International Journal on Advances in Networks and Services

2018 vol. 11 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 11, Number 3 & 4, 2018

Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Christos Bouras, University of Patras, Greece
Mahmoud Brahimi, University of Msila, Algeria
Marco Bruti, Telecom Italia Sparkle S.p.A., Italy
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Maria-Dolores Cano, Universidad Politécnica de Cartagena, Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Eduardo Cerqueira, Federal University of Para, Brazil
Bruno Chatras, Orange Labs, France
Marc Cheboldaeff, Deloitte Consulting GmbH, Germany
Kong Cheng, Vencore Labs, USA
Dickson Chiu, Dickson Computer Systems, Hong Kong
Andrzej Chydzinski, Silesian University of Technology, Poland
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Noelia Correia, University of the Algarve, Portugal
Noël Crespi, Institut Telecom, Telecom SudParis, France
Paulo da Fonseca Pinto, Universidade Nova de Lisboa, Portugal
Orhan Dagdeviren, International Computer Institute/Ege University, Turkey
Philip Davies, Bournemouth and Poole College / Bournemouth University, UK
Carlton Davis, École Polytechnique de Montréal, Canada
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
João Henrique de Souza Pereira, University of São Paulo, Brazil
Javier Del Ser, Tecnalia Research & Innovation, Spain
Behnam Dezfouli, Universiti Teknologi Malaysia (UTM), Malaysia
Daniela Dragomirescu, LAAS-CNRS, University of Toulouse, France
Jean-Michel Dricot, Université Libre de Bruxelles, Belgium
Wan Du, Nanyang Technological University (NTU), Singapore
Matthias Ehmann, Universität Bayreuth, Germany
Wael M El-Medany, University Of Bahrain, Bahrain
Imad H. Elhajj, American University of Beirut, Lebanon
Gledson Elias, Federal University of Paraíba, Brazil
Joshua Ellul, University of Malta, Malta
Rainer Falk, Siemens AG - Corporate Technology, Germany
Károly Farkas, Budapest University of Technology and Economics, Hungary
Huei-Wen Ferng, National Taiwan University of Science and Technology - Taipei, Taiwan
Gianluigi Ferrari, University of Parma, Italy
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Ulrich Flegel, HFT Stuttgart, Germany
Juan J. Flores, Universidad Michoacana, Mexico
Ingo Friese, Deutsche Telekom AG - Berlin, Germany
Sebastian Fudickar, University of Potsdam, Germany
Stefania Galizia, Innova S.p.A., Italy
Ivan Ganchev, University of Limerick, Ireland / University of Plovdiv “Paisii Hilendarski”, Bulgaria
Miguel Garcia, Universitat Politecnica de Valencia, Spain

Emiliano Garcia-Palacios, Queens University Belfast, UK
Marc Gilg, University of Haute-Alsace, France
Debasis Giri, Haldia Institute of Technology, India
Markus Goldstein, Kyushu University, Japan
Luis Gomes, Universidade Nova Lisboa, Portugal
Anahita Gouya, Solution Architect, France
Mohamed Graiet, Institut Supérieur d'Informatique et de Mathématique de Monastir, Tunisie
Christos Grecos, University of West of Scotland, UK
Vic Grout, Glyndwr University, UK
Yi Gu, Middle Tennessee State University, USA
Angela Guercio, Kent State University, USA
Xiang Gui, Massey University, New Zealand
Mina S. Guirguis, Texas State University - San Marcos, USA
Tibor Gyires, School of Information Technology, Illinois State University, USA
Keijo Haataja, University of Eastern Finland, Finland
Gerhard Hancke, Royal Holloway / University of London, UK
R. Hariprakash, Arulmigu Meenakshi Amman College of Engineering, Chennai, India
Go Hasegawa, Osaka University, Japan
Eva Hladká, CESNET & Masaryk University, Czech Republic
Hans-Joachim Hof, Munich University of Applied Sciences, Germany
Razib Iqbal, Amdocs, Canada
Abhaya Induruwa, Canterbury Christ Church University, UK
Muhammad Ismail, University of Waterloo, Canada
Vasanth Iyer, Florida International University, Miami, USA
Imad Jawhar, United Arab Emirates University, UAE
Aravind Kailas, University of North Carolina at Charlotte, USA
Mohamed Abd rabou Ahmed Kalil, Ilmenau University of Technology, Germany
Kyoung-Don Kang, State University of New York at Binghamton, USA
Sarfraz Khokhar, Cisco Systems Inc., USA
Vitaly Klyuev, University of Aizu, Japan
Jarkko Kneckt, Nokia Research Center, Finland
Dan Komosny, Brno University of Technology, Czech Republic
Ilker Korkmaz, Izmir University of Economics, Turkey
Tomas Koutny, University of West Bohemia, Czech Republic
Evangelos Kranakis, Carleton University - Ottawa, Canada
Lars Krueger, T-Systems International GmbH, Germany
Kae Hsiang Kwong, MIMOS Berhad, Malaysia
KP Lam, University of Keele, UK
Birger Lantow, University of Rostock, Germany
Hadi Larijani, Glasgow Caledonian Univ., UK
Annett Laube-Rosenpflanzer, Bern University of Applied Sciences, Switzerland
Gyu Myoung Lee, Institut Telecom, Telecom SudParis, France
Shiguo Lian, Orange Labs Beijing, China
Chiu-Kuo Liang, Chung Hua University, Hsinchu, Taiwan
Wei-Ming Lin, University of Texas at San Antonio, USA
David Lizcano, Universidad a Distancia de Madrid, Spain

Chengnian Long, Shanghai Jiao Tong University, China
Jonathan Loo, Middlesex University, UK
Pascal Lorenz, University of Haute Alsace, France
Albert A. Lysko, Council for Scientific and Industrial Research (CSIR), South Africa
Pavel Mach, Czech Technical University in Prague, Czech Republic
Elsa María Macías López, University of Las Palmas de Gran Canaria, Spain
Damien Magoni, University of Bordeaux, France
Ahmed Mahdy, Texas A&M University-Corpus Christi, USA
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Gianfranco Manes, University of Florence, Italy
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Moshe Timothy Masonta, Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa
Hamid Menouar, QU Wireless Innovations Center - Doha, Qatar
Guowang Miao, KTH, The Royal Institute of Technology, Sweden
Mohssen Mohammed, University of Cape Town, South Africa
Miklos Molnar, University Montpellier 2, France
Lorenzo Mossucca, Istituto Superiore Mario Boella, Italy
Jogesh K. Muppala, The Hong Kong University of Science and Technology, Hong Kong
Katsuhiro Naito, Mie University, Japan
Deok Hee Nam, Wilberforce University, USA
Sarmistha Neogy, Jadavpur University- Kolkata, India
Rui Neto Marinheiro, Instituto Universitário de Lisboa (ISCTE-IUL), Instituto de Telecomunicações, Portugal
David Newell, Bournemouth University - Bournemouth, UK
Ngoc Tu Nguyen, Missouri University of Science and Technology - Rolla, USA
Armando Nolasco Pinto, Universidade de Aveiro / Instituto de Telecomunicações, Portugal
Jason R.C. Nurse, University of Oxford, UK
Kazuya Odagiri, Yamaguchi University, Japan
Máirtín O'Droma, University of Limerick, Ireland
Jose Oscar Fajardo, University of the Basque Country, Spain
Constantin Paleologu, University Politehnica of Bucharest, Romania
Eleni Patouni, National & Kapodistrian University of Athens, Greece
Harry Perros, NC State University, USA
Miodrag Potkonjak, University of California - Los Angeles, USA
Yusnita Rahayu, Universiti Malaysia Pahang (UMP), Malaysia
Yenumula B. Reddy, Grambling State University, USA
Oliviero Riganelli, University of Milano Bicocca, Italy
Antonio Ruiz Martinez, University of Murcia, Spain
George S. Oreku, TIRDO / North West University, Tanzania/ South Africa
Sattar B. Sadkhan, Chairman of IEEE IRAQ Section, Iraq
Husnain Saeed, National University of Sciences & Technology (NUST), Pakistan
Addisson Salazar, Universidad Politecnica de Valencia, Spain
Sébastien Salva, University of Auvergne, France
Ioakeim Samaras, Aristotle University of Thessaloniki, Greece
Luz A. Sánchez-Gálvez, Benemérita Universidad Autónoma de Puebla, México
Teerapat Sanguankotchakorn, Asian Institute of Technology, Thailand
José Santa, University Centre of Defence at the Spanish Air Force Academy, Spain

Rajarshi Sanyal, Belgacom International Carrier Services, Belgium
Mohamad Sayed Hassan, Orange Labs, France
Thomas C. Schmidt, HAW Hamburg, Germany
Véronique Sebastien, University of Reunion Island, France
Jean-Pierre Seifert, Technische Universität Berlin & Telekom Innovation Laboratories, Germany
Dimitrios Serpanos, Univ. of Patras and ISI/RC ATHENA, Greece
Roman Y. Shtykh, Rakuten, Inc., Japan
Salman Ijaz Institute of Systems and Robotics, University of Algarve, Portugal
Adão Silva, University of Aveiro / Institute of Telecommunications, Portugal
Florian Skopik, AIT Austrian Institute of Technology, Austria
Karel Slavicek, Masaryk University, Czech Republic
Vahid Solouk, Urmia University of Technology, Iran
Peter Soreanu, ORT Braude College, Israel
Pedro Sousa, University of Minho, Portugal
Cristian Stanciu, University Politehnica of Bucharest, Romania
Vladimir Stantchev, SRH University Berlin, Germany
Radu Stoleru, Texas A&M University - College Station, USA
Lars Strand, Nofas, Norway
Stefan Strauβ, Austrian Academy of Sciences, Austria
Álvaro Suárez Sarmiento, University of Las Palmas de Gran Canaria, Spain
Masashi Sugano, School of Knowledge and Information Systems, Osaka Prefecture University, Japan
Young-Joo Suh, POSTECH (Pohang University of Science and Technology), Korea
Junzhao Sun, University of Oulu, Finland
David R. Surma, Indiana University South Bend, USA
Yongning Tang, School of Information Technology, Illinois State University, USA
Yoshiaki Taniguchi, Kindai University, Japan
Anel Tanovic, BH Telecom d.d. Sarajevo, Bosnia and Herzegovina
Rui Teng, Advanced Telecommunications Research Institute International, Japan
Olivier Terzo, Istituto Superiore Mario Boella - Torino, Italy
Tzu-Chieh Tsai, National Chengchi University, Taiwan
Samyr Vale, Federal University of Maranhão - UFMA, Brazil
Dario Vieira, EFREI, France
Lukas Vojtech, Czech Technical University in Prague, Czech Republic
Michael von Riegen, University of Hamburg, Germany
You-Chiun Wang, National Sun Yat-Sen University, Taiwan
Gary R. Weckman, Ohio University, USA
Chih-Yu Wen, National Chung Hsing University, Taichung, Taiwan
Michelle Wetterwald, HeNetBot, France
Feng Xia, Dalian University of Technology, China
Kaiping Xue, USTC - Hefei, China
Mark Yampolskiy, Vanderbilt University, USA
Dongfang Yang, National Research Council, Canada
Qimin Yang, Harvey Mudd College, USA
Beytullah Yildiz, TOBB Economics and Technology University, Turkey
Anastasiya Yurchyshyna, University of Geneva, Switzerland
Sergey Y. Yurish, IFSA, Spain

Jelena Zdravkovic, Stockholm University, Sweden
Yuanyuan Zeng, Wuhan University, China
Weiliang Zhao, Macquarie University, Australia
Wenbing Zhao, Cleveland State University, USA
Zibin Zheng, The Chinese University of Hong Kong, China
Yongxin Zhu, Shanghai Jiao Tong University, China
Zuqing Zhu, University of Science and Technology of China, China
Martin Zimmermann, University of Applied Sciences Offenburg, Germany

International Journal on Advances in Networks and Services
Volume 11, Numbers 3 & 4, 2018

CONTENTS

pages: 152 - 160

Tracking of Vehicles by Almost Everyone

Markus Ullmann, Federal Office for Information Security & University of Applied Sciences Bonn-Rhine-Sieg, Germany
Gerd Nolden, Federal Office for Information Security, Germany
Timo Hoss, Federal Office for Information Security, Germany

Time-Critical Data Delivery for Emergency Applications in Vehicle-to-Vehicle Communication

Saleem Raza*, Erik Neuhaus†, and Mesut Güneş*
*Communication and Networked Systems (ComSys), Faculty of Computer Science
Otto-von-Guericke University Magdeburg, Germany
saleem.raza, mesut.guenes{@ovgu.de}
†Communication and Networked Systems (ComSys), Faculty of Computer Science
Westfälische-Wilhelms University Münster, Germany
erikneuhaus@uni-muenster.de

***Abstract*—Transmission of time-critical messages in accident situations is of paramount importance for safety applications in VANET. These messages always require very low latency, which is an important metric for these applications. In particular, they impose real-time requirements. The MAC layer is an important place to satisfy multitude of performance metrics and can be greatly exploited to achieve low latency. In this paper, we exploit the existing VeMAC protocol and modify its TDMA frame structure to improve its performance for time-critical emergency traffic. We introduce additional emergency slots for transmission of emergency messages so that vehicles with time-critical emergency messages do not have to wait for their turn for transmission of such messages. The modified version of the VeMAC protocol results in improved performance for transmission of emergency traffic. The proposed protocol is evaluated through simulations. The results show great improvements and achieve lower latency in different scenarios.**

Index Terms*—*Vehicular Ad-hoc Network* (VANET); *Vehicle-to-Vehicle* (V2V) *Communication*; *Medium Access Control* (MAC); *Time-Critical Traffic*; *Emergency Applications

I. Introduction

The transmission of time-critical emergency traffic in Vehicle-to-Vehicle (V2V) network is imperative to safeguard safety of drivers and passengers, in this regard an emergency optimized Medium Access Control (MAC) protocol has been proposed in [1]. A Vehicular Ad-hoc Network (VANET) [2], [3] is a network of moving vehicles, where the vehicles, equipped with sufficient sensing, computation, and communication capabilities dynamically form an ad-hoc network without any mandatory infrastructure. The sensing, computation, and communication capabilities are housed into a unit referred to as On Board Unit (OBU). VANETs are a special class of Mobile Ad-hoc Networks (MANETs) [4], but having unique characteristics such as high mobility of nodes, dynamic network topology, varying communication environment, varying number of nodes, varying node distribution.

VANETs are designed for the purpose to exchange traffic or accidental information between Vehicle-to-Vehicle (V2V) and Vehicle-to-Road Side Unit (V2RSU) or Vehicle-to-Infrastructure (V2I) networks as shown in Figure 1.

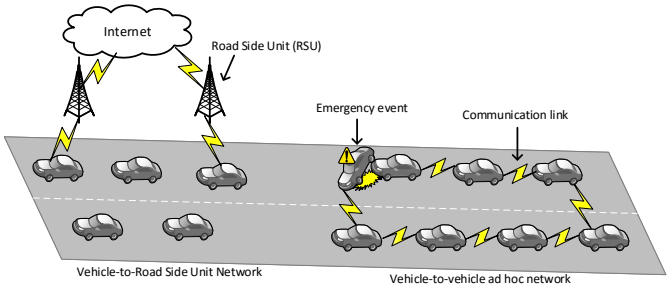


Figure 1. A Vehicular Ad-hoc Network (VANET) showing an emergency event caused by an accident between two vehicles.

The V2V allows the direct communication among vehicles through their OBUs, whereas the V2RSU involves vehicles to communicate with the RSU or vice versa. Generally, the RSUs are simply stationary network nodes that are mounted on traffic lights, street lights, road signs, etc. [5]. The cellular base stations, which are already prevalent can serve as RSUs and can be utilized to support V2RSU communication. An other option may be to use LTE base stations to support V2RSU communication [6].

VANETs have received tremendous attention due to plethora of applications they support such as *intelligent transportation system* (ITS), traffic information dissemination, infotainment, and the Internet connectivity on the go [7] [8]. Among these, the potential application of VANET is ITS, where the core objective is to control accidents, reduce traffic congestion, and improve driving safety in urban areas. Owing to importance of VANETs and the multitude of applications supported by the technology, several efforts were taken to standardize it. Federal Communications Commission (FCC) allocated $75\,\mathrm{MHz}$ spectrum in the $5.9\,\mathrm{GHz}$ band for Dedicated Short-Range Communication (DSRC) [11] solely for the purpose of V2V and V2RSU communication. DSRC is

Table I
LATENCY REQUIREMENTS FOR CERTAIN EMERGENCY SERVICES IN VANETS [9] [10].

Service	Latency requirement
Collision warning	$\leq 100\,\mathrm{ms}$
Pre-crash sensing	$\leq 20\,\mathrm{ms}$
Lane change warning	$\leq 100\,\mathrm{ms}$
Transit vehicle signal priority	$\leq 100\,\mathrm{ms}$

widely recognized as the IEEE 802.11p [12] Wireless Access in Vehicular Environments (WAVE) and is considered the *de-facto* standard for VANETs, it is based on IEEE 802.11 MAC and IEEE 802.11a Physical (PHY) layer [13].

The prime goal of VANETs is to disseminate safety and emergency messages, the timely transmission of such messages is critical to smooth operation of safety applications. The prominent examples of safety and emergency applications are primarily related to road accidents that cause loss of life of the drivers and passengers in vehicles. Other emergency related applications are intersection collisions warning, lane change assistance, overtaking vehicle warning, emergency vehicle warning, pre-crash warning, wrong way driving warning, signal violation warning, hazardous location warning, etc. [14].

In case of an emergency situation such as an accident as depicted in Figure 1, it is imperative to timely communicate such information to nearby vehicles so as to ensure safety of other nearby vehicles. But if such safety message and warning encounter longer delays, it becomes less effective to prevent such accidental situation for nearby vehicles. There is a life-time associated with the safety messages, which requires them to be transmitted timely otherwise they become ineffective. Table I depicts typical latency requirements for various emergency services in VANETs applications.

As latency is an important performance metric for safety/emergency applications and can be controlled through the Medium Access Control (MAC) layer so this requires for efficient medium sharing. Thus, an efficient MAC protocol should ensure high reliability, low end-to-end latency, and high throughput. Therefore, we analyze and exploit the MAC layer in reducing latency for safety/emergency messages in the context of V2V communication.

In this paper, we propose the emergency enhanced VeMAC (EEVeMAC) protocol, which is a variant of the VeMAC [15] protocol. VeMAC is a multichannel Time Division Multiple Access (TDMA) MAC protocol, which is based on ADHOC MAC [16]. The EEVeMAC protocol modifies the slotframe structure of VeMAC and uses emergency slots to transmit time-critical emergency messages in case of road accidents or collisions among vehicles in VANETs. In this way, it targets to meet the real-time requirements of the safety applications. The EEVeMAC achieves low latency for emergency messages under different scenarios and is evaluated by simulation. With reference to our earlier work [1], the main contribution of this work are as follows:

- We extend the EEVeMAC superframe structure to four emergency slots for the transmission of time-critical messages.
- We extensively evaluate the EEVeMAC by simulations with two emergency slots as well as with four emergency slots.
- We demonstrate the protocol in the dense urban scenario, and show how the addition of more emergency slots impact the protocol behavior and latency.
- We analyze the effect of adding more slots on collision rate and compare it with collision rate of two emergency slots.

The remainder of the paper proceeds as follows. In Section II, we give a background overview VeMAC, its working principle, and frame structure. It talks about possible collisions that can occur and explains slot divisions into disjunct sets. We highlight the drawbacks of VeMAC for low latency aspects. Subsequently, Section III gives overview of the desired changes in VeMAC to achieve low latency for transmission of time-critical messages. In Section IV, we describe the evaluation details of our proposed MAC protocol through simulation. We discuss different real life scenarios for which the protocol is evaluated. We also present details of the simulation environment and the associated components that were used to conduct simulations. Section V discusses the results of simulation and shows latency improvements through box plots. The results are presented for various scenarios considered. Finally, a conclusion is drawn in Section VI.

II. BACKGROUND

In this section, we present background details of the VeMAC protocol. We thoroughly explain the VeMAC, its frame structure, and working mechanism.

VeMAC frame structure: VeMAC (no abbreviation) [15] is a multi-channel TDMA protocol for VANETs, which utilizes two radios. One of the radios is always tuned to the control channel c_0, while the other radio can be tuned to one of the service channels. Each node should acquire exactly one slot on the control channel. The node holds onto this slot until it does not need it anymore or until a merging collision occurs. These collisions occur if two nodes, with the same slot, enter the same two-hop-neighborhood due to their mobility. To reduce the number of collisions, the slots are divided into disjunct sets L, R, and F as shown in Figure 2. The frame structure is split in two disjunct sets based on the general direction of movement of the vehicles. If a node travels in general eastern direction, so $0 - 180^\circ$ degrees of a compass

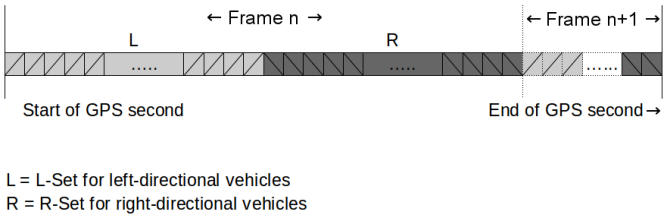


Figure 2. Frame structure of VeMAC protocol [15].

as shown in Figure 3, it would be in the R-subset (colored in dark grey in Figure 2), the rest in the L-subset (colored in light gray in Figure 2). F is an optional set for RSUs, which has no direction of movement. That way, vehicles driving in opposing directions are not competing for the same slot and it reduces the relative speed of nodes competing for the same slots and thereby increases the network topology persistence within these sets.

The directions are provided by the GPS unit that each vehicle is mandatory to be equipped with. With the GPS unit, it is possible to synchronize the frames through the pulse per second (PPS) signal provided by each GPS receiver. A frame should start at the beginning of each GPS second.

The VeMAC protocol proposes a time division in a periodical frame structure of fixed duration. One frame consists of 100 slots, where the length of one slot is of $1\,\mathrm{ms}$ duration, hence a frame length of $100\,\mathrm{ms}$. Each node should transmit periodically one message per frame in its allocated slot. The message consists of a header field, two fields to organize the allocation of slots on the service channels as well as one field for exchange of information for high-priority short applications.

Each node should have a unique random ID to identify the node. The header of the message of node x includes, amongst others, the set $N(x)$, which is the set of IDs of the one-hop neighbors of node x on channel c_0, from which node x has received packets on channel c_0 [15] in the previous 100 slots.

With the sets $N(y)$ of each one-hop neighbor y, the node is able to determine which slots are used by its two-hop neighborhood. These slots, that the node must not use in the next 100 time slots, are denoted by $T_0(x)$. With this information, the node builds the set of available slots $A(x) = \overline{T_0(x)}$ respectively with regard to the directional division, e.g., $A(x) = \overline{T_0(x)} \cap R$ for vehicles driving in eastern direction. These sets are the respective complementary set to $T_0(x)$, a node can use any slot that is not explicitly marked as used by its two-hop neighborhood. With the provided information, the node is able to solve the hidden-terminal problem.

Node x also determines whether or not all of its one-hop neighbors received its last broadcast by looking for its ID in the right slot in all $N(y)$. It thereby constitutes a reliable broadcast mechanism. Due to the regular transmitting, there exists an upper bound for transmission of messages of $100\,\mathrm{ms}$. However, $100\,\mathrm{ms}$ is a long time in high mobility scenarios.

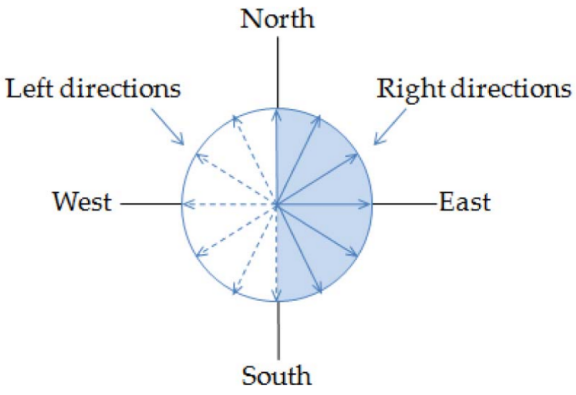


Figure 3. Division of node per direction [15] showing the distinction for the L set and R set. All vehicles driving in $0 - 180°$ degrees of a compass, will be in the R-subset for right directions, the rest in the L-subset for left directions.

Limitation of VeMAC for emergency messages: In $100\,\mathrm{ms}$, a car traveling on the highway with the recommended speed of $130\,\mathrm{km/h}$ already covers a distance of $3.6\,\mathrm{meter}$ and many cars drive considerably faster on the highway in Germany. While the $100\,\mathrm{ms}$ limit should be sufficient in normal use, it might be too long for emergency situations where fast responses are crucial.

III. Emergency Enhanced VeMAC (EEVeMAC) protocol

To reduce the latency in emergency situations, in this paper, we propose Emergency Enhanced VeMAC (EEVeMAC), which is the variant of the VeMAC protocol, by introducing emergency slots (colored in orange) at the beginning of the L set in slot 0 and R set in slot 50 as shown in Figure 4. They are evenly distributed across the frame structure to reduce average distance to any other slot. The slots are based on the principle of Carrier Sense Multiple Access (CSMA) for the transmission of time-critical emergency data.

In case of an emergency, a vehicle wants to send time-critical data to notify other vehicles of its situation. In this way, instead of waiting for its next allocated slot, the vehicle can use these additional emergency slots to quickly transmit the messages and avoid catastrophic situations. With additional slots, vehicles have three possible slots instead of one to transmit their data during emergency situations, effectively bringing down the upper bound latency to $50\,\mathrm{ms}$. While the upper bound latency is $50\,\mathrm{ms}$, the median average is further reduced since a slot is able to choose from three possible slots for emergency transmission instead of one.

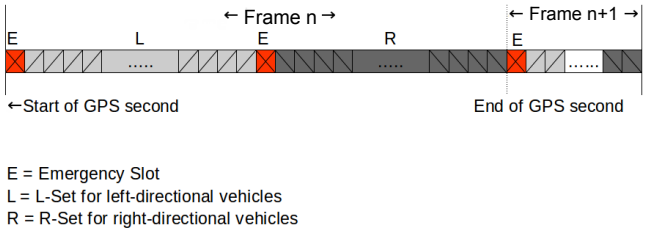


Figure 4. Frame structure of VeMAC with emergency slots. The emergency slots are set at the beginning of the L set respectively the R set.

While the original VeMAC protocol does not define the exact nature of $N(x)$ for node x, we implemented them in both VeMAC and EEVeMAC as pair of ID and slot number to preserve the reliable broadcast mechanism. Through this modification, an ID can be twice in a set. A receiving node then thereby acknowledges the reception of an emergency message by including the ID of the sending node in the emergency slot number in which it received the emergency message. This implementation decision will extend the length of the regular message by a maximum of 100 bytes (88 bits total, 7 bits for representation of numbers up to 128, rounded up to 8, multiplied by 100 slots).

IV. Implementation and Performance Evaluation

The EEVeMAC protocol is evaluated through simulation in OMNeT++ [17] simulation environment together with Veins [18] and SUMO [19]. Veins is an open source simulation framework for vehicular network simulation. It bi-directionally couples two softwares: OMNeT++ is utilized for network simulation and the open source traffic suite SUMO of the German Aerospace Center provides the traffic simulation data. SUMO has several car-following-models and lane-changing-models to reproduce realistic traffic behavior. Veins integrates MiXiM [20] for modeling physical layer effects and provides realistic interference models. For our simulation we use the two-ray-interference model provided by Veins [21]. The simulation parameters are listed in Table II whereas the scenario parameters are given in Table III.

Scenarios:

Two scenarios "straight" and "interchange" with reduced road traffic and normal road traffic were tested to examine the influence of node numbers on collisions.

Straight and interchange scenarios:

In the straight scenario, only traffic from northern and southern directions was present; in the interchange scenario vehicles started from each direction. The highway interchange Münster south, Germany was created in SUMO as shown in Figure 5 and provided with traffic statistic of the state office for road construction NRW [22] to achieve a realistic traffic

Table II
Simulation Parameters

Layer	Parameters	Value
APP	Field size	6000 x 6000 m
	Network	Dynamic
	Number of nodes	Varying number
	Distance between nodes	varying
	Data rate	18 Mbps
PHY	Radio tx output power	20 mW
	Propagation model	Two-ray model
	Frequency spectrum	2.4 GHz
	Sensitivity	−89 dB m
	Thermal noise	−110 dB m
MAC	Frame duration	100 ms
	Slot duration	1 ms

Table III
Configuration Parameters in the Two Scenarios

Parameters	Value	
Scenario	Straight	Interchange
Traffic flow from direction	North/South	North/East/South/West
vehiclesPerHour (total value)	4645	9476
Use of Emergencyslots	False/True	False/True
Emergency in slot	1/25/49	1/25/49
Replications	50	50
Simulation duration	80 sec	80 sec

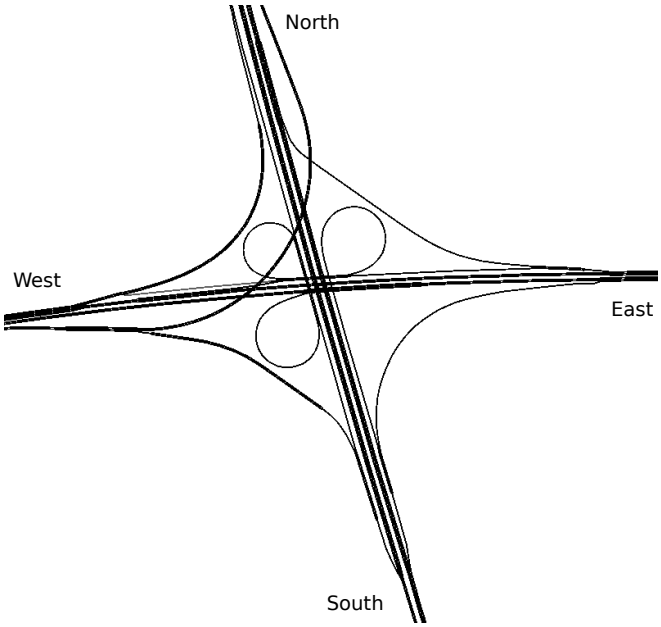


Figure 5. Interchange Münster south, the car with the emergency tries to travel from south direction to west direction and breaks down in the clover interchange line.

scenario. To get two different but still realistic scenarios from the traffic statistic, the crossing traffic from eastern/western directions was left out in the straight scenario. Hence, only traffic from northern and southern directions was present in the straight scenario. In the interchange scenario vehicles started from each direction. In each scenario, 20% of cars were presumed to change from one highway to the other highway with 10% in each direction of the highway. The road traffic was implemented with the traffic flow functionality of SUMO, which regularly introduces vehicles based on the number of vehicles per hour. The scenario consists of a car that drives on the highway in northern direction and wants to change the highway in western direction on the interchange. It breaks down on the clover interchange lane and sends an emergency message. The car drove in north-west direction and hence it has a regular slot in the first half of the frame structure. In each scenario, the emergency was set to three different slots. To slot 1, directly after an emergency slot, to slot 25, in the middle between two emergency slots, and to slot 49, right before an emergency slot. Each configuration was run with 50 repetitions to achieve a good confidence interval.

In addition to the aforementioned scenarios, we conducted a scenario "dense traffic" with additional cars to simulate extremely dense traffic as it would be expected in urban traffic. We conducted it with the same parameters as the "Interchange" scenario, but increased the numbers of vehicles to 13 600 vehicles per hour.

V. Results

In this section, we present the results of two different evaluations of the protocol. First, we discuss the results obtained from using two emergency slots and then we show the results with four emergency slots.

A. Results with two emergency slots

For the evaluation of EEVeMAC protocol, we measured two values. The latency from the moment the emergency occurred to the moment the one-hop neighbors receiving the emergency message. The second evaluation value consists of the occurrence of collisions, which were calculated to the arithmetic average per node. The results showed an overall improvement of the latency as further explained below.

1) Latency in straight scenario:
In the straight scenario, there were 21 nodes in transmitting range of the emergency vehicle at the moment of the emergency situation. The emergency message took a median time of 69.99 ms to reach the one-hop neighbors of the emergency vehicle in the original VeMAC. With EEVeMAC, with the addition of emergency slots, this value was reduced to 16.57 ms as shown in Figure 6. If the emergency occurred

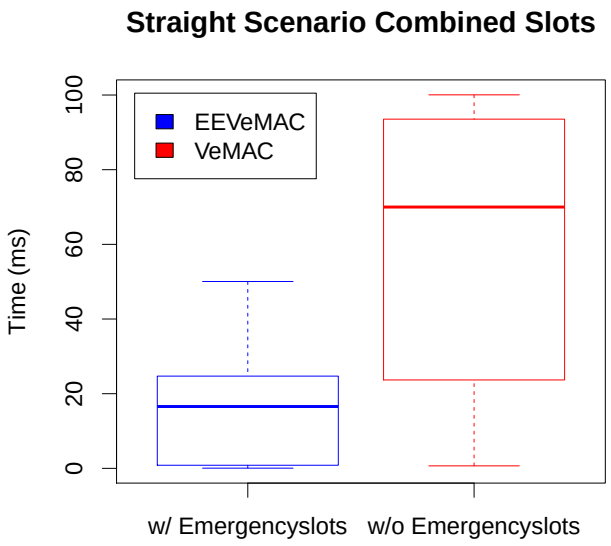
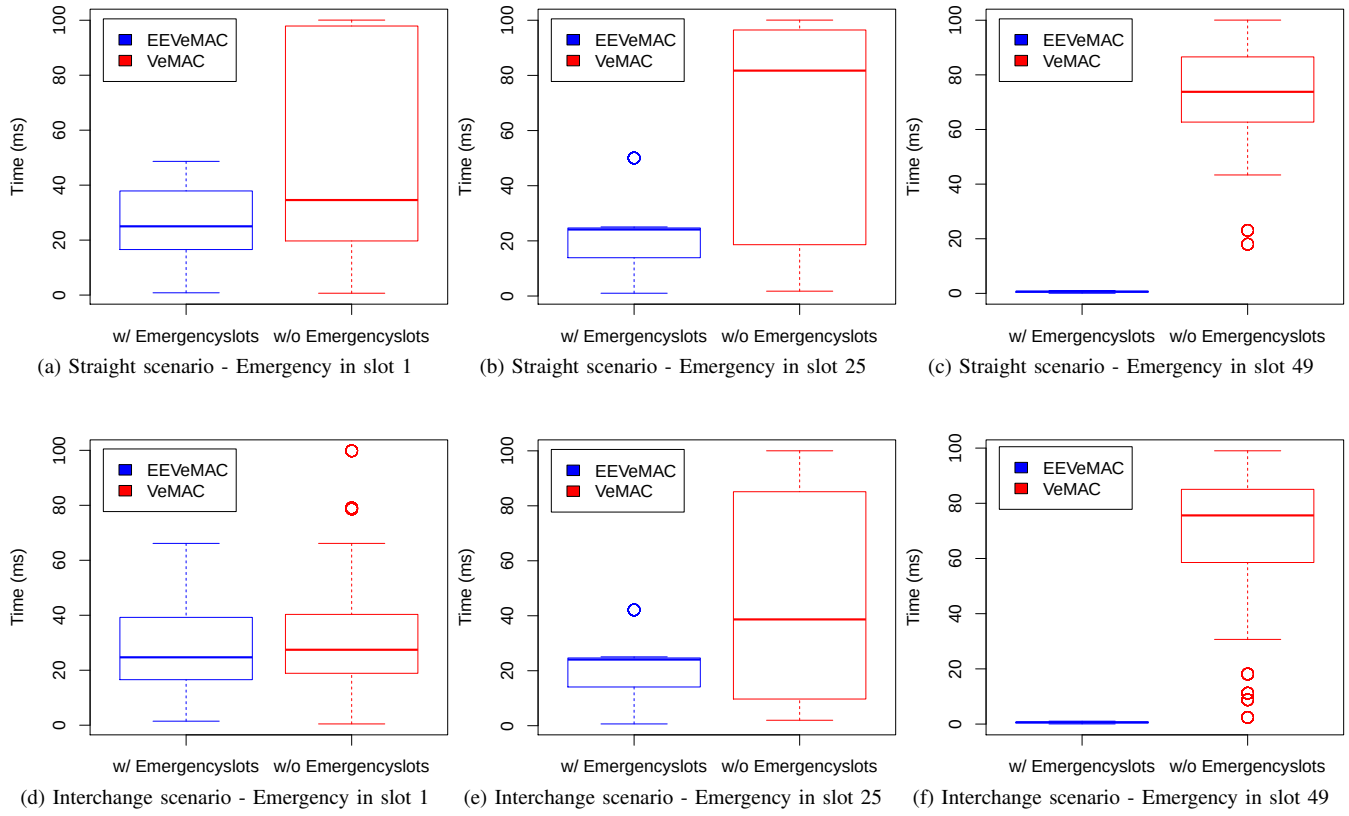


Figure 6. Evaluation results of straight scenario: On the left the latency results of EEVeMAC with emergency slots (w/Emergencyslots), on the right the latency results of original VeMAC without emergency slots (w/oEmergencyslots).

in the first slot after an emergency slot, the median latency was closest to the original protocol with 34.58 ms (VeMAC) vs. 25.01 ms (EEVeMAC) as depicted in Figure 7 (a), since there is a good chance that the regular slot of the emergency vehicle is between the slot in which the emergency occurs and the next emergency slot.

If there is a regular slot in between the emergency and an emergency slot, there is no difference between both protocols as they would both transmit the emergency message in the regular slot. The improvement occurs in the cases where the emergency slot is used. The biggest improvement could be measured with the emergency in slot 49, directly in front of an emergency slot with 73.79 ms (VeMAC) vs. 0.61 ms (EEVeMAC) as shown in Figure 7 (c). Without the emergency slots, the emergency vehicle has to wait at least 50 ms if it does not have slot 49 as its regular slot. It can not transmit in the slot numbers 50-99 since the emergency vehicle is driving in north western direction and hence prefers a slot in the L-set in slot numbers 0-49 of the frame. With the emergency right between two emergency slots, the median latency was improved by 57.61 ms from 81.73 ms in the original VeMAC to 24.12 ms in EEVeMAC with emergency slots as depicted in Figure 7 (b).

2) Latency in interchange scenario: The results of the interchange scenario with traffic flow from each direction showed similar improvements as shown in Figure 8. In this scenario, 35 nodes were present in transmitting range during the emergency situation. The median latency was



(a) Straight scenario - Emergency in slot 1 (b) Straight scenario - Emergency in slot 25 (c) Straight scenario - Emergency in slot 49

(d) Interchange scenario - Emergency in slot 1 (e) Interchange scenario - Emergency in slot 25 (f) Interchange scenario - Emergency in slot 49

Figure 7. Overview of evaluation of latency results for the straight and the interchange scenarios with the emergency in slots 1, 25, and 49. The small red circles in the figures indicate outliers.

improved by factor 3 from 48.73 ms (VeMAC) to 14.66 ms (EEVeMAC). The biggest improvement could be once again measured if the emergency occurred in the slot right before an emergency slot 75.61 ms in the original VeMAC vs. 0.61 ms in the EEVeMAC as depicted in Figure 7 (f), the smallest improvement with the emergency right behind an emergency slot 27.45 ms vs. 24.7 ms (Figure 7 (d)). When the emergency occurred in the middle between two emergency slots, the median latency still shows an improvement of 14.6 ms with 38.67 ms measured in the VeMAC and 24.07 ms in the EEVeMAC as shown in Figure 7 (e).

3) Latency in dense traffic scenario:

In the interchange scenario with additional traffic, 63 nodes were present in range of the emergency vehicle. The results showed overall similar results as in the normal interchange scenario. The median latency was measured slightly higher with 16.05 ms (EEVeMAC) vs. 75.65 ms (VeMAC) as shown in Figure 9.

In the straight scenario with dense traffic configuration the latencies are closer together with 21.99 ms (EEVeMAC) vs. 52.69 ms as shown in Figure 10. The respective results of both scenarios in dense configurations for the different emergency slot placements can be seen in Figure 11. In this scenario, there were 159 nodes in range of the emergency vehicle. In contrast between these two configurations one can see that the latencies remain quite stable in EEVeMAC with 27.05 ms in the straight dense scenario vs. 24.05 ms in the interchange dense scenario for accident slot number 1, 24.08 ms vs. 22.07 ms for slot number 25, and 0.69 ms vs 0.70 ms for slot number 49. In VeMAC, the data is more heterogeneous since they do not have always the same (emergency-) slots to transmit the data. The exact numbers are 59.65 ms vs. 33.05 ms for accident in slot number 1, 51.15 ms vs 82.77 ms for accident in slot number 25, and 46.29 ms vs 85.05 ms for accident slot number 49.

4) Collisions:

The reservation of two slots for transmission of emergency messages results in a higher expectation of collisions. Instead

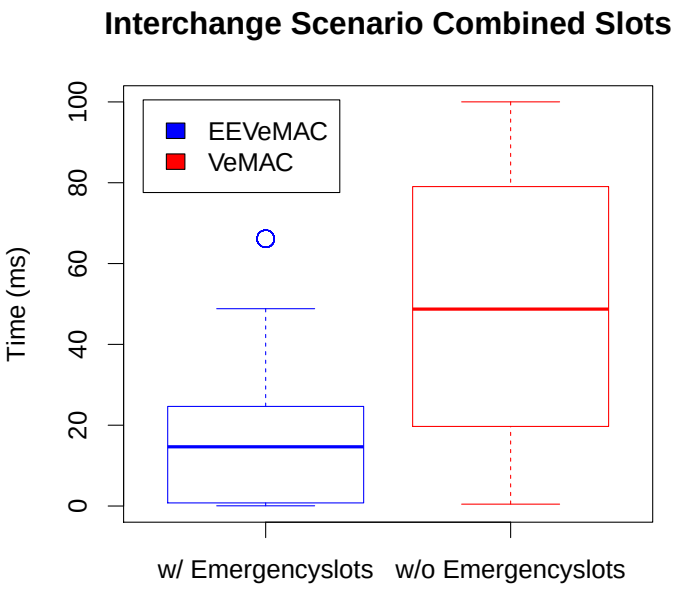


Figure 8. Evaluation results of interchange scenario: On the left the latency results of the EEVeMAC, on the right the latency results of original VeMAC.

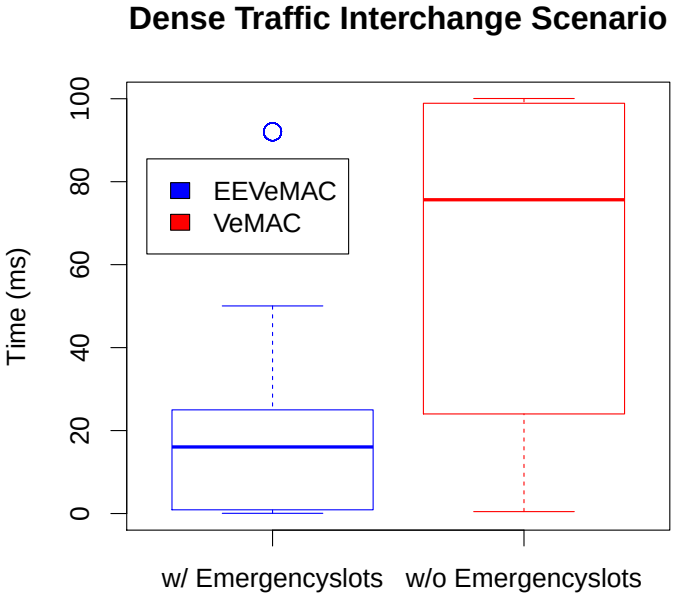


Figure 9. Evaluation results of dense traffic interchange scenario: On the left the latency results of the EEVeMAC, on the right the latency results of original VeMAC. The slots of the three configurations are combined in this diagram.

of 100 slots for transmission of their regular message, the nodes only have a maximum of 98 slots to choose from. Therefore, we also measured the number of collisions. As a measurement, we took the average number of collisions per node. The number of collisions increases in the straight scenario from 0.045 average collisions per node in the original VeMAC to 0.047 average collisions per node in the EEVeMAC. The results of the second scenario show that the effect is negligible compared to the effect the number of nodes have. The VeMAC had 0.294 collisions per node whereas the EEVeMAC had 0.290 collisions per node on average. EEVeMAC having lower collisions per node shows that the randomization has a bigger impact on the collisions than the protocol changes in this traffic density. The average number of collisions increases further with additional traffic in the dense traffic scenario. In the simulation runs with VeMAC, 0.528 collisions occurred whereas EEVeMAC measured 0.663 collisions.

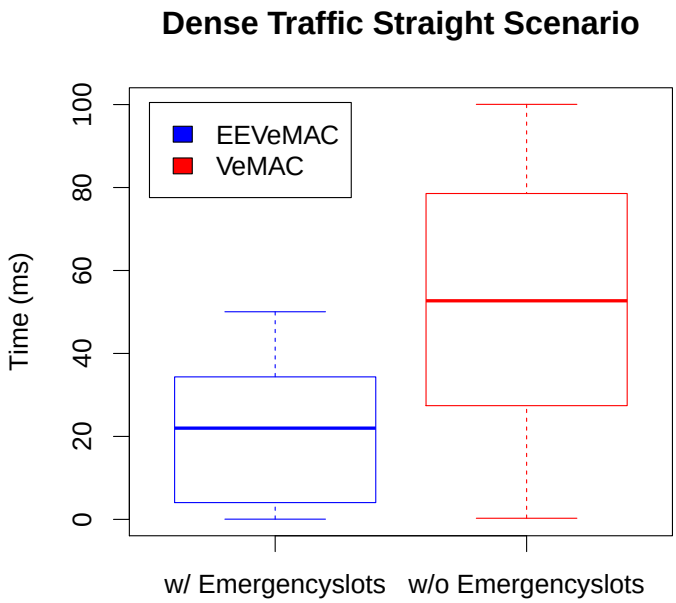


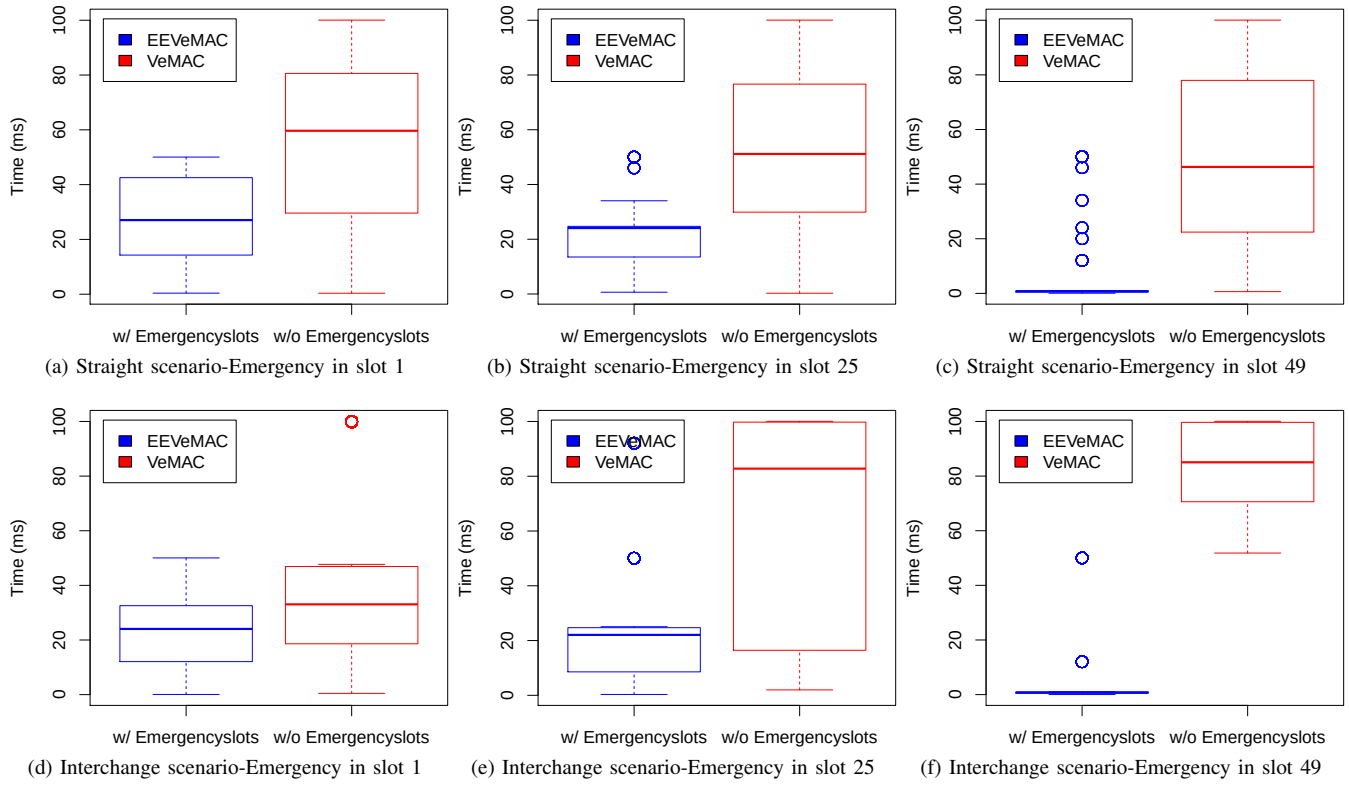
Figure 10. Evaluation results of dense traffic straight scenario: On the left the latency results of the EEVeMAC, on the right the latency results of original VeMAC. The slots of the three configurations are combined in this diagram.

B. Results with four emergency slots

To evaluate the impact of the number of emergency slots on latency, we increased the number of emergency slots to four. These additional slots were added in the middle between the existing emergency slots on slot number 25 and 75. Each simulation configuration was run with the same simulation parameters as before, i.e., emergency occurrence in slot 1, slot 25, and slot 49. The result diagrams are shown in Figure 12.

1) Latency in straight scenario:

In the straight scenario with the emergency in slot 1, the median latencies of the configurations with emergency slots were quite close together at 23.14 ms (2 Slots) and 22.81 ms for EEVeMAC (4 Slots). The simulation for VeMAC without emergency slots had a much longer median latency at



(a) Straight scenario-Emergency in slot 1 (b) Straight scenario-Emergency in slot 25 (c) Straight scenario-Emergency in slot 49

(d) Interchange scenario-Emergency in slot 1 (e) Interchange scenario-Emergency in slot 25 (f) Interchange scenario-Emergency in slot 49

Figure 11. Overview of evaluation of latency results for the straight and the interchange scenarios in dense traffic configurations with the emergency in slots 1, 25, and 49. The small red circles in the figures indicate outliers.

58.15 ms as shown in Figure 12 (a). There were several outliers in the variant with only 2 emergency slots as the next possible transmission slot is further away when cars miss the first emergency slots due to collisions or other reasons.

A similar result can be seen in the next scenario Figure 12 (b), with the emergency happening in slot 25, with the median latency of 22.30 ms (2 Slots) and 21.22 ms (4 Slots) for EEVeMAC and 60.09 ms for VeMAC respectively.

The configuration in Figure 12 (c) showed no big difference between both EEVeMAC variants with median latencies around 1 ms (1.00 ms vs. 0.99 ms) as the emergency is right before an emergency slot in both configurations. The VeMAC took longer with a median latency of 58.54 ms.

Additionally, we also run some simulations with the dense straight traffic scenario to examine how the traffic density affects the latency in this case. The results were pretty similar with 21.99 ms combined median latency with two Emergency Slots vs. 20.14 ms combined median latency with four Emergency Slots. The combined results are shown Figure 13.

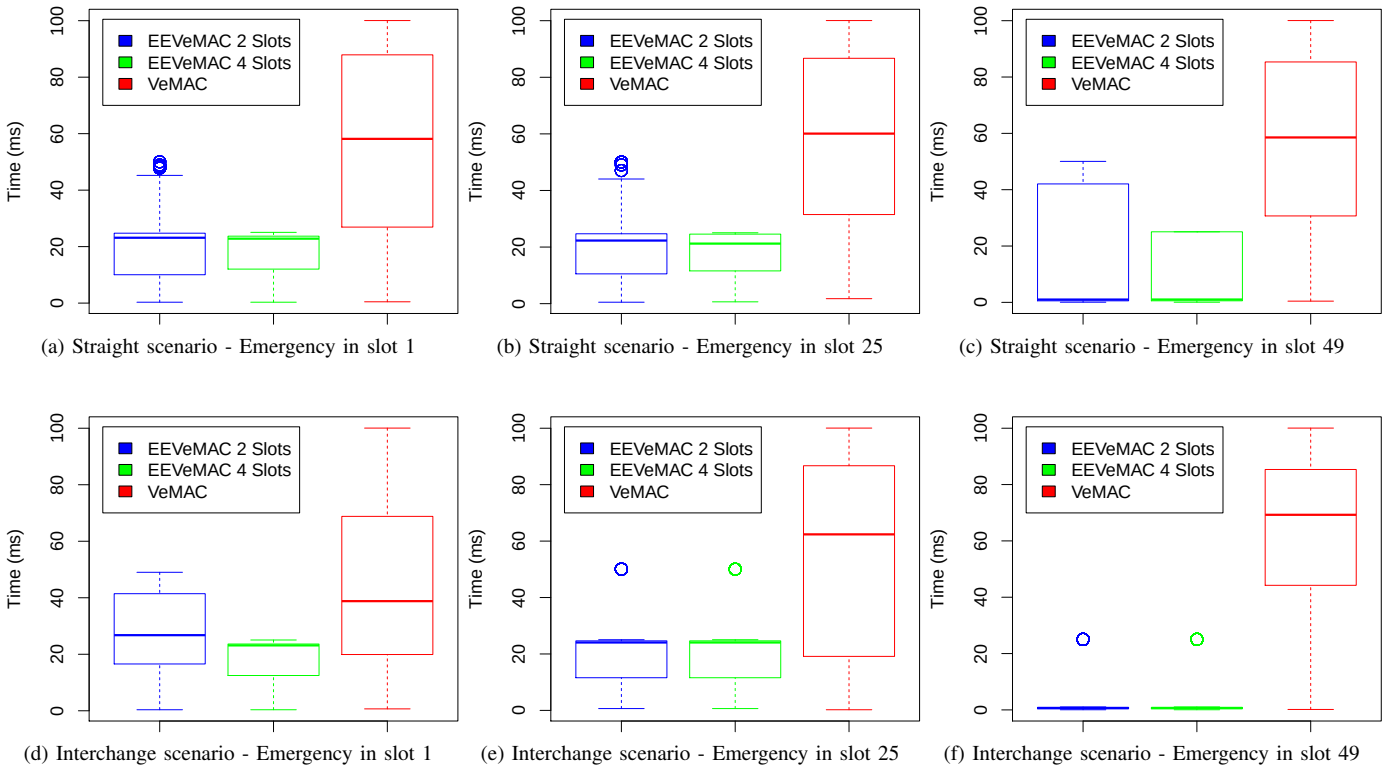
2) Latency in interchange scenario:

The simulation configuration in Figure 12 (d) with four emergency slots showed a reduction of the latency for the EEVeMAC. The median latency of four slots EEVeMAC was measured to be 23.13 ms and as much as expected lower than 25 ms maximum needed to the next emergency slot. This also shows in the absence of latency measurements higher than 25 ms. With only two emergency slots, this median rises to 26.75 ms, but in case of VeMAC without emergency slots, it is the highest at 38.80 ms.

