) Shard sequence: {2.7, 0.3, 2.7, 0.3, . . . }

Figure 5. Number of symbols spanned by shards. Alternating fixed-size shards of sizes 0.3 and 2.7 symbols, with $v_1 = v_2 = 0.5$.



(a) Shard of size J with y being the fractional part of J

(b) Random shard placement

Figure 6. Number of symbols that a randomly placed shard of size J spans.

Note that the pdf for $\delta \in (0.3, 1)$ is that determined by (19). Also, the pdf for $\epsilon \in (0.7, 1)$ is that determined by (20).

We now observe that the pdf determined by (20) is different from that determined by (19). Moreover, both of them, are different from that determined by (18) for the case of a segregated shard placement. Therefore, from the above, we deduce that the pdf $\{p_j\}$ of the number of symbols K that an arbitrary shard spans not only depends on the percentage of the various shard sizes in a sequence, as specified in (15), but also on their actual placement.

3) Random Shard Placement: According to this placement, the starting position ϵ $(0 < \epsilon < 1)$ of the first shard within the first symbol is uniformly distributed in $(0, 1)$. Successive shard sizes are assumed to be identically distributed, according to the distribution given in (15), but not necessarily independent. Note that this relaxes the assumption made in [1] of independent and identically distributed (i.i.d) successive shard sizes.

Let us consider a randomly chosen shard. Let also J denote its size, as shown in Figure 6(a), and y its fractional part, that is, $y = J - \lfloor J \rfloor$. Owing to the random placement of the first shard, the chosen shard, too, is randomly placed, such that it spans either $\lfloor J \rfloor + 1$ or $\lfloor J \rfloor + 2$ symbols, as depicted by the red and the blue shards shown in Figure 6(b), respectively. Let X denote the distance between the starting position of the shard and the left boundary z of the first symbol that the shard spans. Owing to the random placement of the shard, the random variable X is uniformly distributed between 0 and 1. Furthermore, when $X \leq 1 - y$, the shard spans $\lfloor J \rfloor + 1$ symbols whereas when $X > 1 - y$, the shard spans $\lfloor J \rfloor + 2$ symbols. Consequently, the probability that the shard spans $\lfloor J \rfloor + 1$ symbols is

$$P(K = \lfloor J \rfloor + 1) = \int_0^{1-y} dx = 1 - y \,, \tag{21}$$

which implies that the probability that the shard spans $\lfloor J \rfloor + 2$ symbols is

$$P(K = \lfloor J \rfloor + 2) = 1 - P(K = \lfloor J \rfloor + 1) \overset{(21)}{=} y \,. \tag{22}$$

Therefore, and given that $y = J - \lfloor J \rfloor = fr(J)$, it holds that

$$P(K = i) = p_i = \begin{cases} 1 - fr(J) \,, & \text{for } \; i = \lfloor J \rfloor + 1 \\ fr(J) \,, & \text{for } \; i = \lfloor J \rfloor + 2 \\ 0 \,, & \text{otherwise} \,. \end{cases} \tag{23}$$

From (23), and using (9), it follows that the mean number $E(K)$ of symbols that a shard of size J spans is

$$E(K) = (\lfloor J \rfloor + 1)P(K = \lfloor J \rfloor + 1) + (\lfloor J \rfloor + 2)P(K = \lfloor J \rfloor + 2) = (\lfloor J \rfloor + 1)\,[1 - fr(J)] + (\lfloor J \rfloor + 2)\,fr(J) = J + 1\,. \tag{24}$$

From (15), (23), and (24), it follows that the pdf and the average number of symbols K that an arbitrary shard spans are determined by

$$P(K = i) = p_i = \begin{cases} [1 - fr(J_j)]\,v_j\,, & \text{for } \; i = \lfloor J_j \rfloor + 1 \\ fr(J_j)\,v_j\,, & \text{for } \; i = \lfloor J_j \rfloor + 2 \\ 0\,, & \text{otherwise,} \end{cases} \quad \text{for } j = 1, 2, \ldots, E_s\,, \tag{25}$$

and

$$E(K) = \sum_{j=1}^{E_s} (\,J_j + 1)\,v_j = E(J) + 1\,. \tag{26}$$

Remark 1: For two different shard-size values, say $J_m \neq J_n$, for which it holds that $\lfloor J_m \rfloor = \lfloor J_n \rfloor = j$, the corresponding probabilities of the number of symbols K that these shards span are determined additively, that is, $P(K = j + 1) = [1 - fr(J_m)]\,v_m + [1 - fr(J_n)]\,v_m$ and $P(K = j + 2) = fr(J_m)\,v_m + fr(J_n)\,v_n$. Similarly, if $\lfloor J_m \rfloor + 1 = \lfloor J_n \rfloor = j$, then it holds that $P(K = j + 1) = fr(J_m)\,v_m + [1 - fr(J_n)]\,v_n$.

Remark 2: From (17) and (25), it follows that the pdfs of the number of symbols K that an arbitrary shard spans in the segregated and the random shard placement cases are the same. This is also the pdf that corresponds to the alternating shard placement case when the first shard is randomly placed. For the discrete bimodal distribution considered in Section IV-B2 for the alternating shards of sizes 0.3 and 2.7 symbols, when the first shard is randomly placed, that is, when the variables ϵ and δ are uniformly distributed in $(0, 1)$, combining (19) and (20) yields a pdf that is the same as that derived in (18) for the segregated shard placement case. Clearly, in the segregated and alternating shard placement cases successive shard sizes are dependent.

V. Derivation of MTTDL, EAFEL, and EAFEDL

The MTTDL, EAFEL, and EAFEDL reliability metrics are derived using the direct-path-approximation methodology presented in [2-6] and extend it to assess the effect of lazy rebuilds [21] in the presence of correlated symbol errors [5].

At any point in time, the system is in one of two modes: non-rebuild or rebuild mode. Note that part of the non-rebuild mode is the normal mode of operation where all devices are operational and all data in the system has the original amount of redundancy. Upon device failures, a rebuild process attempts to restore the lost data, which eventually leads the system either to a Data Loss (DL) with probability $P_{\rm DL}$ or back to the original normal mode by restoring initial redundancy, with probability $1 - P_{\rm DL}$. The MTTDL metric is then obtained by [6, Eq. (5)]:

$$\text{MTTDL} \approx \frac{E(T)}{P_{\rm DL}}\,, \tag{27}$$

where $P_{\rm DL}$ is determined by (49) and $E(T)$ denotes the expected duration, expressed in years, of a typical interval of normal operation until the rebuild process of failed devices is triggered, which is determined by Eq. (12) of [21] as follows:

$$E(T) = \left(\sum_{u=0}^{d} \frac{1}{\tilde{n}_u} \right) \Big/ \lambda\,, \quad \text{where } \tilde{n}_0 \triangleq n\,, \tag{28}$$

where $1/\lambda$ is the mean time to failure of a device and $\tilde{n}_u$ are determined by (35), (38), or (41), depending on the data placement scheme.

The EAFEL metric is obtained by Eq. (16) of [7] as follows:

$$\text{EAFEL} \approx \frac{E(Y)}{E(T) \cdot N_E}\,, \tag{29}$$

that is, as the ratio of the expected number $E(Y)$ of lost entities, normalized to the number N_E of entities in the system, to the expected duration $E(T)$ expressed in years. The number N_E of entities in the system is determined by

$$N_E \approx \frac{U}{E(e_s)} \overset{(1)}{=} \frac{n}{m} \cdot \frac{l\,c}{E(e_s)} \overset{(16)}{=} \frac{n}{m} \cdot \frac{c}{E(J)\,s}\,, \tag{30}$$

and $E(T)$ and $E(Y)$ are determined by (28) and (64).

Analogous to Eq. (9) of [6], the EAFEDL is obtained as the ratio of the expected amount $E(\breve{Q})$ of lost user data at the entity level, normalized to the amount U of user data, to the expected duration of $E(T)$ expressed in years:

$$\text{EAFEDL} \approx \frac{E(\breve{Q})}{E(T) \cdot U} \overset{(1)}{=} \frac{m\,E(\breve{Q})}{n\,l\,c\,E(T)}\,, \tag{31}$$

where $E(T)$ and $E(\breve{Q})$ are determined by (28) and (84).

A. Reliability Analysis

The EAFEL and EAFEDL are evaluated in parallel with MTTDL using the theoretical framework presented in [7]. The system is at exposure level u $(0 \leq u \leq \tilde{r})$ when there are codewords that have lost u symbols owing to device failures, but there are no codewords that have lost more symbols. These codewords are referred to as the *most-exposed* codewords. Transitions to higher exposure levels are caused by device failures, whereas transitions to lower ones are caused by successful rebuilds. We denote by C_u the number of most-exposed codewords upon entering exposure level u, $(u \geq 1)$. Upon the first device failure it holds that

$$C_1 = C\,, \tag{32}$$

where C is determined by (2).

A certain portion of the device bandwidth is reserved for read/write data recovery during the rebuild process, and the remaining bandwidth is used to serve user requests. Let b denote the actual average reserved rebuild bandwidth per device. Lost symbols are rebuilt in parallel using the rebuild bandwidth b available on each surviving device. The amount of data corresponding to the number C_u of symbols to be rebuilt at exposure level u is written at an average rate b_u $(\leq b)$ to selected device(s). For the time X required to read (or write) an amount c of data from (or to) a device it holds that

$$E(X) = c/b\,. \tag{33}$$

The results in this article hold for *highly reliable* storage devices, which satisfy the following condition [5][7]

$$\mu \int_0^\infty F_\lambda(t)[1 - F_X(t)]dt \ll 1, \quad \text{with} \quad \frac{\lambda}{\mu} \ll 1 \,. \tag{34}$$

This condition expresses the fact that the ratio of the mean time $1/\mu$ to read all contents of a device (which typically is on the order of tens of hours) to the mean time to failure of a device $1/\lambda$ (which is typically at least on the order of thousands of hours) is very small, and in particular the fact that it is very unlikely that a given device fails during a rebuild period.

At exposure level u, the number $\tilde{n}_u$ of devices whose failure causes an exposure level transition to level $u+1$ and the fraction V_u of the C_u most-exposed codewords that have symbols stored on any given such device depend on the codeword placement scheme. In particular, for the symmetric and declustered data placement, at each exposure level u, for $u = 1, \cdots, \tilde{r}-1$, it holds that [2][3]

$$\tilde{n}_u^{\text{sym}} = \; k - u \,, \qquad \text{for} \;\; u = 1, \ldots, \tilde{r} \tag{35}$$

$$b_u^{\text{sym}} = \frac{\min((k-u)\, b, B_{\max})}{l+1} \,, \quad \text{for} \;\; u = d+1, \ldots, \tilde{r} \tag{36}$$

$$V_u^{\text{sym}} = \; \frac{m-u}{k-u} \,, \qquad \text{for} \;\; u = 1, \ldots, \tilde{r} \,, \tag{37}$$

where $B_{\max}$ is the maximum network rebuild bandwidth.

The corresponding parameters $\tilde{n}_u^{\text{declus}}$, b_u^{declus}, and V_u^{declus} for the declustered placement are derived from (35), (36), and (37) by setting $k = n$ as follows:

$$\tilde{n}_u^{\text{declus}} = \; n - u \,, \qquad \text{for} \;\; u = 1, \ldots, \tilde{r} \tag{38}$$

$$b_u^{\text{declus}} = \frac{\min((n-u)\, b, B_{\max})}{l+1} \,, \quad \text{for} \;\; u = d+1, \ldots, \tilde{r} \tag{39}$$

$$V_u^{\text{declus}} = \; \frac{m-u}{n-u} \,, \qquad \text{for} \;\; u = 1, \ldots, \tilde{r} \,. \tag{40}$$

For the clustered placement, it holds that [2][3]

$$\tilde{n}_u^{\text{clus}} = \; m - u \,, \qquad \text{for} \;\; u = 1, \ldots, \tilde{r} \tag{41}$$

$$b_u^{\text{clus}} = \; \min(\, b \,, B_{\max}/l \,) \,, \qquad \text{for} \;\; u = d+1, \ldots, \tilde{r} \tag{42}$$

$$V_u^{\text{clus}} = \; 1 \,, \qquad \text{for} \;\; u = 1, \ldots, \tilde{r} \,. \tag{43}$$

Also, for the rebuild time R_u of the most-exposed codewords at exposure level u and for its fraction α_u still left when another device fails, causing the exposure level transition $u \to u+1$, it holds that [21, Eq. (49)]

$$R_{d+1} \approx \left(\prod_{j=1}^{d} V_j \right) \frac{b}{b_{d+1}} \, X \,, \tag{44}$$

with the convention that for any integer j and for any sequence δ_i , $\prod_{i=j}^{0} \delta_i \triangleq 1$.

For $u \leq d$, no rebuild is performed and therefore $\alpha_u = 1$. For $u > d$, α_u is approximately uniformly distributed in $(0,1)$ such that [21, Eq. (8)],

$$\alpha_u \;\approxeq\; \begin{cases} 1 \,, & \text{for} \;\; u = 1, \ldots, d \\ U(0,1) \,, & \text{for} \;\; u = d+1, \ldots, \tilde{r}-1 \,. \end{cases} \tag{45}$$

TABLE II. NOTATION OF RELIABILITY METRICS AT EXPOSURE LEVELS

Parameter	Definition
u	exposure level
P_u	probability of entering exposure level u
P_{UF_u}	probability of data loss due to unrecoverable symbol errors at exposure level u
P_{UF}	probability of data loss due to unrecoverable symbol errors
P_{DF}	probability of data loss due to $\tilde{r}$ successive device failures
P_{DL}	probability of data loss
q_u	probability that, at exposure level u, a codeword that has lost u symbols can be restored
$\hat{q}_u$	probability that, under instantaneous transitions from exposure level 1 to exposure level u, all of the C_u most-exposed codewords, which have lost u symbols, can be restored
$\tilde{q}_u$	probability that, at exposure level u, an arbitrary entity is lost
$\breve{q}_u$	expected amount of lost user data of an arbitrary entity at exposure level u
$\tilde{q}_{s,u}(x)$	the probability of loss, at exposure level u, of an entity whose shard size expressed in symbols is x

Furthermore, it holds that [21, Eq. (10)]

$$C_u \;\approx\; C \prod_{i=1}^{u-1} V_i \, \alpha_i \,, \quad \text{for} \;\; u = 1, \ldots, \tilde{r} \,, \tag{46}$$

The reliability metrics of interest are derived using the *direct path approximation*, which considers only transitions from lower to higher exposure levels [2-6]. This implies that each exposure level is entered only once. At any exposure level u $(u = d+1, \ldots, \tilde{r}-1)$, data loss may occur during rebuild owing to one or more unrecoverable failures, which is denoted by the transition $u \to \text{UF}$. Moreover, at exposure level $\tilde{r}-1$, data loss occurs owing to a subsequent device failure, which leads to the transition to exposure level $\tilde{r}$. Consequently, the direct paths that lead to data loss are the following:

$\overrightarrow{UF_u}$: the direct path of successive transitions $1 \to 2 \to \cdots \to u \to \text{UF}$, for $u = d+1, \ldots, \tilde{r}-1$, and

$\overrightarrow{\text{DF}}$: the direct path of successive transitions $1 \to 2 \to \cdots \to \tilde{r}-1 \to \tilde{r}$,

with corresponding probabilities P_{UF_u} and P_{DF}, respectively. The notation for the probabilities of the events that lead to data loss is summarized in Table II.

1) Data Loss: It holds that

$$P_{\text{UF}_u} = P_u \; P_{u \to \text{UF}} \,, \quad \text{for} \;\; u = d+1, \ldots, \tilde{r}-1 \,, \tag{47}$$

where P_u is the probability of entering exposure level u determined by [21, Eq. (17)]:

$$P_u \approx \frac{(\lambda\, c \prod_{j=1}^{d} V_j)^{u-d-1}}{(u-d-1)!} \, \frac{E(X^{u-d-1})}{[E(X)]^{u-d-1}} \prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i} \, V_i^{u-1-i} \,, \tag{48}$$

and $P_{u \to \text{UF}}$ is the probability of encountering an unrecoverable failure during the rebuild process at this exposure level.

In [10], it was shown that P_{DL} is accurately approximated by the probability of all direct paths to data loss. Therefore,

$$P_{\text{DL}} \approx \; P_{\text{DF}} + \sum_{u=d+1}^{\tilde{r}-1} P_{\text{UF}_u} \,. \tag{49}$$

Approximate expressions for the probabilities of data loss P_{UF_u} and P_{DF} are subsequently obtained by the following proposition.

Proposition 1: For $u = d+1, \ldots, \tilde{r}-1$, it holds that

$$P_{\mathrm{UF}_u} \approx -\left(\lambda\, c \prod_{j=1}^{d} V_j\right)^{u-d-1} \frac{E(X^{u-d-1})}{[E(X)]^{u-d-1}} \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i}\, V_i^{u-1-i}\right) \cdot \log(\hat{q}_u)^{-(u-d-1)} \left(\hat{q}_u - \sum_{i=0}^{u-d-1} \frac{\log(\hat{q}_u)^i}{i!}\right), \quad (50)$$

where

$$\hat{q}_u \triangleq q_u^{f_{\mathrm{cor}}\, C \prod_{j=1}^{u-1} V_j}, \quad (51)$$

$$q_u = 1 - \sum_{j=\tilde{r}-u}^{m-u} \binom{m-u}{j} P_s^j (1-P_s)^{m-u-j}, \quad (52)$$

$$P_{\mathrm{DF}} \approx \frac{(\lambda\, c \prod_{j=1}^{d} V_j)^{\tilde{r}-d-1}}{(\tilde{r}-d-1)!} \frac{E(X^{\tilde{r}-d-1})}{[E(X)]^{\tilde{r}-d-1}} \prod_{i=d+1}^{\tilde{r}-1} \frac{\tilde{n}_i}{b_i}\, V_i^{\tilde{r}-1-i}. \quad (53)$$

Proof: Immediate by combining Proposition 1 of [5] and Proposition 1 of [21], and by also taking into account the effect of correlated latent errors via the variable f_{cor}, which is determined by (6), as discussed in Appendix A. ■

Remark 3: For small values of P_s, and a according to Remark 1 of [5], it holds that

$$q_u \approx \begin{cases} 1 - \binom{m-u}{\tilde{r}-u} P_s^{\tilde{r}-u}, & \text{for } P_s \ll \binom{m-u}{\tilde{r}-u}^{-\frac{1}{\tilde{r}-u}} \\ 0\,, & \text{for } P_s \gg \binom{m-u}{\tilde{r}-u}^{-\frac{1}{\tilde{r}-u}}, \end{cases} \quad (54)$$

$$\hat{q}_u \approx \begin{cases} 1 - Z_u\, P_s^{\tilde{r}-u}, & \text{for } P_s \ll P^*_{s,u} \\ 0\,, & \text{for } P_s \gg P^*_{s,u}, \end{cases} \quad (55)$$

where

$$Z_u \triangleq f_{\mathrm{cor}}\, C \left(\prod_{j=1}^{u-1} V_j\right) \binom{m-u}{\tilde{r}-u}, \quad (56)$$

and $P^*_{s,u} = Z_u^{-\frac{1}{\tilde{r}-u}}$.

Corollary 1: For $u = d+1, \ldots, \tilde{r}-1$, it holds that

$$P_{\mathrm{UF}_u} \approx \begin{cases} A_u\, P_s^{\tilde{r}-u}, & \text{for } P_s \ll P^{(\tilde{r})}_{\tilde{s},u} \\ P_u\,, & \text{for } P_s \gg P^{(\tilde{r})}_{\tilde{s},u}, \end{cases} \quad (57)$$

where

$$A_u \triangleq f_{\mathrm{cor}}\, C \binom{m-u}{\tilde{r}-u} (\lambda\, c)^{u-d-1} \frac{\left(\prod_{j=1}^{d} V_j\right)^{u-d}}{(u-d)!} \cdot \frac{E(X^{u-d-1})}{[E(X)]^{u-d-1}} \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i}\, V_i^{u-i}\right), \quad (58)$$

P_u is determined by (48), and

$$\boxed{P^{(\tilde{r})}_{\tilde{s},u} \triangleq \left[\frac{u-d}{f_{\mathrm{cor}}\, C \binom{m-u}{\tilde{r}-u} \prod_{i=1}^{u-1} V_i}\right]^{\frac{1}{\tilde{r}-u}}}. \quad (59)$$

Proof: See Appendix A. ■

Remark 4: It follows from (59) that $P^{(\tilde{r})}_{\tilde{s},u}$ is dominated by the large value of C. Consequently, it holds that

$$0 = P^{(\tilde{r})}_{\tilde{s},\tilde{r}} < P^{(\tilde{r})}_{\tilde{s},\tilde{r}-1} < \cdots < P^{(\tilde{r})}_{\tilde{s},d+2} < P^{(\tilde{r})}_{\tilde{s},d+1}\,. \quad (60)$$

Remark 5: Note that P_{DL}, as a function of P_s, exhibits $\tilde{r}-d$ plateaus at levels P_u in the intervals $(P^{(\tilde{r})}_{\tilde{s},u}, P^{(\tilde{r})}_{\hat{s},u})$, for $u = d+1, \cdots, \tilde{r}$, respectively, where $P^{(\tilde{r})}_{\hat{s},d+1} \triangleq 1$ and $P^{(\tilde{r})}_{\hat{s},u}$ is determined by (59). Also, $[0, P^{(\tilde{r})}_{\hat{s},u})$ is the range of values of P_s for which it holds that $P_{\mathrm{UF}_{u-1}} \ll P_{\mathrm{UF}_u}$. It follows from approximation (57) that $P^{(\tilde{r})}_{\hat{s},u}$ satisfies equation $A_{u-1}\,(P^{(\tilde{r})}_{\hat{s},u})^{\,\tilde{r}-u+1} = P_u$, which, using (2) and (48), yields

$$\boxed{P^{(\tilde{r})}_{\hat{s},u} \triangleq \left[\frac{\lambda\, s\, E(X^{u-d-1})\, \tilde{n}_{u-1}}{f_{\mathrm{cor}} \binom{m-u+1}{\tilde{r}-u+1} E(X)\, E(X^{u-d-2})\, b_{u-1}}\right]^{\frac{1}{\tilde{r}-u+1}}}. \quad (61)$$

Note also that when $P^{(\tilde{r})}_{\tilde{s},u} > P^{(\tilde{r})}_{\hat{s},u}$, the interval $(P^{(\tilde{r})}_{\tilde{s},u}, P^{(\tilde{r})}_{\hat{s},u})$ as well as the corresponding plateau vanish.

Remark 6: From (61), and given that the term in the bracket is quite small, it follows that

$$P^{(\tilde{r})}_{\hat{s},\tilde{r}} < P^{(\tilde{r})}_{\hat{s},\tilde{r}-1} < \cdots < P^{(\tilde{r})}_{\hat{s},d+2} < P^{(\tilde{r})}_{\hat{s},d+1} = 1\,. \quad (62)$$

The methodology presented in [10] that considers the most probable path to data loss yields an approximate function for P_{DL}. This function is obtained analytically by Corollary 1, and Remarks 4, 5, and 6, and has the shape shown in a log-log plot in Figure 7 along with the plateaus and corresponding intervals.

Remark 7: The plateaus derived in the case where $d = 0$ are in agreement with those determined in [5].

Remark 8: According to Remarks 5 and 6, P_{DL} and, by virtue of (27), MTTDL is affected when

$$\boxed{P_s \gg P^{(\tilde{r})}_{\hat{s},\tilde{r}} \overset{(3)(61)}{=} \frac{\lambda\, s\, E(X^{\tilde{r}-d-1})\, \tilde{n}_{\tilde{r}-1}}{f_{\mathrm{cor}}\, l\, E(X)\, E(X^{\tilde{r}-d-2})\, b_{\tilde{r}-1}}}. \quad (63)$$

2) Entity Loss: We proceed to derive the number of lost entities during rebuild. Let Y be the number of lost entities. Let also Y_{DF} and Y_{UF_u} denote the number of lost entities associated with the direct paths $\overrightarrow{\mathrm{DF}}$ and $\overrightarrow{UF_u}$, respectively. Then, it holds that [7, Eqs. (37), (38), (41)]

$$E(Y) \approx E(Y_{\mathrm{DF}}) + \sum_{u=d+1}^{\tilde{r}-1} E(Y_{\mathrm{UF}_u}) \approx E(Y_{\mathrm{DF}}) + E(Y_{\mathrm{UF}})\,, \quad (64)$$

where Y_{UF} denotes the number of lost entities due to unrecoverable failures with its mean given by

$$E(Y_{\mathrm{UF}}) \approx \sum_{u=d+1}^{\tilde{r}-1} E(Y_{\mathrm{UF}_u})\,. \quad (65)$$

Proposition 2: For $u = d+1, \ldots, \tilde{r}-1$, it holds that

$$E(Y_{\mathrm{UF}_u}) \approx \frac{C}{E(J)} \frac{P_u}{u-d} \left(\prod_{j=1}^{u-1} V_j\right) \tilde{q}_u\,, \quad (66)$$

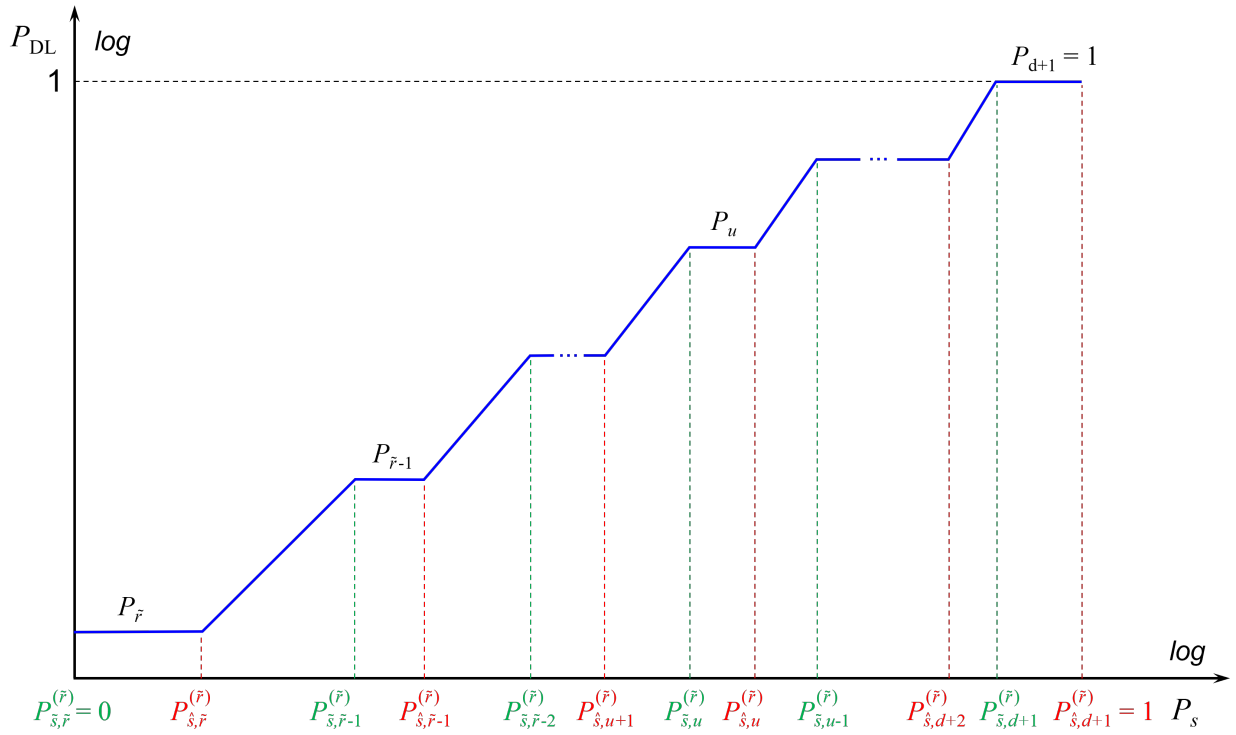


Figure 7. Approximate P_{DL} vs. P_s considering the most probable path to data loss.

where $\tilde{q}_u$, which denotes the probability that an arbitrary entity is lost, is determined by

$$\tilde{q}_u = \sum_{j=1}^{E_s} \tilde{q}_{s,u}\left(\frac{e_{s,j}}{l\,s}\right) v_j\,, \quad \text{for } u = d+1,\ldots,\tilde{r}\,, \tag{67}$$

with the probability $\tilde{q}_{s,u}(x)$ of loss, at exposure level u, of an entity whose shard size expressed in symbols is x, determined by

$$\tilde{q}_{s,u}(x) \triangleq 1 - [1 - fr(x)]\, q_u^{f_{\text{cor}}(\lfloor x \rfloor+1)} - fr(x)\, q_u^{f_{\text{cor}}(\lfloor x \rfloor+2)}\,, \tag{68}$$

and the probability q_u that a codeword that has lost u symbols can be restored, determined by (52).

It also holds that

$$E(Y_{\text{DF}}) \approx \frac{C}{E(J)}\,\frac{P_{\text{DF}}}{\tilde{r}-d}\prod_{j=1}^{\tilde{r}-1} V_j\,, \tag{69}$$

where C is determined by (2), P_s is determined by (5), $fr(x)$ is determined by (9), $E(J)$ is determined by (16). Also, the probability P_u of entering exposure level u is determined by (48).

Proof: Equation (66) is obtained in Appendix B. Equation (69) is obtained from (66) by setting $u = \tilde{r}$ and recognizing that $q_{\tilde{r}} = 0$, $\tilde{q}_{s,\tilde{r}}(x) = 1$, $\forall\, x \in \mathcal{R}$, $\tilde{q}_{\tilde{r}} = 1$, and $P_{\tilde{r}} = P_{\text{DF}}$. ■

Remark 9: For $u = d+1,\ldots,\tilde{r}-1$ and for small values of P_s, it follows from (68) and (54) that

$$\tilde{q}_{s,u}(x) \approx \begin{cases} f_{\text{cor}}\,(x+1)\binom{m-u}{\tilde{r}-u} P_s^{\tilde{r}-u}, & \text{for } P_s \ll P^{\#}_{s,u}(x) \\ 1\,, & \text{for } P_s \gg P^{\#}_{s,u}(x)\,, \end{cases} \tag{70}$$

where $P^{\#}_{s,u}(x)$ is obtained from the approximation (70), $\tilde{q}_{s,u}(x) \approx f_{\text{cor}}\,(x+1)\binom{m-u}{\tilde{r}-u} P^{\#}_{s,u}(x)^{\tilde{r}-u} = 1$, as follows: $P^{\#}_{s,u}(x) \triangleq \left[f_{\text{cor}}\,(x+1)\binom{m-u}{\tilde{r}-u}\right]^{-\frac{1}{\tilde{r}-u}}$. Also, for $u = \tilde{r}$, and given that $q_{\tilde{r}} = 0$, it follows from (68) that

$$\tilde{q}_{s,\tilde{r}}(x) = 1\,, \quad \forall\, x \in \mathcal{R}\,. \tag{71}$$

From (67), and using (70) and (71), it follows that

$$\tilde{q}_u \approx f_{\text{cor}}\left(\frac{E(e_s)}{l\,s}+1\right)\binom{m-u}{\tilde{r}-u} P_s^{\tilde{r}-u}, \quad \text{for } P_s \ll P^{(\tilde{r})}_{\dot{s},u}\,, \tag{72}$$

where

$$P^{(\tilde{r})}_{\dot{s},u} \triangleq P^{\#}_{s,u}\left(\frac{e_{s,E_s}}{l\,s}\right) = \left[f_{\text{cor}}\,\left(\frac{e_{s,E_s}}{l\,s}+1\right)\binom{m-u}{\tilde{r}-u}\right]^{-\frac{1}{\tilde{r}-u}}, \tag{73}$$

and, for $u = \tilde{r}$,

$$\tilde{q}_{\tilde{r}} = 1\,. \tag{74}$$

Remark 10: Let $P^{(\tilde{r})}_{\ddot{s},u}$ be the value of P_s for which it holds that $E(Y_{\text{UF}_{u-1}}) \approx E(Y_{\text{UF}_u})$, for $u = d+2,\ldots,\tilde{r}$. It follows from (48), (66), and (72) that

$$P^{(\tilde{r})}_{\ddot{s},u} \triangleq \frac{\lambda\, c\,(\tilde{r}-u+1)\,\tilde{n}_{u-1}}{(m-u+1)\,(u-d)\,b_{u-1}} \cdot \frac{E(X^{u-d-1})}{E(X)\,E(X^{u-d-2})} \cdot \prod_{i=1}^{u-1} V_i\,. \tag{75}$$

Corollary 2: For deterministic rebuild time distributions, it holds that

$$P^{(\tilde{r})}_{\ddot{s},\tilde{r}} < P^{(\tilde{r})}_{\ddot{s},\tilde{r}-1} < \cdots < P^{(\tilde{r})}_{\ddot{s},d+1}\,. \tag{76}$$

Proof: See Appendix C. ■

For Weibull rebuild time distributions, including exponential ones, relation (76) does not necessarily hold. Nevertheless, the following relation always holds.

Corollary 3: For Weibull rebuild time distributions, it holds that

$$P_{\ddot{s},\tilde{r}}^{(\tilde{r})} < P_{\ddot{s},u}^{(\tilde{r})} , \qquad \text{for } \; u = d+2, \ldots, \tilde{r}-1 . \tag{77}$$

Proof: See Appendix D. ■

Remark 11: Note that for $d=0$, the $P_{\ddot{s},u}^{(\tilde{r})}$ obtained from (75) is equal to $P_{\ddot{s},u}^{(\tilde{r})}$, as determined by Eq. (54) of [5]. Therefore, by considering Remarks 8 and 9 of [5] and Corollaries 2 and 3, we deduce that for the deterministic and Weibull rebuild time distributions and for any value of d, $[0, P_{\ddot{s},u}^{(\tilde{r})})$ is the range of values of P_s for which it holds that $E(Q_{\text{UF}_{\tilde{r}}}) \ll E(Q_{\text{UF}_u})$, for $u = 1, \ldots, \tilde{r}-1$, where Q is the (not effective) amount of lost user data. Consequently, EAFDL is affected when

$$P_s \gg P_{\ddot{s},\tilde{r}}^{(\tilde{r})} = P_{\ddot{s},\tilde{r}}^{(\tilde{r})} \tag{78}$$

Remark 12: According to (71) and given that $\tilde{q}_{\tilde{r}} = 1$, for $u = \tilde{r}$, approximation (72) yields $\tilde{q}_{\tilde{r}}(\text{approximation}) \approx f_{\text{cor}}\left(\frac{E(e_s)}{l\,s}+1\right) > 1 = \tilde{q}_{\tilde{r}}$. This in turn implies that $E(Y_{\text{UF}_{\tilde{r}}})/E(Y_{\text{DF}}) \approx f_{\text{cor}}\left(\frac{E(e_s)}{l\,s}+1\right) > 1$, given that $f_{\text{cor}} \geq 1$. Moreover, let $P_{\ddot{s},\tilde{r}}^{(\tilde{r})}$ be the value of P_s for which it holds that $E(Y_{\text{UF}_{\tilde{r}-1}}) \approx E(Y_{\text{DF}})$. From the above, and using (48), (53), (66), (69), (72), and (75), it follows that

$$P_{\grave{s},\tilde{r}}^{(\tilde{r})} \approx \frac{P_{\ddot{s},\tilde{r}}^{(\tilde{r})}}{f_{\text{cor}}\left(\frac{E(e_s)}{l\,s}+1\right)} < P_{\ddot{s},\tilde{r}}^{(\tilde{r})} . \tag{79}$$

Remark 13: For the deterministic and Weibull rebuild time distributions, inequalities (76), (77), and (79) imply that

$$P_{\grave{s},\tilde{r}}^{(\tilde{r})} < P_{\ddot{s},u}^{(\tilde{r})} , \qquad \text{for } \; u = d+1, \ldots, \tilde{r}-1 . \tag{80}$$

Therefore, for values of P_s in the interval $[0, P_{\grave{s},\tilde{r}}^{(\tilde{r})})$, it holds that $E(Y_{\text{UF}_u}) \ll E(Y_{\text{DF}})$, for $u = d+1, \ldots, \tilde{r}-1$. Consequently, from (64), $E(Y)$ and, by virtue of (29), EAFEL are affected when

$$P_s \gg P_{\grave{s},\tilde{r}}^{(\tilde{r})} , \tag{81}$$

where $P_{\grave{s},\tilde{r}}^{(\tilde{r})}$ is obtained using (3), (75) and (79) as follows:

$$\boxed{P_{\grave{s},\tilde{r}}^{(\tilde{r})} \triangleq \frac{\lambda\, c\;\tilde{n}_{\tilde{r}-1}\;\; E(X^{\tilde{r}-d-1})}{f_{\text{cor}}\left(\frac{E(e_s)}{l\,s}+1\right)\; l\;(\tilde{r}-d)\; b_{\tilde{r}-1}\; E(X)\, E(X^{\tilde{r}-d-2})} \prod_{i=1}^{\tilde{r}-1} V_i .} \tag{82}$$

Corollary 4: For small values of P_s such that $P_s \ll \min\left(P_{\ddot{s},u}^{(\tilde{r})}, P_{\grave{s},u}^{(\tilde{r})}\right)$, the following relation holds

$$E(Y_{\text{UF}_u}) \;\approx\; \frac{E(J)+1}{E(J)}\; P_{\text{UF}_u} . \tag{83}$$

Proof: See Appendix E. ■

3) Effective Amount of Data Loss: We proceed to derive the effective amount of lost user data during rebuild. Let $\breve{Q}$ be the amount of user data contained in the Y lost entities, which is permanently lost, too. Let also $\breve{Q}_{\text{DF}}$ and $\breve{Q}_{\text{UF}_u}$ denote the amount of lost user data associated with the direct paths $\overrightarrow{\text{DF}}$ and $\overrightarrow{UF_u}$, respectively.