In both of the next two simulation configurations in Figure 12 (e) and Figure 12 (f), the results of the two slots EEVeMAC and the four slots EEVeMAC were the same as they had the same configurations of emergency slots from their respective position onwards, whereas the VeMAC has higher latency as expected.

3) Collisions:

In the straight scenario, the mean number of collisions was measured lower in the four slots scenario than in the two slots scenario. Since the results are very close together with 0.0098 collisions per node in the two slots scenario in comparison to



(a) Straight scenario - Emergency in slot 1 (b) Straight scenario - Emergency in slot 25 (c) Straight scenario - Emergency in slot 49

(d) Interchange scenario - Emergency in slot 1 (e) Interchange scenario - Emergency in slot 25 (f) Interchange scenario - Emergency in slot 49

Figure 12. Overview of evaluation of latency results for the straight and the interchange scenarios with the emergency in slots 1, 25, and 49. The small red circles in the figures indicate outliers.

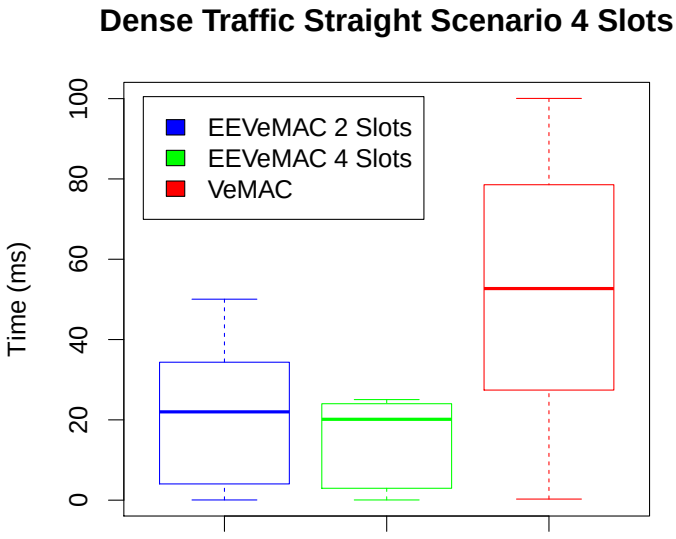


Figure 13. Evaluation results of dense straight traffic scenario: On the left the latency results of the EEVeMAC with two Emergency slots, in the middle the results of the EEVeMAC with four Emergency slots, on the right the latency results of original VeMAC. The slots of the three configurations are combined in this diagram.

0.0092 collisions per node in the four slots scenario.

It is assumed that it is an outcome from the randomness and does not have the cause in the changed slot configuration. Although the configuration with no emergency slots showed fewer mean collisions per node with 0.0068. The traffic flow had a much bigger impact on the collision rate in the interchange scenario.

The mean number of collisions increased by two magnitude up to 0.681 collisions per node in the two slot scenario and 0.685 collisions with four slots scenario. Without emergency slots, it had still a similar magnitude with 0.615 average collisions per node.

VI. Conclusion

The introduction of emergency slots in VeMAC shows great improvements for the transmission of high-priority emergency messages. Instead of median latencies of up to 80+ ms we achieved in our simulation experiments a maximum of median latencies smaller than 25 ms. The latencies were reduced by factor of 3-4. The median and average latencies were improved in each study configuration. The reduction of available slots for regular transmission through the reservation of emergency

slots had negligible effects on the rate of collisions. In situations where several vehicles try to send out an emergency message at the same time, competition emerges and latency increases as the vehicles fail to acquire the emergency slot. The vehicles can still use their normal slots to transmit the emergency message, which means that the average latency converges to the maximum latency of VeMAC, e.g., $100\,\mathrm{ms}$.

Further, with the introduction of four emergency slots, median latencies were almost consistent with the two emergency slots configurations. Overall, EEVeMAC with emergency slots shows great improvements for transmission of time-critical traffic compared to VeMAC.

REFERENCES

[1] Erik Neuhaus, Saleem Raza, and Mesut Güneş. Emergency optimized low latency MAC protocol for VANETs based on VeMAC. In *The Sixth International Conference on Advances in Vehicular Systems, Technologies and Applications*, Nice, France, July 2017.

[2] Sherali Zeadally, Ray Hunt, Yuh-Shyan Chen, Angela Irwin, and Aamir Hassan. Vehicular ad hoc networks (VANETS): status, results, and challenges. *Telecommunication Systems*, 50(4):217–241, Aug 2012.

[3] Saif Al-Sultan, Moath M Al-Doori, Ali H Al-Bayatti, and Hussien Zedan. A comprehensive survey on vehicular ad hoc network. *Journal of network and computer applications*, 37:380–392, 2014.

[4] M. Conti and S. Giordano. Mobile ad hoc networking: milestones, challenges, and new research directions. *IEEE Communications Magazine*, 52(1):85–96, January 2014.

[5] Marthinus J Booysen, Sherali Zeadally, and G-J Van Rooyen. Survey of media access control protocols for vehicular ad hoc networks. *IET Communications*, 5(11):1619–1631, 2011.

[6] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou. Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions. *IEEE Communications Surveys Tutorials*, 17(4):2377–2396, Fourthquarter 2015.

[7] P. Papadimitratos, A. D. La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza. Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation. *IEEE Communications Magazine*, 47(11):84–95, November 2009.

[8] Maria Kihl. *Vehicular Network Applications and Services*. Vehicular Networks: Techniques, Standards, and Applications. CRC Press, 2009.

[9] Yamen Nasrallah. *Enhanced IEEE 802.11. p-Based MAC Protocols for Vehicular Ad hoc Networks*. PhD thesis, Université d'Ottawa/University of Ottawa, 2017.

[10] Mohamad Yusof Bin Darus and Kamarulnizam Abu Bakar. Congestion control framework for disseminating safety messages in vehicular adhoc networks (VANETs). *International Journal of Digital Content Technology and its Applications*, 5(2):173–180, 2 2011.

[11] J. B. Kenney. Dedicated short-range communications (DSRC) standards in the united states. *Proceedings of the IEEE*, 99(7):1162–1182, July 2011.

[12] IEEE standard for information technology– local and metropolitan area networks– specific requirements– part 11: Wireless lan medium access control (MAC) and physical layer (PHY) specifications amendment 6: Wireless access in vehicular environments. *IEEE Std 802.11p-2010 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, IEEE Std 802.11n-2009, and IEEE Std 802.11w-2009)*, pages 1–51, July 2010.

[13] D. Jiang and L. Delgrossi. IEEE 802.11p: Towards an international standard for wireless access in vehicular environments. In *VTC Spring - IEEE Vehicular Technology Conference*, pages 2036–2040, May 2008.

[14] Georgios Karagiannis, Onur Altintas, Eylem Ekici, Geert Heijenk, Boangoat Jarupan, Kenneth Lin, and Timothy Weil. Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions. *IEEE Communications Surveys & Tutorials*, 13(4):584–616, 2011.

[15] Hassan Aboubakr Omar, Weihua Zhuang, and Li Li. VeMAC: A TDMA-Based MAC Protocol for Reliable Broadcast in VANETs. *IEEE Transactions on Mobile Computing*, 12(9), September 2013.

[16] Flaminio Borgonovo, Antonio Capone, Matteo Cesana, and Luigi Fratta. ADHOC MAC: New MAC architecture for ad hoc networks providing efficient and reliable point-to-point and broadcast services. *Wireless Networks*, 10(4):359–366, 2004.

[17] András Varga and Rudolf Hornig. An overview of the omnet++ simulation environment. In *Proc. of the 1st international conference on SIMTOOLS*, page 60, 2008.

[18] Veins: The open source vehicular network simulation framework. http://veins.car2x.org. Date last visited 28.04.2018.

[19] SUMO – Simulation of Urban MObility. http://sumo.dlr.de/. Date last visited 02.04.2018.

[20] Andreas Köpke et al. Simulating wireless and mobile networks in OMNeT++ the MiXiM vision. In *Proceedings of the 1st international conference on SIMTOOLS*, page 71, 2008.

[21] C. Sommer, S. Joerer, and F. Dressler. On the applicability of two-ray path loss models for vehicular network simulation. In *2012 IEEE Vehicular Networking Conference (VNC)*, pages 64–69, Nov 2012.

[22] Strassen.NRW. https://www.strassen.nrw.de/. Date last visited 14.03.2018.

ClusterWIS Revisited

An Updated Look at the Decentralized Forest Information and Management System

Jürgen Roßmann, Michael Schluse, Martin Hoppen
Institute for Man-Machine Interaction
RWTH Aachen University
Aachen, Germany
email: {rossmann,schluse,hoppen}@mmi.rwth-aachen.de

Christoph Averdung
CPA ReDev GmbH
Siegburg, Germany
email: averdung@supportgis.de

Gregor Nägele, Tobias Marquardt
Department Robot Technology
RIF Institute for Research and Transfer e.V.
Dortmund, Germany
email: {gregor.naegele,tobias.marquardt}@rt.rif-ev.de

Werner Poschenrieder, Fabian Schwaiger
Chair of Forest Growth and Yield Science
Technical University of Munich
Freising, Germany
email:
{werner.poschenrieder,fabian.schwaiger}@lrz.tum.de

***Abstract*—The cluster forestry and wood's major challenges are its structural complexity and heterogeneity, its many stakeholders, and its decentralized processes. The aim of the ClusterWIS approach is to overcome these challenges. Its core idea is the development of a novel forest information system based on a decentralized infrastructure integrating new planning and consulting methods and interconnecting existing decentralized work processes. It provides end-to-end encrypted communication to run the various processes and to supply them with data while using international standards throughout the system and keeping participation requirements low. The paper at hand gives details on the ClusterWIS communication infrastructure, its apps and services, and the reference applications as well as their realization in a comprehensive demonstrator.**

Keywords-forest information system; sustainable feedstock management; decentralized data management; secure communication; demonstrator.

I. INTRODUCTION

In [1], we introduced the idea of ClusterWIS, a completely decentralized forest information and management system that conveys principles of Industry 4.0 to the cluster forestry and wood. The conceptual framework of ClusterWIS has now been further developed in its communication infrastructure and implemented as a demonstrator for practice oriented test and promotion.

The cluster forestry and wood is the economic sector comprising all stakeholders from forest owners to forestal service providers and the woodworking industry. Its major challenges are its structural complexity and heterogeneity, a huge number of stakeholders with often contrary objectives, and decentralized processes. In the federal state of North Rhine-Westphalia (Germany) alone, 150,000 private forest owners own two-thirds of the forest (90% of which own less than 5 ha), and many small service providers (for planning, tending, logging, etc.) exist [2]. Furthermore, the "production plant" forest provides not only wood as its main product (used for building, paper or as a fuel) but also serves as a long-term CO_2 reservoir or as a recreation area. Altogether, this renders process optimization far more complex than in classical manufacturing industry.

Thus, for a sustainable feedstock management and an efficient wood and biomass mobilization throughout the cluster, the increasing demand for wood from sustainably cultivated forests need to be aligned with the requirements of climate change and resilience, environmental protection and society in general. For that purpose, the research project ClusterWIS (WIS for German "Waldinformationssystem" – Forest Information System) introduces novel planning and consulting methods. In order to implement these processes within a decentralized infrastructure, ClusterWIS adopts, refines and widely interconnects the crucial working processes within the wood and forestry sector.

Often, centralized approaches have been used to resolve the cluster's structural weakness. However, such approaches, contradict the highly decentralized organization that is typical for the whole sector. In particular, many conservative forest owners do not accept an obligatory centralized data management for reasons of data privacy (especially in Germany). For this reason, the foundation of the ClusterWIS approach is a novel, decentralized infrastructure based on standards for data modeling and data exchange. It provides end-to-end encrypted communication to run the various processes and to supply them with highly topical inventory and process data. To provide for the cluster's heterogeneity, it keeps the participation requirements for third party systems low. Furthermore, international standards are used throughout the system like Open Geospatial Consortium (OGC) web service standards, Geography Markup Language (GML) for data exchange in general, ForestGML [3] for *n*D temporal inventory data, ELDATsmart [4] for timber logistics data, StanForD [5] for forest machine data, or papiNet [6] for communication with the paper industry. However, the approach is open to any other formats and standards.

A ClusterWIS network (Figure 1) comprises applications and services as its nodes, connected by means of the secure communication infrastructure. Stakeholders use specialized desktop and mobile applications (e.g., for forest information, forest inventory, production planning etc.) or web applications (e.g., for forest owners, service providers etc.) to access this network. Specialized services, e.g., for processing remote sensing data or forest growth simulations, perform computationally expensive and data intensive tasks for a broad user group even on thin clients like mobile devices. Finally, services for administrative tasks like communication, cloud storage or registration build the network's backbone.

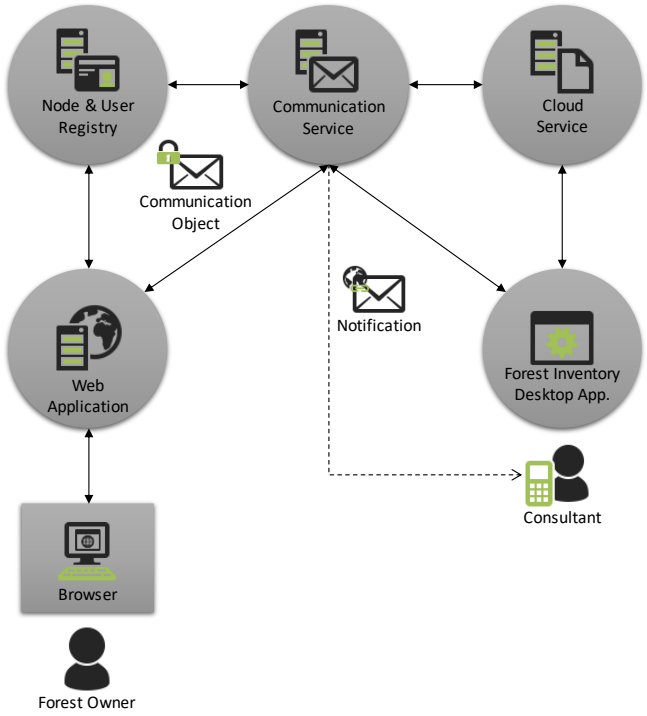


Figure 1. A ClusterWIS network consists of service and application nodes.

For the first time, this decentralized network allows for process optimization across the cluster. In the research project, eight interdependent reference processes (like forest information, planning, consulting, timber trade and production) are analyzed in detail.

The ClusterWIS concept can also be described in terms of four layers:

ClusterWIS Documentation: This layer provides the documentation of the service infrastructure comprising data standards, a common communication standard, and a common public key infrastructure (PKI). Furthermore, it contains best practices for data exchange (e.g., which formats to use), the usage of software services (e.g., for remote sensing data processing), and for offering services (e.g., consultation or forest inventory).

ClusterWIS Directory: While *ClusterWIS Documentation* yields a "common language", *ClusterWIS Directory* provides the necessary directory (or registry) of users and nodes. This includes the infrastructure for their registration, retrieval, and authentication – all needed for secure communication and data exchange.

ClusterWIS Notifications: In principle, the previous two layers suffice to realize and operate a ClusterWIS network. This third layer adds a communication infrastructure to further simplify data exchange between actors within the forestal domain (where senders and receivers are often offline) by buffering messages and by publishing notifications for them.

ClusterWIS Applications: This layer consists of the software solutions built on top of the ClusterWIS infrastructure. The focus is on extending existing software. For that purpose, the participation requirements are kept low (see previous levels). Such systems comprise locally installed management software (e.g., for forest inventory), web-based management software, platforms for data exchange or trading, and mobile applications.

The rest of this paper is structured as follows: Section II presents work related to our own and motivates the development of the ClusterWIS approach. Sections III and IV give more details on the ClusterWIS infrastructure and its communication approach. Sections V and VI introduce the ClusterWIS applications (desktop, mobile, Web) and specialized services while Section VII gives an overview of the reference processes analyzed in the research project. In Section VIII, we add a detailed look at the recently realized demonstrator and the yielded response. Finally, Section IX concludes this paper.

II. Related Work

ClusterWIS is built on its project partners' preliminary work. Important results come from the research project series Virtual Forest in general, its commercial spin-offs, the underlying SupportGIS technology, and the forest growth simulator SILVA [7]. ClusterWIS aims at making these results available to the whole cluster. Besides summarizing this work, this section introduces similar approaches developed by others.

A. The Virtual Forest

The ClusterWIS approach is built on the methods of the "*n*D Forest Management System Virtual Forest" [8], developed in the research project series "Virtual Forest". It provides the necessary technological framework as well as the basis for data modeling, management and distribution.

The idea of the Virtual Forest is a central database that manages all forestal data. It provides various applications for remote sensing data processing (tree species classification [9], stand attributes evaluation [10], or single tree delineation and attribution [11]), forest inventory, planning in biological and technical production, forest machine simulation for training, and support of the logging process.

The technological basis of the central database is the SupportGIS technology [12]. It is widely used for GIS related applications, is based on the standards of OGC and ISO, and powered by object-relational databases. It efficiently manages large amounts of data and supports exchange by standard OGC web services. Furthermore, data can be managed in *n* spatiotemporal dimensions [3], allowing to track and analyze forestal data over time.

The Virtual Forest uses ForestGML [3], a GML-based modeling language, to model forestal data on a consistent, OGC compliant basis. This facilitates its widespread usage and allows for the usage of OGC web services.

Central parts of the Virtual Forest system are already in use by two German state enterprises. While the Virtual Forest focuses on the usage in such large, homogenous enterprises, ClusterWIS aims at making these results available to the whole cluster by decentralizing the approach.

B. Forest Growth Simulator SILVA

Silviculture today has to consider a wide range of ecosystem services (ES) that earlier were considered a by-product of traditional forestry. Moreover, against the background of climate change, forest management has to maintain climatic resilience and stability through provision of an adequate forest structure. Thus, forest consulting increasingly applies forest simulation models to estimate the effect of various silvicultural pathways on productivity, quality and further ES [13] [14]. Such ES are carbon sequestration, biodiversity, recreation, and groundwater recharge. Yet, they typically stand within the particular focus of state forestry. However, private forest stakeholders today also advocate to foster the adaptation of such services by private forestry based on financial incentives [15]. The forest ecosystem model [13] is a preferential tool to take into account ES synergies and tradeoffs and to optimize among various silvicultural objectives. It allows forest consulting to compare scenarios that adhere to a sensible preselection of silvicultural pathways and to direct forest management towards the most effective subset of them. Such simulation models have primarily been applied within state forestry institutions that maintain the necessary IT infrastructure.

Within that scope, the forest growth simulator SILVA has been integrated into the aforementioned Virtual Forest system. SILVA implements the paradigm of a service oriented architecture (SOA). Its kernel is an independent application that does not expose any specific tasks but rather a wide collection of services that may be coupled and assembled to provide specific simulations or evaluations. Moreover, it provides its services through various types of interfaces that use international standards, such as the Simple Object Access Protocol (SOAP). Thus, it fits into a distributed environment as well as into a strictly local one. Within the Virtual Forest project, several scenarios that integrate SILVA both locally and as a remote service into the larger environment were implemented and tested.

C. Similar Approaches

Until now, others developed approaches similar to ClusterWIS. Some proprietary solutions are available, e.g., online platforms like “IHB Holzbörse” for timber trade [16] or the “Branchenbuch Wald und Forst” as a business directory for consultants [17]. The internet marketplace “CoSeDat" offers the possibility to exchange data and electronically signed PDF documents [18]. In Finland, UPM Paper offers “UPM Customer Online” [19], a digital service channel for customers. In summary, these approaches focus on specific aspects of the complex process chain, only. Hence, a permeability of shared data between the different processes is not given. Often, the idea was to develop centralized systems such as “virtual enterprises” [20], the “FOCUS-Platform” [21]. Another centralized approach is presented in the research project “GeProOpt_Holz” [22]. Similar to ClusterWIS, it aims at optimizing business processes in the cluster. However, as mentioned in the introduction, such centralized approaches are not accepted by many cluster actors.

Software solutions like “WaldPlaner” [23] already deliver functionality for planning and decision-making regarding sustainable forest management, but on their own they lack the necessary communication infrastructure and integration into larger processes. Approaches like the “Scottish Forest and Timber Technologies initiative”, supported by enterprises and industry, promote knowledge exchange and cooperation between enterprises in the sector [24]. They are successfully able to connect regional actors, but the knowledge is not shared more widely. The web portals “Wald in Österreich” [25] in Austria and “WaldSchweiz” [26] in Switzerland serve the exchange of information in the sector.

Thus, existing approaches already cover several of the cluster’s requirements. However, they lack an approach to connect its complex and decentralized structure and do not support a comprehensive execution of all the essential business processes from request to invoicing. This motivates the development of the ClusterWIS approach as introduced in Section I.

III. INFRASTRUCTURE

The cluster’s achievable efficiency is strongly related to the way its actors communicate. This requires a framework that does not unnecessarily restrict an actor’s professional view or its organization’s structure. Thus, the ClusterWIS infrastructure is based on secure networking of so-called ClusterWIS nodes. These nodes can either be applications (Section V), specialized web services (Section VI) or services for administrative tasks.

To use its services and applications, any actor can register and participate in the ClusterWIS network. Well-established methods of IT security are employed to guarantee the safety of connections and exchanged data between actors, applications and web services. Client-side Hypertext Transfer Protocol Secure (HTTPS) is used for authentication and secure connections. It is integrated into a PKI that allows for end-to-end data encryption. Finally, authorization is based on GeoXACML (Geospatial eXtensible Access Control Markup Language) providing user rights on data and methods.

The administration of the ClusterWIS network is reduced to few central services:

- A node and user registry for all participating actors and nodes (applications and web services) accessed via Lightweight Directory Access Protocol (LDAP).
- A Web Feature Service (WFS)-based communication service as a mediator between sender and receiver of so-called communication objects.
- A cloud service used to buffer communication objects, as a general data storage for the network, and as a platform to initialize and run OGC compliant web

services (WFS, Web Map Service (WMS) and Web Map Tile Service (WMTS)) on its stored data.

This lean infrastructure (combined with its communication approach presented in the next section) also keeps the participation requirements for third party systems low.

IV. COMMUNICATION

Three basic rules apply to communication within a ClusterWIS network: Data and (service) requests are always transferred by secure connections and encrypted by the public key of the recipient. Furthermore, recipients account for conformant data usage inside their domain. Finally, it has to be assumed that many communication partners and systems are regularly offline (e.g., when being in the forest with bad reception).

A. Communication Object and Encryption

The aforementioned communication objects are a means for the secure transfer of data and corresponding requests. As shown in Figure 2, a communication object consists of two parts: The non-encrypted transport information (top) and the encrypted secured message (bottom).

While not encrypted, the transport information is still secured by transport level security using HTTPS. It comprises information on the sender, the receivers, the utilized encryption (see below), and optional metadata used to flexibly add further information needed on transport level, e.g., to identify that a message is a response to another one.

The encrypted message comprises the intention in terms of a request, which can be, e.g., a service call, a service call response or a user-to-user message. It is complemented by parameters, e.g., for service calls (name of the service to be executed, surface under investigation, alphanumerical parameters), service call responses (a computation's result), or subject and body of a user-to-user message. Furthermore, embedded files, links to files on cloud services, or metadata describing files can be added.

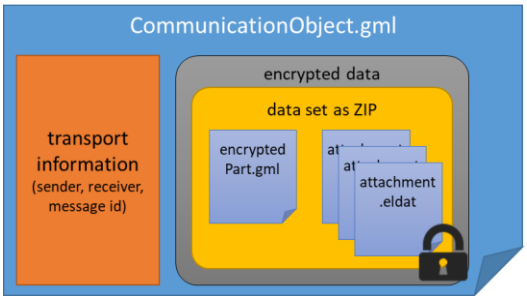


Figure 2. Implementation of the ClusterWIS communication object.

The communication object is implemented as follows (Figure 2): The to-be-secured message is stored as a GML file ("encryptedPart.gml"). The data set comprising this GML file together with the (optional) file attachments is stored in a ZIP file that is subsequently encrypted.

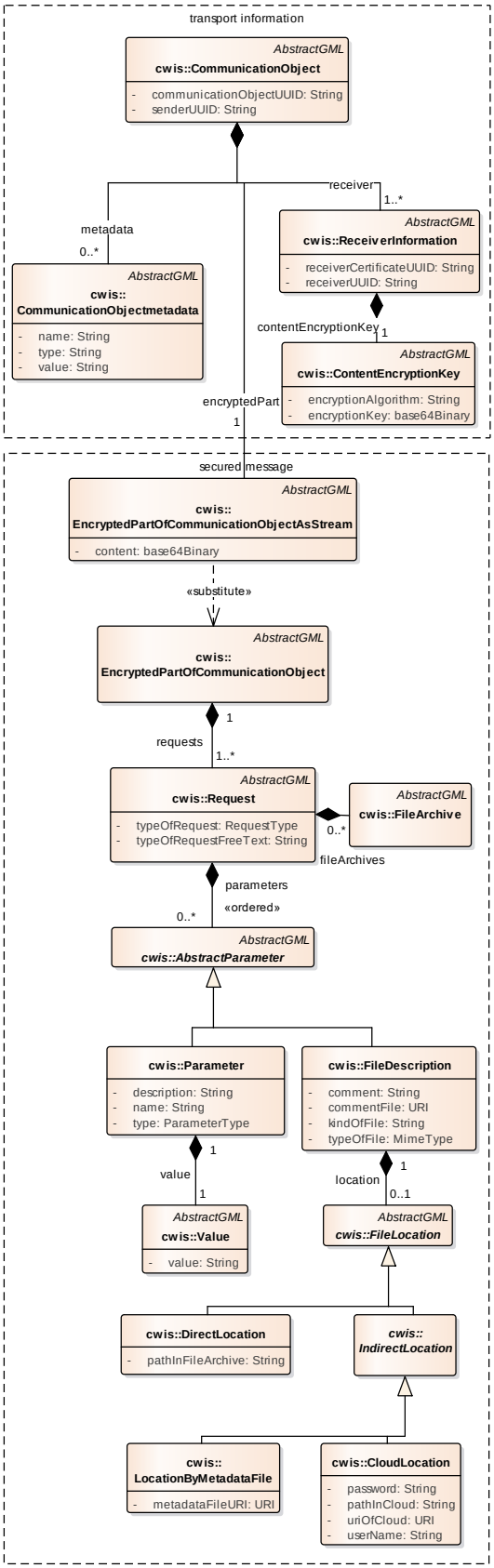


Figure 3. Structure of the ClusterWIS communication object.

The unencrypted transport information combined with the encrypted data is also stored as a GML file ("CommunicationObject.gml"). Thus, only the transport information remains readable.

Figure 4 illustrates the encryption of the communication object in more detail. First, the zipped data set is encrypted by an automatically generated symmetric key that is disposable and only used for this single data interchange. In turn, this key is encrypted by the receiver's public key and stored together with the user id of the receiver within the transport information.

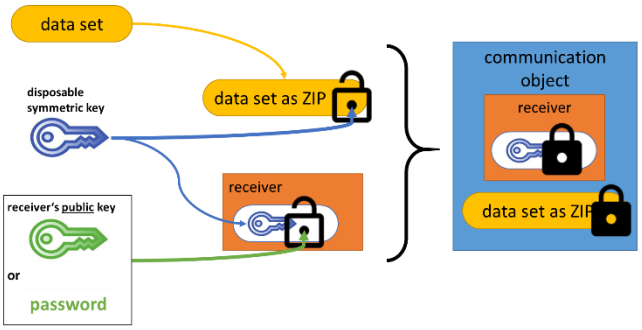


Figure 4. Encryption process of the ClusterWIS communication object.

This combination of encryption methods has three advantages:

- First of all, symmetric encryption is much more efficient on larger data sets.
- Secondly, the same data set can be encrypted for multiple receivers. For that purpose, the same disposable symmetric key is separately encrypted for each receiver of the communication object using the receiver's public key (Figure 5) and embedded into the transport information.
- At last, for the sake of keeping requirements low for participants in the ClusterWIS network, even users without a public/private key pair can participate. In this case, the disposable symmetric key is encrypted using a password agreed upon by the sender and the receiver(s).

For decrypting a message (Figure 5), the receiver uses his own private key only he knows (or the agreed upon password).

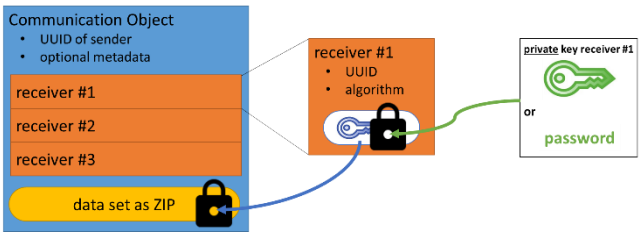


Figure 5. Decryption process of the ClusterWIS communication object.

This yields the disposable symmetric key needed to decrypt the ZIP archive containing the actual data set. Finally, the received data can be accessed.

B. Communication Service

The communication service operates the transfer of communication objects. Using HTTPS, a second layer of security is added to the encryption of the communication objects. This guarantees that neither unauthorized third parties nor the cloud service or even the communication service itself get access to the data.

The registry service provides information on the dispatch method for the communication objects, which may be different with regard to the receiving user (or node). Following dispatch methods are available:

- Notification by mail or smartphone push message including a download link to the communication object.
- Direct delivery using a WFS-like interface of the recipient.
- Actively pulling the list of new communication objects from the communication service.

The communication service provides a WFS interface that uses the unencrypted part of the communication object schema (Figure 3). Thus, a new communication object is sent to the communication service by using a WFS transaction (insert). For each communication object, the communication service creates a lightweight acknowledgement object (Figure 6) for message management. Here, any update to a communication object's state is logged (received, notification sent, communication object fetched …). Using these acknowledgement objects, the receiver can query a list of new messages (communication objects) from the communication service using a standard WFS query. Subsequently, this list can be used to fetch the actual communication objects themselves.

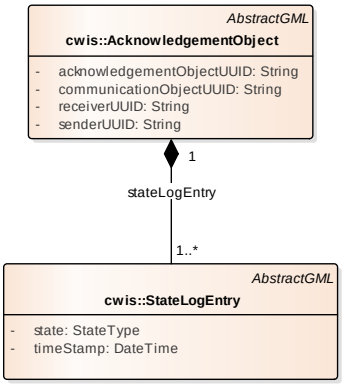


Figure 6. Acknowledgement objects provide an interface to fetchable communication objects.

A communication example is shown in Figure 1 (note that the disposable key is omitted in this sequence): A forest owner uses a browser to access the web application (Section V.B) and create a communication object with a request for forest inventory. It is encrypted (using the public key of the consultant) and sent to the communication service, which buffers it and sends a notification to the recipient (consultant). The latter starts its Forest Inventory application (Section

V.A.2), which asks the communication service for new communication objects that are subsequently fetched. Finally, the consultant decrypts the communication object with his private key and processes the message.

Note that all connections between nodes are additionally secured by client-side HTTPS, which is also used for authentication.

C. Open Platform Communications - Unified Architecture

The decentralized ClusterWIS approach is similar to those approaches subsumed as "Industry 4.0" in the manufacturing industry. Thus, it can be counted among current intentions relating to an adaption for the cluster forestry and wood dubbed "Forestry and Wood 4.0".

Furthermore, ClusterWIS communication not only takes place between actors but also from and to forest machinery. This motivates the integration of standard Industry 4.0 protocols into the network. As a well-established standard, Open Platform Communications – Unified Architecture (OPC UA) [27] is advisable for this purpose. Especially, as it provides a decentralized client server architecture without the need for central servers, it integrates well into the ClusterWIS PKI, it is an open and vendor-independent standard, it is robust, and it supports participants being temporarily offline.

Thus, to complement the aforementioned WFS-based approach, ClusterWIS nodes may also be equipped with an OPC UA client and server component allowing the exchange of communication objects.

V. Applications

An important part of the ClusterWIS approach are the user and scenario specific portals the actors can use to access the network. These comprise desktop, mobile and web applications.

A. Desktop and Mobile Applications

Desktop and mobile end-user applications provide online as well as offline access to ClusterWIS features. They can be used by actors like forest owners, service providers, or contractors to view, gather, modify, and exchange forestal data. In the context of the research project, applications are based on the Virtual Forest prototypes. They use the VEROSIM framework [28] that combines an integrated runtime database with subject-specific modules to create adapted applications for diverse scenarios.

Four different applications are being developed and refined to meet the requirements of the project's reference processes as described in Section VII:

1) Forest Information

The Forest Information application acts as an information portal to the data managed by ClusterWIS. Its primary functions are visualization, combination and analysis of geographic and business data, e.g., orthophotos, satellite imagery, LiDAR, cadaster, inventory, or regulatory data. This data may be available locally via files and databases or provided by OGC-compliant web services (WMS, WMTS, WFS) within the ClusterWIS network.

2) Forest Inventory

This application supports the forest inventory process. It allows a service provider to work with data made available by the commissioning forest owner and provides tools to record relevant stand attributes and single tree information. As this data is typically gathered on-site, the software also offers assistance for spatial localization during the process.

3) Forest Planning

The Forest Planning application provides a user-friendly and efficient interface to forest growth simulation. This comprises input parameterization as well as result analysis and visualization. The computationally intensive simulation itself is sourced out to a service (see Section VI.B).

4) Technical Production

This application supports the technical production process in its different phases, namely preparation of work assignments, assistance of forest workers and machine operators with instructions, and practical guidance as well as documentation of the harvesting operations and its results.

B. Web Applications

Web applications are ideally suited to provide a low-threshold access to the ClusterWIS network. They do not need client-side installation and can be used on both desktop and mobile devices alike. Capacity and performance scaling is easy and new features can be provided to users with no effort. Finally, web applications easily support operation in secure networks.

The browser-based GIS SGJ GeoHornet is one example of such a web application that is used in ClusterWIS. It has already successfully been employed in the Virtual Forest project as well as its commercial spin-offs. Various data sources like ForestGML databases or web services can be accessed and embedded. For example, registered forest owners can get an overview of their entire property. GeoHornet also provides methods to plot maps and enhance these plots with own graphical and textual annotations. It can create, send and receive communication objects, e.g., to send a request to another actor in the ClusterWIS network or to commission a service-based computation. GeoHornet can be customized for the user's demands.

VI. Specialized Services

Based on the backbone services of ClusterWIS (Sections III-IV), a large variety of specialized services for supporting particular business processes can arbitrarily be interconnected. Examples for such services are the processing of remote sensing data or the simulation of forest growth. As a single requirement, a specialized service has to comply with the ClusterWIS communication approach and its PKI.

Within a typical scenario, an application might call a service for data processing and store the results either locally or in a ClusterWIS cloud service. Subsequently, it might run two complementing simulation services based on this preprocessed data set. Finally, the application might use various evaluation services for evaluating the resulting scenario data.

A. Remote Sensing Data Processing

Often, the data necessary for a sustainable feedstock management can only be made available using remote sensing methods like tree species classification [9], stand attributes evaluation [10], or single tree delineation and attribution [11]. However, such methods usually need to access, process and store vast amounts of raw geo data, unfeasible, e.g., for mobile apps. Furthermore, existing methods need to be enhanced to easily incorporate stakeholders to refine the data with their expert knowledge (e.g., provide tree samples to optimize local tree species classification results). Thus, a goal of the ClusterWIS project is to make these methods available as services to allow the usage of suitable hardware on server side and to provide service interfaces for user provided calibration data.

B. Forest Growth Simulation

Forest growth simulators - beyond scenarios of stand development - provide further services that are closely connected to a simulator's core function. Such services are virtual tree generation based on stand structure attributes and computation of assortments using individual tree data. Hence, one relevant task within ClusterWIS is to extend existing data formats, such as ForestGML, to comply with the time-related data content that is specific to simulation models.

SILVA provides stand development as a result of rule-based management plans. That way, the simulator may provide scenarios that put emphasis on a specific subset of ecosystem services or that promote the development of specific stand structures and species mixtures. The seamless and manifold integration of SILVA [7] into the ClusterWIS infrastructure enables its coupling to any other service that might receive data from the simulator or provide essential basic data to it. That objective is particularly important against the background of ecosystem service provision. Ecosystem services are typically linked by mutual synergies and tradeoffs. Therefore, one relevant coupling scenario is the linkage between SILVA and vegetation distribution models. Such specialized land surface models [14] represent processes of vegetation growth, seed dissemination and disturbance. They may thus provide valuable results about the establishment of regeneration trees and individual young trees to forest growth simulators. Moreover, as vegetation models often use a simplified representation of main stand development, they might straightforwardly integrate individual tree data provided by the growth simulator.

C. Evaluation Services

While simulation services such as SILVA may provide built-in components for result evaluation, ClusterWIS may also complement such simulation services with dedicated evaluation services to provide them in any environment desired. Evaluation services may be readily implemented by forest and wood scientific staff within statistical environments such as R [29] that do not presuppose profound experience in software development. For example, R based evaluation services may simply be exposed by embedding them into ClusterWIS compliant service wrappers that call the R runtime with script name and parameters. As R provides packages for convenient XML parsing, a wrapper may straightforwardly pass ForestGML data to and from the script being called. Thus, an evaluation service may run on simulation results or survey data that has been generated on a remote system by services hosted on a different platform.

VII. Reference Processes

ClusterWIS not only provides an infrastructure, protocols and applications. It also specifies processes for a sustainable feedstock management realized on this foundation, which will be tested and demonstrated in actual forest stands. An important aspect of ClusterWIS is that these processes do no longer take place in a parallel and unrelated manner but start to interact with each other. A selection of practically relevant reference processes is considered within the project and briefly introduced below:

A. Forestal Data Provision

Sustainable natural resource management requires information and planning. For that purpose, up-to-date, highly qualitative, and detailed (geo) data is needed. Data is usually compiled of various data sources (ForestGML-structured data, third party spatial base data and business specific data). Currently, the comprehensive provision of such data to the cluster is an unresolved problem. Thus, this process describes the provision within the ClusterWIS network.

B. Forest Information

This process describes an actor's access to the provisioned forestal data of a specific area in the right time at the right place, comprising visualization, analysis, and editing.

C. Forest Inventory

Forest inventory is the acquisition and management of environmental data in forestry. Thus, the purpose of this process is to provide the cluster with always up-to-date, detailed and high-quality data. An important aspect in this context is to automatically and logically connect different data sources and, if applicable, different timestamps (for trend analysis) within the *n*D forest information system.

D. Planning and Consulting

The comprehensive data provided by the ClusterWIS network enables consultants to give forest owners efficient and goal-orientated advice on how to manage their forests. In particular, they can use simulation tools to demonstrate how different management alternatives result in different future outcomes.

E. Timber Trade

The ClusterWIS network opens new ways for getting in contact. By providing all relevant information to all actors involved in the process, a more efficient communication between sellers and buyers can be established. Thus, ClusterWIS provides the framework for a more efficient timber trade and contributes to a more efficient wood and biomass mobilization.

F. Sustainable Harvesting

Integrated into the aforementioned processes within the ClusterWIS network, the technical production process can access a vast number of relevant data. This allows for the planning of more sustainable harvesting measures. It comprises the (simulated) determination and visualization of wood assortments, harvesting costs, accessibility and harvesting routes, average skidding distances, as well as aspects of nature conservation. Besides planning, this process also comprises the execution of planned measures and their documentation, where the latter can again be used in downstream processes.

VIII. DEMONSTRATOR

In order to implement and test the developed concepts and to showcase them to potential users, a demonstrator has been created. It serves as a proof of concept and includes prototypical implementations of different applications, services and processes as described above.

The current demonstrator contains two demo scenarios focusing on different aspects of the aforementioned reference processes.

A. Demo Scenario: Forest Inventory and Planning

This scenario contains aspects of the reference processes “Forestal Data Provision”, “Forest Information”, “Forest Inventory” as well as “Planning and Consulting”. It addresses the private forest owner’s demand to easily get up-to-date stand attributes and data evaluations for his forest as a basis for further planning.

The forest owner uses an Android app (Figure 7) with a GIS user interface to connect to the ClusterWIS network.

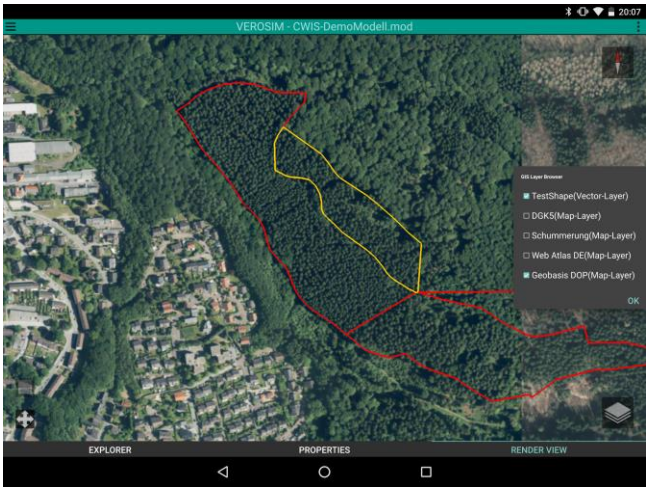


Figure 7. ClusterWIS Android app for the forest owner.

Various freely available WMS are integrated and provide geo data like orthophotos or LiDAR images. Using a remote service for remote sensing data processing (see Section VI.A), the user can retrieve automatically calculated forest inventory data for his or her forest areas just by selecting the desired areas and invoking the service by a simple push of a button (Figure 8). The service receives the work order packaged in an encrypted communicating object that also contains the surface to process on via the communication service. It uses remote sensing data to calculate stand attributes (tree species, area, age, stock volume, top height etc.) and creates a new communication object containing the encrypted results, which are sent back to the forest owner’s client app using the communication service, as well.

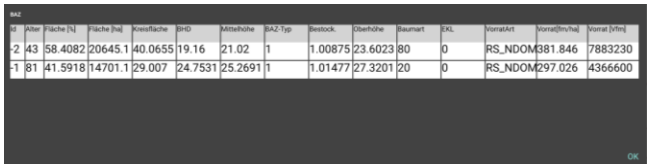


Figure 8. Exemplary stand attributes result from ClusterWIS remote sensing data processing service in the Android app for the forest owner.

In addition, several R-based evaluation services (see Section VI.C) have been integrated into the demonstrator. Based on the previously generated stand attributes, they calculate values like groundwater recharge [30] or key indicators of biodiversity [31]. The results are visualized in the form of diagrams (Figure 9).

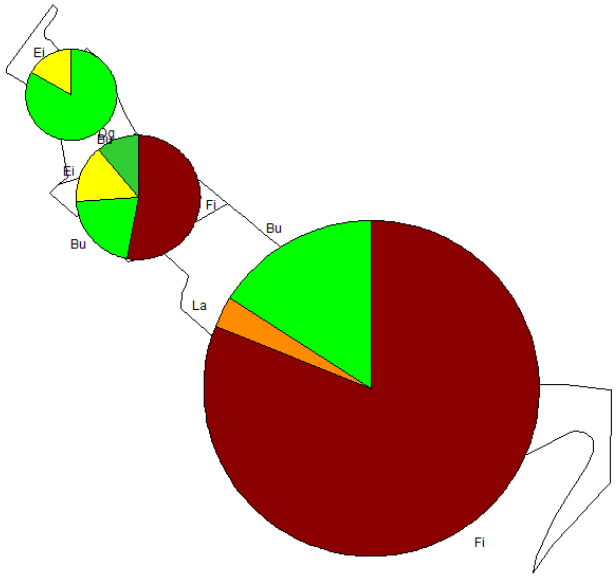


Figure 9. Stand inventory data aggregated to the forest owner specific resolution level (volume share per species and stand) by the ClusterWIS evaluation service and presented in the Android app for forest owners; tree species distinguished by color and abbreviation (Bu: European beech, Ei: oak, Dg: Douglas fir, Fi: spruce, La: larch).

The service is used the same way:

- The user selects surfaces,
- invokes the service call,
- an encrypted communication object containing both is sent to the service via the communication service,
- results are calculated, encrypted and sent back in a new communication object via the communication service, and

- the user's client fetches the new communication object and presents the results.

By using these two services alone, the forest owner already gets a considerable amount of information about his forest on demand and without the need for locally stored masses of geo data, for massive local computing power or for doing site surveys. This information serves as a basis for further planning of tending and felling measurements. The forest owner usually relies on other experts to get efficient and goal-oriented planning advice. To support this process, the mobile app gives access to a "yellow pages" feature, which is based on the registry service, where the user can search, e.g., for consultants registered in the ClusterWIS network and correspond with them directly via his app. In addition to a written request, other relevant data can be attached to the message that is forwarded to the recipient via the communication infrastructure described in Section IV. That way, forest consultants may offer support in forest planning based on the data within the secure communication object. Such planning, in turn, may be backed by scenarios from simulation services like SILVA (see Section VI.B).

B. *Demo Scenario: Timber Trade*

This scenario builds on the previous one and addresses the reference process "Timber Trade". Having planned or already conducted a felling, the forest owner wants to sell the harvested wood to a selling agent or directly to a saw mill.

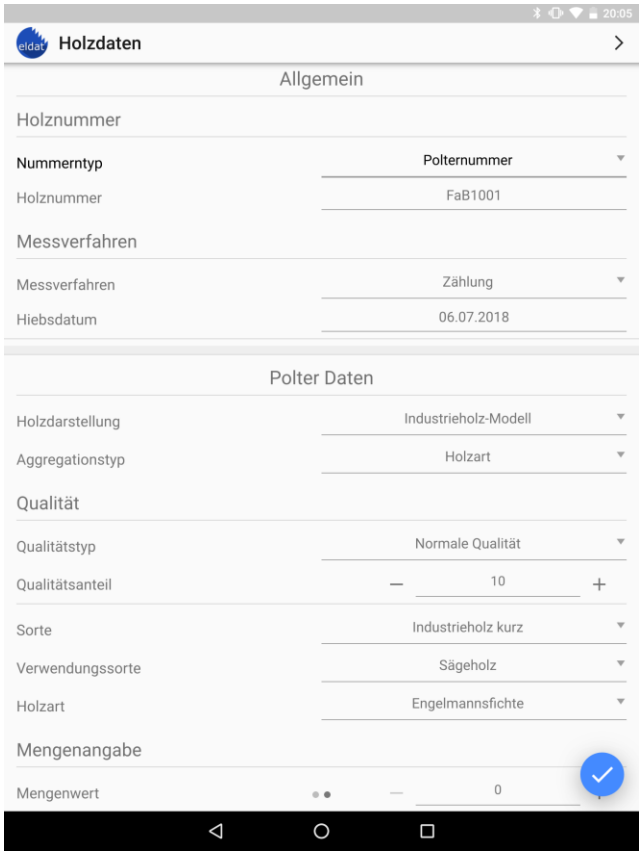


Figure 10. ClusterWIS Android app with a timber logistics user interface for creating ELDATsmart jobs.

For that purpose, he can use the Android app (Figure 10) to create the appropriate ELDATsmart-based timber logistics job for wood allocation. Again, the "yellow pages" feature integrated into the ClusterWIS network helps him find a registered selling agent or saw mill operator. Subsequently, he can send the generated wood allocation job to the chosen recipient using the ClusterWIS communication infrastructure.

The communication partner on the side of the selling agent or saw mill retrieves the message using his own client application, extracts the ELDATsmart file and can continue to use it in his usual workflow.

A similar user interface for timber logistics was also created for a prototype of a ClusterWIS desktop app (Figure 11). Here, the user can also derive ELDATsmart-based timber logistics files, e.g., for wood allocation, which is (in parts, where applicable) derived from inventory data and can subsequently be sent using a ClusterWIS communication user interface within the app.

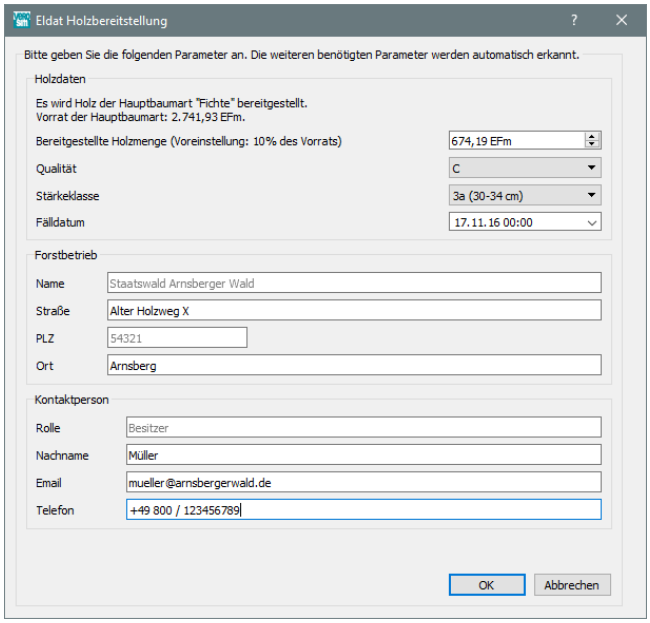


Figure 11. ClusterWIS desktop app with a timber logistics user interface for creating an ELDATsmart wood allocation job.

C. *Achievements*

The demo scenarios implemented in the current demonstrator showcase a variety of concepts and their practical application in real world scenarios. They demonstrate the communication between the cluster's actors, easy and inexpensive information retrieval and processing via user-specific applications and the integration of service providers into the network.

Furthermore, the demonstrator is a proof of concept for the technical aspects of ClusterWIS. For example, different types of nodes have been implemented and connected to a network:

- Desktop and mobile applications tailored to the needs of private forest owners.

- Automated services that calculate stand attributes and evaluations on demand based on remote sensing data.
- A user registry service providing lookup functionality using LDAP.
- A communication service transporting user-to-user and user-to-service messages.

The nodes communicate via HTTPS utilizing the ClusterWIS PKI and communication objects as described above. Client applications and services implement an interface to ForestGML and are able to serialize and deserialize data like surface geometries, stand attributes and evaluation results. In addition, the evaluation service based on SILVA demonstrates the successful integration of existing software into the ClusterWIS network.

D. Presentation and Feedback

The ClusterWIS project and its current demonstrator were successfully presented at the INTERFORST 2018 [32], an international leading trade fair for forestry and forest technology in Munich, Germany. Private forest owners, freelance forestal experts and service companies showed a particular interest. In general, most visitors expressed the willingness to participate in a network like ClusterWIS.

Especially small forest owners see a great advantage in the ability to get forest inventory data in an inexpensive way as such data is usually not easily available without a traditional forest inventory measurement. There is also a demand for lightweight apps and web portals as costs and complexity of traditional desktop software are not feasible for many forest owners. Easy to use and intuitive user interfaces are a must for such apps and portals.

Another aspect favored by some forest owners was the straightforward communication with other actors, which improve their independence and avoids fees for intermediation by third parties.

While the consistent use of end-to-end-encryption was well received, some visitors expressed security concerns regarding the public accessibility of the network. Some kind of "supervisory body" to review registered users was suggested. Some forest owners also fear that the structures of ClusterWIS might promote the dictate of pricing by buyers of timber like saw mills. These concerns deserve particular attention and will be considered in the further course of the ClusterWIS system.

IX. Conclusion

The cluster forestry and wood is an important economic sector. Yet, its major challenges (structural complexity and heterogeneity, huge number of stakeholders, and decentralized processes) are insufficiently addressed in current IT solutions. The ClusterWIS approach can resolve these problems by providing a decentralized, secure, and lean infrastructure for communication and data management. Based on this infrastructure, services and applications are orchestrated to realize novel, interconnected, and sustainable processes for feedstock management among the cluster's actors.

As presented in the paper at hand, in its current phase, most ClusterWIS services and apps are implemented and integrated into a comprehensive demonstrator. The feedback to this demonstrator already shows the high interest of different actors (forest owners, service providers …) in ClusterWIS-based solutions.

Next steps within the project comprise the integration of a payment infrastructure to allow service providers to offer payed services and the implementation of a tree species classification service based on user-provided local samples. Subsequently, the demonstrator will be complemented by further aspects of the reference processes.

Acknowledgment


The research project ClusterWIS is co-financed by the European Union and the German federal state of North Rhine-Westphalia: European Union - Investing in our Future - European Regional Development Fund (EFRE-0800036, EFRE-0800037, EFRE-0800038, EFRE-0800088).

References

[1] J. Rossmann et al., "ClusterWIS - A Decentralized Forest Information and Management System for the Cluster Forestry and Wood," in *The Tenth International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2018) IARIA*, March 25-29, 2018, Rome, Italy, pp. 29-35, ISBN 978-1-61208-617-0, 2018.

[2] Waldbauernverband NRW e.V. (English: Wood owner association of North Rhine-Westphalia), "Waldbauernverband NRW e.V.," URL: http://www.waldbauernverband.de/2016/ [accessed: 2018-11-08].

[3] M. Hoppen, M. Schluse, J. Rossmann, and C. Averdung, "A New nD Temporal Geodata Management Approach using GML," in *GEOProcessing 2015 - The Seventh International Conference on Advanced Geographic Information Systems, Applications, and Services*, Lisbon, Portugal, 2015, pp. 110–116. ISBN 978-1-61208-383-4, Permalink https://www.thinkmind.org/index.php?view=article &articleid=geoprocessing_2015_6_20_30086 [accessed: 2018-11-08]

[4] Kuratorium für Waldarbeit und Forsttechnik e.V. (KWF, English: German Center for Forest Work and Technology), "ELDATsmart," URL: http://www.eldatstandard.de [accessed: 2018-11-08].

[5] Skogforsk, "StanForD," URL: http://www.skogforsk.se/english/projects/stanford [accessed: 2018-11-08].

[6] papiNet Europe/NA, "The intelligent choice…..papiNet," URL: http://www.papinet.org [accessed: 2018-11-08].

[7] M. Kahn and H. Pretzsch, "Parametrisierung und Validierung des Wuchsmodells SILVA 2.2 für Rein- und Mischbestände aus Fichte, Tanne, Kiefer, Buche, Eiche und Erle (English: Parameterisation and validation of the growth model SILVA 2.2. for pure and mixed stands of spruce, fir, pine, beech, oak and alder)," in *Jahrestagung der DVFFA Sektion Ertragskunde*, Kevelaer , 1998.

[8] J. Rossmann, M. Hoppen, and A. Buecken, "Semantic World Modelling and Data Management in a 4D Forest Simulation and Information System," in *ISPRS 8th 3DGeoInfo Conference & WG II/2 Workshop, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Istanbul, 2013, vol. XL-2/W2, pp. 65–72.

[9] P. Krahwinkler, "Machine learning based classification for semantic world modeling : support vector machine based decision tree for single tree level forest species mapping," PhD thesis, RWTH Aachen, 2013.

[10] A. Buecken and J. Rossmann, "Mining for the Timber-Volume for a State-Wide Forest Information System," Intl. LiDAR Mapping Forum 2017, Denver, pp. 1-4, 2017

[11] A. Buecken and J. Rossmann, "Modelling of Forest Landscapes from Remote Sensing LiDAR Data and Aerial Photos," in *Capturing Reality - 3D, Laser Scanning and LiDAR Technologies Forum 2015, 23-25 November, 2015, Salzburg, Austria, pp. 1-6*, 2016.

[12] CPA, "CPA ReDev GmbH," URL: http://www.cpa-redev.de/index.php?lang=e [accessed: 2018-11-08].

[13] P. Biber et al., "How Sensitive are Ecosystem Services in European Forest Landscapes to Silvicultural Treatment?," Forests, 6(5), 1666–1695, 2015. doi: 10.3390/f6051666

[14] S. Hudjetz et al., "Modeling Wood Encroachment in Abandoned Grasslands in the Eifel National Park – Model Description and Testing," PLoS One 9:e113827, 2014. doi: 10.1371/journal.pone.0113827, 2014.

[15] I. Prokofieva, "Payments for Ecosystem Services—the Case of Forests," Current Forestry Reports, 2(2), pp. 130–142, 2014. doi: 10.1007/s40725-016-0037-9

[16] Fordaq, "IHB," URL: http://www.ihb.de [accessed: 2018-11-08].

[17] Wald-wird-mobil.de gGmbH, "Branchenbuch Wald und Forst (English: Yellow Pages Wood and Forest)," URL: http://www.waldhilfe.de [accessed: 2018-11-08].

[18] EGGER, "Der Internet-Marktplatz CoSeDat (English: The internet marketplace CoSeDat)," URL: http://www.cosedat.com [accessed: 2018-11-08].

[19] The Biofore Company UPM, "UPM Customer Online" URL: http://www.upmpaper.com [accessed: 2018-11-08].

[20] H. Jacke, "Abschlussbericht zur Pre-Feasibility-Study "Holztransport und Logistik / Virtueller Betrieb Forst und Holz NRW (English: Final report of the pre feasibility study "wood transport and logistics / virtual enterprise forest and wood NRW")"," Göttingen, 2001.

[21] "Vision, FOCUS - The Project," URL: http://focusnet.eu/about-focus/project-vision [accessed: 2018-11-08].

[22] J.-L. Payeur-Poirier and T. Ahrenholz, "Die Wertschöpfungskette Holz: optimiert vom Wald ins Werk. Ein Holz-Logistikprojekt für den Kleinprivatwald (English: The valued added chain wood: optimized from forest to plant. A wood logistics project for small forest owners)," proWALD, 2(2018), pp. 9-11, 2018.

[23] Nordwestdeutsche Forstliche Versuchsanstalt (NW-FVA, English: Northwest German Forest Research Institute), "Softwareprogramme und Webapplikationen der NW-FVA (English: Software programs and web applications of NW-FVA)," URL: http://www.nw-fva.de/index.php?id=3 [accessed: 2018-11-08].

[24] Scottish Forest & Timber Technologies, "The Scottish Forest and Timber Technologies initiative" URL: http://www.forestryscotland.com [accessed: 2018-11-08].

[25] Wald in Österreich (English: Forest in Austria), "Das Portal zu Wald und Holz (English: Portal to forest and wood)" URL: http://www.wald-in-oesterreich.at [accessed: 2018-11-08].

[26] WaldSchweiz, Verband der Waldeigentümer (English: Suisse wood owner association), "WaldSchweiz" URL: http://www.waldschweiz.ch [accessed: 2018-11-08].

[27] J. Lange, F. Iwanitz, and T. Burke, "OPC From Data Access to Unified Architecture," 2010. ISBN 978-3-8007-3242-5

[28] J. Rossmann, M. Schluse, C. Schlette and R. Waspe, "A New Approach to 3D Simulation Technology as Enabling Technology for eROBOTICS," in 1st International Simulation Tools Conference & EXPO 2013 (SIMEX'2013), 2013, pp. 39-46.

[29] R Development Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0, URL: http://www.R-project.org [accessed: 2018-11-08]

[30] F. Schwaiger, W, Poschenrieder, T. Rötzer, P. Biber, and H. Pretzsch, „Groundwater recharge algorithm for forest management models," Ecological Modelling, 385, 154-164, 2018. doi: 10.1016/j.ecolmodel.2018.07.006

[31] A. Toraño-Caicoya, P. Biber, W. Poschenrieder, F. Schwaiger, and H. Pretzsch, "Forestry projections for species diversity-oriented management: an example from Central Europe," Ecological Processes, 7(23). doi: 10.1186/s13717-018-0135-7

[32] Messe München (English: Trade fair Munich), "INTERFORST | Key Trade Fair for Forestry and Forest Technology," URL: https://www.interforst.com/index-2.html [accessed: 2018-11-08].

Using Dijkstra and Fusion Algorithms to Provide a Smart Proactive mHealth Solution for Saudi Arabia's Emergency Medical Services

Khulood Alghamdi
Medical Education Department, College of Medicine
Information Technology Department, College of Computer and Information Sciences
King Saud University (KSU), Riyadh, KSA
kmalghamdi@ksu.edu.sa

Shada Alsalamah
Information Systems Department
College of Computer and Information Sciences, KSU
Riyadh, KSA
saalsalamah@ksu.edu.sa

Ghada Al-Hudhud
Information Technology Department
College of Computer and Information Sciences, KSU
Riyadh, KSA
galhudhud@ksu.edu.sa

Thamer Nouh
Trauma and Acute Care Surgery Unit
Department of Surgery, College of Medicine, KSU
Riyadh, KSA
tnouh@ksu.edu.sa

Ibrahim Alyahya
Emergency Medical Services
Saudi Red Crescent Authority
Riyadh, KSA
imyahya@srca.org.sa

Sakher AlQahtani
Department of Pediatric Dentistry and Orthodontics
College of Dentistry, KSU
Riyadh, KSA
asakher@ksu.edu.sa

Abstract*—Ambulance diversion in overcrowded hospitals and Emergency Departments (EDs) can have negative consequences on patients' health and safety. Saudi Arabia is thriving to enhance its healthcare sector through an ongoing digitization transition as part of Vision 2030. However, existing infrastructure is still ill-equipped to integrate the distributed siloed systems involved in hospital bed and Emergency Medical Services (EMSs) ecosystem leading to delayed decisions and poor outcomes based on incomplete and inaccurate information. This article addresses this problem by proposing, mEmergency, a smart proactive mobile healthcare solution that connects EMS ambulances with EDs in Riyadh trauma centers to facilitate making real-time informed decisions in a timely fashion to save lives. First, mEmergency predicts inpatient bed capacity using Fusion Algorithm and prioritizes the nearest and most equipped ED for the patient's condition to paramedics using a mHealth application. Second, it uses Dijkstra's Algorithm to route the ambulance path taken to the right ED, while the patient's vital signs are being recorded into the solution along with physical assessment. Finally, mEmergency disseminates the paramedic's assessment information to the ED prior to arrival to optimize time and resources for continuous readiness. mEmergency is an effective solution that should optimize EMS resources, reduce ED crowding, and increase the quality of urgent care services to help stakeholders regain trust in Saudi urgent care service delivery.

Keywords- mHealth; Emergency medical services; Emergency department; Dijkstra algorithm; Fusion algorithm.

I. INTRODUCTION

Ambulance diversion is a strategy often used by overcrowded hospitals and Emergency Departments (EDs) with unavailable beds or resources. EDs that do not have available inpatient beds for emergency patients often have to divert ambulances to other hospitals. This adds to the total time it takes the ambulance to transport the patient, and thus, can have negative consequences on the patient's health and safety, especially in cases of urgent care when every second is crucial and could mean life or death. Shen and Hsia, in the Journal of the American Medical Association [2], report that heart attack victims are more likely to die when their nearest hospital has high diversion rates for more than 12 hours per day, in contrast to other heart attack victims who happen to be near a hospital without diversion or with diversion for less than 12 hours per day [2]. Therefore, to reduce the time it takes to reach the nearest available emergency department, there is an urgent need for a continuous, stable process that connects both the ED and the ambulance services together.

Furthermore, an arriving patient may be declined or placed on hold by Medical Care Utilization (MCU) if there is no bed available. In such cases, the patient may be subject to similar health risks due to the necessity of finding an alternative health care provider or waiting until a bed becomes available. Similarly, an arriving patient may be accepted for treatment but can be accommodated in another MCU (e.g., an arriving obstetrics patient can be boarded in neurosurgery), and hence there is the chance that it may not be possible to move admitted patients from the ED due to an inpatient bed problem. This forces the ED to board admitted patients until inpatient beds are available, effectively reducing the ED's capacity to care for new patients. Boarding of inpatients in the ED has also been cited as the most important determinant of ambulance diversion.

In addition to time and systems integration factors, cyber attacks on healthcare infrastructure for a deliberate interruption of healthcare services could leave hospitals no choice but to divert ambulances. Global attacks on electronic service providers in healthcare, business, and government sectors have been witnessed in recent times, and there has been a wave of deliberate attacks using the WannaCry virus. WannaCry is a Trojan malware virus known as "ransomware." The virus holds the infected computer hostage and demands that the victim pays a ransom to regain access to the files on his/her computer. Ransomware, like WannaCry, works by encrypting most or all of the files on a user's computer. demanding that a ransom be paid in order to have the files decrypted. The episodes of Ransom ware attacks [3] have had a massive scale effect on roughly 150 countries, targeting not only home computers but also healthcare, communications infrastructure, logistics, and government entities for financial gain. This has challenged healthcare providers' deployment of their computerized complex systems to maintain patient records and patient diversion data at critical times, so as to paralyze healthcare services in some targeted health trusts in European countries, such as England's National Health Service (NHS) [3]. Hospitals across England's National Health Service reported that the cyberattack was causing huge problems to their services by affecting X-ray imaging systems, pathology test results, phone systems and patient administration systems" [3].

This article is an extended version of preliminary work published in [1] that aims to address ambulance diversion challenges in Emergency Medical Services (EMSs) in Saudi Arabia by connecting siloed information systems to make informed decisions in a timely fashion. The preliminary work [1] identified the decentralization challenges faced in Saudi's EMS ecosystem due to the fragmentation of data, a lack of communication among stakeholders, and a lack of interoperability between the information systems. This article proposes mEmergency, a practical solution that can reduce ambulance diversion and the time it takes to reach the nearest appropriate ED: one with available inpatient beds and the right resources for that patient's care.

The rest of this paper is organized as follows. Section II introduces the background, while Section III discusses related work. Section IV describes our methodology and structure, and Section V draws our conclusions.

II. BACKGROUND

The Saudi Arabian government has accorded high priority to healthcare services. According to the Saudi Arabian "Basic Law of Governance" [4], the government guarantees the right to healthcare for its citizens and their families. It is responsible for providing public health care services to all Saudi citizens. In recent years, healthcare services in Saudi Arabia have improved tremendously in terms of quantity and quality. This is evident in the literature and government white papers in general, and reflected in the total number of hospitals in Saudi Arabia. The governmental sector owns nearly 70 percent of hospitals, with the rest being operated by the private sector [5]. The number of government hospitals in Saudi increased by 49 hospitals between 2011 and 2015 with a capacity of 69,394 beds.

A. ED Crowding

ED crowding and inpatient bed capacity are pressing problems facing healthcare worldwide. ED Crowding is defined as [6]:

> "*A situation in which the ED function is impeded by the number of patients waiting to be seen, undergoing assessment and treatment, or waiting for departure, exceeding the physical or staffing capacity of the department.*"

Ed crowding should not be managed as a standalone problem. Contributing factors must be examined in order to eliminate the problem. The lack of hospital inpatient bed capacity could lead to ED boarding, which is a significant cause of ED crowding [7]. ED boarding is the practice of keeping patients in the ED waiting area due to the lack of available inpatient beds, even after their admission to the hospital. This results in many issues, including ambulance diversion, extended patient waiting periods, delays in treatment, and longer waiting times for other patients who do not require admission to be treated [8]. Finally, there is an urgent need for a solution to eliminate ED crowding and ambulance diversions resulting from unavailable inpatients bed capacity and ED boarding. This is to provide emergency patients with fast and reliable healthcare and reduce the time it takes the ambulance to reach the nearest available emergency department. To address the above issues, there is a need for a continuous, stable process that requires system-wide support among all healthcare related parties, in order to connect and work with both the emergency departments and the ambulance services together.

B. Hospital Bed-Ecosystem

A hospital bed is not simply a piece of furniture; it is a vital part of hospital infrastructure that enables patient treatment. A hospital bed is an ecosystem that enables delivery of care via trained professionals, managerial staff, equipment and pharmaceuticals (see Figure 1).

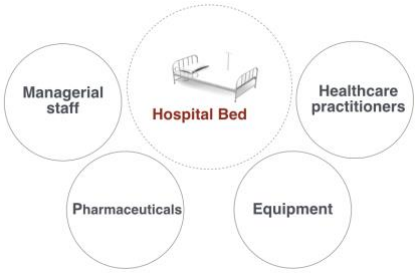


Figure 1. Hospital Bed EcoSystem.

C. Hospital Bed Capacity Planning

Hospital capacity planning is crucial in healthcare. It is essential for managing hospital resources and hospital staff and personnel. In addition, it could be the deciding factor between a patient's life and death. In some countries, such as Finland, New Zealand and Germany, the unit for measuring hospital care and capacity is bed occupancy rate [9], which is defined as "the number of hospital beds occupied by patients expressed as a percentage of the total beds available in the hospital" [10]. This rate remains an essential unit in hospital capacity planning. Nevertheless, using bed numbers and occupancy as a measurement in hospital capacity planning will not foretell the hospital's future demand; neither will it provide a valid estimate of hospital services [9].

In research published by the World Health Organization, researchers propose using strategies that focus on the benefits of using systematic processes in hospital capacity planning. They argue that it is not beneficial to look at the hospital from the perspective of beds and occupancy rates, but rather it is necessary to focus on processes and the path taken by the patients inside the hospital. One of the strategies mentioned is to design hospital flows around "care pathways" instead of counting beds; the strategy works by identifying the variety of pathways the patients take inside the hospital, as well as the factors that can cause delays in patients' treatment, thus identifying bottlenecks [11].

Therefore, the key to successful capacity planning is to try to eliminate any possible future cause of bottlenecks. Sometimes this could be the number of inpatients' available beds, ineffective allocation of existing patients among different medical service units. and sometimes it could be other hospital departments attempting to enhance their performance without realizing how such actions might affect others. Guaranteeing that there are as few bottlenecks as possible will, in turn, result in minimizing delays in patient treatment, separating patients into two streams based on complexity rather than urgency, and as a result creating a fast track for patients who can be treated and discharged more or less immediately [10].

D. International Hospital Statistics

Despite hospital planning strategies, this section highlights some international statistics for acute hospital bed shortages around the globe. In Austria [12], hospital beds have an Average Length of Stay (ALoS) of 18 days or less. This includes some daycare beds. In Germany, acute hospital beds are the beds other than psychiatric and long-term beds. It does not include any daycare beds. In Iceland, acute hospital beds are calculated from bed-days, assuming 90% occupancy rate; beds in medicine and surgeries of leading hospitals and mixed facilities are available in small hospitals that do not include any daycare beds. In Italy, acute hospital beds include inpatient beds of psychiatric hospitals and in-patient beds of psychiatric wards of other hospitals; these do not include any daycare beds. In Spain, acute hospital beds include general hospitals, maternity, other specialized hospitals, and health centers -- no daycare beds. In the UK, acute hospital beds include NHS acute medical, surgical and maternity beds (excluding Northern Ireland) [12] (see Table I below for others country).

TABLE I. DEFINITIONS OF HOSPITAL BEDS IN SELECTED COUNTRIES [12]

Country	Content
Austria	Beds in hospitals with average length of stay of 18 days or less
Germany	Beds other than psychiatric and long term beds
Italy	In-patient beds of psychiatric hospitals and in-patient
United Kingdom	National Health Service acute medical, surgical and maternity beds (excluding Northern Ireland)
Spain	General hospitals, maternity, other specialized hospitals, health centres
Sweden	Beds for short term care run by county councils and three independent communities (short-term includes medical, surgical, miscellaneous medicine/surgery, admission department and intensive care)
Turkey	Public hospitals, health centres, maternity hospitals, cardiovascular and thoracic surgical centres, orthopaedic surgery hospitals

III. LITERATURE REVIEW

A. Bed Capacity Planning Approach

Like any other industry, healthcare faces enormous pressure to improve efficiency and reduce costs. A study by McKinsey [6] points out that the U.S. spends at least $600 - $850 billion on healthcare annually. One area that can be leveraged in healthcare is the support of informed decision-making processes. This is to allow the end user (namely, hospital administrators or clinical managers) to assess the efficiency of existing healthcare delivery systems.

Discrete-Event Simulation (DES) is a widely used technique for the analysis of systems with complex behaviors [14]. DES has been widely applied in healthcare services [14] to study the interrelationships between admission rates, hospital occupancy, and several different policies for allocating beds to MCUs. Lewis D [15] studied bed management in Germany's hospitals, and decision support systems were presented depending on mathematical approaches and computer-based assistance designed to improve the efficiency and effectiveness of admission planning and bed assignment. The study starts by interviewing professionals in bed management to identify aspects that must be respected when developing the decision system, ensuring that patients' treatment priorities and individual preferences are respected [15]. Patient admission and assignment are based on up-to-date and flexible Length of Stay (LoS) estimates, being taken into aggregated contingents of hospital beds, treatment priorities, patient preferences, and a linkage between clinics and wards [15].

The main reason for using DES for modeling a healthcare clinic instead of other mathematical modeling tools (such as linear programming and Markov chain analysis) is the ability to simulate complex patient flows through healthcare clinics, and to play "what if" games by changing the patient flow

rules and policies [15]. Such flows are usually in emergency rooms, where patients can be seen without appointments and require treatment for various sets of ailments and conditions [15]. These disorders can range from mild injuries to serious medical emergencies. Although the number of patients is unpredictable, medical staff can control the treatment by minimizing patient waiting times and increasing staff utilization rates [16].

In emergency rooms, to reduce waiting times of low priority patients, Schmidt et al. [16] analyzed the effects of using a fast track lane. As emergency rooms are prioritized according to the level of patient sickness, low priority patients may have to wait for exceedingly long periods of time [16]. A simulation model is used to classify daily occupancy distributions; it helps in studying the swapping of overflow and bed capacity levels, and it investigates the effects of various changes. ED overcrowding principally results from the incapability of admitted patients to be transferred toward beds in a timely manner [17]. Most experts agree that a greater inpatient capacity is required in order to relieve access block and decrease ED overcrowding [16]. The authors in [16] collected real data from a single month at a single hospital, and a computer model was developed to examine the relationship between admissions, discharges and ED overcrowding (the number of hours admitted patients waited in ED before transfer to an inpatient bed). Meanwhile, authors in [16] proposed a facility location model to locate ER services on a network and determine their respective capacity levels, such that the probability of diverting patients is not larger than a particular threshold.

Author Chin-I Lin in [17] presents a conceptual model of ED overcrowding to help administrators, researchers, and policymakers. The ED conceptual model recognizes at least three general categories of care delivered in the ED: Emergency care, Unscheduled urgent care, and Safety net care. The outputs are patients who are unable to obtain follow-up care and often return to the ED if their condition does not improve or deteriorates [17]. The throughput component of the model identifies the patient length of stay in the ED as a potential contributing factor to ED crowding. The ED crowding is then measured based on two phases; the first phase includes triage, room placement, and the initial provider evaluation. The triage phase is used to objectively identify patients suitable for treatment by emergency nurse practitioners [17]. Since emergency nurse practitioners show high diagnostic accuracy, the emergency nurse practitioner model of care is considered an essential strategy in reducing the LoS of ED patients and may prevent ED crowding [17]. Meanwhile, the second phase of the throughput component includes diagnostic testing and ED treatment.

The input-throughput-output conceptual model of ED crowding may be useful for organizing research, policy, and operations management agenda to alleviate the problem [17]. This model illustrates the need for a systems approach with integrated, rather than piecemeal, solutions for ED crowding [17]. In the study, there are four general areas of ED crowding that require future research.

First, research must consider developing valid and reliable measures of ED crowding. These measures should be sensitive to changes throughout time. Second, research in the field should identify the most critical causes of ED crowding from each component of the model. Third, the effect of ED crowding on the quality of patient care must be assessed. Finally, interventions to reduce ED crowding need to be evaluated.

Ineffective allocation of existing bed capacity among different medical service units can lead to service quality problems for the patients, along with operational and financial inefficiencies for the hospitals [17]. Most health care managers apply relatively simple approaches, such as the use of target occupancy level with an average length of stay, to forecast bed capacity required for a hospital or an MCU [17]. The failure to adequately consider uncertainties associated with patient arrivals and the time needed to treat patients by using such simple approaches may result in bed capacity configurations where a large proportion of patients may have to be turned away. The application of queuing theory allows for the evaluation of the expected (long-run) performance measure of a system by solving the associated set of flow balance equations [17].

Other research considers the hierarchical relationship between care units. For example, after a mother-to-be delivers her child in the labor and delivery unit, she should be moved to the postpartum unit for recovery [17]. If the capability downstream is insufficient, patients must stay within the current care units with typically more costly equipment, thereby reaching capacity limits at these upstream care units [17].

To take the interactions among care units in a hospital into account, C Lin [17] first applies queuing network methodology (without blocking) to discover a balanced bed allotment, which is obtained through trial-and-error work. Then, they use simulation analysis to estimate the blocking behavior and patient sojourn times. The authors [17] develop a mathematical programming formulation to address this problem, and the system uses a different approach by integrating results from queuing theory into an optimization framework. Specifically, the model with each MCU in a hospital has an M/M/c/c queuing system to estimate the probability of rejection when there are beds.

B. International Bed Management System (BMS)

Hospitals should use a BMS that provides a real-time display of hospital occupied beds, along with the available beds, as well as the current status of each one [17]. Therefore, by using a BMS system, the hospital staff would be able to view the status of each bed, whether the bed was occupied, vacant or being prepared for a patient. Also, the hospital staff would be able to view the patient's status; if the patient was going to be discharged, had already left, or was being transferred. Moreover, by using the system, nurses can utilize more of their time in caring for the patient, instead of handling bed assignment tasks manually. This Bed Management Unit used at Alexandra Hospital has provided critical benefits to both patients and staff; indeed, by using the system, patients' waiting times have been decreased by 30 percent [18]. Singapore General Hospital (SGH) is another hospital that has benefited from the use of technology to manage hospital beds.

The SGH uses BMS technology to help improve hospital capacity and care. BMS is a web-based system that allows hospital staff and administration to access information related to patient flow anywhere in the hospital. The BMS user interface has been configured to display the location of patients in the hospital as well as the primary physician. The system also allows the hospital staff to view and track information, records and any specific actions related to the patient's needs, as well as any follow-up movements required. The system gives nurses and bed management staff a full overview of the bed status and patient needs, which allows them to take immediate action.

The way this system works is as follows: when the patient is admitted to the hospital, he/she will receive a Radio Frequency Identification (RFID) tag with a unique identifier that will identify and track the patient's location during his/her stay in the hospital. The system will then search for a bed that best fits the needs of the patient, based on his/her condition before assigning that bed to the patient. The system also uses a real-time location system to help identify the location of the patient through the tag and display; this creates a workflow related to the patient's movements. Hospital departments are also provided with LCD panels that display the BMS dashboard with real-time patient RFID location, which shows the patient's' information and bed statuses automatically in real-time. Once the patient is discharged, the BMS system will notify the housekeeping staff through their PDA to clean and prepare the vacated bed, and once the bed is ready, the housekeeping staff update the bed status [19].

C. Nature of Trauma

In physical medicine, major trauma is an injury or damage to a biological organism caused by physical harm from an external source [29]. Major trauma is also an injury that can potentially lead to long-term severe outcomes, such as chronic pain or other lifelong ailments. There are different types of trauma [29]:

1. Birth trauma: an injury to the infant during the process of being born. In some psychiatric theories, the psychic shock is produced in an infant by the experience of being born [29].
2. Psychic trauma: a psychologically upsetting experience that produces an emotional or mental disorder or otherwise has lasting negative effects on a person's thoughts, feelings, or behavior [29].
3. Risk for trauma: a nursing diagnosis accepted by the North American Nursing Diagnosis Association, defined as accentuated risk of accidental tissue injury, such as a wound, burn, or fracture.

The initial evaluation of a trauma patient is challenging and time-critical, as every minute could be the difference between life and death. Over the past 50 years, the assessment of trauma patients has evolved because of an improved understanding of the distribution of mortality and the mechanisms that contribute to morbidity and mortality in trauma [28]. On the one hand, early deaths may occur in the minutes or hours after the injury. These patients frequently arrive at a hospital before death, which usually occurs because of hemorrhage and cardiovascular collapse [28]. On the other hand, late trauma mortality peaks in the days and weeks after the injury and is primarily due to sepsis and multiple organ failure [28]. Therefore, systems supporting trauma care focus on the treatment of a patient from early trauma mortality, whereas critical care is designed to prevent late trauma mortality. This is the reason why it is essential to find beds prior to trauma cases [29], and therefore, our proposal strives to address this matter.

D. Hospital Bed Capacity Globally and Saudi Vision 2030

Hospitals' provisions for accommodating the increasing number of emergency admissions is a matter of considerable public and political concern and has been the subject of widespread debate [26]. When discussing a hospital's bed capacity, a number of questions are often raised. First, what is a hospital bed? As discussed earlier, a bed is more than an item of furniture on which a patient can lie. For a bed to make any meaningful contribution to a healthcare facility's ability to treat someone, it must be accompanied by an appropriate hospital infrastructure, including trained professional and managerial staff, equipment and pharmaceuticals [13].

For several years, hospital managers have been under pressure to reduce bed capacity and increase occupancy rates for operational efficiency, especially in the Hajj season [26]. More recently, public concern has arisen in cases where patients could not gain access to a local hospital or were subjected to extended delays for the availability of vacant beds [21]. Many countries now struggle to provide cost-effective, quality healthcare services to their citizens [24]. Saudi Arabia has experienced high costs along with concerns about the quality of care in its public facilities [27]. To address these issues, Saudi Arabia is currently restructuring its healthcare system to privatize public hospitals and introduce insurance coverage for both its citizens and foreign workers [27]. These changes provide an exciting and insightful case for the challenges faced when radically changing a country's healthcare system. The situation also demonstrates a unique case in the Middle East for greater reliance on the private sector to address rapidly escalating healthcare costs and deteriorating quality of care [27]. The complexity of changing a healthcare system is discussed with the many challenges associated with the change.

According to Saudi Vision 2030 [23], the healthcare system has benefited from substantive investment in recent decades. As a result, we now have 2.2 hospital beds for every 1,000 people, world-class medical specialists and an average life expectancy rising from 66 years to 74 years in the past three decades [25]. Work is currently underway to build and develop 38 news hospitals with a total capacity of 9,100 beds, in addition to two medical sites accommodating 2,350 beds [25]. During the current fiscal year, 1437/1438, 23 new hospitals (4,250 beds) in various regions across the Kingdom were built [25].

E. Availability of Beds in Saudi Hospital ED

Implementing a Saudi Arabia-wide system would allow a patient's referral from one healthcare provider/facility to another. This includes the ability to electronically transfer patient-related data in either a structured (namely, organized

and well-maintained information that can be obtained in a simple click) or non-structured fashion (namely, data which is laborious to handle). Alternatively, pointers to eHealth accessible data could be used, including patient diagnosis and treatment, referral notes, medication lists, laboratory test results, radiology reports, digital images, audio and video files. This solution would enable integration of information on the availability of the facility, bed, provider or specialty. In addition, such a solution supports optimizing the search for best-fit resource utilization. The proposed work supports:

- Riyadh hospital's bed management program, including automated interfaces with a hospital information system (HIS) "as an element of health informatics that focuses mainly on the administration needs of hospitals."
- Centralized query capabilities for Headquarter and Regional Administrators.
- Operational support for the hospital and PHC practitioners providing patient referrals.
- Support for full inpatient bed management cycle-interface with multiple systems, including registries, HIS and communication systems. Generation of messages to hospital housekeeping.
  Emergency bed requirements and other hospital departments to inform of status - full reporting and analytical capability [25] (see Table II below).

TABLE II. HOSPITALS ESTABLISHED IN SAUDI ARABIA (2010-2013) [25]

Regions	No. of Hospitals Established	No. of Beds
Riyadh	8	1,400
Makkah	8	2,386
Eastern	7	1,150
Al Madinah	5	650
Hail	3	180
Qassim	4	475
Northern Border	3	400
Asir	8	800

F. Regionalization Vs. Bed Capacity

In a healthcare service facility, when an ER is full or all intensive care beds are occupied, hospitals send out a divert status. When a hospital is on divert status, incoming patients might be sent to hospitals which are farther away or kept at the hospitals where they are currently that may not be able to provide adequate service. For a critical trauma victim, the consequence of diverting status can be the difference between life and death. The healthcare service facility attempts to construct a facility location model, which simultaneously determines the number of facilities opened and their particular locations, as well as the capacity levels of the facilities so that the probability that all servers in a facility are busy does not exceed a pre-determined level. In other words, we want to locate ER services on a network and determine their respective capacity levels such that the probability of diverting patients is not larger than a particular threshold. To address the issue, some papers incorporate queuing systems into facility location models to consider the chance of availability of servers and focus on reducing the demand lost due to the shortage of capacity or system congestion.

It is worth mentioning that we try to find similar solutions in this field to compare with our proposal. mEmergency is expected to enhance performance of emergency service as well as to reduce number of deaths number of trauma diversion.

IV. METHODOLOGY

This research uses a mixture of qualitative research methods for data collection and system design. First, a set of semi-structured interviews with different stakeholders (ED and EMS) were conducted to collect data about: a) all Trauma Centers in Riyadh, b) the number of states in emergency and c) data assessment sheet that needs to be sent to hospitals with notification along with patient ID number. After conducting interviews, information about the challenges faced by the ED was collected. Second, soft systems methodology was used to design the proposed system and produce an architectural design.

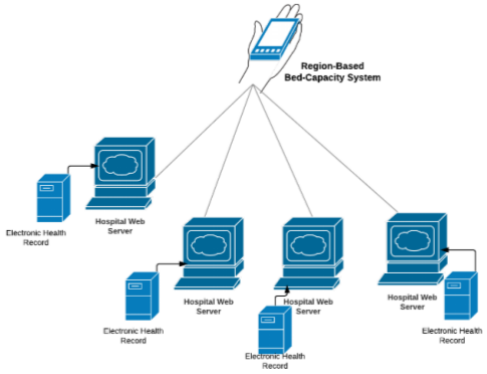


Figure 2. Prposed System Overview.

A. mEmergency System Framework

After interviewing five Saudi Red Crescent Authority representatives (with information technology and medical backgrounds), we came to learn that there is substantial evidence that supports the assertion that hospital beds can reduce deaths during emergencies, and enable treatment of many patients in an emergency by developing the bed capacity system. Therefore, this research decided to develop web services (see in Figure 2) that connect the proposed mobile application with any Electronic Health Record (EHR) in any hospital to show the bed capacity. Furthermore, to learn more about ED process, we met with Dr. Thamer Nouh [30], and discussed eight questions, so that we could then decide criteria relating to emergency department bed capacity in trauma care. We started by basing the criteria on one of the large tertiary hospitals in Riyadh (namely, King Khalid University Hospital).

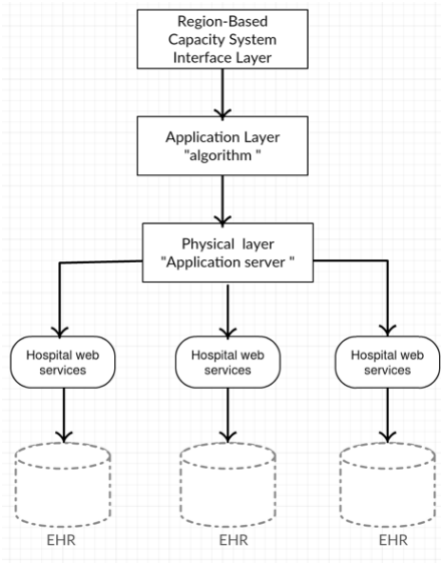


Figure 3. Proposed System Architecture

B. mEmergency System Architecture

mEmergency solution is structured as a multi-layered application (See in Figure 3) consisting of:

- **Interface layer:** This is the top level of the application. The presentation tier displays information related to nearest hospitals. It communicates with the next layer (application layer) by which it puts out the results to the EMS rescuer/ paramedic. In simple terms, it is a layer which users can access directly system's GUI.
- **Application layer:** The algorithm layer is pulled out from the Interface Layer and, as its own layer, it controls an application's functionality by performing detailed processing through follow criteria in section C.
- **Physical layer:** The physical layer includes the data persistence mechanisms (database hospitals servers) and the data access layer that encapsulates the persistence mechanisms and exposes the data. An API is generated by the data access layer to the application layer to find out different methods for managing the available data without creating dependencies on the data storage mechanism.

1) mEmergency Functional Requirements

Based on two interviews with a trauma surgeon who works in King Khalid hospital and workers in EMS [31], the following solutions were proposed:

a) The EMS rescuer/paramedic can view information about hospitals across the city of Riyadh.

b) The information available in the application must be dynamic and in real-time, which means that the paramedic can see the available and current

c) Anywhere across Riyadh.

d) The EMS rescuer/paramedic can view each hospital's exact location, bed capacity, and the available medical resources to suit a patient's case.

e) The paramedic can view available inpatient bed capacity in real-time for the hospitals.

f) The paramedic can find the nearest suitable hospital location for a patient's case with the right medical services for this patient's emergency situation, to guarantee a transfer in a speedy manner that would prevent negative implications to the patient and, at the same time, provide the right health care at the right time in a speedy manner.

g) The Emergency Department Support Officer (EDSO) [32] receives notification of the new patient and can view his/her assessment information.

h) The administrator of EMS can view reports about every EMS rescuer/paramedic case and patient information.

C. mEmergency Workflow Design

Based on the collected data, four criteria are used to decide on bed availability:

1) *Available bed in Radiology department*
2) *Available bed in Intensive Care Unit (ICU)*
3) *Available bed for inpatient*
4) *Available bed in OR*

Another criterion is the availability of specialists in one of the three fields (Orthopedic surgery, Neurosurgery, Emergency surgery; however, the EMS cannot decide which patient needs to go to the radiology department for examination unless the specialist visits the radiology department. Hence, the system will be in green mode, and then the rescuer/paramedic can choose a suitable hospital. Accordingly, ER can be available if there are enough beds in the Radiology department, ICU, inpatient and Operation Room (OR) (see Figure 2).

The system has two components:

1) In the first part of the system, an attempt is made to construct module web services integration with each internal hospital system to decide ability of ER to receive the new patient; this depends on the criteria that are mentioned above.

2) On the other hand, EMS regenerates bed capacity of Riyadh hospitals in trauma cases to decide on a suitable hospital, depending on the following:

- Shortest way from the hospital location that guarantees a speedy transfer that would prevent negative implications.
- Hospitals have the medical resources to treat the patient's case.

- Hospitals have specialized personnel to treat the patient's case.

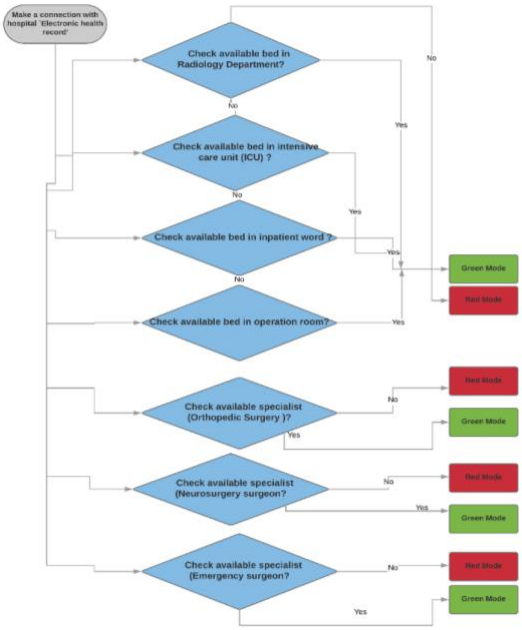


Figure 4. System Framework

D. mEmergency Process Algorithm

The goal of this phase is to build a decision support model that is powerful, robust, comprehensible, optimal, and effective. There is a large number of different search algorithms that can be chosen to run the best diction support model (for example, Fusion [33] and Dijkstra's [34]). The best algorithm to be chosen depends on the collected data.

Fusion Algorithm is used in many tracking and surveillance systems. One method for the design of such systems is to employ a number of sensors (perhaps of different types) and to fuse the information obtained from all these sensors on a central processor. Past efforts to solve this problem required the organization of feedback from the central processor to local processor units. This can be used by the hospitals for proper surveillance of accidents on the roads. It can also reduce the time necessary for bringing the patient to the hospital [32].

Dijkstra's Algorithm is an algorithm for finding the shortest path from a starting node to a target node in a weighted graph. Dijkstra's Algorithm is a graph search algorithm that solves the single-source shortest path problem for a graph with nonnegative edge path costs, producing a shortest path tree. This algorithm is used in routing and as a subroutine in other graph algorithms. It can also be used for finding the costs of shortest paths from a single vertex to a single destination vertex by stopping the algorithm once the shortest path to the destination vertex has been determined. For example, if the vertices of the graph represent cities and edge path costs represent driving distances between pairs of cities connected by a direct road, then Dijkstra's Algorithm can be used to find the shortest route between one city and all other cities. It can also be used by hospitals for ambulances in case of emergency to find the shortest available path. The algorithm creates a tree of shortest paths from the starting vertex, the source, to all other points in the graph. The algorithm exists in many variants; Dijkstra's original variant found the shortest path between two nodes, but a more common variant fixes a single node as the "source" node and finds the shortest paths from the source to all other nodes in the graph, producing a shortest-path tree [33]. For this, the following are points which are necessary for Dijkstra's Algorithm in hospitals [34]:

- The existence of a widespread road system that connects all parts of the city.
- Availability of sufficient VANET modules in the routes in order to detect traffic congestion.
- Infrastructure, such as GPS, communication links and two-way radio are provided.
- Presence of a Dispatch Centre (DC) that serves the purpose of information exchange.
- The existence of an updated database of the roads and hospitals.
- The existence of Road Side Units (RSU) at suitable locations which might be inaccessible due to restrictions for the propagation of the signal.

Algorithm of Decision Support Model

```
Initialize arrivaltime to time start to enter
data
Initialize senddatatime to time submit
assessment sheet
Initialize responsetime to time hospital receive
data
Initialize location to location of patient
Initialize Empty Nearestlist
Initialize Empty Availablelist
Initialize Empty Availablespecialistlist
   For each hospital around Location value
   Collocate distance
Set all Hospital to Nearestlist Item arrange by
Short distance
   For each Hospital In Nearestlist
   Connect with Bed Capacity Web Service
Call and set value of Available Bed
Call and set value of  available  Orthopedic
Surgery specialist
Call and set value of  available  Neuro surgery
specialist
Call and set value of available Emergency
surgeon specialist
   For each hospital in Nearestlist
        If Available Bed is Not False
        Set Hospital in Availablelist
        Print Availablelist
        Delete Hospital from Nearestlist
   For each Hospital In Availablelist
        If available Orthopedic Surgery
        specialist OR available Neuro surgery
        specialist OR Emergency surgeon
        specialist
        Set Hospital in Availablespecialistlist
        Delete Hospital from
        Availablespecialistlist
        Print Availablespecialistlist
        Else
        Print Nearestlist
```

Algorithm of Bed Capacity Web Service

```
Initialize available bed to False
Initialize available Orthopedic Surgery
specialist to False
Initialize available Neuro surgery specialist to
False
Initialize available Emergency surgeon
specialist to False
Connect Electronic Health Record
Call Available bed in Radiology department
Call Available bed in  intensive care unit (ICU)
Call Available bed for inpatient
Call Available bed in OR
Call available Orthopedic Surgery specialist
Call available Neuro surgery specialist
Call available Emergency surgeon specialist
   If Available bed in Radiology department is
   Not False AND Available bed in intensive care
   unit (ICU) is Not False AND
   Available bed for inpatient is Not False AND
   Available bed in OR is Not False
   Set available bed True
   Print available bed
   Else
   Set available bed False
   Print available bed
   If Orthopedic Surgery specialist is Not False
   Set Orthopedic Surgery specialist True
   Print available Orthopedic Surgery specialist
   Else
   Set Orthopedic Surgery specialist False
   Print available Orthopedic Surgery specialist
   If Neuro surgery specialist is Not False
   Set Neuro surgery specialist True
   Print available Neuro surgery specialist
   Else
   Set Neuro surgery specialist False
   Print Neuro surgery specialist
   If Emergency surgeon specialist is Not False
   Set Emergency surgeon specialist True
   Print available Emergency surgeon specialist
   Else
   Set Emergency surgeon specialist False
Print available Emergency
```

E. mEmergency Interface Design

In the proposed mHealth application, there are three potential users: rescuer or paramedic front-end, EDSO, and EMS administrator for the back-end. The rescuer or paramedic is the primary user of the mHealth application. They can view a list of suitable nearest hospitals then choose one of them; they can also send the trauma victims assessment sheet containing vital information on patients to the chosen hospital. An EMS dashboard administrator can view reports about every rescue/paramedics, and we can also search by rescue/paramedics name, code number or hospital name (see Figure 5). On the other hand, the EDSO can view the assessment sheet of patients before they arrive (see Figure 6).

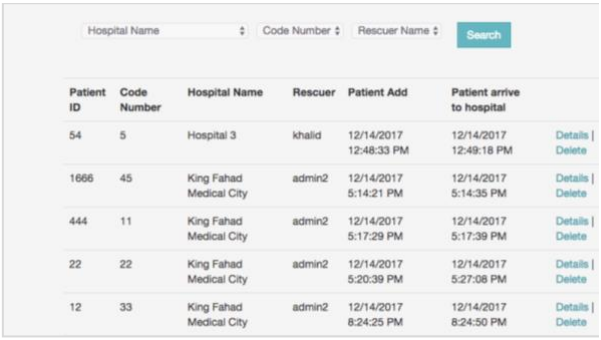


Figure 5. Hospital Bed Capacity Webpage Interface.

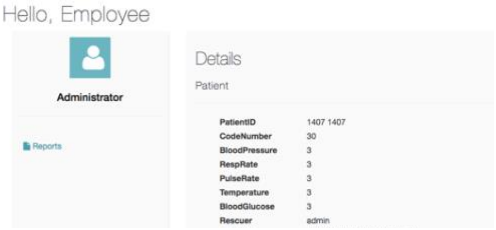


Figure 6. mEmergency Dashboard Interface.

An aggregated result illustrated in Figure 7 below shows all suitable hospitals prioritization by location and available beds. This is a color-coded scheme of availability: Red icon: Not Available, and Green icon: Available.

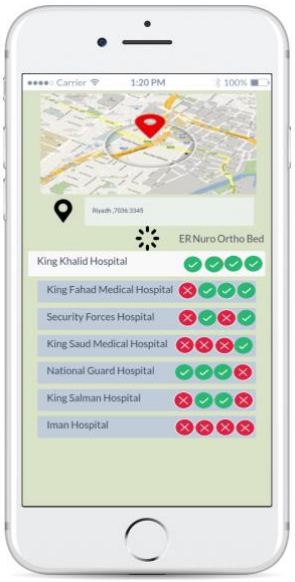


Figure 7. mEmergency prioritisation of nearest and most-equipped trauma EDs to the patient's case.

- First icon from the right is: Available bed.
- Second icon from right is: Available Orthopedic Surgery specialist.
- Third icon from right is: Available Neuro Surgery specialist.
- Fourth icon from right is: Available Emergency Surgery specialist.

The screen in Figure 8 appears after choosing the suitable hospital in Figure 7, so the EMS paramedics can record the patient's assessment data for sharing with the selected hospital prior to arrival to optimize its resources and enhance its readiness.

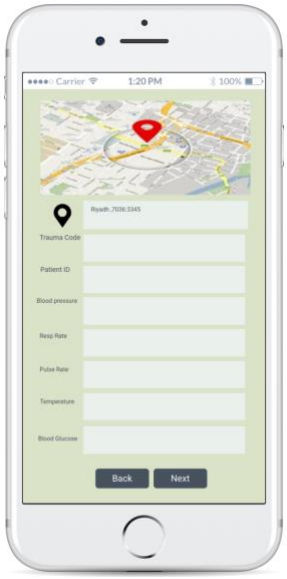


Figure 8. mEmergency records the patient's assessment data for sharing with the selected hospital prior to arrival.

- A "fields set" of patient's attributes should be completed.
- A Next button transfers the user to the next page (confirmed page).
- A Back button transfers the user to the previous screen.

V. Discussion and Conclusion

Overcrowded hospitals with limited resources in Saudi Arabia are left with no option but to divert ambulances, which can be the leading cause of death in some trauma cases. The work presented in this article proposes an effective mEmergency solution to this problem. First, it links Emergency Medical Services (EMSs) with trauma centers in Riyadh to help the paramedic deliver patients to the nearest trauma center's Emergency Department (ED) with available resources based on informed decisions in the shortest possible time to save their life. This is achieved by predicting inpatient bed capacity in real-time using Fusion Algorithm, while routing the EMS ambulance path taken using Dijkstra's Algorithm. In addition, mEmergency, optimizes the selected trauma center's resources and enhances its readiness by the sharing of patient's vital signs and physical assessment data prior his/her arrival to reduce the time for preparation prior to trauma victim arrival. The evaluation results of the solution show the capability of providing regional-based availabilities of resources in nearest hospitals in order to avoid ED crowding and shortages in inpatient bed capacity. Ultimately, mEmergency should help optimize EMSs resources in Riyadh city to improve the ambulance diversion issue in particular and the quality of urgent care service delivery to Saudis in general.

References

[1] K. Alghamdi, S. Alsalamah, G. Al-Hudhud,T. Nouh, I. Alyahya and S. Alqahtani. "Region-Based Bed Capacity mHealth Application for Emergency Medical Services: Saudi Arabia Case Study", *eTELEMED*, 2018, p.114

[2] Y. Shen, "Association Between Ambulance Diversion and Survival Among Patients With Acute Myocardial Infarction", *JAMA*, vol. 305, no. 23, p. 2440, 2011.

[3] "WannaCry ransomware: Everything you need to know", CNET, 2018. [Online]. Available: https://www.cnet.com/news/wannacry-wannacrypt-uiwix-ransomware-everything-you-need-to-know/. [Accessed: 23- Nov- 2018].

[4] Ministry of Foreign Affairs. "The Basic Law of Governance" Kingdom of Saudi. 1992.

[5] Ministry of Health, "Kingdom of Saudi Arabia - Ministry of Health Portal", 2017. [Online]. Available: http://www.moh.gov.sa/en/Ministry/Statistics/book/Pages/default.aspx. [Accessed: 29- Nov- 2018].

[6] Indexmundi. "Hospital bed density - Country Comparison", 2017. [Online]. Available: http://www.indexmundi.com/g/r.aspx?v=2227. [Accessed: 29- Nov- 2018].

[7] Van der Linden, M. C. "Emergency department crowding: Factors influencing flow", 9789461089007, 2015.

[8] National Association of EMS Physici, "Ambulance Diversion and Emergency Department Offload Delay", Prehospital Emergency Care, vol. 15, no. 4, pp. 543-543, 2011.

[9] Encyclopedia. "bed occupancy - Dictionary definition of bed occupancy", Encyclopedia.com, 2017. [Online]. Available: http://www.encyclopedia.com/caregiving/dictionaries-thesauruses-pictures-and-press-releases/bed-occupancy. [Accessed: 23- Nov- 2018].

[10] ACEP, "Definition of Boarded Patient", 2017. [Online]. Available: https://www.acep.org/Clinical---Practice-Management/Definition-of-Boarded-Patient-2147469010/. [Accessed: 23- Nov- 2018].

[11] B. Rechel, S. Wright, J. Barlow and M. McKee, "Hospital capacity planning: from measuring stocks to modelling flows," Bulletin of the World Health Organization, vol. 88, no. 8, pp. 632-636, 2010.

[12] World Health Organization, "Millennium Development Goals (MDGs), 2017. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs290/en/. [Accessed: 23- Nov- 2018].

[13] M. McKee, "What are the lessons learnt by countries that have had dramatic reductions of their hospital bed capacity?" HEN synthesis report on reduction of hospital beds", Health Evidence Network (HEN), 2003.

[14] Jun, S. Jacobson and J. Swisher, "Application of Discrete-Event Simulation in Health Care Clinics: A Survey", The Journal of the Operational Research Society, vol. 50, no. 2, p. 109, 1999.

[15] Dawn M. Lewis," A Qualitative Case Study: Hospital Emergency Preparedness Coordinators' Perspectives Of Preparing For And Responding To Incidents", Capella University, 2015

[16] R. Schmidt, S. Geisler and C. Spreckelsen, "Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources", BMC Medical Informatics and Decision Making, vol. 13, no. 1, 2013.

[17] Chin-I Lin. "Optimization Models For Capacity Planning In Health Care Delivery." (Doctoral Dissertation). University of Florida, Gainesville, USA. 2008.

[18] B. Asplin, D. Magid, K. Rhodes, L. Solberg, N. Lurie and C. Camargo, "A conceptual model of emergency department

crowding", Annals of Emergency Medicine, vol. 42, no. 2, pp. 173-180, 2003.

[19] Cisco Systems. "Singapore Hospital Modernizes Bed Management." [Accessed: 7- Nov- 2018] [Online]. Available: http://www.cisco.com/c/dam/en_us/solutions/industries/docs/healthcare/AlexandraHospital.pdf. 2006.

[20] Groundbreaking Bed Management Technology Helps SGH Improves Capacity and Care. [Online]. Available: https://www.sgh.com.sg/about-us/newsroom/news-release/2010/Pages/GroundbreakingBedManagementTechnologyHelpsSGHImprovesCapacityandCare.aspx. [Accessed: 30- Nov- 2018].

[21] M. Ribbe, G. Ljunggren, K. Steel, E. Topinkova, C. Hawes, N. Ikegami, J. Henrard and P. JONnson, "Nursing Homes in 10 Nations: A Comparison Between Countries and Settings", *Age and Ageing*, vol. 26, no. 2, pp. 3-12, 1997

[22] Kerr E, Sibrand V. Health care systems in transition: Belgium. Copenhagen, European Observatory on Health Care Systems, 2000.

[23] S. Hussain, "Full Text of Saudi Arabia's Vision 2030 | Saudi Gazette", Saudi Gazette, 2017. [Online]. Available: http://saudigazette.com.sa/saudi-arabia/full-text-saudi-arabias-vision-2030/. [Accessed: 23- Nov- 2018].

[24] Ministry of Finance. "Ministry of Finance 2017 Budget of Kingdom of Saudi Arabia." [Online]. Available: https://www.mof.gov.sa/en/budget2017/Documents/The_National_Budget.pdf. [Accessed: 23- Nov- 2018].

[25] Ministry of Health. "Achievement of Ministry of health 2013", 2017. [Online]. Available: http://www.moh.gov.sa/en/ministry/statistics/book/documents/1433.pdf [Accessed: 21- Nov- 2018].

[26] Ministry of Health. "News - Ajyad Hospital in Makkah Shows Continuous Readiness for the Hajj Season", Moh.gov.sa, 2017. [Online]. Available: http://www.moh.gov.sa/en/hajj/news/pages/news-2014-09-22-002.aspx. [Accessed: 23- Nov- 2018].

[27] S. Walston, Y. Al-Harbi and B. Al-Omar, "The Changing Face of Healthcare in Saudi Arabia", Annals of Saudi Medicine, vol. 28, no. 4, pp. 243-250, 2008.

[28] A. Bagust, M. Place and J. Posnett, "Dynamics of bed use in accommodating emergency admissions: stochastic simulation model", BMJ, vol. 319, no. 7203, pp. 155-158, 1999.

[29] Emedicine.medscape. "Initial Evaluation of the Trauma Patient: Overview, Triage and Organization of Care, Initial Assessment", Emedicine.medscape.com, 2017. [Online]. Available: http://emedicine.medscape.com/article/434707-overview. [Accessed: 23- Nov- 2018].

[30] Dr. Thamer Nouh, King Khaild Hospital,Riyadh, 2017.

[31] M. Aouad,Saudi Red Crescent Authority, Riyadh, 2017.

[32] "Emergency Department Support Officer", Nswhealth.erecruit.com.au, 2017. [Online]. Available: http://nswhealth.erecruit.com.au/ViewPosition.aspx?Id=270427. [Accessed: 23- Nov- 2018].

[33] Yuemin Li, Renbiao Wu, Tao Zhang, "Multi-scale fusion and estimation for multi-resolution sensors", Signal Processing (ICSP) 2012 IEEE 11th International Conference on, vol. 1, pp. 193-197, 2012, ISSN 2164-523X.

[34] P. Singal and R. R.S.Chhillar, "Dijkstra Shortest Path Algorithm using Global Position System", International Journal of Computer Applications, vol. 101, no. 6, pp. 12-18, 2014.

Virtual Network Function Use Cases Implemented on SONATA Framework

Cosmin Conțu, Andra Țapu, Eugen Borcoci
University POLITEHNICA of Bucharest – UPB
Bucharest, Romania
Emails: cosmin.contu@elcom.pub.ro, andratapu@elcom.pub.ro, eugen.borcoci@elcom.pub.ro

***Abstract*—Network Function Virtualization (NFV) is a recent and powerful technology capable to support the development of flexible and customizable virtual networks and services, including sliced networks, in multi-tenant, multi-operator and multi-domain environment. Open research issues still exist for architectural, interoperability, design and also related to implementation and experimental aspects. Among several NFV-oriented projects, Service Programming and Orchestration for Virtualized Software Networks (SONATA) is a representative framework. This paper is an extended version of a previous work and develops several Virtual Network Functions (VNF) on SONATA framework. The examples of VNFs include virtual hosts, HTTP server, firewall and a graph of virtual routers. They have been integrated in separate topologies and then chained together into a more complex topology and have been tested using SONATA basis.**

Keywords-Network Function Virtualization; Software Defined Networking; Cloud computing; SONATA; Containernet; Docker.

I. INTRODUCTION

This paper is an extended version of the work [1] [2] and is dedicated to further develop several use cases implementations of *Network Virtual Functions (VNF)* in NFV architecture, illustrating the capability of chaining different functions.

After 2014, *Network Functions Virtualization* (NFV) started to be investigated and developed; today it is recognized as a powerful concept, as well as architecture and technology. It aims to solve some of the current telecommunication world limitations, problems and challenges, like large number of proprietary hardware appliances dedicated to specific services, lack of flexibility and dynamicity, low interoperability, high capital and operational expenditures: capital expenditure (CAPEX), operational expenditure (OPEX), energy consumption and installation space issues [3][4]. Currently, NFV is also seen as a supporting technology in cloud/edge computing domains. NFV decouples the hardware appliances from the network functions that are running over them, by using generic hardware (servers, storage and switches) and running the network functions over virtual machines installed on this generic equipment.

Based on virtualization technologies, NFV allows faster development and deployment (compared to traditional approach) of services composed of network functions that can be implemented in virtualized way. Different virtualized network functions can be defined, instantiated, deployed or moved, while sharing the same infrastructure. They can be created, modified and deleted without needing to physically visit a site to change the hardware supporting those network functions.

The operators' CAPEX and OPEX can be reduced, due to software development (taking advantage of the growing IT industry). Energy consumption reduction is also possible, if a clever power management and migration plan for the virtual machines (VM) is designed.

Software Defined Networking (SDN) [5] is a complementary technology to NFV. The main SDN concept of separating the control plane from the data/user plane creates high flexibility, programmability and network technology abstraction. This approach offers powerful capabilities for the management and control functions. While independent of each other, SDN and NFV can cooperate in order to construct powerful and flexible systems in cloud computing and networking areas. In a general view, NFV and SDN can be seen as "orthogonal": while SDN separates the control and data plane, NFV approach can be used both in the control or data plane to implement different control or forwarding functions as virtual ones.

According to ETSI [6][7], the NFV architecture (see Figure 1) is divided in two main parts: operational and management. The operational part is composed of the functional blocks: *Network Function Virtualization Infrastructure (NFVI)* which contains the physical resources and their abstraction (virtual resources constructed by a virtualization layer); Virtual Network Functions (VNF) which defines different functions that can be composed in services; Operations and Business Support Systems (OSS/BSS). The management part is represented by the Management and Orchestration (MANO) which provides the orchestration and the Life Cycle Management (LCM) of the network functions and infrastructure. Each of the three subsystems in the operational part has a corresponding manager in MANO.