Similar to (64), it holds that

$$E(\breve{Q}) \approx E(\breve{Q}_{\text{DF}}) + \sum_{u=d+1}^{\tilde{r}-1} E(\breve{Q}_{\text{UF}_u}) \approx E(\breve{Q}_{\text{DF}}) + E(\breve{Q}_{\text{UF}}) , \tag{84}$$

where $\breve{Q}_{\text{UF}}$ denotes the amount of user data lost due to unrecoverable failures with its mean given by

$$E(\breve{Q}_{\text{UF}}) \approx \sum_{u=d+1}^{\tilde{r}-1} E(\breve{Q}_{\text{UF}_u}) . \tag{85}$$

Proposition 3: For $u = d+1, \ldots, \tilde{r}-1$, it holds that

$$E(\breve{Q}_{\text{UF}_u}) \;\approx\; \frac{C}{E(J)}\;\frac{P_u}{u-d}\;\left(\prod_{j=1}^{u-1} V_j\right)\breve{q}_u , \tag{86}$$

where the expected amount $\breve{q}_u$ of lost user data of an arbitrary entity is determined by

$$\breve{q}_u \;=\; \sum_{j=1}^{E_s} e_{s,j}\;\tilde{q}_{s,u}\left(\frac{e_{s,j}}{l\,s}\right) v_j . \tag{87}$$

It also holds that

$$E(\breve{Q}_{\text{DF}}) \;\approx\; \frac{C}{E(J)}\;\frac{P_{\text{DF}}}{\tilde{r}-d}\;\left(\prod_{j=1}^{\tilde{r}-1} V_j\right)\breve{q}_{\tilde{r}} , \tag{88}$$

where C is determined by (2), $E(J)$ is determined by (16), $\tilde{q}_{s,u}(x)$ is determined by (68), P_u is determined by (48), P_{DF} is determined by (53), and V_j are determined by (37), (40), and (43).

Proof: Equations (86) and (87) are obtained in Appendix B. Equation (88) is obtained from (86) by setting $u = \tilde{r}$ and recognizing that $P_{\tilde{r}} = P_{\text{DF}}$. ■

Remark 14: For $u = d+1, \ldots, \tilde{r}-1$ and for small values of P_s, it follows from (87), and using (70), that

$$\breve{q}_u \;\approx f_{\text{cor}}\left(\frac{E({e_s}^2)}{l\,s}+E(e_s)\right)\binom{m-u}{\tilde{r}-u} P_s^{\tilde{r}-u}, \;\; P_s \ll P_{\grave{s},u}^{(\tilde{r})} , \tag{89}$$

where $P_{\grave{s},u}^{(\tilde{r})}$ is determined by (73). Moreover, for $u = \tilde{r}$ and using (13) and (71), (87) yields

$$\breve{q}_{\tilde{r}} \;=\; E(e_s) . \tag{90}$$

Also, from (72) and (89), it follows that

$$\breve{q}_u \;\approx\; f(e_s)\, E(e_s)\; \tilde{q}_u , \tag{91}$$

where

$$f(e_s) \;\triangleq\; \frac{\frac{E({e_s}^2)}{l\,s}+E(e_s)}{\left(\frac{E(e_s)}{l\,s}+1\right)E(e_s)} \;\geq\; 1 , \tag{92}$$

with the inequality being deduced from the fact that for any random variable X, it holds that $E(X^2) \geq E(X)^2$.

Combining (66), (86), and (91) yields

$$E(\breve{Q}_{\text{UF}_u}) \approx f(e_s)\, E(e_s)\, E(Y_{\text{UF}_u}) , \tag{93}$$

Also, from (69), (88), and (90), it follows that

$$E(\breve{Q}_{\text{DF}}) \approx E(Y_{\text{DF}})\, E(e_s) . \tag{94}$$

Remark 15: From Remark 10 and (93), it follows that $E(\breve{Q}_{\text{UF}_u}) \approx E(\breve{Q}_{\text{UF}_{u-1}})$ for $P_s = P_{\breve{s},u}^{(\tilde{r})}$, which is determined by (75).

Remark 16: According to (90) and given that $\breve{q}_{\tilde{r}} = E(e_s)$, for $u = \tilde{r}$, approximation (89) yields $\breve{q}_{\tilde{r}}(\text{approximation}) \approx f_{\text{cor}}\left(\frac{E({e_s}^2)}{l\,s} + E(e_s)\right) > E(e_s) = \tilde{q}_{\tilde{r}}$. This in turn implies that $E(\breve{Q}_{\text{UF}_{\tilde{r}-1}})/E(\breve{Q}_{\text{DF}}) \approx f_{\text{cor}}\left(\frac{E({e_s}^2)}{l\,s} + E(e_s)\right)/E(e_s) > 1$, given that $f_{\text{cor}} \geq 1$. Moreover, let $P_{\breve{s},\tilde{r}}^{(\tilde{r})}$ be the value of P_s for which it holds that $E(\breve{Q}_{\text{UF}_{\tilde{r}-1}}) \approx E(\breve{Q}_{\text{DF}})$. From the above, and using (48), (53), (75), (79), (86), (88), (89), and (92), it follows that

$$P_{\breve{s},\tilde{r}}^{(\tilde{r})} \approx \frac{E(e_s)\, P_{\breve{s},\tilde{r}}^{(\tilde{r})}}{f_{\text{cor}}\left(\frac{E({e_s}^2)}{l\,s} + E(e_s)\right)} \approx \frac{P_{\grave{s},\tilde{r}}^{(\tilde{r})}}{f(e_s)} \leq P_{\grave{s},\tilde{r}}^{(\tilde{r})} < P_{\breve{s},\tilde{r}}^{(\tilde{r})} . \tag{95}$$

Remark 17: For the deterministic and Weibull rebuild time distributions, inequalities (76), (77), and (98) imply that

$$P_{\breve{s},\tilde{r}}^{(\tilde{r})} < P_{\breve{s},u}^{(\tilde{r})} , \qquad \text{for } u = d+1, \ldots, \tilde{r}-1 . \tag{96}$$

Therefore, for values of P_s in the interval $[0, P_{\breve{s},\tilde{r}}^{(\tilde{r})})$, it holds that $E(\breve{Q}_{\text{UF}_u}) \ll E(\breve{Q}_{\text{DF}})$, for $u = d+1, \ldots, \tilde{r}-1$. Consequently, from (84), $E(\breve{Q})$ and, by virtue of (31), EAFEDL are affected when

$$P_s \gg P_{\breve{s},\tilde{r}}^{(\tilde{r})} , \tag{97}$$

where $P_{\breve{s},\tilde{r}}^{(\tilde{r})}$ is obtained using (3), (75) and (79) as follows:

$$\boxed{P_{\breve{s},\tilde{r}}^{(\tilde{r})} \triangleq \frac{E(e_s)\ \left(\prod_{i=1}^{\tilde{r}-1} V_i\right) \lambda\, c\ \tilde{n}_{\tilde{r}-1}\ E(X^{\tilde{r}-d-1})}{f_{\text{cor}}\left(\frac{E({e_s}^2)}{l\,s} + E(e_s)\right) l\ (\tilde{r}-d)\ b_{\tilde{r}-1}\ E(X)\, E(X^{\tilde{r}-d-2})}} . \tag{98}$$

Corollary 5: For small values of P_s such that $P_s \ll P_{\breve{s},\tilde{r}}^{(\tilde{r})}$, the fraction of lost entities $E(Y)/N_E$ reflects the fraction of lost user data $E(\breve{Q})/U$ and therefore it holds that EAFEL $\approx$ EAFEDL, which is determined by

$$\text{EAFEDL} \approx \frac{m\, P_{\text{DF}}}{n\ (\tilde{r}-d)\ E(T)} \left(\prod_{j=1}^{\tilde{r}-1} V_j\right) , \ \text{ for } \ P_s \ll P_{\breve{s},\tilde{r}}^{(\tilde{r})} . \tag{99}$$

Moreover, the common value of the EAFEL and EAFEDL reliability metrics does not depend on the entity sizes nor the symbol size.

Proof: From (64), (84), and according to Remark 16, we deduce that, for $P_s \ll P_{\breve{s},\tilde{r}}^{(\tilde{r})}$, it holds that $E(Y) \approx E(Y_{\text{DF}})$ and $E(\breve{Q}) \approx E(\breve{Q}_{\text{DF}})$. Consequently, combining (29), (30), (31), and (94), yields $E(Y)/N_E \approx E(\breve{Q})/U$ and EAFEL $\approx$ EAFEDL. Moreover, substituting (88) into (31), and using (2) and (16), yields (99). From (28), (37), (40), (43), and (53), we deduce that all variables involved in (99) are independent of the symbol size s and the entity sizes $e_{s,1}, \cdots, e_{s,E_s}$. ■

TABLE III. PARAMETER VALUES

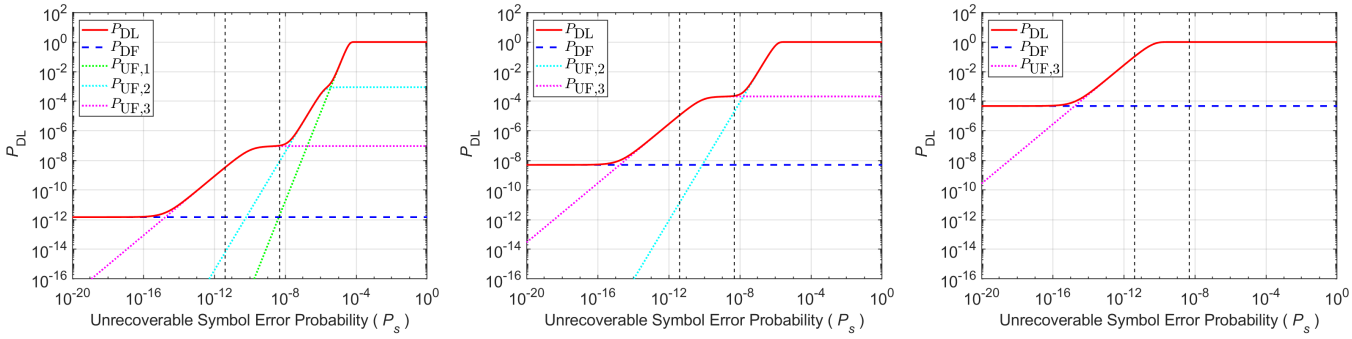
Parameter	Definition	Values
n	number of storage devices	64
c	amount of data stored on each device	20 TB
s	symbol (sector or data set) size	512 B, 5 MB
λ^{-1}	mean time to failure of a storage device	876,000 h
b	rebuild bandwidth per device	100 MB/s
m	symbols per codeword	16
l	user-data symbols per codeword	13, 14, 15
d	lazy rebuild threshold ($0 \leq d < m-l$)	0, 1, 2
U	amount of user data stored in the system	1.04 to 1.2 PB
μ^{-1}	time to read an amount c of data at a rate b from a storage device	55.5 h

VI. NUMERICAL RESULTS

Here, we assess the reliability of the clustered and declustered placement schemes for the system and the parameter values considered in [7], as listed in Table III. The system is comprised of $n = 64$ devices (HDDs), it is protected by MDS erasure codes with $m = 16$ and $l = 13$, 14, 15 and employs a lazy rebuild scheme with a threshold $d = 0$, 1, and 2. Each HDD stores an amount of $c = 20$ TB with a sector (symbol) size s of 512 bytes. The value for the parameter λ^{-1} is chosen to be $876,000$ h (100 years) that corresponds to an AFR of 1%. Also, for an average reserved rebuild bandwidth b of 100 MB/s, the mean rebuild time of a device is $E(X) = c/b = 55.5$ h, such that $\lambda/\mu = 6.3 \times 10^{-5} \ll 1$, which, according to (34), is a condition that ensures the accuracy of the reliability results obtained. Moreover, it is assumed that the maximum network rebuild bandwidth is sufficiently large ($B_{\max} \geq n\,b = 6.4$ GB/s), that the rebuild time distribution is deterministic, such that $E(X^k) = [E(X)]^k$, and that sector errors are correlated with $\bar{B} \approx 1$. From (6), it follows that $f_{\text{cor}} \approx 1$, which implies that the obtained results also apply to the case of independent sector errors.

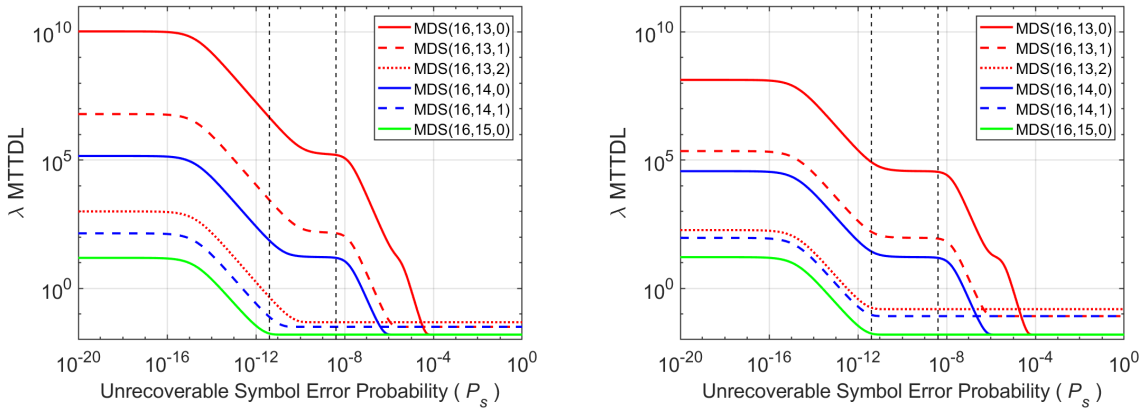
The probability of data loss P_{DL}, which does not depend on the entity size, is determined by (49) as a function of P_s and shown in Figure 8 for the declustered placement scheme ($k = n = 64$) for various MDS-coded configurations with $m = 16$, $l = 13$, and varying values of d. The probabilities P_{UF_u} and P_{DF} are also shown, as obtained from (50) and (53), respectively. We observe that P_{DL} increases monotonically with P_s and, according to Remark 5, exhibits a number of $\tilde{r} - d$ plateaus. For $d = 0$, the four plateaus are obtained from (59) and (61) as follows: $[0, 1.75\times10^{-15})$, $(1.1\times10^{-10}, 1.58\times10^{-8})$, $(1.54\times10^{-6}, 3.68\times10^{-6})$, and $(3.83\times10^{-5}, 1]$. For $d = 1$, the 3 plateaus are $[0, 1.75\times10^{-15})$, $(7.33\times10^{-11}, 1.58\times10^{-8})$, and $(1.09\times10^{-6}, 1]$. For $d = 2$, the two plateaus are $[0, 1.75\times10^{-15})$, and $(3.66\times10^{-11}, 1]$. In the interval $[4.096\times10^{-12}, 4.096\times10^{-9}]$ of practical importance for P_s, which is indicated between the two vertical dashed lines, the probability of data loss P_{DL} and, by virtue of (27), the MTTDL are degraded by one order of magnitude.

Next, we assess the reliability for the declustered placement scheme ($k = n = 64$) for the MDS-coded configurations



(a) $d = 0$ (b) $d = 1$ (c) $d = 2$

Figure 8. Probability of data loss P_{DL} vs. P_s for $d = 0, 1, 2$; $m = 16$, $l = 13$, $(\tilde{r} = 4)$, $n = k = 64$, $\lambda/\mu = 0.00006$, $c = 20$ TB, and $s = 512$ B.



(a) $k = 64$ (declustered data placement scheme) (b) $k = 16$ (clustered data placement scheme)

Figure 9. Normalized MTTDL vs. P_s for various MDS(m, l, d) codes; $n = 64$, $\lambda/\mu = 0.00006$, $c = 20$ TB, and $s = 512$ B.

considered in [7] with $m = 16$ and varying values of l and d. These configurations are denoted by MDS(m,l,d) and the corresponding results are shown in Figures 9, 10, and 11 by solid lines for $d = 0$ (no lazy rebuild employed), dashed lines for $d = 1$ and dotted lines for $d = 2$. Six configurations are considered: MDS(16,13,0), MDS(16,13,1), MDS(16,13,2), MDS(16,14,0), MDS(16,14,1), and MDS(16,15,0), for each of the declustered and clustered data placement schemes. In particular, for the clustered placement scheme, the MDS(16,15,0) and MDS(16,14,0) configurations correspond to the RAID-5 and RAID-6 systems.

The normalized λ MTTDL measure, which does not depend on the entity size, is obtained from (27) as a function of P_s and shown in Figure 9(a) for the declustered data placement scheme. The MTTDL for the MDS(16,13,0), MDS(16,13,1), and MDS(16,13,2) configurations is depicted by the red curves and is obtained from the probability of data loss shown in Figure 8. We observe that MTTDL decreases monotonically with P_s and, according to Remark 5, exhibits $\tilde{r} - d$ plateaus. In the interval of interest for P_s, MTTDL is degraded by orders of magnitude. Increasing the number of parities (reducing l) improves reliability by orders of magnitude. By contrast, employing lazy rebuild degrades reliability by orders of magnitude. Moreover, for equivalent systems, such as MDS(16,15,0), MDS(16,14,1) and MDS(16,13,2), MTTDL increases as d increases. We call *equivalent systems* those that employ a given codeword length m and have the same number $m - l - d$ of exposure levels at which the rebuild process is active.

The normalized λ MTTDL measure for the clustered data placement scheme is shown in Figure 9(b). We observe that the declustered placement scheme achieves a significantly higher MTTDL than the clustered one.

The normalized EAFEL$/\lambda$ reliability metric corresponding to the declustered data placement scheme is obtained from (29) and shown in Figure 10(a) for a fixed entity size of $e_s = 10$ GB. In the interval $[10^{-15}, 10^{-12}]$ of interest for P_{b}, EAFEL is degraded by orders of magnitude. Note that in the case of fixed-size entities, the values of the EAFEL and EAFEDL metrics are the same, because the fraction of lost entities reflects the fraction of lost user data.

Next, we consider the case of a discrete bimodal distribution for the entity size, with $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB, and probabilities $v_1 \approxeq 0.99$ and $v_2 \approxeq 0.01$ chosen such that the average entity size $E(e_s)$ is $v_1\, e_{s,1} + v_2\, e_{s,2} = 10$ GB, which is the same as the entity size e_s in the fixed-entity-size case considered previously. From (16), it follows that the average shard size $E(J)$ remains the same, which, according to (30), implies that the number N_E of entities in the system remains the same as in the fixed-entity-size case. The resulting EAFEL is shown in Figure 10(b). Comparing the case of bimodal entity sizes with that of fixed entity sizes, we observe that, for $P_{\text{b}} < 10^{-14}$, reliability remains essentially the same, whereas for higher values of P_{b}, EAFEL is reduced. The reason for that

is the following. For very small values of P_{b}, there can be at most one codeword lost, which results in one lost entity. Thus, the fraction of lost entities is $1/N_E$ in both cases. However, the lost entity in the fixed case has a size of 10 GB which is different from that of the lost entity in the bimodal case, which is either 1 MB or 1 TB. In fact, the size of the lost entity in the bimodal case is almost surely 1 TB, because the probability of this event is $v_2\, e_{s,2}/E(e_s) \approx 1$. Consequently, the size of 1 TB of the lost entity in the bimodal case is 100 times larger than that of 10 GB of the entity lost in the fixed case. This is reflected in Figure 10(c) that shows the EAFEDL metric. Note that for $P_{\mathrm{b}} = 10^{-15}$, indicated by the left vertical dashed line, EAFEDL is about 100 times larger than EAFEL. Consequently, in the case of variable size entities, it is more appropriate to consider the EAFEDL rather than the EAFEL metric, because it captures the amount of lost user data. Also, Figures 10(b) and 10(c) confirm Corollary 5 according to which, for small values of P_{b} such that $P_{\mathrm{b}} \ll P_{\check{s},\tilde{r}}^{(\tilde{r})}/s$, the EAFEL and EAFEDL metrics tend to the same value. This holds because, when $P_{\mathrm{b}} = 0$, the fraction of lost entities reflects the fraction of lost user data.

Clearly, the vulnerability of entities to loss increases with their size, which implies that lost entities are most likely large rather than small. For the case of the bimodal entity sizes, and for $v_2 \cong 0.01$, the number of the large 1-TB entities is significantly smaller than that of the 1-MB entities. We therefore deduce that the fraction of lost entities in the bimodal case is smaller than that for the fixed case, and this is more pronounced for larger values of P_{b}, as it is reflected by the EAFEL metric. By contrast, EAFEDL is larger in the bimodal case compared to the fixed case for the entire range of bit error rates. We therefore deduce that increasing the variability of the entity sizes, while keeping their average constant, results in degraded EAFEDL, but improved EAFEL, which is misleading. Clearly, the EAFEL metric that assesses the fraction of lost entities does not account for their size and the corresponding amount of lost user data and this led us to introduce the EAFEDL metric.

By observing Figures 11(a), 11(b) 11(c) that show the reliability results for the case of clustered placement, we arrive to the same conclusions. From the above discussion, it follows that in the case of variable size entities, it is important to consider the EAFEDL rather than the EAFEL metric.

The expected fraction of lost entities $E(Y)/N_E$ is obtained from (64) and shown in Figure 12 for the declustered placement scheme ($k = n = 64$) for various MDS-coded configurations with $m = 16$, $l = 13$, and varying values of d. The expected fractions of lost entities $E(Y_{\mathrm{UF}_u})/N_E$ and $E(Y_{\mathrm{DF}})/N_E$ are also shown as determined by (65) and (69), respectively. We observe that each of the $E(Y_{\mathrm{UF}_u})/N_E$ curves exhibits two plateaus owing to the bimodal nature of the entity sizes. According to Remark 13, $E(Y)$ and EAFEL degrade when P_s is greater than $P_{\check{s},\tilde{r}}^{(\tilde{r})}$, which for deterministic rebuild times and in the absence of network rebuild bandwidth constraints, by virtue of (82), is equal to 1.3×10^{-13}. For a symbol size of 512 bytes, the corresponding unrecoverable bit error probability is $P_{\mathrm{b}} \approx P_{\check{s},\tilde{r}}^{(\tilde{r})} \,/\, (512 \times 8) = 1.3 \times 10^{-13} \,/\, 4096 = 3.18 \times 10^{-17}$. This is depicted by the red curves in Figures 12 and 10(b).

The expected fraction of the effective amount of lost user data $E(\breve{Q})/U$ is obtained from (84) and shown in Figure 13. The expected fractions of the effective amounts of lost user data $E(\breve{Q}_{\mathrm{UF}})/U$ and $E(\breve{Q}_{\mathrm{DF}})/U$ are also shown as determined by (85) and (88), respectively. Despite the bimodal nature of the entity sizes, we observe that in this case each of the $E(\breve{Q}_{\mathrm{UF}})/U$ curves exhibits only a single plateau. According to Remark 17, $E(\breve{Q})$ and EAFEDL degrade when P_s is greater than $P_{\check{s},\tilde{r}}^{(\tilde{r})}$, which for deterministic rebuild times and in the absence of network rebuild bandwidth constraints, by virtue of (98), is equal to 1.3×10^{-15}. For a symbol size of 512 bytes, this degradation occurs when the unrecoverable bit error probability P_{b} is greater than $P_{\check{s},\tilde{r}}^{(\tilde{r})} \,/\, (512 \times 8) = 1.3 \times 10^{-15} \,/\, 4096 = 3.18 \times 10^{-19}$. This is depicted by the red curves in Figures 13 and 10(c). Note also that for extremely small values of P_{b}, such that $P_{\mathrm{b}} \ll 3.18\times10^{-19}$, and according to Corollary 5, it holds that $E(\breve{Q})/U \approx E(Y)/N_E$. This also holds when $P_{\mathrm{b}} \to 1$.

The effect of symbol size on reliability is assessed by considering the case of a large 5-MB symbol size. The corresponding normalized EAFEL$/\lambda$ and EAFEDL$/\lambda$ reliability metrics are shown in Figures 14 and 15. As expected, comparing these results with those shown in Figures 10 and 11, system reliability degrades compared to the case of a smaller symbol size. This degradation applies to both the EAFEL and EAFEDL reliability metrics.

Next, we assess the system reliability for the CERN file size distribution [22] that was considered in [23] and listed in Table IV shown in Appendix B. For the file sizes uniformly distributed within the bins, the mean is 843 MB, the standard deviation is 2.8 GB, the second moment is 8.9 GB^2 and the coefficient of variation is equal to 3.39. It turns out that the reliability metrics are extremely well approximated by considering the file sizes $e_{s,j}$ to be the bin mean sizes, such that $E_s = 38$. In this case, the mean is 843 MB, the standard deviation is 2.8 GB, the second moment is 8.5 GB^2 and the coefficient of variation is 3.37. The corresponding reliability results are shown in Figures 16 and 17. In all cases considered, the reliability level achieved by the declustered data placement scheme is higher than that of the clustered one.

VII. Real-World Erasure Coding Schemes

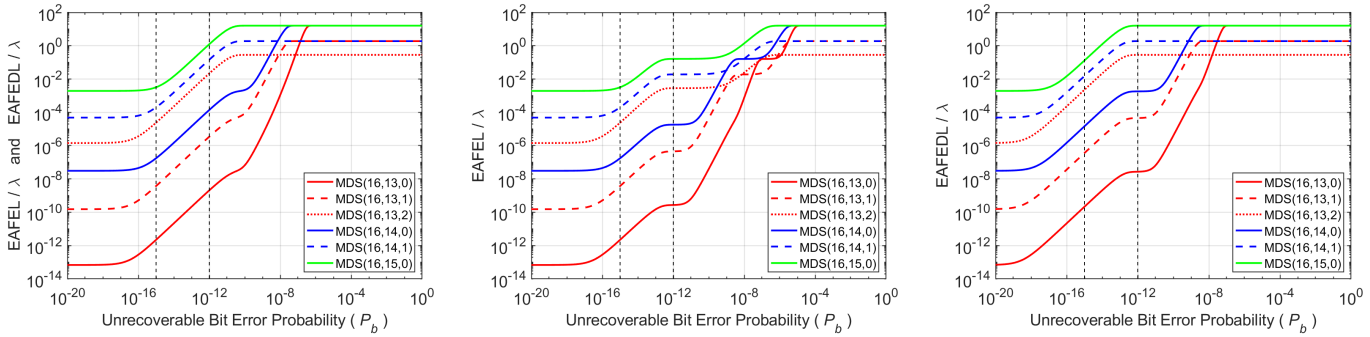
Here we assess the reliability of systems that store files whose size is distributed according to the CERN distribution listed in Table IV and shown in Figure 24(a). In particular, we assess the reliability of the practical systems considered in [4] that store an amount of $U = 1.2$ PB user data on devices (disks) whose capacity is $c = 20$ TB. This amount of user data can therefore be stored on $U/c = 60$ devices. The system comprises n devices, where n is determined using (1) as follows:

$$n = \frac{U}{c}\,\frac{m}{l} = 60\,\frac{m}{l}\,. \tag{100}$$

Subsequently, we consider the following real-world erasure coding schemes:

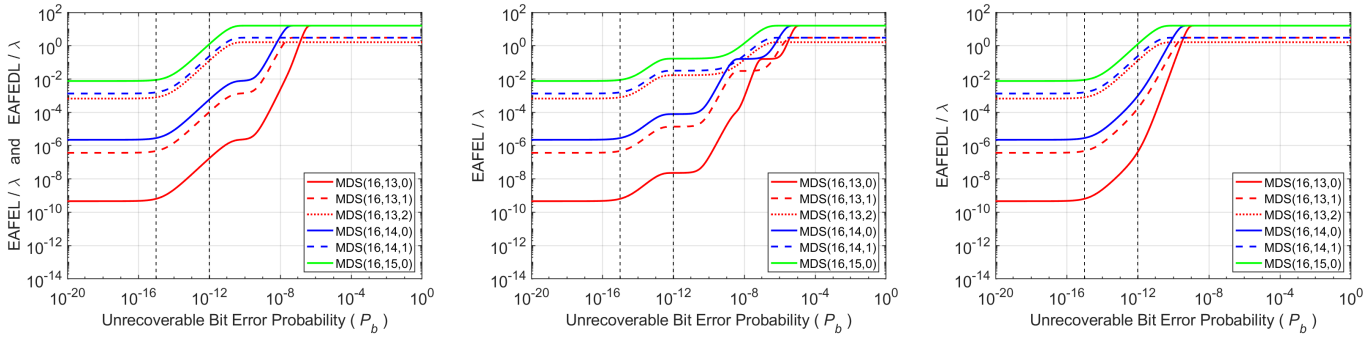
1) the 3-way replication (triplication) scheme that was initially used by Google's GFS, Microsoft® Azure[1], and

[1]Microsoft is a trademark of Microsoft Corporation in the United States, other countries, or both.



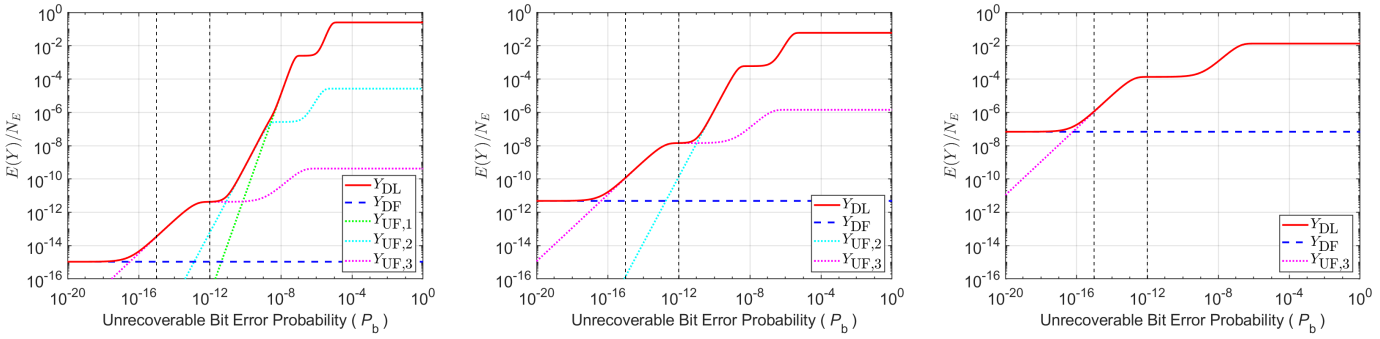
(a) Fixed Entity Size: $e_s = 10$ GB (b) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB (c) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB

Figure 10. Normalized EAFEL and EAFEDL vs. P_{b} for various MDS(m, l, d) codes; symbol size $s = 512$ B, declustered data placement.



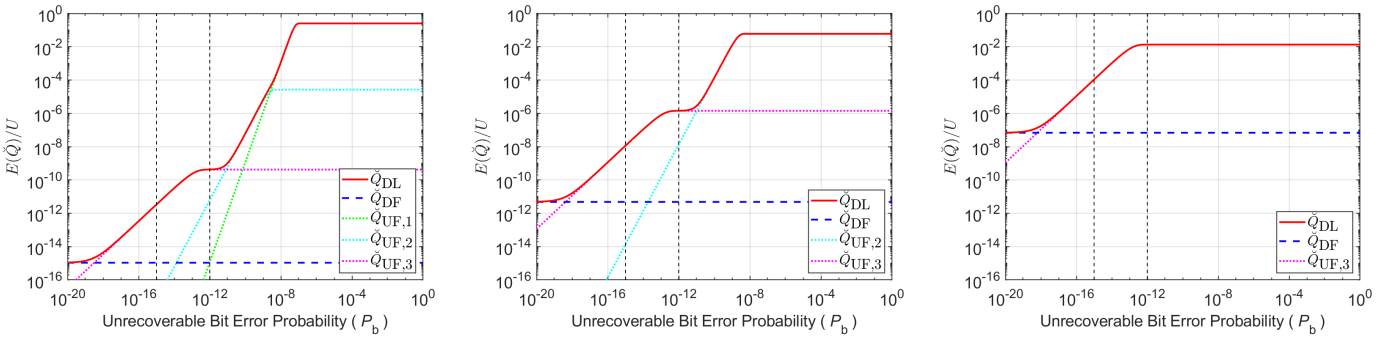
(a) Fixed Entity Size: $e_s = 10$ GB (b) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB (c) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB

Figure 11. Normalized EAFEL and EAFEDL vs. P_{b} for various MDS(m, l, d) codes; symbol size $s = 512$ B, clustered data placement.



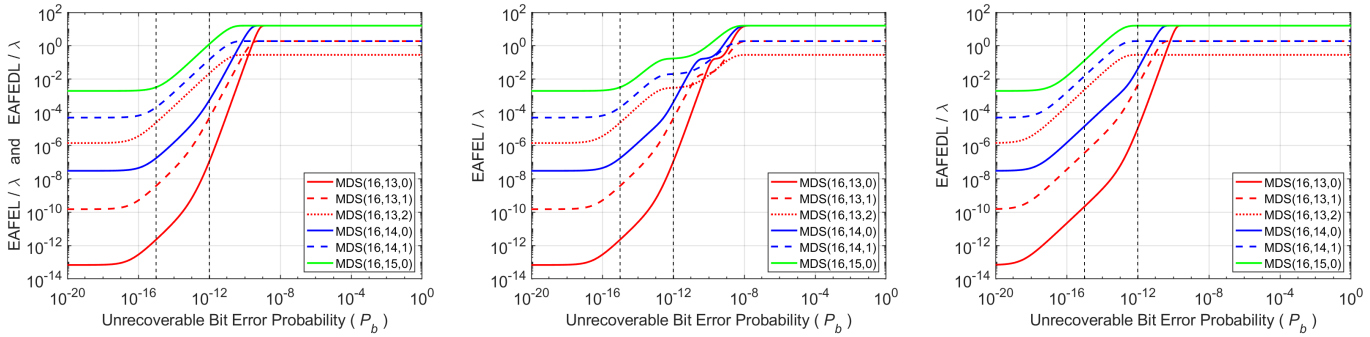
(a) $d = 0$ (b) $d = 1$ (c) $d = 2$

Figure 12. Normalized $E(Y)$ vs. P_s for $d = 0, 1, 2$; $m = 16$, $l = 13$, $(\tilde{r} = 4)$, $n = k = 64$, $c = 20$ TB, and $s = 512$ B, bimodal entity sizes.



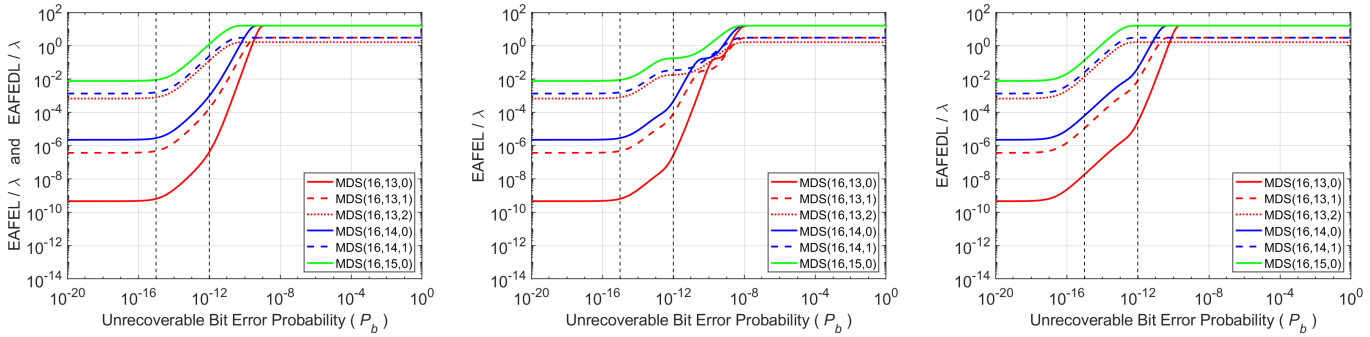
(a) $d = 0$ (b) $d = 1$ (c) $d = 2$

Figure 13. Normalized $E(\breve{Q})$ vs. P_s for $d = 0, 1, 2$; $m = 16$, $l = 13$, $(\tilde{r} = 4)$, $n = k = 64$, $c = 20$ TB, and $s = 512$ B, bimodal entity sizes.