Numerous studies, realizations, projects, proofs of concepts and demos are currently developed in NFV, SDN areas [8][9]. However, many still open research issues exist for such technologies related to different aspects: architectural but also related to use cases, service creation and composition, manageability and resource allocation, virtualization methods, performance obtained in dynamic and mobile environment, scalability, implementation aspects and selection of the software technologies applicable, multi-tenant and multi-domain capabilities and security.

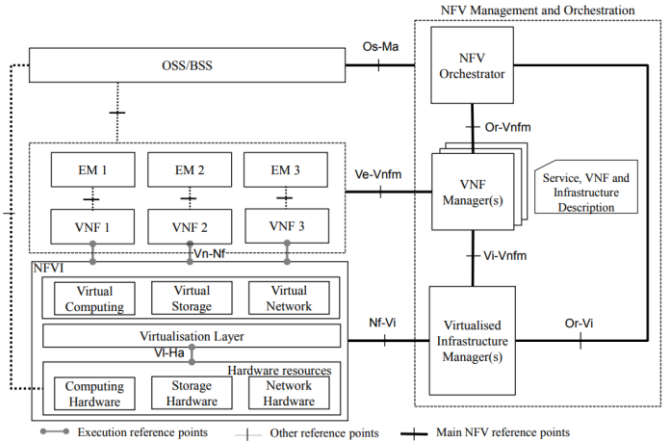


Figure 1. ETSI NFV reference architectural framework [6]

In terms of Development and Operations (DevOps), several problems are recognized to exist [10], like: SDN/NFV infrastructures are not yet stable; Virtual Network Functions (VNFs) are not sufficiently interoperable with orchestrators; multi-vendor environments are not certified; the number of services for which the SDN/NFV framework brings very strong benefits in marketplaces is not yet so large; SDN/NFV combination is difficult and does not offer easy E2E multi-site support; frequently, there is a need for some additional development; key features like network slicing are not yet completely clarified; auto VNF scalability, SP recursiveness, VNF intelligent placements, security, etc., are other open research issues.

Therefore, more extensive experiments with SDN/NFV frameworks are necessary to further clarify different development aspects.

The EU H2020 project Service Programming and Orchestration for Virtualized Software Networks (SONATA) [11] is a relevant example and offers a framework allowing DevOps oriented to SDN/NFV area.

The main purpose of this paper is to further develop experiments started in [1][2], based on SONATA framework in order to more deeply understand the capabilities of the framework, to test its scalability for using it to develop and test some custom VNFs. This work is an additional step to achieve the goal of creating a network service package which will be uploaded and deployed on SONATA platform.

The paper is organized as follows. Section II is an overview of related work. Section III shortly presents the architecture of SONATA framework. Section IV contains the results of the experiments done with SONATA framework and all the steps taken. Section V presents conclusions and future work.

II. RELATED WORK

This section shortly presents a selective view on some related work dedicated to service development and orchestration in virtualized networks and its relation to SONATA architecture, when applicable. It is split in brief overview firstly on EU-funded collaborative projects, open-source solutions and commercial solution provided.

UNIFY [12] (*EU-funded Collaborative Projects*) proposes an architecture which is similar to those of ETSI-MANO and *Open Networking Foundation* (ONF)-SDN. Its objective is to reduce operational costs by removing the need for costly onsite hardware upgrades, taking advantage of SDN and NFV. Across the infrastructure one can develop networking, storage and computing components, through a service abstraction model. The UNIFY global orchestrator consists of algorithms used for optimization of elementary service components across the infrastructure. The project exposes the fact that all resource orchestration related functionalities existing in a distributed way in the MANO SONATA framework, can be logically centralized, when there is an abstraction combination of compute, network and storage resources.

Even if the main idea of a recursive service platform is applicable both for UNIFY and SONATA, the implementation is different. First, the recursiveness in UNIFY is obtained as a repeatable orchestration layer for each infrastructure design, while within SONATA is implemented as a repeated deployment of a complete SONATA platform. Another difference is related to the service specific functionality: in UNIFY it is added by developer inside a Control Network Function (NF), as a dedicated part of the Service Graph, running in the infrastructure; in SONATA the service functionality is obtained using plugins in the service platform which means that it is not mandatory to be on the same infrastructure where the Virtual Network Function (VNF) is running.

OpenStack [13][26] is an *open source project*, mainly written in Python, that provides an *Infrastructure as-a-Service* solution through a variety of loosely coupled services. Each service offers an API that facilitates the integration. Due to its variety of components, the current version of the OpenStack not only provides a pure *Virtual Infrastructure Manager* (VIM) implementation, but spans various parts of the ETSI-NFV architecture. OpenStack is made up of many different moving parts. Because of its open nature, additional components can be joined to OpenStack in order to meet specific needs. *OpenStack Keystone* [14], for instance, offers authentication and authorization not only to the VIM part, but it can be integrated to other services as well. *OpenStack Ceilometer* [15] provides a pluggable monitoring infrastructure that consolidates various monitoring information from various sources and makes the available to OpenStack users and other services. *OpenStack Tacker* [16] aims at the management and orchestration functionality described by ETSI-NFV.

The overall architecture relies on message buses to interconnect the various OpenStack components. To this end, OpenStack uses the *Advanced Message Queuing Protocol* (AMQP) [17] as messaging technology and an AMQP broker, namely either RabbitMQ [18] or Qpid [19], which sits between any two components and allows them to communicate in a loosely coupled fashion. More precisely, OpenStack components use *Remote Procedure Calls*

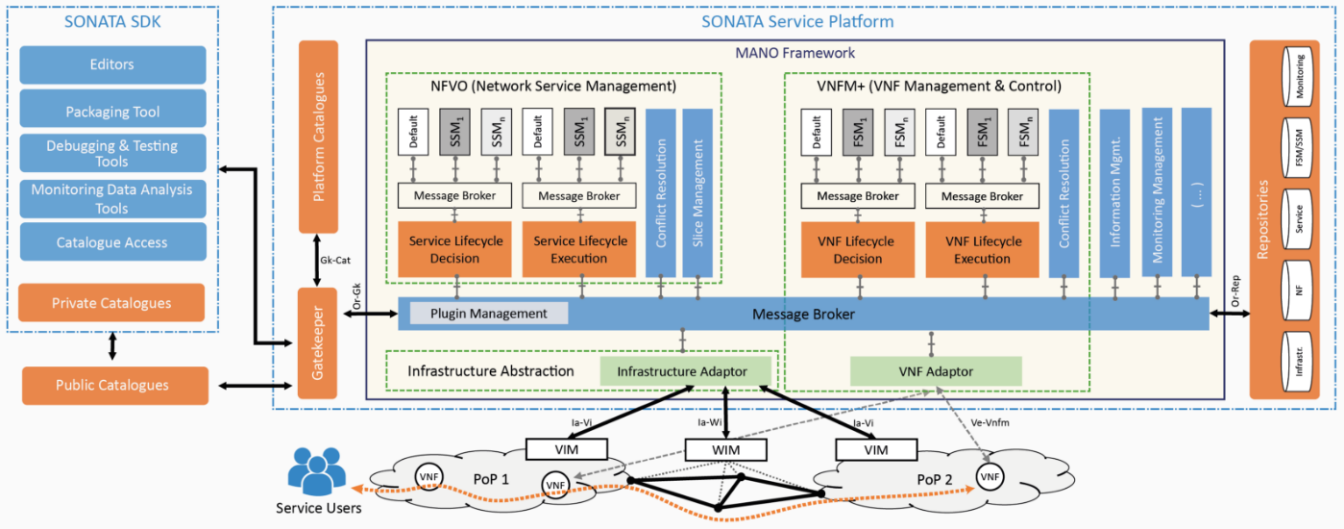


Figure 2. SONATA Framework [21]

(RPCs) to communicate to one another. The OpenStack architecture has been proven to be scalable and flexible. Therefore, it could act as a blueprint for the SONATA architecture.

From SONATA's perspective, OpenStack is used as being supportive and complementary. For the SONATA developers there is the need to have access to a running OpenStack installation to use the capabilities of a VIM for running services from the Service Platform.

Another option for service developers when it comes to SONATA is the SONATA's emulation platform to locally prototype and test complete network service chains in realistic end-to-end scenarios. The emulator of SONATA supports OpenStack-like API endpoints to allow carrier-grade MANO stacks (SONATA, Open Source MANO) to control the emulated VIMs.

To raise their NFV holding**,** commercial vendors have started to market solutions for the orchestration layer. Even if they created their own NFV context, the first generation of NFV Orchestrators (NFVO) is based off ETSI MANO specifications. But there are also several orchestration solutions developed by established network vendors to further expand a larger NFV ecosystem [20].

From SONATA's perspective, the NFV orchestration concept meets the commercial solutions from the following points: to the complete VNF and network service lifecycles, including onboarding, test and validation, scaling, assurance and maintenance. Vendor marketing material and white papers present their upcoming products as holistic solutions for both service and network orchestration, compatible with current ETSI MANO specifications.

These orchestration solutions are commonly part of a fully integrated NFV management platform, including NFVO, VNFM and extended services such as enhanced monitoring and analytics. For example, IBM's SmartCloud Orchestrator can be integrated with its counterpart solutions, SmartCloud Monitoring and IBM Netcool Network Management System, providing an end-to-end offering.

For this paper, SONATA framework was chosen due to its platform which follows the DevOps approach as well as for its Software Development Tools which help the developers to design, create, debug and analyze network services.

III. SONATA FRAMEWORK

In order to make this paper enough self-contained, this section very shortly presents the SONATA framework architecture [25] along with its objectives, use cases and features along with its correspondence with ETSI NFV framework.

SONATA main goal is to develop a NFV framework that provides to third party developers a programming model and a suite of tool for virtualized services integrated with an orchestration system. SONATA allows the developers to achieve a lower time-to-market of networked services, to optimize and reduce the costs of network services (NS) deployment and to speed-up the integration of software networks in telecommunication industry.

Figure 2 presents the general architecture of SONATA framework which complies with and builds upon the ETSI reference architecture for NFV MANO, contains the following components [27]:

- Service Platform (SP)
- Software Development Kit (SDK)
- Catalogues containing different system artefacts.

The Service Platform (SP) is responsible for management and control of network functions and services and it has the same role as MANO block from ETSI model. It is a modular and customizable environment in which the platform operators can create specific platforms appropriate for their business model, by replacing components of MANO plugins. SP has the following core components: gatekeeper, catalogues and repository, MANO Framework (NFVO and VNFM) and infrastructure abstraction.

MANO framework is the core of SP and provides the management for complex NSs for their entire lifecycle. Same way as developed in ETSI, MANO consists of NFVO and VNFM blocks.

The operations from SONATA involved into lifecycle management are split in two types: service-level and function-level operations, which together represent functionalities of NFVO and VNFM from ETSI's reference architecture. The NFVO in SONATA, as in ETSI model, orchestrates the NFVI resources and manages the lifecycles of network services. The VNFM manages the lifecycle of one or multiple VNF instances of same or different types.

Both in SONATA and ETSI, the VIM controls and manages the virtualized resources (network, storage and compute) in an operator's infrastructure domain. (NFVI-PoP). A VIM can handle a specific or multiple type of NFVI resources. Generally the VIM contains the following functional blocks:

A specialized VIM, called WAN infrastructure Manager (WIM) is used to provide connectivity between endpoints in different NFVI-PoPs.

The catalogues and repositories of SONATA consist of network function and services information like code, executables, configuration data and other requirements. These catalogues are divided into:

1) private (located in SDK and used to store locally developed network services per developer or per project).

2) service platform (holds the data which operates and run network services that can be instantiated using SP. Actually in SP there are same types of catalogues and repositories as defined in ETSI (NS and VNF catalogues; NFV instances and NFVI resources repositories) but also two extra: Service specific managers (SSMs) / Function Specific Managers (FSMs) / catalogue and SSF/FSM repository which offers flexibility for service developers to customize their own services.

3) public (representing the third-party network services which are ready to be used by the service developers and SP operators).

In addition, to MANO block of the ETSI model, in SONATA there can be found the gatekeeper component which is responsible of validating the network services posted into SP in a form of packages by mediating between development and operational tasks [30].

Therefore, the service platform follows the ETSI NFV reference design (see Figure 3) but in the same time adds its own extensions which facilitates multi tenancy support by allowing resource slicing which can be mapped to tenants exclusively.

Comparing to ETSI model, SONATA adds SDK as a new important architectural component. The SDK helps the third-party developers to create complex services composed of multiple VNFs, with a set of software tools and also supports service providers to deploy and manage their created NSs on multiple SONATA SPs.

SDK contains different tools to: generate network functions; emulates trail of services; debug and monitoring; support for DevOps operations of network services [28].

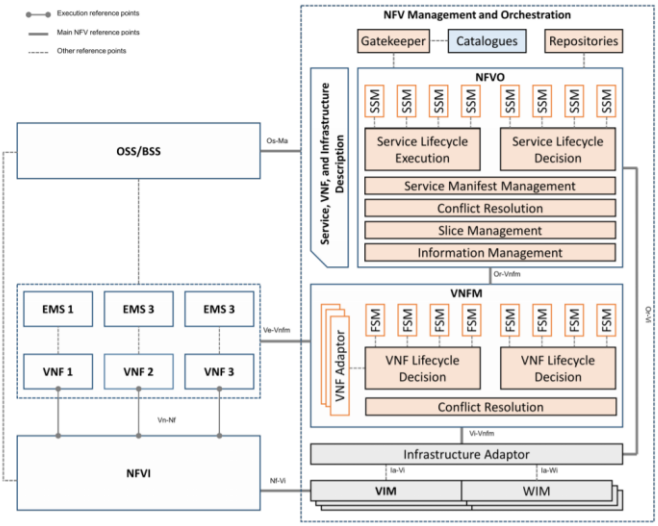


Figure 3. Comparative view: ETSI and SONATA [21]

IV. EXPERIMENTS WITH SONATA

This section presents NFV experiments whose purpose is to test the functionality of different VNFs in various topologies using SONATA framework.

These topologies are represented as custom emulated networks which use Docker [22] containers as compute instances to run VNFs. Moreover, these experiments are developed around SONATA framework and using some specific tools as:

a) Virtual Machine (VM): the experiments are running on a VM of 80GB storage on a 64 -bit Ubuntu distribution ready to use which has been downloaded from SONATA repository [20]

b) Containernet [23]: it is a ramification of Mininet network emulator which allows the developer to create network topologies using Docker containers.

c) Open-source utilities: to create and test the VNFs needed in the proposed topologies, the following collection of utilities has been used: *"iptables"[24], "iproute", "bridge-utils", "traceroute", "inetutils-ping"; "curl"; "squid"; "apache"*.

d) SONATA emulator (son-emu): this is a part of SONATA SDK and it is based on MeDICINE emulation platform. MeDICINE is intended for service developers who can create network service chains and then test them in realistic emulated environments.

A. *(UC1) Simple Virtual Hosts Experiment*

a) *Main objectives:* create two virtual hosts and test their inter-communication.

b) Topology: the topology depicted in Figure 4 contains data centers (DC) in terms of point of presence (PoP) which can be defined as specific emulated hardware by installing docker images which contain the VNFs. In this simple experiment two DCs have been defined, created and used as following:

- Two hosts (dc1 and dc2)

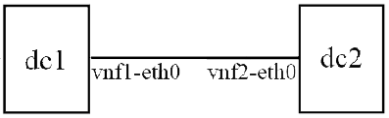


Figure 4. vHosts Experiment Topology

c) *Configuration, tests and results:* first step was to deploy the topology and then instantiate and start the VNFs on each DC (see Figure 5).

```
root@demo:/home/sonata# son-emu-cli compute list
+--------------+-------------+--------------------+------------------+
| Datacenter   | Container   | Image              | Interface list   |
+==============+=============+====================+==================+
| dc2          | vnf2        | vhost-iptables-img | vnf2-eth0        |
+--------------+-------------+--------------------+------------------+
| dc1          | vnf1        | vhost-iptables-img | vnf1-eth0        |
+--------------+-------------+--------------------+------------------+
```

Figure 5. vHosts Experiment compute list

First, using *ifconfig* command on both dc1 and dc2, it can be seen that the IP addresses are in the same network. The IP addresses were set for each data center during the instantiation (see Figures 6 and 7). Afterwards, the "*son-emu-cli network add*" command is invoked, in order to establish the connection between the two datacenters.

```
containernet> vnf1 ifconfig vnf1-eth0
vnf1-eth0 Link encap:Ethernet  HWaddr ca:45:be:12:cb:6f
          inet addr:10.0.0.1  Bcast:10.255.255.255  Mask:255.0.0.0
```

Figure 6. vHosts Experiment ifconfig command on vnf1

```
containernet> vnf2 ifconfig vnf2-eth0
vnf2-eth0 Link encap:Ethernet  HWaddr 32:af:20:7c:4d:2b
          inet addr:10.0.0.2  Bcast:10.255.255.255  Mask:255.0.0.0
```

Figure 7. vHosts Experiment ifconfig command on vnf2

To check the connectivity between the two data centers, the "ping" command is used from both datacenters (see Figures 8 and 9).

```
containernet> vnf1 ping -c3 10.0.0.2
PING 10.0.0.2 (10.0.0.2): 56 data bytes
64 bytes from 10.0.0.2: icmp_seq=0 ttl=64 time=41.388 ms
64 bytes from 10.0.0.2: icmp_seq=1 ttl=64 time=20.545 ms
64 bytes from 10.0.0.2: icmp_seq=2 ttl=64 time=21.240 ms
--- 10.0.0.2 ping statistics ---
3 packets transmitted, 3 packets received, 0% packet loss
round-trip min/avg/max/stddev = 20.545/27.724/41.388/9.666 ms
```

Figure 8. vHosts Experiment ping command from vnf1 to vnf2

```
containernet> vnf2 ping -c3 10.0.0.1
PING 10.0.0.1 (10.0.0.1): 56 data bytes
64 bytes from 10.0.0.1: icmp_seq=0 ttl=64 time=21.580 ms
64 bytes from 10.0.0.1: icmp_seq=1 ttl=64 time=21.156 ms
64 bytes from 10.0.0.1: icmp_seq=2 ttl=64 time=21.199 ms
--- 10.0.0.1 ping statistics ---
3 packets transmitted, 3 packets received, 0% packet loss
round-trip min/avg/max/stddev = 21.156/21.312/21.580/0.191 ms
```

Figure 9. vHosts Experiment ping command from vnf2 to vnf1

B. (UC2) Virtual HTTP Server Experiment

a) *Main objectives:* create a virtual HTTP server which can be accessed from a different host from the same network.

b) Topology: in this topology two DCs have been used (see Figure 10) as following:

- One host (dc1)
- HTTP server (dc2)

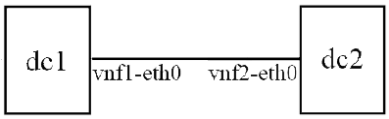


Figure 10. vHTTP_Server Experiment Topology

The two datacenter are in the same network (10.0.0.0/8), dc1 has the IP address 10.0.0.1/8 and dc2 has 10.0.0.2/8. Using apache, a HTTP Server VNF was installed on dc2. Dc1 was used as a virtual host.

c) *Configuration, tests and results:* first step was to deploy the topology and then instantiate and start the VNFs on each DC as can be seen in Figure 11.

```
root@demo:/home/sonata# son-emu-cli compute list
+--------------+-------------+---------------+------------------+
--+
| Datacenter   | Container   | Image         | Interface list   |
  |
+==============+=============+===============+==================+
==+
| dc2          | vnf2        | vhttp-img-new | vnf2-eth0        |
  |
+--------------+-------------+---------------+------------------+
--+
| dc1          | vnf1        | vhost-img-new | vnf1-eth0        |
  |
+--------------+-------------+---------------+------------------+
```

Figure 11. vHTTP_Server Experiment compute list

After the instantiation, the ping command is called between vnf1 and vnf2 in order to verify the connectivity between the two datacenters (see Figure 12).

```
containernet> vnf1 ping -c3 10.0.0.2
PING 10.0.0.2 (10.0.0.2): 56 data bytes
64 bytes from 10.0.0.2: icmp_seq=0 ttl=64 time=21.548 ms
64 bytes from 10.0.0.2: icmp_seq=1 ttl=64 time=20.832 ms
64 bytes from 10.0.0.2: icmp_seq=2 ttl=64 time=20.361 ms
--- 10.0.0.2 ping statistics ---
3 packets transmitted, 3 packets received, 0% packet loss
round-trip min/avg/max/stddev = 20.361/20.914/21.548/0.488 ms
```

Figure 12. vHTTP_Server Experiment ping command from vnf1 to vnf2

On the HTTP Server a resource (html file) will be created in order to be accessed by the virtual host using http protocol. Using *"more"* command on HTTP Server, it can be seen the content of the html file (see Figure 13).

```
containernet> vnf2 more /var/www/site/example.html
root@vnf2:/#
<!DOCTYPE html>
<html>
<head>
        <title>HTTP Call Succeded!</title>
</head>
<body>
<p>HTTP Server Call worked!</p>
</body>
</html>
```

Figure 13. vHTTP_Server Experiment checking the local file from vnf2

With the *"curl"* command called from vnf1, the resource is fetched from vnf2 (see Figure 14) and it can be seen that it the same html file that was previously created on vnf2, as expected.

```
containernet> vnf1 curl 10.0.0.2:80/example.html
root@vnf1:/#
<!DOCTYPE html>
<html>
<head>
        <title>HTTP Call Succeded!</title>
</head>
<body>
<p>HTTP Server Call worked!</p>
</body>
</html>
```

Figure 14. vHTTP_Server Experiment curl command from vnf1 to vnf2 (http server)

C. (UC3) Virtual Firewall Experiment

a) *Main objectives:* create a virtual firewall which has the purpose to block –if requested - some particular traffic flows between two hosts.

b) Topology (Figure 15): it consists of three DCs have been used as following:

- Two hosts (dc1 and dc2)
- Firewall (dc3)

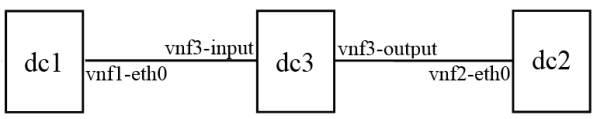


Figure 15. vFw Experiment Topology

The subnet 10.0.0.0/8 has been used together with the *"bridge-utils"* utility on dc3 to make the communication between dc1 and dc2 possible. Utility *"iptables"* has been used to create the *"DROP"* rule for the traffic which is forwarded by dc3.

c) *Configuration, tests and results:* first step was to deploy the topology and then instantiate and start the VNFs on each DC as can be seen in Figure 16.

```
root@demo:/home/sonata# son-emu-cli compute list
+--------------+-------------+-----------------+-----------------+
| Datacenter   | Container   | Image           | Interface list  |
+==============+=============+=================+=================+
| dc2          | vnf2        | ubuntu:trusty   | vnf2-eth0       |
+--------------+-------------+-----------------+-----------------+
| dc3          | vnf3        | vfw-iptables-img | input,output    |
+--------------+-------------+-----------------+-----------------+
| dc1          | vnf1        | ubuntu:trusty   | vnf1-eth0       |
+--------------+-------------+-----------------+-----------------+
```

Figure 16. vFw Experiment compute list

Further, the "DROP" rule has been added for vnf3 and the connectivity between the two hosts (vnf1 with 10.0.0.7 on interface vnf1-eth0 and vnf2 with 10.0.0.5 on interface eth2) has been tested.

If the "DROP" rule is removed, it can be seen in Figure 17 that the two hosts can communicate with each other:

```
containernet> vnf3 iptables -D FORWARD -j DROP
containernet> vnf1 ping -c3 vnf2
PING 10.0.0.5 (10.0.0.5) 56(84) bytes of data.
64 bytes from 10.0.0.5: icmp_seq=1 ttl=64 time=61.7 ms
64 bytes from 10.0.0.5: icmp_seq=2 ttl=64 time=62.0 ms
64 bytes from 10.0.0.5: icmp_seq=3 ttl=64 time=62.0 ms

--- 10.0.0.5 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 61.781/61.947/62.045/0.118 ms
```

Figure 17. vFw Experiment ping without "DROP" rule

When "DROP" rule is added then the whole traffic between the 2 hosts does not exist anymore. This rule is exposed in Figure 18.

```
containernet> vnf3 iptables -A FORWARD -j DROP
containernet> vnf1 ping -c3 vnf2
PING 10.0.0.5 (10.0.0.5) 56(84) bytes of data.
--- 10.0.0.5 ping statistics ---
3 packets transmitted, 0 received, 100% packet loss, time 2025ms
```

Figure 18. vFw Experiment ping with "DROP" rule

D. (UC4) Virtual Routers Graph Experiment

a) *Main objectives:* create a small network of virtual routers which will forward traffic through a network graph between three hosts from three different subnets.

b) Topology (Figure 19): it consists of six DCs using two different docker images, one for the virtual routers and another for virtual hosts.

- Three hosts (dc1, dc2 and dc3)
- Three routers (dc4, dc5 and dc6)

Routing tables (containing static routes) have been made for the entire topology using *"iproute"* utility. The hosts are assigned within the subnets 11.0.0.0/8, 12.0.0.0/8, 13.0.0.0/8 and the subnets between routers are 10.0.0.0/8 (dc4-dc5), 20.0.0.0/8 (dc5-dc6) and 30.0.0.0/8 (dc4-dc6).

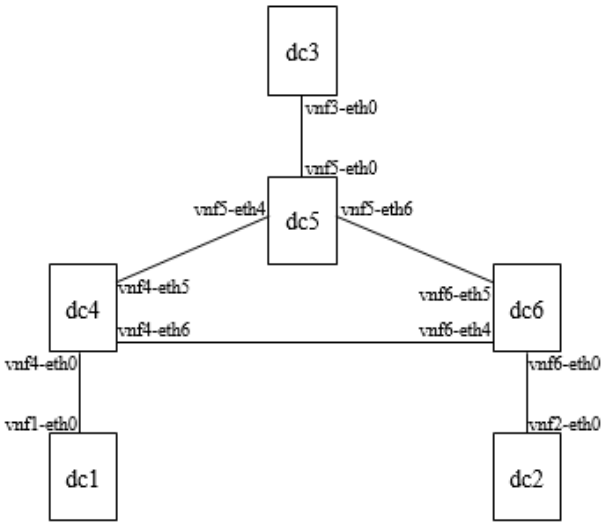


Figure 19. vRouters Graph Experiment Topology

c) *Configuration, tests and results:* after deploying the topology, the VNFs were instantiated and started on each DC and the links between them were also added as illustrated in Figure 20.

```
root@demo:/home/sonata# son-emu-cli compute list
+-------------+------------+------------------+-----------------------------+
---------+
| Datacenter  | Container  | Image            | Interface list              |
        |
+=============+============+==================+=============================+
=========+
| dc6         | vnf6       | vnat-iptables-img | vnf6-eth0,vnf6-eth5,vnf6-eth4 |
.s1-eth5 |
+-------------+------------+------------------+-----------------------------+
---------+
| dc4         | vnf4       | vnat-iptables-img | vnf4-eth0,vnf4-eth5,vnf4-eth6 |
.s1-eth5 |
+-------------+------------+------------------+-----------------------------+
---------+
| dc5         | vnf5       | vnat-iptables-img | vnf5-eth0,vnf5-eth4,vnf5-eth6 |
.s1-eth6 |
+-------------+------------+------------------+-----------------------------+
---------+
| dc2         | vnf2       | vhost-iptables-img | vnf2-eth0                  |
        |
+-------------+------------+------------------+-----------------------------+
---------+
| dc3         | vnf3       | vhost-iptables-img | vnf3-eth0                  |
        |
+-------------+------------+------------------+-----------------------------+
---------+
| dc1         | vnf1       | vhost-iptables-img | vnf1-eth0                  |
        |
+-------------+------------+------------------+-----------------------------+
```

Figure 20. vRouters Graph Experiment compute list

Another way to visualize, as in Figure 21, and monitor the state of the topology and output of *son-emu-cli* is through web-based emulator dashboard.

In this example, the dc4 vRouter has two routes to dc6, with different generic metrics:

- via interface vnf4-eth5, with metric 20;

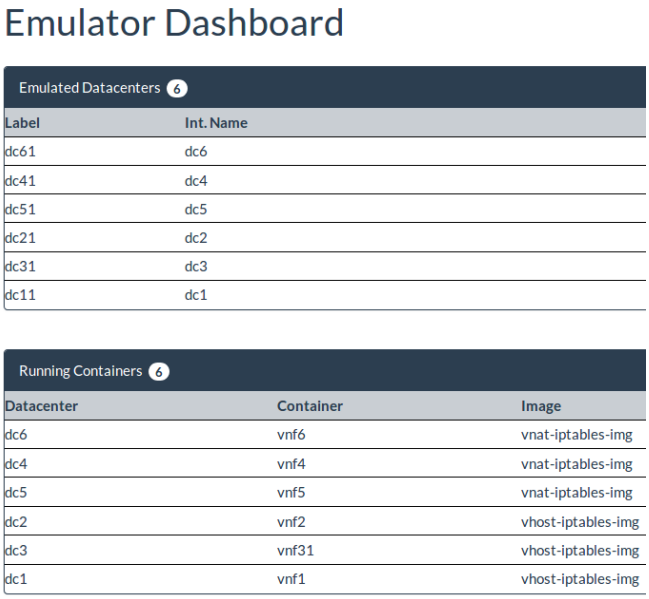


Figure 21. vRouter Graph Experiment emulator dashboard (partial view)

- via vnf4-eth6 with metric 10 (same settings were made respectively on dc6 since static routing is in place).

A shortest path route selection is supposed.

To verify the functionality of the experiment, a traceroute between dc1 and dc2 hosts has been made and it can be seen in Figure 22 that the traffic has been forwarded through the route with the lowest metric (10).

```
containernet> vnf1 traceroute vnf2
traceroute to 12.0.0.1 (12.0.0.1), 30 hops max, 60 byte packets
1  11.0.0.2 (11.0.0.2)  20.589 ms  20.560 ms  20.552 ms
2  30.0.0.2 (30.0.0.2)  82.123 ms  82.116 ms  82.109 ms
3  12.0.0.1 (12.0.0.1)  123.580 ms  123.574 ms  123.569 ms
```

Figure 22. vRouters Graph Experiment traceroute metric 10

If the interface vnf6-eth4 is down and the link between dc4 and dc6 is stopped, it can be observed in Figure 23 that traffic will be forwarded through the route with metric 20 (the only one now remained) when a traceroute between dc1 and dc2 is made again.

```
containernet> vnf6 ifconfig vnf6-eth4 down
containernet> vnf1 traceroute vnf2
traceroute to 12.0.0.1 (12.0.0.1), 30 hops max, 60 byte packets
1  11.0.0.2 (11.0.0.2)  22.001 ms  22.025 ms  22.028 ms
2  10.0.0.2 (10.0.0.2)  43.172 ms  43.165 ms  43.157 ms
3  20.0.0.1 (20.0.0.1)  84.176 ms  84.168 ms  84.160 ms
4  12.0.0.1 (12.0.0.1)  125.179 ms  125.171 ms  125.163 ms
```

Figure 23. vRouters Graph Experiment traceroute metric 20

Although the above experiments are rather simple, they illustrate a complete implementation successful sequence of steps, i.e., to define, instantiate and then run VNF-based

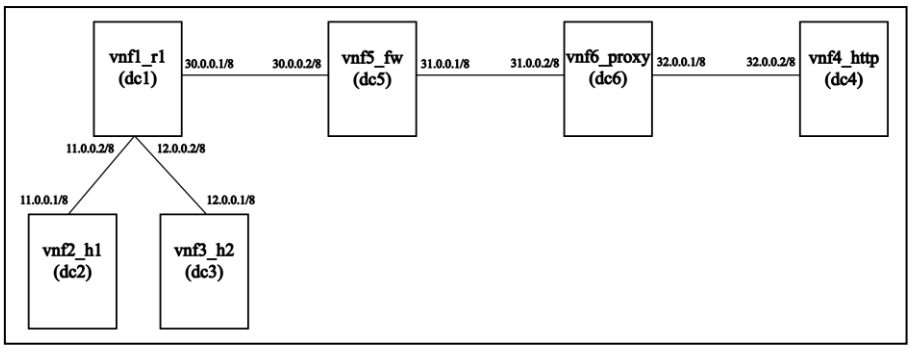


Figure 24. SONATA Framework [21]

topologies on the complex SONATA framework. Modification of the operational parameters is also demonstrated.

E. Multiple Chained VNFs Experiment

1) Main objectives: create a network topology whose purpose is to instantiate a chain of VNFs with SONATA platform. These VNFs roles are : hosts, routers, firewall, proxy, http server, all virtual.

2) Topology: the topology contains data centers (DC) in terms of point of presence (PoP) which can be defined as specific emulated hardware by installing docker images which contain the VNFs. In this experiment (Figure 24) there have been used six DCs as following:

a) Two hosts (vnf2_h1 and vnf3_h2).
b) One router (vnf1_r1).
c) One firewall (vnf5_fw).
d) One proxy server (vnf6_proxy).
e) One http server (vnf4_http).

Routing tables (containing static routes) have been made for the entire topology using *"iproute"* utility. The hosts are assigned within the subnets 11.0.0.0/8, 12.0.0.0/8; the subnet between router and firewall is 30.0.0.0/8; between firewall and proxy is 31.0.0.0/8 and between proxy and http server is 32.0.0.0/8.

3) Configuration, tests and results.

a) First step was to deploy the topology and then instantiate and start the VNFs on each DC as can be seen on Figure 25.

```
root@demo:/home/sonata# son-emu-cli compute list
+--------------+------------+-----------------+-----------------------------+
| Datacenter   | Container  | Image           | Interface list              |
+==============+============+=================+=============================+
| dc6          | vnf6_proxy | vproxy-img-new  | vnf6-eth5,vnf6-eth4         |
+--------------+------------+-----------------+-----------------------------+
| dc4          | vnf4_http  | vhttp-img-new   | vnf4-eth6                   |
+--------------+------------+-----------------+-----------------------------+
| dc5          | vnf5_fw    | vfw-iptables-img| vnf5-eth1,vnf5-eth6         |
+--------------+------------+-----------------+-----------------------------+
| dc2          | vnf2_h1    | vhost-img-new   | vnf2-eth0                   |
+--------------+------------+-----------------+-----------------------------+
| dc3          | vnf3_h2    | vhost-img-new   | vnf3-eth0                   |
+--------------+------------+-----------------+-----------------------------+
| dc1          | vnf1_r1    | vnat-iptables-img | vnf1-eth2,vnf1-eth3,vnf1-eth5 |
+--------------+------------+-----------------+-----------------------------+
```

Figure 25. Topology compute list

b) Second step is meant to prove that the routing is working, and it has been tested with "traceroute" utility between vnf2_h1 and vnf6_proxy (Figure 26).

```
root@vnf2_h1:/# traceroute 31.0.0.2
traceroute to 31.0.0.2 (31.0.0.2), 30 hops max, 60 byte packets
 1  11.0.0.2 (11.0.0.2)  22.552 ms  22.884 ms  22.872 ms
 2  30.0.0.2 (30.0.0.2)  146.628 ms  146.606 ms  146.582 ms
 3  31.0.0.2 (31.0.0.2)  186.763 ms  186.741 ms  186.719 ms
```

Figure 26. Traceroute command between vnf2_h1 and vnf6_proxy

c) Third step represents the functionality of the VNF proxy squid which acts as an intermediary passing the clients (vnf2_h1 and vnf3_h2) requests to the http server (vnf4_http). In the presence of the proxy server, there is no direct communication between the clients h1 and h2 and the http server (Figure 27).

```
root@vnf2_h1:/# ping -c3 32.0.0.2
PING 32.0.0.2 (32.0.0.2): 56 data bytes
--- 32.0.0.2 ping statistics ---
3 packets transmitted, 0 packets received, 100% packet loss
```

Figure 27. Ping command between vnf2_h1 and vnf4_http

Instead, the client connects to the proxy server and sends requests for a resource file that resides on http server (Figure 28).

```
root@vnf2_h1:/# ping 31.0.0.2
PING 31.0.0.2 (31.0.0.2): 56 data bytes
64 bytes from 31.0.0.2: icmp_seq=0 ttl=62 time=187.837 ms
64 bytes from 31.0.0.2: icmp_seq=1 ttl=62 time=185.059 ms
64 bytes from 31.0.0.2: icmp_seq=2 ttl=62 time=185.054 ms
^C--- 31.0.0.2 ping statistics ---
3 packets transmitted, 3 packets received, 0% packet loss
round-trip min/avg/max/stddev = 185.054/185.983/187.837/1.311 ms
```

Figure 28. Ping command between vnf2_h1 and vnf6_proxy

The proxy server handles this request by fetching (with the *"curl"* command) the required resource (proxy_test.html file) from the http server and forwarding the same to the client (Figure 29).

```
root@vnf3_h2:/# curl -x http://31.0.0.2:3128 32.0.0.2:80/proxy_test.html
<html>
<body>
<p>Success! HTTP Server call using Proxy VNF worked!</p>
</body>
</html>
```

Figure 29. Curl command from vnf3_h2 to vnf6_proxy

d) Last step is meant to show the functionality of firewall VNF as it blocks the TCP traffic between h2 (12.0.0.1) and http server through proxy but it allows the rest of traffic, for example ICMP (Figures 30, 31 and 32).

```
root@vnf5_fw:/# iptables -L
Chain INPUT (policy ACCEPT)
target     prot opt source               destination

Chain FORWARD (policy ACCEPT)
target     prot opt source               destination
DROP       tcp  --  12.0.0.1             anywhere

Chain OUTPUT (policy ACCEPT)
target     prot opt source               destination
```

Figure 30. Drop rule added on vnf5_fw

```
root@vnf3_h2:/# curl -x 31.0.0.2:3128 32.0.0.2:80/proxy_test.html
curl: (7) Failed to connect to 31.0.0.2 port 3128: Connection timed out
```

Figure 31. Curl command from vnf3_h2 to vnf6_proxy after adding the Drop rule

```
root@vnf3_h2:/# ping 31.0.0.2
PING 31.0.0.2 (31.0.0.2): 56 data bytes
64 bytes from 31.0.0.2: icmp_seq=0 ttl=62 time=224.957 ms
64 bytes from 31.0.0.2: icmp_seq=1 ttl=62 time=224.172 ms
^C--- 31.0.0.2 ping statistics ---
2 packets transmitted, 2 packets received, 0% packet loss
round-trip min/avg/max/stddev = 224.172/224.565/224.957/0.392 ms
```

Figure 32. Ping command from vnf3_h2 to vnf6_proxy after adding the Drop rule

The above experiment proves the capability of SONATA to emulate a multiple chained VNFs complex topology. All VNFs successfully communicated with each other, creating a functional, complex network topology.

V. CONCLUSIONS

This paper presented the results of multiple experiments with different VNFs treated separately and a more complex topology containing a multiple chained VNFs using the emulator (part of the SDK) from the SONATA framework.

The single-VNF experiments: virtual hosts which demonstrates the connectivity between two datacenters, virtual http server from which a html file was fetched, virtual firewall to filter the traffic between two hosts and virtual routers that are able to route traffic between several networks, were successfully completed.

Using all the above VNFs, together with a proxy VNF, a more complex topology was created, having as a goal to prove that the functionality of all the VNFs from the VNF chain are preserved and can work together.

The tests have successfully proved that the access to http server through proxy server worked without a known route and also that firewall filtered the inbound traffic to proxy by blocking a certain network.

This paper accomplished the proposed objective to successfully test various single-VNF and multiple chained VNF topologies using the emulator from SONATA SDK.

As future work, new experiments will be done by creating "network service packages" containing multiple chained VNFs and having a lifecycle management, coordinated by the MANO framework. The network service packages will be uploaded and tested with the Service Platform, part of SONATA framework.

Further work should be developed in attempt to solve still existing, specific open issues of the complex SONATA framework, like those mentioned in [29]: NFV Orchestration development- including LCMs for virtual functions and Service Specific Managers (SSM); interfacing the SDK to the Service Platform; scalability and flexibility of the monitoring framework; network slicing capabilities of SONATA (concept still not clear); cooperation of the service platform with recursive architectures, service function chaining via Docker- based VIM (not yet mature); continuous integration and delivery (CI/CD) methodology and others.

REFERENCES

[1] A. Ţapu, C. Conţu, E. Borcoci, "Network Function Virtualization Experiments using SONATA Framework", The International Symposium on Advances in Software Defined Networking and Network Functions Virtualization SOFTNETWORKING 2018.

[2] A. Ţapu, C. Conţu, E. Borcoci, "Multiple Chained Virtual Network Functions Experiments with SONATA Emulator", The Twelfth International Conference on Communications COMM 2018.

[3] NFV White paper: "Network Functions Virtualisation, An Introduction, Benefits, Enablers, Challenges & Call for Action. Issue 1". Available from: https://portal.etsi.org/NFV/NFV_White_Paper.pdf [retrieved: February, 2018].

[4] R. Mijumbi et al., "Network function virtualization: State-of-the-art and research challenges", IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 236-262, 1st Quart. 2016.

[5] B. N. Astuto, M. Mendonca, X. N. Nguyen, K. Obraczka, and T. Turletti, "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks", Communications Surveys and Tutorials, IEEE Communications Society, (IEEE), 2014, 16 (3), pp. 1617 – 1634.

[6] NFV White paper: "Network Functions Virtualisation (NFV) ,Network Operator Perspectives on Industry Progress. Issue 1".Available from: https://portal.etsi.org/NFV/NFV_White_Paper2.pdf [retrieved: February, 2018].

[7] ETSI GS NFV 002: "Network Functions Virtualisation (NFV); Architectural Framework". Available from: http://www.etsi.org/deliver/etsi_gs/NFV/001_099/002/01.02.01_60/gs_NFV002v010201p.pdf [retrieved: February, 2018].

[8] S. Van Rossem et al, "Deploying elastic routing capability in an sdn/nfv-enabled environment", 2015 IEEE Conference on Network Function Virtualization and Software Defined Network, pp. 22-24, 2015.

[9] ETSI Plugtests Report: "1st ETSI NFV Plugtests, Madrid, Spain, 23rd January–3rd February". Available from: https://portal.etsi.org/Portals/0/TBpages/CTI/Docs/1st_ETSI_NFV_Plugtests_Report_v1.0.0.pdf [retrieved: February, 2018].

[10] J.Martrat, "SONATA approach towards DevOps in 5G Networks", SDN World Congress, 2017, Hague. Available

from: http://sonata-nfv.eu/content/sonata-approach-towards-devops-5g-networks-0 [retrieved: February, 2018].

[11] S. Dräxler, H. Karl, M. Peuster, H. R. Kouchaksaraei, M. Bredel, J. Lessmann, T. Soenen, W. Tavernier, S. Mendel-Brin, and G. Xilouris, "Sonata: Service programming and orchestration for virtualized software networks," in 2017 IEEE International Conference on Communications Workshops (ICC Workshops), May 2017, pp. 973–978

[12] Mario Kind et al. "Deliverable 2.2: Final Architecture". Available from: https://www.fp7-unify.eu/files/fp7-unify-eu-docs/Results/Deliverables/UNIFY%20Deliverable%202.2%20Final%20Architecture.pdf [retrieved: February, 2018].

[13] The OpenStack Project. OpenStack: The Open Source Cloud Operating System. Available from: http://www.openstack.org/ [retrieved: February, 2018].

[14] The OpenStack Project. Openstack keystone developer. Available from: http://www.openstack.org/developer/keystone [retrieved: February, 2018].

[15] The OpenStack Project. Openstack ceilometer developer. Available from: http://docs.openstack.org/developer/ceilometer [retrieved: February, 2018].

[16] The OpenStack Project. Openstack tacker: An open nfv orchestrator on top of openstack. Available from: https://wiki.openstack.org/wiki/Tacker [retrieved: February, 2018].

[17] OASIS. Advanced messaging queuing protocol. Available from: https://www.amqp.org/ [retrieved: February, 2018].

[18] Pivotal Software. RabbitMq - Messaging. Available from: https://www.rabbitmq.com [retrieved: February, 2018].

[19] Apache Software Foundation. Qpid.Available from: https://qpid.apache.org/ [retrieved: February, 2018].

[20] Containernet and SONATA Emulator Demo. Available from: https://github.com/sonata-nfv/son-tutorials/tree/master/upb-containernet-emulator-summerschool-demo [retrieved: February, 2018].

[21] SONATA. D2.2 Architecture Design.Available from: http://sonata-nfv.eu/sites/default/files/sonata/public/content-files/pages/SONATA_D2.2_Architecture_and_Design.pdf [retrieved: February, 2018].

[22] Docker - Build, Ship, and Run Any App, Anywhere. Available from: https://www.docker.com/ [retrieved: February, 2018].

[23] Containernet. Available from: https://containernet.github.io/ [retrieved: February, 2018].

[24] The netfilter.org "iptables" project.Available from: http://netfilter.org/projects/iptables/ [retrieved: February, 2018].

[25] M. Peuster, H. Karl, and S. v. Rossem: "MeDICINE: Rapid Prototyping of Production-Ready Network Services in Multi-PoP Environments". IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN), Palo Alto, CA, USA, pp. 148-153. doi: 10.1109/NFV-SDN.2016.7919490. (2016)

[26] Daniel Grzonka, "The Analysis of OpenStack Cloud Computing Platform: Features and Performance" in Journal of Telecommunicationation Technology, 3/2015, pp. 52-57.

[27] Sevil Draxle et al., "SONATA: Service Programming and Orchestration for Virtualized Software Networks", 2017 IEEE International Conference on Communications Workshops (ICC Workshops).

[28] Steven van Rossem et al., "A Network Service Development Kit Supporting the End-to-End Lifecycle of NFV-based Telecom Services", 2017 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN).

[29] T. Soenen, S.Van Rossem, W.Tavernier, F.Vicensy, D.Valocchiz, et al., "Insights from SONATA: Implementing and Integrating a Microservice-based NFV Service Platform with a DevOps Methodology", https://biblio.ugent.be/publication/8562744

[30] The SONATA Gatekeeper, Available from: http://sonata-nfv.eu/sites/default/files/sonata/public/content-files/article/SONATA_Gatekeeper_SDNWorld_3.pdf [retrieved: August, 2018].

Reliability Evaluation of Erasure Coded Systems under Rebuild Bandwidth Constraints

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

***Abstract*—Modern storage systems employ erasure coding redundancy and recovering schemes to ensure high data reliability at high storage efficiency. The widely used replication scheme belongs to this broad class of erasure coding schemes. The effectiveness of these schemes has been evaluated based on the Mean Time to Data Loss** (MTTDL) **and the Expected Annual Fraction of Data Loss** (EAFDL) **metrics. To improve the reliability of data storage systems, certain data placement and rebuild schemes reduce the rebuild times by recovering data in parallel from the storage devices. It is often assumed that there is sufficient network bandwidth to transfer the data required by the rebuild process at full speed. In large-scale data storage systems, however, the network bandwidth is constrained. This article obtains** MTTDL **and** EAFDL **of erasure coded systems analytically for arbitrary rebuild time distributions and for the symmetric, clustered, and declustered data placement schemes under network rebuild bandwidth constraints. The resulting reliability degradation is assessed and the results obtained establish that the declustered placement scheme offers superior reliability in terms of both metrics. Efficient codeword configurations that achieve high reliability in the presence of network rebuild bandwidth constraints are identified.**

Keywords–Storage; Reliability; Data placement; MTTDL; EAFDL; RAID; MDS codes; Information Dispersal Algorithm; Prioritized rebuild; Repair bandwidth; Network bandwidth constraint.

I. INTRODUCTION

In today's large-scale data storage systems, data redundancy is introduced to ensure that data lost due to device and component failures can be recovered. Appropriate redundancy schemes are deployed to prevent permanent loss of data and, consequently, enhance the reliability of storage systems [1]. The effectiveness of these schemes has been evaluated based on the Mean Time to Data Loss (MTTDL) [2-21] and, more recently, the Fraction of Data Loss Per Year (FDLPY) [22] and the equivalent Expected Annual Fraction of Data Loss (EAFDL) reliability metrics [23-25]. Analytical reliability expressions for MTTDL were obtained predominately using Markovian models, which assume that component failure and rebuild times are independent and exponentially distributed. In practice, however, these distributions are not exponential. To cope with this issue, system reliability was assessed in [17][19][24][25] using an alternative methodology that does not involve Markovian analysis and considers the practical case of non-exponential failure and rebuild time distributions. Moreover, the misconception reported in [26] that MTTDL derivations based on Markovian models provide unrealistic results was dispelled in [27] by invoking improved MTTDL derivations that yield satisfactory results, and also by drawing on prior work that obtains MTTDL analytically without involving Markovian analysis.

Earlier works have predominately considered the MTTDL metric, whereas recent works have also considered the EAFDL metric [23][24][25]. The introduction of the latter metric was motivated by the fact that Amazon S3 considers the durability of data over a given year [28], and, similarly, Facebook [29], LinkedIn [30] and Yahoo! [31] consider the amount of data lost in given periods.

To protect data from being lost and to improve the reliability of data storage systems, replication-based storage systems spread replicas corresponding to data stored on each storage device across several other storage devices. To improve the low storage efficiency associated with the replication schemes, erasure coding schemes that provide a high data reliability as well as a high storage efficiency are deployed. Special cases of such codes are the Redundant Arrays of Inexpensive Disks (RAID) schemes, such as RAID-5 and RAID-6, that have been deployed extensively in the past thirty years [2][3].

State-of-the-art data storage systems [32-35] employ more general erasure codes that affect the reliability, performance, and the storage and reconstruction overhead of the system. In this article, we focus on the reliability assessment of erasure coded systems in terms of the MTTDL and EAFDL metrics. These metrics were analytically derived in [25] for the symmetric, clustered, and declustered data placement schemes under the assumption that there is sufficient network bandwidth to transfer the data required by the rebuild process at full speed. For instance, in the case of a declustered placement, redundant data associated with the data stored on a given device is placed across all remaining devices in the system. In this way, the rebuild process can be parallelized, which in turn results in short rebuild times. The restoration time can be minimized provided there is sufficient network rebuild bandwidth available. In large-scale erasure coded storage systems, however, the rebuild operations generate a significant amount of network traffic that interferes with user-generated network traffic [36]. A common practice to cope with growing network traffic is to throttle the network bandwidth available for recovery operations, which leads to the network bandwidth being constrained. This in turn results in a reliability degradation, the extent of which is minimized by employing a prioritized rebuild process that first rebuilds the most-exposed to failure data [25][32].

The effect of network rebuild bandwidth constraints on the reliability of replication-based storage systems was studied in [9][16]. It was found that system reliability was significantly

reduced when replicas are spread over a higher number of devices than what the network rebuild bandwidth can support at full speed during a parallel rebuild process. The reliability of erasure coded systems in the absence of bandwidth constraints was assessed in [25]. The MTTDL and EAFDL metrics were obtained analytically for the symmetric, clustered, and declustered data placement schemes based on a general framework and methodology. In this article, we recognize that this methodology also holds in the case of network rebuild bandwidth constraints and apply it to derive enhanced closed-form reliability expressions for the MTTDL and EAFDL metrics for these placement schemes in the presence of such rebuild bandwidth constraints. Subsequently, we provide insight into the effect of the placement schemes and the impact of the available network rebuild bandwidth on system reliability. The validity of this methodology for accurately assessing the reliability of storage systems was confirmed by simulations in several contexts [15][16][17][19][23]. It was demonstrated that theoretical predictions for the reliability of systems comprised of highly reliable storage devices are in good agreement with simulation results. Consequently, the emphasis of the present work is on the theoretical assessment of the effect of network rebuild bandwidth constraints on the reliability of erasure coded systems. Moreover, this work extends the reliability results obtained in [16] for the special case of replication-based storage systems to the more general case of erasure coded systems.

The key contributions of this article are the following. We consider the reliability of erasure coded systems under network rebuild bandwidth constraints that was assessed in our earlier work [1] for deterministic rebuild times. In this study, we extend our previous work by also considering arbitrary rebuild times. We show that the codeword lengths that optimize the MTTDL and EAFDL metrics are similar. Furthermore, we derive the asymptotic analytic expressions for the MTTDL and EAFDL reliability metrics when the number of devices becomes large. We then obtain analytically the optimal codeword lengths corresponding to large storage systems. We subsequently establish theoretically that, for large storage systems that use a declustered placement scheme, both metrics are optimized when the codeword length is about 60% of the storage system size, regardless of the rebuild bandwidth constraints.

The remainder of the article is organized as follows. Section II describes the storage system model and the corresponding parameters considered. Section III presents the adaptation of a general framework and methodology for deriving the MTTDL and EAFDL metrics analytically for the case of erasure coded systems under network rebuild bandwidth constraints. Closed-form expressions for the symmetric, clustered, and declustered placement schemes are derived. In Section IV, the data placement schemes that offer the best reliability are identified and the resulting optimal system configurations are specified in Section V. Section VI presents numerical results demonstrating the effectiveness of the erasure coding redundancy schemes for improving system reliability. It also assesses the sensitivity to the network rebuild bandwidth constraints under various codeword configurations. Section VII provides a discussion of the applicability of the results obtained. Finally, we conclude in Section VIII.

II. Storage System Model

Modern data storage systems use erasure coded schemes to protect data from device failures. When devices fail, the redundancy of the affected data is reduced and eventually lost. To avoid irrecoverable data loss, the system performs rebuild operations that use the data stored in the surviving devices to reconstruct the temporarily lost data, thus maintaining the initial data redundancy. We proceed by briefly reviewing the basic concepts of erasure-coding and data-recovery procedures of such storage systems. To assess their reliability, we consider the model used in [25], and adopt and extend the notation. More precisely, the storage system considered here comprises n storage devices (nodes or disks), with each device storing an amount c of data, such that the total storage capacity of the system is $n\,c$.

A. Redundancy

User data is divided into blocks (or symbols) of a fixed size (e.g., sector size of 512 bytes) and complemented with parity symbols to form codewords. We consider (m, l) maximum distance separable (MDS) erasure codes, which are a mapping from l user-data symbols to a set of m ($> l$) symbols, called a codeword, having the property that any subset containing l of the m symbols of the codeword can be used to decode (reconstruct, recover) the codeword. The corresponding storage efficiency s_{eff} is given by

$$s_{\text{eff}} = \frac{l}{m} \,. \tag{1}$$

Consequently, the amount U of user data stored in the system is given by

$$U = s_{\text{eff}}\, n\, c \;=\; \frac{l\, n\, c}{m} \,. \tag{2}$$

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.

The m symbols of each codeword are stored on m distinct devices, such that the system can tolerate any $\tilde{r} - 1$ device failures, but $\tilde{r}$ device failures may lead to data loss, with

$$\tilde{r} = m - l + 1 \,. \tag{3}$$

From the above, it follows that

$$1 \le l < m \qquad \text{and} \qquad 2 \le \tilde{r} \le m \,. \tag{4}$$

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 $s_{\text{eff}}^{(\text{replication})} = 1/r$.
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 $s_{\text{eff}}^{(\text{RAID-6})} = (N - 2)/N$.
Reed–Solomon: It is based on (m, l) MDS erasure codes.