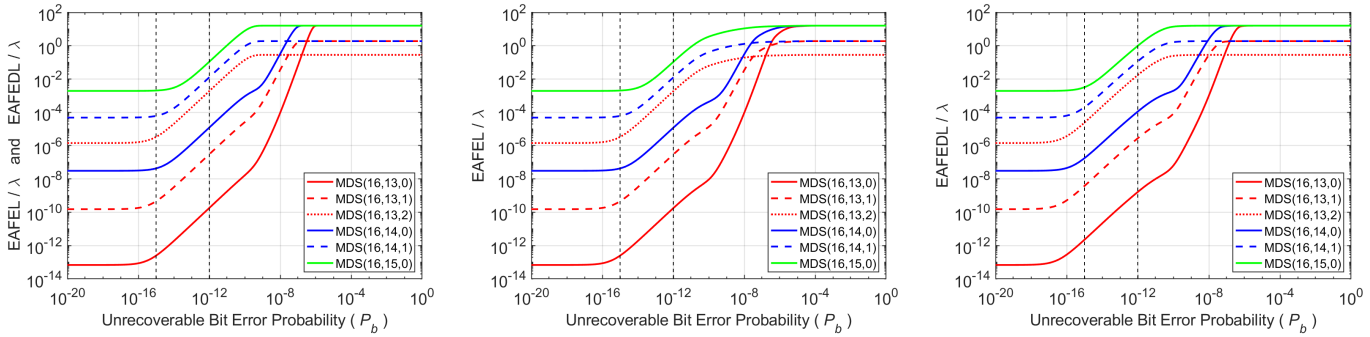
(a) Fixed Entity Size: $e_s = 10$ GB (b) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB (c) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB

Figure 14. Normalized EAFEL and EAFEDL vs. P_{b} for various MDS(m, l, d) codes; symbol size $s = 5$ MB, declustered data placement.



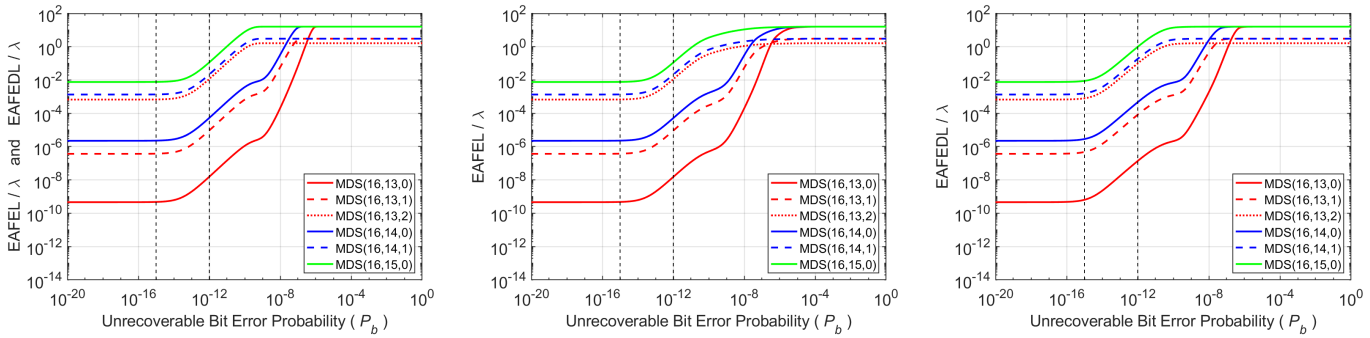
(a) Fixed Entity Size: $e_s = 10$ GB (b) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB (c) Bimodal Entity Sizes: $e_{s,1} = 1$ MB, $e_{s,2} = 1$ TB

Figure 15. Normalized EAFEL and EAFEDL vs. P_{b} for various MDS(m, l, d) codes; symbol size $s = 5$ MB, clustered data placement.



(a) Fixed File Size: $e_s = 843$ MB (b) CERN File Sizes; $E(e_s) = 843$ MB (c) CERN File Sizes; $E(e_s) = 843$ MB

Figure 16. Normalized EAFEL and EAFEDL vs. P_{b} for various MDS(m, l, d) codes; symbol size $s = 512$ B, declustered data placement.



(a) Fixed File Size: $e_s = 843$ MB (b) CERN File Sizes; $E(e_s) = 843$ MB (c) CERN File Sizes; $E(e_s) = 843$ MB

Figure 17. Normalized EAFEL and EAFEDL vs. P_{b} for various MDS(m, l, d) codes; symbol size $s = 512$ B, clustered data placement.

Facebook. In this case, $m = 3$, $l = 1$, with a corresponding storage efficiency of $s_{\text{eff}} = 33\%$. According to (100), this scheme requires the employment of $n = 180$ devices.

2) the RS(9,6) erasure coding scheme employed by Google's GFS as well as QFS [24][25], which for $m = 9$ and $l = 6$ achieves a storage efficiency of $s_{\text{eff}} = 66\%$ and requires a number of $n = 90$ devices.
3) the MDS(16,12) erasure coding scheme akin to the LRC(16,12) code used by Microsoft® Azure [26], which for $m = 16$ and $l = 12$ achieves a storage efficiency of $s_{\text{eff}} = 75\%$ and requires a number of $n = 80$ devices.
4) the RS(14,10) erasure coding scheme employed by Facebook [27], which for $m = 14$ and $l = 10$ achieves a storage efficiency of $s_{\text{eff}} = 71\%$ and requires a number of $n = 84$ devices.

We proceed to assess the reliability of the four erasure coding schemes assuming a 512-B symbol size and for the declustered data placement scheme, which achieves a superior data reliability. The results for the EAFEL and EAFEDL reliability metrics are shown in Figures 18 and 19, respectively. We observe that, in all cases, EAFEDL is larger than EAFEL.

First, we assess the reliability of the 3-way replication (triplication) scheme. Figure 19(a) shows that, in the interval of interest for P_{b}, EAFEDL ranges between 10^{-12} and 10^{-9}. In particular, when P_{b} is larger than 10^{-14}, EAFEDL is larger than 10^{-11}, the durability of eleven nines (11 9s) targeted by the Amazon S3 [28]. Employing the MDS(9,6) coding scheme, improves reliability by orders of magnitude. Figure 19(b) shows that, in the interval of interest for P_{b}, EAFEDL ranges between 10^{-16} and 10^{-13}. Further reliability improvement is achieved by employing the MDS(16,12) coding scheme. According to Figure 19(c), in the interval of interest for P_{b}, EAFEDL ranges between 10^{-20} and 10^{-17}. Superior reliability is achieved by employing the MDS(14,10) coding scheme. Figure 19(d) shows that, in the interval of interest for P_{b}, EAFEDL ranges between 10^{-21} and 10^{-18}.

Also, Figures 18 and 19 confirm Corollary 5 according to which, for small values of P_{b} such that $P_{\text{b}} \ll P_{\check{s},\check{r}}^{(\check{r})}/s$, the EAFEL and EAFEDL metrics tend to the same value. However, in the interval of interest for P_{b}, by employing the 3-way replication, MDS(9,6), and MDS(16,12) coding schemes, both EAFEL and EAFEDL improve by four orders of magnitude, successively. By contrast, employing the MDS(14,10) coding scheme results in a reliability improvement of only one order of magnitude of that achieved by the MDS(16,12) coding scheme.

We proceed to assess the reliability of the four erasure coding schemes for the declustered data placement scheme by considering the case of a large 5-MB symbol size. The results for the EAFEL and EAFEDL reliability metrics are shown in Figures 20 and 21, respectively. We observe that, in all cases, EAFEDL is larger than EAFEL.

First, we assess the reliability of the 3-way replication scheme. Figure 21(a) shows that, in the interval of interest for P_{b}, EAFEDL ranges between 10^{-12} and 10^{-7}. In particular, when P_{b} is larger than 10^{-14}, EAFEDL is larger than 10^{-11}, the durability of eleven nines (11 9s) targeted by the Amazon S3. Comparing Figures 18(a) and 20(a) as well as Figures 19(a) and 21(a), we observe that the increased symbol size affects EAFEL and EAFEDL only for P_{b} values in the interval $(10^{-14}, 10^{-7})$.

Employing the MDS(9,6) coding scheme, improves reliability by orders of magnitude. Figure 21(b) shows that, in the interval of interest for P_{b}, EAFEDL ranges between 10^{-16} and 10^{-10}. In particular, when P_{b} is larger than 8×10^{-13}, EAFEDL is larger than 10^{-11}, the durability of eleven nines (11 9s) targeted by the Amazon S3 [28]. Comparing Figures 18(b) and 20(b) as well as Figures 19(b) and 21(b), we observe that the increased symbol size affects EAFEL and EAFEDL only for P_{b} values in the intervals $(10^{-15}, 10^{-5})$ and $(10^{-15}, 10^{-6})$, respectively.

Further reliability improvement is achieved by employing the MDS(16,12) coding scheme. According to Figure 19(c), in the interval of interest for P_{b}, EAFEDL ranges between 10^{-20} and 10^{-13}. Comparing Figures 18(c) and 20(c) as well as Figures 19(c) and 21(c), we observe that the increased symbol size affects EAFEL and EAFEDL only for P_{b} values in the intervals $(10^{-15}, 10^{-5})$ and $(10^{-15}, 10^{-6})$, respectively.

Superior reliability is achieved by employing the MDS(14,10) coding scheme. Figure 21(d) shows that, in the interval of interest for P_{b}, EAFEDL ranges between 10^{-21} and 10^{-13}. Comparing Figures 18(d) and 20(d) as well as Figures 19(d) and 21(d), we observe that the increased symbol size affects EAFEL and EAFEDL only for P_{b} values in the interval $(10^{-15}, 10^{-5})$.

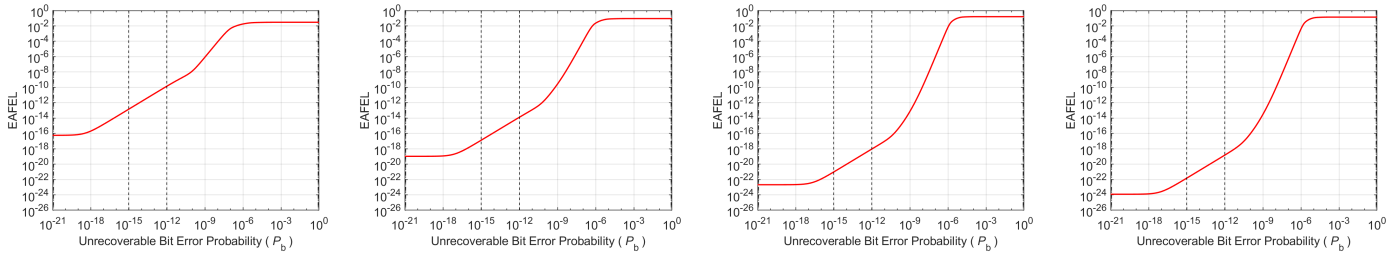
Figures 20 and 21 confirm Corollary 5 according to which, for small values of P_{b} such that $P_{\text{b}} \ll P_{\check{s},\check{r}}^{(\check{r})}/s$, the EAFEL and EAFEDL metrics tend to the same value. However, for $P_{\text{b}} = 10^{-15}$, by employing the 3-way replication, MDS(9,6), and MDS(16,12) coding schemes, both EAFEL and EAFEDL improve by four orders of magnitude, successively. By contrast, employing the MDS(14,10) coding scheme results in a reliability improvement of only one order of magnitude of that achieved by the MDS(16,12) coding scheme. Also, for $P_{\text{b}} = 10^{-12}$, by employing the 3-way replication, MDS(9,6), and MDS(16,12) coding schemes, both EAFEL and EAFEDL improve by three orders of magnitude, successively. By contrast, employing the MDS(14,10) coding scheme does not achieve any reliability improvement compared to the MDS(16,12) coding scheme.

A. Reliability Improvement

The improvement of the EAFEL and EAFEDL reliability metrics achieved by the erasure coding schemes considered over the initial 3-way replication is shown in Figures 22 and 23.

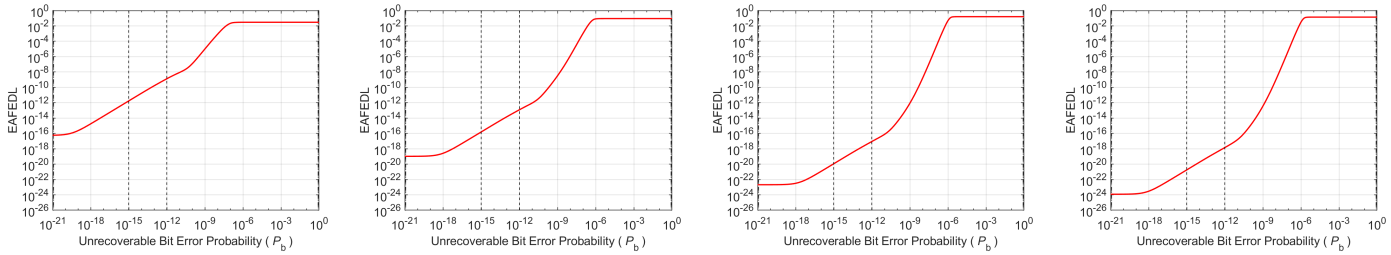
For the declustered data placement and for a symbol size of 512 B, Figure 22 demonstrates that in the interval of interest, the MDS(9,6) erasure coding scheme improves reliability by four orders of magnitude, the MDS(16,12) erasure coding scheme improves reliability by eight orders of magnitude, whereas the MDS(14,10) erasure coding scheme improves reliability by nine orders of magnitude.

For the declustered data placement and for a symbol size of 5 MB, Figure 23 demonstrates that in the interval of interest, the reliability improvement achieved by the erasure coding schemes considered varies. In particular, for $P_{\text{b}} = 10^{-15}$, the



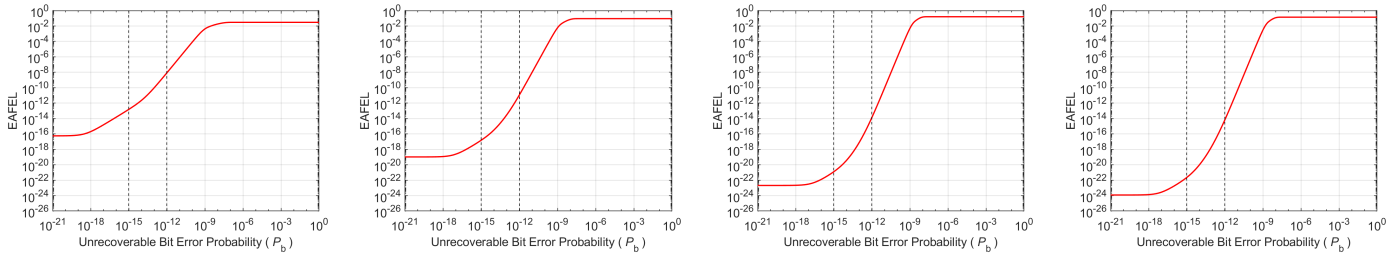
(a) MDS(3,1) (b) MDS(9,6) (c) MDS(16,12) (d) MDS(14,10)

Figure 18. EAFEL vs. P_s for various MDS coding schemes; symbol size $s = 512$ B, declustered data placement.



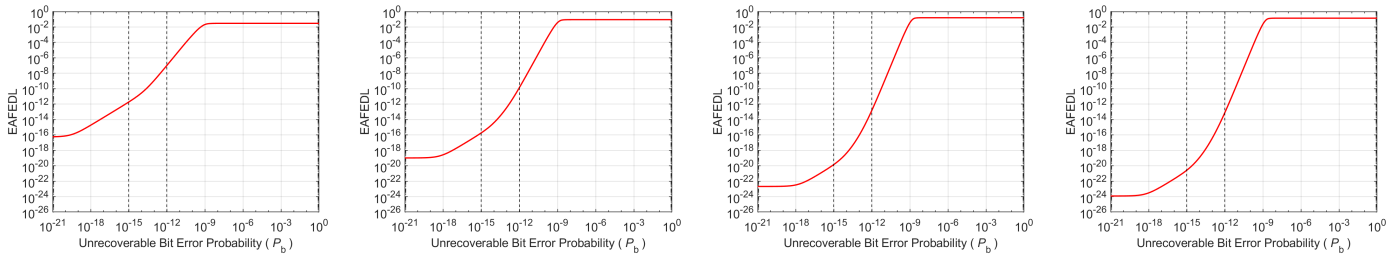
(a) MDS(3,1) (b) MDS(9,6) (c) MDS(16,12) (d) MDS(14,10)

Figure 19. EAFEDL vs. P_s for various MDS coding schemes; symbol size $s = 512$ B, declustered data placement.



(a) MDS(3,1) (b) MDS(9,6) (c) MDS(16,12) (d) MDS(14,10)

Figure 20. EAFEL vs. P_s for various MDS coding schemes; symbol size $s = 5$ MB, declustered data placement.



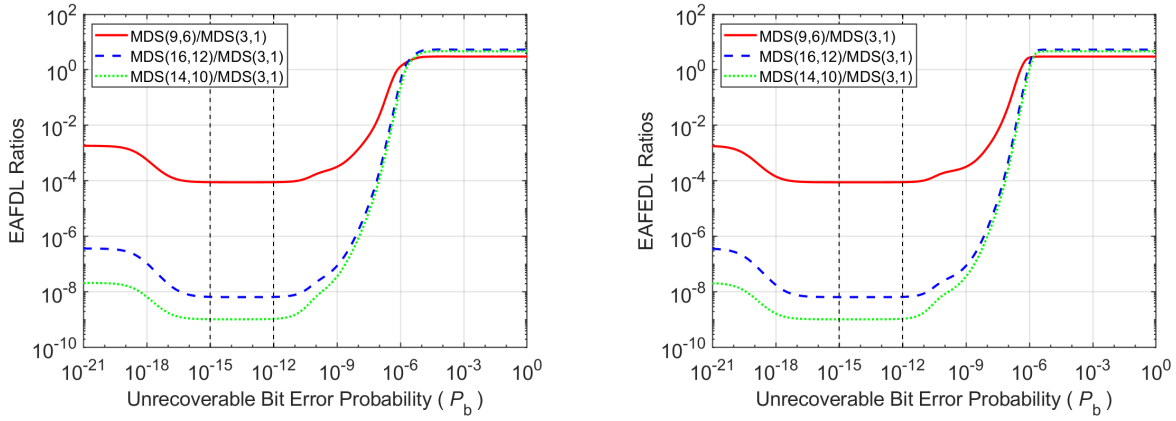
(a) MDS(3,1) (b) MDS(9,6) (c) MDS(16,12) (d) MDS(14,10)

Figure 21. EAFEDL vs. P_s for various MDS coding schemes; symbol size $s = 5$ MB, declustered data placement.

MDS(9,6) erasure coding scheme improves reliability by four orders of magnitude, the MDS(16,12) erasure coding scheme improves reliability by eight orders of magnitude, whereas the MDS(14,10) erasure coding scheme improves reliability by nine orders of magnitude. However, for $P_b = 10^{-12}$, the MDS(9,6) erasure coding scheme improves reliability by three orders of magnitude, whereas the MDS(16,12) and MDS(14,10) erasure coding scheme improve reliability by six orders of magnitude.

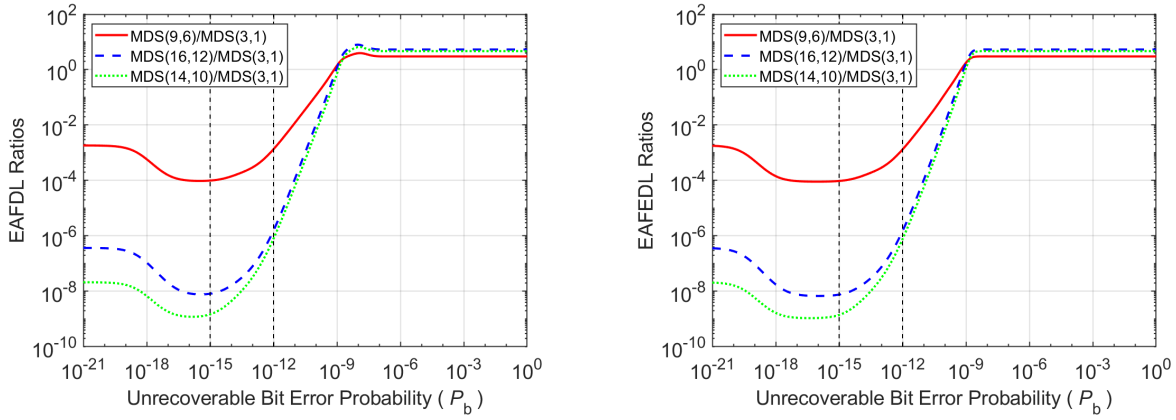
VIII. Conclusions

The Expected Annual Fraction of Entity Loss EAFEL metric assesses the durability of data storage systems at an entity, say file, object, or block level. Contrary to the Mean Time to Data Loss (MTTDL) metric, EAFEL is affected by the distribution of the number of codewords that entities span. The distribution of this number was obtained analytically in closed form for the segregated and the random entity placement cases as a function of the size of the entities and the frequency of their occurrence. It was also demonstrated that, in certain cases



(a) EAFEL ratios (b) EAFEDL ratios

Figure 22. Ratios of the EAFEL and EAFEDL metrics for the MDS(9,6), MDS(16,12), and MDS(14,10) schemes to those corresponding to the 3-way replication scheme; symbol size $s = 512$ B, declustered data placement.



(a) EAFEL ratios (b) EAFEDL ratios

Figure 23. Ratios of the EAFEL and EAFEDL metrics for the MDS(9,6), MDS(16,12), and MDS(14,10) schemes to those corresponding to the 3-way replication scheme; symbol size $s = 5$ MB, declustered data placement.

of deterministic entity placements of variable-size entities, this distribution also depends on their actual placement.

To evaluate the durability of storage systems in the case of variable-size entities, a new reliability metric was introduced, the Expected Annual Fraction of Effective Data Loss (EAFEDL), which assesses the fraction of lost user data annually at the entity level. The MTTDL, EAFEL, and EAFEDL metrics were obtained analytically for erasure-coding redundancy schemes and for the clustered, declustered, and symmetric data placement schemes. Closed-form expressions capturing the effect of unrecoverable latent errors and lazy rebuilds were derived. We established that the reliability of storage systems is adversely affected by the presence of latent errors and that the declustered data placement scheme offers superior reliability. We demonstrated that an increased variability of entity sizes results in improved EAFEL, but degraded EAFEDL. We also established that EAFEL and EAFEDL are adversely affected by the symbol size. We considered several real-world erasure coding schemes and demonstrated their efficiency. The analytical reliability results obtained enable the identification of erasure-coded redundancy schemes that ensure a desired level of reliability.

This work has the potential to be applied for further studies of data storage reliability and it is particularly relevant for tape storage reliability, which is a subject of further investigation [29].

Appendix A

Proof of Corollary 1.

From Eqs. (57) and (65) of [21], (61) of [21] yields

$$P_{\mathrm{UF}_u}(R_{d+1}) \approx -(\lambda b_{d+1} R_{d+1})^{u-d-1} \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i} V_i^{u-1-i} \right) \cdot \cdot \left(\sum_{j=1}^{\infty} \frac{\log(\hat{q}_u)^j}{(u-d-1+j)!} \right). \quad (101)$$

From (55) and for $P_s \ll P^*_{s,u}$, it follows that $\hat{q}_u \approx 1$. Furthermore, $\log(\hat{q}_u) = -(1-\hat{q}_u) + O((1-\hat{q}_u)^2) \approx -(1-\hat{q}_u)$. Consequently, by virtue of (55), it holds that $\log(\hat{q}_u) \approx -Z_u P_s^{\tilde{r}-u}$. For small values of P_s, all the terms of the summation in (101) are negligible compared with the first one.

Therefore, from the above, it follows that

$$P_{\text{UF}_u}(R_{d+1}) \;\approx\; (\lambda b_{d+1} R_{d+1})^{u-d-1} \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i}\, V_i^{u-1-i} \right) \cdot \cdot \frac{Z_u\, P_s^{\tilde{r}-u}}{(u-d)!} \,. \tag{102}$$

Unconditioning (102) on R_{d+1}, and using (33) and (44), yields

$$P_{\text{UF}_u} \;\approx\; \frac{(\lambda\, c \prod_{j=1}^{d} V_j)^{u-d-1}}{(u-d)!} \left(\prod_{i=d+1}^{u-1} \frac{\tilde{n}_i}{b_i}\, V_i^{u-1-i} \right) \cdot \cdot \frac{E(X^{\tilde{r}-d-1})}{[E(X)]^{\tilde{r}-d-1}}\, Z_u\, P_s^{\tilde{r}-u} \,. \tag{103}$$

Consider the direct path $\overrightarrow{UF_u} = 1 \to 2 \to \cdots \to u \to \text{UF}$. Then the probability $P_{\text{UF}_u}(R_{d+1}, \vec{\alpha}_{u-1})$ of entering exposure level u through vector $\vec{\alpha}_{u-1} \triangleq (\alpha_1, \ldots, \alpha_{u-1})$ and encountering an unrecoverable failure during the rebuild process at this exposure level, given a rebuild time R_{d+1}, is determined by [21, Eq. (46)]

$$P_{\text{UF}_u}(R_{d+1}, \vec{\alpha}_{u-1}) = P_u(R_{d+1}, \vec{\alpha}_{u-2}) \cdot P_{u\to\text{UF}}(R_{d+1}, \vec{\alpha}_{u-1}) \,. \tag{104}$$

where P_u is the probability of entering exposure level u and $P_{u\to\text{UF}}$ is the probability of encountering an unrecoverable failure during the rebuild process at this exposure level. We now proceed to calculate $P_{u\to\text{UF}}(R_{d+1}, \vec{\alpha}_{u-1})$. Upon entering exposure level u, the rebuild process attempts to restore the C_u most-exposed codewords, each of which has $m-u$ remaining symbols. The probability q_u that a codeword can be restored is determined by (52). Note that, if a codeword is corrupted, then at least one of its l user-data symbols is lost. When symbol errors are independent, codewords are independently corrupted. Consequently, the conditional probability $P_{\text{UF}|C_u}$ of encountering an unrecoverable failure during the rebuild process of the C_u codewords is determined by $1 - q_u^{C_u}$ [21, Eq. (58)]. In the case of correlated symbol errors, $P_{\text{UF}|C_u}$ is determined by $1 - q_u^{f_{\text{cor}}\, C_u}$ [5, Eq. (98)]. Consequently, it holds that

$$P_{\text{UF}|C_u} = 1 - q_u^{f_{\text{cor}}\, C_u} \,, \quad \text{for } u = d+1, \ldots, \tilde{r} \,. \tag{105}$$

Substituting (46) into (105) and using (51) yields

$$P_{u\to\text{UF}}(R_{d+1}, \vec{\alpha}_{u-1}) \approx 1 - q_u^{C \prod_{j=1}^{u-1} V_j\, \alpha_j} = 1 - \hat{q}_u^{\prod_{j=1}^{u-1} \alpha_j} \,. \tag{106}$$

Substituting (106) into (104) yields

$$P_{\text{UF}_u}(R_{d+1}, \vec{\alpha}_{u-1}) \;\approx\; P_u(R_{d+1}, \vec{\alpha}_{u-2}) \left[1 - \hat{q}_u^{\prod_{j=1}^{u-1} \alpha_j} \right] . \tag{107}$$

From (55) and for $P_s \gg P^*_{s,u}$, it follows that $\hat{q}_u \approx 0$, which by virtue of (106) implies that $P_{u\to\text{UF}}(R_1, \vec{\alpha}_{u-1}) \approx 1$. Consequently, it follows from (104) that $P_{\text{UF}_u} \approx P_u$. Also, substituting (56) into (103) yields (57), with the variable A_u determined by (58). In particular, $P_{\tilde{s},u}^{(\tilde{r})}$ is obtained from the approximation (57) $P_{\text{UF}_u} \approx A_u\, (P_{\tilde{s},u}^{(\tilde{r})})^{\tilde{r}-u} = P_u$ and using (2), (48), and (58).

□

APPENDIX B

Proof of Proposition 2.

Upon entering exposure level u ($u \geq d+1$), there are C_u most-exposed codewords to be recovered. As a shard size of s_s corresponds to J symbols, an entity size e_s corresponds to J codewords. Therefore, the average entity of size $E(e_s)$ determined by (13) corresponds to $E(J)$ codewords, with $E(J)$ determined by (16). Consequently, for the number E_u of entities to be recovered it holds that

$$E_u \;\approx\; \frac{C_u}{E(J)} \,, \quad \text{for } u = d+1, \ldots, \tilde{r}-1 \,. \tag{108}$$

Let K ($K \geq 1$) denote the number of codewords that an entity of size e_s spans or, equivalently, the number of symbols that a shard of size s_s spans. The entity is lost if any of these K codewords is permanently lost. Therefore, according to Eq. (98) of [5], the probability of recovering the entity is $q_u^{f_{\text{cor}}\, K}$, where q_u is the probability of restoring a codeword and is determined by (52), and f_{cor} accounts for the correlation of latent errors and is determined by Eq. (29) of [5]. Consequently, the probability $\tilde{q}_u|K$ of loss of an entity that spans K codewords is determined by

$$\tilde{q}_u|K \;=\; 1 - q_u^{f_{\text{cor}}\, K} \,. \tag{109}$$

Unconditioning (109) on K using (23) yields the probability $\tilde{q}_{s,u}(J)$ that the entity (for the shard size J) is lost, where $\tilde{q}_{s,u}(x)$ is determined by (68). Thus, using (4), the probability $\tilde{q}_u(e_s)$ that the entity is lost is determined by

$$\tilde{q}_u(e_s) \;=\; \tilde{q}_{s,u} \left(\frac{e_s}{l\, s} \right) \,. \tag{110}$$

For this entity, the expected amount $\breve{q}_u(e_s)$ of lost user data is

$$\breve{q}_u(e_s) \;=\; e_s\, \tilde{q}_u(e_s) \,. \tag{111}$$

From (12), the probability $\tilde{q}_u$ that an arbitrary entity is lost is

$$\tilde{q}_u = \sum_{j=1}^{E_s} \tilde{q}_u(e_{s,j})\, v_j \,, \tag{112}$$

which, using (110), yields (67).

Similarly, from (12), it follows that the expected amount $\breve{q}_u$ of lost user data of an arbitrary entity is determined by

$$\breve{q}_u = \sum_{j=1}^{E_s} \breve{q}_u(e_{s,j})\, v_j \,, \tag{113}$$

which, using (110) and (111), yields (87).

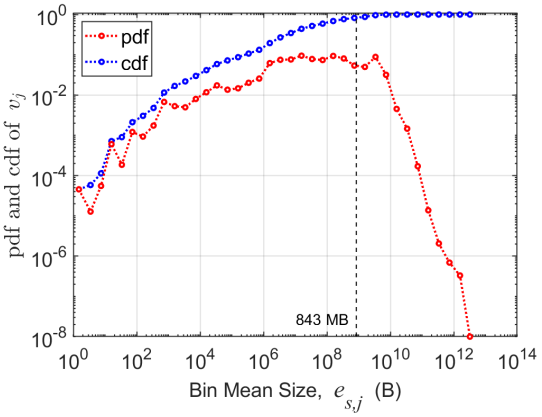
Remark 18: Note that (108) holds when $C_u \gg E(J)$. In the case where $C_u \ll E(J)$, it holds that $E_u = 1$, that is, a single entity is to be recovered. Let $\hat{e}_s$ denote its size. From the pdf of the lifetime of sampled intervals [30], we deduce that the pdf $\{\hat{v}_j\}$ of the size $\hat{e}_s$ of the sampled entity is no longer the typical $\{v_j\}$ pdf, but is determined by

$$\hat{v}_j \;=\; P(\hat{e}_s = e_{s,j}) = \frac{e_{s,j}\, v_j}{E(e_s)} \,, \quad \text{for } j = 1, 2, \ldots, E_s \,. \tag{114}$$

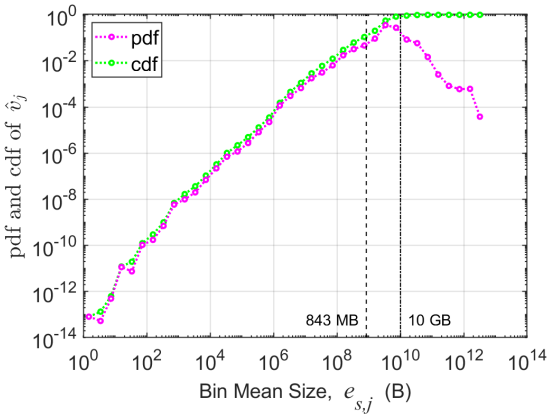
The $\{\hat{v}_j\}$ pdf of $\hat{e}_s$ is listed in Table IV and shown in Figure 24(b). For the file sizes uniformly distributed within the bins,

TABLE IV. CERN FILE SIZE e_s AND SAMPLED FILE SIZE $\hat{e}_s$ DISTRIBUTIONS

j	Bins			Bin Mean Size $e_{s,j}$	pdf v_j	pdf $\hat{v}_j$
1	1 B	–	2 B	2 B	0.00004559	0.000000000000081
2	2 B	–	5 B	4 B	0.00001275	0.000000000000053
3	5 B	–	10 B	8 B	0.00005533	0.000000000000492
4	10 B	–	22 B	16.0 B	0.00060401	0.000000000011464
5	22 B	–	46 B	34.0 B	0.00018569	0.000000000007489
6	46 B	–	100 B	73.0 B	0.00121244	0.000000000104989
7	100 B	–	215 B	157.5 B	0.00093013	0.000000000173774
8	215 B	–	464 B	339.5 B	0.00174431	0.000000000702464
9	464 B	–	1 KB	732.0 B	0.00675513	0.000000005865509
10	1 KB	–	2.154 KB	1.577 KB	0.00530524	0.000000009924249
11	2.154 KB	–	4.642 KB	3.398 KB	0.00496005	0.000000019992649
12	4.642 KB	–	10 KB	7.321 KB	0.00800625	0.000000069528117
13	10 KB	–	21.544 KB	15.772 KB	0.01174913	0.000000219813008
14	21.544 KB	–	46.416 KB	33.980 KB	0.01738480	0.000000700735281
15	46.416 KB	–	100 KB	73.208 KB	0.01359001	0.000001180155486
16	100 KB	–	215.443 KB	157.721 KB	0.01471745	0.000002753495549
17	215.443 KB	–	464.159 KB	339.801 KB	0.02018806	0.000008137296681
18	464.159 KB	–	1 MB	732.079 KB	0.02566358	0.000022286219101
19	1 MB	–	2.154 MB	1.577 MB	0.06221012	0.000116389428894
20	2.154 MB	–	4.642 MB	3.398 MB	0.07519022	0.000303072948937
21	4.642 MB	–	10 MB	7.321 MB	0.07654035	0.000664675346806
22	10 MB	–	21.544 MB	15.772 MB	0.09501620	0.001777665788444
23	21.544 MB	–	46.416 MB	33.980 MB	0.07847651	0.003163191377566
24	46.416 MB	–	100 MB	73.208 MB	0.07416942	0.006440862144930
25	100 MB	–	215.443 MB	157.721 MB	0.09371673	0.017533538933119
26	215.443 MB	–	464.159 MB	339.801 MB	0.08093624	0.032623369369260
27	464.159 MB	–	1 GB	732.079 MB	0.05399279	0.046887264039909
28	1 GB	–	2.154 GB	1.577 GB	0.04992384	0.093402916675691
29	2.154 GB	–	4.642 GB	3.398 GB	0.08871583	0.357591270942897
30	4.642 GB	–	10 GB	7.321 GB	0.03182476	0.276365775047813
31	10 GB	–	21.544 GB	15.772 GB	0.00452804	0.084715467164424
32	21.544 GB	–	46.416 GB	33.980 GB	0.00146156	0.058911819675084
33	46.416 GB	–	100 GB	73.208 GB	0.00017060	0.014814880370463
34	100 GB	–	215.443 GB	157.721 GB	0.00001375	0.002568882068470
35	215.443 GB	–	464.159 GB	339.801 GB	0.00000206	0.000829598407954
36	464.159 GB	–	1 TB	732.079 GB	0.00000069	0.000599130022577
37	1 TB	–	2.154 TB	1.577 TB	0.00000033	0.000616531696433
38	2.154 TB	–	4.310 TB	3.230 TB	0.00000001	0.000038314523896



(a) CERN file size distribution



(b) CERN sampled file size distribution

Figure 24. CERN file size distributions v_j and $\hat{v}_j$.

the mean $E(\hat{e}_s)$ is equal to 10.5 GB, the standard deviation is 53.1 GB, the second moment is 2,935 GB^2, and the coefficient of variation is equal to 5.05. By considering the file sizes $e_{s,j}$ to be the bin mean sizes, the mean $E(\hat{e}_s)$ is equal to 10.4 GB, the standard deviation is 52.6 GB, the second moment is 2,873 GB^2, and the coefficient of variation is equal to 5.03.

From the above discussion, and analogous to (112), it follows that the probability $\tilde{q}_{\hat{u}}$ that the single entity is lost is determined by

$$\tilde{q}_{\hat{u}} = \sum_{j=1}^{E_s} \tilde{q}_u(e_{s,j})\, \hat{v}_j \ . \tag{115}$$

Similarly, and analogous to (113), the expected amount $\breve{q}_{\hat{u}}$ of lost user data of the single entity is determined by

$$\breve{q}_{\hat{u}} = \sum_{j=1}^{E_s} \breve{q}_u(e_{s,j})\, \hat{v}_j \ . \tag{116}$$

Let Y_{U} be the number of lost entities and $\breve{Q}_{\mathrm{U}}$ the amount

of lost user data at exposure level u during the rebuild process of the C_u codewords. Then it holds that

$$E(Y_{\mathrm{U}}|C_u) \;=\; E_u\;\tilde{q}_u \;\overset{(108)}{\approx}\; \frac{C_u}{E(J)}\;\tilde{q}_u\;, \tag{117}$$

and

$$E(\breve{Q}_{\mathrm{U}}|C_u) \;=\; E_u\;\breve{q}_u \;\overset{(108)}{\approx}\; \frac{C_u}{E(J)}\;\breve{q}_u\;. \tag{118}$$

Note that $E(Y_{\mathrm{U}}|C_u)$, as determined by (117), can be obtained from Eq. (71) of [7] by replacing the shard size J with its average value $E(J)$. Consequently, (66) and (69) are obtained from the corresponding Eqs. (42) and (44) of [7] by replacing the shard size J with its average value $E(J)$.

Note also that $E(\breve{Q}_{\mathrm{U}}|C_u)$, as determined by (118), can be obtained from (117) by replacing the probability $\tilde{q}_u$ that an arbitrary entity is lost with its expected amount $\breve{q}_u$ of lost user data. Consequently, (86) is obtained from (66) by replacing $\tilde{q}_u$ with $\breve{q}_u$.

Remark 19: According to Remark 18, in the case where $C_u \ll E(J)$, it holds that $E_u = 1$, that is, a single entity is to be recovered. In this case, and considering (115), (116), (117), and (118), we have

$$E(Y_{\mathrm{U}}|C_u) \;=\; \tilde{q}_{\hat{u}}\;, \quad \text{for } C_u \ll E(J)\;. \tag{119}$$

and

$$E(\breve{Q}_{\mathrm{U}}|C_u) \;=\; \breve{q}_{\hat{u}}\;, \quad \text{for } C_u \ll E(J)\;. \tag{120}$$

According to (46), it holds that $C_u \;\approx\; C\,\prod_{i=1}^{u-1} V_i\,\alpha_i$. Consequently, condition $C_u \;\ll\; E(J)$ holds when the α_i variables take very small values. Note that, according to (45), these variables are approximately either equal to 1 or uniformly distributed in $(0, 1)$. Therefore, the region that corresponds to very small values of these variables is negligible. Consequently, Eqs. (66) and (69), which are obtained exclusively based on (117) and (118) without taking into consideration (119) and (120), are good approximations.

□

APPENDIX C

Proof of Corollary 2.

For a deterministic rebuild time distribution, it holds that $E(X^k) = [E(X)]^k$. Consequently, for $u = d+2, \ldots, \tilde{r}$, and from (75), it follows that

$$f_u \triangleq \frac{P^{(\tilde{r})}_{\ddot{s},u+1}}{P^{(\tilde{r})}_{\ddot{s},u}} = \frac{(\tilde{r}-u)(m-u+1)\,(u-d)\,\tilde{n}_u\,b_{u-1}\,V_u}{(\tilde{r}-u+1)(m-u)(u-d+1)\,\tilde{n}_{u-1}\,b_u}\;. \tag{121}$$

We shall now show that $f_u < 1$.

For the symmetric placement scheme, and using (36), (121) yields

$$\begin{aligned} f_u &= \frac{(\tilde{r}-u)(m-u+1)\,(u-d)\,(k-u)\frac{\min((k-u+1)\,b,B_{\max})}{l+1}\,\frac{m-u}{k-u}}{(\tilde{r}-u+1)(m-u)(u-d+1)\,(k-u+1)\frac{\min((k-u)\,b,B_{\max})}{l+1}} \\ &= \frac{(\tilde{r}-u)(m-u+1)\,(u-d)\;\min((k-u+1)\,b,B_{\max})}{(\tilde{r}-u+1)(k-u+1)(u-d+1)\;\min((k-u)\,b,B_{\max})} \\ &= \frac{\tilde{r}-u}{\tilde{r}-u+1}\;\frac{u-d}{u-d+1}\;\frac{m+1-u}{k-u}\;\frac{\min(b,\frac{B_{\max}}{k+1-u})}{\min(b,\frac{B_{\max}}{k-u})}\;. \end{aligned} \tag{122}$$

The fact that $\frac{B_{\max}}{k+1-u} < \frac{B_{\max}}{k-u}$ implies that $\min(b, \frac{B_{\max}}{k+1-u}) \leq \min(b, \frac{B_{\max}}{k-u})$ and therefore the last fraction is less than or equal to 1. Also, given that $k \geq m+1$, the third fraction is less than or equal to 1. Moreover, each of the first two fractions is less than 1. Consequently, $f_u < 1$.

For the clustered placement scheme, and using (42), (121) yields

$$\begin{aligned} f_u &= \frac{(\tilde{r}-u)(m-u+1)\,(u-d)\,(m-u)\;\min(b\,,\frac{B_{\max}}{l})}{(\tilde{r}-u+1)(m-u)(u-d+1)\,(m-u+1)\;\min(b\,,\frac{B_{\max}}{l})} \\ &= \frac{\tilde{r}-u}{\tilde{r}-u+1}\;\frac{u-d}{u-d+1}\;<\;1\;. \end{aligned} \tag{123}$$

□

APPENDIX D

Proof of Corollary 3.

For $u = d+2, \ldots, \tilde{r}$, and from (75), it follows that

$$g_u \triangleq \frac{P^{(\tilde{r})}_{\ddot{s},\tilde{r}}}{P^{(\tilde{r})}_{\ddot{s},u}} = \frac{(m-u+1)\,(u-d)\;\tilde{n}_{\tilde{r}-1}\,b_{u-1}}{(\tilde{r}-u+1)\,(m-\tilde{r}+1)\,(\tilde{r}-d)\;\tilde{n}_{u-1}\,b_{\tilde{r}-1}} \cdot \frac{E(X^{u-d-2})\,E(X^{\tilde{r}-d-1})}{E(X^{u-d-1})\,E(X^{\tilde{r}-d-2})} \cdot \prod_{i=u}^{\tilde{r}-1} V_i\;. \tag{124}$$

For a Weibull rebuild time distribution, with probability density and cumulative distribution functions

$$\begin{aligned} f_X(x;\eta,\Lambda) \;&=\; \frac{\eta}{\Lambda}\left(\frac{x}{\Lambda}\right)^{\eta-1} e^{-(x/\Lambda)^\eta} \\ F_X(x;\eta,\Lambda) \;&=\; 1 - e^{-(x/\Lambda)^\eta}\;, \end{aligned} \tag{125}$$

it holds that

$$E(X^k) = \Lambda^k\,\Gamma(1+k/\eta)\;. \tag{126}$$

Note that this distribution provides a continuous spectrum between the deterministic distribution (for $\eta \to \infty$) and the exponential distribution (for $\eta = 1$). Let us introduce the variable h_u defined as follows:

$$h_u \triangleq \frac{E(X^{u-d-2})\,E(X^{\tilde{r}-d-1})}{E(X^{u-d-1})\,E(X^{\tilde{r}-d-2})}\;. \tag{127}$$

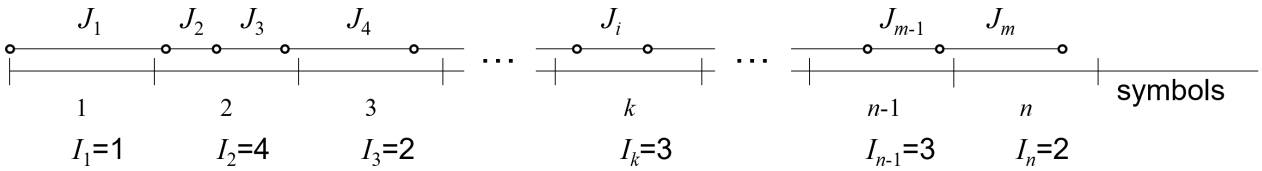


Figure 25. Number of shards that symbols span.

By virtue of (126), (127) yields

$$h_u \triangleq \frac{\Lambda^{u-d-2}\,\Gamma\left(1+\frac{u-d-2}{\eta}\right)\Lambda^{\tilde{r}-d-1}\,\Gamma\left(1+\frac{\tilde{r}-d-1}{\eta}\right)}{\Lambda^{u-d-1}\,\Gamma\left(1+\frac{u-d-1}{\eta}\right)\Lambda^{\tilde{r}-d-2}\,\Gamma\left(1+\frac{\tilde{r}-d-2}{\eta}\right)} = \frac{\Gamma\left(1+\frac{u-d-2}{\eta}\right)\Gamma\left(1+\frac{\tilde{r}-d-1}{\eta}\right)}{\Gamma\left(1+\frac{u-d-1}{\eta}\right)\Gamma\left(1+\frac{\tilde{r}-d-2}{\eta}\right)} . \quad (128)$$

From (126) and for $n \to \infty$, it holds that $E(X^k) = \Lambda^k = [E(X)]^k$ and consequently $h_u = 1$. For $n = 1$, it holds that $E(X^k) = \Lambda^k\,\Gamma(1+k) = k!\,\Lambda^k = k!\,[E(X)]^k$ and consequently $h_u = (\tilde{r}-d-1)/(u-d-1)$. As the function $\Gamma(x)$ is convex, it holds that h_u decreases with increasing η, such that

$$1 \le h_u \le \frac{\tilde{r}-d-1}{u-d-1} , \quad \text{for } 1 \le \eta < \infty . \quad (129)$$

For the symmetric placement scheme, and using (37) and the fact that $k \ge m+1$, it holds that

$$\prod_{i=n_1}^{n_2} V_i = \prod_{i=n_1}^{n_2} \frac{m-u}{k-u} < \prod_{i=n_1}^{n_2} \frac{m-u}{m+1-u} = \frac{m-n_2}{m+1-n_1} . \quad (130)$$

For the symmetric placement scheme, and using (35), (36) and (127), (124) yields

$$g_u = \frac{(m-u+1)\,(u-d)\,(k-\tilde{r}+1)\,\frac{\min((k-u+1)\,b,B_{\max})}{l+1}}{(\tilde{r}-u+1)(m-\tilde{r}+1)\,(\tilde{r}-d)\,(k-u+1)\,\frac{\min((k-\tilde{r}+1)\,b,B_{\max})}{l+1}} \cdot \frac{E(X^{u-2})\,E(X^{\tilde{r}-1})}{E(X^{u-1})\,E(X^{\tilde{r}-2})} \cdot \prod_{i=u}^{\tilde{r}-1} V_i = \frac{(m-u+1)\,(u-d)\,\min(b,\frac{B_{\max}}{k-u+1})}{(\tilde{r}-u+1)(m-\tilde{r}+1)\,(\tilde{r}-d)\,\min(b,\frac{B_{\max}}{k-\tilde{r}+1})}\; h_u \prod_{i=u}^{\tilde{r}-1} V_i . \quad (131)$$

Given that $u \le \tilde{r}-1 < \tilde{r}$, it holds that $\frac{B_{\max}}{k-u+1} < \frac{B_{\max}}{k-\tilde{r}+1}$ and therefore $\min(b,\frac{B_{\max}}{k-u+1}) < \min(b,\frac{B_{\max}}{k-\tilde{r}+1})$. Consequently, using (129) and (130), (131) yields

$$g_u < \frac{(m-u+1)\,(u-d)}{(\tilde{r}-u+1)(m-\tilde{r}+1)\,(\tilde{r}-d)}\;\frac{\tilde{r}-d-1}{u-d-1}\;\frac{m-\tilde{r}+1}{m+1-u} = \frac{u-d}{\tilde{r}-d}\;\frac{\tilde{r}-d-1}{(\tilde{r}-u+1)(u-d-1)} < \frac{\tilde{r}-d-1}{(\tilde{r}-u+1)(u-d-1)} . \quad (132)$$

Given that $d+2 \le u < \tilde{r}$, it holds that $[u-(d+2)](u-\tilde{r}) \le 0$ or, equivalently, $\tilde{r}-d-1 \le (\tilde{r}-u+1)(u-d-1)$. Consequently, it follows from (132) that $g_u < 1$.

For the clustered placement scheme, and using (41), (42), (43), and (127), (131) yields

$$g_u = \frac{(m-u+1)\,(u-d)\,(m-\tilde{r}+1)\,h_u\,\min(b\,,\frac{B_{\max}}{l})}{(\tilde{r}-u+1)(m-\tilde{r}+1)\,(\tilde{r}-d)\,(m-u+1)\,\min(b\,,\frac{B_{\max}}{l})} \overset{(129)}{\le} \frac{u-d}{\tilde{r}-d}\;\frac{\tilde{r}-d-1}{(\tilde{r}-u+1)(u-d-1)} < \frac{\tilde{r}-d-1}{(\tilde{r}-u+1)(u-d-1)} . \quad (133)$$

As the last term of (133) is the same as the last term of (132), which is less or equal to 1, it follows that $g_u < 1$.

□

APPENDIX E

Proof of Corollary 4.

Immediate from Corrolary 1, (16), (66), and (72).

Relation (83) can alternatively be obtained as follows. At exposure level u and for very small values of P_s, an entity failure is most likely caused by a single corrupted codeword that loses $\tilde{r}$ symbols. Let I be the number of shards that have parts stored in a symbol of this codeword. Then the expected number $E(Y_{\text{UF}_u})$ of lost entities associated with the direct path $\overrightarrow{UF_u}$ is determined by

$$E(Y_{\text{UF}_u}) \approx E(I)\; P_{\text{UF}_u} , \quad (134)$$

where P_{UF_u} is the probability of data loss due to unrecoverable symbol errors at exposure level u.

We proceed to show that

$$E(I) = 1 + \frac{1}{E(J)} . \quad (135)$$

Let us consider m successive shards stored in n symbols, as depicted in Figure 25, with the shard boundaries indicated by the circles and the symbol boundaries indicated by the vertical lines. Let J_i denote their size ($i = 1, \ldots, m$) and let I_k ($k = 1, \ldots, n$) denote the number of shards that have parts stored in the k-th symbol. For large values of m and n, it holds that

$$\sum_{i=1}^{m} J_i \approx n , \quad (136)$$

such that

$$\lim_{m\to\infty} \frac{\sum_{i=1}^{m} J_i}{n} = 1 . \quad (137)$$

It also holds that

$$E(J) = \frac{\sum_{i=1}^{m} J_i}{m} , \quad (138)$$

and

$$E(I) = \frac{\sum_{k=1}^{n} I_k}{n} \ , \tag{139}$$

Combining (137) and (138) yields

$$\lim_{m\to\infty} \frac{m}{n} = \frac{1}{E(J)} \ . \tag{140}$$

Note that the number of shards that have parts stored in a symbol decreased by one is equal to the number of shard boundaries within the symbol. For instance, regarding the kth symbol, there are three shards that have parts stored in this symbol, namely the $(i-1)$th, ith, and $(i+1)$th shard, such that $I_k = 3$, which decreased by one yields the two shard boundaries within this symbol. Consequently, considering the n symbols and the corresponding m boundaries, we have

$$\sum_{k=1}^{n} (I_k - 1) = m \tag{141}$$

or

$$\sum_{k=1}^{n} I_k = n + m \tag{142}$$

Substituting (142) into (139), and using (140), yields (135).

An alternative proof for the case where $J_j \geq 1$, for $j = 1, 2, \ldots, E_s$, is the following. Let us consider an arbitrary symbol and let $\hat{J}$ be the size of the shard that is stored at the beginning of the symbol. As this shard is a sampled shard, the pdf of its size $\hat{J}$ is determined by (114), that is,

$$P(\hat{J} = J_j) = P(\hat{e}_s = e_{s,j}) = \hat{v}_j \ , \quad \text{for } \ j = 1, 2, \ldots, E_s \ . \tag{143}$$

Let y be the size from the beginning of the sampled shard to the beginning of the symbol. Then y is uniformly distributed in the interval $(0, \hat{J})$. For y in the interval $(0, \hat{J}-1)$, the symbol only contains a part of the sampled shard, that is, it contains a part of a single shard. The probability p of this event is

$$p = \int_0^{\hat{J}-1} \frac{1}{\hat{J}} \, dx = \frac{\hat{J}-1}{\hat{J}} \ . \tag{144}$$

On the other hand, for y in the interval $(\hat{J}-1, 1)$, the symbol contains parts of the sampled shard, as well as of the subsequent shard, that is, the symbol contains parts of two shards. The probability of this event is $1-p$. Consequently, the expected number $E(I|\hat{J})$ of shards that have parts stored in the symbol is

$$E(I|\hat{J}) = 1 \cdot p + 2 \cdot (1-p) = 2 - p \overset{(144)}{=} \frac{\hat{J}+1}{\hat{J}} \ . \tag{145}$$

Unconditioning (145) on $\hat{J}$ and using (143) yields

$$E(I) = \sum_{j=1}^{E_s} E(I|J_j) \, P(\hat{J} = J_j) \overset{(145)}{=} \sum_{j=1}^{E_s} \frac{J_j + 1}{J_j} \, \hat{v}_j$$
$$\overset{(14)(16)(114)}{=} \sum_{j=1}^{E_s} \frac{J_j + 1}{J_j} \cdot \frac{J_j \, v_j}{E(J)} = \frac{E(J)+1}{E(J)} = 1 + \frac{1}{E(J)} \ , \tag{146}$$

which is relation (135).

Substituting (135) into (134) yields (83).

□

REFERENCES

[1] I. Iliadis, "Relations between entity sizes and error-correction coding codewords and data loss," in Proceedings of the 17th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), May 2024, pp. 1–11.

[2] I. Iliadis and V. Venkatesan, "Reliability evaluation of erasure coded systems," Int'l J. Adv. Telecommun., vol. 10, no. 3&4, Dec. 2017, pp. 118–144.

[3] I. Iliadis, "Reliability evaluation of erasure coded systems under rebuild bandwidth constraints," Int'l J. Adv. Networks and Services, vol. 11, no. 3&4, Dec. 2018, pp. 113–142.

[4] ——, "Reliability of erasure-coded storage systems with latent errors," Int'l J. Adv. Telecommun., vol. 15, no. 3&4, Dec. 2022, pp. 23–41.

[5] ——, "Reliability evaluation of erasure-coded storage systems with latent errors," ACM Trans. Storage, vol. 19, no. 1, Jan. 2023, pp. 1–47. [Online]. Available: https://doi.org/10.1145/3568313

[6] I. Iliadis and V. Venkatesan, "Expected annual fraction of data loss as a metric for data storage reliability," in Proceedings of the 22nd Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2014, pp. 375–384.

[7] I. Iliadis, "Expected annual fraction of entity loss as a metric for data storage durability," in Proceedings of the 16th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2023, pp. 1–11.

[8] G. A. Jaquette, "LTO: A better format for mid-range tape," IBM J. Res. Dev., vol. 47, no. 4, Jul. 2003, pp. 429–444.

[9] I. Iliadis, R. Haas, X.-Y. Hu, and E. Eleftheriou, "Disk scrubbing versus intradisk redundancy for RAID storage systems," ACM Trans. Storage, vol. 7, no. 2, Jul. 2011, pp. 1–42.

[10] I. Iliadis and V. Venkatesan, "Most probable paths to data loss: An efficient method for reliability evaluation of data storage systems," Int'l J. Adv. Syst. Measur., vol. 8, no. 3&4, Dec. 2015, pp. 178–200.

[11] A. Dholakia, E. Eleftheriou, X.-Y. Hu, I. Iliadis, J. Menon, and K. Rao, "A new intra-disk redundancy scheme for high-reliability RAID storage systems in the presence of unrecoverable errors," ACM Trans. Storage, vol. 4, no. 1, May 2008, pp. 1–42.

[12] I. Iliadis, "Reliability modeling of RAID storage systems with latent errors," in Proceedings of the 17th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2009, pp. 111–122.

[13] T. J. E. Schwarz, Q. Xin, E. L. Miller, D. D. E. Long, A. Hospodor, and S. Ng, "Disk scrubbing in large archival storage systems," in Proceedings of the 12th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Oct. 2004, pp. 409–418.

[14] A. Oprea and A. Juels, "A clean-slate look at disk scrubbing," in Proceedings of the 8th USENIX Conference on File and Storage Technologies (FAST), Feb. 2010, pp. 57–70.

[15] B. Schroeder, S. Damouras, and P. Gill, "Understanding latent sector errors and how to protect against them," ACM Trans. Storage, vol. 6, no. 3, Sep. 2010, pp. 1–23.

[16] S. A. Chamazcoti, B. Safaei, and S. G. Miremadi, "Can erasure codes damage reliability in ssd-based storage systems?" IEEE Transactions on Emerging Topics in Computing, vol. 7, no. 3, 2019, pp. 435–446.

[17] M. Kishani, S. Ahmadian, and H. Asadi, "A modeling framework for reliability of erasure codes in ssd arrays," IEEE Transactions on Computers, vol. 69, no. 5, 2020, pp. 649–665.

[18] M. Zhang, S. Han, and P. P. C. Lee, "SimEDC: A simulator for the reliability analysis of erasure-coded data centers," IEEE Trans. Parallel Distrib. Syst., vol. 30, no. 12, 2019, pp. 2836–2848.

[19] M. Silberstein, L. Ganesh, Y. Wang, L. Alvisi, and M. Dahlin, "Lazy means smart: Reducing repair bandwidth costs in erasure-coded distributed storage," in Proceedings of the 7th ACM International Systems and Storage Conference (SYSTOR), Jun. 2014, pp. 15:1–15:7.

[20] Tape Roadmap, Information Storage Industry Consortium (INSIC) Report, 2019. [Online]. Available: https://www.insic.org/wp-content/uploads/2019/07/INSIC-Applications-and-Systems-Roadmap.pdf [retrieved: November, 2024]

[21] I. Iliadis, "Effect of lazy rebuild on reliability of erasure-coded storage systems," in Proceedings of the 15th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2022, pp. 1–10.

[22] G. Cancio et al., "Tape archive challenges when approaching exabyte-scale," 2010, Presentation at CHEP 2010, available online.

[23] I. Iliadis, L. Jordan, M. Lantz, and S. Sarafijanovic, "Performance evaluation of automated tape library systems," in Proceedings of the 29th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Nov. 2021, pp. 1–8.

[24] D. Ford, F. Labelle, F. I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan, "Availability in globally distributed storage systems," in Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Oct. 2010, pp. 61–74.

[25] M. Ovsiannikov, S. Rus, D. Reeves, P. Sutter, S. Rao, and J. Kelly, "The quantcast file system," in Proceedings of the 39th International Conference on Very Large Data Bases (VLDB), vol. 6, no. 11. VLDB Endowment, Aug. 2013, pp. 1092–1101.

[26] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin, "Erasure coding in Windows Azure Storage," in Proceedings of the USENIX Annual Technical Conference (ATC), Jun. 2012, pp. 15–26.

[27] S. Muralidhar et al., "f4: Facebook's Warm BLOB Storage System," in Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Oct. 2014, pp. 383–397.

[28] Amazon Simple Storage Service (Amazon S3), 2024. [Online]. Available: http://aws.amazon.com/s3/ [retrieved: November, 2024]

[29] I. Iliadis and M. Lantz, "Reliability evaluation of automated tape library systems," in Proceedings of the 32nd IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Oct. 2024, pp. 80–87.

[30] L. Kleinrock, Queueing Systems, Volume 1: Theory. New York: Wiley, 1975.

Enhancing Path Reliability in Contact Graph Routing via Improved Hop Time Estimations

Ricardo Lent
Engineering Technology
University of Houston
Houston, Texas, USA
rlent@uh.edu

***Abstract*—Delay-Tolerant Networking (DTN) enables the forwarding of data bundles over space networks that experience extended link disruptions and path disconnections. Routing in such environments is challenging yet crucial for efficient end-to-end data delivery. Contact Graph Routing (CGR) is the standard routing method adopted for space DTNs. This study enhances CGR by exploring the potential inclusion of a Cognitive Element (CE) that leverages a data-driven approach. The CE is anticipated to use machine learning to estimate average single-hop bundle delivery times based on selected inputs. These estimates then replace the propagation delay that is used as the sole decision metric in CGR's shortest-path algorithm, improving the accuracy of the average one-hop delivery time predictions by allowing consideration of significant factors such as Convergence Layer Adapter (CLA) behavior, configuration parameters, packet drop probabilities, and unreliable contacts. The result is enhanced routing performance. The paper evaluates the CE extension in a simulated Earth-Moon network, assuming implementation independence and examining the effects of contact plan size (defined as a look-ahead window) and potential performance degradation from reduced prediction accuracy due to partial data or model limitations. Insights into the practical benefits of this approach are provided with a focus on realistic contact features and unreliable links.**

Keywords-delay-tolerant networking; routing; reliability; performance evaluation; cognitive networking.

I. Introduction

This paper extends the exploration of the integration of a Cognitive Element (CE) into Contact Graph Routing (CGR) to enhance the routing performance [1] of space Delay-Tolerant Networks (DTNs). A DTN provides crucial services in facilitating the communication among spacecraft, rovers, orbiters, landers, and ground stations in space exploration missions that often times involve significant signal propagation delays because of the long-distance of the communication links and periods of signal disruption due to celestial bodies obstructing line-of-sight communication paths and other factors. DTN plays a key role in LunaNet, NASA's proposed lunar communications and navigation network [2], [3] for the Artemis program, which demands systems capable of efficiently managing the unique challenges of space networking. LunaNet will provide transparent networking services that establish end-to-end data paths through a disconnected and time-varying topology. Routing is a key component of space DTNs that determines the store-carry-and-forward communication path for data bundles. The pre-planned nature of these networks simplifies the routing task, as contact opportunities can be anticipated from the expected positions of nodes as derived from orbital calculations. These calculations not only identify link obstructions but also provide the information required for a link budget analysis. CGR leverages the available future contact information to compute the optimal next-hop for bundles achieving minimum latency to the destination.

However, it is relevant to point out that, despite the deterministic assumption of contacts in scheduled DTNs, variations can still arise due to a multitude of factors. For instance, cloud coverage can bring large signal attenuation at high radio frequencies and in free-space optical links that can disrupt expected contacts between an orbiter and a ground node. Node malfunction and antenna misalignment issues may also occur randomly preventing contact realizations. Moreover, operational priorities may dynamically change resulting in the re-assignment of expected contacts to a different application.

These observations are aligned with the evolution of Opportunistic Contact Graph Routing (OCGR), which explores the potential utilization of non-scheduled contacts associated with a calculated confidence level. OCGR introduces a shift in the path search methodology of CGR, allowing the discovery of the k-shortest paths and the assessment of path reliability. Extending this concept further, it can be assumed that all contacts in a DTN have an opportunistic nature, including scheduled contacts, as they may randomly fail as discussed. Therefore, at least the path searching part of OCGR can be widely applicable to optimize unreliable DTN scenarios, provided each contact can be associated with a confidence level.

One limitation of CGR and OCGR is that the time progression step of each bundle forwarding within the path search algorithm assumes ideal transmission conditions that are determined solely by the link propagation delay or one-way light time. Buffering information is considered unavailable beyond the links leading to neighboring nodes, therefore not fully accounting for transmission and queuing delays. Additionally, protocol dynamics, including the convergence-layer adapter (CLA), particularly concerning the handling of packet losses through retransmissions, are overlooked. These factors contribute to differences between the calculated times within the CGR path search algorithm and the actual bundle

This work was supported by grant #80NSSC22K0259 from the National Aeronautics and Space Administration (NASA).

forwarding performance, which may impact routing optimally specially with network congestion.

Building on the initial exploration of performance gains from integrating a CE into CGR [1], this study extends the evaluation by incorporating additional options. The core concept remains that the CE can generate more accurate estimates of average one-hop bundle delivery times, helping CGR's shortest-time algorithm to identify optimal paths by factoring in predicted network performance metrics. The main contributions of this work include:

1) The concept of using a CE to forecast average single-hop bundle delivery times, which is utilized in the time progression step of CGR, is further developed. The core idea is to introduce a data-driven approach that aids in identifying the best paths by considering factors such as specific CLA behavior, configuration parameters, packet drops, and unreliable contacts. The CE could be trained either offline using an analytical model or historical data, or online with real-time measurements to achieve accurate predictions. This approach eliminates the need for modifications to the CGR algorithm to handle uncertain contacts and random factors, thereby removing the requirement for searching for the k-shortest paths as done in OCGR. Additionally, this study eliminates the constraint that bundle arrivals must coincide within a contact duration by noting the average nature of the delivery time estimations. Network congestion metrics are also used to filter the contact plan before computing the path search.
2) This study evaluates the impact of the contact plan size on routing optimality in CGR and the CE extension. Since CGR relies on building a graph where nodes represent future available contacts, a shorter contact plan can speed up path computation and improve efficiency. However, this also risks insufficient network connectivity for path computation. The study assesses how the size of the contact plan, defined as the look-ahead time window, affects CGR routing performance and the performance of the CE extension.
3) An evaluation of the performance impact of the limitations of the CE in producing accurate average bundle time estimations. The CE provides a function that maps the known network state to forecast the time required for a bundle to reach the next hop. The limitations of the method are therefore related to the accuracy of the network state knowledge, particularly because the required information may not necessarily be available at the nodes. This study provides an implementation-agnostic assessment of the performance advantages and limitations of the CE, identifying the performance bounds of the method across three variations regarding the severity of assumptions involving the network state. In the first scenario, only local state information, normally assumed available in the standard CGR, is assumed. The second and third scenarios require global knowledge, i.e., information external to the node, with differences in how they predict transmission hop times. The evaluation is conducted within the context of an Earth-Moon network [2], employing approximately realistic values for contact features and considering unreliable contacts. The evaluation provides insight into the impact of imperfect CE model predictions on end-to-end bundle routing performance.

The remainder of the paper is structured as follows: Section II reviews related works relevant to this study. Section III provides a detailed explanation of the Cognitive Element (CE) method. Section IV describes the evaluation scenario and simulation assumptions. Section V presents the results demonstrating the CE's effectiveness in optimizing bundle flow across an Earth-Moon network. Lastly, Section VI offers concluding remarks.

II. Related Works

The reliability of DTN protocols remains a dynamic area of research with application to many ambitious missions [2], [4]. A feature that characterizes space DTNs is the use of scheduled contacts, commonly used jointly with the Bundle Protocol (BP) [5], [6] and Contact Graph Routing (CGR) [7], [8], which begins by constructing a graph, where vertices denote active contacts and links represent logical transitions between contacts—where one contact's endpoint aligns with the next contact's starting point, feasible within a defined time frame. While this process incorporates factors, such as transmission time, propagation delay, and network disruptions, buffering delays are typically overlooked due to the distributed nature of the algorithm, as this information is normally inaccessible. CGR is commonly implemented by adapting Dijkstra or Yen's algorithms, with the latter method raising scalability concerns [9].

The performance of CGR in scenarios involving unreliable links has been explored in various contexts, including satellite constellations [10] and random networks [11]. Reliability has been mainly addressed by BP custody [12] and CLA design via retransmissions, e.g., the Licklider Transmission Protocol (LTP). For experimental results, see for example [13]–[15]. These studies have shed light on CGR's vulnerabilities concerning contact failure rates and random losses. An extension known as Opportunistic CGR (OCGR) [16] investigates the potential integration of nonscheduled contacts—either discovered or predicted—into CGR's standard path search algorithm, assigning them a confidence level. OCGR maintains a record of the contact history of nonscheduled contacts to predict future contacts, alongside their associated properties and confidence levels, calculated based on available contact history [16]. Discovered contacts are assigned a unit confidence [17] and the resulting route is assigned a delivery confidence derived from the product of the confidence levels of the contacts involved. In recent iterations, the implementation of OCGR [18] evaluates path candidates based on their arrival confidence, considering a predefined margin from the highest confidence level.

Additional related methods to this work include CGR extensions to support multigraphs [19], better handling of capacity constraints [20], [21], a variety of networks (and Roaming DTN (RDTN) [22] that integrates roaming nodes with unpredictable motion, Best Routing Under Failures (BRUF) [23], where the routing process is conceptualized as a Markov Decision Process, with certain state transitions becoming probabilistic due to the limited reliability of specific contacts and Routing under Uncertain Contact Plans (RUCoP) [24], [25] that introduces a multiple-copy Markov Decision Process. Also related, is the Cognitive Space Gateway (CSG) [26] where routing decisions are delegated to a Spiking Neural Network which is continually trained after the bundle transmissions using a reinforcement learning approach.

This paper presents an alternative approach to enhance CGR performance, a method known for its computational efficiency and practicality, but limited in handling random factors impacting single-hop bundle transmissions, such as packet losses and contact failures. Unlike previous approaches, this method modifies the conventional one-hop bundle time calculations. Specifically, it introduces the idea of using a cognitive element designed to accurately predict average bundle transmission times. While the implementation of this cognitive element is expected to utilize a neural network or similar structure, this study evaluates its limitations without specifying a particular technology. Instead, it offers widely applicable findings focused on determining performance bounds based on assumptions regarding available network state information used as inputs to the CE.

III. Cognitive Extension and CGR

CGR defines a decentralized approach in which each node calculates the path to the destination node, but using only next-hop information to forward bundles. This method requires access to the contact plan for all future contacts, which is distributed to the DTN nodes in advance.

A. Standard Mechanisms

A contact plan consists of a sequence of entries of the following form: $(\mathcal{I}_i, \mathcal{F}_i, \mathcal{T}_i, \mathcal{S}_i, \mathcal{E}_i, \mathcal{R}_i, \mathcal{O}_i, r_i)$ and that includes a contact identifier $\mathcal{I}_i$, the sending $\mathcal{F}_i$ and receiving node $\mathcal{T}_i$ identifiers, the start $\mathcal{S}_i$ and end $\mathcal{E}_i$ times, the transmission rate $\mathcal{R}_i$ and the propagation delay or one-way light time $\mathcal{O}_i$ that depends on the distance between the nodes. The term r_i, $0 \leq r_i \leq 1$, is the contact confidence as used by O-CGR [16].

To determine routes for each desired destination, CGR builds a contact graph $G = (V, E)$ using each contact entry of the plan as a vertex minus the entries containing excluded nodes (e.g., known failed nodes). A contact graph is a directed acyclic graph where an edge exists when two contacts are logically connected, which happens when the destination node of the first contact matches the sending node of the second contact and the latter expires after the first. The target contact of an edge is called the proximate of the first contact. The contact graph is considered directed as transmissions in the reverse direction of a given contact may not be possible or may occur with different transmission parameters, e.g., different transmission rate, due to transceiver limitations. A start time of the proximate that is later than the current time while using a contact brings forced data buffering due to the corresponding link disruption.

CGR derives the path to the destination node by calculating the shortest path on the contact graph between two auxiliary vertices that are attached to represent the root and terminal contacts. The root is the node executing CGR. These auxiliary contacts involve a zero-cost to the relevant proximates. Starting from the root, a graph traversal iteratively tracks the bundle transmission progress in the network by estimating its arrival time as it is forwarded over contacts. That is, if t_i represents the bundle arrival time calculated at vertex i of the contact graph, the algorithm evaluates the proximate vertices j and greedily chooses the one offering the smallest t_j. Specifically, the evaluation of the proximate vertex j, yields the following arrival time.

$$t_j = \begin{cases} t_i + \mathcal{O}_j & \mathcal{S}_i \leq t_i \\ \mathcal{S}_j + \mathcal{O}_j & \mathcal{S}_i > t_i \end{cases} \quad (1)$$

The calculation does not include transmission times, but that metric is utilized to determine the remaining data volume for transmissions. This additional step enables consideration of whether given contacts are likely to be already fully booked. However, this assessment is restricted to contacts leading to neighboring nodes, as information beyond that scope is unavailable. The output of the algorithm is the path $P = v_0, v_1, \ldots, v_k$, where $v_i \in V$ is a contact and v_0, v_k are the auxiliary contact entries for the source and sink nodes respectively. If t_k is the estimated time to deliver the bundle to the end contact based on (1) for each step, the objective of the algorithm is to minimize t_k among all possible paths from v_0 to v_k in G.