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
s	symbol 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
s_{eff}	storage efficiency of redundancy scheme ($s_{\text{eff}} = l/m$)
U	amount of user data stored in the system ($U = s_{\text{eff}}\, n\, c$)
$\tilde{r}$	minimum number of codeword symbols lost that lead to an irrecoverable data loss ($\tilde{r} = m - l + 1$ and $2 \leq \tilde{r} \leq m$)
N_b	maximum number of devices from which rebuild can occur at full speed in parallel ($N_b = B_{\max}/b$)
ϕ	bandwidth constraint factor $\left(\phi = \min\left(\frac{N_b}{k}, 1\right)\right)$
B_{eff}	effective average network rebuild bandwidth
$f_X(.)$	probability density function of X ($f_X(.) = F'_X(.)$)
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 ($\lambda^{-1} = \int_0^\infty [1 - F_\lambda(t)]dt$)

Note that the RAID-10 and RAID-01 storage systems are non-MDS in that they can sustain a single disk failure and potentially a second one. Similarly, the nested two-dimensional RAID-5 systems, such as RAID 51, use non-MDS erasure codes in that they can sustain any three device failures, but also certain other subsets of more than three device failures [21].

B. Symmetric Codeword Placement

According to a symmetric codeword placement, each codeword is stored on m distinct devices with one symbol per device. In a large storage system, the number of devices n is usually much larger than the codeword length m. Therefore, there are many ways in which a codeword of m symbols can be stored across a subset of the n devices. For each device in the system, the *redundancy spread factor* k denotes the number of devices over which the codewords stored on that device are spread [19]. The system effectively comprises n/k disjoint groups of k devices. Each group contains an amount U/k of user data, with the corresponding codewords placed on the corresponding k devices in a distributed manner. Each codeword is placed entirely in one of the n/k groups. Within each group, all $\binom{k}{m}$ possible ways of placing m symbols across k devices are used equally to store all the codewords in that group.

In such a symmetric placement scheme, within each of the n/k groups, the $m-1$ codeword symbols corresponding to the data on each device are spread *equally* across the remaining $k-1$ devices, the $m-2$ codeword symbols corresponding to the codewords shared by any two devices are spread equally across the remaining $k-2$ devices, and so on. Note also that the n/k groups are logical and therefore need not be physically located in the same node/rack/datacenter.

From the above, it follows that

$$l < m \leq k \leq n\ . \tag{5}$$

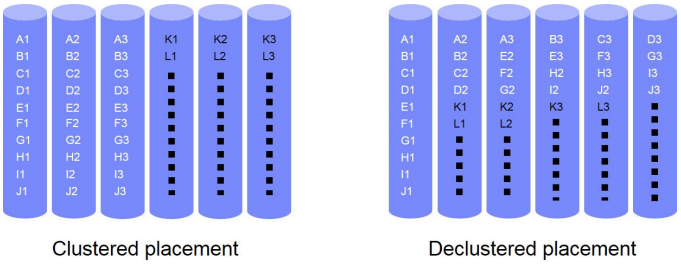


Figure 1. Clustered and declustered placement of codewords of length $m = 3$ on $n = 6$ devices. X1, X2, X3 represent a codeword (X = A, B, C, ... , L).

We proceed by considering the clustered and declustered placement schemes, which are special cases of symmetric placement schemes for which k is equal to m and n, respectively. This results in n/m groups for clustered and one group for declustered placement schemes.

1) Clustered Placement: The n devices are divided into disjoint sets of m devices, referred to as *clusters*. According to the *clustered* placement, each codeword is stored across the devices of a particular cluster, as shown in Figure 1. In such a placement scheme, it can be seen that no cluster stores the redundancies that correspond to the data stored on another cluster. The entire storage system can essentially be modeled as consisting of n/m independent clusters. In each cluster, data loss occurs when $\tilde{r}$ devices fail successively before rebuild operations can be completed successfully.

2) Declustered Placement: In this placement scheme, all $\binom{n}{m}$ possible ways of placing m symbols across n devices are used equally to store all the codewords in the system, as shown in Figure 1.

The clustered and declustered placement schemes represent the two extremes in which the symbols of the codewords associated with the data stored on a failing device are spread across the remaining devices and hence the extremes of the degree of parallelism that can be exploited when rebuilding this data. For declustered placement, the symbols are spread equally across *all* remaining devices, whereas for clustered placement, the symbols are spread across the smallest possible number of devices.

C. Codeword Reconstruction

When storage devices fail, codewords lose some of their symbols, and this leads to a reduction in data redundancy. The system attempts to maintain its redundancy by reconstructing the lost codeword symbols using the surviving symbols of the affected codewords. We assume that failures are detected instantaneously, which immediately triggers the rebuild process.

When a declustered placement scheme is used, as shown in Figure 2, spare space is reserved on each device for temporarily storing the reconstructed codeword symbols before they are transferred to a new replacement device. The rebuild process used to restore the data lost by failed devices is assumed to be both *prioritized* and *distributed*. As discussed in [25], a prioritized (or intelligent) rebuild process always attempts first to rebuild the *most-exposed* codewords, namely, the codewords that have lost the largest number of symbols. The prioritized rebuild process recovers one of the symbols that each of the

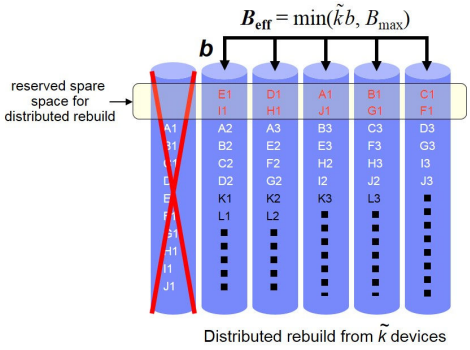


Figure 2. Rebuild under declustered placement.

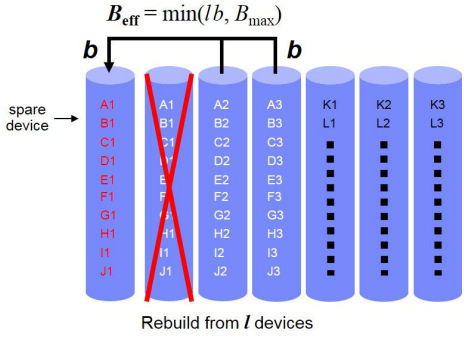


Figure 3. Rebuild under clustered placement.

most-exposed codewords has lost by reading $m - \tilde{r} + 1$ of the remaining symbols. In a distributed rebuild process, the codeword symbols lost by failed devices are reconstructed by reading surviving symbols from a number, say $\tilde{k}$, of surviving devices and storing the recovered symbols in the reserved spare space of the $\tilde{k}$ surviving devices, as shown in Figure 2.

A certain proportion of the device bandwidth is reserved for data recovery during the rebuild process, where b denotes the actual average reserved rebuild bandwidth per device. This bandwidth is usually only a fraction of the total bandwidth available at each device, the remaining bandwidth being used to serve user requests. Thus, the lost symbols are rebuilt in parallel using the rebuild bandwidth b available on each surviving device. During this process, it is desirable to reconstruct the lost codeword symbols on devices in which another symbol of the same codeword is not already present. Assuming that the system is at exposure level u (as described in Section II-D below), b_u $(\leq b)$ denotes the average rate at which the amount of data that needs to be rebuilt (repair traffic) is written to selected device(s). Also, denote the cumulative distribution function of the time X required to read (or write) an amount c of data from (or to) a device by $F_X(.)$ and its corresponding probability density function by $f_X(.)$. The kth moment of X, $E(X^k)$, is then given by

$$E(X^k) = \int_0^\infty t^k f_X(t)dt\ , \quad \text{for } k = 1, 2, \dots\ . \tag{6}$$

In particular, $1/\mu$ denotes the average time required to read (or write) an amount c of data from (or to) a device, given by

$$\frac{1}{\mu} \triangleq E(X) = \frac{c}{b}\ . \tag{7}$$

In a distributed rebuild process involving $\tilde{k}$ devices, performing a rebuild at full speed consumes an average network bandwidth of $\tilde{k}\,b$. Let $B_{\max}$ $(\geq b)$ denote the available average network bandwidth for rebuilds. Then, the effective average network rebuild bandwidth used by rebuilds, $B_{\text{eff}}(\tilde{k})$, cannot exceed $B_{\max}$ and is therefore given by

$$B_{\text{eff}}(\tilde{k}) = \min(\tilde{k}\,b, B_{\max}) = \min(\tilde{k}, N_b)\,b\ , \tag{8}$$

where N_b specifies the effective maximum number of devices from which rebuild can occur in parallel at full speed, and is given by

$$N_b \triangleq \frac{B_{\max}}{b}\ . \tag{9}$$

Note that N_b may not be an integer; it only represents the *effective* maximum number of devices from which distributed rebuild can occur at full speed.

A similar reconstruction process is used for other symmetric placement schemes within each group of k devices, except for clustered placement. When clustered placement is used, the codeword symbols are spread across all $k = m$ devices in each group (cluster). Therefore, reconstructing the lost symbols on the surviving devices of a group will result in more than one symbol of the same codeword on the same device. To avoid this, the lost symbols are reconstructed directly in spare devices as shown in Figure 3. In these reconstruction processes, decoding and re-encoding of data are assumed to be done on the fly, so the reconstruction time is equal to the time taken to read and write the required data to the devices. Note also that alternative erasure coding schemes have been proposed to reduce the amount of data transferred over the storage network during reconstruction (see [37][38] and references therein).

D. Exposure Levels and Amount of Data to Rebuild

At time t, $D_j(t)$ denotes the number of codewords that have lost j symbols, where $0 \leq j \leq \tilde{r}$. The system is at exposure level u $(0 \leq u \leq \tilde{r})$, where

$$u = \max_{D_j(t)>0} j. \tag{10}$$

The system is at exposure level u if there are codewords with $m - u$ symbols left, but there are no codewords with fewer than $m - u$ symbols left in the system, that is, $D_u(t) > 0$, and $D_j(t) = 0$, for all $j > u$. These codewords are referred to as the *most-exposed* codewords. At $t = 0$, $D_j(0) = 0$ for all $j > 0$, and $D_0(0)$ is the total number of codewords stored in the system. Device failures and rebuild processes cause the values of $D_1(t), \cdots, D_{\tilde{r}}(t)$ to change over time, and when a data loss occurs, $D_{\tilde{r}}(t) > 0$. Device failures cause transitions to higher exposure levels, whereas rebuilds cause transitions to lower ones. Let t_u denote the time of the first transition from exposure level $u - 1$ to exposure level u, and t_u^+ the instant immediately after t_u. Then, the number C_u of most-exposed codewords when entering exposure level u, $u = 1, \dots, \tilde{r}$, is given by $C_u = D_u(t_u^+)$. For $u = 1$, according to [25, Equation (8)], it holds that $C_1 = c/s$, where s denotes the symbol size. For $u \geq 2$, according to [25, Equations (6) and (27)], the

expected value of C_u is given by

$$E(C_u \,|\, \alpha_1, \ldots, \alpha_{u-1}) = \frac{c}{s} \prod_{j=1}^{u-1} V_j \, \alpha_j \;, \quad \text{for} \;\; u = 2, \ldots, \tilde{r} \;, \tag{11}$$

where V_j represents the fraction of the most-exposed codewords at exposure level j that have symbols stored on a newly failed device that causes the exposure level transition $j \rightarrow j+1$. Note that this fraction is dependent only on the codeword placement scheme. Also, α_j denotes the fraction of rebuild time R_j still left when another device fails causing the exposure level transition $j \rightarrow j+1$. Unconditioning (11) on $\alpha_1, \ldots, \alpha_{u-1}$ yields

$$E(C_u) = \frac{c}{s} \left(\prod_{j=1}^{u-1} V_j \right) \, E \left(\prod_{j=1}^{u-1} \alpha_j \right) \;, \quad \text{for} \;\; u = 2, \ldots, \tilde{r} \;. \tag{12}$$

Analytic expressions for the reliability metrics of interest were derived in [25] using the direct path approximation, which considers only transitions from lower to higher exposure levels [15][17][19]. This implies that each exposure level is entered only once.

E. Failure and Rebuild Time Distributions

We adopt the model and notation considered in [25]. The lifetimes of the n devices are assumed to be independent and identically distributed, with a cumulative distribution function $F_\lambda(.)$ and a mean of $1/\lambda$. We consider real-world distributions, such as Weibull and gamma, as well as exponential distributions that belong to the large class defined in [17]. Note that, although the model considered here does not account for correlated device failures, their effect can be assessed by enhancing the model according to the approach presented in [14]. This issue, however, is beyond the scope of this article. The storage devices are characterized to be *highly reliable* in 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, that is,

$$\frac{\lambda}{\mu} = \frac{\lambda \, c}{b} \ll 1 \;. \tag{13}$$

We consider storage devices the cumulative distribution function F_λ satisfies the condition

$$\mu \int_0^\infty F_\lambda(t)[1 - F_X(t)]dt \ll 1, \quad \text{with} \; \frac{\lambda}{\mu} \ll 1 \;, \tag{14}$$

such that the MTTDL and EAFDL reliability metrics of erasure coded storage systems tend to be insensitive to the device failure distribution, that is, they depend only on its mean $1/\lambda$, but not on its density $F_\lambda(.)$. They are, however, sensitive to the distribution $F_X(.)$ of the device rebuild times [25].

III. Derivation of MTTDL and EAFDL

The MTTDL metric assesses the expected amount of time until some data can no longer be recovered and therefore is irrecoverably lost, whereas the EAFDL assesses the fraction of stored data that is expected to be lost by the system annually. The EAFDL is obtained as the ratio of the expected amount of user data lost normalized to the amount of user data to the mean time to data loss [23, Equation (6)]:

$$\text{EAFDL} = \frac{E(H)}{U \cdot \text{MTTDL}} \;, \tag{15}$$

where H denotes the amount of user data lost, given that a data loss has occurred, and with the MTTDL expressed in years.

The MTTDL(B_{max}) and EAFDL(B_{max}) metrics are derived as a function of B_{max} based on the framework and methodology presented in [25]. More specifically, this methodology uses the direct path approximation and does not involve Markovian analysis. It holds for general failure time distributions, which can be exponential or non-exponential, such as the Weibull and gamma distributions that satisfy condition (14). Note that this framework is general because it is also valid in the case where the network rebuild bandwidth is constrained. The only parameters that are affected by the network rebuild bandwidth constraint are the rebuild rates and, accordingly, those parameters that are dependent on them, such as the rebuild times. Analytic expressions for the two metrics of interest were derived in [25, Equations (49) and (50)] as follows:

$$\text{MTTDL}(B_{\text{max}}) \approx \frac{1}{n\,\lambda} \, \frac{(\tilde{r}-1)!}{(\lambda\, c)^{\tilde{r}-1}} \, \frac{[E(X)]^{\tilde{r}-1}}{E(X^{\tilde{r}-1})} \, \prod_{u=1}^{\tilde{r}-1} \frac{b_u(B_{\text{max}})}{\tilde{n}_u} \, \frac{1}{V_u^{\tilde{r}-1-u}} \;, \tag{16}$$

and

$$\text{EAFDL}(B_{\text{max}}) \approx m\,\lambda\,(\lambda\, c)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \, \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}} \, \prod_{u=1}^{\tilde{r}-1} \frac{\tilde{n}_u}{b_u(B_{\text{max}})} \, V_u^{\tilde{r}-u} \;, \tag{17}$$

where $\tilde{n}_u$ represents the number of devices at exposure level u whose failure before the rebuild of the most-exposed codewords causes an exposure level transition to level $u+1$. As mentioned above, b_u, the average rate at which the amount of data that needs to be rebuilt at exposure level u is written to selected device(s), is dependent on B_{max}, the upper limitation of the average network rebuild bandwidth.

The expected amount $E(Q)$ of data lost upon a first-device failure is given by [25, Equation (47)]

$$E(Q) \approx \, l\, c\, (\lambda\, c)^{\tilde{r}-1} \, \frac{1}{\tilde{r}\,!} \, \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}} \, \prod_{u=1}^{\tilde{r}-1} \frac{\tilde{n}_u}{b_u} \, V_u^{\tilde{r}-u} \;, \tag{18}$$

where $E(X^{\tilde{r}-1})$ is obtained by (6). Subsequently, the expected amount $E(H)$ of data lost, given that a data loss has occurred, is given by [25, Equation (48)]

$$E(H) \approx \left(\frac{l}{\tilde{r}} \, \prod_{u=1}^{\tilde{r}-1} V_u \right) c \;. \tag{19}$$

Central to the derivation of $E(Q)$ and $E(H)$ is the assessment of the amount H of user data lost, that is, the amount of user data stored in the most-exposed codewords when entering exposure level $\tilde{r}$ that can no longer be recovered and therefore is irrecoverably lost. In [25, Equation (22)] it was assumed that

each of the most-exposed codewords loses all its l symbols of user data, that is,

$$H = C_{\tilde{r}}\, l\, s\ , \tag{20}$$

where $C_{\tilde{r}}$ is the number of the most-exposed codewords when entering exposure level $\tilde{r}$, and s is the symbol size. This clearly overestimates the amount of data lost, especially when the number of user-data symbols l in each of the most-exposed codewords is greater than the number of lost symbols $\tilde{r}$. We proceed to revise the derivation of H. Let n_l denote the number of user-data symbols irrecoverably lost in a typical most-exposed codeword. As devices are equally likely to fail, and given that the fraction of user-data symbols in a codeword is equal to l/m, a moment's reflection reveals that the expected number of user-data symbols irrecoverably lost is given by

$$E(n_l) = \frac{l}{m}\,\tilde{r}\ =\ \frac{\tilde{r}}{m}\,l\ . \tag{21}$$

Consequently, for a number $C_{\tilde{r}}$ of most-exposed codewords, the expected amount $E(H\,|\,C_{\tilde{r}})$ of user data lost is given by

$$E(H\,|\,C_{\tilde{r}}) = C_{\tilde{r}}\,E(n_l)\,s\ \overset{(21)}{=}\ C_{\tilde{r}}\,\frac{\tilde{r}}{m}\,l\,s\ , \tag{22}$$

which in turn, by unconditioning on $C_{\tilde{r}}$, yields

$$E(H) = E(C_{\tilde{r}})\,\frac{\tilde{r}}{m}\,l\,s\ . \tag{23}$$

From the above, it follows that in each of the most-exposed codewords, the expected fraction f_l of the user-data symbols that are lost is given by

$$f_l\ \triangleq\ E\left(\frac{n_l}{l}\right) = \frac{E(n_l)}{l} = \frac{\tilde{r}}{m}\ \overset{(3)}{=}\ \frac{m-l+1}{m}\ . \tag{24}$$

The resulting expressions for $E(Q)$, EAFDL, and $E(H)$ can now be obtained by multiplying (18), (17), and (19) with f_l, which yields

$$E(Q) \approx\ \frac{l}{m}\,c\,(\lambda\,c)^{\tilde{r}-1}\ \frac{1}{(\tilde{r}-1)\,!}\ \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}}\ \prod_{u=1}^{\tilde{r}-1}\frac{\tilde{n}_u}{b_u}\ V_u^{\tilde{r}-u}\ , \tag{25}$$

where $E(X^{\tilde{r}-1})$ is obtained by (6),

$$\text{EAFDL}(B_{\max}) \approx$$
$$\lambda\,(\lambda\,c)^{\tilde{r}-1}\ \frac{1}{(\tilde{r}-1)\,!}\ \frac{E(X^{\tilde{r}-1})}{[E(X)]^{\tilde{r}-1}}\ \prod_{u=1}^{\tilde{r}-1}\frac{\tilde{n}_u}{b_u(B_{\max})}\ V_u^{\tilde{r}-u}\ , \tag{26}$$

and

$$E(H) \approx \left(\frac{l}{m}\ \prod_{u=1}^{\tilde{r}-1} V_u\right) c\ . \tag{27}$$

Remark 1: From (16), (26), and (27), and given that $E(X) = c/b$, it follows that MTTDL and EAFDL are dependent on the $(m-l)$th moment of the rebuild time distribution. Furthermore, given that $E(X^{m-l}) \geq [E(X)]^{m-l}$, random rebuild times result in lower MTTDL and higher EAFDL values than deterministic rebuild times do. In contrast, the expected amount $E(H)$ of user data lost, given that a data loss has occurred, is not dependent on λ, b and c nor on the rebuild time distribution. Moreover, $E(H)$ is not dependent on b_u and therefore is not affected by the limitation on the network rebuild bandwidth; it is only dependent on the storage efficiency and data placement scheme. Moreover, MTTDL is dependent on n, but EAFDL and $E(H)$ are not.

Remark 2: The analytic expressions for the MTTDL and EAFDL reliability metrics were derived in [25] in the absence of network rebuild bandwidth constraints. Consequently, they correspond to the case of $B_{\max} = \infty$, where the two metrics are denoted by MTTDL(∞) and EAFDL(∞), respectively.

From (16) and (17), or the enhanced expression (26), it follows that

$$\frac{\text{MTTDL}(B_{\max})}{\text{MTTDL}(\infty)} = \frac{\text{EAFDL}(\infty)}{\text{EAFDL}(B_{\max})} = \theta\ , \tag{28}$$

where θ represents the *reliability reduction factor* that assesses the reliability degradation due to a network rebuild bandwidth constraint, and is given by

$$\theta\ \triangleq\ \prod_{u=1}^{\tilde{r}-1}\frac{b_u(B_{\max})}{b_u(\infty)}\ . \tag{29}$$

Equation (28) suggests that the reliability reduction factor for EAFDL is the same as the one for MTTDL. At first glance, and given the different nature of the MTTDL and EAFDL metrics, this seems to be counterintuitive. The reason for this result is that network rebuild bandwidth constraints effectively prolong the duration of the rebuild times, which equally affects the MTTDL and EAFDL metrics. More specifically, at exposure level u, and according to [25, Equations (43) and (44)], the transition probability $P_{u\rightarrow u+1}$ from exposure level u to $u+1$ is proportional to the rebuild time R_u, which in turn is inversely proportional to the average rebuild rate b_u. Thus, constraining the average rebuild rates b_u increases the probability of data loss P_{DL}. Consequently, the reliability metrics are equally affected given that, according to [25, Equations (14) and (15)], MTTDL and EAFDL are inversely proportional and proportional to P_{DL}, respectively. Note also that the corresponding amount of data lost H is dependent only on the data placement scheme and is therefore not affected by the prolongation of the rebuild times.

Remark 3: From (29), and given that $b_u(B_{\max})$ decreases with decreasing $B_{\max}$, it follows that θ decreases with increasing $\tilde{r}$ or decreasing $B_{\max}$.

Remark 4: From (12), (23), and (27), it follows that

$$E\left(\prod_{j=1}^{\tilde{r}-1}\alpha_j\right) = \frac{1}{\tilde{r}}\ . \tag{30}$$

Note that the variables $\alpha_1,\ldots,\alpha_{\tilde{r}-1}$ are generally independent and approximately uniformly distributed between 0 and 1 such that $E(\alpha_u) \approx 1/2$, $u = 1,\ldots,\tilde{r}-1$ [23, 25]. In this context, however, this assumption would lead to the erroneous result $E(\prod_{j=1}^{\tilde{r}-1}\alpha_j) = 1/2^{\tilde{r}-1}$, which is smaller than the correct one by a factor of $2^{\tilde{r}-1}/\tilde{r}$. This is analogous to the factor of $2^{r-1}/r$ that was derived for replication-based storage systems in Section V.E of [23]. It turns out that when a data loss has occurred, the variables $\alpha_1,\ldots,\alpha_{\tilde{r}-1}$ are not distributed identically. Further insight regarding this subtle issue is provided in the relevant discussion of that section.

Assuming that the system has reached exposure level u, we deduce from (30) that

$$E\left(\prod_{j=1}^{u-1}\alpha_j\right)=\frac{1}{u}\,,\quad \text{for}\ \ u=2,\ldots,\tilde{r}\,. \tag{31}$$

A. Symmetric Placement

We consider the case where the redundancy spread factor k is in the interval $m<k\leq n$. The special case $k=m$, which corresponds to the clustered placement scheme, has to be considered separately for the reasons discussed in Section II-B1. As discussed in [25, Section IV-B], the *prioritized* rebuild process at each exposure level u recovers one of the u symbols that each of the most-exposed codewords has lost by reading $m-\tilde{r}+1$ of the remaining symbols from the $\tilde{n}_u$ surviving devices in the affected group. According to [25, Equation (51)], it holds that

$$\tilde{n}_u^{\text{sym}}=k-u\,. \tag{32}$$

Furthermore, in the absence of a network rebuild bandwidth constraint, the total write bandwidth, which is also the average rebuild rate b_u, is given by [25, Equation (52)]

$$b_u^{\text{sym}}(\infty)=\frac{\tilde{n}_u^{\text{sym}}}{m-\tilde{r}+2}\,b\overset{(3)}{=}\frac{\tilde{n}_u^{\text{sym}}\,b}{l+1}\,,\quad u=1,\ldots,\tilde{r}-1\,. \tag{33}$$

However, in the presence of a network rebuild bandwidth constraint B_{max} and according to (8) with $\tilde{k}=\tilde{n}_u=\tilde{n}_u^{\text{sym}}$, the average rebuild rate b_u is given as a function of B_{max} by

$$b_u^{\text{sym}}(B_{\text{max}})=\frac{B_{\text{eff}}(\tilde{n}_u)}{l+1}=\frac{\min(\tilde{n}_u\,b,B_{\text{max}})}{l+1}=\frac{\min(\tilde{n}_u,N_b)\,b}{l+1}\overset{(32)}{=}\frac{\min(k-u,N_b)\,b}{l+1}\,,\ \text{for}\ \ u=1,\ldots,\tilde{r}-1\,. \tag{34}$$

Substituting (33) and (34) into (29) yields

$$\theta^{\,\text{sym}}=\prod_{u=1}^{\tilde{r}-1}\frac{\min(k-u,N_b)}{k-u}\,. \tag{35}$$

Note that when $N_b\geq k-1$, the system reliability is not affected because all rebuilds are performed at full speed, and therefore the factor θ is equal to 1. However, when $N_b<k-1$, it may not be possible for some of the rebuilds to be performed at full speed, and therefore the factor θ will be less than 1, which affects the system reliability. Consequently, the reliability reduction factor θ depends on the *bandwidth constraint factor* ϕ which is given by

$$\phi\triangleq\min\left(\frac{N_b}{k},1\right)\overset{(9)}{=}\min\left(\frac{B_{\text{max}}}{k\,b},1\right)\,,\ \text{with}\ 0\leq\phi\leq 1\,. \tag{36}$$

From (34), (35), and (36), and recognizing that $\min(k-u,N_b)=\min(\min(k-u,k),N_b)=\min(k-u,\min(k,N_b))=\min(k\min(1,N_b/k),k-u)=\min(k\,\phi,k-u)$, it follows that

$$b_u^{\text{sym}}(B_{\text{max}})=\frac{\min\left(\frac{\phi}{1-\frac{u}{k}},1\right)(k-u)\,b}{l+1}\,,\ \text{for}\ \ u=1,\ldots,\tilde{r}-1\,, \tag{37}$$

and

$$\theta^{\,\text{sym}}=\prod_{u=1}^{\tilde{r}-1}\min\left(\frac{\phi}{1-\frac{u}{k}},1\right)\,. \tag{38}$$

Using (38) and the fact that MTTDL(∞) is given by [25, Equation (54)], (28) yields

$$\text{MTTDL}_k^{\text{sym}}(B_{\text{max}})\approx\frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l}(m-l)!\,\frac{[E(X)]^{m-l}}{E(X^{m-l})}\prod_{u=1}^{m-l}\left(\frac{k-u}{m-u}\right)^{m-l-u}\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{k}},1\right)\,, \tag{39}$$

where B_{max} is expressed via ϕ given by (36).

Note that, for a deterministic rebuild time distribution, for which it holds that $E(X^{m-l})=[E(X)]^{m-l}$, and for a replication-based system, for which $m=r$ and $l=1$, and by virtue of (35) and (38), Equation (39) is in agreement with Equation (24) of [16], where $c/b=1/\mu$.

Using (38) and the fact that EAFDL(∞) is obtained by multiplying [25, Equation (55)] with the expected fraction f_l of the user-data symbols that are contained in the irrecoverable codewords and are lost, by virtue of (24), (28) yields

$$\text{EAFDL}_k^{\text{sym}}(B_{\text{max}})\approx\lambda\left[\frac{(l+1)\,\lambda\,c}{b}\right]^{m-l}\frac{1}{(m-l)!}\,\frac{E(X^{m-l})}{[E(X)]^{m-l}}\prod_{u=1}^{m-l}\left(\frac{m-u}{k-u}\right)^{m-l+1-u}\Bigg/\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{k}},1\right), \tag{40}$$

where B_{max} is expressed via ϕ given by (36).

Moreover, $E(H)$ can be obtained by multiplying [25, Equation (56)] with f_l, which, according to (24), yields

$$E(H)_k^{\text{sym}}\approx\left(\frac{l}{m}\prod_{u=1}^{m-l}\frac{m-u}{k-u}\right)c \tag{41}$$

$$=\frac{l\,(m-1)!\,(k-m+l-1)!}{m\,(k-1)!\,(l-1)!}\,c\,. \tag{42}$$

For given l, m, n, and ϕ, the redundancy spread factors or, equivalently, the optimal group sizes that maximize MTTDL$^{\text{sym}}$, EAFDL$^{\text{sym}}$, and $E(H)^{\text{sym}}$ are given by the following propositions.

Proposition 1: For given l, m, n, and ϕ, and for any λ, c, b, and rebuild time distribution of X, the value of k $(m+1\leq k\leq n)$, denoted by $\hat{k_s}$, that maximizes MTTDL$_k^{\text{sym}}$ is given by

$$\hat{k_s}=\begin{cases}\text{any } j\in[m+1,n]\cap D_n\,,\ \text{which includes } j=k_m\,,\\ \qquad\text{for } m-l=1 \text{ and } \phi\geq 1-\frac{1}{n}\\ \text{any } j\in[m+1,\frac{1}{1-\phi}]\cap D_n\,,\ \text{which includes } j=k_m\,,\\ \qquad\text{for } m-l=1 \text{ and } 1-\frac{1}{k_m}\leq\phi<1-\frac{1}{n}\\ k_m\,,\qquad\text{for } m-l=1 \text{ and } \phi<1-\frac{1}{k_m}\\ n\,,\qquad\text{for } m-l\geq 2\,,\end{cases} \tag{43}$$

where D_n is the set of the divisors (factors) of n, that is,

$$D_n \triangleq \{\, j \,:\, j \,|\, n \} \;\equiv\; \left\{ j \,:\, j \in \mathbb{N} \quad \text{and} \quad \frac{n}{j} \in \mathbb{N} \right\} , \quad (44)$$

and k_m is the smallest integer in the interval $I_k = [m+1, n]$ that divides n, that is,

$$k_m \triangleq \min_j \{\, j \in I_k \cap D_n \} . \quad (45)$$

Proof: See Appendix A. ■

Proposition 2: For given l, m, n, and ϕ, and for any c, λ, and rebuild time distribution of X, $\text{EAFDL}_k^{\text{sym}}$ and $E(H)_k^{\text{sym}}$ are decreasing in k and are therefore minimized when $k = n$.

Proof: Considering l, m, and n to be fixed, it follows from (40) that $\text{EAFDL}_k^{\text{sym}}$ is inversely proportional to the function B_k given by

$$B_k \triangleq \prod_{u=1}^{m-l} (k-u)^{m-l-u} \;\; \min(k\,\phi,\, k-u) . \quad (46)$$

Note that each of the terms in the product is increasing in k, which implies that B_k is also increasing in k and, consequently, $\text{EAFDL}_k^{\text{sym}}$ is decreasing in k. Furthermore, it follows from (41) that $E(H)_k^{\text{sym}}$ is also decreasing in k. ■

Remark 5: From the preceding two propositions it follows that, for $l+1 < m < k \leq n$, $\text{MTTDL}_k^{\text{sym}}$ is maximized and $\text{EAFDL}_k^{\text{sym}}$ and $E(H)_k^{\text{sym}}$ are minimized by the declustered placement scheme, that is, when $k = n$.

An approximate expression for the reliability reduction function is given by the following lemma.

LEMMA 1: For large values of k, m, l, and $m-l$, $\theta^{\text{sym}}_{\text{approx}}$ can be approximated as follows:

$$\log\left(\theta^{\text{sym}}_{\text{approx}}\right) \approx \left[\log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right] k - \frac{1}{2}\log(1-\widehat{\phi}) , \quad (47)$$

where $\widehat{\phi}$ is given by

$$\widehat{\phi} \;\triangleq\; \min(1-\phi\,, hx) , \quad (48)$$

h is given by

$$h \;\triangleq\; 1 - s_{\text{eff}} = 1 - \frac{l}{m} , \quad (49)$$

and x by

$$x \triangleq \frac{m}{k} . \quad (50)$$

Proof: See Appendix B. ■

Approximate expressions for the reliability metrics of interest are given by the following propositions.

Proposition 3: For large values of n, k, m, l, and $m-l$, $\text{MTTDL}^{\text{sym}}$ normalized to $1/\lambda$ can be approximated as follows:

$$\begin{aligned}\log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(B_{\max})\right) \approx\; & \log\left(\frac{k}{n}\right) \\ & + k^2\,\frac{W(h,x)}{2} + k\left\{hx\,\log\left(\frac{hx\sqrt{x}\,k\,b}{e\,[(1-h)x\,k+1]\,\lambda\,c}\right)\right. \\ & \left. + \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right\} \\ & - \frac{1}{8}\left[h(1-x) - \log\left(\frac{1-h}{1-hx}\right)\right] + \log\left(\sqrt{\frac{2\pi hx}{k}}\right) \\ & - \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{[E(X)]^{hxk}}{E(X^{hxk})}\right) , \end{aligned} \quad (51)$$

where $W(h,x)$ is given by

$$W(h,x) \triangleq hx(1-x) - \log\left(\frac{\left[(1-h)^{(1-h)^2}\,x^{h^2}\right]^{x^2}}{(1-hx)^{(1-hx)^2}}\right) , \quad (52)$$

and h, x, and $\widehat{\phi}$ are given by (49), (50), and (48), respectively.

Proof: It follows from (28) that

$$\text{MTTDL}^{\text{sym}}_{\text{approx}}(B_{\max}) \;=\; \text{MTTDL}^{\text{sym}}_{\text{approx}}(\infty)\;\theta^{\,\text{sym}}_{\text{approx}} , \quad (53)$$

or, equivalently,

$$\begin{aligned}&\log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(B_{\max})\right) = \\ &\qquad \log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(\infty)\right) + \log\left(\theta^{\,\text{sym}}_{\text{approx}}\right) . \end{aligned} \quad (54)$$

Substituting the analytic expression obtained in [25, Equation (62)] for the term $\log\left(\lambda\,\text{MTTDL}^{\text{sym}}_{\text{approx}}(\infty)\right)$, and (47) into (54) yields (51). ■

Proposition 4: For large values of k, m, l, and $m-l$, $\text{EAFDL}^{\text{sym}}$ normalized to λ can be approximated as follows:

$$\begin{aligned}\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(B_{\max})/\lambda\right) \approx\; & -k^2\,\frac{W(h,x)}{2} \\ & + k\left\{hx\,\log\left(\frac{e\,[(1-h)x\,k+1]\,\lambda\,c}{h\,\sqrt{x}\,k\,b}\right) + \log\left(\frac{(1-hx)^{1-hx}}{(1-h)^{(1-h)x}}\right)\right. \\ & \left. - \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) - \widehat{\phi}\right\} \\ & + \frac{1}{8}\,h(1-x) + \log\left(\sqrt{\frac{1}{2\pi hxk}}\left(\frac{1-h}{1-hx}\right)^{\frac{3}{8}}\right) \\ & + \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{E(X^{hxk})}{[E(X)]^{hxk}}\right) , \end{aligned} \quad (55)$$

where $B_{\max}$ is expressed via ϕ given by (36), and h, x, $W(h,x)$, and $\widehat{\phi}$ are given by (49), (50), (52), and (48), respectively.

Proof: It follows from (28) that

$$\text{EAFDL}^{\text{sym}}_{\text{approx}}(B_{\max}) \;=\; \frac{\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)}{\theta^{\,\text{sym}}_{\text{approx}}} . \quad (56)$$

or, equivalently,

$$\begin{aligned}&\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(B_{\max})/\lambda\right) = \\ &\qquad \log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)/\lambda\right) - \log\left(\theta^{\,\text{sym}}_{\text{approx}}\right) . \end{aligned} \quad (57)$$

The term $\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)$ is obtained by multiplying the corresponding term obtained in [25] with the expected fraction f_l of the user-data symbols that are contained in the irrecoverable codewords and are lost. Consequently, the term $\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)/\lambda\right)$ is obtained by adding to the analytic expression obtained in [25, Equation (64)] the term $\log(f_l)$, which, according to (24), and using (49) and (50), is given by

$$f_l = \frac{hxk+1}{xk} \;, \quad \text{or} \quad \log(f_l) = \log\left(\frac{hxk+1}{xk}\right) \;. \tag{58}$$

Substituting the outcome for the resulting enhanced term $\log\left(\text{EAFDL}^{\text{sym}}_{\text{approx}}(\infty)/\lambda\right)$ and (47) into (57) yields (55). ■

Proposition 5: For large values of k, m, l, and $m-l$, $E(H)^{\text{sym}}$ normalized to c can be approximated as follows:

$$\log\left(E(H)^{\text{sym}}_{\text{approx}}/c\right) \approx \log\left((1-h)\sqrt{\frac{1-h}{1-hx}}\right) + k\,V(h,x) \;, \tag{59}$$

where $V(h,x)$ is given by

$$V(h,x) \triangleq \log\left(\frac{x^x\,(1-hx)^{1-hx}}{[(1-h)x]^{(1-h)x}}\right) \;, \tag{60}$$

and h and x are given by (49) and (50), respectively.

Proof: Immediate by multiplying the term $\text{EAFDL}^{\text{sym}}_{\text{approx}}$ with f_l or, equivalently, by adding to [25, Equation (58)] the term $\log(f_l)$ given by (58). ■

B. Clustered Placement

In the clustered placement scheme, the n devices are divided into disjoint sets of m devices, referred to as *clusters*. According to *clustered* placement, each codeword is stored across the devices of a particular cluster. At each exposure level u, the rebuild process recovers one of the u symbols that each of the C_u most-exposed codewords has lost by reading $m-\tilde{r}+1$ of the remaining symbols. Note that the remaining symbols are stored on the $m-u$ surviving devices in the affected group. According to [25, Equation (65)], it holds that

$$\tilde{n}^{\text{clus}}_u = m-u \;. \tag{61}$$

In the case of clustered placement, the rebuild process recovers the lost symbols by reading l symbols from l of the $\tilde{n}_u$ surviving devices of the affected cluster. In the absence of a network rebuild bandwidth constraint, the symbols are read from each of the l devices at an average rate of b such that the average effective network rebuild bandwidth is equal to $B_{\text{eff}} = l\,b$. Subsequently, the lost symbols are computed on-the-fly and written to a spare device at an average rate of $B_{\text{eff}}/l = b$. Consequently, it holds that

$$b^{\text{clus}}_u(\infty) = b \,, \quad u = 1,\ldots,\tilde{r}-1 \;. \tag{62}$$

However, in the presence though of a network rebuild bandwidth constraint $B_{\max}$ the effective average network rebuild bandwidth is equal to $B_{\text{eff}} = \min(l\,b, B_{\max})$, which implies that the lost symbols are written to a spare device at an average rate of B_{eff}/l. Thus, the average rebuild rate b_u is given as a function of $B_{\max}$ by

$$b^{\text{clus}}_u(B_{\max}) = \frac{B_{\text{eff}}(B_{\max})}{l} = \frac{\min(l\,b, B_{\max})}{l} = \frac{\min(l, N_b)\,b}{l} \;, \quad \text{for } u = 1,\ldots,\tilde{r}-1 \;. \tag{63}$$

Remark 6: Note that, as far as the data placement is concerned, the clustered placement scheme is a special case of a symmetric placement scheme for which k is equal to m. However, its reliability assessment cannot be directly obtained from the reliability results derived in Section III-A for the symmetric placement scheme by simply setting $k=m$. The reason for that is the difference in the rebuild processes. In the case of a symmetric placement scheme, recovered symbols are written to the spare space of existing devices, whereas in the case of a clustered placement scheme, recovered symbols are written to a spare device. This results in different rebuild bandwidths, which are given by (34) and (63), respectively.

Substituting (62) and (63) into (29) yields

$$\theta^{\,\text{clus}} = \left(\frac{\min(l, N_b)}{l}\right)^{\tilde{r}-1} \;. \tag{64}$$

As $l < m$, it holds that $\min(l, N_b) = \min(\min(l,m), N_b) = \min(\min(N_b, m), l) = \min(m\min(N_b/m, 1), l) = \min(m\phi, l)$, where, analogously to (36), and with $k=m$,

$$\phi \triangleq \min\left(\frac{N_b}{m}, 1\right) \overset{(9)}{=} \min\left(\frac{B_{\max}}{m\,b}, 1\right) \;, \quad \text{where } 0 \le \phi \le 1 \;. \tag{65}$$

Consequently, (63) and (64) yield

$$b^{\text{clus}}_u(B_{\max}) = \min\left(\frac{m}{l}\,\phi\,, 1\right) b \;, \quad \text{for } u = 1,\ldots,\tilde{r}-1 \;, \tag{66}$$

and

$$\theta^{\,\text{clus}} = \min\left(\frac{m}{l}\,\phi\,, 1\right)^{\tilde{r}-1} \;, \tag{67}$$

respectively.

Remark 7: It follows from (67) that for $m\,\phi/l \ge 1$ or, equivalently, for $\phi \ge s_{\text{eff}} = l/m$, $\theta^{\,\text{clus}}$ is equal to 1, which implies that the bandwidth constraint does not affect the system reliability.

Using (3) and (67), and the fact that $\text{MTTDL}(\infty)$ is given by [25, Equation (68)], (28) yields

$$\text{MTTDL}^{\text{clus}}(B_{\max}) \approx \frac{1}{n\,\lambda}\left(\frac{\min(\frac{m\,\phi}{l}, 1)\,b}{\lambda\,c}\right)^{m-l} \frac{1}{\binom{m-1}{l-1}}\,\frac{[E(X)]^{m-l}}{E(X^{m-l})} \;, \tag{68}$$

where $B_{\max}$ is expressed via ϕ given by (65).

Note that for a RAID-5 array comprised of N devices, such that $n=k=m=N$ and $l=N-1$, for $\phi \ge s_{\text{eff}} = l/m$ and by virtue of (7), (68) yields

$$\text{MTTDL}^{\text{clus}}_{\text{RAID-5}} \approx \frac{\mu}{N(N-1)\lambda^2} \;, \tag{69}$$

which is the same result as that reported in [2]. Note that when $m-l=1$, MTTDL is insensitive to the rebuild time distribution.

For an exponential rebuild time distribution, for which it holds that $E(X^{m-l}) = (m-l)!\,[E(X)]^{m-l}$, and for a RAID-6 array comprised of N devices, such that $n = k = m = N$ and $l = N-2$, for $\phi \geq s_{\text{eff}} = l/m$ and by virtue of (7), (68) yields

$$\text{MTTDL}^{\text{clus, exp}}_{\text{RAID-6}} \;\approx\; \frac{\mu^2}{N(N-1)(N-2)\lambda^3}\;, \tag{70}$$

which is the same result as that reported in [3]. That result was derived using a continuous-time Markov chain (CTMC) model with the repair rate equal to μ, which is not dependent on the number of failed devices. This is analogous to our model where lost symbols are written to a spare device at an average rate of b, which is fixed and is not dependent on the number of failed devices, and the rebuild time distribution is exponential.

In contrast, in [39], the Mean Time Between Failures (MTBF) was derived using a CTMC model and assuming that the repair rate of each failed device is fixed, which implies that the total repair rate is proportional to the number of failed devices. In the case where $\lambda \ll \mu$, [39, Equation (1.1)] with k replaced by l and n by N yields

$$\text{MTBF} \;\approx\; \frac{1}{l\,\lambda\,\binom{N}{l}\,(\lambda/\mu)^{N-l}}\;. \tag{71}$$

For a deterministic rebuild time distribution, for which it holds that $E(X^{m-l}) = [E(X)]^{m-l}$, for $\phi \geq s_{\text{eff}} = l/m$ and for a RAID-6 array, (68) yields

$$\text{MTTDL}^{\text{clus, det}}_{\text{RAID-6}} \;\approx\; \frac{2\,\mu^2}{N(N-1)\lambda^3}\;, \tag{72}$$

and for arbitrary l values ($l < m = k = n = N$), (68) yields

$$\begin{aligned}\text{MTTDL}^{\text{clus, det}} \;&\approx \frac{1}{N\,\lambda}\,\frac{1}{(\lambda/\mu)^{N-l}}\,\frac{1}{\binom{N-1}{l-1}}\\ &= \frac{1}{l\,\lambda\,\binom{N}{l}\,(\lambda/\mu)^{N-l}}\;,\end{aligned} \tag{73}$$

which is the same result as in (71), despite some strikingly different characteristics in the operation of the underlying systems. MTBF is obtained assuming exponential rebuild times (Markovian behavior) and repair rates proportional to the number of failed devices, whereas our model yields the same result assuming deterministic rebuild times (non-Markovian behavior) and a fixed repair rate independent of the number of failed devices.

We now proceed to derive EAFDL and $E(H)$. Using (3) and (38), and the fact that EAFDL(∞) is obtained by multiplying [25, Equation (69)] with the expected fraction f_l of the user-data symbols that are contained in the irrecoverable codewords and are lost, by virtue of (24), (28) yields

$$\begin{aligned}&\text{EAFDL}^{\text{clus}}(B_{\max}) \approx\\ &\qquad \lambda\left(\frac{\lambda\,c}{\min(\frac{m\,\phi}{l},1)\,b}\right)^{m-l}\binom{m-1}{l-1}\,\frac{E(X^{m-l})}{[E(X)]^{m-l}}\;,\end{aligned} \tag{74}$$

where $B_{\max}$ is expressed via ϕ given by (65).

Moreover, $E(H)$ can be obtained by multiplying [25, Equation (70)] with f_l, which, according to (24), yields

$$E(H)^{\text{clus}} \approx \frac{l}{m}\,c\,. \tag{75}$$

Approximate expressions for the reliability metrics of interest are given by the following propositions.

Proposition 6: For large values of n, m, l, and $m-l$, $\text{MTTDL}^{\text{clus}}$ normalized to $1/\lambda$ and $\text{EAFDL}^{\text{clus}}$ normalized to λ can be approximated as follows:

$$\begin{aligned}\lambda\,\text{MTTDL}^{\text{clus}}_{\text{approx}}(B_{\max}) \;&\approx\; \sqrt{\frac{2\pi hx}{(1-h)n}}\\ &\left[\left(\frac{h\min\left(\frac{\phi}{1-h},1\right)b}{\lambda\,c}\right)^{h}(1-h)^{1-h}\right]^{xn}\,\frac{[E(X)]^{hxn}}{E(X^{hxn})}\;,\end{aligned} \tag{76}$$

$$\begin{aligned}\text{EAFDL}^{\text{clus}}_{\text{approx}}(B_{\max})/\lambda \;&\approx\; \sqrt{\frac{1-h}{2\pi hxn}}\\ &\left[\left(\frac{h\min\left(\frac{\phi}{1-h},1\right)b}{\lambda\,c}\right)^{h}(1-h)^{1-h}\right]^{-xn}\,\frac{E(X^{hxn})}{[E(X)]^{hxn}}\;,\end{aligned} \tag{77}$$

where

$$x = \frac{m}{n}\;, \tag{78}$$

$B_{\max}$ is expressed via ϕ given by (65), and h is given by (49).

Proof: It follows from (28) that

$$\text{MTTDL}^{\text{clus}}_{\text{approx}}(B_{\max}) \;=\; \text{MTTDL}^{\text{clus}}_{\text{approx}}(\infty)\;\theta^{\,\text{clus}}\;, \tag{79}$$

and

$$\text{EAFDL}^{\text{clus}}_{\text{approx}}(B_{\max}) \;=\; \frac{\text{EAFDL}^{\text{clus}}_{\text{approx}}(\infty)}{\theta^{\,\text{clus}}}\;. \tag{80}$$

It follows from (67), and using (3), (49), and (78) that

$$\theta^{\,\text{clus}} \;=\; \min\left(\frac{\phi}{1-h},1\right)^{hxn}\;. \tag{81}$$

Substituting the analytic expression obtained in [25, Equation (71)] for the term $\lambda\,\text{MTTDL}^{\text{clus}}_{\text{approx}}(\infty)$, and (81) into (79) yields (76). Subsequently, substituting (75) and (76) into (15) and using (49) and (78) yields (77). ■

C. Declustered Placement

The declustered placement scheme is a special case of a symmetric placement scheme in which k is equal to n. Consequently, for $k = n$, (39), (40), and (41) yield

$$\begin{aligned}\text{MTTDL}^{\text{declus}}(B_{\max}) \;\approx\; &\frac{1}{n\,\lambda}\left[\frac{b}{(l+1)\,\lambda\,c}\right]^{m-l}(m-l)!\\ &\frac{[E(X)]^{m-l}}{E(X^{m-l})}\prod_{u=1}^{m-l}\left(\frac{n-u}{m-u}\right)^{m-l-u}\prod_{u=1}^{m-l}\min\left(\frac{\phi}{1-\frac{u}{n}},1\right)\;,\end{aligned} \tag{82}$$

$$\text{EAFDL}^{\text{declus}}(B_{\max}) \approx \lambda \left[\frac{(l+1)\,\lambda\, c}{b}\right]^{m-l} \frac{1}{(m-l)!} \frac{E(X^{m-l})}{[E(X)]^{m-l}} \prod_{u=1}^{m-l} \left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Bigg/ \prod_{u=1}^{m-l} \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right), \tag{83}$$

where $B_{\max}$ is expressed via ϕ given by (36) with $k = n$, and

$$E(H)^{\text{declus}} \approx \left(\frac{l}{m} \prod_{u=1}^{m-l} \frac{m-u}{n-u}\right) c \tag{84}$$

$$= \frac{l\,(m-1)!\,(n-m+l-1)!}{m\,(n-1)!\,(l-1)!}\, c\,. \tag{85}$$

Approximate expressions for the reliability metrics of interest are given by the following propositions.

Proposition 7: For large values of n, m, l, and $m-l$, $\text{MTTDL}^{\text{declus}}$ normalized to $1/\lambda$ can be approximated as follows:

$$\begin{aligned}\log\left(\lambda\,\text{MTTDL}^{\text{declus}}_{\text{approx}}(B_{\max})\right) \approx\ & + n^2\,\frac{W(h,x)}{2} + n\left\{hx\,\log\left(\frac{hx\sqrt{x}\,n\,b}{e\,[(1-h)x\,n+1]\,\lambda\,c}\right)\right. \\ & \left. + \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right\} \\ & - \frac{1}{8}\left[h(1-x) - \log\left(\frac{1-h}{1-hx}\right)\right] + \log\left(\sqrt{\frac{2\pi hx}{n}}\right) \\ & - \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{[E(X)]^{hxn}}{E(X^{hxn})}\right)\,,\end{aligned} \tag{86}$$

where $B_{\max}$ is expressed via ϕ given by (36) with $k = n$, and h, x, $W(h,x)$, and $\widehat{\phi}$ are given by (49), (78), (52), and (48), respectively.

Proof: Immediate from Proposition 3 by replacing k with n and using (78). ■

Proposition 8: For large values of n, m, l, and $m-l$, the $\text{EAFDL}^{\text{declus}}$ normalized to λ can be approximated as follows:

$$\begin{aligned}\log\left(\text{EAFDL}^{\text{declus}}_{\text{approx}}(B_{\max})/\lambda\right) \approx\ & - n^2\,\frac{W(h,x)}{2} \\ & + n\left\{hx\,\log\left(\frac{e\,[(1-h)x\,n+1]\,\lambda\,c}{h\,\sqrt{x}\,n\,b}\right) + \log\left(\frac{(1-hx)^{1-hx}}{(1-h)^{(1-h)x}}\right)\right. \\ & \left. - \log\left(\phi^{\widehat{\phi}}\,(1-\widehat{\phi})^{1-\widehat{\phi}}\right) - \widehat{\phi}\right\} \\ & + \frac{1}{8}\,h(1-x) + \log\left(\sqrt{\frac{1}{2\pi hxn}}\left(\frac{1-h}{1-hx}\right)^{\frac{3}{8}}\right) \\ & + \frac{1}{2}\log(1-\widehat{\phi}) + \log\left(\frac{E(X^{hxn})}{[E(X)]^{hxn}}\right)\,,\end{aligned} \tag{87}$$

where $B_{\max}$ is expressed via ϕ given by (36) with $k = n$, and h, x, $W(h,x)$, and $\widehat{\phi}$ are given by (49), (78), (52), and (48), respectively.

Proof: Immediate from Proposition 4 by replacing k with n and using (78). ■

Proposition 9: For large values of n, m, l, and $m-l$, $E(H)^{\text{declus}}$ normalized to c can be approximated as follows:

$$\log\left(E(H)^{\text{declus}}_{\text{approx}}/c\right) \approx \log\left((1-h)\sqrt{\frac{1-h}{1-hx}}\right) + n\,V(h,x)\,, \tag{88}$$

where h, x, and $V(h,x)$ are given by (49), (78), and (60), respectively.

Proof: Immediate from Proposition 5 by replacing k with n and using (78). ■

IV. Reliability Optimization

For given l, m, n, and ϕ, we identify the placement scheme that offers the best reliability in terms of the MTTDL, EAFDL, and $E(H)$ metrics. In Section III, we identified the optimal placement scheme within the class of symmetric placement schemes when $m < k \leq n$. The corresponding reliability achieved should be compared with the one achieved by the clustered placement scheme when $k = m < n$. For the comparison to be meaningful, there should be at least two clustered groups, which implies that $m \leq n/2$, or, by also using (3) and (4),

$$1 \leq l < m \qquad \text{and} \qquad 1 \leq m-l < m \leq \frac{n}{2}\,. \tag{89}$$

A. *Maximizing* MTTDL

To obtain the optimal MTTDL value and identify the corresponding optimal placement, we consider the following two cases. If m does not divide n, then the optimal MTTDL value is equal to the $\text{MTTDL}^{\text{sym}}_{\hat{k_s}}$ value obtained by a symmetric placement where $k = \hat{k_s}$. If m divides n, then we need to compare the $\text{MTTDL}^{\text{sym}}_{\hat{k_s}}$ value with the $\text{MTTDL}^{\text{clus}}$ value obtained by the clustered placement with $k = m$. From (39) and (68), it follows that the ratio $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ of these two values is given by

$$\begin{aligned} r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} &\triangleq \frac{\text{MTTDL}^{\text{sym}}_{\hat{k_s}}}{\text{MTTDL}^{\text{clus}}} \\ &\approx \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \Bigg/ \min\left(\frac{m\,\phi}{l}, 1\right)^{m-l} \\ &\quad \prod_{u=1}^{m-l} \left(\frac{\hat{k_s}-u}{m-u}\right)^{m-l-u} \min\left(\frac{\phi}{1-\frac{u}{\hat{k_s}}}, 1\right). \end{aligned} \tag{90}$$

Remark 8: It follows from (90) that the placement that maximizes MTTDL is not dependent on λ, c, or the rebuild time distribution of X, but is dependent on b and $B_{\max}$ only through ϕ, that is, the ratio $B_{\max}/b$.

The optimal placement is given by the following proposition.