B. Cognitive Element

The central idea of this paper is to enhance the route selection quality in CGR by refining the accuracy of the single-step bundle forwarding time calculation. This involves substituting the computation outlined in (1) with the output of a cognitive element (CE) designed to accurately predict the average time needed to deliver a bundle to the next hop, accounting for the segmentation, transmission and retransmission times of the convergence-layer adapter, buffering delays, and the reliability of contacts, among other factors:

$$t_j = \begin{cases} t_i + y_j & \mathcal{S}_i \leq t_i \\ \mathcal{S}_j + y_j & \mathcal{S}_i > t_i \end{cases} \quad (2)$$

where $y_j = f_\theta(x)$ represents the output of a function f_θ given the specified system state x and the model parameters θ.

A second observation concerns the interpretation of t_j, which now represents the average time to reach the next hop, rather than the precise definition in CGR. This change is required to properly take into account probabilistic factors, such as transmission errors and contact failures. The idea is that these probabilistic factors will affect the one-hop bundle

delivery time along the path adding uncertainty into the calculation of the final delivery time. With this reinterpretation of t_j, the shortest path algorithm of CGR requires no modification. It continues to identify the route with the smallest average time of arrival t_k (instead of precise time), but now able to accommodate random factors affecting the paths.

In this study, we keep the concept of introducing a CE to CGR separated from its implementation on purpose, recognizing that diverse techniques may be used to define this element. Possible mechanisms encompass a range of neural network architectures, including multi-layer feedforward, convolutional, generative adversarial, recurrent (such as Long Short-Term Memory Networks), autoencoders, graph neural networks, and more. These mechanisms can be implemented using either continuous activation or spiking neurons. Given the potential variations in prediction accuracy resulting among different techniques, our focus is in assessing the performance bounds attained with the introduction of the CE concept and understanding the performance implications of using imperfect inputs for $f_\theta(x)$.

In particular, we focus on studying three variations for $f_\theta(x)$. The first case, which is labeled CE-1, considers $f_\theta(x)$ providing the average one-hop bundle time that aggregates the propagation delay, and additional buffering time required with imperfect contact reliability. The second case, CE-2, improves CE-1 estimation by also aggregating the estimated bundle transmission time. The last case, CE-3, includes in addition to the elements of CE-2 the expected buffering time.

We note that CE-1 and CE-3 use the same expressions as in [1]. However, unlike the previous work, we remove the constraint in the path search algorithm that requires bundle transmissions to be completed within a contact duration. In this study, we consider transmissions as simple approximations of the average time required to deliver a bundle with reliability constraints. We note that this change tends to improve throughput as observed in the evaluation scenario.

Regarding the training of the models, CE-1 and CE-2 are comparatively the easiest to train since they require only local state information, which is available in the standard contact plan, i.e., transmission rate, propagation delay, and confidence level. The latter parameter can be understood as an estimate of contact reliability. CE-3, however, requires predicting the global state, as buffer occupancy levels dynamically change. CE-2 and CE-3 could be further improved if the channel bit-error-rate value could be estimated as this value determines the extended times required for retransmissions. To maintain the study's focus on evaluating the effectiveness of the CE concept rather than discussing specific approaches, we omit further details of the training phase for these models.

C. Model Approximation

CE-based predictions address the effect of contact reliability and are calculated using the following models [27], which provides the average time required to deliver a packet over contact j:

CE-1:

$$y_j = \mathcal{O}_j + \frac{1-r}{r}C \tag{3}$$

CE-2:

$$y_j = \frac{L}{R_j} + \mathcal{O}_j + \frac{1-r_j}{r_j}C \tag{4}$$

CE-3:

$$y_j = \frac{L+B}{R_j} + \mathcal{O}_j + \frac{1-r_j}{r_j}C \tag{5}$$

where L is the average bundle size, R_j the link rate, P_j the propagation delay, C the average time between contacts, B the buffer occupancy and r_j the contact's reliability. Note that the subscript j emphasizes that the parameters is per-contact and that both B and C refer to values associated with the next-hop node of proximate j. The value y_j is then used to advance time in the iteration of the shortest path calculation as used in 2. Assuming that the contact plan includes r_j as done with O-CGR, all of the inputs can be directly extracted from the contact plan except for L. However, L can be iteratively estimated from bundle arrivals through exponential averaging, that is, each arriving bundle of size l allows updating L as follows: $L \leftarrow \alpha l + (1-\alpha)L$, where α, $0 \leq \alpha \leq 1$, is a hyperparameter.

IV. Evaluation Scenario

Consider a communication scenario where a node located on the lunar surface regularly emits messages to a terrestrial sink. This scenario corresponds to a typical space exploration communication model, such as a rover collecting scientific data that is then sent for analysis or a hub aggregating data from various sources before transmitting it to Earth. The time required to deliver the data, i.e., the response time, and the risk of data loss provide sensible assessment of the benefits of introducing the CE extension.

To implement the scenario two simulators were developed. The first simulator generates the contact plan by estimating the locations of nodes from orbital calculations that accounts for both Earth's and Moon's rotation and translation, which helps to determine transmission opportunities based on the line-of-sight between nodes. The starting separation distance between the nodes is used to define the one-way light time that appears in the contact plan for each entry. The second simulator evaluates routing performance by using the generated contact plan and implementing an event-driven simulation of the bundle transmission dynamics, including buffering and drops, while considering potential contact failures that prolong buffering times.

The traffic originates from a Lunar node positioned on the southern far side of the Moon at -19.94, -200.07 in the selenographic coordinate system. As such, the node is not visible from Earth, but three orbital relays are available to forward the data: LO1, LO2, and LO3. For simplicity, Keplerian assumptions were used to model their orbits. The orbits are characterized by inclinations of 10, 40, and -40 degrees, and Right Ascension of the Ascending Node (RAAN) values of 4.462, 90, and 40 respectively. It is relevant to note

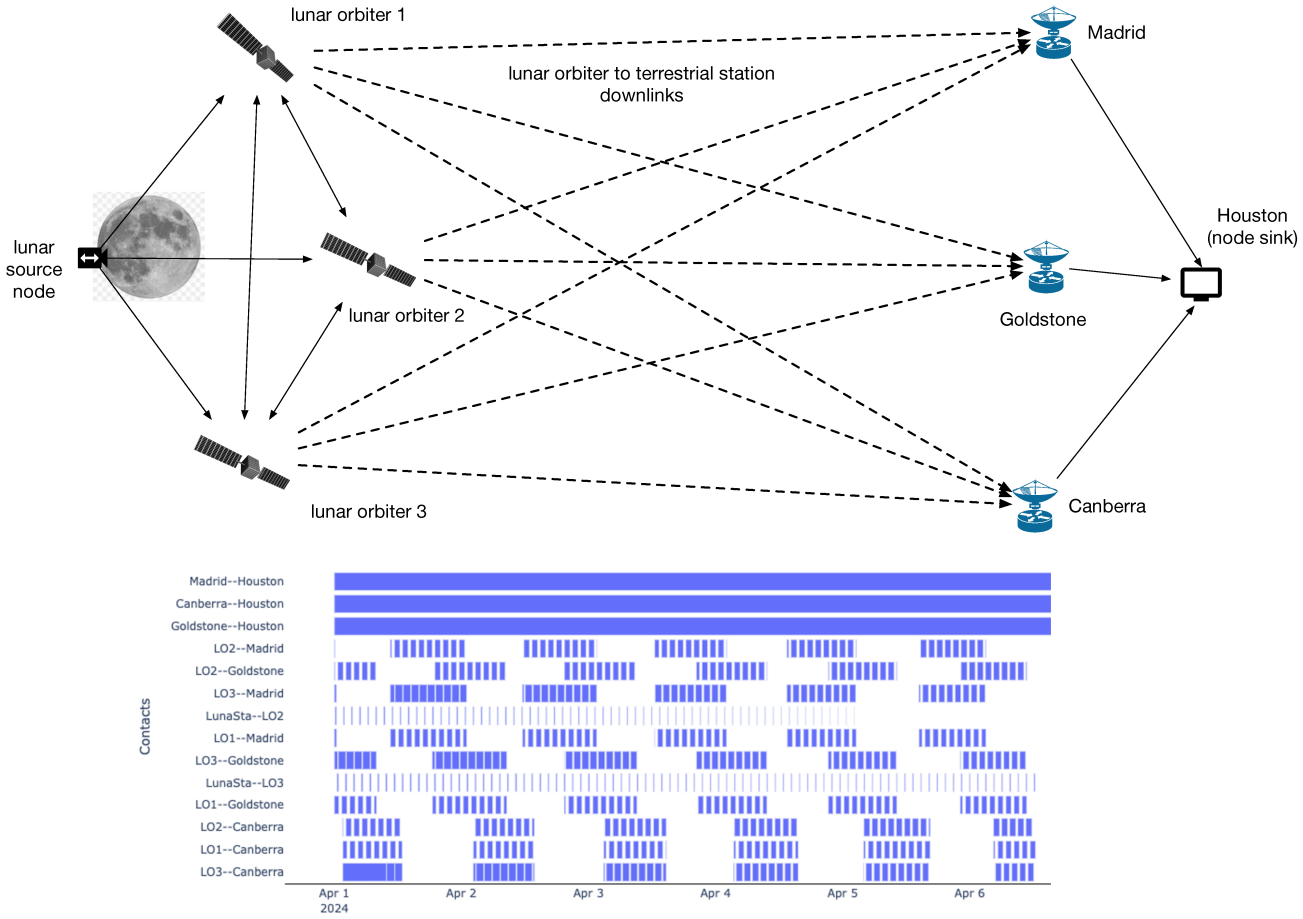


Figure 1. The evaluation scenario involves a source located on the surface of the Moon that generates traffic addressed to a terrestrial sink. The traffic can be routed through contacts provided by three lunar orbiters and three ground stations, as depicted in the figure. The dashed lines represent links that may be affected by unreliable contacts. The lower part of the figure shows the contact pattern between the nodes.

that these orbits were not modeled after existing lunar satellites but were defined to facilitate the establishment of contacts of varying durations with the rover and Earth stations. However, we consider the scenario to be close enough to that of a real mission.

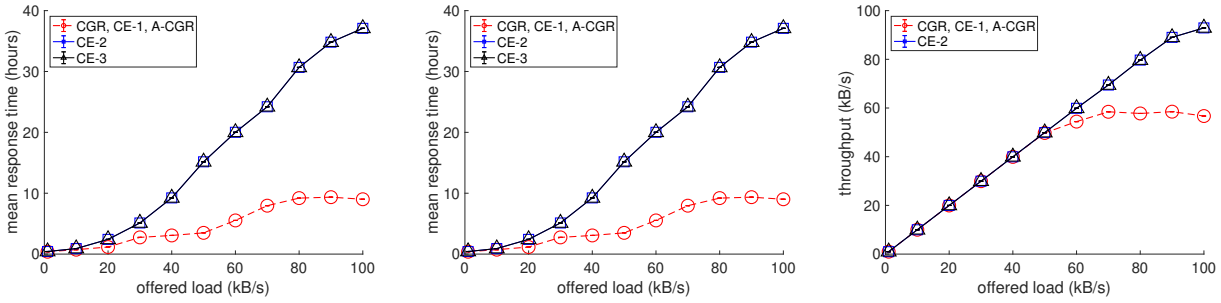
The terrestrial ground stations are modeled to match with the locations of the Deep-Space Communication (DSN) complexes in Canberra, Madrid, and Goldstone. The sink is assumed to be situated in Houston, with a permanent link established from each DSN location to Houston. In reality, these connections would traverse one or more networks. To account for this, the propagation delays for the terrestrial connections were calculated based on the distance between the involved nodes as the direct as-the-crow-flies distance with an additional 20% margin to account for network overhead.

The critical part of the system lies in the communication between the lunar node and the lunar orbiters, as well as between the lunar orbiters and the terrestrial stations, where the links provide on-off capacity. The evaluation topology is depicted in Figure 1 and Table I summarizes the features of these contacts, including the average and standard deviation of the contact duration and period (i.e., the inter-contact time). The contact pattern extracted from the orbital model is shown in the bottom chart of Figure 1. Note that the topology allows for sending orbital crosslink traffic if required by the routing protocol and provided contacts are available based on the orbital model.

TABLE I. AVERAGE (μ) AND STANDARD DEVIATION (σ) OF CONTACT DURATIONS AND PERIOD LENGTHS (TIME BETWEEN CONSECUTIVE CONTACTS) FOR THE EARTH-MOON EVALUATION NETWORK.

Contact type	**Duration μ**	**Duration σ**	**Period μ**	**Period σ**
Rover to LO1	13.0	1.9	91.4	3.4
Rover to LO2	10.4	3.8	90.6	10.3
Rover to LO3	13.3	2.5	91.1	6.5
LO1 to Madrid	55.0	10.8	149.7	186.7
LO1 to Canberra	53.7	12.8	160.9	217.7
LO1 to Goldstone	55.3	9.5	147.1	185.2
LO2 to Madrid	56.7	19.0	147.5	181.6
LO2 to Canberra	61.2	12.0	170.4	229.9
LO2 to Goldstone	58.9	16.6	147.9	189.4
LO3 to Madrid	69.8	15.1	152.3	196.4
LO3 to Canberra	75.7	67.4	178.6	239.4
LO3 to Goldstone	68.8	16.7	146.6	194.4

The transmission rate for the terrestrial (wired) links was



(a) Mean bundle response time. (b) Bundle loss ratio. (c) Throughput.

Figure 2. Flow performance metrics in a network with reliable contacts.

fixed at 2 Mbps, while all wireless transmissions were set to 100 Kbps. In all cases, the links are assumed to have negligible bit error rates (BER). It is also assumed that all contacts are reliable, possibly except for the ones between the orbiters and ground stations, due to the long distances involved. To this end, we evaluate three scenarios: (1) where none of the downlinks are affected, (2) where all downlinks from the three lunar orbiters are affected with reliability factors of 0.95, 0.85, and 0.5, and (3) where the downlinks from a single lunar orbiter have a limited reliability of 0.5 while the others remain fully reliable. While these values have been chosen arbitrarily, they are intended to illustrate a good range of possible scenarios that could occur in a real system.

In this context, the CGR agent running in the lunar node decides which orbiter will handle the bundle forwarding to Earth. The selected orbiter then determines which terrestrial station will receive the bundle before forwarding it to the sink. Also, the bundles are not associated with a finite deadline and the node buffers are assumed to be large enough to ignore the impact of buffer overflows, so bundles that miss any given contact simply continue waiting in the buffer for future service.

However, bundle drops are still possible under certain conditions. For computational efficiency, the length of the contact plan used by the CGR for route determination is limited to a pre-selected look-ahead window. That is, the contact plan is filtered to contain unexpired contacts whose starting time is not later than the current time plus the look-ahead window. If it is not possible to determine a path for the bundle using the information in the contact plan–for example, if all contacts are expected to be busy–the bundle is considered lost for the purposes of this study. This is measured by the "routing miss" metric. With the use of varied reliability factors and look-ahead window sizes, it is possible to identify the level of impact and assess the robustness of the routing scheme under various conditions.

V. Results and Discussion

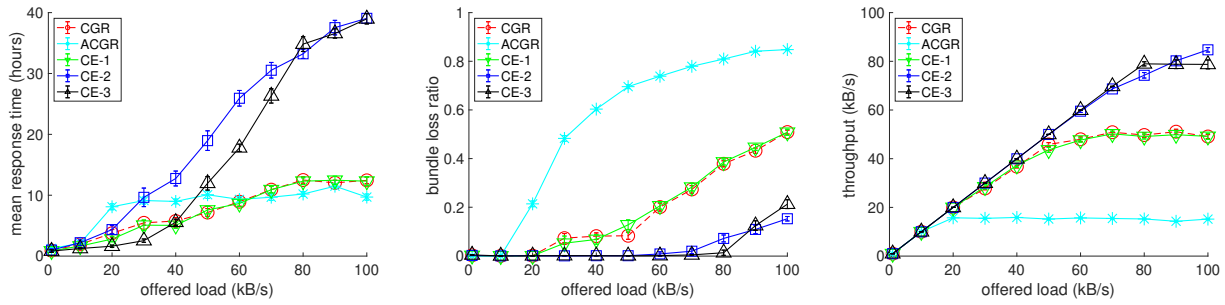
The routing performance is evaluated through observations of the response time, bundle loss ratio, and throughput of a test flow, and studied under simulation conditions where buffer capacities are uncapped, bit error rates (BER) are negligible, and no deadlines are imposed on bundle delivery times. Throughout is calculated as the product of the offered load and the delivery ratio (one minus the bundle loss ratio), both of which can be directly measured in the simulator. In the tests, bundles are generated at a constant rate of one every 100 seconds, while the bundle size is varied as an experimental parameter to adjust the offered load of the flow. The results reported in the next sections show in all cases the 95% confidence interval of the acquired samples for each experimental factor.

A. Reliable Contacts

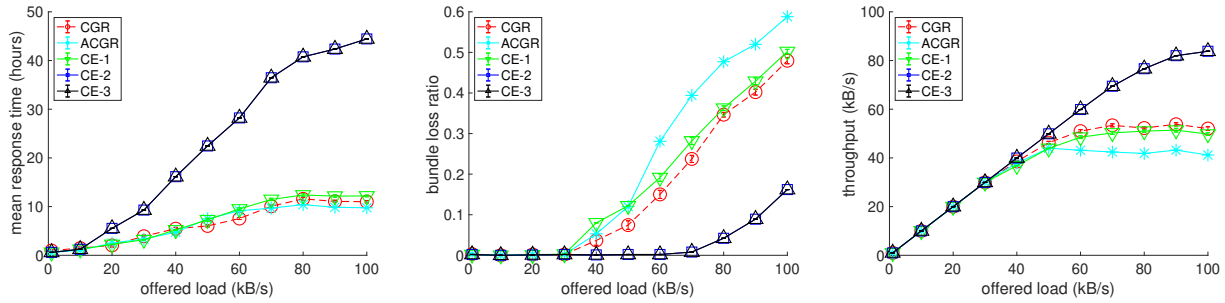
It is initially assumed that all contacts are reliable. Figure 2 (a) shows the average bundle response time that was observed with such conditions as a function of traffic load. The response time metric includes both transmission and buffering times, with the latter determined by the time required for bundles to reach the head of the transmission buffers. Buffering time is influenced by factors such as traffic load, transmission rates, and the waiting time for contacts along the selected path. The response time of a bundle is measured as the difference between its arrival time at the sink and its generation time by the simulator.

It can be observed that the average time required to transmit small files is around one hour or less. This duration is primarily determined by the waiting times for the next contact opportunities, as transmission times are short and, under light traffic conditions, buffers tend to remain empty. With increasing file sizes, there is a corresponding rise in both storage and transmission demands, leading to an increase in the average response time. The results provided by CGR and CE-1 are identical given the reliable contacts assumption. Also A-CGR, which will be discussed in the next section, yields identical results. CE-2 and CE-3 produced about the same response times in this scenario.

The simulations were run with a look-ahead time window of 6 hours. It was observed that the standard CGR and CE-1 started to have difficulties in determining the path for bundles with loads above 50 kB/s using the contacts contained within that window duration. This was not the case with CE-2 and CE-3 as can be observed in Figure 2 (b). Lower bundle losses



(a) Case 1: Mean response time. (b) Case 1: Bundle loss ratio. (c) Case 1: Throughput.



(d) Case 2: Mean response time. (e) Case 2: Bundle loss ratio. (f) Case 2: Throughput.

Figure 3. Flow metrics obtained in a network with unreliable contacts in the lunar orbiter to terrestrial station downlinks: (a), (b) and (c) with all downlinks unreliable; (d), (e) and (f) downlinks from one lunar orbiter unreliable.

benefits throughput as shown in Figure 2 (c) and increase the average response time as a result given that a large number of bundles are kept in the network, which explains the larger response time of CE-2 and CE-3 compared to the other methods.

B. Unreliable Contacts

We consider two cases as representative of the broad spectrum of possibilities involving unreliable contacts affecting the downlinks from the lunar orbiters to a terrestrial station. Case 1 assumes that the downlinks originated from each of the three lunar orbiters are affected by reliability factors of 0.95, 0.85, and 0.5. Case 2 assumes that all contacts are reliable except for the downlinks from one lunar orbiter, which have a contact reliability factor of 0.5. Figure 3 depicts the results.

In addition to CGR and the three cognitive extensions, we introduce A-CGR as a basic routing method that attempts to improve the route computation by enforcing the use of contacts with a reliability factor that is above a predefined threshold. This threshold was set to 0.9 in the experiments. This basic logic makes A-CGR functionally related to O-CGR, which evaluates the reliability of paths when making routing decisions. Although not identical to O-CGR, A-CGR provides reasonable baseline performance.

The charts on the left part of Figure 3, i.e., the ones labeled (a), (c) and (d), correspond to Case 1. It can be observed that A-CGR produced the lowest throughput given that the action of removing contacts despite offering limited reliability, also removes network capacity leading to a higher level of bundle drops. The response time of A-CGR, CE-1 and the standard CGR was observed to be very close. CE-2 and CE-3 achieved the highest throughput of all methods as both consider the impact of buffering delays when building the contact graph, with CE-3 producing lower response times than CE-2 by also considering buffering delays within the shortest path calculation. The results for Case 2 in the right part of Figure 3 follow a similar pattern although less penalized as only one link was affected by unreliable contacts. Also, the results with CE-2 and CE-3 were very close.

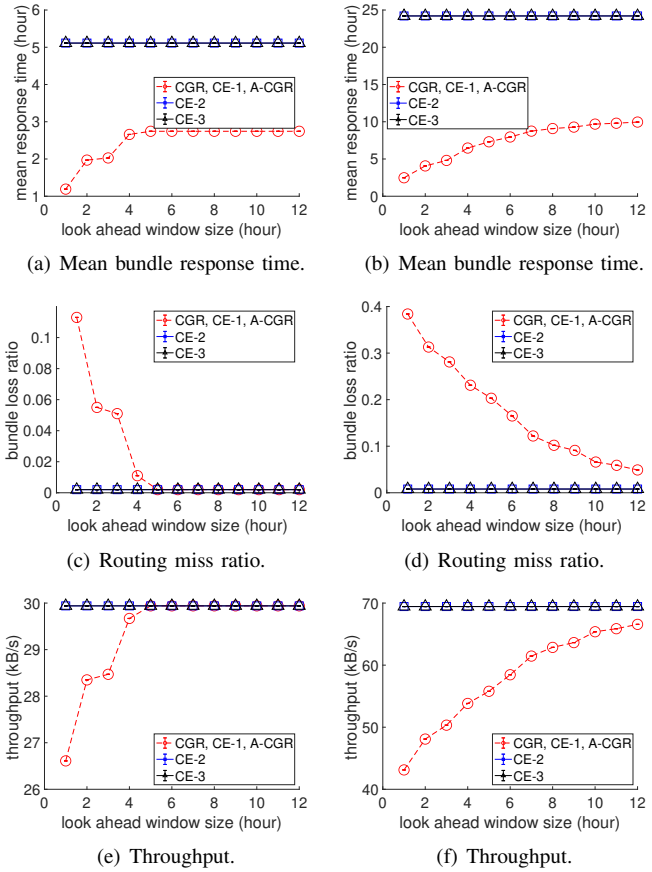
C. Impact of the Look-Ahead Window Size

The previous results were obtained with the look-ahead window size set at 6 hours. The next set of results evaluates the impact of selecting different look-ahead window sizes on routing performance. Observations were collected for two traffic load points: low at 30 kB/s and high at 70 kB/s. Based on prior results, the response time difference at high load is approximately twice as much as at low load, allowing us to observe performance differences between these two cases.

Figure 4 presents the observations collected in a network with reliable contacts. The figures in the left column show the results for low load, while those in the right column show results for high load. In each case, the flow metrics are presented in terms of average response time, bundle loss ratio, and bundle throughput. The results indicate that, for both cases, CE-2 and CE-3 yield similar outcomes across the range of look-ahead window sizes tested. However, CGR, CE-1, and A-CGR showed sensitivity to the look-ahead window size. In a

reliable network, these three methods produce identical results. At low traffic load, the CGR, CE-1, and A-CGR methods showed throughput improvement up to a look-ahead window size of approximately five hours, though with higher mean response times. At high traffic load, they follow a similar trend but over a much larger span.

Figures 5 and 6 report the same metrics for both traffic loads but with the tests running on a network with unreliable contacts. As before, two cases were observed: the first with all downlinks from the lunar orbiters affected by limited contact reliability at 0.95, 0.85, and 0.5 per orbiter (Figure 5), and another affecting a single orbiter at 0.5 (Figures 6). It can be observed that the trend initially noted in the reliable network continues in both scenarios with unreliable contacts. As the look-ahead window increases, CGR, CE-1, and A-CGR show larger mean response times, lower loss, and higher throughput. Performance differences between these methods are evident, with A-CGR tending to yield higher response times and lower throughput than the other methods in the first scenario.

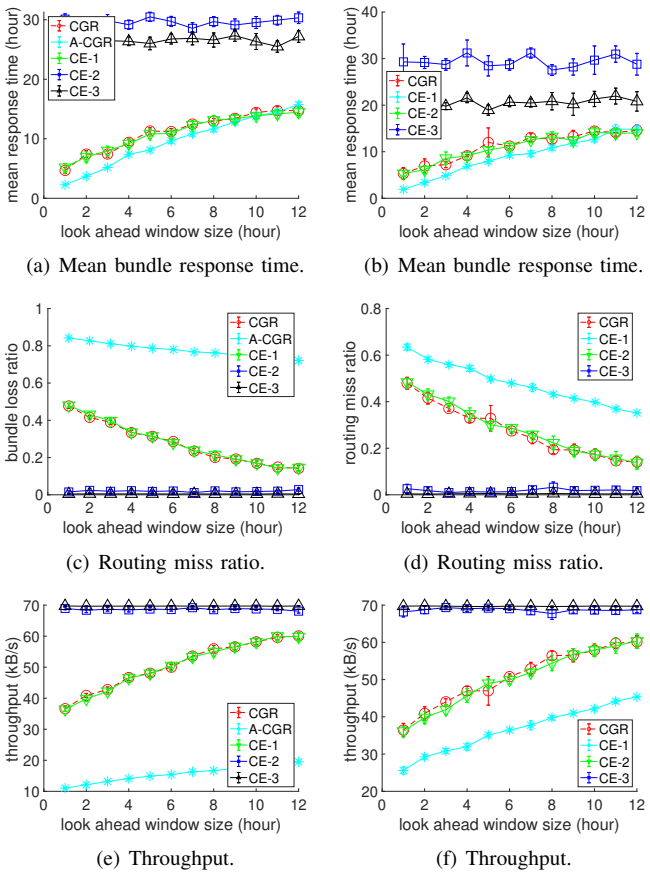


(a) Mean bundle response time. (b) Mean bundle response time.

(c) Routing miss ratio. (d) Routing miss ratio.

(e) Throughput. (f) Throughput.

Figure 4. Impact of the look ahead window size with reliable contacts for traffic loads of 30 kB/s (left column) and 70 kB/s (right column)



(a) Mean bundle response time. (b) Mean bundle response time.

(c) Routing miss ratio. (d) Routing miss ratio.

(e) Throughput. (f) Throughput.

Figure 5. Impact of the look ahead window size with with all downlinks affected by unreliable contacts (0.95, 0.85, 0.5 per orbiter) for traffic loads of 30 kB/s (left column) and 70 kB/s (right column)

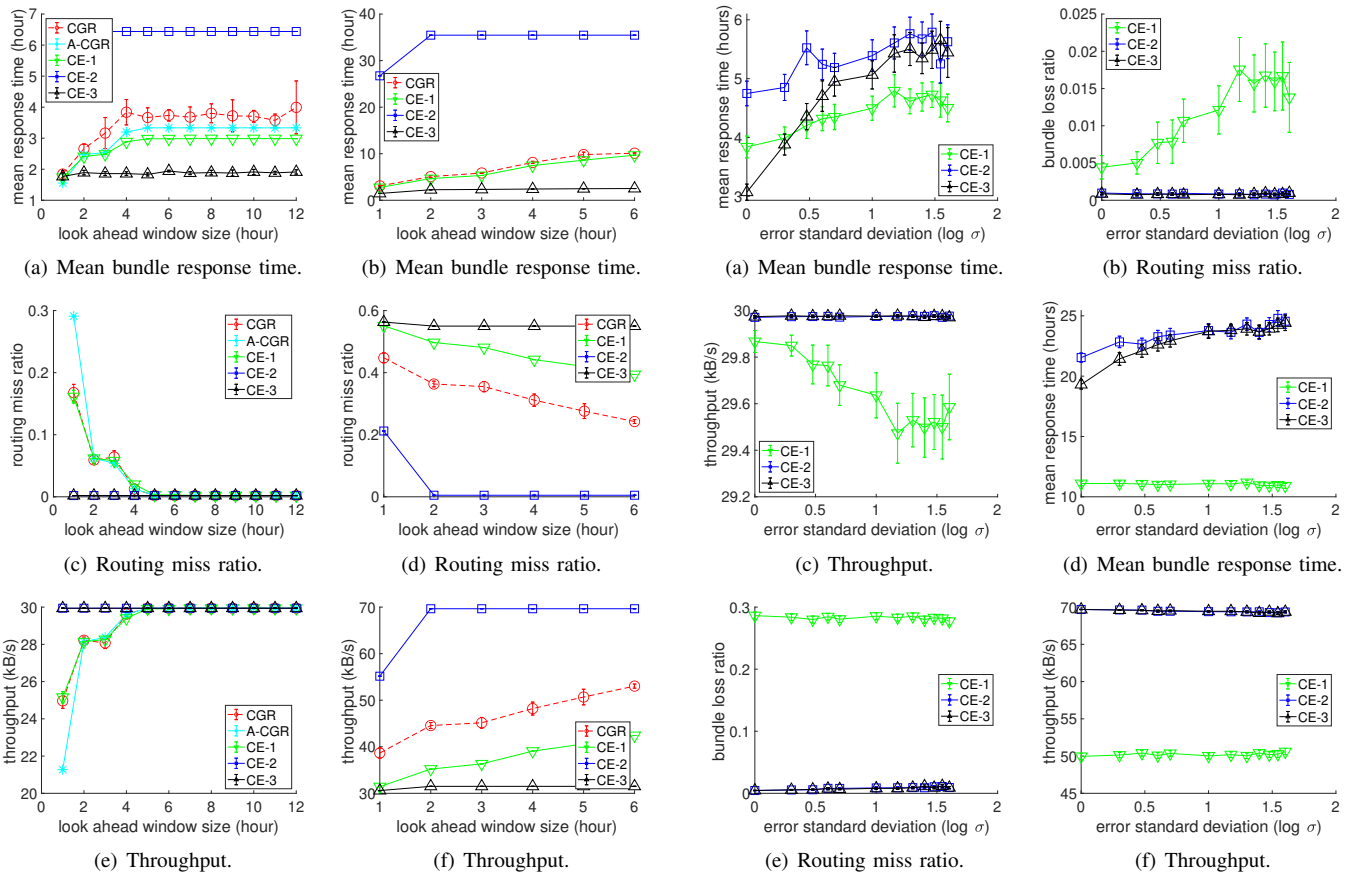
D. Impact of Prediction Errors in the Reliability Factor

The exact mechanism to determine reliability factors is left unspecified in this study, but it is interesting to observe how errors in this estimation would affect the performance metrics of the traffic flow. To achieve control over the error level, deviations where introduced to the model approximation y_j given by 3, 4, or 5, as follows:

$$y'_j = \max\{y_{min}, y_j \times \mathcal{N}(1, \sigma_e)\} \quad (6)$$

where y_{min} is a selected lower bound (0.1 in the tests) and $\mathcal{N}(1, \sigma_e)$ is a sample from a normal distribution with unit mean and standard deviation σ_e. The value y'_j is used in place of y_j when evaluating the effect of contact reliability in expression 2.

Observations were collected for two reference traffic load points at 30 kB/s and 70 kB/s as before for one of the scenarios to illustrate the impact of the estimation error given by parameter σ_e. Figure 7 shows the mean response time and throughput as a function of factor σ_e in the network with unreliable contacts. As before, assuming reliability factors of $0.95, 0.85, 0.5$ for the downlinks originating from each lunar orbiter. The top row containing figures (a), (b), (c) depicts the resulting flow metrics for low traffic, whereas the bottom row with figures (d), (e), (f) depicts the same metrics under high traffic. For low traffic loads (30 kB/s), all three extensions experience an increase in the response time with larger values of σ_e which is expected as the accuracy of the information



(a) Mean bundle response time. (b) Mean bundle response time.

(c) Routing miss ratio. (d) Routing miss ratio.

(e) Throughput. (f) Throughput.

Figure 6. Impact of the look ahead window size with one downlink affected by unreliable contacts r = 0.5 for traffic loads of 30 kB/s (left column) and 70 kB/s (right column)

(a) Mean bundle response time. (b) Routing miss ratio.

(c) Throughput. (d) Mean bundle response time.

(e) Routing miss ratio. (f) Throughput.

Figure 7. Impact of the prediction error σ_e with all downlinks originating from the three lunar orbiters affected by unreliable contacts $(0.95, 0.85, 0.5)$ and for traffic loads of 30 kB/s (top row) and 70 kB/s (bottom row).

available for routing becomes corrupted. However, both CE-2 and CE-3 demonstrated good resiliency in terms of throughout unlike CE-1 given the low bundle loss levels. With high traffic (70 kB/s), CE-2 and CE-3 show an increase in the response time with σ_e, but not CE-1 with the loss and throughput metrics of all methods unchanged. These results indicate that at the selected traffic level, the buffers of all downlinks reach a level of saturate that makes little difference choosing one downlink over another.

VI. Conclusion

In conclusion, this study assesses the integration of a Cognitive Element into CGR and its impact on routing performance. With the use of a data-driven methodology, the CE is expected to predict average single-hop bundle delivery times, accounting for latency-related factors such as CLA protocol behavior (e.g., retransmission dynamics), configuration parameters, and random variables like packet drops and unreliable contacts. This paper builds on prior work by removing the constraint in the path search algorithm that required bundle arrivals to fall within the bounds of a given contact. The use of average values from CE estimations allows better handling of the effects of unreliable contacts in the path search. Additionally, the study explores filtering contacts from the contact plan based on their predicted availability after factoring in expected buffer occupancies in the network. This approach results in shorter graphs, faster computation, and enhanced routing performance.

Comprehensive simulations conducted within an Earth-Moon network simulated context, assuming realistic contact features and accounting for unreliable contacts, show significant improvements in routing performance with the inclusion of a CE compared to the conventional CGR approach. This was evident when considering both regular network information available at a DTN node, i.e., the information contained in the contact plan, and extending this information to include network-wide buffer occupancies, i.e., global information. Unsurprisingly, the latter assumption yielded significant throughput improvements, particularly for traffic loads exceeding 50 kB/s in the tests, i.e., under congestion. For those cases, the simulations also showed lower requirements for the length of the contact plan.

We note that this study provides an implementation-agnostic assessment of the proposed approach using an analytical definition of the CE and its prediction errors. In practice, the CE is expected to be implemented using a neural network or related mechanism, with its structure, training algorithm,

and data quality affecting its prediction accuracy. The study addressed the potential performance degradation due to prediction errors. In particular, the results indicate that the CE method shows sensitivity to small errors, which can lead to delays of up to twice as much, although throughput remains largely unaffected. These findings highlight the advantages of using a cognitive networking approach to optimize space DTN performance and point to the importance of designing an accurate CE. Future research will focus on developing practical applications of this concept.

REFERENCES

[1] R. Lent, "Enhanced path reliability of Contact Graph Routing through a cognitive extension," in *The Sixteenth International Conference on Advances in Satellite and Space Communications (SPACOMM 2024)*, June 2024.

[2] D. J. Israel, K. D. Mauldin, C. J. Roberts, J. W. Mitchell, A. A. Pulkkinen, L. V. D. Cooper, M. A. Johnson, S. D. Christe, and C. J. Gramling, "Lunanet: a flexible and extensible lunar exploration communications and navigation infrastructure," in *2020 IEEE Aerospace Conference*, 2020, pp. 1–14.

[3] R. Dudukovich, D. Gormley, S. Kancharla, K. Wagner, R. Short, D. Brooks, J. Fantl, S. Janardhanan, and A. Fung, "Toward the development of a multi-agent cognitive networking system for the lunar environment," *IEEE Journal of Radio Frequency Identification*, vol. 6, pp. 269–283, 2022.

[4] C. Caini and R. Firrincieli, "Application of Contact Graph Routing to LEO satellite DTN communications," in *2012 IEEE International Conference on Communications (ICC)*, 2012, pp. 3301–3305.

[5] K. Scott and S. C. Burleigh, "Bundle Protocol Specification," RFC 5050, Nov. 2007. [Online]. Available: https://www.rfc-editor.org/info/rfc5050

[6] S. Burleigh, K. Fall, and E. J. Birrane, "Bundle Protocol Version 7," RFC 9171, Jan. 2022. [Online]. Available: https://www.rfc-editor.org/info/rfc9171

[7] S. Burleigh, C. Caini, J. J. Messina, and M. Rodolfi, "Toward a unified routing framework for delay-tolerant networking," in *2016 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE)*, 2016, pp. 82–86.

[8] J. Segui, E. Jennings, and S. Burleigh, "Enhancing Contact Graph Routing for delay tolerant space networking," in *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*, 2011, pp. 1–6.

[9] O. De Jonckère, J. A. Fraire, and S. Burleigh, "On the tractability of Yen's algorithm and contact graph modeling in Contact Graph Routing," in *2023 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE)*, 2023, pp. 80–86.

[10] J. A. Fraire, P. Madoery, S. Burleigh, M. Feldmann, J. Finochietto, A. Charif, N. Zergainoh, R. Velazco, and S. Céspedes, "Assessing Contact Graph Routing performance and reliability in distributed satellite constellations," *J. Comput. Netw. Commun.*, vol. 2017, jan 2017.

[11] P. G. Madoery, F. D. Raverta, J. A. Fraire, and J. M. Finochietto, "Routing in space delay tolerant networks under uncertain contact plans," in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.

[12] K. Zhao, R. Wang, S. C. Burleigh, A. Sabbagh, W. Wu, and M. De Sanctis, "Performance of bundle protocol for deep-space communications," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 52, no. 5, pp. 2347–2361, 2016.

[13] R. Wang, X. Liu, L. Yang, Y. Xi, M. De Sanctis, K. Zhao, H. Yang, and S. C. Burleigh, "A study of DTN for reliable data delivery from space station to ground station," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 5, pp. 1344–1358, 2024.

[14] C. Caini, T. de Cola, A. Shrestha, and A. Zappacosta, "Ltp performance on near-earth optical links," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 59, no. 6, pp. 9501–9511, 2023.

[15] J. Liang, X. Liu, R. Wang, L. Yang, X. Li, C. Tang, and K. Zhao, "Ltp for reliable data delivery from space station to ground station in the presence of link disruption," *IEEE Aerospace and Electronic Systems Magazine*, vol. 38, no. 9, pp. 24–33, 2023.

[16] M. S. Net and S. Burleigh, "Evaluation of opportunistic Contact Graph Routing in random mobility environments," in *2018 6th IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE)*. IEEE, dec 2018.

[17] A. Berlati, S. Burleigh, C. Caini, F. Fiorini, J. Messina, S. Pozza, M. Rodolfi, and G. Tempesta, "Implementation of (o-)cgr in the one," in *2017 6th International Conference on Space Mission Challenges for Information Technology (SMC-IT)*, 2017, pp. 132–135.

[18] The Interplanetary Overlay Network (ION) software distribution, "ION-DTN," https://sourceforge.net/projects/ion-dtn. Retrieved: 04-01-2024.

[19] M. Moy, R. Kassouf-Short, N. Kortas, J. Cleveland, B. Tomko, D. Conricode, Y. Kirkpatrick, R. Cardona, B. Heller, and J. Curry, "Contact multigraph routing: Overview and implementation," in *2023 IEEE Aerospace Conference*, 2023, pp. 1–9.

[20] H. Liang, X. Xu, Y. Li, and Y. Yao, "An optimized Contact Graph Routing algorithm in deep space communication," in *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp. 147–151.

[21] S. Dhara, C. Goel, R. Datta, and S. Ghose, "CGR-SPI: A new enhanced Contact Graph Routing for multi-source data communication in deep space network," in *2019 IEEE International Conference on Space Mission Challenges for Information Technology (SMC-IT)*, 2019, pp. 33–40.

[22] D. Ta, R. Menon, J. Taggart, A. Tettamanti, S. Feaser, P. Torrado, and J. Smith, "Roaming DTN: Integrating unscheduled nodes into contact plan based DTN networks," in *2023 IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW)*, 2023, pp. 1–9.

[23] F. D. Raverta, R. Demasi, P. G. Madoery, J. A. Fraire, J. M. Finochietto, and P. R. D'Argenio, "A Markov decision process for routing in space DTNs with uncertain contact plans," in *2018 6th IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE)*, 2018, pp. 189–194.

[24] F. D. Raverta, J. A. Fraire, P. G. Madoery, R. A. Demasi, J. M. Finochietto, and P. R. D'Argenio, "Routing in delay-tolerant networks under uncertain contact plans," *CoRR*, vol. abs/2108.07092, 2021.

[25] P. R. D'Argenio, J. Fraire, A. Hartmanns, and F. Raverta, "Comparing statistical, analytical, and learning-based routing approaches for delay-tolerant networks," in *ACM Transactions on Modeling and Computer Simulation*. New York, NY, USA: Association for Computing Machinery, May 2024. [Online]. Available: https://doi.org/10.1145/3665927

[26] R. Lent, "Implementing a cognitive routing method for high-rate delay tolerant networking," in *2023 IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW)*, 2023, pp. 1–6.

[27] R. Lent, "Assessing DTN routing performance in the presence of unreliable contacts," in *GLOBECOM 2024–2024 IEEE Global Communications Conference*, December 2024.

Development of UAV-aided Information-Centric Wireless Sensor Network Platform in mmWaves for Smart-City Deployment

Shintaro Mori
Department of Electronics Engineering and Computer Science
Fukuoka University
8-19-1 Nanakuma, Jonan-ku, Fukuoka 814-0180, Japan
E-mail: smori@fukuoka-u.ac.jp

Abstract—**This paper presents an information-centric, wireless sensor network-based ecosystem for smart-city applications. The proposed scheme targets the integration of a non-terrestrial wireless network using unmanned aerial vehicles, leveraging higher frequency bands for future broadband wireless communication in disaster-resilient smart cities. To demonstrate the feasibility of the scheme, we conducted a preliminary evaluation of computer-calculation capabilities and network performance, including throughput and jitter in the application and TCP layers. In addition, as part of a scenario for disaster-information sharing systems, we conducted a video-streaming test through an on-site experiment and developed a prototype device for edge-side node device, aimed at exploring new wireless networking technologies in promising mmWave bands.**

Keywords—Information-centric wireless sensor network; mmWave communications; unmanned aerial vehicle

I. INTRODUCTION

Emerging technologies, such as the Internet of Things (IoT), the metaverse, and Artificial Intelligence (AI), are widely utilized and applied in our daily lives, contributing to making cities smarter. For example, IoT can improve legacy barriers, AI can assist in decision-making, and the metaverse can accelerate non-face-to-face communications. Thanks to the massive amount of valuable information they provide, several problems related to urbanization and social life can be mitigated. Smart cities represent a new paradigm that can lead to the provision of smart services centered around healthcare, transportation, energy, and natural disasters, making cities greener, safer, and friendlier for residents [2]. At the same time, natural disasters (earthquakes, typhoons, hurricanes, floods, and other geologic processes) can potentially cut or destroy the existing territorial wireless network infrastructure. Typical (smart) cities separately construct IoT systems to provide daily and disaster-related applications. On the other hand, the concept of this paper will coexist with those systems, designed as disaster-resistant smart cities. Namely, disaster-related information is shared using the same system when a disaster occurs. This scheme brings two advantages: economic efficiency (i.e., eliminating the need for an exclusive disaster communication and networking system) and improved availability (i.e., the system can be more available in emergencies because it is already in place for daily operations).

A Wireless Sensor Network (WSN) is an essential foundational technology supporting these application services. In WSNs, many sensors and actuators are heterogeneously interconnected with edge, cloud, and user networks to collect and distribute sensing data, such as text-based sensing, real-time streaming, and 3D high-capacity data. Specifically, the data includes various protocols, demands, and priorities, all of which should be accommodated in the same system, particularly in the Physical and Medium Access Control (PHY/MAC) layers. To address this situation, we use three technologies: Unmanned Aerial Vehicles (UAVs), millimeter-wave (mmWave), and Information-Centric Networking (ICN).

UAVs have been widely utilized in various fields, such as commercial uses, industrial uses, and hobbies. For example, aerial photography and video-streaming technologies with high-resolution cameras provide a well-known multi-purpose solution in UAV use cases. However, UAVs are known as aerial base stations in future wireless telecommunications systems and as a component of non-terrestrial networks, along with satellites (geostationary, quasi-zenith, and low-earth-orbit) and high-altitude platforms (small planes). As previously described, UAVs can expand wireless coverage to disaster-stricken or rural areas and serve as new data producers and carriers [3][4].

MmWave communications have been recognized as a revolutionary new research domain in future mobile networking technologies, capable of accommodating various data streams. MmWaves can support a wider bandwidth compared with current mainstream spectrums, such as ultra-high frequency and microwave bands. Due its vast spectrum bandwidth, mmWaves enable multi-gigabit data transfer [5], and the spectrum is globally assigned (for example, in 28, 38, and 60 GHz in cellular networks utilized in 3GPP-FR2 [6]). Therefore, mmWave communication is positioned at the forefront of the global frontier and is an essential element in discussions on next-generation wireless communications.

ICN is a remarkable candidate for a future network architecture that shifts the focus from host locations to content data [7]. At the network layer, the protocol suite must be designed on the basis of an autonomous and decentralized network architecture. ICN has an advantage in data-intensive applications optimized for content retrieval in an autonomous, decentralized ad-hoc network environment. The data are named instead of the address, enabling end-users to discover and obtain the data via names, resulting in a location-free

structure. Another vital feature of ICN is in-network caching. Namely, the data are copied and stored in cache memories on network nodes to facilitate further data retrieval. In addition, the data are handled separately by individual content units, i.e., the data can be self-certified and encrypted by their producer, contributing to improved security.

Applying ICN to WSNs yields an Information-Centric Wireless Sensor Network (ICWSN) [8], which positively affects network performance by boosting data delivery and improving data fetching delay. ICWSNs have the potential to address challenges arising from cases where most WSN devices are resource-constrained with radio frequency, processing resource, energy, and memory limitations. In addition, the data abstraction resulting from ICN design contributes to easy data spreading and simplifies management of such systems, including the network-transport layer protocol suite.

In our study, we integrate these technologies by implementing a UAV-aided ICWSN in mmWaves to effectively collect and distribute the data. As previously stated, mmWaves have the characteristic of propagating straightforwardly and therefore being significantly attenuated by penetration, atmosphere (oxygen), heavy rain, and moisture-containing material. Fortunately, UAVs can establish more reliable Line-of-Sight (LoS) links for ground nodes, leading to a better communication channel. In our baseline paper [1], we provided the blueprint of the proposed scheme and several fundamental evaluation results. This paper further investigates an ecosystem to support application services for disaster-resilient smart cities. At the same time, we need to ensure the proposed scheme can provide a high data rate and low latency with stable connectivity to establish a new sustainable smart-city ecosystem.

Consequently, in this paper, after briefly surveying the covered areas, we describe the development of testbed devices and constructed test fields in our project. Using these facilities, we evaluate whether the proposed ICWSN platform can function effectively for the assumed smart-city deployment scenario. These evaluations contain mmWave identification, including fundamental communication characteristics and network performances, for future application services. In addition, we demonstrate a practical implementation of an integrated air-to-ground mmWave ICWSN platform in the test field and address the feasibility of operating wideband, real-time applications in disaster-resilient smart cities. Furthermore, we reveal that the proposed system is ready for deployment in an actual city with prototype implementations of both WSN and infrastructure-side devices.

The remainder of this paper is organized as follows. Section II discusses related work. Section III provides a brief overview of the development of the proposed ICWSN test field. Section IV describes the proposed scheme. Section V presents the evaluation results and discussion. Finally, Section VI concludes this paper with a brief summary and mention of future work.

II. Related Work

For smart-city application platforms, Malik et al. [9] presented a comprehensive assessment of the smart-city concept, surveying its possible applicability in upcoming technological growth on the basis of an exhaustive existing literature investigation regarding smart cities. Vera-Panez et al. [10] investigated the design and implementation of the WSN platform for the fog-computing paradigm. For evacuation during emergencies, Ahanger et al. [11] presented an intelligent evacuation framework by integrating IoT, edge, and cloud-computing paradigms. The proposed framework utilizes IoT technology to collect ambient data and track occupant movement on the basis of location.

For the foundational technologies of mmWaves in physical and MAC layers, Pan et al. [12] proposed a cooperative communication scheme using network coding for vehicular ad-hoc networks to enhance resilience to transmission errors. The scheme was designed based on a graph-theoretic approach, considering the directionality of mmWaves and the effect transmission redundancy. In the 5G networks, traditional approaches overlook critical handover issues related to interference and channel intermittency in dense network environments. Ganapathy et al. [13] investigated the handover technologies. For future wireless networking technologies, including mobility and ad-hoc networks, Luo et al. [14] proposed a joint communication and positioning technique for resource allocation algorithms.

In network and transport layers, Zhang et al. [15] revealed fundamental issues under highly variable links for end-to-end mmWave applications through a comprehensive simulation-based study of various congestion control algorithms in TCP. Khorov et al. [16] classified TCP schemes and investigated the performance of the promising QUIC method in high-frequency bands. Poorzare et al. [17] revealed that mmWaves 5G could provide high data rates with low latencies. However, maintaining a reliable end-to-end connection throughout 5G mmWave networks is challenging due to the fluctuation of high-frequency channels, primarily because TCP, the main protocol exploited by the transport layer, cannot perform sufficiently under these conditions. Through actual experiments, Yang et al. [18] examined and discussed the performance of several TCP congestion control algorithms in the 60-GHz band and inspected improvements in TCP performance over mmWave hybrid networks using TCP proxies in single-flow and multi-flow scenarios.

Regarding the studies of IEEE 802.11 ad/ay and 5G cellular networks, Wang et al. [19] conducted an experiment using a testbed to exploit the high gain of mmWave RF and flexible configuration of embedded systems. Validation and field tests show that the developed testbed could provide up to a 2.3-Gbps network layer data rate in a single channel with low latency and support point-to-multi-point transmission aided by relay. Aldubaikhy et al. [20] investigated a fixed wireless access system, including unlicensed Wi-Fi and licensed 5G networks. This paper described a comprehensive review of the considered new protocol specifications and design elements of the DNs and provided a case study proposing a low-complexity concurrent transmission protocol to enhance the network performance while mitigating the interference.

For air-to-ground integrated networks, Tuan Do et al. [21] surveyed recent approaches to UAV-aided communication

networks in mmWaves and presented their main characteristics based on intelligent learning-based methods. Dabiri et al. [22] proposed a backhaul architecture in mmWaves using fixed-wing UAVs to maximize the average channel capacity, including a modelization of a single relay fixed-wing UAV-based communication system that considers realistic physical parameters and investigates crucial channel parameters' effects, such as antenna pattern gain and flight path, on the system's performance. Sanchez et al. [23] formulated a stochastic channel model for mmWave UAV communications under hovering conditions. Cheng et al. [24] proposed a new three-dimensional channel model for air-to-ground mmWave communication environments based on the ray tracing theory.

As the foundational technologies for comprehensive mmWave communication systems, Xiao et al. [25] surveyed beamforming techniques as an essential technology in UAV-aided mmWaves. Zhao et al. [26] proposed a geometric analysis method to detect blockage in multi-UAV communication systems and addressed a user-scheduling formulation and its efficient algorithm to enhance spectral efficiency. Chang et al. [27] proposed a new integrated scheduling method of sensing, communication, and control in mmWave and THz communications within UAV networks, including an analysis of interactions among these functionalities and providing a new definition from a motion control perspective, i.e., the relationship between sensing-control pattern activation and data rate.

III. MmWave Wireless Information-Centric Networking and Test-Field Development

This section provides an overview of mmWave communication and information-centric wireless networking technologies, and we describe a test-field development.

A. MmWave communication system

A global standardized communication system, like Wireless Local Area Networks (WLANs) enables local network connectivity for various devices such as computers, tablets, smartphones, and IoT devices. The IEEE 802.11 family (Wi-Fi) is globally recognized as a WLAN standard, and IEEE 802.11 ay is the latest version of mmWave communications on unlicensed 60-GHz bands, aiming to improve upon IEEE 802.11 ad while guaranteeing backward compatibility for legacy users. In contrast to other systems, future mobile (cellular) networks, like local or private Fifth Generation (5G) are also ready to leverage mmWaves. However, IEEE 802.11 ay has the advantage of widespread user terminals, which yields economic benefits in common device usage during smart-city deployment phases.

When we deploy the mmWave WLANs using IEEE 802.11 ay-compliant systems, they can operate under the point-to-point and point-to-multi-point topologies in both indoor and outdoor environments. In other words, the network can be constructed on the basis of meshed-network technology, providing a cost-efficient broadband wireless solution to replace fiber optical networks in city areas. In addition, meshed networks can find the most efficient path for information en route under a dynamic network environment with multi-hop wireless communications. Namely, if an intermediate node fails, another can immediately take over its role, thereby improving the network's availability. This feature is suitable for a network that supports disaster-resilient smart cities.

B. MmWave communications platform

To deploy an IEEE 802.11 ay-compliant meshed network, Distribution Nodes (DNs) and Client Nodes (CNs) are used, i.e., multiple DNs are interconnected to form a backhaul network, enabling end-users access via CNs. As a commercial product, Meta (Facebook) offers Terragraph (TG) as an IEEE 802.11 ad/ay-compliant meshed network [28]. TG aims to provide operators with an alternative low-cost solution to provide a similar cellular network or regional Internet service, such as a metropolitan area network.

In the network layer, the TG network communicates via multi-hop transmissions with a maximum of 15 hops, and the router node supports the Open/R routing protocol. In the Medium Access Control (MAC) layer, TG is compatible with the IEEE 802.11 ad/ay specification but uses time-division access instead of contention-based carrier sense for system-complexity improvement. In the PHY layer, TG only provides single-carrier modulation among several IEEE 802.11 ad/ay-available PHY methods. Specifically, the TG system selects the pair of modulation and coding schemes among Binary Phase-Shift Keying (BPSK), Quadrature PSK (QPSK), and 16-Quadrature Amplitude Modulation (QAM), and a code rate of R = 1/2, 5/8, 3/4, or 13/16, as a rate (link) adaptation technique based on the received Signal-to-Noise-Ratio (SNR) and Packet Error Ratio (PER). In addition, the TG antenna uses a beamforming technique that adaptively modifies to maximize signal quality when the radio link is disconnected or the system turns on, enabling half-duplex physical data rates up to 4.6 Gbit/s.

C. Information-centric networking and its platforms

In the ICN system, users obtain data from the nearest node without servers or clouds, thereby detaching the data from its original location and reducing network congestion and latency. To enable the ICN mechanism, intermediate nodes maintain three databases: Pending Interest Table (PIT), Forwarding Information Base (FIB), and Content Store (CS). In PIT, interest packets (data-request messages), including in/out interfaces and names, are registered, enabling requested data to be forwarded to the data requester by back-tracing on the basis of the PIT information. The interest packet is further forwarded to the next node based on FIB when the data has not been stored in the cache memory. Namely, if the node has the data, it replies with it; otherwise, it relays the interest packet. The cached data is managed using CS, enabling nodes to inquire whether the requested data has been stored.

As ICN platforms, Data-Oriented Network Architecture (DONA) was the first ICN framework to use flat names in place of hierarchical addresses. Content-Centric Networking (CCN) adopts a hierarchical naming scheme and serves as a fundamental design in the ICN platform. Named-Data Networking (NDN) is the first CCN-based ICN framework and one of the renowned ICN platforms for research. In

another branch of CCN, PURSUIT features a hierarchical routing structure and uses the distributed hash table-based routing scheme. NetInf uses a publish-subscribe scheme and maps names to locators. CCNx, the latest CCN-based ICN framework has a standardized protocol.

D. Development of ICWSN framework

We have been developing a testbed device and test field to evaluate an mmWave ICWSN framework. For the hardware of the mmWave TG communication system, we used the BeMap MLTG-360 as a DN and MLTG-CN as a CN [29]. According to the catalog-based specification, DNs can transmit up to distances of about 300 m, and all wireless links must be LoS with no foliage, walls, or other obstacles between antennas. The maximum transmission Effective Isotropic Radiated Powers (EIRP) are 43 dBm (DN) and 38 dBm (CN), and the antenna gains are 28 dBi (DN) and 22 dBi (CN). Each antenna consists of a phased array with 64 elements, and the steering angular ranges are [−45°, 45°] in the azimuth plane (φ) and [−25°, 25°] in the elevation plane (θ). In the beamforming scheme, an index value representing the antenna direction can be selected among predefined beamforming patterns. In Japan's regulation of the Radio Act, TG is assigned the unlicensed 60-GHz band (57–66 GHz) with four channels, consisting of 58.32, 60.48, 62.64, and 64.80 GHz (central) frequency bands each with a 2.16-GHz bandwidth.

For the middleware of the ICN platform, we used Cefore [30]. Cefore is an open-source CCNx-based ICN platform available on Linux (Ubuntu). Its deamon processes include cefnetd, which exchanges the data and forwards interest packets based on PIT and FIB, and csmgrd, which provides an in-network caching scheme based on CS. Note that cefnetd also provides a simple on-memory caching scheme. To integrate Cefore into the system, we can register and obtain the data from the application software (program) using the following commands: cefputfile and cefgetfile for sending and receiving static data, respectively, and cefputstream and cefgetstream for broadcasting and receiving real-time streaming data, respectively.

The proposed scheme can manage a unified ICWSN distributed over a wide area. The test fields were located at the KOIL mobility field (Kashiwa, Chiba), the baseball field in Advantech Japan (Nogata, Fukuoka), and Fukuoka University (Fukuoka). These fields are inter connected via a broker deployed on a cloud server and logically placed on the same network segment through virtual private network connections [31][32][33]. In this paper, we focus more on the baseball field as this is where we conducted the experiment. The ICWSN is composed of a group of Sensor Nodes (SNs), Relay Nodes (RNs), and a Private (self-operated) Base Station (PBS). RNs include a Ground RN (GRN) and an Aerial RN (ARN) equipped on a land-based access point and a low-altitude UAV, respectively. Figure 1 shows an overview of the test-field sites. The testbed has two external network connections for network-connective availability: a primary TG network and a secondary cellular network.

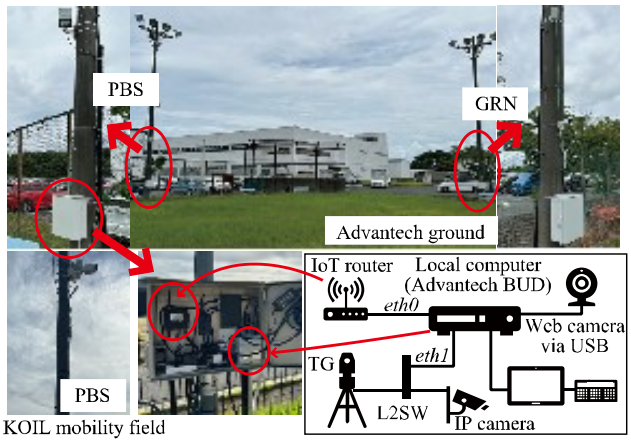


Figure 1. Overview of the test field developed in [31][32][33].

IV. Computer-Calculation Capability

In this section, we present the statistical characteristics of sensing data in smart-city applications. Using the formulated model, we evaluate the computational processing capability with respect to data compression and encryption.

A. Statistical modelization of real sensing data

As a statistical model of sensing data, we used a river-monitoring system developed in our previous study for disaster-resilient smart-city applications [34]. The system estimates river flow direction and velocity to prevent internal flooding caused by typhoons and heavy rain. The system was experimentally placed in the Onga River, classified as a first-class river, and passed through Nogata (Fukuoka, Japan). We captured photographs of the water surface and ambient-area scenery as sensing data. We prepared 25 datasets containing 20 images, totaling of 500 data, for statistical model modelization. Note that each (image) data has the same conditions, such as location, camera angle, and recording time, but differed slightly in recording time due to continuous shooting.

From the datasets, we obtained a statistical model of the sensing data representing a Cumulative Distribution Function (CDF) of byte-by-byte codewords (`0x00` to `0xFF`). We calculate the frequency of 256 different codewords for each data and sorted the occurrence probabilities in descending order. Figure 2 shows the results of the calculated CDFs drawn as superimposed lines. The red line represents the case where the codewords occur equally. The results show that the sensing data that were captured in the on-site field indicate a particular bias because the lines should overlap the red line if codewords were equally distributed. Despite varying capture conditions, the statistical characteristics of the data are similar. Subsequent evaluations will utilize this statistical model.

B. Hardware selection and analysis data preparation

In ICWSN systems, named data, in which sensing data are encapsulated, are primitively compressed and encrypted for efficiency and reliability. ICNs provide secure mechanisms for individual data. In the rest of this section, we aim demonstrate the feasibility of our proposed scheme's edge-side network installation and evaluate its fundamental capability in terms of processing time; we discuss the effect on the network in terms of delay. In this paper, we consider three types of node devices: user-terminal nodes, edge nodes, and SNs. In the next section, we will assess the performance of each categorized node using real hardware devices, providing a benchmark and feasible conditions to construct the network and deploy it in smart cities. For the hardware, we used a MacBook Pro as a user-terminal node, an Advantech AIR-020 as an edge node, and Advantech EPC-S202 and Raspberry Pi 4B as SNs, respectively. The specifications of these devices are listed in Table I. In particular, the Advantech AIR-020 and EPC-S202 are both highly reliable embedded computers for industrial usage, and their overview are shown in Figure 3.

For the computer-capability evaluation, we generated and prepared 100 randomized datasets with 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1,000, 5,000, and 10,000 kbytes. These datasets were generated using the computer program implemented in the C++ language and were based on the statistical model that was modelized in the previous section. The variations of the datasets, particularly minimum and maximum sizes, were assumed to be text-based sensing data with an ICN header (and footer) and a typical ICN chunk size, respectively. Note that the named data exceeds the predefined size; it is divided into short-length data units called chunks, similar to packet fragmentation in IP networks. In addition, we used randomized data to ensure it was free of legal copyright and portrait rights.

C. Experimental results

In this section, to evaluate the performance of edge-side node devices, we first measured the processing time for data compression and decompression processes. In particular, we used the deflate algorithm and the Lempel-Ziv-Markov-Chain Algorithm (LZMA) as data (de) compression algorithms. The deflate method uses a sliding dictionary and Huffman coding techniques. It is widely used for file archiving implementation and is distributed with operating systems in the ZIP file format. The LZMA method is enhanced with a range coding technique to improve compression performance. In addition, it is one of the most highly efficient data-compression methods. Both are lossless data-compression schemes and are supported by several embedded devices. The LZMA method generally has a faster decoding speed than the deflate method.

Figure 4 shows the results, including the processing time for encoding and decoding using the deflate and LZMA methods, respectively. Note that both the vertical and horizontal axes use logarithmic scales. The processing time of data compression and decompression was practically the same in the region where the data size was less than 100 kbytes, with the deflate and LZMA methods requiring 10 ms or less for up to 100 kbytes, regardless of the device. In other words,

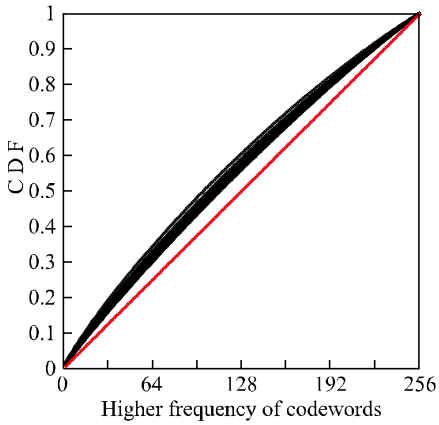


Figure 2. Statistical characteristics of IoT data: CDF characteristics for frequency of byte-based codeword occurrences

TABLE I. TEST DEVICE SPECIFICATIONS

Terms	Parameters			
	MacBook	***AIR-020***	***EPC-S202***	***Raspberry Pi 4B***
CPU	M3 Pro (11 cores)	ARM (6 cores)	Intel Atom (4 cores)	ARM A72 (4 cores)
RAM	18 GB	8 GB	4 GB	4 GB
OS	Sonoma	Ubuntu 18.04 with JetPack	Ubuntu 20.04 LTS	Raspbian 6.6

(a) (b)

Figure 3. Overview of the industial-use reliable computers: Advantech (a) AIR 020 and (b) EPC-S202

this indicates a minimum processing delay required for data (de) compressing, including processing overhead, such as initialization and termination of the methods, which is unavoidable. In the region with more than 100-kbyte data, the processing time increased proportionally to the data size.

As a data encryption scheme, we used the Advanced Encryption Standard (AES) method, which is widely utilized for various computer systems and is the most popular common-key (symmetric-key) cryptosystem. Systems with an AES-based encryption scheme select a key length, such as 128, 192, or 256 bits, with longer key lengths offering stronger cryptographic strength. In this paper, we selected 128 and 256 bits since these are widely used in general IoT devices. In the experiment, we used OpenSSL [35], an open-source

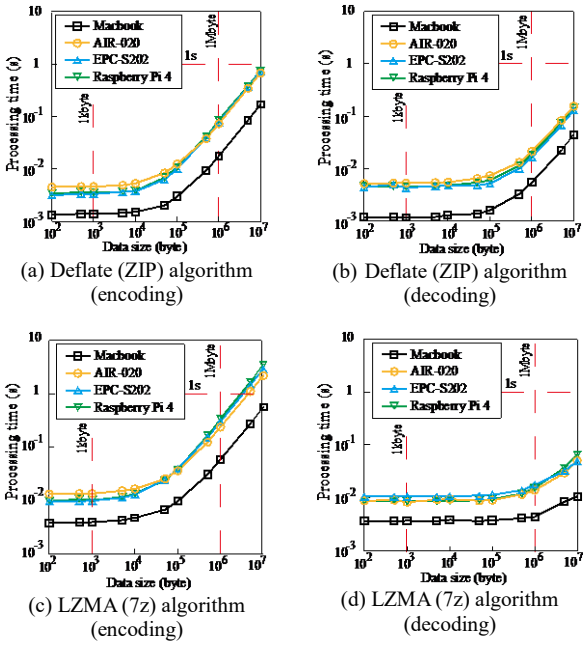
software library providing secure communications on Internet servers, such as HTTP over TLS/SSL protocols.

Figure 5 shows the results, including the processing time for encoding and decoding using AES methods with 128 and 256 bits, respectively. The results show that the processing times for data encryption follow the same trend to those for data compression. However, the MacBook performed better than other devices. Note that the vertical axis is logarithmic. While the AIR-020 was equipped with a Jetson platform, its AI optimization mechanism did not effectively work for general-purpose processing, like data compression and encryption. The MacBook demonstrated the best performance among the tested devices as it is equipped with powerful end-user processors. Conversely, edge-side WSN devices must consider outdoor use, and heat dissipation and stability issues, leading to compromised specifications with an approximate processing delay of 100 ms.

D. Discussions

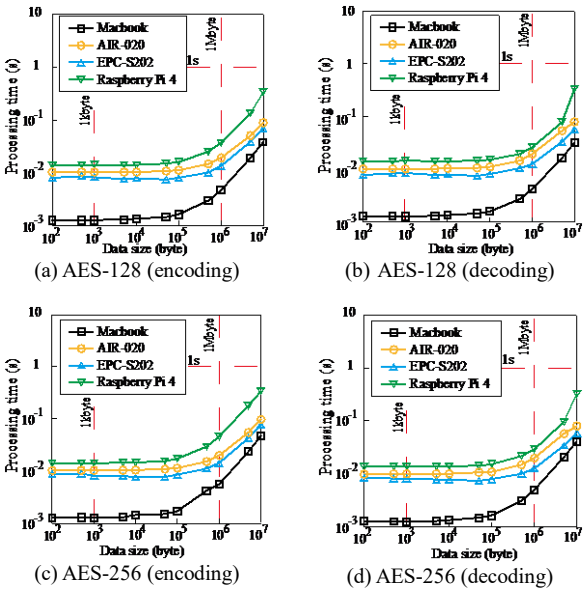
In the previous section, we found that the processing delay time of the edge-side node devices was under approximately 100 ms. In this section, we will discuss the effects of these results on ICWSN deployment in smart-city applications. For application services where delay is not a significant concern, such as those with hourly (or daily) data-collection intervals like delay-tolerant networking-based application services, the delays are in significant. For example, river-monitoring systems for disaster-resilient smart-city applications require sensing data every hour under normal weather conditions; however, this interval may decrease to every 20 minutes or less when a disaster occurs [34].

The proposed ICWSN can be used in such scenarios, but we should acknowledge the limitations regarding wireless network construction. Regarding network topology, low-power wide-area networks (a well-known IoT networking platform) directly connect BSs and SNs (star-type networks). The limitations are negligible if the BS has sufficient capacity to accommodate a sufficient number of SNs. Alternatively, networks constructed on the basis of relay or ad-hoc network technologies, IEEE 802.11 ay experiences significant delays due to multi-hop wireless transmission accumulating to several hundred milliseconds per wireless section. Despite this potential second delay, the ICWSN remains applicable. However, delay-sensitive application services, such as those that include actuator (or actor) control and real-time streaming solutions, require more complex mechanisms for overall layers, which is a consideration for future work and out of scope in this paper.



(a) Deflate (ZIP) algorithm (encoding)
(b) Deflate (ZIP) algorithm (decoding)
(c) LZMA (7z) algorithm (encoding)
(d) LZMA (7z) algorithm (decoding)

Figure 4. Results of data compression and decompression performance



(a) AES-128 (encoding)
(b) AES-128 (decoding)
(c) AES-256 (encoding)
(d) AES-256 (decoding)

Figure 5. Results of data encryption and decryption performance

V. MmWave Communication Performance

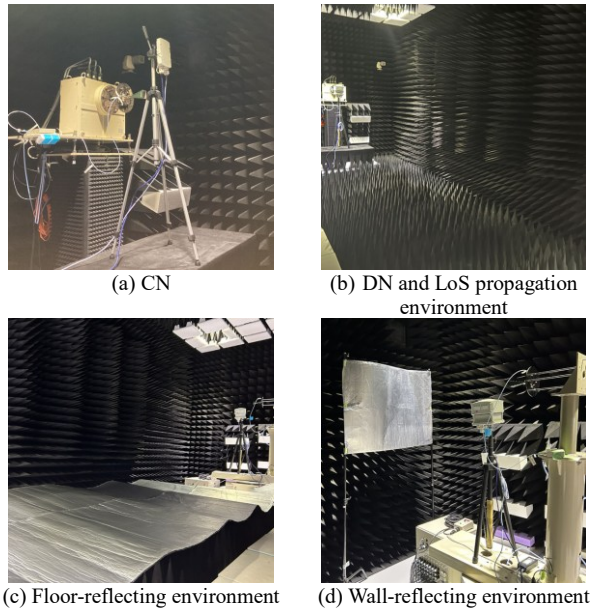
In this section, we investigate the network performance in which mmWaves affect the upper layers, such as network and application layers.

A. Experiment environment

As previously mentioned, we use BeMap's TG [29] for the mmWave communication system, and the detailed specification are listed in Table II. The experiment was conducted in an anechoic chamber (shielded room) assumed to be an ideal environment for mmWave propagations. Figures 6 and 7 show the scenery and configurations of the experimental environment, respectively. As shown in Figure 7(a), a DN and CN were placed face-to-face at a distance of 6.9 m, and the radio wave planes remained horizontally oriented; the DN and CN are shown in Figures 6(a) and (b), respectively. Both antenna surfaces were vertically oriented to the ground, and the direct line between them was kept obstacle-free. The experiments were conducted

TABLE II. TG SPECIFICATIONS

Terms	Parameters	
	DN	***CN***
Size	20x20x20 cm	18x11x4.3 cm
Weight	3.9 kg	1.1 kg
Maximum energy consumption	75 W (4 radios) 30 W (1 radio)	15 W
Frequency	Fc = 58.32 GHz (57.0-59.4 GHz)	
Tx power	43 dBm	38 dBm
Antenna	Gain: 28 dBi	Gain: 22 dBi
	Phased array antenna with 64 elements Azimuth range: -45° to +45° Elevation range: -25° to +25°	
LAN	Gigabit Ethernet (1x port)	
PHY/MAC	IEEE 802.11 ad/ay	
Modulation method	OFDM with BPSK, QPSK, 16QAM	
Required RSSI	-66 dBm (MCS9), -61 dBm (MCS12)	
Theoretical range	150 m (MCS9), 100 m (MCS12)	



(a) CN (b) DN and LoS propagation environment (c) Floor-reflecting environment (d) Wall-reflecting environment

Figure 6. Experimental environment in an anechoic chamber

assuming three propagation environments: LoS propagation, floor-reflecting, and wall-reflecting, as shown in Figure 7(b). In the LoS propagation environment, the floor and walls were filled with radio wave absorbers (mmWave-qualified), and we assumed a condition with no radio waves other than direct waves, as shown in Figure 6 (b). For the floor-reflecting and wall-reflecting environments, the floor and walls were covered with aluminum sheets that can reflect radio waves, and we assumed the conditions under which direct and reflected waves could arrive, as shown in Figures 6(c) and (d), respectively.