Proposition 10: For given l, m, n, and ϕ, and for any λ, c, b, and rebuild time distribution of X, the value of k ($m \leq$

$k \leq n$), denoted by $\hat{k}$, that maximizes MTTDL_k is given by

$$\hat{k} = \begin{cases} m\,, & \text{for } m-l=1 \text{ and } n = j\,m\,, \text{ for some } j \in \mathbb{N} \\ \hat{k_s}\,, & \text{for } m-l=1 \text{ and } n \neq j\,m\,, \text{ for all } j \in \mathbb{N} \\ m\,, & \text{for } l=1,\ m=3,\ n=3\,j \text{ with } 2 \leq j \leq 11\,, \\ & \qquad \text{and } \phi < \frac{2\sqrt{n-2}}{n} \\ m\,, & \text{for } l=2,\ m=4,\ n=8,\ \phi < \frac{3\sqrt{3}}{8} = 0.649 \\ m\,, & \text{for } l=2,\ m=4,\ n=12,\ \phi < \frac{\sqrt{5}}{4} = 0.559 \\ m\,, & \text{for } l=3,\ m=5,\ n=10,\ \phi < \frac{4\sqrt{2}}{5\sqrt{3}} = 0.653 \\ m\,, & \text{for } l=1,\ m=4,\ n=8,\ \phi < \frac{1}{4}\sqrt[3]{\frac{15}{7}} = 0.322 \\ n\,, & \text{otherwise}\,, \end{cases} \tag{91}$$

where $\hat{k_s}$ is given by (43).

Proof: See Appendix C. ■

B. Minimizing EAFDL

To obtain the optimal EAFDL value and identify the corresponding optimal placement, we consider the following two cases. If m does not divide n, then, according to Proposition 2, the optimal EAFDL value is obtained by the declustered placement ($k = n$). If m divides n, then we need to compare the $\text{EAFDL}^{\text{declus}}$ value with the $\text{EAFDL}^{\text{clus}}$ value obtained by the clustered placement with $k = m$. From (83) and (74), it follows that the ratio $r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}}$ of these two values is given by

$$\begin{aligned} r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}} &\triangleq \frac{\text{EAFDL}^{\text{declus}}}{\text{EAFDL}^{\text{clus}}} \\ &\approx (l+1)^{m-l}\ \frac{(l-1)!}{(m-1)!}\ \min\left(\frac{m\,\phi}{l}, 1\right)^{m-l} \\ &\quad \prod_{u=1}^{m-l} \left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Bigg/ \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right). \end{aligned} \tag{92}$$

Remark 9: It follows from (92) that the placement that minimizes EAFDL is not dependent on λ or c. Moreover, the ratio $r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}}$ is dependent on b and B_{max} only through ϕ, that is, the ratio B_{max}/b.

The optimal placement is given by the following proposition.

Proposition 11: For any l, m, n ($n > m$), ϕ, λ, c, b, and rebuild time distribution of X, the value of k ($m \leq k \leq n$) that minimizes the EAFDL_k is equal to n, which corresponds to the declustered placement scheme.

Proof: See Appendix D. ■

C. Minimizing $E(H)$

To obtain the optimal $E(H)$ value and identify the corresponding optimal placement, we consider the following two cases. If m does not divide n, then, according to Proposition 2, the optimal $E(H)$ value is obtained by the declustered placement ($k = n$). If m divides n, then we need to compare the $E(H)^{\text{declus}}$ value with the $E(H)^{\text{clus}}$ value obtained by the clustered placement with $k = m$.

From (75) and (84) and using (89), it follows that

$$r^{\text{declus,H}}_{\text{clus,H}} \triangleq \frac{E(H)^{\text{declus}}}{E(H)^{\text{clus}}} \approx \prod_{u=1}^{m-l} \frac{m-u}{n-u} < 1\,. \tag{93}$$

Remark 10: It follows from (92) that the placement that minimizes $E(H)$ is not dependent on λ, c, b, B_{max} (or, consequently, ϕ), nor on the rebuild time distribution of X.

The optimal placement is given by the following proposition.

Proposition 12: For any l, m, n ($n > m$), ϕ, λ, c, b, and rebuild time distribution of X, the value of k ($m \leq k \leq n$) that minimizes the $E(H)_k$ is equal to n, which corresponds to the declustered placement scheme.

Proof: Immediate from (93). ■

V. Optimal System Configuration

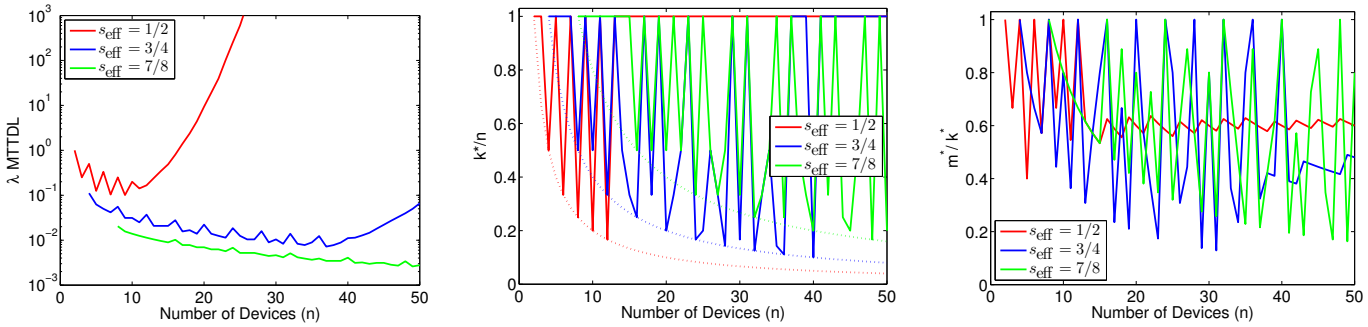
We address the issue of maximizing the reliability of a system storing an amount U of user data under a given storage efficiency s_{eff} and bandwidth constraint factor ϕ. The required number of devices n is then determined by (2). Consequently, the parameters to be specified are l, m, and k. However, these parameters are dependent. More specifically, according to (1), $l = s_{\text{eff}}\,m$. Also, given l and m, the optimal value $\hat{k}$ of k was obtained in Section IV. Consequently, to maximize system reliability, it suffices to determine the appropriate value m^* of m for the optimal codeword length. Then the optimal value k^* for the parameter k is obtained by Propositions 10, 11, and 12. Next, using a specific example, we will show that for MTTDL, we may find that $m^* \leq k^* < n$, which implies that optimality may be achieved by multiple groups, whereas for EAFDL and $E(H)$, optimality is always achieved by a single group as expressed by the following proposition.

Proposition 13: For any n, λ, c, b, B_{max} (and, consequently, ϕ), and rebuild time distribution of X, the optimal value k^* for the parameter k that minimizes the EAFDL or the $E(H)$ is equal to n.

Proof: Let us first consider the EAFDL metric, where m^* and k^* are the values that minimize it. We consider the following two cases for m^*. If $m^* < n$, then, invoking Proposition 11 with $m = m^*$, the value of k ($m^* \leq k \leq n$) that minimizes the EAFDL_k is equal to n, which implies that $k^* = n$. If $m^* = n$, then, owing to (5), it follows that $m^* = k^* = n$. Similarly, from Proposition 12, it follows that the optimal value k^* for the parameter k that minimizes the $E(H)$ is equal to n. ■

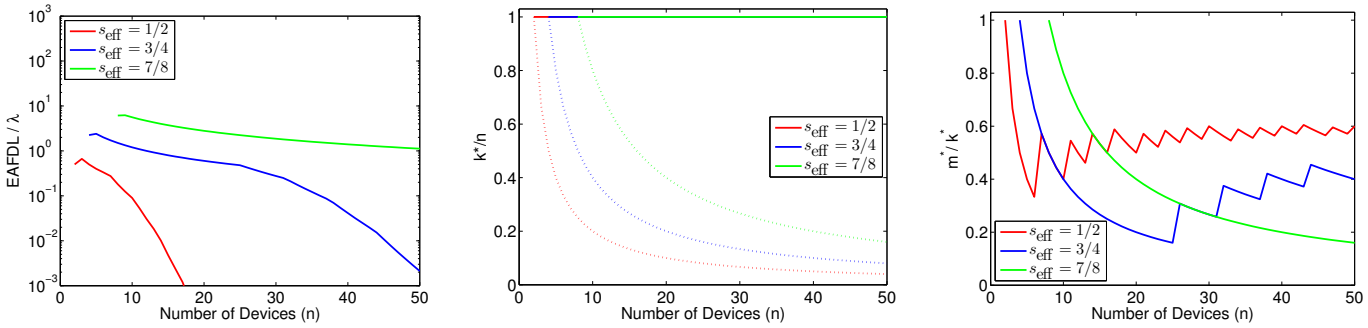
Consequently, for EAFDL and $E(H)$, the optimal placement is always the clustered or declustered one, whereas for MTTDL it may also be the symmetric one.

An alternative way to determine the optimal values m^* and k^* for the parameters m and k, respectively, is first to determine the optimal codeword length m^*_k for any given k. Note that from (39), (40), and (41), it follows that m^*_k depends on k, but not on the storage system size n. Subsequently, the optimal value of k^* can be determined by considering all possible values for k, along with the corresponding values m^*_k, and identifying the pair (k, m^*_k) that optimizes the reliability metric.



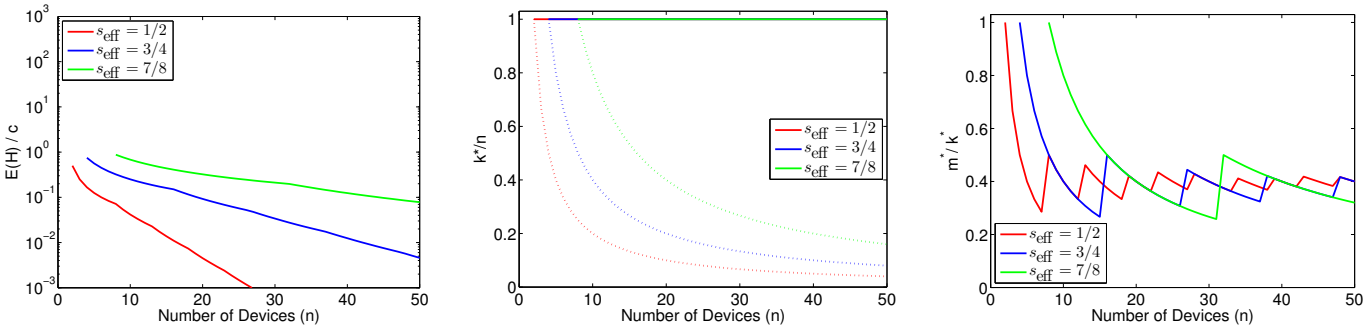
(a) Optimal MTTDL (b) k^*/n (c) m^*/k^*

Figure 4. Optimization of MTTDL vs. number of devices for $s_{\text{eff}} = 1/2, 3/4,$ and $7/8$; $\lambda/\mu = 0.001$, $\phi = 0.001$ and deterministic rebuild times.



(a) Optimal EAFDL (b) k^*/n (c) m^*/k^*

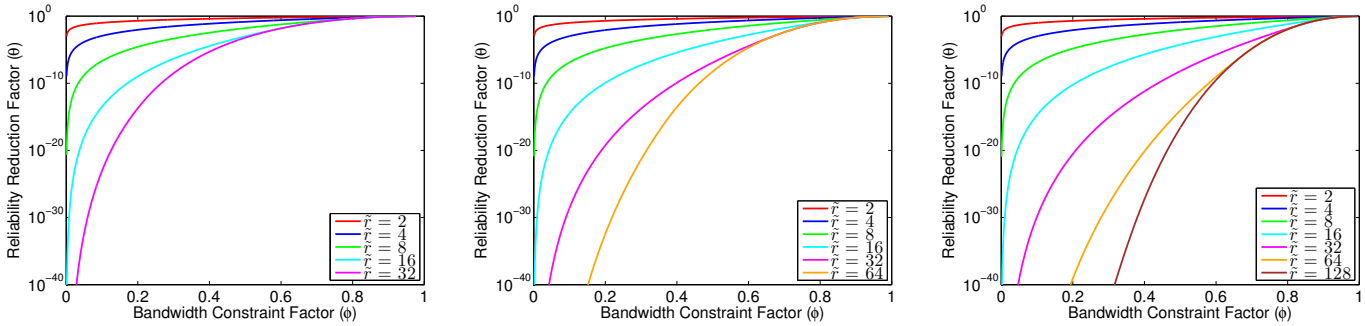
Figure 5. Optimization of EAFDL vs. number of devices for $s_{\text{eff}} = 1/2, 3/4,$ and $7/8$; $\lambda/\mu = 0.001$, $\phi = 0.001$ and deterministic rebuild times.



(a) Optimal $E(H)$ (b) k^*/n (c) m^*/k^*

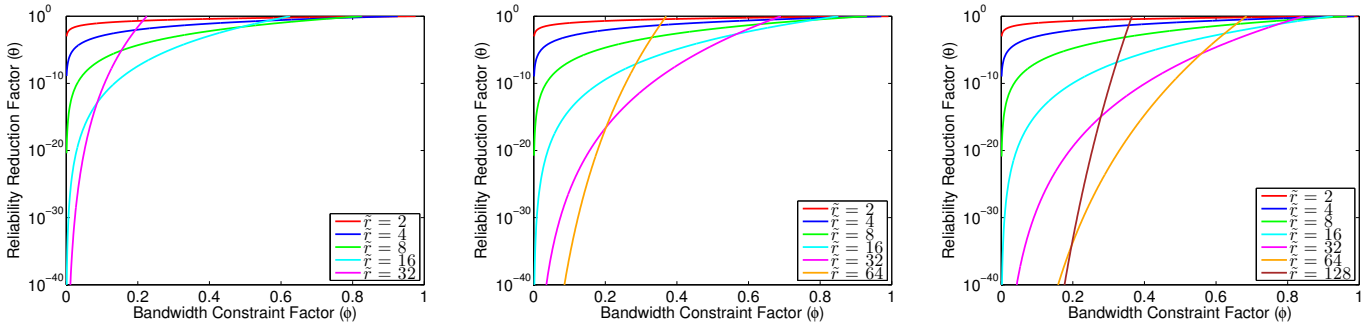
Figure 6. Optimization of $E(H)$ vs. number of devices for $s_{\text{eff}} = 1/2, 3/4,$ and $7/8$; $\lambda/\mu = 0.001$ and $\phi = 0.001$.

Next, we consider a storage system for which it holds that $\lambda/\mu = \lambda c/b = 0.001$ and $\phi = 0.001$. We identify the optimal group sizes k^* and the optimal codeword lengths m^* that optimize the MTTDL metric for various system sizes, assuming that the rebuild time distribution is deterministic. The optimal normalized λMTTDL, EAFDL$/\lambda$, and $E(H)/c$ values along with the corresponding normalized values k^*/n and m^*/k^* are shown in Figures 4, 5, and 6, respectively, as a function of system size. From Figure 4(a) we observe that, for a given storage efficiency s_{eff}, MTTDL initially decreases and then increases as n increases. For $s_{\text{eff}} = 7/8$, MTTDL starts increasing when $n \geq 115$, which is not shown in the figure. Figure 4(b) shows the ratio of k^* to n. Given that $k^* \leq n$, the maximum value of this ratio is equal to 1. Also, k^* cannot be less than the minimum codeword length, which is equal to 2, 4 and 8, for $s_{\text{eff}} = 1/2, 3/4$ and $7/8$, respectively. Therefore, k^*/n cannot be less than $2/n$, $4/n$ and $8/n$, as indicated by the dotted lines for $s_{\text{eff}} = 1/2, 3/4$ and $7/8$, respectively. Note that when a point lies on a dotted line, that is, when k^* is equal to the minimum codeword length, then the optimal codeword length m^*, which according to (5) cannot exceed k^*, is also equal to the minimum codeword length. This implies that $k^* = m^*$ and the optimal placement is the clustered one. In this case, the ratio m^*/k^* is equal to 1, as shown in Figure 4(c). For instance, for $n = 8$ and $s_{\text{eff}} = 3/4$, $k^*/n = 0.5$, that is, $k^* = m^* = 4$, and MTTDL is maximized when we consider two groups with a clustered placement within each group. However, we see in Figure 4(b) that, for $n = 10$ and $s_{\text{eff}} = 3/4$, k^*/n is still equal to 0.5, which means that the optimal group size is now equal to 5, and



(a) $k = 40$ (b) $k = 100$ (c) $k = 200$

Figure 7. Reliability reduction factor vs. bandwidth constraint factor for various values of $\tilde{r}$; symmetric placement.



(a) $m = 40$ (b) $m = 100$ (c) $m = 200$

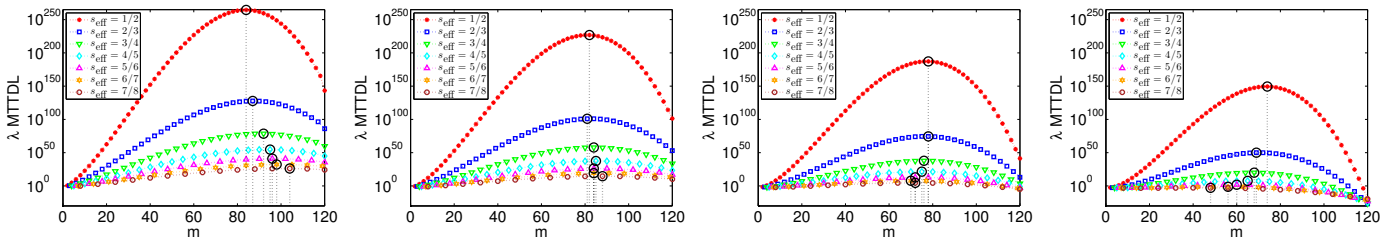
Figure 8. Reliability reduction factor vs. bandwidth constraint factor for various values of $\tilde{r}$; clustered placement.

the optimal codeword length remains equal to 4. In this case it holds that $m^*/k^* = 0.8$, as shown in Figure 4(c), and MTTDL is maximized when we consider two groups with a symmetric placement within each group. Note also that for $s_{\text{eff}} = 1/2$ and for system sizes that contain an even number of devices not exceeding 12, MTTDL is maximized by considering group sizes of two under a clustered placement $(k^* = m^* = 2)$. By contrast, for an odd number of devices, MTTDL is maximized by considering a single group and the declustered placement $(k^* = n)$. In particular, the optimal codeword lengths are $m^* = 2, 2, 4, 6$, and 6 for $n = 3, 5, 7, 9$, and 11, respectively. But when the number of devices exceeds 12, MTTDL is maximized by considering a single group under declustered placement and codewords whose lengths are about 60% of the system size $(k^* = n$ and $m^* \approx 0.6\, n)$. For $s_{\text{eff}} = 7/8$ and $n = 42$, it turns out that $k^* = 14$ or, equivalently, $k^*/n = 1/3$, and $m^* = 8$ or, equivalently, $m^*/k^* = 4/7$. Thus, MTTDL is maximized by considering three groups with a group size of 14 and a symmetric placement of codewords of length 8 containing seven user-data symbols each. Considering $l = 7$, $m = 8$, and $n = 42$, we confirm the optimal value of k by invoking (91), which in this case yields $\hat{k} = \hat{k_s}$, then using (43), which yields $\hat{k_s} = k_m$, and finally using (45), which yields $k_m = 14$. In general, the declustered placement is optimal, except in the cases of small n and ϕ where another placement may be optimal. However, this does not happen in the case of minimizing the EAFDL and $E(H)$ metrics. According to Proposition 13, and as shown in Figures 5(b) and 6(b), for all values of n, EAFDL and $E(H)$ are minimized by a single group $(k^* = n)$ and the clustered or declustered placement depending on whether n is equal to or exceeds the minimum codeword length, respectively.

VI. Numerical Results

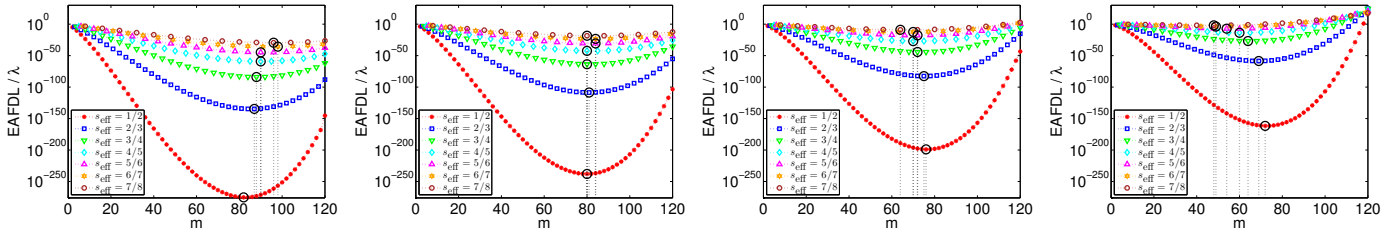
First, we assess the reduction in reliability owing to bandwidth constraints. The reliability reduction factor θ is obtained by (38) and (67) for the symmetric and clustered placements, respectively, and shown in Figures 7 and 8 as a function of the bandwidth constraint factor. For a symmetric placement scheme, Figure 7 demonstrates that as the group size k increases, the reliability reduction factor θ decreases and the magnitude of the reduction is more pronounced for larger values of $\tilde{r}$. Clearly, if codewords are spread over a higher number of devices than what the network rebuild bandwidth can support at full speed during a parallel rebuild process, the system reliability is affected and a drastic reliability degradation occurs as the system size increases. In contrast, according to Remark 7, the reliability of a clustered placement scheme remains unaffected for $\phi \geq l/m = (m - \tilde{r} + 1)/m$. This is due to the fact that the effective rebuild bandwidth is significantly smaller because the rebuilds are not distributed, but performed directly on a spare device. However, as Figure 8 demonstrates for $\phi < l/m$, the reliability reduction factor drops sharply, especially for large values of $\tilde{r}$.

Next, we consider a storage system of a given size and assess its reliability for various codeword configurations, storage efficiencies, and network rebuild bandwidth constraints. In particular, we consider a system containing 120 devices under a declustered placement scheme $(k = n = 120)$, which according to Remark 5 is optimal within the class of symmetric schemes. The amount U of user data stored is determined by



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 9. Normalized MTTDL vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $n = k = 120$, $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

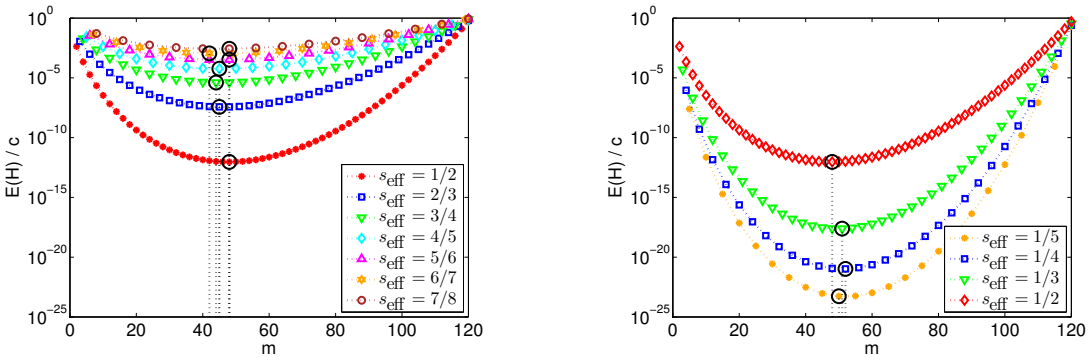
Figure 10. Normalized EAFDL vs. codeword length for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $n = k = 120$, $\lambda/\mu = 0.001$ and deterministic rebuild times.

the storage efficiency s_{eff} via (2). As discussed in Section II-E, the analytical reliability results obtained are accurate when the storage devices are highly reliable, that is, when the ratio λ/μ of the mean rebuild time $1/\mu$ to the mean time to failure of a device $1/\lambda$ is very small. We proceed by considering systems for which it holds that $\lambda/\mu = \lambda\, c/b = 0.001$ and the distribution of the rebuild time X is deterministic, that is, $E(X^{m-l}) = [E(X)]^{m-l}$.

The combined effect of the network rebuild bandwidth constraint and the storage efficiency on the normalized λ MTTDL measure is obtained by (82) and shown in Figure 9 as a function of the codeword length. In particular, when the codeword length is equal to the system size $(m = k = n)$, the placement becomes clustered and the normalized λ MTTDL measure is obtained by (68). Four cases for the network rebuild bandwidth constraint were considered: $\phi = 1$ corresponds to the case where there is no network rebuild bandwidth constraint given that $N_b \geq k = 120$ or, equivalently, $B_{\max} \geq k\, b = 120\, b$; $\phi = 0.1$, $\phi = 0.01$, and $\phi = 0.001$ correspond to the cases where $N_b = 0.1\, k = 12$, $N_b = 0.01\, k = 1.2$, and $N_b = 0.001k = 0.12$ or, equivalently, $B_{\max} = 0.1\, k\, b = 12\, b$, $B_{\max} = 0.01\, k\, b = 1.2\, b$, and $B_{\max} = 0.001\, k\, b = 0.12\, b$, respectively. The values for the storage efficiency are chosen to be fractions of the form $z/(z+1)$, $z = 1, \ldots, 7$, such that the first point of each of the corresponding curves is associated with the single-parity $(z, z+1)$-erasure code, and the second point of each of the corresponding curves is associated with the double-parity $(2z, 2z+2)$-erasure code.

For all values of ϕ considered, we observe that MTTDL increases as the storage efficiency s_{eff} decreases. This is because, for a given m, decreasing s_{eff} implies decreasing l, which in turn implies increasing the parity symbols $m - l$ and consequently improving the MTTDL. Furthermore, for a given storage efficiency s_{eff}, MTTDL decreases by orders of magnitude as the maximum permitted network rebuild bandwidth decreases. We now proceed to identify the optimal codeword length m^* that maximizes MTTDL for a given bandwidth constraint and storage efficiency. The optimal codeword length is dictated by two opposing effects on reliability. On the one hand, larger values of m imply that codewords can tolerate more device failures, but on the other hand, they result in a higher exposure degree to failure as each of the codewords is spread across a larger number of devices. In Figure 9, the optimal values m^* are indicated by the circles, and the corresponding codeword lengths are indicated by the vertical dotted lines. By comparing Figures 9(a), (b), (c), and (d), we deduce that as ϕ decreases, so do the optimal codeword lengths. For example, in the case of $s_{\text{eff}} = 3/4$ and $\phi = 1$, the maximum MTTDL value of $4{\times}10^{78}$ is obtained when $m = m^* = 92$. However, in the case of $\phi = 0.1$, the maximum MTTDL value of $6{\times}10^{57}$ is obtained for $m^* = 84$. The reason for the reduction of the optimal codeword length is that $\tilde{r}$ increases with increasing m for a given value of s_{eff}, which, according to Remark 3, results in a lower reliability reduction factor. Thus, the reliability reduction factor corresponding to $m = 92$ is lower than the one corresponding to $m = 84$, which in turn causes MTTDL for $m = 92$ to no longer be optimal as it becomes lower than the one for $m = 84$. Note that for $m = 84$ and $s_{\text{eff}} = 3/4$, it follows from (1) and (3) that $l = 63$ and $\tilde{r} - 1 = 21$. From (38), and given that $u \leq \tilde{r} - 1 = 21 \ll k = 120$, such that $\phi/(1 - u/k) \approx \phi$, it now follows that $\theta \approx \phi^{\tilde{r}-1} = 0.1^{21} = 10^{-21}$, which implies that the reliability is reduced by 21 orders of magnitude. In the cases of $\phi = 0.01$ and $\phi = 0.001$, the maximum MTTDL values of $6{\times}10^{37}$ and $8{\times}10^{19}$ are obtained for $m^* = 76$ and $m^* = 68$, respectively.

The combined effect of the network rebuild bandwidth constraint and the storage efficiency on the normalized $\text{EAFDL}^{\text{declus}}/\lambda$ measure is obtained by (74) and (83), and



(a) $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$

(b) $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$

Figure 11. Normalized $E(H)$ vs. codeword length; $n = k = 120$.

shown in Figure 10 as a function of the codeword length. We observe that EAFDL increases as the storage efficiency s_{eff} decreases. Furthermore, for a given storage efficiency s_{eff}, EAFDL increases by orders of magnitude as the maximum permitted network rebuild bandwidth decreases. In fact, for $\phi = 0.01$ and $\phi = 0.001$, Figures 10(c) and (d) show that the EAFDL can be greater than 1.

Remark 11: Although the fraction of data loss never exceeds 1, EAFDL can exceed 1 because it expresses the annual fraction of data loss, which also takes into account the frequency of data losses.

Similarly to the case of MTTDL, by comparing Figures 10(a), (b), (c), and (d), we observe that as ϕ decreases, so do the optimal codeword lengths. For example, in the case of $s_{\text{eff}} = 3/4$ and $\phi = 1$, the minimum EAFDL value of 10^{-84} is obtained when $m = m^* = 88$. However, in the case of $\phi = 0.1$, the minimum EAFDL value of $2{\times}10^{-64}$ is obtained for $m^* = 80$, which implies that the reliability is reduced by 20 orders of magnitude. In the cases of $\phi = 0.01$ and $\phi = 0.001$, the minimum EAFDL values of $7{\times}10^{-45}$ and 10^{-27} are obtained for $m^* = 72$ and $m^* = 64$, respectively. By comparing Figures 9 and 10, we deduce that in general the optimal codeword lengths m^*_{MTTDL} (for MTTDL) and m^*_{EAFDL} (for EAFDL) are similar.

The effect of the storage efficiency on the normalized $E(H)/c$ measure is obtained by (75) and (84), and shown in Figure 11 as a function of the codeword length. Note that according to Remark 1, neither the network rebuild bandwidth constraint nor the rebuild time distribution affects this metric. We observe that $E(H)$ increases as the storage efficiency s_{eff} increases.

Remark 12: From (27), and recalling (1), (5) and the fact that V_u is a fraction, we deduce that $E(H)/c \leq s_{\text{eff}} < 1$, which implies that $E(H) < c$.

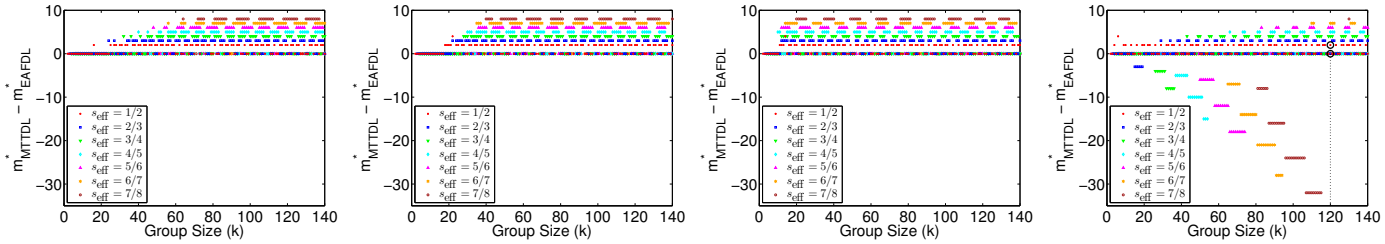
Reducing $B_{\max}$ or, equivalently, ϕ affects the optimal codeword lengths for MTTDL and EAFDL as follows.

Proposition 14: For given n, k, and s_{eff}, and for the MTTDL and EAFDL reliability metrics, the optimal codeword length m^* decreases with decreasing ϕ.

Proof: Consider two bandwidth constraint factors ϕ_1 and ϕ_2 with $\phi_1 > \phi_2$. Let m_1^* and m_2^* be the corresponding optimal codeword lengths for the MTTDL metric. We shall now show that $m_1^* \geq m_2^*$.

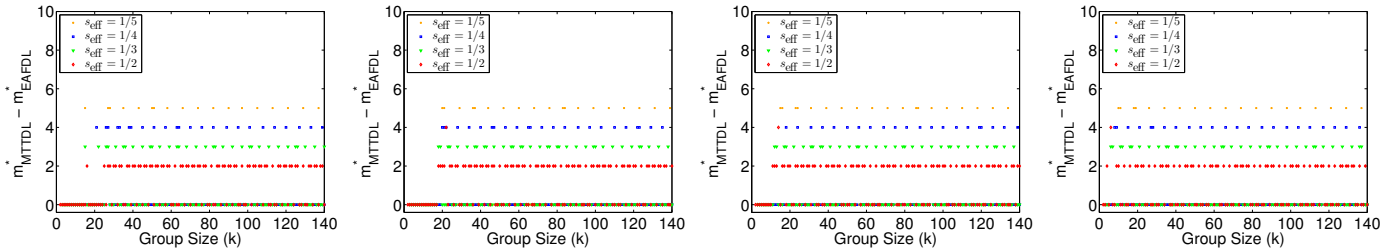
As m_1^* is the optimal codeword length for ϕ_1, it holds that $\text{MTTDL}(\phi_1, m) \leq \text{MTTDL}(\phi_1, m_1^*)$ for all $m \geq m_1^*$. Also, from (1) and (3), it holds that $\tilde{r} = (1 - s_{\text{eff}})\, m + 1$, which implies that as m increases, so does $\tilde{r}$. From (29), it follows that $\theta^{(2)}/\theta^{(1)} = \prod_{u=1}^{\tilde{r}-1} \frac{b_u(\phi_2)}{b_u(\phi_1)}$, which, owing to the fact that $b_u(\phi_2) \leq b_u(\phi_1)\ \forall\, u$, decreases with increasing $\tilde{r}$ or, equivalently, m. Consequently, $\theta_m^{(2)}/\theta_m^{(1)} \leq \theta_{m_1^*}^{(2)}/\theta_{m_1^*}^{(1)}$ for all $m \geq m_1^*$. Also, it follows from (28) that $\text{MTTDL}(\phi_2,m)/\text{MTTDL}(\phi_1,m) = \theta_m^{(2)}/\theta_m^{(1)}$ for all values of m. From the preceding, it follows that $\text{MTTDL}(\phi_2,m)/\text{MTTDL}(\phi_1,m) = \theta_m^{(2)}/\theta_m^{(1)} \leq \theta_{m_1^*}^{(2)}/\theta_{m_1^*}^{(1)} = \text{MTTDL}(\phi_2,m_1^*)/\text{MTTDL}(\phi_1,m_1^*) \leq \text{MTTDL}(\phi_2,m_1^*)/\text{MTTDL}(\phi_1,m)$ for all $m \geq m_1^*$. Thus, $\text{MTTDL}(\phi_2,m) \leq \text{MTTDL}(\phi_2,m_1^*)$ for all $m \geq m_1^*$, which in turn implies that $m_2^* \leq m_1^*$. The proof for EAFDL is similar to that for MTTDL and is therefore omitted. ■

Figures 12 and 13 show the difference between the optimal codeword lengths for MTTDL and EAFDL. They demonstrate that the optimal codeword length for MTTDL is generally greater than or equal to that for EAFDL, with the difference being equal either to $z + 1$, the denominator of the storage efficiency fraction, or to 0. This implies that the optimal codeword lengths m^*_{EAFDL} for EAFDL are either equal to or slightly smaller than and adjacent to the optimal codeword lengths m^*_{MTTDL} for MTTDL. However, for small values of ϕ, such as $\phi = 0.001$, m^*_{MTTDL} can be smaller than m^*_{EAFDL}, as observed in Figure 12(d). This occurs only for certain group sizes that are smaller than 120, whereas for $k \geq 120$, the optimal codeword lengths follow the general trend discussed above. For example, in the case of $k = 120$, $\phi = 0.001$, and $s_{\text{eff}} = 1/2$, Figure 9(d) shows that the maximum value of MTTDL is achieved when the codeword length m is equal to 74, which implies that $m^*_{\text{MTTDL}} = 74$. Also, Figure 10(d) shows that the minimum value of EAFDL is achieved when the codeword length m is equal to 72, which implies that $m^*_{\text{EAFDL}} = 72$. The value of 72 is adjacent to 74 because when $s_{\text{eff}} = 1/2$, m cannot be equal to 73. Consequently, the difference of the optimal codeword lengths for EAFDL and MTTDL is given by 74 – 72 = 2, indicated by a circle in Figure 12(d). Similarly, for $k = 120$ and $s_{\text{eff}} = 2/3$, Figures 9(d) and 10(d) show that both the optimal MTTDL and the



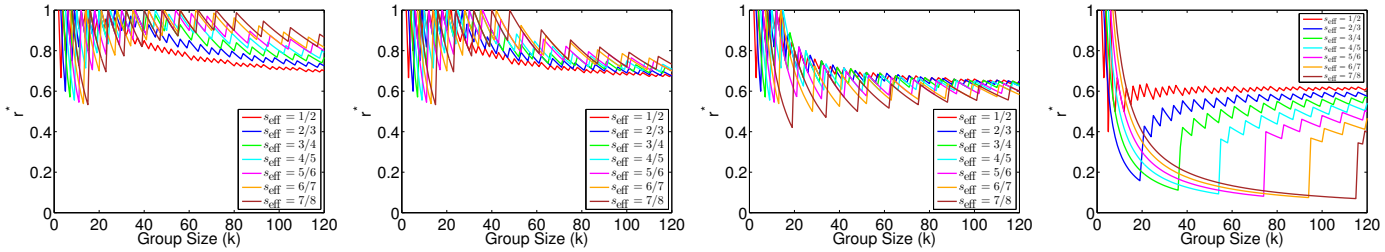
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 12. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for various storage efficiencies; $\lambda/\mu = 0.001$ and deterministic rebuild times.



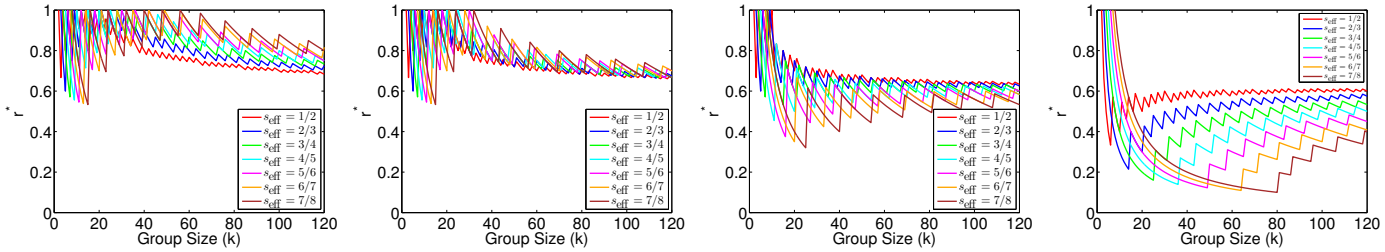
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 13. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 14. r^* for MTTDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

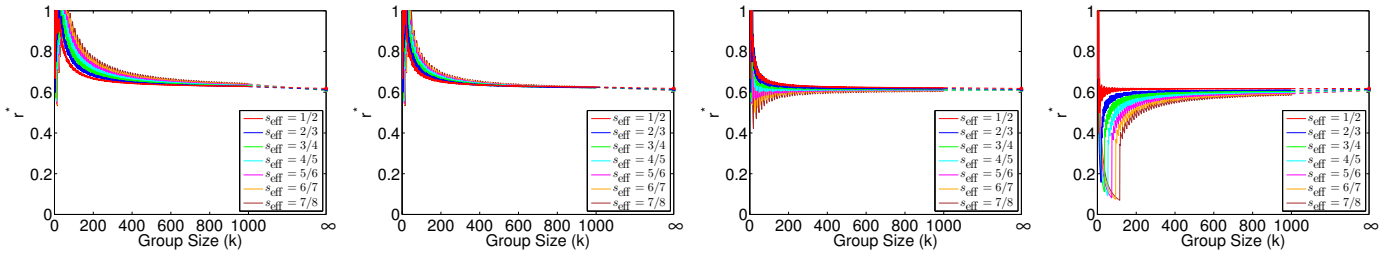
Figure 15. r^* for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.

optimal EAFDL are obtained when the codeword length is equal to 69, that is, $m^*_{\text{MTTDL}} = m^*_{\text{EAFDL}} = 69$. In this case, the difference of the optimal codeword lengths for EAFDL and MTTDL is equal to 0 , indicated by a circle in Figure 12(d).

To investigate the behavior of the optimal codeword length m^*_k with increasing group size k, we proceed by considering the normalized optimal codeword length r^*, namely, the ratio of m^*_k to k:

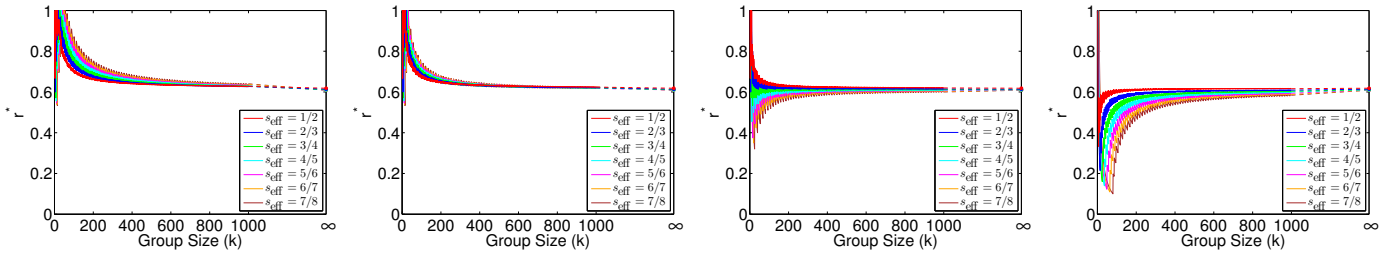
$$r^* \triangleq \frac{m^*_k}{k} . \tag{94}$$

The r^* values for the MTTDL and EAFDL metrics are shown in Figures 14 and 15, respectively, for various storage efficiencies and network rebuild bandwidth constraints. According to Proposition 14, for any storage efficiency s_{eff} and for any given group size k, the optimal codeword lengths and, consequently, the r^* values decrease with decreasing ϕ. Also, when the bandwidth constraint factor ϕ is small, the r^* values first decrease and then gradually increase with increasing k. The initial decrease is due to the fact that the optimal codeword length m^* remains fixed and equal to $z + 1$, which is the minimum possible codeword length for the storage efficiency



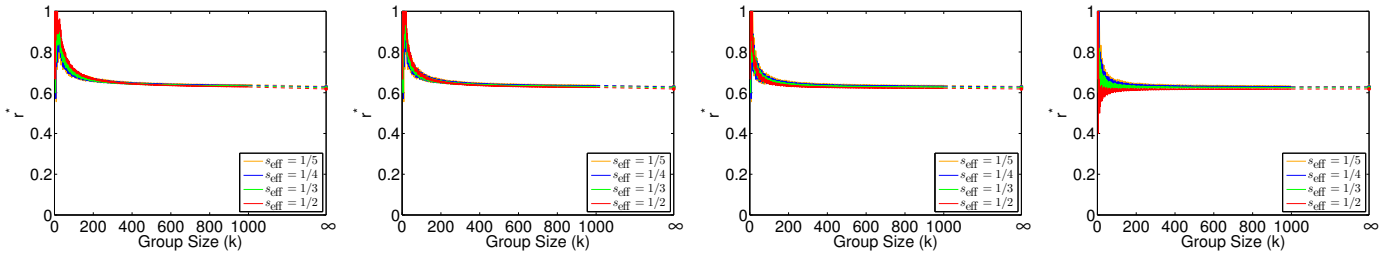
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 16. r^* for MTTDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



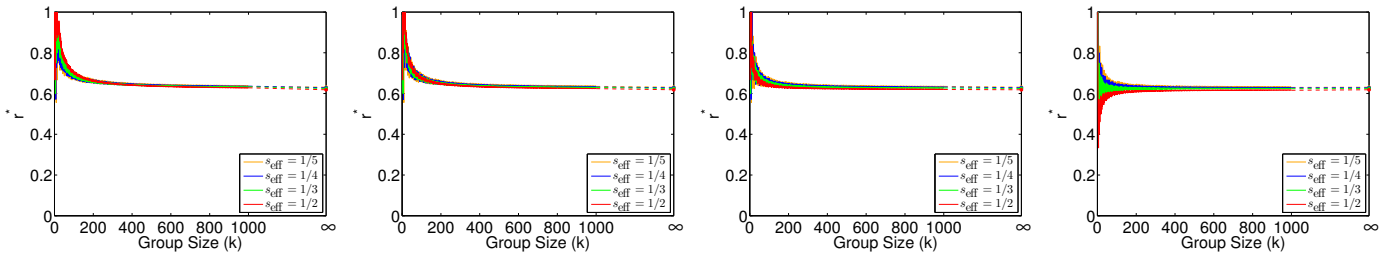
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 17. r^* for EAFDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 18. r^* for MTTDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 19. r^* for EAFDL vs. group size $k \to \infty$, $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and deterministic rebuild times.

fractions $z/(z+1)$, $z = 1, \ldots, 7$. For example, in the case of $s_{\text{eff}} = 7/8$ and $\phi = 0.001$, $m^* = 8$ for $k \leq 115$ in the case of MTTDL, or for $k \leq 80$ in the case of EAFDL, as shown in Figures 14(d) and 15(d), respectively.

The r^* values for the MTTDL and EAFDL metrics for various values of the storage efficiency s_{eff} and for large values of k are shown in Figures 16, 17, 18, and 19. We observe that, for a given storage efficiency and as k increases, the r^* values for MTTDL and EAFDL approach a common value, denoted by r^*_∞ and indicated by a small bullet. The r^*_∞ value is given by the following proposition.

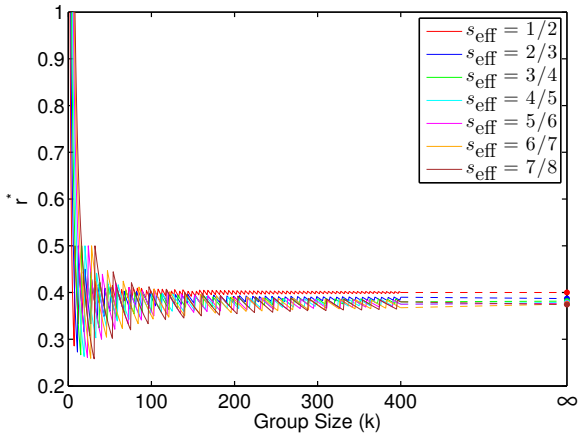
Proposition 15: As k increases, the r^* values for MTTDL and EAFDL approach r^*_∞ that satisfies the following equation:

$$Q(h, r^*_\infty) = 0 \,, \tag{95}$$

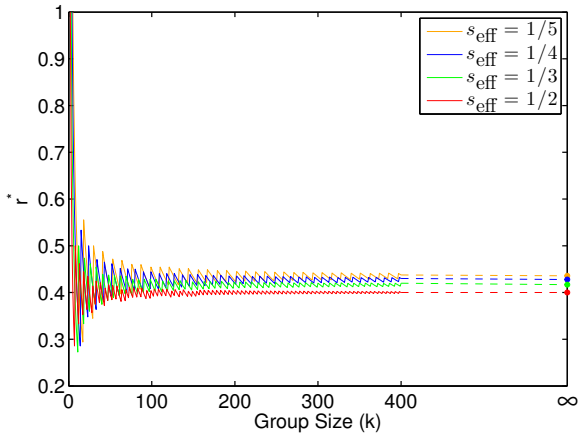
where $Q(h, x)$ is given by

$$Q(h,x) \triangleq hx + \log\left([(1-h)^{(1-h)^2} x^{h^2}]^x (1-hx)^{h(1-hx)}\right) , \tag{96}$$

and h and x are given by (49) and (50), respectively.



(a) $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5,\ 5/6, 6/7,$ and $7/8$

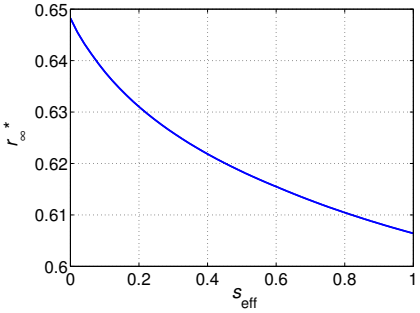


(b) $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$

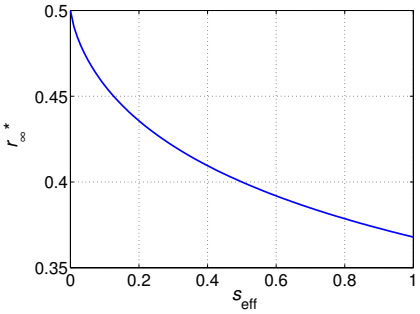
Figure 20. r^* for $E(H)$ vs. group size $k \to \infty$.

TABLE II. r^*_∞ VALUES FOR VARIOUS s_{eff}

s_{eff}			r^*_∞	
			MTTDL and EAFDL	$E(H)$
0	=	0	0.648419	0.5
10^{-4}	=	0.0001	0.648404	0.499795
10^{-3}	=	0.001	0.648265	0.498520
10^{-2}	=	0.01	0.646985	0.490770
10^{-1}	=	0.1	0.637940	0.456298
1/8	=	0.125	0.636043	0.450268
1/7	=	0.142857	0.634788	0.446383
1/6	=	0.166667	0.633224	0.441637
1/5	=	0.2	0.631212	0.435664
1/4	=	0.25	0.628500	0.427826
1/3	=	0.333333	0.624638	0.416889
1/2	=	0.5	0.618499	0.4
2/3	=	0.666667	0.613720	0.387097
3/4	=	0.75	0.611679	0.381625
4/5	=	0.8	0.610543	0.378586
5/6	=	0.833333	0.609818	0.376650
6/7	=	0.857143	0.609316	0.375307
7/8	=	0.875	0.608946	0.374322
$1 - 10^{-1}$	=	0.9	0.608440	0.372971
$1 - 10^{-2}$	=	0.99	0.606713	0.368368
$1 - 10^{-3}$	=	0.999	0.606549	0.367928
$1 - 10^{-4}$	=	0.9999	0.606532	0.367884
1	=	1	0.606531 = $1/\sqrt{e}$	0.367879 = $1/e$



(a) MTTDL and EAFDL



(b) $E(H)$

Figure 21. r^*_∞ vs. s_{eff}.

Proof: It follows from (51) that, for large values of k, $k^2\, W(h,x)/2$ is the dominating term. Thus, MTTDL is maximized when $W(h,x)$ is maximized. According to the arguments in Appendix F of [25], it therefore holds that [25, Equation (165)]

$$r^*_\infty = \arg \max_{0<x\leq 1} W(h,x)\ . \tag{97}$$

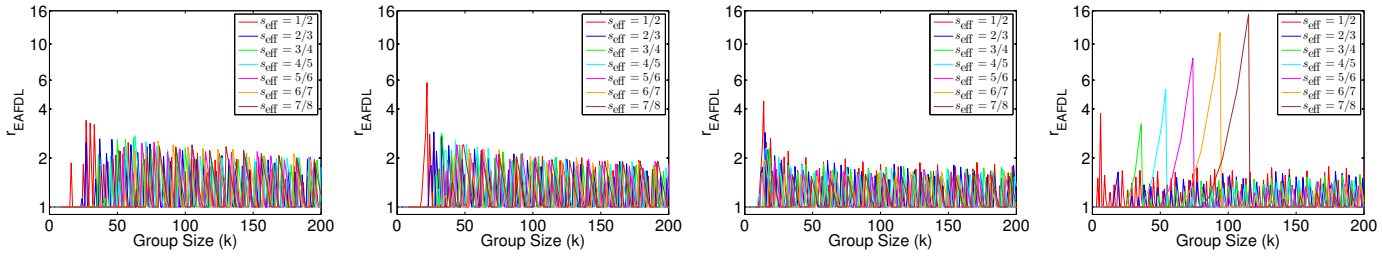
Consequently, r^*_∞ is obtained as the unique root of the equation $Q(h,x) = 0$, with respect to x, in the interval $(0,1]$, that is, [25, Equation (176)]

$$Q(h, r^*_\infty) = 0\ , \quad \text{with}\ \ r^*_\infty \in (0,1]\ , \tag{98}$$

where $Q(h,x)$ is given by (96) [25, Equation (105)]. From (55), it follows that the same rationale applies in the case of EAFDL. ■

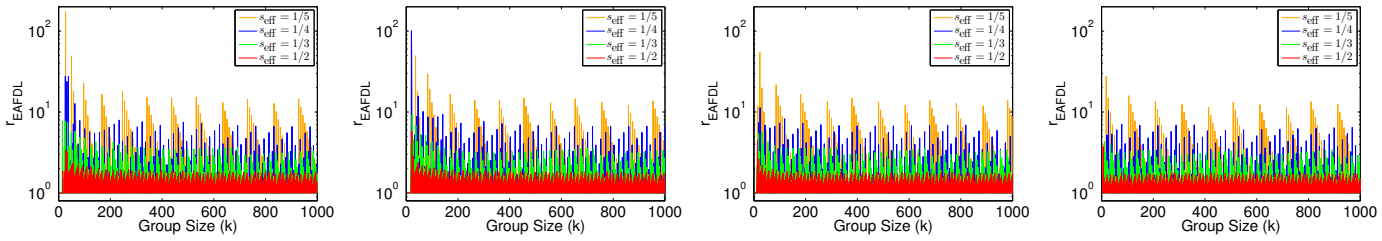
The r^* values for the $E(H)$ metric are shown in Figure 20 for various storage efficiencies and also for large group sizes. Clearly, the optimal codeword lengths for $E(H)$ are generally significantly shorter than those for MTTDL and EAFDL. We observe that, for a given storage efficiency and as k increases, the r^* values for $E(H)$ oscillate and approach a common value, denoted by r^*_∞ and indicated by a small bullet.

Remark 13: The r^*_∞ values are not affected by the bandwidth constraint factor ϕ and depend only on the storage efficiency s_{eff}. This is because the resulting reliability reduction factor θ, according to (47), is of the order $O(k)$, whereas the MTTDL and EAFDL reliability metrics, according to (51) and (55), are of the order $O(k^2)$, which is higher. Note that this also holds when the average network rebuild bandwidth is upper limited by $B_{\max}$, such that the bandwidth constraint factor ϕ is no longer constant, but, for large values of k and according



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 22. The EAFDL efficiency ratio r_{EAFDL} vs. group size; $\lambda/\mu = 0.001$ and deterministic rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 23. The EAFDL efficiency ratio r_{EAFDL} vs. group size; $\lambda/\mu = 0.001$ and deterministic rebuild times.

to (36), is inversely proportional to k. In this case, according to (119), the resulting reliability reduction factor θ is of the order $O(k\,\log(k))$, which is still smaller than the order $O(k^2)$. Also, according to Remark 1, the $E(H)$ metric is not affected by the bandwidth constraint factor, which implies that the r^*_∞ value for $E(H)$ is given by [25, Equation (107)]

$$r^*_\infty = \frac{1}{h + (1-h)^{-\frac{1-h}{h}}}\ , \tag{99}$$

with h given by (49).

The r^*_∞ values for the reliability metrics considered were initially derived in [25] and are included in this paper in Table II and Figure 21 for completeness. Note that the r^*_∞ values for the MTTDL and EAFDL are in the interval $[e^{-1/2} = 0.606, 0.648]$, whereas those for $E(H)$ are in the interval $[e^{-1} = 0.368, 0.5]$. Also, the r^*_∞ values decrease with increasing storage efficiency s_{eff}.