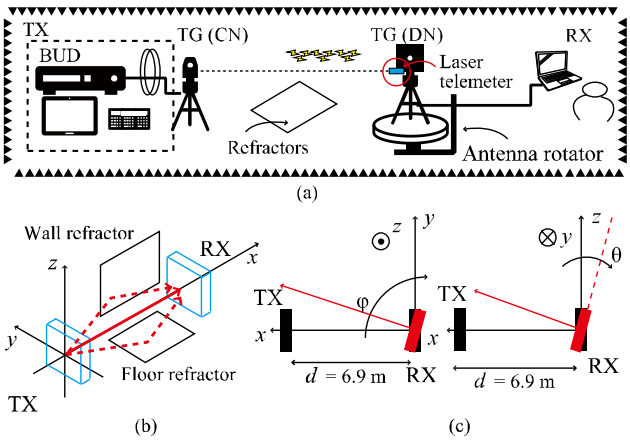


Figure 7. Configuration of experimental environment: (a) Diagram of experimental equipment, (b) Radio propagation conditions between the transmitter and receiver, and (c) relationship between antenna surfaces of transmitter and receiver antenna

To reconfigure the beamforming directions to face each other, we restarted the CN device. Note that the BeMap's TG devices were reset to the beamforming configuration upon rebooting or when the signal becomes unclear. During the experiment, we continuously monitored the PHY information to ensure that beamforming was not reset. To measure network performance, the CN was fixed, and the DN was placed on the antenna base and rotated using an antenna rotator, as shown in Figure 7(a). The antenna rotator can move within the steering angular ranges of [0°, 90°] in the azimuth plane (φ) and [0°, 20°] in the elevation plane (θ), as shown in Figure 7 (c).

B. Results for network performacne

To evaluate the network performance of the scheme, we measured TCP throughput using iPerf3 and ICN throughput using Cefore. Figure 8 shows the experimental results, including TCP and ICN throughput for the LoS propagation, floor-reflecting, and wall-reflecting conditions. The TCP throughput is an average (mean) value from three rounds, measured every second for 30 s. ICN throughput is an average (mean) value, measured when retrieving three different data. In particular, the ICN platform used Cefore, a ccnx-compliant protocol stack previously mentioned in Section III.D. We installed Cefore only on the control computer of ARN and PC; the data can be exchanged via the cefnetd and csmgrd daemon processes from the application program. ICN throughput was calculated on the basis of the time intervals between provider commitments using the cefputfile command and receiver retrievals using the cefgetfile command. To mitigate the effect of in-network caching on throughput, we retrieved different data files each time to avoid repeated requests for the same data.

Figure 8(a) shows the result of TCP throughput. The radio link remained connected regardless of angles θ and φ. Except for φ > 90°, where it was disconnected, and θ > 20°, where the wireless link was unstable, preventing continuous TCP

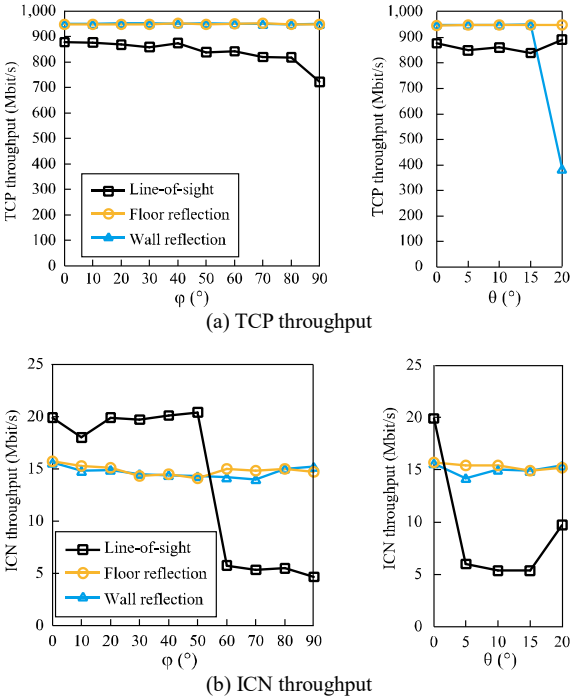


Figure 8. Results of network performance in mmWaves: (a) TCP throughput and (b) ICN throughput

throughput measurement. Compared with the LoS propagation conditions, TCP throughput improved by 13.2% for the case with reflected radio waves (in the floor- and wall-reflecting conditions). This result indicates improved receiver environment due to reflected and direct waves, which occur in the anechoic chamber. However, practical throughput may not be achievable at $\varphi > 45°$ for TGs deployed in real cities because of complex radio propagations, including reflection, interference, scattering, and shadowing.

Figure 8(b) shows the result of ICN throughput. Unlike TCP throughput, ICN throughput for direct waves only was higher than that of the reflected waves at $\varphi < 50°$, but worse with that of both the floor- and wall-reflecting conditions. This is because ICN layer performance improves when received with the reflected waves' assistance. Similarly, this is also the case with $\theta > 5°$. Although TCP throughput performs reasonably well, ICN throughput has different characteristics, i.e., its maximum values are significantly worse than those of TCP throughput.

These are several conceivable reasons for these observations: mismatched beamforming of mmWaves, significant bugs in the Cefore platform, or TCP algorithm inefficiency for mmWave communications. Therefore, we should seriously consider integrating the overall protocol stack across application, network (transport), and physical layers as it is beyond the scope of this paper.

VI. Demonstration of Air-to-ground Integrated ICWSN

In this section, we present the experimental results, including application demonstrations, for our air-to-ground integrated ICWSN system. We implemented an aerial node device and evaluated its network performance using mmWaves, demonstrating the video-streaming app in our test fields using the implemented device. In addition, as an SN device, we present a prototype testbed designed to function as a zero-touch edge-side node for reliable and self-organization capabilities.

A. Development of ARN device

The ARN device consists of a control computer, camera, and CN mounted on the UAV, as shown in Figure 9. Note that we used an industrial drone with a payload capacity of several kilograms. The placement of each component was balanced and adjusted for uninterrupted flight operation. The control signals for drone flight used the licensed VUHF band radio rather than using mmWave communications. The control computer used the Advantech Brain Unit for Drone (BUD) device (two-core 1.8 GHz Intel Atom E3930 CPU, 4 GB RAM, and Ubuntu 20.04 OS). The camera and CN were connected to the computer via USB and Ethernet (wired LAN) cables, respectively. As shown in Figure 10, due to Japan's Radio Act and Civil Aeronautics Act regulations, the UAV flew captive flights (not free), anchored by a mooring rope with an integrated LAN cable for PoE to the CN. Figure 11 shows the network model of the experiment. For the end-user terminal, a PC (two-core 1.3 GHz Intel Core i5U CPU, 8 GB RAM, and Ubuntu 20.04 OS) was directly connected to the DN on the PBS, and static IP addresses were assigned to the ARN and PC.

B. Experimental results

Let d denote the distance between the ARN and PBS. To establish communication, UAVs hovered at a location where d = 10 m and at an altitude of 5 m, matching PBS height with the antenna surfaces facing each other. Under these conditions, the TG link could be reconstructed, including the beamforming direction. TCP throughput was measured every 1 s for 30 s using iPerf3. The ICN throughput was the mean value of the three measurements for three different file fetches. The same methodology was used as in Section V.

Figure 12 shows the results of network performance. As shown in Figure 12(a), the average TCP throughput was 891 Mbit/s (median value) and 735, 787, and 899 Mbit/s (in mean value) in the cases where d = 10, 20, and 30 m, respectively. Note that in the physical layer, the TG can support data transfer rates up to 1,925–4,620 Mbit/s, but as the devices only support Gigabit Ethernet (GbE), this causes a bottleneck. Figure 12(b) shows the variance of TCP throughput. The standard deviation decreases when d increases because the UAV moves vertically and horizontally (including roll and pitch), even if it is stably hovering in a fixed position. This movement affects the mmWave feature (i.e., straight radio propagation and directional beamforming), which can be relatively small for far distances of d.

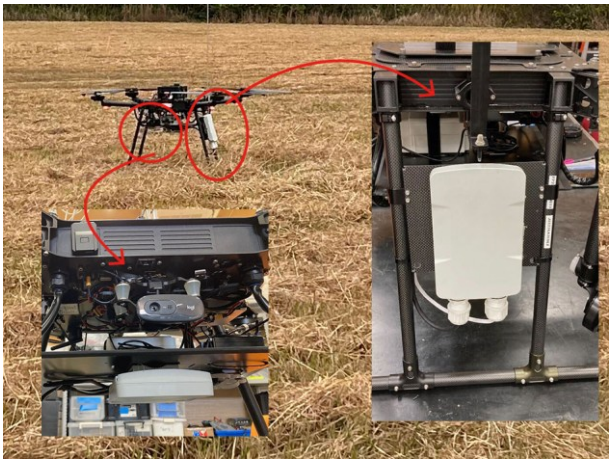


Figure 9. Overview of developed ARN device

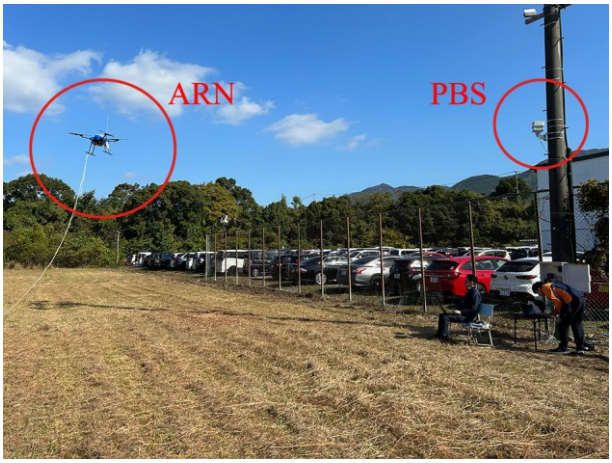


Figure 10. Field view of experimental site

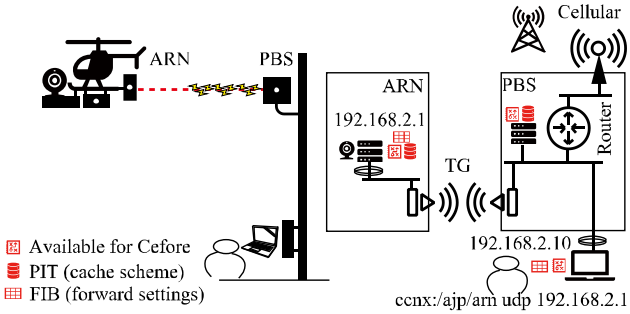


Figure 11. Network model of experimental site

As shown in Figures 12(c) and (d), the average ICN throughputs are 12.2, 13.0, and 14.6 Mbit/s, and the average jitters are 712, 669, and 583 µs for d = 10, 20, and 30 m, respectively. These results have the same characteristics as those of TCP evaluations shown in Figures 12(a) and (b). The ICN throughput is much lower than that of TCP because the

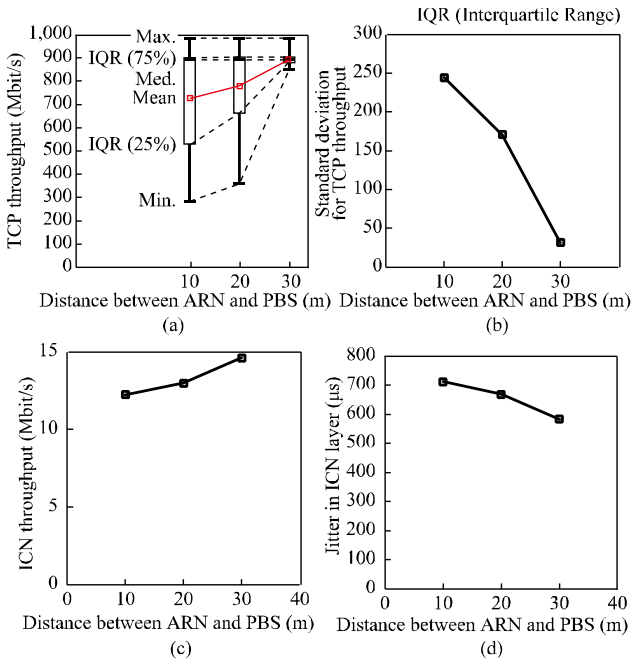


Figure 12. Experimental results of network performance: (a) TCP throughput. (b) Standard deviation for TCP throughput. (c) ICN throughput. (d) Jitter versus distance between nodes

latency causing mmWave propagations affects the ICN layer, and Cefore cannot optimally work, which is for wired LANs.

C. Demonstration of video-streaming application

To demonstrate information provisioning for disaster-stricken areas, the ARN performed live video broadcasting from the sky to the PC (connected to the ground PBS). Figure 13 shows a screenshot of the streaming video image as received by the PC. The ICN platform Cefore supports both stored file transmissions (including cached data) and real-time transmissions. On the basis of the literature [36], the ARN used the "cefputstream" command for live video broadcasting from the UAV-mounted camera, while the PC used the "cefgetstream" command to receive the video.

Obtaining a bird's-eye view of disaster areas from the sky is crucial, and seamless real-time reception is vital. Figure 13 shows screenshot of the PC during the demonstration, demonstrating this capability, although motion cannot be depicted in a static image. Some lag was observed during the video-streaming-delivery experiment between the air and ground in the test field and test devices, despite no such issues being present in the preliminary ground-based delivery experiments.

D. Development of SN testbed device for future deployment

As mentioned in the previous sections, the proposed ICWSN framework can be deployed in real cities as a smart-city-as-a-service platform. Toward this goal, we also implemented a prototype SN device not limited to a specific application service but designed on the basis of a reliable and

zero-touch design, as shown in Figure 14. Note that those who install or manage the system may not necessarily know the detailed system structures, i.e., it should be a self-organized mechanism when the SNs are placed in on-site fields. In addition, the device requires a commercial power source, which seems consistent with the outdoor environment assumed for the placement location. However, we believe this is not a serious problem because the node will be placed where the monitoring system is using a commercial power source for the central system. For example, in a water-gate control system in a disaster-resilient system and a plastic greenhouse in a smart agriculture system, the actuator and robots need a commercial supply that is sufficiently large compared with that for the SN device.

As shown in Figure 14(a), the developed SN device was designed to be waterproof since it would be placed in extreme outdoor environments, such as greenhouses and disaster-strike areas. As shown in Figure 14(b), the device adopts a zero-touch design; namely, all that is required is pressing the power button after connecting the device to a commercial power supply. Figure 14(c) shows the internal view of the device. We avoided installing mechanical structures, such as motor-driven systems and air-cooling fans to improve the system's reliability. The control computer used an EPC-S020, which was presented in Section IV. The other components were the power supply unit and internal network layer-2 switching hub. As shown in Figure 14(d), the sensors and other modules were mounted outside the case. The device can be equipped with sensor connections having serial connections, such as RS-232/485 (Modbus), making it a general-purpose design for compatibility and scalability. As shown in Figure 14(e), the inside and outside of the device can be connected via a waterproof connector.

VII. Conclusion

In this paper, we presented a development of test fields and test devices as part of an ecosystem of UAV-aided mmWave ICWSNs aimed at deploying smart cities. We illustrated the computer-calculation capabilities and fundamental characteristics in mmWaves for feasibility evaluations based on network performance. In addition, we demonstrated a wideband video-streaming application for distributing disaster-related information and illustrated a prototype SN device for smart-city deployment. In future work, we plan to construct stable mmWave networks in an actual city and deploy the proposed ecosystem on it.

Acknowledgment

This work was partly supported by NICT Japan, Grant No. JPJ012368C05601. We are grateful to Dr. Kenji Kanai for helpful discussions, and to Advantech Japan, BeMap, Haft, Panasonic, and TEAD for their help with the experiments.

References

[1] S. Mori, "MmWave UAV-assisted information-centric wireless sensor network for disaster-resilient smart cities: Preliminary evaluation and demonstration," *Proc. IARIA ICN 2024*, pp. 1–4, Barcelona, Spain, May 2024.

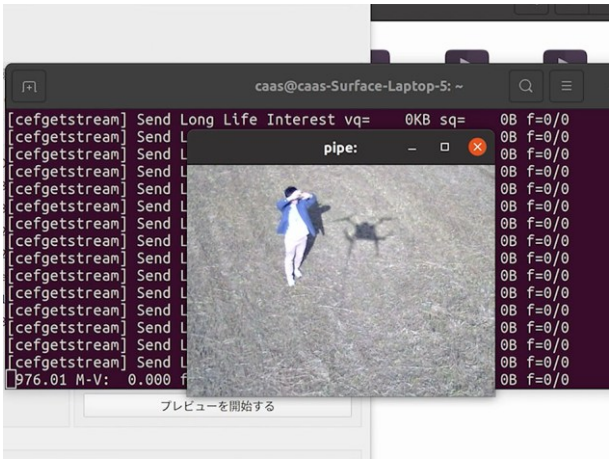
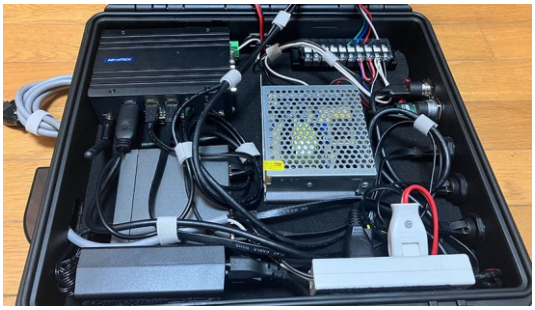


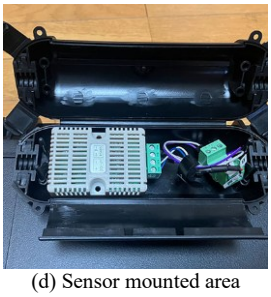
Figure 13. Demonstration of video-streaming application

(a) Top view of SN device

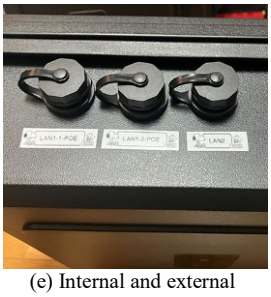
(b) Power-supply button



(c) Internal view of SN device



(d) Sensor mounted area



(e) Internal and external connector-connection area

Figure 14. Testbed device of SN device for smart-city deployment

[2] P. Mishra and G. Singh, "6G-IoT framework for sustainable smart city: Vision and challenges," *IEEE Consumer Electeric Mag.*, pp. 1–8, Aug. 2023.

[3] Q. T. Do, D. S. Lakew, A. T. Tran, D. T. Hua, and S. Cho, "A review on recent approaches in mmWave UAV-aided communication networks and open issues," *Proc. ICOIN 2023*, Bangkok, Thailand, Feb. 2023, pp. 728–731, doi: 10.1109/ ICOIN56518.2023.10049043.

[4] M. T. Dabiri, M. Hasna, N. Zorba, T. Khattab, and K. A. Qaraqe, "Enabling long mmWave aerial backhaul links via fixed-wing UAVs: Performance and design," *IEEE Trans. Communications*, vol. 71, no. 10, pp. 6146–6161, Oct. 2023.

[5] K. Aldubaikhy, W. Wu, N. Zhang, N. Cheng, and X. Shen, "MmWave IEEE 802.11 ay for 5G fixed wireless access," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 88–95, Apr. 2020.

[6] M. Cudak, A. Ghosh, and J. Andrews, "Integrated access and backhaul: A key enabler for 5G millimeter-wave deployments," *IEEE Communications Mag.*, vol. 59, no. 4, pp. 88–94, Apr. 2021.

[7] B. Ahlgren, C. Dannewitz, C. Imbrenda, D. Kutscher, and B. Ohlman, "A survey of information-centric networking," *IEEE Communications Mag.*, vol. 50, no. 7, pp. 26–36, July 2012.

[8] L. C. M. Hurali and A. P. Patil, "Application areas of information-centric networking: State-of-the-art and challenges," *IEEE Access*, vol. 10, pp. 122431–122446, Nov. 2022.

[9] P. K. Malik et al., "Smart cities monitoring using Internet of things: Opportunities and challenges," *Proc. ICESC 2023*, Coimbatore, India, July 2023, pp. 450–455, doi: 10.1109/ ICESC57686. 2023.10192958.

[10] M. Vera-Panez, K. Cuadros-Claro, M. Castillo-Cara, and L. Orozco-Barbosa, "BeeGOns!: A wireless sensor node for fog-computing in smart city applications," *IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 1, pp. 171–175, Jan. 2024.

[11] T. A. Ahanger, U. Tariq, A. Aldaej, A. Almehizia, and M. Bhatia, "IoT-inspired smart disaster evacuation framework," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 12885–12892, Apr. 2024.

[12] S. Pan and X. M. Zhang, "Cooperative gigabit content distribution with network coding for mmWave vehicular networks," *IEEE Trans. Mobile Computing*, vol. 23, no. 2, pp.1863–1877, Feb. 2024.

[13] V. R. Gannapathy, R. Nordin, N. F. Abdullah, and A. Abu-Samah, "A smart handover strategy for 5G mmWave dual connectivity networks," *IEEE Access*, vol. 11, pp. 134739–134759, Nov. 2023.

[14] X. Luo, X. Lu, B. Yin, and K. Yang, "Resource allocation for joint communication and positioning in mmWave ad-hoc networks," *IEEE Trans. Vehicular Technology*, vol. 73, no. 2, pp. 2187–2201, Feb. 2024.

[15] M. Zhang et al., "Will TCP work in mmWave 5G cellular networks?," *IEEE Communications Mag.*, vol. 57, no. 1, pp. 65–71, Jan. 2019.

[16] E. Khorov, A. Krasilov, M. Susloparov, and L. Kong, "Boosting TCP & QUIC performance in mmWave, terahertz, and lightwave wireless networks: A survey," *IEEE Communications Surveys and Tutorials*, vol. 25, no. 4, pp. 2862–2891, Aug. 2023.

[17] R. Poorzare and A. C. Auge, "How sufficient is TCP when deployed in 5G mmWave networks over the urban deployment?," *IEEE Access*, vol. 9, pp. 36342–36355, Mar. 2021.

[18] W. Yang et al., "A measurement study of TCP performance over 60 GHz mmWave hybrid networks," *Proc. IEEE WoWMoM 2022*, Belfast, United Kingdom, Aug. 2022, pp. 300–305, doi: 10.1109/WoWMoM54355.2022.00057.

[19] C. Wang, M. Pang, D. Zhong, Y. Cui, and W. Wang, "A mmWave communication testbed based on IEEE 802.11ad with scalable PtMP configuration," *China Communications*, vol. 19, no. 4, pp. 44–56, Apr. 2022.

[20] K. Aldubaikhy, W. Wu, N. Zhang, N. Cheng, and X. Shen, "MmWave IEEE 802.11ay for 5G fixed wireless access," *IEEE Wireless Commun.*, vol. 27, no. 2, pp. 88–95, Apr. 2020.

[21] Q. Tuan Do, D. Shumeye Lakew, A. Tien Tran, D. Thien Hua, and S. Cho, "A review on recent approaches in mmWave UAV-aided communication networks and open issues," *Proc. ICOIN 2023*, Bangkok, Thailand, Jan. 2023, pp. 728–731, doi: 10.1109/ICOIN56518.2023.10049043.

[22] M. T. Dabiri, M. Hasna, N. Zorba, T. Khattab, and K. A. Qaraqe, "Enabling long mmWave aerial backhaul links via fixed-wing UAVs: Performance and design," *IEEE Trans. Communications*, vol. 71, no. 10, pp. 6146–6161, Oct. 2023.

[23] S. G. Sanchez, S. Mohanti, D. Jaisinghani, and K. R. Chowdhury, "Millimeter-wave base stations in the sky: An experimental study of UAV-to-ground communications," *IEEE Trans. Mobile Computing*, vol. 21, no. 2, pp. 644–662, Feb. 2022.

[24] L. Cheng et al., "Modeling and simulation for UAV air-to-ground mmWave channels," Proc. EuCAP 2020, Copenhagen, Denmark, Mar. 2020, pp. 1–5, doi: 10.23919/EuCAP48036. 2020.9136077.

[25] Z. Xiao et al., "A survey on millimeter-wave beamforming enabled UAV communications and networking," *IEEE Communications Survey and Tutorials*, vol. 24, no. 1, pp. 557–610, Nov. 2022.

[26] J. Zhao, J. Liu, J. Jiang, and F. Gao, "Efficient deployment with geometric analysis for mmWave UAV communications," *IEEE Wireless Communications Letter*, vol. 9, no. 7, pp. 1115–1119, July 2020.

[27] B. Chang, W. Tang, X. Yan, X. Tong, and Z. Chen, "Integrated scheduling of sensing, communication, and control for mmWave/THz communications in cellular connected UAV networks," *IEEE J. Sel. Areas in Communications*, vol. 40, no. 7, pp. 2103–2113, July 2022.

[28] A. Nordrum, "Facebook pushes networking tech: The company's Terragraph technology will soon be available in commercial gear," *IEEE Spectrum*, vol. 56, no. 4, pp. 8–9, Apr. 2019.

[29] BeMap, https://www.bemap.co.jp/ (retrieved: Nov. 2024).

[30] Cefore, https://cefore.net/ (retrieved: Nov. 2024).

[31] S. Mori, "Information-centric wireless sensor networks for smart-city-as-a service: Concept proposal, testbed develop¬ment, and fundamental evaluation," *Proc. IEEE CCNC 2023*, Las Vegas, NV, USA, Jan. 2023, pp. 945–946, doi: 10.1109/CCNC51644.2023.10060577.

[32] S. Mori, "Energy-efficient cooperative caching scheme for green ICWSN: Preliminary analysis and testbed development," *Proc. ACM MobiCom 2023 WS NET4us*, Madrid, Spain, Oct. 2023, pp. 207–212, doi: 10.1145/3615991.3616406.

[33] S. Mori, "Test-field development for ICWSNs and preliminary evaluation for mmWave-band wireless communications," *Proc IEEE CCNC 2024*, Las Vegas, NV, USA, Jan. 2024, pp. 1–2, doi: 10.1109/CCNC51664.2024.10454799.

[34] S. Mori, "Prototype development of river velocimetry using visual particle image velocimetry for smart cities and disaster area networks," *Proc. ISCIT 2021*, Tottori, Japan, Oct. 2021, pp. 169–2171, doi: 10.1109/ISCIT52804.2021.9590602.

[35] Open SSL, https://www.openssl.org/ (retrieved: Nov. 2024).

[36] K. Matsuzono and H. Asaeda, "NMRTS: Content name-based mobile real-time streaming," *IEEE Communications Mag.*, vol. 54, no. 8, pp. 92–98, Aug. 2016.

Deep Reinforcement Learning Enabled Adaptive Virtual Machine Migration Control in Multi-Stage Information Processing Systems

Yukinobu Fukushima
Faculty of Environmental, Life, Natural Science and Technology
Okayama University
Okayama, Japan
fukusima@okayama-u.ac.jp

Yuki Koujitani
Graduate School of Natural Science and Technology
Okayama University
Okayama, Japan
ppxw4aek@s.okayama-u.ac.jp

Kazutoshi Nakane
Graduate School of Information Science
Nagoya University
Nagoya, Japan
nakane@net.itc.nagoya-u.ac.jp

Yuya Tarutani
Graduate School of Engineering
Osaka University
Osaka, Japan
tarutani@comm.eng.osaka-u.ac.jp

Celimuge Wu
Graduate School of Informatics and Engineering
The Univ. of Electro-Commun.
Tokyo, Japan
celimuge@uec.ac.jp

Yusheng Ji
Information Systems Architecture Research Division
National Institute of Informatics
Tokyo, Japan
kei@nii.ac.jp

Tokumi Yokohira
Faculty of Interdisciplinary Science and Engineering in Health Systems
Okayama University
Okayama, Japan
yokohira@okayama-u.ac.jp

Tutomu Murase
Graduate School of Information Science
Nagoya University
Nagoya, Japan
tom@itc.nagoya-u.ac.jp

***Abstract*—This paper tackles a Virtual Machine (VM) migration control problem to maximize the progress (accuracy) of information processing tasks in multi-stage information processing systems. The conventional methods for this problem are effective only for specific situations, such as when the system load is high. In this paper, in order to adaptively achieve high accuracy in various situations, we propose a VM migration method using a Deep Reinforcement Learning (DRL) algorithm. It is difficult to directly apply a DRL algorithm to the VM migration control problem because the size of the solution space of the problem dynamically changes according to the number of VMs staying in the system while the size of the agent's action space is fixed in DRL algorithms. To cope with this difficulty, the proposed method divides the VM migration control problem into two problems: the problem of determining only the VM distribution (i.e., the proportion of the number of VMs deployed on each edge server) and the problem of determining the locations of all the VMs so that it follows the determined VM distribution. The former problem is solved by a DRL algorithm, and the latter by a heuristic method. This approach makes it possible to apply a DRL algorithm to the VM migration control problem because the VM distribution is expressed by a vector with a fixed number of dimensions and can be directly outputted by the agent. The simulation results confirm that our proposed method can adaptively achieve quasi-optimal accuracy in various situations with different link delays, types of the information processing tasks and the number of VMs.**

***Keywords*-Multi-stage information processing system; VM migration control; Deep reinforcement learning; Deep Deterministic Policy Gradient (DDPG)**

I. Introduction

This paper is an extended and improved version of an earlier paper presented at the IARIA International Conference on Networks (ICN 2024) [1] in Barcelona, Spain.

In recent years, ultra-real-time services, such as Cross Reality (XR) and automated driving, are expected to appear. In these services, information processing tasks requested by clients need to be executed immediately (e.g., on the order of milliseconds) and the progress (accuracy) of the processing results should be as high as possible.

A multi-stage information processing system [2] [3] is one of the promising candidates for the edge computing infrastructures for ultra-real-time services. In the system, information processing tasks requested by clients are executed in parallel by an edge server and a data center. The edge server prioritizes responsiveness over accuracy; it returns the highly responsive but low accurate processing results to the clients while the data center prioritizes accuracy over responsiveness; it return the highly accurate but low responsive processing results to the clients. When operating ultra-real-time services in a multi-stage information processing system, it is important to maximize the accuracy of information processing tasks executed by the edge servers while satisfying the responsiveness requested by clients.

Previous researches on multi-stage information processing systems focused on improving the accuracy of information processing tasks executed by edge servers through Virtual Machine (VM) migration control [2] [3]. VM migration control dynamically migrates VMs, which execute the information processing tasks requested by clients on edge servers, among multiple edge servers, which leads to effective use of CPU resources on edge servers, appropriate adjustment of CPU times allocated to the tasks and reduction of the communication delay between clients and VMs, thereby improving the

accuracy of the tasks. In the previous researches, as heuristic methods for VM migration control, VM sweeping method [3], VM number averaging method [3], early-blooming type priority processing method [2], and late-blooming type priority processing method [2] were proposed and their effectiveness were confirmed. These methods are, however, effective only in specific situations, such as when the system load is high and the type of information processing tasks is the late-blooming type. Since the system load and the type of information processing tasks change dynamically, VM migration control that can adaptively achieve high accuracy in a wide variety of situations is needed.

In this paper, in order to adaptively achieve high accuracy in a variety of situations, we propose a VM migration method using a Deep Reinforcement Learning (DRL) algorithm. DRL algorithms are expected to adaptively achieve a quasi-optimal performance in a variety of situations through interactions between a learning agent and a dynamically changing environment. On the other hand, it is difficult to directly apply a DRL algorithm to the VM migration control problem because, in the problem, the size of the solution space dynamically changes according to the dynamic changes in the number of VMs staying in the system while the size of the agent's action space is fixed in DRL algorithms, and consequently it is difficult for the agent to directly output an solution for the problem. To cope with this difficulty, in this paper, we divide the VM migration control problem into two problems: the problem of determining only the VM distribution (i.e., the proportion of the number of VMs deployed on each edge server) and the problem of determining the locations of all the VMs so that it follows the determined VM distribution. The former problem is solved by a DRL algorithm, and the latter by a heuristic method. This approach makes it possible to apply a DRL algorithm with a fixed action space size to the VM migration control problem because the VM distribution is expressed by a vector with a fixed number of dimensions and can be directly outputted by the agent.

The rest of this paper is organized as follows. Section II introduces related work on VM migration control. Section III describes the multi-stage information processing system and the VM migration control problem. In Section IV, we propose a VM migration method using a DRL algorithm. In Section V, we evaluate the effectiveness of our proposed method with computer simulations. In Section VI, we summarize the paper.

II. Related Work

The previous researches in [4]–[11] tackle VM migration control problems in server migration services, and propose heuristic methods [4] [6], mathematical programming methods [5], [7]–[9], [11], and Q-learning methods [10]. These methods, however, aim at improving the communication quality between clients and VMs and reducing network power consumption, and do not consider the accuracy of information processing tasks.

The previous researches in [2] [3] tackle VM migration control problems in multi-stage information processing systems,

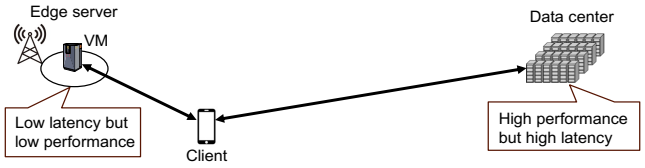


Figure 1. Multi-stage information processing systems.

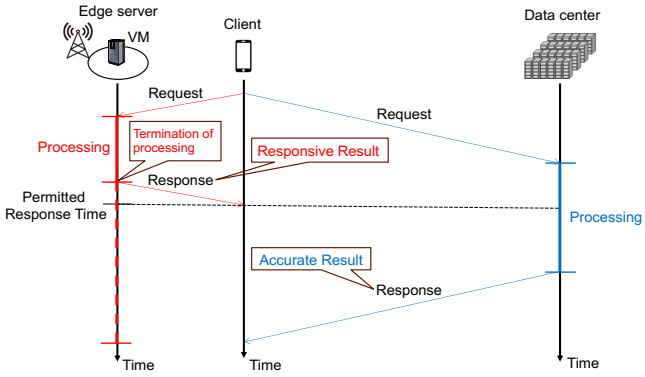


Figure 2. Flow of information processing in a multi-stage information processing system.

and propose the heuristic methods; VM sweeping method [3], VM number averaging method [3], early-blooming type priority processing method [2], and late-blooming type priority processing method [2]. These methods are, however, effective only in specific situations. For example, the VM sweeping method is shown to be effective only in situations where the system load is high and the type of information processing tasks is the late-blooming type. Since the system load and the type of information processing tasks change dynamically, VM migration control that can adaptively achieve high accuracy in a wide variety of situations is needed.

The previous researches in [12] [13] tackle VM migration control problems in mobile edge computing, and propose VM migration methods using Deep Q-Network (DQN) [14], which is a kind of DRL algorithms. These methods, however, can only be applied to VM migration control problems with a single VM because the size of an agent's action space is fixed in DQN, and cannot be applied to VM migration control problems with multiple VMs.

III. Multi-Stage Information Processing Systems

As shown in Figure 1, a multi-stage information processing system consists of edge servers located proximate (e.g., base stations) to clients and data centers located distant from them. The system provides clients with both highly responsive and highly accurate processing results by executing information processing tasks in parallel at the edge servers and the data centers.

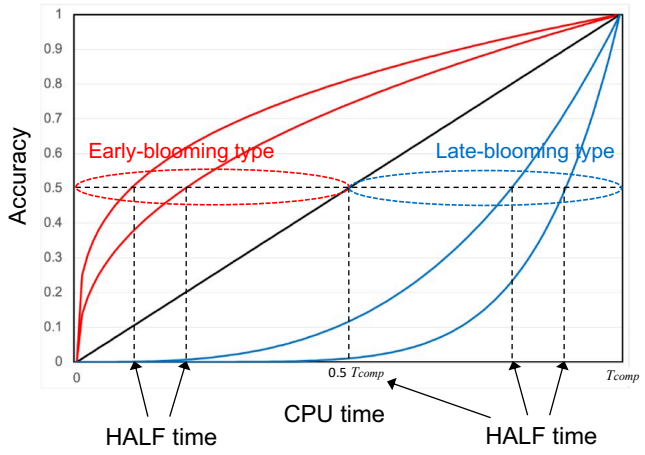


Figure 3. Relationship between CPU time allocated to a task and accuracy of the task.

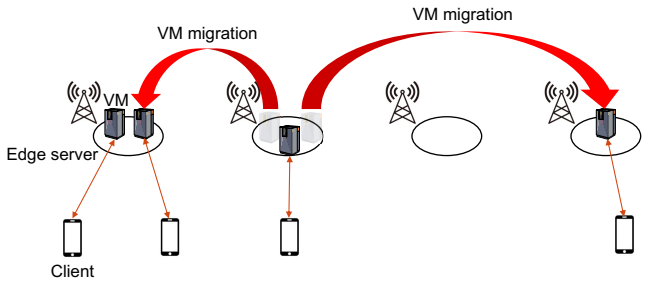


Figure 4. VM migration control in a multi-stage information processing system.

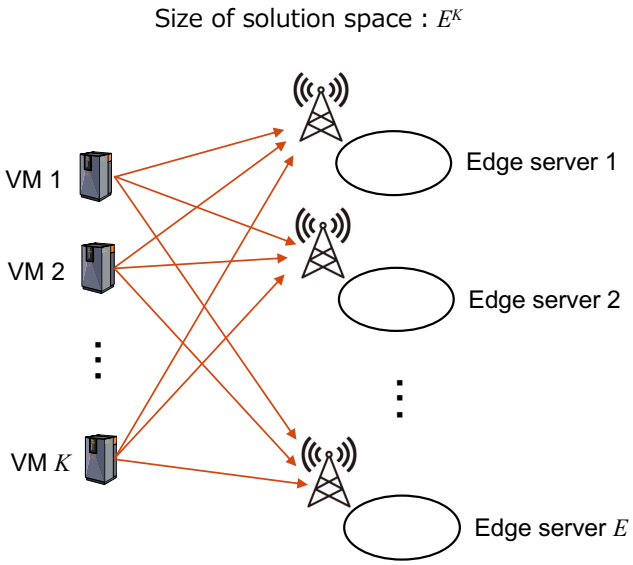


Figure 5. Size of solution space in VM migration control problem.

Figure 2 shows the flow of information processing in a multi-stage information processing system. A client requests both an edge server and a data center to process its task in parallel. When the response time permitted by the client approaches, the edge server terminates its processing to meet the permitted response time and returns the highly responsive processing result to the client. The data center, on the other hand, accomplishes its processing and returns the highly accurate processing result to the client.

In this paper, we adopt the accuracy model (i.e., the relationship between the CPU time (t_{CPU}) allocated to a task and the accuracy ($f(t_{CPU})$) of the task) in [3]. Figure 3 shows the accuracy model. In the model, the accuracy of the task is calculated as follows.

$$f(t_{CPU}) = \left(\frac{t_{CPU}}{T_{comp}}\right)^{\frac{log(0.5)}{log(\frac{HALFtime}{T_{comp}})}} \quad (1)$$

where T_{comp} represents the time for the task to be completed (i.e., accuracy reaches 1.0) and HALF time represents the time for the task to reach accuracy of 0.5. Tasks are classified based on their HALF time. The tasks with HALF time shorter than 0.5 T_{comp} are classified into early-blooming type, those with HALF time of 0.5 T_{comp} are classified into linear type, and those with HALF time longer than 0.5 T_{comp} are classified into late-blooming type.

In this paper, we tackle a VM migration control problem among multiple edge servers for maximizing the accuracy of information processing tasks executed by edge servers while satisfying the responsiveness requested by the clients (Figure 4). The objective of the problem is to maximize the sum of accuracies of all the information processing tasks. VM migration enables effective use of CPU resources on edge servers, appropriate adjustment of CPU times allocated to the tasks and reduction of the communication delay between clients and VMs, thereby improving the accuracy of the tasks. On the other hand, VM migration stops the execution of the tasks during the VM migration time, which may decrease the accuracy of the tasks. We need to carefully determine the locations of the VMs with consideration of the pros and cons of VM migration.

IV. Proposed Method

In this paper, in order to adaptively achieve high accuracy in a variety of situations, we propose a VM migration method using a Deep Reinforcement Learning (DRL) algorithm. In reinforcement learning, an agent learns policy (i.e., how to map a situation to an action) from interactions with an environment in discrete timesteps. At each timestep t, the agent observes state s_t of the environment, takes action a_t and receives reward r_t. The objective of the agent is to acquire the policy that maximizes the discounted cumulative reward $R_t = \sum_{i=t}^{T} \gamma^{i-t} r_i$ where $\gamma \in [0, 1]$ is the discount rate. We believe that DRL is promising for VM migration control because the agent can adaptively learn an appropriate policy in accordance with the dynamically changing environment.

With regard to applying a DRL algorithm to a VM migration control problem in multi-state information processing systems,

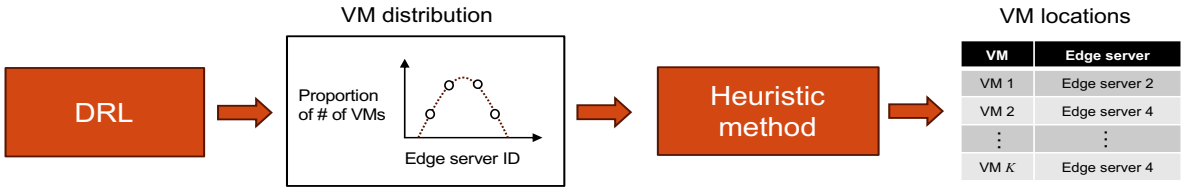


Figure 6. Outline of our proposed method.

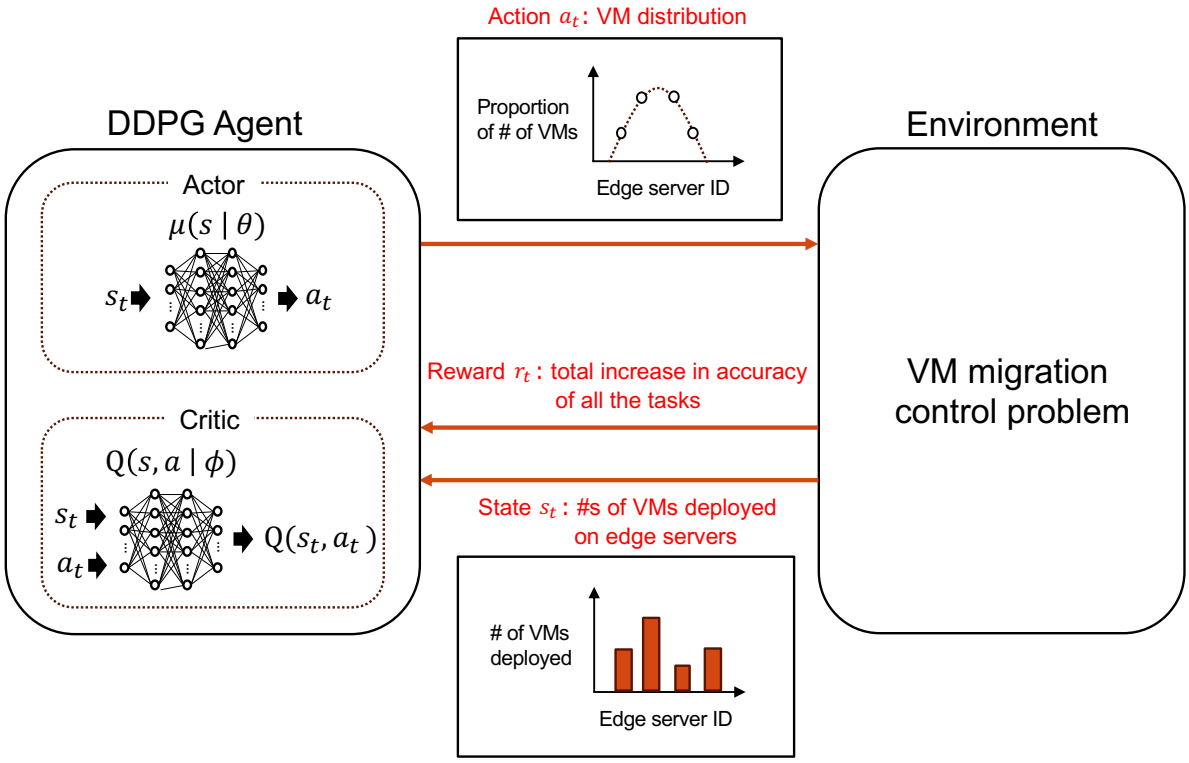


Figure 7. Interaction between the DDPG agent and the environment.

it should be noted that the size of the solution space (i.e., the total number of all possible solutions) of the problem dynamically changes according to the dynamic changes in the number of VMs staying in the system. As shown in Figure 5, the size of the solution space is E^K where E is the number of edge servers and K is the number of VMs, and the size of the solution space E^K dynamically changes according to the number of VMs K. On the other hand, the size of the agent's action space in DRL algorithms is fixed. For example, an agent in Deep Deterministic Policy Gradient (DDPG) [15] outputs a vector with a fixed number of dimensions. Therefore, it is difficult for an agent to directly output an solution for the VM migration control problem.

To cope with the dynamic change in the size of solution space, we divide the VM migration control problem into two problems (Figure 6): the problem of determining only the VM distribution (i.e., the proportion of the number of VMs deployed on each edge server) and the problem of determining the locations of all the VMs so that it follows the determined VM distribution. The former problem is solved by a DRL algorithm, and the latter problem is solved by a heuristic method. This approach makes it possible to apply a DRL algorithm with a fixed action space size to the VM migration control problem because the VM distribution is expressed by a vector with a fixed number of dimensions and can be directly outputted by an agent.

We adopt DDPG [15] as a DRL algorithm. DDPG approximates both a policy function $\mu(s|\theta)$ (Actor) and an action-value function $Q(s, a|\phi)$ (Critic) with deep neural networks. Actor $\mu(s|\theta)$ maps a given state to an action to be taken. Critic $Q(s, a|\phi)$ maps a given state-action pair to the expected value of the discounted cumulative reward if the action is taken under the state. During the training phase, Critic $Q(s, a|\phi)$ and Actor $\mu(s|\theta)$ are updated using experiences, which are

expressed with the tuple (s_t, a_t, r_t, s_{t+1}), obtained in interactions with the environment. As for Critic $Q(s, a|\phi)$, weights ϕ of $Q(s, a|\phi)$ are updated with a gradient decent method so that the following loss L is minimized:

$$L = \mathbb{E}\Big[(y_t - Q(s, a|\phi))^2\Big] \tag{2}$$

where $y_t = r_t + \gamma Q(s_{t+1}, \mu(s_{t+1}|\theta)|\phi)$. As for Actor $\mu(s|\theta)$, weights θ of $\mu(s|\theta)$ are updated with a gradient ascent method so that the performance (J) of the actor (i.e., expected value of the discounted cumulative reward) is maximized. In the gradient ascent method, the policy gradient $\nabla_\theta J$ is calculated by applying the chain rule to J with respect to weights θ as follows.

$$\begin{aligned}\nabla_\theta J &\approx \mathbb{E}\Big[\nabla_\theta Q(s, a|\phi)\Big] \\ &= \mathbb{E}\Big[\nabla_a Q(s, \mu(s|\theta)|\phi)\nabla_\theta \mu(s|\theta)\Big]\end{aligned} \tag{3}$$

In DDPG, Actor can output the VM distribution (i.e., the proportion of the number of VMs deployed on each edge server) as an action because it can operate over continuous action space. As well as DQN [14], DDPG adopts experience replay and target network techniques in order to train Actor and Critic in a stable and robust way.

Figure 7 depicts an interaction between a DDPG agent and an environment, which corresponds to the VM migration control problem. When applying a DRL algorithm to the VM migration control problem, we need to define action, state, and reward in accordance with the problem. Action a_t of the agent is defined as the VM distribution (i.e., the proportion of the number of VMs deployed on each edge server), and is expressed with the following equation.

$$a_t = (p_1, p_2, \ldots, p_E) \tag{4}$$

where p_i is the proportion of the number of VMs deployed on edge server i. State s_t of the environment is defined as the numbers of VMs deployed on edge servers for observing the load of each edge server, and is expressed with the following equation.

$$s_t = (d_1, d_2, \ldots, d_E) \tag{5}$$

where d_i is the number of VMs deployed on edge server i. Reward r_t is defined as the total increase in accuracy of all the tasks during the period from the last VM migration control to the current one.

Algorithm 1 in Figure 8 shows the procedure of our proposed method. In line 1, we generate Actor $\mu(s|\theta)$ and Critic $Q(s, a|\phi)$, and randomly initialize weights θ and ϕ. In lines 2 and 3, we generate the target networks of Actor and Critic, initialize their weights with those of Actor and Critic, and initialize replay buffer R, which stores a set of experiences for experience replay. The procedures in lines 4 to 17 and those in lines 7 to 16 are repeated for each episode and for each timestep of the episode, respectively. In lines 8 to 11, the agent selects action a_t (i.e., VM distribution), determines locations of all the VMs by the heuristic, observes reward r_t and next state s_{t+1}, and stores the obtained experience (s_t, a_t, r_t, s_{t+1}) to the replay buffer. Please note that a random noise $\mathcal{N}$ is added to the output by Actor for action exploration. In lines 12 to 15, we train Actor, Critic and target networks. In line 13, we update weights ϕ of Critic $Q(s, a|\phi)$ with a gradient descent method. Please note that we use target networks $Q'(s, a|\phi')$ and $\mu'(s|\theta')$ instead of Critic $Q(s, a|\phi)$ and Actor $\mu(s|\theta)$ for calculating target value y_i. In line 14, we update weights θ of Actor $\mu(s|\theta)$ with a gradient ascent method. In line 15, we update weights of target networks.

Algorithm 1 Procedure of our proposed method

1: Randomly initialize weights θ of Actor $\mu(s|\theta)$ and weights ϕ of Critic $Q(s, a|\phi)$
2: Initialize weights of Actor's target network $\mu'(s|\theta')$ and Critic's target network $Q'(s, a|\phi')$: $\theta' \leftarrow \theta$, $\phi' \leftarrow \phi$
3: Initialize replay buffer R
4: **for** episode = 1, M **do**
5: Initialize a random noise $\mathcal{N}$ for action exploration
6: Observe initial state s_1 from the environment
7: **for** t = 1, T **do**
8: Select VM distribution $a_t = \mu(s_t|\theta) + \mathcal{N}_t$ as action
9: Determine locations of all the VMs by the heuristic method among the VM locations that follow the determined VM distribution a_t, and migrates the VMs
10: Observe reward r_t and the next state s_{t+1}
11: Store experience (s_t, a_t, r_t, s_{t+1}) in R
12: Sample a random minibatch of N experiences (s_i, a_i, r_i, s_{i+1}) from R
13: Learning of Critic:
Calculate target value y_i:
$y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta')|\phi')$
Update weights ϕ with a gradient descent method so that loss $L = \frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\phi))^2$ is minimized
14: Learning of Actor:
Calculate policy gradient $\nabla_\theta J$:
$\nabla_\theta J \propto \frac{1}{N}\sum_i \nabla_a Q(s_i, \mu(s_i|\theta)|\phi)\nabla_\theta \mu(s_i|\theta)$
Update weights θ with a gradient ascent method so that performance of Actor J is maximized
15: Update weights of target networks:
$\theta' \leftarrow \tau\theta + (1-\tau)\theta'$
$\phi' \leftarrow \tau\phi + (1-\tau)\phi'$
16: **end for**
17: **end for**

Figure 8. Procedure of our proposed method.

After determining the VM distribution, we determine the locations of all the VMs by a heuristic method so that it follows the determined VM distribution. In this paper, we adopt a minimum client-VM delay method as the heuristic method. The minimum client-VM delay method selects the VM location with the minimum sum of the delays between clients and VMs in a brute force manner among the VM

locations that follow the VM distribution determined by the DDPG agent.

V. Performance Evaluation

In this section, we evaluate our proposed method with computer simulations. Section V.A explains the simulation model. Section V.B shows the evaluation results.

A. Simulation Model

We developed the VM migration control simulator and the DDPG agent with OpenAI Gym [16] and Keras-rl [17], respectively. Table I summarizes the parameter settings as to the DDPG agent. We adopt the same parameter values as those used by the DDPG agent in Keras-rl [17] because the previous research [15] reports that a DDPG agent with the same parameter setting successfully solved various physics tasks.

The left side of Figure 9 shows the network model. The network consist of four edge servers, which are connected in a full mesh manner. An edge server equally allocates its CPU time to all the VMs located on it. A VM is individually generated for each client, that is, the number of VMs is equal to the number of clients. We set the response time permitted by a client to 110 [ms] and the completion time of an information processing task (T_{comp}) to 110 [ms].

In order to evaluate whether our proposed method can adaptively cope with various situations, we change 1) link delay, 2) task type, and 3) total number of clients as follows.

1) link delay
   We assume that the delays of all the links are identical. We set the delay of each link to one of the following values: 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 [ms].
2) task type
   We assume that task type (i.e., HALF time) of all the information processing tasks are identical. We set HALF time of each task to either (1) 11 [ms] (= 0.1 $\times T_{comp}$) assuming the task type is the early-blooming type, (2) 55 [ms] (= 0.5 $\times T_{comp}$) assuming the task type is the linear type, or (3) 99 [ms] (= 0.9 $\times T_{comp}$) assuming the task type is the late-blooming type.
3) total number of clients
   We set the total number of clients that join the system (and the corresponding VMs) to either four or eight.

During an episode of the simulation, the following events occur (right side of Figure 9). When an episode starts, four or eight clients join the system in turn at time 0.1 [ms] with the interval of 0.1 [ms]. The locations of all the clients are fixed at edge server 1 during the episode. The initial locations of all the VMs are set to edge server 1. At time 3 [ms], we perform the first VM migration control. Then, at time 103 [ms], we perform the second VM migration control. Lastly, all the clients leave the system in turn at time 110.1 [ms] with the interval of 0.1 [ms]. The first VM migration control aims at determining the locations of the VMs during the episode and the second VM migration control aims at obtaining the reward and the experience for training the DDPG agent.

TABLE I
Parameter settings

Parameter	Value
Number of training episodes (M)	10,000
Discount rate (γ)	0.99
Number of hidden layers	Actor : 2, Critic : 5
Number of neurons in a hidden layer	Actor : 256, 256, Critic : 16, 32, 32, 256, 256
Activation function of hidden layers	Actor : relu, Critic : relu
Learning rate (α)	Actor : 0.001, Critic : 0.002
Noise process for action exploration ($\mathcal{N}$)	Ornstein-Uhlenbeck process
Size of replay buffer	10,000
Minibatch size (N)	64
Weights of updated parameters when updating the weights of target networks (τ)	0.005

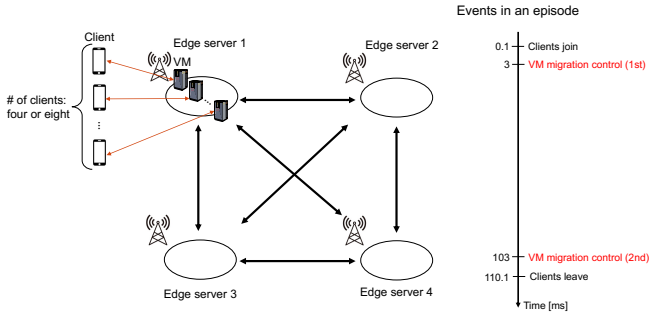


Figure 9. Network model and events in an episode.

We compare our proposed method with the following methods.

- VM sweeping method [3]
  It classifies all the edge servers into a congested edge server and working edge servers. On each of the working edge servers, a single VM with higher accuracy increase rate is individually deployed so that the VM can occupy the CPU time on the edge server. On the congested edge server, the remaining VMs are aggregated.
- VM number averaging method [3]
  It equally distributes all the VMs to all the edge servers for load balancing.
- Non-migration method
  It fixes all the VMs at their initial edge server (i.e., edge server 1).
- Minimum client-VM delay method
  It locates each of the VMs on the edge server most proximate to its client.

B. Evaluation Results

Figure 10 shows the average accuracy as a function of link delay for all the VM migration methods when the task type is the early-blooming type (HALF time = 11 [ms]) and the total number of clients (and the corresponding VMs) is four. The average accuracy of our proposed method (DDPG + Minimum client-VM delay method) is plotted with 95% confidence interval of 50 trials because it varies trial-by-trial

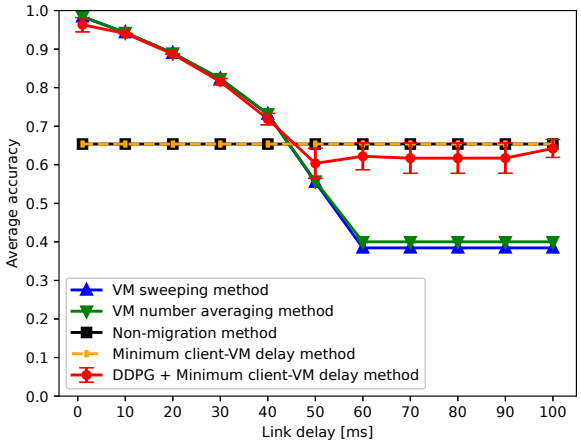


Figure 10. Average accuracy as a function of link delay (HALF time: 11 [ms], Total number of clients: 4).

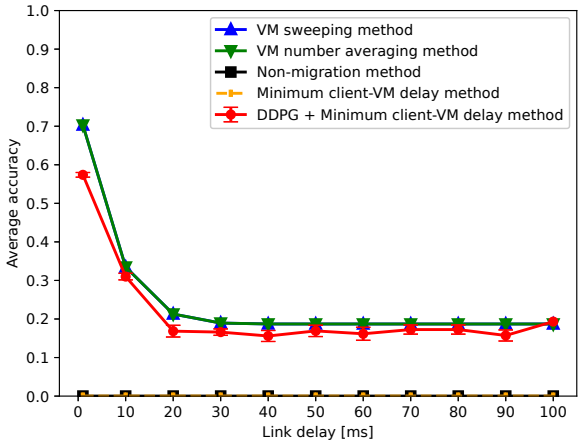


Figure 12. Average accuracy as a function of link delay (HALF time: 99 [ms], Total number of clients: 4).

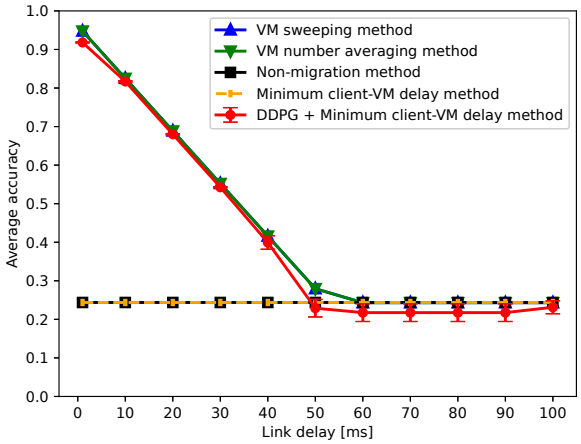


Figure 11. Average accuracy as a function of link delay (HALF time: 55 [ms], Total number of clients: 4).

depending on the initial weights of Actor and Critic, and the noises for action exploration.

Both non-migration method and minimum client-VM delay method show the constant accuracy of about 0.65 regardless of the link delay. This is because these methods always fix all the VMs at their initial edge server (edge server 1) regardless of the link delay. Both VM sweeping method and VM number averaging method achieve the maximum accuracy of about 0.98 when the link delay is 1 [ms], and the accuracy decreases as the link delay increases. This is explained as follows. These methods always distribute the VMs to all the edge servers so that a VM is individually located at an edge server regardless of the link delay. As the link delay increases, the VM migration time and the communication delay between the client and the VM increases, and consequently the CPU time allocated to the task decreases after completing the VM migration.

We compare the performances of non-migration method, minimum client-VM delay method, VM sweeping method, and VM number averaging method. When the link delay is shorter than or equal to 40 [ms], VM sweeping method and VM number averaging method achieve 12 to 50% higher accuracy than non-migration method and minimum client-VM delay method. Therefore, in this case, it is desirable to distribute all the VMs to different edge servers. When the link delay is longer than or equal to 50 [ms], non-migration method and minimum client-VM delay method achieve 17 to 70 % higher accuracy than VM sweeping method and VM number averaging method. Therefore, in this case, it is desirable to fix all the VMs at their initial edge server.

We focus on the performance of our proposed method. When the link delay is shorter than or equal to 40 [ms], our proposed method 1) achieves 10 to 49% higher accuracy than non-migration method and minimum client-VM delay method, and 2) achieves almost as high accuracy (at most 2% lower accuracy) as VM sweeping method and VM number averaging method, by successfully learning the policy that distributes all the VMs to all the edge servers similarly to VM sweeping method and VM number averaging method in most trials. When the link delay is longer than or equal to 50 [ms], our proposed method 1) achieves 9 to 68% higher accuracy than VM sweeping method and VM number averaging method, and 2) achieves almost as high accuracy (at most 6% lower accuracy) as non-migration method and minimum client-VM delay method, by successfully learning the policy that fixes all the VMs at their initial edge servers similarly to non-migration method and minimum client-VM delay method in most trials.

Figure 11 shows the average accuracy as a function of link delay for all the VM migration methods when the task type is changed to the linear type (HALF time = 55 [ms]) and the total number of clients is four. All of the conventional methods

select the same VM locations as those in Figure 10 because they determine VM locations without considering the task type; non-migration method and minimum client-VM delay method fix all the VMs at their initial locations while VM sweeping method and VM number averaging method distribute all the VMs to all the edge servers.

When the link delay is shorter than or equal to 50 [ms], VM sweeping method and VM number averaging method achieve 14 to 289% higher accuracy than non-migration method and minimum client-VM delay method because distributing the VMs to all the edge servers leads to more efficient use of CPU resources of all the edge servers thanks to short VM migration time. When the link delay is longer than or equal to 60 [ms], all the conventional methods shows the identical accuracy because only the VMs fixed at the initial edge server can execute the tasks due to long VM migration time, and those VMs achieve the identical accuracy in total regardless of their numbers for the linear model.

Our proposed method achieves almost as high accuracy (at most 16% lower accuracy) as VM sweeping method and VM number averaging method when the link delay is shorter than or equal to 50 [ms]. In most trials, our proposed method successfully learns the policy that distributes all the VMs to all the edge servers similarly to VM sweeping method and VM number averaging method. In addition, our proposed method achieves almost as high accuracy (at most 9% lower accuracy) as all the conventional methods when the link delay is longer than or equal to 60 [ms]. Although our proposed method learns various policies that determine different VM locations, most policies deploy at least one VM on the initial edge server, which leads to achieving the comparable accuracy as all the conventional methods.

Figure 12 shows the average accuracy as a function of link delay for all the VM migration methods when the task type is changed to the late-blooming type (HALF time = 99 [ms]) and the total number of clients is four. All of the conventional methods select the same VM locations as those in Figures 10 and 11.

Both non-migration method and minimum client-VM delay method show the accuracy close to zero regardless of link delay. This is because these methods fix all of the four VMs at the initial edge server and each of the VM is assigned only one fourth of the CPU time of the edge server, which is not enough for the late-blooming tasks to increase the accuracy. Both VM sweeping method and VM number averaging method achieve the accuracy of 0.70 when the link delay is 1 [ms], and the accuracy decreases as the link delay increases because the accuracies achieved by the three VMs distributed to edge servers 2, 3 and 4 decrease due to longer VM migration time. Our proposed method achieves almost as high accuracy (at most 18% lower accuracy) as VM sweeping method and VM number averaging method by successfully learning the policy that distributes all the VMs to all the edge servers similarly to VM sweeping method and VM number averaging method in most trials.

The average accuracy as a function of link delay for all the VM migration methods when the total number of clients is changed to eight are depicted in Figures 13, 14 and 15. Please note that VM sweeping method selects VM locations different from those by VM number averaging method; VM sweeping method distributes a single VM to each of edge servers 2, 3 and 4 and fixes the remaining five VMs at the initial edge server while VM number averaging method distributes two VMs to each of edge servers 2, 3 and 4 and fixes the remaining two VMs at the initial edge server. Non-migration method and minimum client-VM delay method fix all the VMs at their initial edge server.

In Figure 13 where the task type is the early-blooming type (HALF time = 11 [ms]), when the link delay is shorter than or equal to 20 [ms], VM number averaging method achieves higher accuracy than VM sweeping method because the VMs distributed to edge servers 2, 3 and 4 in the former method gain higher accuracy than the VMs fixed at the initial edge server in the latter method. For example, when the link delay is 1 [ms], all the VMs gain accuracy of about 0.80 in VM number averaging method while the five VMs fixed at the initial edge server gain only accuracy of about 0.61 and the three VMs distributed to edge servers 2, 3 and 4 gain accuracy of about 0.98 in VM sweeping method. When the link delay is longer than or equal to 40 [ms], VM sweeping method conversely achieves higher accuracy than VM number averaging method. This is explained as follows. As the link delay gets longer, the accuracy gained by the VMs distributed to edge servers 2, 3 and 4 decrease and the accuracy is dominated by those gained by the VMs fixed at the initial edge server. Because VM number averaging method fixes more VMs than VM number averaging method, the former method achieves higher accuracy than the latter.

Our proposed method achieves almost as high accuracy (at most 2% lower accuracy) as the best conventional methods by successfully learning the same policies as 1) VM number averaging method when the link delay is shorter than or equal to 30 [ms], 2) VM sweeping method when the link delay is 40 [ms], and 3) non-migration method and minimum client-VM delay method when the link delay is longer than or equal to 50 [ms], in most trials.

In Figure 14 where the task type is the linear type (HALF time = 55 [ms]), our proposed method also achieves almost as high accuracy (at most 14% lower accuracy) as the best conventional methods by successfully learning the same policies as them in most trials.

In Figure 15 where the task type is the late-blooming type (HALF time = 99 [ms]), VM sweeping method achieves higher accuracy than other conventional methods when the link delay is shorter than or equal to 10 [ms]. This is because only the VM that is individually deployed at an edge server and occupies the CPU time on it can obtain high accuracy for the late-blooming type tasks. In VM sweeping method, each of the three VMs distributed to edge servers 2, 3 and 4 can occupy the CPU time while in other conventional methods, no VM occupies the CPU time. In VM sweeping method, the accuracy decreases as the link delay increases due to longer

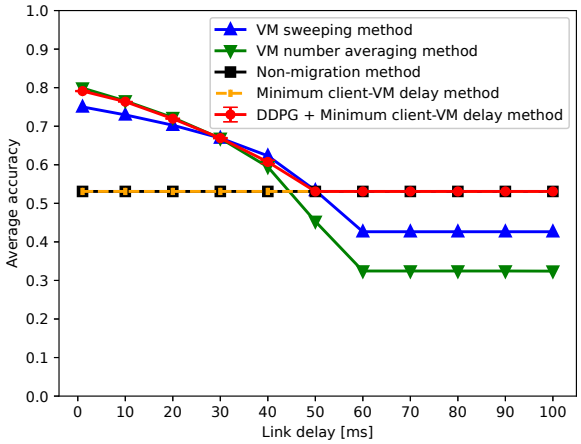


Figure 13. Average accuracy as a function of link delay (HALF time: 11 [ms], Total number of clients: 8).

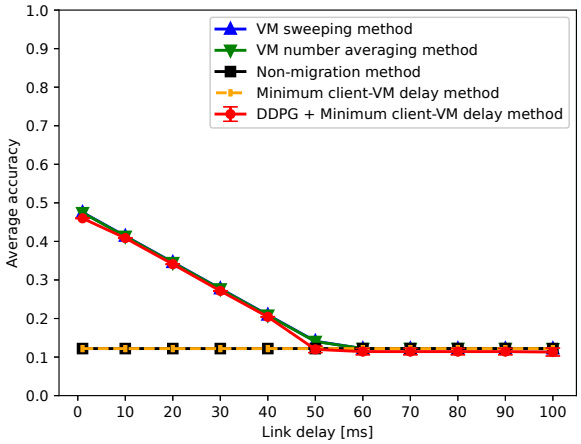


Figure 14. Average accuracy as a function of link delay (HALF time: 55 [ms], Total number of clients: 8).

VM migration time.

When the link delay is 1 [ms], our proposed method achieves almost as high accuracy (about 23% lower accuracy) as VM sweeping method by learning the policies that individually deploy a single VM on three of all the four edge servers in most trials. When the link delay is longer than or equal to 10 [ms], our proposed method achieves higher accuracy of about 0.12 by learning the policies that fix only a single VM at the initial edge server and making it occupy the CPU time of it without the VM migration time while all the conventional methods show the accuracy close to zero.

VI. Conclusions

In this paper, we proposed a VM migration method using a DRL algorithm in order to adaptively achieve high accuracy

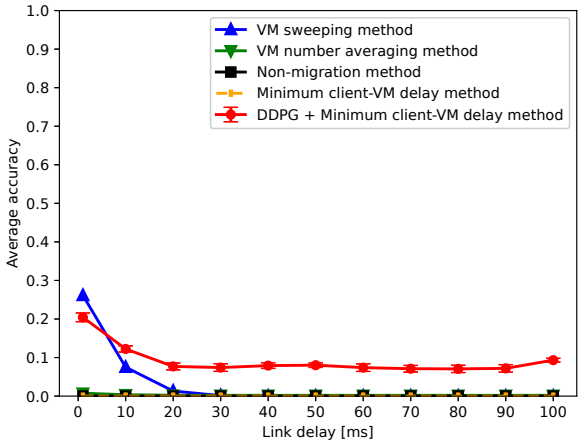


Figure 15. Average accuracy as a function of link delay (HALF time: 99 [ms], Total number of clients: 8).

of information processing tasks in various situations for multi-stage information processing systems. Our proposed method divides the VM migration control problem into two problems: the problem of determining only the VM distribution and the problem of determining the locations of all the VMs so that it follows the determined VM distribution. Our proposed method solves the former problem by a DRL algorithm and the latter problem by the minimum client-VM delay method. In order to evaluate whether our proposed method can adaptively cope with various situations, we performed simulation evaluations with different 1) link delays, 2) types of the tasks and 3) the number of VMs. The simulation results confirm that our proposed method can adaptively achieve quasi-optimal accuracy in those situations.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP23K11065 and JP24K02937.

References

[1] Y. Fukushima, Y. Koujitani, K. Nakane, Y. Tarutani, C. Wu, Y. Ji, T. Yokohira, and T. Murase, "Application of a Deep Reinforcement Learning Algorithm to Virtual Machine Migration Control in Multi-Stage Information Processing Systems," in *Proc. of ICN*, May 2024, pp. 13–18.

[2] K. Nakane, T. Anjiki, J. Xie, Y. Fukushima, and T. Murase, "VM Migration Considering Downtime for Accuracy Improvement in Multi-Stage Information Processing System," in *Proc. of IEEE ICCE*, Jan. 2022, pp. 335–336.

[3] T. Anjiki, K. Nakane, and T. Murase, "Performance Improvement by Controlling VM Migration between Edge Nodes in a Multi-Stage Information Processing System," in *Proc. of WSCE*, Sept. 2022, pp. 53–58.

[4] A. Yamanaka, Y. Fukushima, T. Murase, T. Yokohira, and T. Suda, "Destination selection algorithm in a server migration service," in *Proc. of CFI*, Sept. 2012, pp. 15–20.

[5] Y. Fukushima, T. Murase, T. Yokohira, and T. Suda, "Optimization of server locations in server migration service," in *Proc. of ICNS*, March 2013, pp. 200–206.

[6] Y. Hoshino, Y. Fukushima, T. Murase, T. Yokohira, and T. Suda, "An online algorithm to determine the location of the server in a server migration service," in *Proc. of IEEE CCNC*, Jan. 2015, pp. 740–745.
[7] Y. Fukushima, T. Murase, T. Yokohira, and T. Suda, "Power-Aware Server Location Decision in Server Migration Service," in *Proc. of ICTC*, pp. 150–155, Oct. 2016.
[8] Y. Fukushima, T. Murase, G. Motoyoshi, T. Yokohira, and T. Suda, "Determining Server Locations in Server Migration Service to Minimize Monetary Penalty of Dynamic Server Migration," Journal of Network and Systems Management, Vol. 26, Iss. 4, pp. 993–1033, Oct. 2018.
[9] Y. Fukushima, T. Murase, and T. Yokohira, "Link Capacity Provisioning and Server Location Decision in Server Migration Service," in *Proc. of IEEE CloudNet*, pp. 1–3, Oct. 2018.
[10] R. Urimoto, Y. Fukushima, Y. Tarutani, T. Murase, and T. Yokohira, "A Server Migration Method Using Q-Learning with Dimension Reduction in Edge Computing," in *Proc. of ICOIN*, pp. 301–304, Jan. 2021.
[11] Y. Fukushima, T. Suda, T. Murase, Y. Tarutani, and T. Yokohira, "Minimizing the Monetary Penalty and Energy Cost of Server Migration Service," Transactions on Emerging Telecommunications Technologies, Vol. 33, Iss. 9, pp. 1–34, Sept. 2022.
[12] D. Zeng, L. Gu, S. Pan, J. Cai, and S. Guo, "Resource Management at the Network Edge: A Deep Reinforcement Learning Approach," IEEE Network, Vol. 33, Iss. 3, pp. 26–33, May/June 2019.
[13] C. Zhang and Z. Zheng, "Task Migration for Mobile Edge Computing Using Deep Reinforcement Learning," Future Generation Computer Systems, Vol. 96, pp. 111–118, 2019.
[14] M. Volodymyr, et al. "Playing Atari with Deep Reinforcement Learning," arXiv preprint arXiv:1312.5602, 2013.
[15] T. P. Lillicrap, et al. "Continuous Control with Deep Reinforcement Learning," arXiv preprint arXiv:1509.02971, 2015.
[16] G. Brockman, et al. "OpenAI Gym," arXiv preprint arXiv:1606.01540, 2016.
[17] M. Plappert, "Keras-rl," GitHub repository, 2016.