Next we examine the increase of the EAFDL metric if, instead of the optimal codeword lengths m^*_{EAFDL}, we use the codeword lengths m^*_{MTTDL} that optimize the MTTDL metric. From the preceding, it generally follows that m^*_{MTTDL} is either equal or adjacent to m^*_{EAFDL}, that is, $m^*_{\text{MTTDL}} = m^*_{\text{EAFDL}}+z+1$. We define the EAFDL *efficiency ratio* r_{EAFDL} as the ratio of $\text{EAFDL}(m^*_{\text{MTTDL}})$ to $\text{EAFDL}(m^*_{\text{EAFDL}})$, that is,

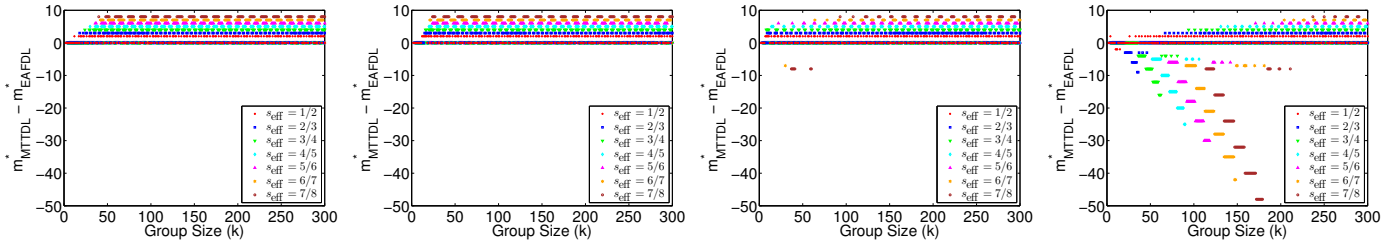
$$r_{\text{EAFDL}} \triangleq \frac{\text{EAFDL}(m^*_{\text{MTTDL}})}{\text{EAFDL}(m^*_{\text{EAFDL}})}\ , \tag{100}$$

where $\text{EAFDL}(m)$ denotes the EAFDL corresponding to a codeword length m.

The EAFDL efficiency ratios r_{EAFDL} as a function of k for various storage efficiencies and network rebuild bandwidth constraints are shown in Figures 22 and 23. We observe that for the storage efficiencies considered in Figure 22 and as k increases, the EAFDL efficiency ratios follow a periodic pattern and, for $\phi = 1$, are always less than a factor of 4. Moreover, as ϕ decreases, the EAFDL efficiency ratios tend to be less than a factor of 2, except in a few cases where they are significantly higher. Nevertheless, in all cases they are less than a factor of 16, which implies that using codewords of length m^*_{MTTDL} yields the maximum possible (optimal) MTTDL and also an EAFDL that is either optimal or of the same order. The maximum value is shown in Figure 22(d) obtained when $\phi = 0.001$, $s_{\text{eff}} = 7/8$, and $k = 115$. In this case, it holds that $m^*_{\text{MTTDL}} = 8$ and $m^*_{\text{EAFDL}} = 40$, such that $\text{EAFDL}(m^*_{\text{EAFDL}})/\lambda = \text{EAFDL}(40)/\lambda = 0.032$ and $\text{EAFDL}(m^*_{\text{MTTDL}})/\lambda = \text{EAFDL}(8)/\lambda = 0.487$, which in turn yields an EAFDL efficiency ratio r_{EAFDL} of 0.487 / 0.032 = 15.2. Also, as the storage efficiency decreases, the EAFDL efficiency ratio r_{EAFDL} increases, as shown in Figure 23. For any given storage efficiency and bandwidth constraint factor, r_{EAFDL} follows a periodic pattern and for $s_{\text{eff}} > 1/4 = 0.25$, r_{EAFDL} tends to be less than a factor of 10. Consequently, using codewords of length m^*_{MTTDL} yields an EAFDL that is of the same order of magnitude as the optimal one.

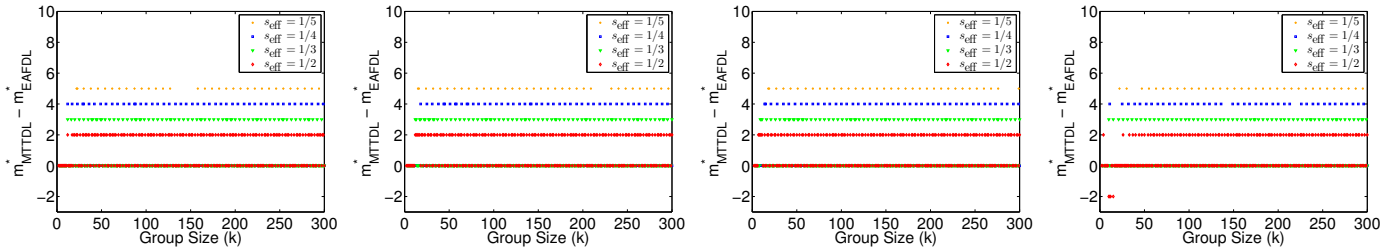
Next, we consider a system where the distribution of the rebuild time X is exponential, for which it holds that $E(X^{m-l}) = (m-l)!\,[E(X)]^{m-l}$. The combined effect of the network rebuild bandwidth constraint, the storage efficiency, and the codeword length on the reliability measures considered is similar to the case of deterministic rebuild times. Furthermore, similar to the case of deterministic rebuild times, the optimal codeword lengths m^*_{EAFDL} for EAFDL are generally either equal to or slightly shorter than and adjacent to the optimal codeword lengths m^*_{MTTDL} for MTTDL, as demonstrated in Figures 24 and 25.

The r^* values for the MTTDL and EAFDL metrics are shown in Figures 26 and 27, respectively, for various storage efficiencies and network rebuild bandwidth constraints. According to Remark 13 and Remark 13 of Appendix F of [25], as k increases, and for any storage efficiency and bandwidth constraint factor, the r^* values for MTTDL and



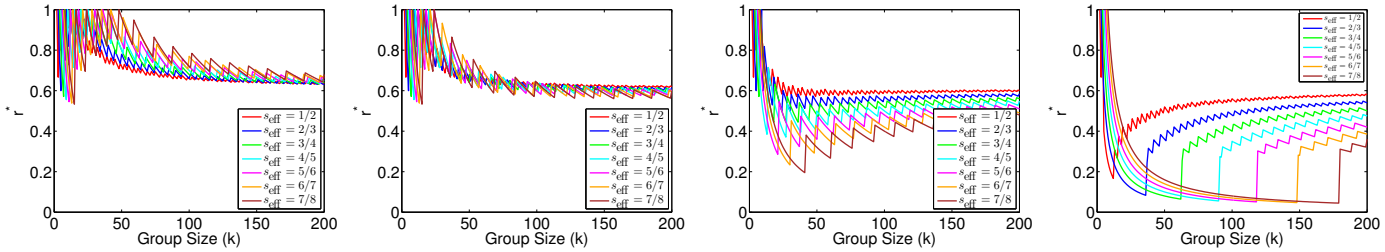
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 24. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for various storage efficiencies; $\lambda/\mu = 0.001$ and exponential rebuild times.



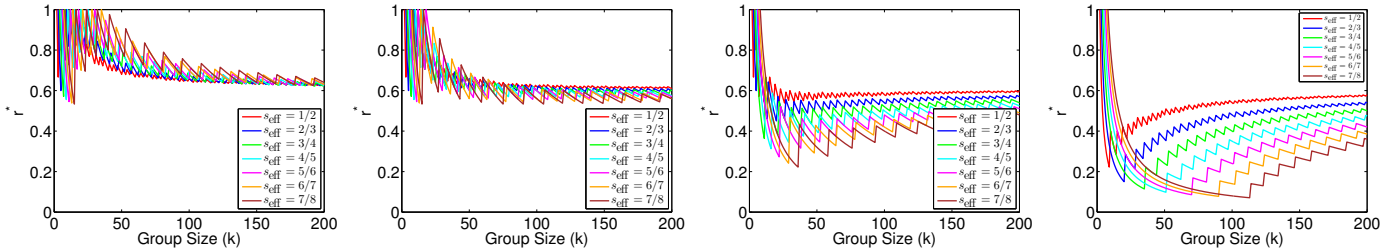
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 25. The difference between m^*_{MTTDL} and m^*_{EAFDL} vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and exponential rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 26. r^* for MTTDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and exponential rebuild times.

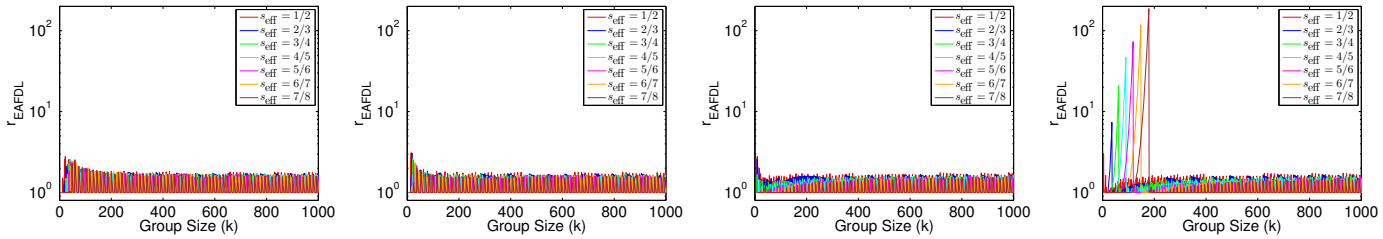


(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 27. r^* for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and exponential rebuild times.

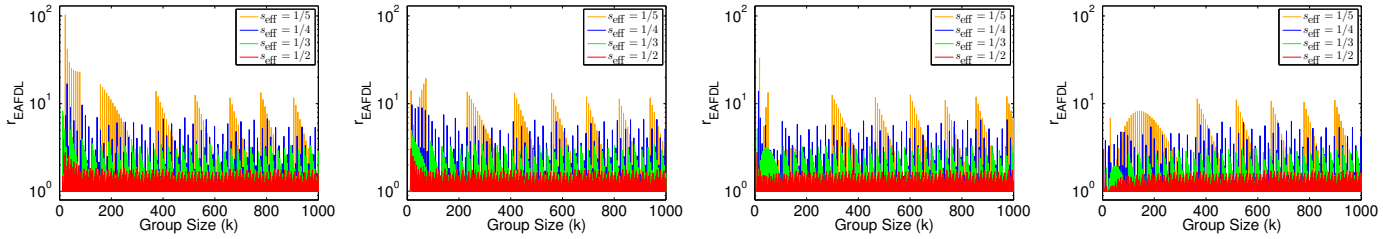
EAFDL approach a common value that is the same as the r^*_∞ value obtained in the case of deterministic rebuild times, which only depends on s_{eff} and is listed in Table II.

The EAFDL efficiency ratios r_{EAFDL} are shown in Figures 28 and 29 as a function of k for various storage efficiencies and network rebuild bandwidth constraints. We observe that as k increases, the EAFDL efficiency ratios follow a periodic pattern, as in the case of deterministic rebuild times. In particular, for the storage efficiencies considered in Figure 28, the EAFDL efficiency ratios tend to be less than a factor of 3, except in a few cases where they are significantly higher. The maximum value is shown in Figure 28(d) obtained when $\phi = 0.001$, $s_{\text{eff}} = 7/8$, and $k = 179$. In this case, it holds that $m^*_{\text{MTTDL}} = 8$ and $m^*_{\text{EAFDL}} = 56$, such that $\text{EAFDL}(m^*_{\text{EAFDL}})/\lambda = \text{EAFDL}(56)/\lambda = 0.00167$ and $\text{EAFDL}(m^*_{\text{MTTDL}})/\lambda = \text{EAFDL}(8)/\lambda = 0.31285$, which in turn yields an EAFDL efficiency ratio r_{EAFDL} of 0.31285 / 0.00167 = 187. Also, as the storage efficiency decreases, the EAFDL efficiency ratio r_{EAFDL} increases, as shown in Figure 29. Nevertheless, for $s_{\text{eff}} > 1/4 = 0.25$, r_{EAFDL} tends to be less than a factor of 10. Consequently, using codewords of length m^*_{MTTDL} yields an EAFDL that is of the same order



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 28. The EAFDL efficiency ratio r_{EAFDL} vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$ and exponential rebuild times.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 29. The EAFDL efficiency ratio r_{EAFDL} vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$ and exponential rebuild times.

of magnitude as the optimal one.

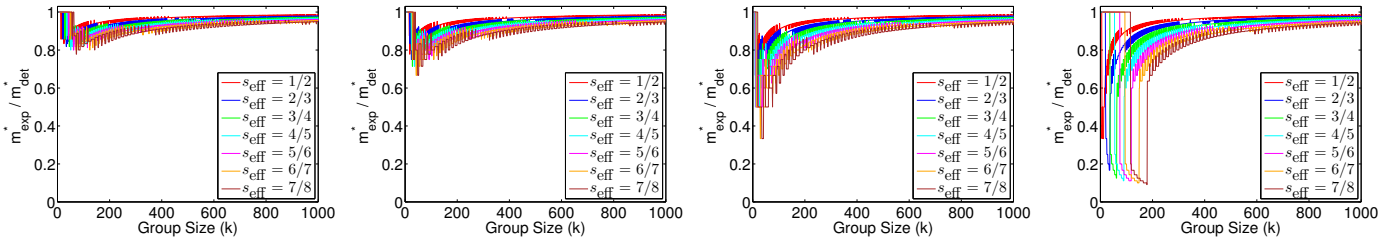
Figures 30 to 33 show the ratio of the optimal codeword length m^*_{exp} for the exponential distribution to the optimal codeword length m^*_{det} for the deterministic distribution for various storage efficiencies and network rebuild bandwidth constraints. We observe that this ratio never exceeds 1 and approaches 1 as k increases. This implies that, regardless of the rebuild bandwidth constraint, the optimal codeword length for the exponential distribution is generally smaller than the optimal codeword length for the deterministic distribution. This can be intuitively explained as follows. As previously mentioned, higher values of m result in a greater exposure degree to failure as each of the codewords is spread across a larger number of devices. The variation of exponentially distributed rebuild times results in increased vulnerability windows and therefore worse reliability. To reduce the exposure degree to failures, codewords should be spread across a shorter number of devices, which implies a shorter optimal codeword length. Also, lower values of the bandwidth constraint factor ϕ result in increased vulnerability windows, which in turn result in shorter optimal codeword lengths. In particular, we observe that the ratio of the optimal codeword lengths generally decreases with decreasing bandwidth constraint factor ϕ. However, when the optimal codeword lengths m^*_{exp} and m^*_{det} reach the value of the minimum codeword length, then the ratio becomes equal to 1, as shown in Figures 30(d) and 31(d) for the case of $k = 50$ and for $s_{\text{eff}} = 5/6, 6/7$ and $7/8$.

VII. Discussion

The symmetric and declustered data placement schemes reduce rebuild times by recovering data in parallel from the storage devices. In particular, for large-scale data storage systems, the rebuild times become extremely short. The model presented copes with this issue by considering the realistic case of network rebuild bandwidth constraints, which effectively prolong the duration of rebuild times.

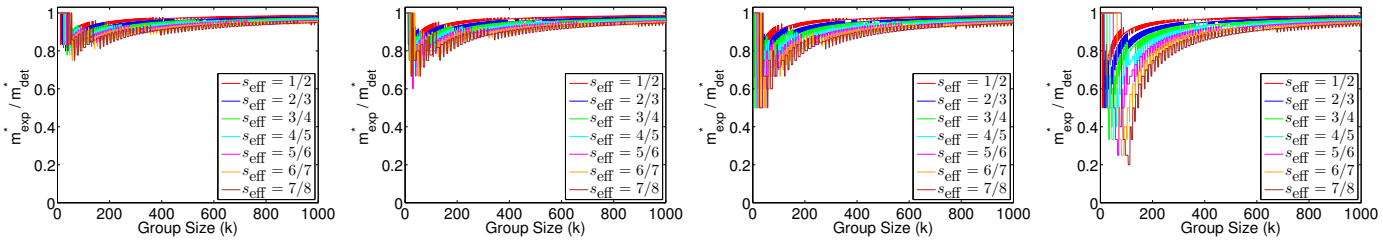
Although erasure coding schemes provide high data reliability and storage efficiency, the rebuild process involves I/O operations and network transfers that increase the consumption of device and network bandwidth. In particular, large MDS codes pose a challenge to the usage of network resources given that a lost symbol is recovered via an (m, l) erasure code by transferring a large number of l symbols from l surviving devices over the network. Although this may not be critical in purely archival tiers, recovering large amounts of data in active tiers results in additional traffic over increased time periods, which has an impact on the latency of the foreground workload and therefore affects system performance. This issue, also known as the *repair bandwidth problem*, has prompted the development of alternative erasure coding schemes that aim to reduce the amount of data transferred over the storage network during reconstruction (see [37][38] and references therein). They result in smaller amounts of data being read from the surviving devices and therefore in shorter rebuild times and higher reliabilities. However, in the case of *functional repairs*, a lost user-data symbol is replaced by an appropriate parity symbol, which now implies that reading such a user-data symbol can no longer be performed directly, but indirectly by accessing l symbols. Although this may be practical for archival tiers, it negatively affects the performance of the workloads encountered in active tiers. Therefore, emphasis is placed on *exact repairs* that preserve user data and maintain the erasure code in systematic form. The effect of these methods on system reliability is beyond the scope of this article and is a subject of further investigation.

The analytical findings of this work are relevant for the case of large erasure-coded data centers where a significant percentage of nodes fail each day [40]. Subsequently, the data recovery operations generate an excessive rebuild traffic that competes with the huge amount of traffic generated by the frequent access of a large number of storage devices [36]. To ensure a desired performance level, the network bandwidth devoted to the repair traffic must be contained. Furthermore,



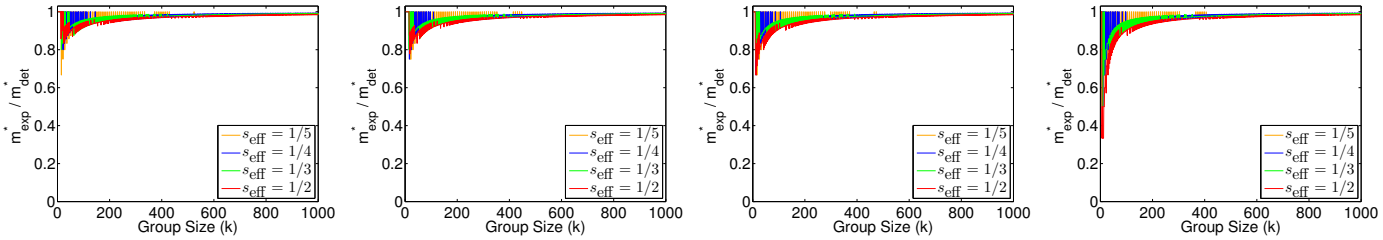
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 30. Ratio m^*_{exp} to m^*_{det} for MTTDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$.



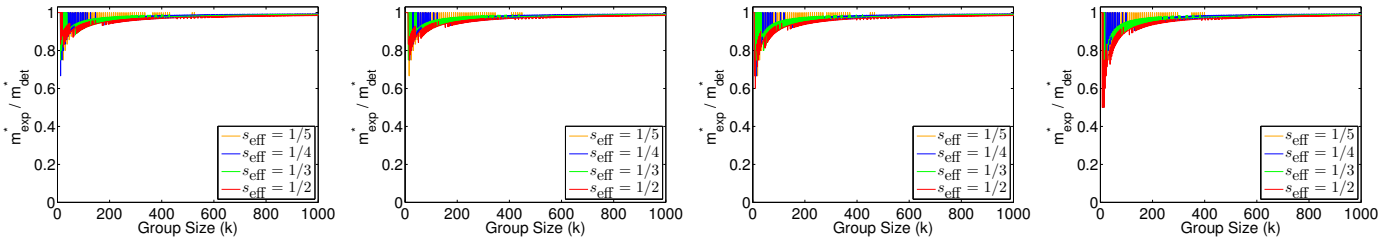
(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 31. Ratio m^*_{exp} to m^*_{det} for EAFDL vs. group size for $s_{\text{eff}} = 1/2, 2/3, 3/4, 4/5, 5/6, 6/7,$ and $7/8$; $\lambda/\mu = 0.001$.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 32. Ratio m^*_{exp} to m^*_{det} for MTTDL vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$.



(a) $\phi = 1$ (b) $\phi = 0.1$ (c) $\phi = 0.01$ (d) $\phi = 0.001$

Figure 33. Ratio m^*_{exp} to m^*_{det} for EAFDL vs. group size for $s_{\text{eff}} = 1/5, 1/4, 1/3,$ and $1/2$; $\lambda/\mu = 0.001$.

for performance reasons, the codeword length should be kept relatively short, otherwise a large number of parity updates will interfere with the normal user traffic, resulting in a performance degradation. More specifically, Google's GFS as well as QFS use an RS(9,6) code that achieves a storage efficiency of 66% [32, 41], Facebook uses an RS(14,10) code that achieves a storage efficiency of 71% [34], and Windows Azure uses an LRC(16,10) code, which is not an MDS code, that achieves a storage efficiency of 75% [33]. Note that these systems initially used a three-way replication by storing three copies of all data, which achieved a storage efficiency of 33%. Consequently, to keep the storage overhead low, the erasure-code parameter values should be chosen such that the storage efficiency is in the range from 0.66 to 0.75.

VIII. Conclusions

Data storage systems use erasure coding schemes to recover lost data and enhance system reliability. However, network rebuild bandwidth constraints may degrade reliability. A general methodology was applied for deriving the Mean Time to Data Loss (MTTDL) and the Expected Annual Fraction of Data Loss (EAFDL) reliability metrics analytically. Closed-form expressions capturing the effect of a network rebuild

bandwidth constraint were obtained for the symmetric, clustered and declustered data placement schemes. We established that the reliability of storage systems is adversely affected by the network rebuild bandwidth constraints. The declustered placement scheme was found to offer superior reliability in terms of both metrics. We subsequently conducted an investigation of the reliability achieved by this scheme under various codeword configurations. The results demonstrated that both metrics are optimized by similar codeword lengths. For large storage systems that use a declustered placement scheme, the optimized codeword lengths are about 60% of the storage system size, independently of the network rebuild bandwidth constraints. The analytical reliability expressions derived can be used to identify redundancy and recovery schemes as well as data placement configurations that can achieve high reliability. The results can also be used to adapt the data placement schemes when the available network rebuild bandwidth or the number of devices in the system changes so that the system maintains a high level of reliability.

Extending the methodology developed to derive the reliability of erasure coded systems in the presence of unrecoverable latent errors is a subject of further investigation. Moreover, owing to the parallelism of the rebuild process, the model considered here yields very short rebuild times for large system sizes. Taking into account the fact that the rebuild times cannot be shorter than the actual failure detection times requires a more sophisticated modeling effort, which is also part of future work.

APPENDIX A
OPTIMAL $\hat{k_s}$ FOR MTTDL$^{\text{sym}}$

Proof of Proposition 1.

As mentioned in Section II-B, the system comprises n/k disjoint groups of k devices. We first obtain the optimal value for k by relaxing the constraint that n/k be an integer, that is, by considering all the integer values for k in the interval $I_k = [m+1, n]$. We subsequently impose the constraint and obtain the optimal value $\hat{k_s}$. Note that the constraint that n/k be an integer translates to $k \in I_k \cap D_n$, where D_n is the set of all integers that divide n, as defined in (44). Also, k_m, as defined in (45), represents the smallest integer in the interval $I_k = [m+1, n]$ that divides n. Thus, $k_m \in D_n$, $n \in I_k$, $n \in D_n$, and therefore $n \in I_k \cap D_n$.

Considering l, m, and n to be fixed, it follows from (39) that $\text{MTTDL}_k^{\text{sym}}$ is approximately proportional to the function A_k given by

$$A_k \triangleq \prod_{u=1}^{m-l} (k-u)^{m-l-u} \prod_{u=1}^{m-l} \min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right) . \quad (101)$$

Consequently, the value of k in the interval $[m+1, n]$ that maximizes $\text{MTTDL}_k^{\text{sym}}$ also maximizes A_k. Depending on the values of m and l, we consider the following two cases:

Case 1: $m - l = 1$. From (101) it follows that $A_k = \min\left(\frac{\phi}{1-\frac{1}{k}}, 1\right) \leq 1$, which is decreasing in k.

Depending on the value of ϕ, the following three subcases are considered:

(a) $\phi \geq 1 - \frac{1}{n} \Leftrightarrow n \leq \frac{1}{1-\phi}$. In this case, A_k achieves its maximum value of 1 for all k for which $\frac{\phi}{1-\frac{1}{k}} \geq 1 \Leftrightarrow k \leq \frac{1}{1-\phi}$, that is, for all k that do not exceed the value of $\frac{1}{1-\phi}$. Thus, A_k is maximized for all k in the interval $I_k = [m+1, n]$. Subsequently, imposing the constraint that n/k be an integer translates to $k \in I_a = I_k \cap D_n$. Note that I_a is not empty because $n \in I$ and $n \in D_n$. Furthermore, $k_m \in I_a$.

(b) $1 - \frac{1}{n} > \phi \geq 1 - \frac{1}{k_m} \Leftrightarrow k_m \leq \frac{1}{1-\phi} < n$. In this case, A_k achieves its maximum value of 1 for all k for which $\frac{\phi}{1-\frac{1}{k}} \geq 1 \Leftrightarrow k \leq \frac{1}{1-\phi}$, that is, for all k that do not exceed the value of $\frac{1}{1-\phi}$. Thus, A_k is maximized for all k in the interval $I = [m+1, \frac{1}{1-\phi}]$, which also includes k_m. Subsequently, imposing the constraint that n/k be an integer translates to $k \in I_b = I \cap D_n$. Note that I_b is not empty because $k_m \in I$ and $k_m \in D_n$, and therefore $k_m \in I_b$.

(c) $1 - \frac{1}{k_m} > \phi \geq 1 - \frac{1}{m+1} \Leftrightarrow m+1 \leq \frac{1}{1-\phi} < k_m$. In this case, A_k achieves its maximum value of 1 for all k for which $\frac{\phi}{1-\frac{1}{k}} \geq 1 \Leftrightarrow k \leq \frac{1}{1-\phi}$, that is, for all k that do not exceed the value of $\frac{1}{1-\phi}$. Thus, A_k is maximized for all k in the interval $I = [m+1, \frac{1}{1-\phi}]$, which does not include k_m. Consequently, none of the values in the interval I divide n. Subsequently, imposing the constraint that n/k be an integer translates to considering values of k that exceed k_m, in which case $A_k = \frac{\phi}{1-\frac{1}{k}} < 1$. As A_k is decreasing in k, we deduce that A_k is maximized when $k = k_m$.

(d) $\phi < 1 - \frac{1}{m+1} \Leftrightarrow m+1 > \frac{1}{1-\phi}$. In this case, it holds that $\frac{\phi}{1-\frac{1}{n}} < \cdots < \frac{\phi}{1-\frac{1}{k}} < \cdots < \frac{\phi}{1-\frac{1}{m+1}} < 1$, for $m+1 < k < n$. Therefore, $A_k = \frac{\phi}{1-\frac{1}{k}}$, $\forall\, k \in [m+1, n]$, and A_k is maximized when $k = m+1$. Subsequently, imposing the constraint that n/k be an integer translates to $k = k_m$.

Case 2: $m - l \geq 2$. It holds that $k > m = (m-l) + l \geq m - l + 1$ and, therefore, $k \geq m - l + 2$. From (101) it follows that

$$A_k = \left[\prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u} \min(k\,\phi,\, k-u)\right] \min\left(\frac{\phi}{1-\frac{m-l}{k}}, 1\right) . \quad (102)$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi \geq 1 - \frac{m-l}{k} \Leftrightarrow k \leq \frac{m-l}{1-\phi}$. In this case, it holds that $\min\left(\frac{\phi}{1-\frac{m-l}{k}}, 1\right) = 1$, and

$$A_k = \prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u} \min(k\,\phi,\, k-u) . \quad (103)$$

Note that each of the terms in the product is increasing in k, which implies that A_k is also increasing in k, for $k \leq k_f$, where $k_f \triangleq \lfloor \frac{m-l}{1-\phi} \rfloor$.

(b) $\phi \leq 1 - \frac{m-l}{k} \Leftrightarrow k \geq \frac{m-l}{1-\phi}$. In this case, it holds that $\min\left(\frac{\phi}{1-\frac{m-l}{k}}, 1\right) = \frac{\phi}{1-\frac{m-l}{k}}$. Also, for $u < m-l$, it holds

that $k-u > k-(m-l) = k\,(1-\frac{m-l}{k}) \geq k\,\phi$, such that $\min(k\,\phi,\, k-u) = k\,\phi$. From (102), it now follows that

$$A_k = \left[\prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u}\; k\,\phi\right] \frac{\phi}{1-\frac{m-l}{k}} = \phi^{m-l}\left[\prod_{u=1}^{m-l-1} (k-u)^{m-l-1-u}\right] \frac{{k}^{m-l}}{k-(m-l)}\,. \quad (104)$$

We proceed by recognizing that the last term in (104) is increasing in k, for $k \geq m-l+2$. This is a direct consequence of the fact that the function $f(x) = \frac{x^{m-l}}{x-(m-l)}$ is increasing in $x \in [\frac{(m-l)^2}{m-l-1}, \infty]$, and the fact that $m-l+2 \geq \frac{(m-l)^2}{m-l-1}$ for $m-l \geq 2$. Moreover, each of the terms in the product is increasing in k, which implies that A_k is also increasing in k, for $k \geq k_c$, where $k_c \triangleq \lceil \frac{m-l}{1-\phi} \rceil$.

To conclude that A_k is increasing in the entire range of k, it suffices to show that $A_{k_f} \leq A_{k_c}$. Clearly, this condition holds when $k_f = k_c$, that is, when $\frac{m-l}{1-\phi}$ is an integer. If $\frac{m-l}{1-\phi}$ is not an integer, then it holds that

$$k_f < \frac{m-l}{1-\phi} < k_c = k_f + 1\,, \quad (105)$$

which in turn implies that

$$1 - \frac{m-l}{k_f} < \phi < \phi_c \triangleq 1 - \frac{m-l}{k_c}\,. \quad (106)$$

Furthermore, using the relation $k_f \geq m-l+2 > m-l-1$, we deduce that $\frac{m-l-1}{k_f} < \frac{m-l}{k_f+1} = \frac{m-l}{k_c} = 1-\phi_c$, that is,

$$\phi_c < 1 - \frac{m-l-1}{k_f}\,. \quad (107)$$

For $u \leq m-l-1$, and combining (106) and (107) yields $k_f - u \geq k_f - (m-l-1) = k_f\,(1-\frac{m-l-1}{k_f}) > k_f\,\phi_c > k_f\,\phi$, such that $\min(k_f\,\phi,\, k_f - u) = k_f\,\phi$. Thus, (103) can be written as follows:

$$\begin{aligned} A_{k_f} &= \prod_{u=1}^{m-l-1} [(k_f-u)^{m-l-1-u} k_f\,\phi] \\ &= \phi^{m-l-1}\left[\prod_{u=1}^{m-l-1} (k_f-u)^{m-l-1-u}\right] {k_f}^{m-l-1} \\ &\overset{(106)}{<} \phi^{m-l}\left[\prod_{u=1}^{m-l-1} (k_f-u)^{m-l-1-u}\right] \frac{{k_f}^{m-l}}{k_f-(m-l)}\,. \end{aligned} \quad (108)$$

Also, from (104) we get

$$A_{k_c} = \phi^{m-l}\left[\prod_{u=1}^{m-l-1} (k_c-u)^{m-l-1-u}\right] \frac{{k_c}^{m-l}}{k_c-(m-l)}\,. \quad (109)$$

From (108) and (109), and given that $k_f < k_c$, it follows that $A_{k_f} < A_{k_c}$.

From the above, we conclude that when $m-l \geq 2$, A_k is increasing in k and, therefore, is maximized when $k = n$. ■

APPENDIX B
APPROXIMATE DERIVATION OF $\theta^{\,\mathrm{sym}}$

Proof of Lemma 1.

From (3) and (38), and using (49) and (50), it follows that

$$\log(\theta^{\,\mathrm{sym}}) = \sum_{u=1}^{hxk} \log\left(\min\left(\frac{\phi}{1-\frac{u}{k}}, 1\right)\right)\,. \quad (110)$$

Given that $\frac{\phi}{1-\frac{u}{k}} \leq 1 \;\Leftrightarrow\; u \leq (1-\phi)k$, (110) yields

$$\begin{aligned} \log(\theta^{\,\mathrm{sym}}) &= \sum_{u=1}^{\widehat{\phi}k} \log\left(\frac{\phi}{1-\frac{u}{k}}\right) \\ &= \sum_{u=1}^{\widehat{\phi}k} \log(\phi) - \sum_{u=1}^{\widehat{\phi}k} \log\left(1-\frac{u}{k}\right) \\ &= \widehat{\phi}\,k\log(\phi) - \sum_{u=1}^{\widehat{\phi}k} \log\left(1-\frac{u}{k}\right)\,, \end{aligned} \quad (111)$$

where $\widehat{\phi} = \min(1-\phi, hx)$ as defined in (48). For large values of k, the preceding summation can be approximated using Lemma 1 of [25], which states that for small values of ϵ, that is, when ϵ approaches 0, and for any function $f(y)$, it holds that [25, Equation (122)]

$$\epsilon \sum_{j=1}^{\alpha/\epsilon} f(j\epsilon) \;\approx\; \int_{\frac{\epsilon}{2}}^{\alpha+\frac{\epsilon}{2}} f(y)\,dy\,, \qquad \forall\,\alpha \in \mathbb{R}\,. \quad (112)$$

For $\alpha = \widehat{\phi}$ and $f(y) = \log(1-y)$, (112) yields

$$\epsilon \sum_{j=1}^{\widehat{\phi}/\epsilon} \log(1-j\epsilon) \;\approx\; \int_{\frac{\epsilon}{2}}^{\widehat{\phi}+\frac{\epsilon}{2}} \log(1-y)\,dy\,. \quad (113)$$

Also, from Equations (128), (129), and (133) of [25], it follows that

$$\begin{aligned} \int_{\frac{\epsilon}{2}}^{\widehat{\phi}+\frac{\epsilon}{2}} \log(1-y)\,dy &= \log\left(\frac{(1-\frac{\epsilon}{2})^{1-\frac{\epsilon}{2}}}{(1-\widehat{\phi}-\frac{\epsilon}{2})^{1-\widehat{\phi}-\frac{\epsilon}{2}}}\right) - \widehat{\phi} \\ &\approx\; \log\left(\frac{1}{(1-\widehat{\phi})^{1-\widehat{\phi}}}\right) - \widehat{\phi} + \frac{\log(1-\widehat{\phi})}{2}\,\epsilon - \frac{\widehat{\phi}}{8(1-\widehat{\phi})}\,\epsilon^2\,. \end{aligned} \quad (114)$$

Substituting (114) into (113), and setting $\epsilon = 1/k$, yields

$$\begin{aligned} \sum_{j=1}^{\widehat{\phi}k} \log\left(1-\frac{j}{k}\right) \;\approx\; &-\left[\log\left((1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right] k \\ &+ \frac{1}{2}\log(1-\widehat{\phi}) - \frac{\widehat{\phi}}{8(1-\widehat{\phi})\,k}\,. \end{aligned} \quad (115)$$

Note that for large values of k, the last term of the right-hand side of (115) is negligible and therefore can be ignored. Substituting (115) into (111) yields the following approximation:

$$\begin{aligned} \log\left(\theta^{\,\mathrm{sym}}_{\,\mathrm{approx}}\right) \approx\; &\widehat{\phi}\,\log(\phi)\,k \\ &+ \left[\log\left((1-\widehat{\phi})^{1-\widehat{\phi}}\right) + \widehat{\phi}\right] k - \frac{1}{2}\log(1-\widehat{\phi})\,, \end{aligned} \quad (116)$$

which yields (47). ■

Remark 14: When the average network rebuild bandwidth is upper limited by $B_{\max}$, the bandwidth constraint factor ϕ is no longer constant, but, for large values of k and according to (36), is given by

$$\phi = \frac{B_{\max}}{k\,b} \overset{(9)}{=} \frac{N_b}{k}\,, \tag{117}$$

which in turn implies that ϕ tends to 0 as k increases. Consequently, for large values of k, (48) yields

$$\widehat{\phi} = hx\,. \tag{118}$$

Substituting (117) and (118) into (116) yields

$$\begin{aligned}\log\left(\theta^{\text{sym}}_{\text{approx}}\right) \approx\ & hx\,\log\left(\frac{N_b}{k}\right)k \\ &+ \left[\log\left((1-hx)^{1-hx}\right) + hx\right]k - \frac{1}{2}\log(1-hx) \\ \approx\ & -hx\,k\,\log(k) + hx\log(N_b)\,k \\ &+ \left[\log\left((1-hx)^{1-hx}\right) + hx\right]k - \frac{1}{2}\log(1-hx)\,.\end{aligned} \tag{119}$$

Appendix C
Optimal $\hat{k}$ for MTTDL

Proof of Proposition 10.

Depending on the values of m and l, we consider the following three cases:

Case 1: $m-l=1$. Substituting $l=m-1$ into (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{m-1}{m}\cdot\frac{\min\left(\frac{\phi}{1-\frac{1}{\hat{k_s}}},1\right)}{\min\left(\frac{\phi}{1-\frac{1}{m}},1\right)} < 1\,. \tag{120}$$

The inequality holds because $\frac{m-1}{m}<1$ and $\min\left(\frac{\phi}{1-\frac{1}{\hat{k_s}}},1\right) \leq \min\left(\frac{\phi}{1-\frac{1}{m}},1\right)$, given that $\hat{k_s} \geq m+1 > m$ and, therefore, $\frac{\phi}{1-\frac{1}{\hat{k_s}}} < \frac{\phi}{1-\frac{1}{m}}$. Consequently, the MTTDL is maximized by the clustered placement scheme.

Case 2: $m-l=2$. This implies that $n/2 \geq m = l+2 \geq 3$. Consequently, $1 \leq \frac{2(m-2)}{(m-1)} = \frac{(m-2)(2m-2)}{(m-1)^2} \leq \frac{(m-2)(n-2)}{(m-1)^2} = \frac{(m-2)n^2}{(n-2)(m-1)^2}\left(1-\frac{2}{n}\right)^2$, which in turn implies that

$$G \leq 1-\frac{2}{n} < 1-\frac{1}{n}\,, \tag{121}$$

where

$$G \triangleq G(m,n) = \sqrt{\frac{n-2}{m-2}}\ \frac{m-1}{n}\,. \tag{122}$$

For $m-l=2$, according to (43), it holds that $\text{MTTDL}^{\text{sym}}_{\hat{k_s}} = \text{MTTDL}^{\text{sym}}_{n}$ and, subsequently, (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-2)(n-1)\min\left(\frac{\phi}{1-\frac{1}{n}},1\right)\min\left(\frac{\phi}{1-\frac{2}{n}},1\right)}{(m-1)^2\ \min\left(\frac{\phi}{1-\frac{2}{m}},1\right)^2}\,. \tag{123}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > G$. In this case, it holds that $\frac{G}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{1}{n}}$ and $\frac{G}{1-\frac{2}{n}} < \frac{\phi}{1-\frac{2}{n}}$. Also, it holds from (121) that $\frac{G}{1-\frac{1}{n}} < \frac{G}{1-\frac{2}{n}} \leq 1$. Consequently, $\min\left(\frac{\phi}{1-\frac{1}{n}},1\right) > \frac{G}{1-\frac{1}{n}}$, and $\min\left(\frac{\phi}{1-\frac{2}{n}},1\right) \geq \frac{G}{1-\frac{2}{n}}$. Moreover, from (123), and using the fact that $\min\left(\frac{\phi}{1-\frac{2}{m}},1\right) \leq 1$, it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > \frac{(m-2)(n-1)\ \frac{G}{1-\frac{1}{n}}\ \frac{G}{1-\frac{2}{n}}}{(m-1)^2} \overset{(122)}{=} 1\,. \tag{124}$$

(b) $\phi \leq G$. In this case, it holds that $\frac{\phi}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{2}{n}} \leq \frac{\phi}{G} \leq 1$. It follows from (123) that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-2)(n-1)\ \frac{\phi}{1-\frac{1}{n}}\ \frac{\phi}{1-\frac{2}{n}}}{(m-1)^2\ \min\left(\frac{\phi}{1-\frac{2}{m}},1\right)^2}\,. \tag{125}$$

Depending on the value of ϕ, the following two subcases are considered:

(i) $\phi \geq 1-\frac{2}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{2}{m}},1\right)=1$, and from (125) it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < \frac{(m-2)(n-1)\ \frac{G}{1-\frac{1}{n}}\ \frac{G}{1-\frac{2}{n}}}{(m-1)^2} \overset{(122)}{=} 1\,. \tag{126}$$

Also, in this case it holds that $1-\frac{2}{m} \leq G$. Note that for $n/2 \geq m$, it holds that

$$\left(\frac{1-\frac{2}{m}}{G}\right)^2 = \frac{(m-2)^3\,n^2}{(m-1)^2\,m^2\,(n-2)} \geq \frac{2\,(m-2)^3}{(m-1)^3}\,. \tag{127}$$

We now deduce that $m \leq 5$, because for $m \geq 6$, it holds that $\frac{2\,(m-2)^3}{(m-1)^3} > 1$ and, therefore, $1-\frac{2}{m} > G$, which is a contradiction. It turns out that for $m=3$ and $m=4$, the ratio $(1-\frac{2}{m})/G$ is less than 1 or, equivalently, $1-\frac{2}{m} < G$, when $n \leq 33$ and $n \leq 12$, respectively. For $m=5$, this ratio is less than 1 when $n=10$.

(ii) $\phi < 1-\frac{2}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{2}{m}},1\right) = \frac{\phi}{1-\frac{2}{m}}$, and (125) yields

$$\begin{aligned} r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} &\approx \frac{(m-2)(n-1)\ \frac{\phi}{1-\frac{1}{n}}\ \frac{\phi}{1-\frac{2}{n}}}{(m-1)^2\ \left(\frac{\phi}{1-\frac{2}{m}}\right)^2} \\ &= \frac{(m-2)^3\,n^2}{(m-1)^2\,m^2\,(n-2)} \overset{(122)}{=} \left(\frac{1-\frac{2}{m}}{G}\right)^2\,. \end{aligned} \tag{128}$$

As argued above, it holds for $m \geq 6$ that $1-\frac{2}{m} > G$ and, therefore, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > 1$. For $3 \leq m \leq 5$, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ may be less than 1. In particular, for $m=3$ and $m=4$, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ is less than 1 when $n \leq 33$ and $n \leq 12$, respectively. For $m=5$, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}}$ is less than 1 when $n=10$.

From the results obtained in the preceding two subcases, we conclude that, when $m-l=2$, the MTTDL is maximized

by the clustered placement scheme ($r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < 1$) only in the following cases:

1) $m = 3$, $n = 3\,j$ with $2 \le j \le 11$, and $\phi < G = \frac{2\sqrt{n-2}}{n}$,
2) $m = 4$, $n = 8$, and $\phi < G = 3\sqrt{3}/8 = 0.649$,
3) $m = 4$, $n = 12$, and $\phi < G = \sqrt{5}/4 = 0.559$, and
4) $m = 5$, $n = 10$, and $\phi < G = 4\sqrt{2}/(5\sqrt{3}) = 0.653$.

In these cases, $\hat{k} = m$, whereas in all other cases, the MTTDL is maximized by the declustered placement scheme ($\hat{k} = n$).

Case 3: $m-l = 3$. This implies that $n/2 \ge m = l+3 \ge 4$. Thus, $n - 3 > m - 2$ or $(n-3)^2/(m-2)^2 > 1$. Also, for $m \ge 4$, it holds that $\frac{(m-3)(2m-1)}{(m-1)(m-2)} > 1$. Consequently, $1 < \frac{(n-3)^2(m-3)(2m-1)}{(m-2)^2(m-1)(m-2)} \le \frac{(n-3)^2(m-3)(n-1)}{(m-1)(m-2)^3} = \frac{(m-3)(n-1)n^3}{(n-3)(m-1)(m-2)^3}\left(1-\frac{3}{n}\right)^3$, which in turn implies that

$$Q < 1 - \frac{3}{n} < 1 - \frac{2}{n} < 1 - \frac{1}{n}\,, \tag{129}$$

where

$$Q \triangleq Q(m,n) = \sqrt[3]{\frac{(n-3)(m-1)}{(m-3)(n-1)}}\;\frac{m-2}{n}\,. \tag{130}$$

For $m-l = 3$, according to (43), it holds that $\text{MTTDL}^{\text{sym}}_{\hat{k_s}} = \text{MTTDL}^{\text{sym}}_{n}$ and, subsequently, (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-3)(n-1)^2(n-2)}{(m-1)(m-2)^3} \cdot \frac{\min\left(\frac{\phi}{1-\frac{1}{n}},1\right)\min\left(\frac{\phi}{1-\frac{2}{n}},1\right)\min\left(\frac{\phi}{1-\frac{3}{n}},1\right)}{\min\left(\frac{\phi}{1-\frac{3}{m}},1\right)^3}\,. \tag{131}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > Q$. In this case, it holds that $\frac{Q}{1-\frac{u}{n}} < \frac{\phi}{1-\frac{u}{n}}$, for $u = 1,2,3$. Also, it holds from (129) that $\frac{Q}{1-\frac{u}{n}} < 1$, for $u = 1,2,3$. Consequently, $\min\left(\frac{\phi}{1-\frac{u}{n}},1\right)$, for $u = 1,2,3$. Moreover, from (131), and using the fact that $\min\left(\frac{\phi}{1-\frac{3}{m}},1\right) < 1$, it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > \frac{(m-3)(n-1)^2(n-2)\;\frac{Q}{1-\frac{1}{n}}\;\frac{Q}{1-\frac{1}{n}}\;\frac{Q}{1-\frac{3}{n}}}{(m-1)(m-2)^3} \overset{(130)}{=} 1\,. \tag{132}$$

(b) $\phi \le Q$. In this case, it holds that $\frac{\phi}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{2}{n}} < \frac{\phi}{1-\frac{3}{n}} < \frac{\phi}{Q} < 1$. It follows from (131) that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-3)(n-1)^2(n-2)\;\frac{\phi}{1-\frac{1}{n}}\;\frac{\phi}{1-\frac{2}{n}}\;\frac{\phi}{1-\frac{3}{n}}}{(m-1)(m-2)^3\;\min\left(\frac{\phi}{1-\frac{3}{m}},1\right)^3}\,. \tag{133}$$

Depending on the value of ϕ, the following two subcases are considered:

(i) $\phi \ge 1-\frac{3}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{3}{m}},1\right) = 1$, and from (133) it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < \frac{(m-3)(n-1)^2(n-2)\;\frac{Q}{1-\frac{1}{n}}\;\frac{Q}{1-\frac{2}{n}}\;\frac{Q}{1-\frac{3}{n}}}{(m-1)(m-2)^3} \overset{(130)}{=} 1\,. \tag{134}$$

Also, in this case it holds that $1-\frac{3}{m} \le Q$. Note that for $n/2 \ge m$, it holds that

$$\left(\frac{1-\frac{3}{m}}{Q}\right)^3 = \frac{(m-3)^4\,(n-1)\,n^3}{m^3\,(m-1)\,(m-2)^3\,(n-3)} \ge \left[4\left(\frac{m-3}{m-2}\right)^3\right]\left[\frac{2(m-3)(2m-1)}{(m-1)(2m-3)}\right]\,. \tag{135}$$

We now deduce that $m = 4$, because for $m \ge 5$, each of the terms in the two brackets is greater than one and, therefore, $1-\frac{3}{m} > Q$, which is a contradiction. It turns out that for $m = 4$, the ratio $(1-\frac{3}{m})/Q$ is less than one or, equivalently, $1-\frac{3}{m} < Q$, only when $n = 8$.

(ii) $\phi < 1-\frac{3}{m}$. Then, it holds that $\min\left(\frac{\phi}{1-\frac{3}{m}},1\right) = \frac{\phi}{1-\frac{3}{m}}$, and (133) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \frac{(m-3)(n-1)^2(n-2)\;\frac{\phi}{1-\frac{1}{n}}\;\frac{\phi}{1-\frac{2}{n}}\;\frac{\phi}{1-\frac{3}{n}}}{(m-1)(m-2)^3\;\min\left(\frac{\phi}{1-\frac{3}{m}},1\right)^3} = \left(\frac{1-\frac{3}{m}}{Q}\right)^3 \overset{(130)}{=} \frac{(m-2)^3\,n^2}{(m-1)^2\,m^2\,(n-2)}\,. \tag{136}$$

As argued above, it holds for $m \ge 5$ that $1-\frac{3}{m} > Q$ and, therefore, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > 1$. In fact, $r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < 1$ only when $m = 4$ and $n = 8$.

From the results obtained in the preceding two subcases, we conclude that, when $m-l = 3$, MTTDL is maximized by the clustered placement scheme ($r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} < 1$) only when $l = 1$, $m = 4$, $n = 8$, and $\phi < Q = \sqrt[3]{15}/(4\sqrt[3]{7}) = 0.322$. In all other cases, MTTDL is maximized by the declustered placement scheme.

Case 4: $m-l \ge 4$. We deduce from (89) that $n-m \ge m$ and

$$n-m+l \ge m+l = m-l+2l \ge 4+2l \ge 6\,. \tag{137}$$

Let us define

$$R_M \triangleq R_M(l,m,n) = \frac{l+1}{n}\left[\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u-1}\right]^{\frac{1}{m-l}}\,. \tag{138}$$

Next, we will show that

$$S_M \triangleq \left(\frac{1-\frac{m-l}{m}}{R_M}\right)^{m-l} = \left(\frac{l}{m\,R_M}\right)^{m-l} > 1\,. \tag{139}$$

Substituting (138) into (139) yields

$$S_M = \left[\frac{l\,n}{(l+1)\,m}\right]^{m-l}\;\prod_{u=1}^{m-l}\left(\frac{n-u}{m-u}\right)^{m-l-u-1}\,. \tag{140}$$

It follows from (89) that $\frac{l}{l+1} \geq \frac{1}{2}$ and $\frac{n}{m} \geq 2$. Consequently, the term in brackets is greater than or equal to 1. Next, we will show that the product is greater than 1. For $m-l \geq 4$, the product can be written as follows:

$$\prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} = \prod_{u=1}^{m-l-4} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} \prod_{u=m-l-3}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} . \tag{141}$$

Clearly, for $n > m$, the first product is greater than or equal to 1. The second product is greater than 1 because it can be written as follows:

$$\begin{aligned} &\prod_{u=m-l-3}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} \\ &= \left[\frac{n-(m-l-3)}{m-(m-l-3)}\right]^2 \cdot \frac{n-(m-l-2)}{m-(m-l-2)} \cdot \frac{m-(m-l)}{n-(m-l)} \\ &= 1 + \frac{(n-m)[l(n-m+l)(n-m+2l+8)-18]}{(l+3)(l+2)(n-m+l)} \\ &\overset{(89)(137)}{\geq} 1 + \frac{(n-m)[6\,(6+l+8)-18]}{(l+3)(l+2)(n-m+l)} > 1 . \end{aligned} \tag{142}$$

Inequality (139) is a direct consequence of (140), (141), and (142).

We deduce from (139) that

$$R_M < \frac{l}{m} = 1 - \frac{m-l}{m} < 1 - \frac{m-l}{n} < \cdots < 1 - \frac{1}{n} . \tag{143}$$

For $m-l \geq 4$, according to (43), it holds that $\text{MTTDL}^{\text{sym}}_{\hat{k}_s} = \text{MTTDL}^{\text{sym}}_{n}$ and, subsequently, (90) yields

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} \approx \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \Bigg/ \min\left(\frac{m\,\phi}{l}, 1\right)^{m-l} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u} \min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right) . \tag{144}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > R_M$. In this case, it holds that $\frac{R_M}{1-\frac{u}{n}} < \frac{\phi}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Also, it holds from (143) that $\frac{R_M}{1-\frac{u}{n}} < 1$, for $u = 1, \ldots, m-l$. Consequently, $\min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right) > \frac{R_M}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Moreover, from (144), and using the fact that $\min\left(\frac{\phi}{1-\frac{m-l}{m}}, 1\right) \leq 1$, it follows that

$$r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} > \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u} \frac{R_M}{1-\frac{u}{n}} \overset{(138)}{=} 1 . \tag{145}$$

(b) $\phi \leq R_M$. In this case, it follows from (143) that $\frac{\phi}{1-\frac{1}{n}} < \cdots < \frac{\phi}{1-\frac{m-l}{n}} < \frac{\phi}{1-\frac{m-l}{m}} = \frac{m\,\phi}{l} < \frac{\phi}{R_M} \leq 1$. Subsequently, (144) yields

$$\begin{aligned} r^{\text{sym,MTTDL}}_{\text{clus,MTTDL}} &\approx \left(\frac{1}{l+1}\right)^{m-l} \frac{(m-1)!}{(l-1)!} \Bigg/ \left(\frac{m\,\phi}{l}\right)^{m-l} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u} \left(\frac{\phi}{1-\frac{u}{n}}\right) \\ &= \left[\frac{l\,n}{(l+1)\,m}\right]^{m-l} \prod_{u=1}^{m-l} \left(\frac{n-u}{m-u}\right)^{m-l-u-1} \\ &\overset{(138)}{=} \left(\frac{l}{m\,R_M}\right)^{m-l} \overset{(139)}{>} 1 . \end{aligned} \tag{146}$$

From the results obtained in the preceding two subcases, we conclude that, when $m-l \geq 4$, MTTDL is maximized by the declustered placement scheme. ■

APPENDIX D
OPTIMAL k FOR EAFDL

Proof of Proposition 11.

Depending on the values of m and l, we consider the following two cases:

Case 1: $m-l = 1$. In this case, (92) yields

$$r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}} \approx \frac{m}{n-1} \cdot \frac{\min\left(\frac{\phi}{1-\frac{1}{m}}, 1\right)}{\min\left(\frac{\phi}{1-\frac{1}{n}}, 1\right)} . \tag{147}$$

It follows from (89) that $\frac{l+1}{l} \leq 2$ and $\frac{m}{n} \leq \frac{1}{2}$, with the equalities holding only when $l = 1$, $m = 2$, and $n = 4$. Consequently,

$$\frac{(l+1)\,m}{l\,n} = \frac{m^2}{(m-1)\,n} \leq 1 , \tag{148}$$

with the equality holding only when $l+1 = m = 2$ and $n = 4$.

We deduce from (148) that

$$\frac{m}{n} \leq \frac{m-1}{m} = 1 - \frac{1}{m} < 1 - \frac{1}{n} . \tag{149}$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > \frac{m}{n}$. In this case, it holds that $\frac{\frac{m}{n}}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{1}{n}}$. Also, it holds from (149) that $\frac{\frac{m}{n}}{1-\frac{1}{n}} < 1$. Consequently, $\min\left(\frac{\phi}{1-\frac{1}{n}}, 1\right) > \frac{\frac{m}{n}}{1-\frac{1}{n}}$. Moreover, from (147), and using the fact that $\min\left(\frac{\phi}{1-\frac{1}{m}}, 1\right) \leq 1$, it follows that

$$r^{\text{declus,EAFDL}}_{\text{clus,EAFDL}} < \frac{m}{n-1} \cdot \frac{1}{\frac{\frac{m}{n}}{1-\frac{1}{n}}} = 1 . \tag{150}$$

(b) $\phi \leq \frac{m}{n}$. In this case, it follows from (149) that $\frac{\phi}{1-\frac{1}{n}} < \frac{\phi}{1-\frac{1}{m}} \leq \frac{\phi}{\frac{m}{n}} \leq 1$. Subsequently, (147) yields

$$r_{\text{clus,EAFDL}}^{\text{declus,EAFDL}} \approx \frac{m}{n-1} \cdot \frac{\frac{\phi}{1-\frac{1}{m}}}{\frac{\phi}{1-\frac{1}{n}}} = \frac{m^2}{(m-1)\,n} \;\leq\; 1\,, \quad (151)$$

with the equality holding only when $l = 1$, $m = 2$, and $n = 4$. In this case, the clustered and declustered placements yield the same reliability.

From the results obtained in the preceding two subcases, we conclude that, when $m - l = 1$, EAFDL is minimized by the declustered placement scheme.

Case 2: $m - l \geq 2$. Let us define

$$R_E \triangleq R_E(l,m,n) = \frac{l+1}{n}\left[\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u}\right]^{\frac{1}{m-l}}. \quad (152)$$

Note that the following inequality holds

$$\left(\frac{R_E}{1-\frac{m-l}{m}}\right)^{m-l} = \left(\frac{m\,R_E}{l}\right)^{m-l} \overset{(152)}{=} \left[\frac{(l+1)\,m}{l\,n}\right]^{m-l}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u} < 1\,, \quad (153)$$

because it holds that $(l+1)\,m/(l\,n) \leq 1$, as shown in Appendix C, and $(m-u)/(n-u) < 1$, for $u = 1, \ldots, m-l-1$, owing to (89).

We deduce from (153) that

$$R_E < \frac{l}{m} = 1 - \frac{m-l}{m} < 1 - \frac{m-l}{n} < \cdots < 1 - \frac{1}{n}\,. \quad (154)$$

Depending on the value of ϕ, the following two subcases are considered:

(a) $\phi > R_E$. In this case, it holds that $\frac{R_E}{1-\frac{u}{n}} < \frac{\phi}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Also, it holds from (154) that $\frac{R_E}{1-\frac{u}{n}} < 1$ for $u = 1, \ldots, m-l$. Consequently, $\min\left(\frac{\phi}{1-\frac{u}{n}}, 1\right) > \frac{R_E}{1-\frac{u}{n}}$, for $u = 1, \ldots, m-l$. Moreover, from (92), and using the fact that $\min\left(\frac{\phi}{1-\frac{m-l}{m}}, 1\right) \leq 1$, it follows that

$$r_{\text{clus,EAFDL}}^{\text{declus,EAFDL}} < (l+1)^{m-l}\,\frac{(l-1)!}{(m-1)!}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Big/ \frac{R_E}{1-\frac{u}{n}} \overset{(152)}{=} 1\,. \quad (155)$$

(b) $\phi \leq R_E$. It follows from (154) that $\frac{\phi}{1-\frac{1}{n}} < \cdots < \frac{\phi}{1-\frac{m-l}{n}} < \frac{\phi}{1-\frac{m-l}{m}} = \frac{m\,\phi}{l} \leq \frac{\phi}{R_E} \leq 1$. Subsequently, (92) yields

$$r_{\text{clus,EAFDL}}^{\text{declus,EAFDL}} \approx (l+1)^{m-l}\,\frac{(l-1)!}{(m-1)!}\left(\frac{m\,\phi}{l}\right)^{m-l}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l+1-u} \Big/ \left(\frac{\phi}{1-\frac{u}{n}}\right) = \left[\frac{(l+1)\,m}{l\,n}\right]^{m-l}\prod_{u=1}^{m-l}\left(\frac{m-u}{n-u}\right)^{m-l-u} \overset{(152)}{=} \left(\frac{m\,R_E}{l}\right)^{m-l} \overset{(153)}{<} 1\,. \quad (156)$$

From the results obtained in the preceding two subcases, we conclude that, when $m - l \geq 2$, EAFDL is minimized by the declustered placement scheme. ■

REFERENCES

[1] I. Iliadis, "Reliability of erasure coded systems under rebuild bandwidth constraints," in Proceedings of the 11th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2018, pp. 1–10.

[2] D. A. Patterson, G. Gibson, and R. H. Katz, "A case for redundant arrays of inexpensive disks (RAID)," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 1988, pp. 109–116.

[3] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Patterson, "RAID: High-performance, reliable secondary storage," ACM Comput. Surv., vol. 26, no. 2, Jun. 1994, pp. 145–185.

[4] M. Malhotra and K. S. Trivedi, "Reliability analysis of redundant arrays of inexpensive disks," J. Parallel Distrib. Comput., vol. 17, Jan. 1993, pp. 146–151.

[5] W. A. Burkhard and J. Menon, "Disk array storage system reliability," in Proceedings of the 23rd International Symposium on Fault-Tolerant Computing, Jun. 1993, pp. 432–441.

[6] K. S. Trivedi, Probabilistic and Statistics with Reliability, Queueing and Computer Science Applications, 2nd ed. New York: Wiley, 2002.

[7] Q. Xin, E. L. Miller, T. J. E. Schwarz, D. D. E. Long, S. A. Brandt, and W. Litwin, "Reliability mechanisms for very large storage systems," in Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST), Apr. 2003, pp. 146–156.

[8] 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.

[9] Q. Lian, W. Chen, and Z. Zhang, "On the impact of replica placement to the reliability of distributed brick storage systems," in Proc. 25th IEEE International Conference on Distributed Computing Systems (ICDCS), Jun. 2005, pp. 187–196.

[10] S. Ramabhadran and J. Pasquale, "Analysis of long-running replicated systems," in Proc. 25th IEEE International Conference on Computer Communications (INFOCOM), Apr. 2006, pp. 1–9.

[11] B. Eckart, X. Chen, X. He, and S. L. Scott, "Failure prediction models for proactive fault tolerance within storage systems," in Proceedings of the 16th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Sep. 2008, pp. 1–8.

[12] A. Thomasian and M. Blaum, "Higher reliability redundant disk arrays: Organization, operation, and coding," ACM Trans. Storage, vol. 5, no. 3, Nov. 2009, pp. 1–59.

[13] K. Rao, J. L. Hafner, and R. A. Golding, "Reliability for networked storage nodes," IEEE Trans. Dependable Secure Comput., vol. 8, no. 3, May 2011, pp. 404–418.

[14] 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.

[15] V. Venkatesan, I. Iliadis, C. Fragouli, and R. Urbanke, "Reliability of clustered vs. declustered replica placement in data storage systems," in Proceedings of the 19th Annual IEEE/ACM International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Jul. 2011, pp. 307–317.

[16] V. Venkatesan, I. Iliadis, and R. Haas, "Reliability of data storage systems under network rebuild bandwidth constraints," in Proceedings of the 20th Annual IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), Aug. 2012, pp. 189–197.

[17] V. Venkatesan and I. Iliadis, "A general reliability model for data storage systems," in Proceedings of the 9th International Conference on Quantitative Evaluation of Systems (QEST), Sep. 2012, pp. 209–219.

[18] J.-F. Pâris, T. J. E. Schwarz, A. Amer, and D. D. E. Long, "Highly reliable two-dimensional RAID arrays for archival storage," in Proceedings of the 31st IEEE International Performance Computing and Communications Conference (IPCCC), Dec. 2012, pp. 324–331.

[19] V. Venkatesan and I. Iliadis, "Effect of codeword placement on the reliability of erasure coded data storage systems," in Proceedings of the 10th International Conference on Quantitative Evaluation of Systems (QEST), Sep. 2013, pp. 241–257.

[20] I. Iliadis and V. Venkatesan, "An efficient method for reliability evaluation of data storage systems," in Proceedings of the 8th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2015, pp. 6–12.

[21] ——, "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.

[22] S. Caron, F. Giroire, D. Mazauric, J. Monteiro, and S. Pérennes, "P2P storage systems: Study of different placement policies," Peer-to-Peer Networking and Applications, Mar. 2013, pp. 1–17.

[23] 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.

[24] ——, "Reliability assessment of erasure coded systems," in Proceedings of the 10th International Conference on Communication Theory, Reliability, and Quality of Service (CTRQ), Apr. 2017, pp. 41–50.

[25] ——, "Reliability evaluation of erasure coded systems," Int'l J. Adv. Telecommun., vol. 10, no. 3&4, Dec. 2017, pp. 118–144.

[26] J. G. Elerath and J. Schindler, "Beyond MTTDL: A closed-form RAID 6 reliability equation," ACM Trans. Storage, vol. 10, no. 2, Mar. 2014, pp. 1–21.

[27] I. Iliadis and V. Venkatesan, "Rebuttal to 'Beyond MTTDL: A closed-form RAID-6 reliability equation'," ACM Trans. Storage, vol. 11, no. 2, Mar. 2015, pp. 1–10.

[28] "Amazon Simple Storage Service." [Online]. Available: http://aws.amazon.com/s3/ [retrieved: November 2017]

[29] D. Borthakur et al., "Apache Hadoop goes realtime at Facebook," in Proceedings of the ACM SIGMOD International Conference on Management of Data, Jun. 2011, pp. 1071–1080.

[30] R. J. Chansler, "Data availability and durability with the Hadoop Distributed File System," ;login: The USENIX Association Newsletter, vol. 37, no. 1, 2013, pp. 16–22.

[31] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop Distributed File System," in Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies (MSST), May 2010, pp. 1–10.

[32] D. Ford et al., "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.

[33] C. Huang et al., "Erasure coding in Windows Azure Storage," in Proceedings of the USENIX Annual Technical Conference (ATC), Jun. 2012, pp. 15–26.

[34] 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.

[35] "IBM Cloud Object Storage." [Online]. Available: www.ibm.com/cloud-computing/products/storage/object-storage/how-it-works/ [retrieved: November 2017]

[36] K. V. Rashmi, N. B. Shah, D. Gu, H. Kuang, D. Borthakur, and K. Ramchandran, "A solution to the network challenges of data recovery in erasure-coded distributed storage systems: A study on the Facebook warehouse cluster," in Proceedings of the 5th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage), Jun. 2013, pp. 1–5.

[37] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network coding for distributed storage," Proc. IEEE, vol. 99, no. 3, Mar. 2011, pp. 476–489.

[38] M. Zhang, S. Han, and P. P. C. Lee, "A simulation analysis of reliability in erasure-coded data centers," in Proceedings of the 36th IEEE Symposium on Reliable Distributed Systems (SRDS), Sep. 2017, pp. 144–153.

[39] J. E. Angus, "On computing MTBF for a k-out-of-n:G repairable system," IEEE Trans. Reliability, vol. 37, no. 3, Aug. 1988, pp. 312–313.

[40] 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.

[41] M. Ovsiannikov et al., "The quantcast file system," in Proc. 39th Int'l Conf. on Very Large Data Bases (VLDB), vol. 6, no. 11. VLDB Endowment, Aug. 2013, pp. 1092–1101.

Adjustment of the QoS Parameters on Routers with Neural Network Implementation

Irina Topalova
Department of Information Technology
University of Telecommunications and Post, Bulgaria
Sofia, Bulgaria
itopalova@abv.bg

Pavlinka Radoyska
College of Energy and Electronics
Technical University - Sofia
Sofia, Bulgaria
pradoiska@abv.bg

***Abstract*—Applying Quality of Service mechanisms to modern communications is essential for the efficiency and for the traffic reliability. The various Quality of Service methods are based on queues management depending on the individual traffic parameters. Choosing Quality of Service parameters on the edge network devices defines the management queue and packet discard/queued parameters on the intermediate devices. The proposed research explores the possibility of automatically adapting to the already selected class based Quality of Service policy of new users added to the backbone of the network. In addition, a method for queue adjustment has been suggested and tested, taking into account the current queue of the added user. A neural network is trained to automatically adapt new end users to the quality of service policy, already set by other end users and accepted by intermediate routers. The obtained results show that the automated adaptation of the Quality of Service parameters to the already set ones is possible for the intermediate routers. A software application, implementing the method in a network segment, is presented. The positive consequences of applying the proposed method are discussed.**

Keywords - traffic congestion, Quality of Service, early detection, queue management, neural network.

I. INTRODUCTION

This publication is based on our research reported at the ICAS 2018 conference [1]. Our research is aimed at creating a mechanism for automatically adjusting Quality of Service (QoS) parameters on routers using Neural Networks (NN). The configuration is based on Differentiated Services Code Point (DSCP) and Weighted Random Early Detection (WRED) queue management. The aim of QoS in communication networks is to guarantee the quality of message delivered by congestion management and congestion avoidance. This is achieved by dividing the traffic in queues and managing the queues individually, based on parameters, configured in any intermediate network device (router or switch). The packets are marked in the endpoint devices, according to the QoS model. Any intermediate device must be configured to create and manage queues, based on this model. Synchronized queue management in all devices is important for quality assurance. The purpose of our work is to find a mechanism by which any new device chooses its configuration parameters for queue management, based on the configuration parameters of the neighboring devices. The various QoS methods are based on queues management depending on individual traffic parameters. The chosen QoS parameters on the edge network devices define the management queue and packet discard/queued parameters on the intermediate devices. The proposed research explores the possibility of automatic adaptation to the already selected class based QoS policy of new users added to the backbone of the network. A NN, defined among many other types of neural networks NNs by Graupe [2] is trained to adapt new end users to the QoS policy, already set by other end users and accepted by the intermediate routers. The WRED method, described in Cisco guide [3], was applied to manage and to define the train and the test NN parameters. Additionally, a queue adjustment in a backbone router is proposed, taking into account the current queue of the added user. The automatic adaptation of additional networking devices to existing infrastructure with already-defined QoS policy would lead to the release of human resources and acceleration of the adaptation of traffic parameters in communication management. Properly tracking and setting the backbone queue in accordance to new added users, would improve the efficiency of the congestion management. The experimental results are presented, discussed and a further continuation of the study is proposed.

The rest of this paper is organized as follows. Section II describes the related to the research works. Section III describes and compares differentiated services and weighted random early detection methods. In Section IV the weighted random early detection with extension of explicit congestion notification is explained. Section V describes the proposed method for parameter adjustment and neural network implementation. Section VI gives the experimental results. Section VII introduces a software application for managing the QoS configuration process with using the NN. The conclusion with discussions close the article.

II. RELATED WORKS

Differentiated Services Code Point in IPv4 and Differentiated Services (DiffServ) in IPv6 are advanced instruments for traffic marking and queue management.

Khater and Hashemi [4] propose to use Differentiated Services Queues at the output port and move the flows between queues to prevent increasing delays of the flows. In another work [5] the authors use TSK fuzzy model to generate Differentiated Services Code Point values dynamically and update them in real time to improve QoS. The authors Sahu and Sar [6] have created an intelligent method to recognize incoming congestion problems earlier. They train a feedforward neural network with parameters equivalent to the total drop, average per packet drop, cumulative per packet drop, maximum packets drop and minimum packets drop, for send and receive features. The final solution is not automatically obtained as a result of the proposed method. It is left to the administrator. The results are not clearly represented and discussed, moreover the authors claim that their developed system missed some points of congestion. Within the model proposed in [7], the transmitted packets/traffic were predicted through a neural network, achieving prediction by alternating the input variables (Bandwidth, Congestion Algorithms, QoS, etc.). In this case, in TCP predictions, where one of the most important factors is related to the limitations of this protocol in both the sender and receiver, congestion improvements or methods for QoS were not considered. The different predictions have validity with respect to the real data, obtaining an average error of 4%. The authors in [8] apply a neural network to predict the actual time needed for transmitting the packet to the destination, depending on the number of hops. As neural network input train parameters, the authors use CWND (Congestion Window) as TCP state variable; Round-Trip Time (RTT) as the length of time it takes for a signal to be sent plus the length of time it takes for an acknowledgement of that signal to be received and the time elapsed from the last loss of a packet. However, this study does not use a method of prioritizing the traffic according to different types of priorities and they do not group traffic into classes according to the priority, given by the end routers/ users.

All mentioned researches do not apply more productive/ efficient methods, such as WRED in conjunction with Class-Based Weighted Fair Queueing (CBWFQ), proposed in Cisco guide [3]. They do not interpret the task we offer - to automatically adapt new end users to the quality of service policy, already set by other end-users and accepted by the intermediate routers.

III. Differentiated Services and Weighted Random Early Detection

Network congestion occurs when the volume of incoming traffic exceeds the bandwidth of the outgoing channel. Congestion avoidance mechanisms are trying to provoke TCP slow-start algorithm (RFC 2001), implemented in end devices. WRED and differentiated services, implemented in routers, become the most effective approach to prevent the congestions.

A. *Active Queue Management congestion avoidance mechanisms*

Congestion avoidance in routers is implemented by Active Queue Management (AQM) congestion avoidance mechanisms. Extra packets coming on the inbound interfaces are queued in buffers. The length of the queue is maintained within defined limits by dropping the packets. One of the first effective AQM mechanism is RED (Random Early Detection), proposed by Floyd and Jacobson [9] in the early 1990s. Two critical thresholds for the queue are defined: minimum queue length (***min***q) and maximum queue length (***max***q) and three queue management phases: no drop, random drop, and full drop, shown in Fig. 1. No drop phase is executed only for queue length from 0 to ***min***q. All packets are buffered. Random drop phase is for queue length from ***min***q to ***max***q. Some packets are dropt. Full drop phase is for queue length above ***max***q. All packets are dropped. The packet drop probability (random drop phase) is calculated based on the average queue length and the MPD (Mark Probability Denominator), Floyd and Jacobson [9]. MPD is the number of dropped packets when the queue size is equal to ***max***q. RED algorithm gives a decision for congestion avoidance problem but has some disadvantages. First, this mechanism does not affect non-TCP protocols. There are risks by insensitive protocols to embezzle the queue. Second, the packets from different TCP sessions are not dropped equally and there is a risk of global synchronization problem. Third, the number of dropped packets sharply jump to 100% when the queue size achieves ***max***q size. Different algorithms for the improvement of active queue management are proposed in [10]. WRED is a kind of class based queue management algorithms. It uses the same parameters as RED, but it has the ability to perform RED on traffic classes individually. Several traffic classes can be defined within a single queue. Each class has a specific level for the ***min***q, ***max***q and MPD. Packets are classified and joined to a specific class. Drop probability for each packet is calculated according to its class parameters. The packets with lowest ***min***q and/or the highest MPD are dropped preferentially. Every class has the same three phases as the RED algorithm. WRED management queue with three classes: AF1, AF2 and AF3 is presented in Fig. 2. AF1 and AF2 have the same ***max***q and MPD parameters. The AF1 ***min***q parameter has a lower value then the AF2 ***min***q parameter. Obviously the most packages are dropped from AF1 class, then from AF2 class and finally from AF3 class. The network traffic is divided in several queues to improve fairness in packet dropping. Each queue is managed by the RED, WRED or a similar algorithm. Weighted Fair Queue (WFQ), discussed by

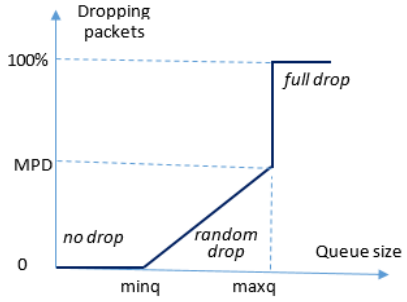


Figure 1. Random Early Detection phases.

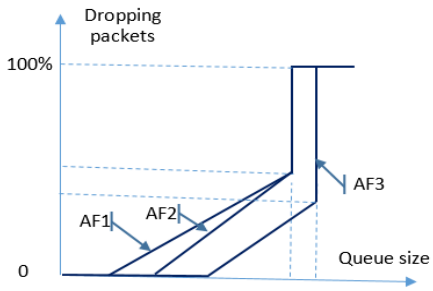


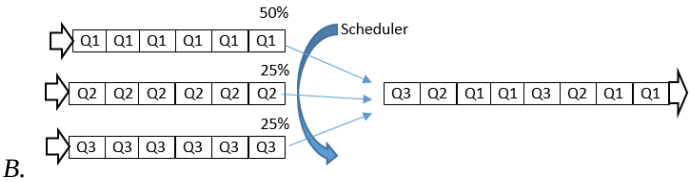
Figure 2. WRED phases.

Vukadinović and Trajković [11] is a data packet scheduling algorithm. All the queues share outbound bandwidth equally or by predefined ratios. The queues are visited one by one in the cycle period. Every queue sends the amount of packets, according to its share part of the outgoing capacity. The simple WFQ example is presented in Fig. 3. Q1 gets 50% of the outgoing capacity, Q2 – 25% and Q3 – 25%. The Scheduler visits Q1 and passes over 2 packets to the output. After that it visits Q2 and passes over 1 packet to the output; visits Q3 and passes over 1 packet to the output, and the cycle is rotated again.

B. Differentiated Services Quality of Service model

There are three main models for providing QoS in a network: Best Effort; Integrated Services (IntServ); Differentiated Services (DiffServ). DiffServ is called soft QoS model and uses WFQ and WRED algorithms. This model is based on user defined service classes and Per-Hop-Behavior (PHB). The flows are aggregated in traffic classes. The network service policies are defined for each class on any single node. Priorities are marked in each packet using DSCP for traffic classification.

The fields Type of Service (ToS) in IPv4 header (RFC 791) and Traffic Class (TC) in IPv6 header (RFC 2460) are predefined as Differentiated Services Field (DS Field) in RFC 2474. The first six bits of the DS field are used as a code point (DSCP) to select the PHB packet experiences at each node. DSCP values are described in RFC 2475. They determine the PHB of a packet. Four conventional PHBs are available: two border marks; Class-Selector PHB and Assured Forwarding (AF). DSCP = 000000 marks best effort behavior. All packets with this mark will be dropped when congestion occurs. This is the default PHB. DSCP = 101110 (46 in decimal) marks Expedited Forwarding (EF). EF PHB provides a virtual leased line and is used for critical traffic class as voice traffic. EF PHB provides low-loss, low-latency, low-jitter and assured bandwidth service. DSCP values of "xxx000" ("xxx" are the class selector bits) mark Class-Selector PHB and are used to assure backward compatibility with IP ToS model. DSCP values of "xxxyy0" mark Assured Forwarding (AF) PHB. "xxx" is for user defined AF class and "yy" is for drop precedence of a packet. "01" denotes low drop precedence, "10" – middle and "11" - high drop precedence. AF PHB classes are the subject of this paper.



B.

Figure 3. Weighted Fair Queue

C. DiffServ model configuration steps

1) Network traffic classification

Performs predominantly on the edge for QoS domain router - Cisco Guide [12]. The traffic type is defined by Access Control Lists (ACL) and joined to the specific AF class. Every class is associated with specific DSCP value. Inbound packets are marked with corresponding DSCP value on the edge routers of QoS domain and it is not recommended to change it in the intermediate routers.

2) Queue building

One or more AF classes can be aggregated in one queue, based on PHB parameters. The Queues can be three types: Strict priority queue (LLQ – Low latency queue); Class based queues (managed by WRED algorithm) end best effort queue.

3) Defining queue parameters

The WRED parameters are defined for every queue. For the Strict priority queue, the defined outbound bandwidth is guaranteed. The rest of outbound bandwidth is distributed between all other queues. For every class based queue, the following parameters are defined:

a) The portion of the bandwidth in percentage;

b) For each AF class (DSCP value) in the queue: min-threshold; max-threshold; MD (Mark-denominator).

Successful congestion avoidance depends on the proper execution of the above three steps. Especially on proper queue management definitions, described in 3) b).

IV. WRED FUNCTIONALITY EXTENDED WITH ECN

WRED drops packets, based on the average queue length exceeding a specific threshold value, to indicate congestion. Explicit Congestion Notification (ECN) (RFC 3168) provide end-to-end lossless communication between two endpoints over an IP routed network as given in [13] [14]. The ECN is an extension to WRED in that ECN marks packets instead of dropping them when the average queue length exceeds the *min-threshold* value. If there is a risk of congestion in a device (*min-threshold* < *queue* < *max-threshold*), instead of dropping the packages, they are marked and forwarded. When a marked packet arrives to the recipient, it sends a confirmation to the sender informing it of the available traffic congestion. As a result, the sender reduces his TCP window and the congestion decreases. This increases the bandwidth of the network because no unnecessary packets are ejected. This mechanism can be built into both - intermediate and end devices. There are also adaptations of ECN to UDP protocol explained in [15] - [17]. Two protocols which support ECN width UDP are defined: Datagram Congestion Control

Protocol (DCCP) (RFC5681) and Stream Control Transmission Protocol (SCTP) (RFC4960). The receiver sends a small special message to the sender, recommending to slow down the sending speed, because of congestion on the route. The next effective congestion avoidance is eXplicit Congestion Control Protocol (XCP), given in [18]. It works on the end and intermediate network devices (switches and routers) width TCP and UDP traffic. In addition, they use end-to-end bandwidth evaluation, to get high congestion estimation. Some of the open problems in Internet congestion control. Are discussed in RFC 6077.

ECN uses two bits - the ECN-capable Transport (ECT) bit and the CE (Congestion Experienced), which are the two least significant bits in the ToS field in the IP header. The four combinations of these bits have the following meaning: “0 0” indicates that a packet is not using ECN, “0 1” and “1 0” are set by the data sender to indicate that the endpoints of the transport protocol are ECN-capable and “1 1” indicates congestion to the endpoints i.e. packets reached a *max-threshold* of a router will be dropped. When ECN is enabled, the packets are treated as given in by Cisco Systems, Congestion Avoidance Configuration Guide, [19] and summarized by us, as follows:

1) If the number of packets in the queue is below the *min-threshold*, they are forwarded, whether or not ECN is enabled, and this is identical to the treatment a packet receives when WRED is only used on the network.

2) If the number of packets in the queue is between the *min-threshold* and the *max-threshold*, one of the following four cases is possible:

a) ECN field is “0 1” or “1 0” on the packet indicates that the endpoints are ECN-capable and the WRED algorithm determines that the packet should have been dropped based on the drop probability. In this case, the ECT and CE bits for the packet are changed to 1 and the packet is transmitted. So that, *the packet gets marked instead of dropped.*

b) *If the ECN field on the packet indicates that neither endpoint is ECN-capable* (that is, the ECT bit is set to 0 and the CE bit is set to 0), the packet may be dropped based on the WRED drop probability. This is the identical packet treatment when WRED is applied without ECN enabled.

c) If the ECN field on the packet indicates that the network is experiencing congestion (that is, both the ECT bit and the CE bit are set to 1), the packet is transmitted. No further marking is required.

3) If the number of packets in the queue is above the *max-threshold*, packets are dropped based on the drop probability. Such a treatment of a package is the same as when the router works only with WRED, without the ECN being set. The properly selected value of *min-threshold* is essential for the proper functioning of the network and congestion avoidance mechanism.

V. Proposed Method for WRED Parameter Adjustment Extended with ECN

In this study, we apply the WRED method for QoS in a network having end routers, a central/backbone router and an ad-hoc “New” router. The first task is to force the new added router to comply with the QoS requirements, which were pre-set in the central router. For this we propose a NN, intended to work in the central router, aiming to adjust the parameters of the “New” to the existing ones. The second task is to propose a method for appropriately determining the average queue and the *min-threshold* in the central router, when applying ECN, taking into account the current average queue of the added “New” router.

A. Investigated topology

We apply the WRED method for QoS, because it gives relation between AF classes and the most important queue traffic parameters. The topology shown in Fig. 4 is considered. It consists of two edge routers (Remotes 1 and 2), an intermediate router(Central) and an edge router "New", which is added later after the QoS parameters are set in the edge routers. WRED is implemented at the central/core routers of a network. Edge routers assign IP precedence to packets as the packets enter the network. With WRED, core routers then use these precedencies to determine how to treat different types of traffic *[18]*. The idea is to train a neural network (NN), implemented in the Central router with WRED parameters: AF class, min-threshold; max-threshold and MD, according to the IOS command random-detect. When an ad-hoc edge router "New" is added with its configured WRED (DSCP) requirements of its network, the already trained NN will approximate/adjust its MD to that already learned by the NN. This adjustment will be performed automatically without the need for any operator intervention. The new added router will have to comply with the pre-set QoS requirements.

B. Neural Network strategy

To conduct the experiment, we chose a neural network of Multi-Layer-Perceptron (MLP) type, training it with a BP (Backpropagation) algorithm. It was trained with the DSCP

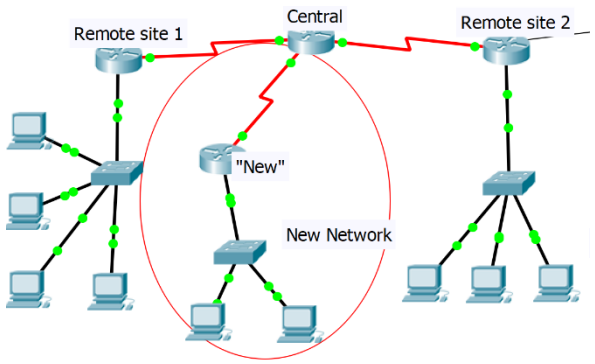


Figure 4. Investigated topology with edge routers (Remote site 1 and 2), intermediate (Central) router and the ‘New’ added router

values, corresponding to AF Classes 1,2,3 and 4, where Class 1 represents the 'worst queue', for low priority traffic and Class 4 – the 'best queue', for high priority traffic as first parameter. The second and third parameters in the input training set are *min-threshold* and *max-threshold*, defined by the command random-detect in the Central router. If the *min-threshold* is reached, Central router randomly drops some packets with the specified IP precedence. If the *max-threshold* is reached, Central router drops all packets with the specified IP precedence. The MLP has one output neuron and it represents the desired *MD*, where *MD* represents the fraction of packets dropped when the average queue depth is at the *max-threshold*. It means that one out of every *MD* packets will be dropped. Table I represents the correspondence between AF classes, DSCP values and drop precedence. After the NN was trained, a combination of different DSCP values with proposed bandwidth percent for each AF class was provided at its input layer, in order to simulate these parameters, send by the 'New' router. According to the "New" requirements/parameters, the Central router generates new *min-threshold* and *max-threshold* and forwards the new information to the NN inputs.

TABLE I. AF CLASSES AND CORRESPONDING DSCP VALUES

Assured Forwarding	Low Drop (DSCP)	Medium Drop (DSCP)	High Drop (DSCP)
Class 4	AF41 (34)	AF42 (36)	AF43 (38)
Class 3	AF31 (26)	AF32 (28)	AF33 (30)
Class 2	AF21 (18)	AF22 (20)	AF23 (22)
Class 1	AF11 (10)	AF12 (12)	AF13 (14)

As a result, the trained NN gives an output with approximated *MD* value, which is near the value defined initially by the Central. In this way, the 'New' router will be forced to "comply" with the chosen QoS policy.

C. Queue adjustment

The average queue size is based on the previous average and current size of the queue, as given in equation (1) [19]:

$$Q_{avr} = \left(old_{avr} * (1 - \frac{1}{2^n})\right) + (curr\_queue\_size * \frac{1}{2^n}), \quad (1)$$

where Q_{avr} is the calculated value of the average queue size, old_{avr} is the previous value of the queue, *curr_queue_size* is the current queue and *n* is the exponential weight factor, a user-configurable value. The analysis of this equation shows that for high values of *n* the previous average queue size becomes more important. At the same time for low values of *n* the average queue size Q_{avr} will closely track the current queue size.

In our case we propose a change in the given equation (1) in order to accommodate C:Q_{avr} of Central router, taking into account also the current queue size $N: curr\_queue\_size$ of

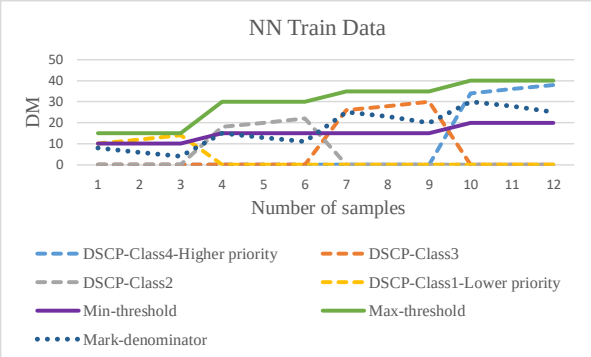


Figure 5. NN train data with initial QoS parameters. The ordinate represents the number of packets in the queue and DM

the New router. We choose the critical moment when the previous value of old_{avr} reaches the *min-threshold* in the Central router, i.e.

$$C: Q_{avr} = \left(C: min\_threshold * (1 - \frac{1}{2^n})\right) + ((C: curr\_queue\_size + N: curr\_queue\_size)/2)) * \frac{1}{2^n}$$

We denote here the parameter ***New Average Queue(NAQ)*** as:

$$\boldsymbol{NAQ} = (C: curr\_queue\_size + N: curr\_queue\_size)/2. \quad (2)$$

As the right choice of *n* is the exponential weight factor and is user-configurable value, we will run the experiment with different *n* values to determine the better option.

VI. EXPERIMENTAL RESULTS

The initially selected MLP network structure is 6-4-1 and is trained to MSE (Mean-Square-Error) = 0.1. The train data are given in Fig. 5. They have 12 input samples as combinations between DSCPs, *min-threshold* and *max-threshold*, defined in Remote 1 and 2. After conducting the test phase with the 'New' data, the obtained *MD* approximation is shown in Fig. 6. The approximation error E_{APPROX} is calculated according to (3), where MD_{RSi} is the initial real system value for the Central router, for i-th input

$$E_{APPROX} = \sqrt{\sum_{i=1}^{n} \frac{(MD_{RS_i} - MD_{NN_i})^2}{n}} \quad (3)$$

combination, MD_{NNi} is the NN response, and *n* is the number of input combinations. The obtained results using this NN topology are given in Fig. 7. In this case, E_{APROX} is 2.56. Obviously, it is necessary to improve the MLP parameters by training a network with an improved structure of 6-6-4-1 and with more iterations, aiming to reach a smaller MSE. In this case, we apply MSE of 0.01. Better obtained results are given in Fig. 8. In this case, E_{APROX} is 0.91. Thus, based on the training of the optimized neural network with the defined AF classes and their initial matching random-detection parameters, we obtain a relatively good MD approximation. Further work is foreseen to test the NN with more combinations of input parameters. For testing the WRED

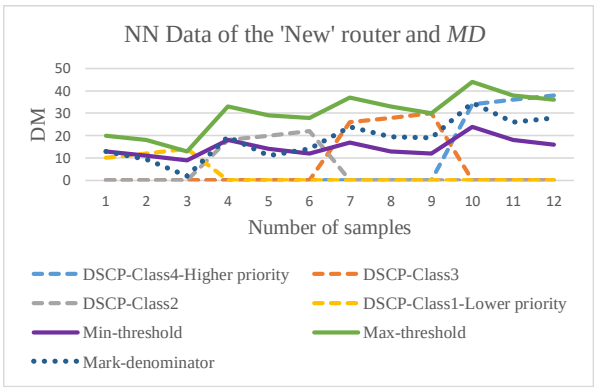


Figure 6. NN 'New' test data with *MD* approximation. The ordinate represents the number of packets in the queue and DM

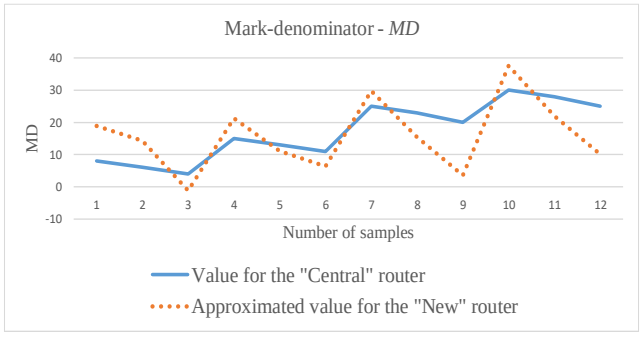


Figure 7. *MD* approximation with MLP – 6-4-1. The ordinate represents the number of packets in the queue and DM

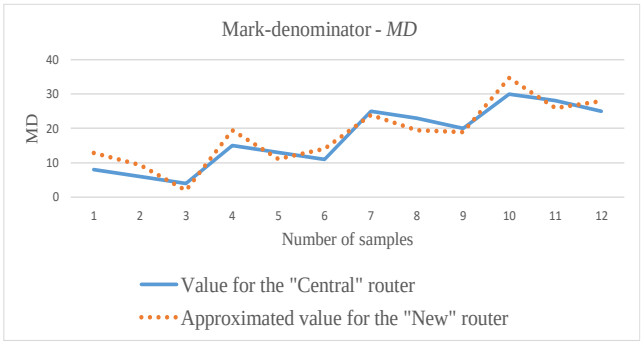


Figure 8. *MD* approximation with MLP – 6-6-4-1 The ordinate represents. the number of packets in the queue and DM

functionality, extended with ECN, corresponding to cases 2)/ a) and 2)/ c), we choose the critical moment when the previous value of old_{avr} reaches the *min-threshold* in the Central router. The goal is to determine the value of C: Q_{avr}, taking into account also the current queue size *N: curr_queue_size* of the New router. We tested how *C:* Q_{var} tracks *NAQ*, depending on its different peak changes, according to (2), and how the exponential weight factor *n* influences the adaptation. Fig. 9, 10, 11 and 12 show the adaptation of *C:* Q_{var} when n=4,3,2,1 correspondingly. The obtained results show that a large factor of n=3 represented in Fig. 10, smooths out the peaks and lowers the queue length. The average queue size will not

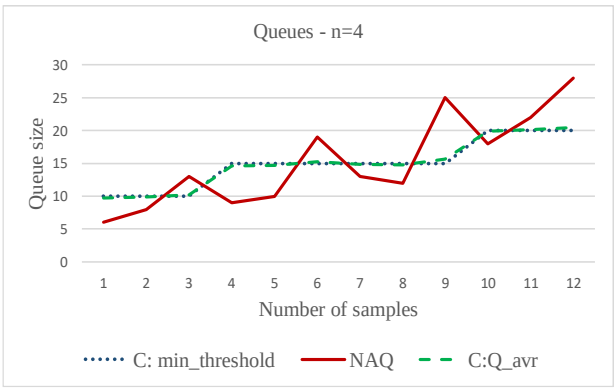


Figure 9. Adaptation of $C: Q_{avr}$ with n=4. The ordinate represents the number of packets in the queues and DM

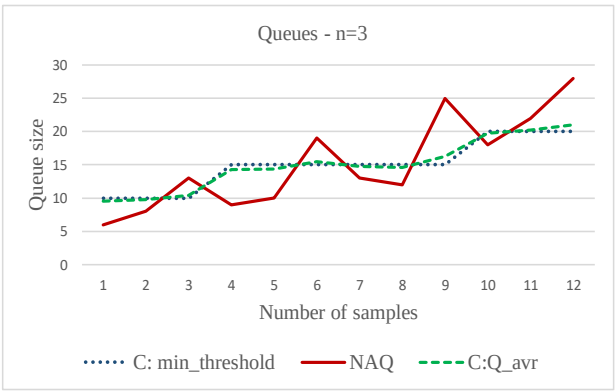


Figure 10. Adaptation of $C: Q_{avr}$ with n=3. The ordinate represents the number of packets in the queues and DM

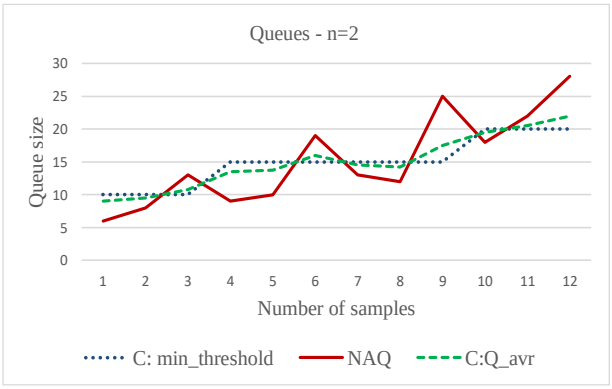


Figure 11. Adaptation of $C: Q_{avr}$ with n=2. The ordinate represents the number of packets in the queues and DM

probably change very quickly, avoiding dramatic fluctuations in size. The WRED process will be slow to start dropping packets and the slow-changing average C: Q_{var} will accommodate temporary peaks in traffic. But if the value of n gets too high (n=4, Fig. 9), WRED will not react to congestion. Packets will be transmitted or dropped as if WRED does not work. For low values of n (n=2, Fig. 11), the C: Qvar tracks closely the current queue size. The resulting average value may fluctuate adequately with the changes in

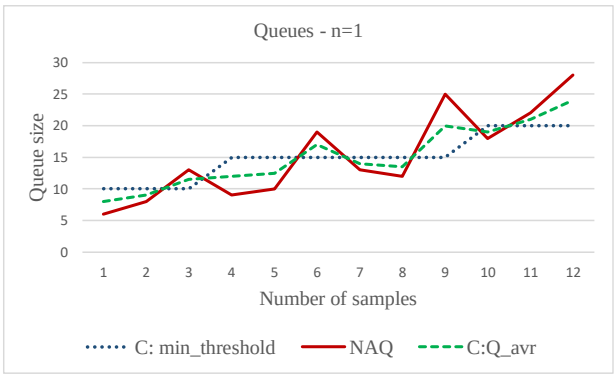


Figure 12. Adaptation of $C: Q_{avr}$ with $n=1$. The ordinate represents the number of packets in the queues

the traffic levels of both Central and New routers (i.e. of NAQ). Once the resulting queue falls below the minimum threshold, the process will stop dropping packets. If the value of n gets too low (n=1, Fig. 12), WRED will overreact to temporary resulting traffic bursts and will drop traffic unnecessarily. Thus, the proposition of n = 2 seems to be the most appropriate in terms of queue efficiency.

VII. REALTIME REMOTE ROUTERS RECONFIGURATION

The purpose of our research is to get better QoS management by synchronizing the queue management parameters on the routers in one network segment without the manual reconfiguration of any new router. Moreover, we try to synchronize the queue management parameters on all routers in network segment after the reconfiguration of only the Central router.

A. Processes and management

A data-flow diagram is shown on Fig. 13. The NN block and The Manager are software blocks that work on an external machine (PC or a laptop). The Manager is responsible for the process navigation. Each router can perform the role of a master (Central) or a subordinate. Router1, Router2 and New on Fig. 13 are subordinate routers.

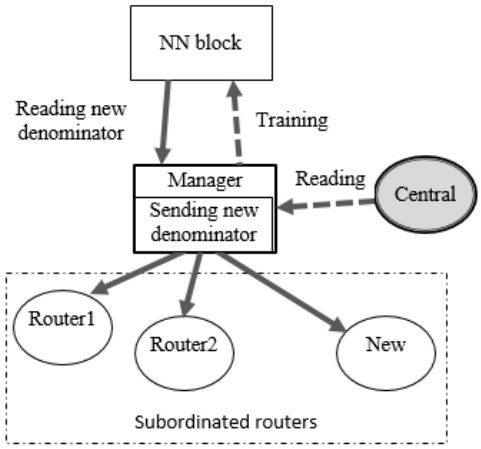


Figure 13. Real-time configuration process

There are two types of processes: training and reconfiguring. The training process includes: reading QoS parameters from the Central router, preparing and sending the training matrix to the NN. The reconfiguring includes: reading the QoS configuration from the subordinate router, preparing and sending a query to the NN and, based on the NN response, prepares the synchronized configuration parameters and sends them to the subordinate router.

There are two possible situations: (1) inclusion of a new router; (2) reconfiguring. In the first situation, the NN is already trained and all routers' configurations are synchronized. A new router with autonomous QoS configuration is included in the network segment. The Manager makes connection to the new router, extracts proper denominator from the NN and reconfigures the new router. The new router starts work in synchronization with all routers in the segment.

In the second situation, all routers' configurations are synchronized but the QoS configuration on the Central router is changed. The Manager makes connection to the Central router, extracts the new queue management parameters and trains the NN. Then it makes connections to all other routers consequently: reads their current QoS parameters, sets them to the NN, gets the new proper denominator and reconfigures the routers. All routers QoS configurations are synchronized again.

The communication between the Manager and the NN block performs in off-line mode being based on computer operating system mechanisms. The communication between the Manager and the routers is accomplished via SSH protocol. Therefore, any router in the management network segment must be registered in the Manager.

B. Manager block implementation

This Manager is written as multithreading Windows application by C# programming language. As hardware devices are used Cisco routers, platforms 2800/2900 with IOS 15.0. The Manager interface has two tabs: Registration and Processes given in Fig. 14 and Fig. 15 respectively. The Central router and the subordinated routers are separated in different blocks for their different roles. The IP address, username and password for SSH connection are saved for each registered router, shown in Fig. 16. The IP address is used as a router identificator.

The training process starts after the button "Train (off-line)" is selected. The result of the training process is displayed in the textbox on the right as shown in Fig. 15. There are two buttons for reconfiguration, according to the two situations mentioned above. Only the selected router is reconfigured after the selection of the button "Reconfigure". All routers, included in the list "Subordinated routers" are reconfigured after the selection of the button "Reconfigure All". The result of this process is displayed in the textbox "Reconfiguration Results" as shown in Fig. 17. The

administrator must troubleshoot the problem in case of appearance of router reconfiguration problems.

All communications in the dataflow diagram, shown in Fig. 13, are of a machine-machine type. The training and the reconfiguration are made automatically but all processes must be started by a person. This approach is appropriate for the first situation, described above – inclusion of a new router. The router registration has to be made manually and the manual start of the reconfiguration should not lead to a significant processing delay. The second situation would be more flexible if the Central router sends a signal to the Manager for the configuration change automatically, thus forcing the training and reconfiguration processes. Solving this problem is a matter of our future research. We need to find a mechanism to alert the Manager about the changes of the Central router configuration. The Manager also must work as a server to listen permanently to that signal.

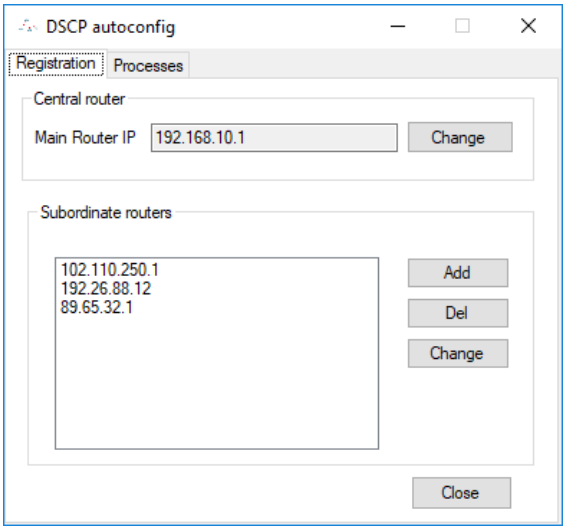


Figure 14. Manager software – Registration tab

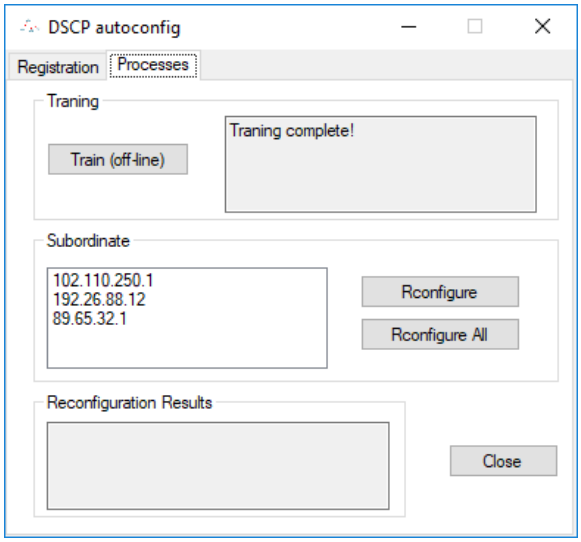


Figure 15. Manager software – Processes tab

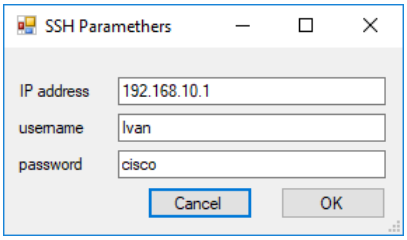


Figure 16. Managing the parameters for SSH connection

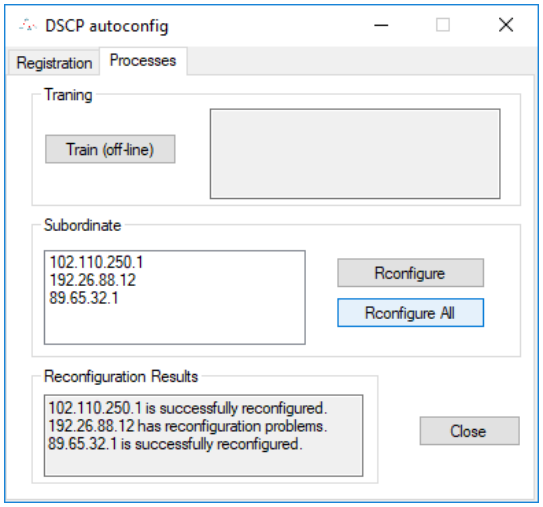


Figure 17. Manager software – Registration tab width reconfiguration results

VIII. CONCLUSION

In this research, a MLP neural network was trained, aiming to automatically adapt new end users to the quality of service policy, already set by other end-users and accepted by the intermediate routers. The WRED method was applied to manage and to define the train and test NN parameters. The proposed method shows good MD approximation results for the tested input set. The main benefit of the automatic adaptation of additional networking devices to existing infrastructure with an already-defined QoS policy would lead to the release of human qualified resources, needed for manual QoS parameter pre-settings. It also would accelerate the traffic parameters adaptation in communication management and in real-time communication. The proposed accommodation of C:Q_{avr} in the Central router, taking into account also the current queue size N: $curr\_queue\_size$ of the New router, choosing the critical moment when the previous value of old_{avr} reaches the *min-threshold* in the Central router, shows good tracking especially when n=2. If the value of n gets too low (n=1, Fig. 12), WRED will overreact to temporary resulting traffic bursts and will drop traffic unnecessarily. Thus, the proposition of n = 2 seems to be the most appropriate in terms of queue efficiency.

A software application was developed to verify the proposed method. It is installed on the external computer system and works as a manager for all processes: reading the initial configuration, preparing the training matrix, starting

the NN training, getting the new proper denominator from already trained NN, reconfiguring the subordinated routers. The verification indicates that the method is applicable.

As further work, the input training and test sets may be increased to generalize the method. The idea is to train the NN with the same standard AF classes but with much more possible/ reasonable combinations of min-max thresholds, together with a proper proposal for the required link bandwidth at the outputs of the NN. The investigated topology given in Fig. 4, may be tested with more Remote routers and many "New" routers, to test the behavior of the Central router. In this case, different NNs could be trained with QoS parameters defined in the different Remotes, and the NN outputs may be combined in input train data for a generalized neural network, to give the final MD proposal. Also, software modules will be developed to integrate the neural network into a module of the Central router operating system, for direct data exchange between the routers. Aiming to achieve/solve this task, we envisage the use of Python programming language, suitable for implementation in networking operating systems.

REFERENCES

[1] I. Topalova, P. Radoyska, "Control of Traffic Congestion with Weighted Random Early Detection and Neural Network Implementation", ICAS 2018, The Fourteenth International Conference on Autonomic and Autonomous Systems, pp. 8-12, Nice, France, 20-24 May 2018

[2] D. Graupe, 'Deep Learning Neural Networks: Design and Case Studies', World Scientific Publishing Co Inc. pp. 57–110, ISBN 978-981-314-647-1, July, 2016.

[3] Cisco IOS Quality of Service Solutions Configuration Guide, Release 12.2, Chapter: Congestion Avoidance Overview https://www.cisco.com/c/en/us/td/docs/ios/12_2/qos/configuration/guide/fqos_c/qcfconav.html#wpxref11086, last accessed 16.11.2018.

[4] A. Khater and M. R. Hashemi, "Dynamic Flow Management Based on DiffServ in SDN Networks," Electrical Engineering (ICEE), Iranian Conference on, Mashhad, 2018, pp. 1505-1510.
doi: 10.1109/ICEE.2018.8472638

[5] J. Li, L. Yang, X. Fu, F. Chao and Y. Qu, "Dynamic QoS solution for enterprise networks using TSK fuzzy interpolation," 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, 2017, pp. 1-6. doi: 10.1109/FUZZ-IEEE.2017.8015711

[6] Y. Sahu and S. K. Sar, 'Congestion analysis in wireless network using predictive techniques', Research Journal of Computer and Information Technology Sciences, ISSN 2320 – 6527 vol. 5(7), pp. 1-4, September, 2017.

[7] A. F. Luque Calderón, E. J. Vela Porras and O. J. Salcedo Parra, 'Predicting Traffic through Artificial Neural Networks', Contemporary Engineering Sciences, vol. 10, no. 24, pp. 1195 - 1209 HIKARI Ltd, www.m-hikari.com https://doi.org/10.12988/ces. 2017.710146, 2017.

[8] S. S. Kumar, K. Dhaneshwar, K. Garima, G. Neha and S. Ayush, 'Congestion Control in Wired Network for Heterogeneous resources using Neural Network', International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013, ISSN: 2277 128X, pp.533-537, 2013.

[9] S. Floyd and V. Jacobson, 'Random Early Detection Gateways for Congestion Avoidance', IEEE/ACM Transactions on Networking, Networking, vol. 1 No. 4, pp. 397-413, August, 1993,Available:http://www.icir.org/floyd/papers/early.twocolumn.pdf: accessed August, 2018.

[10] G. Abbas, Z. Halim and Z. H. Abbas, 'Fairness-Driven Queue Management: A Survey and Taxonomy', IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 324-367, First quarter 2016. doi: 10.1109/COMST.2015.2463121, 2016.

[11] V. Vukadinović and L. Trajković, 'RED with Dynamic Thresholds for improved fairness', Proceedings of the 2004 ACM symposium on Applied computing (SAC '04). ACM, New York, NY, USA, 371-372. DOI: https://doi.org/10.1145/967900.967980, 2004.

[12] QoS: DiffServ for Quality of Service Overview Configuration Guide, Cisco IOS Release 15M&T, January 16, 2013 Available:https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/qos_dfsrv/configuration/15-mt/qos-dfsrv-15-mt-book.html: accessed August, 2018.

[13] B. Trammell, M. Kühlewind, D. Boppart, I. Learmonth, G. Fairhurst, and R. Scheffenegger, "Enabling Internet-Wide Deployment of Explicit Congestion Notification", Passive and Active Measurement. PAM 2015, pp 193-205, March, 2015. DOI: https://doi.org/10.1007/978-3-319-15509-8_15

[14] M. Kühlewind, S. Neuner and B. Trammell, "On the state of ECN and TCP options in the internet", Proceedings of the Passive and Active Measurement, 2013, Hong Kong SAR, China, 2013. DOI: https://doi.org/10.1007/978-3-642-36516-4_14

[15] S. McQuistin and C. Perkins, "Is Explicit Congestion Notification usable with UDP?", IMC '15 Proceedings of the 2015 Internet Measurement Conference, Pages 63-69, Tokyo, Japan — October 28 - 30, 2015 doi: 10.1145/2815675.2815716

[16] E. Stergiou, D. Liarokapis, C. Angelis and F. Vartziotis, "Vigorous Distance Learning Applications Using the Stream Control Transmission Protocol", Science Journal of Education. Vol. 5, No. 6, pp. 262-267, 2017.

[17] S. Saini and A. Fehnker, "Evaluating the Stream Control Transmission Protocol Using Uppaal", EPTCS 244, 2017, pp. 1-13, 2017. DOI: 10.4204/EPTCS.244.1

[18] J. Wang, J. Chen, S. Zhang and W. Wang, "An Explicit Congestion Control Protocol Based on Bandwidth Estimation", IEEE Global Telecommunications Conference - GLOBECOM 2011, Kathmandu, 2011, pp. 1-5. .,doi:10.1109/GLOCOM.2011.6134086, 2011.

[19] QoS: Congestion Avoidance Configuration Guide, Cisco IOS XE Release 3S, Cisco Systems, Inc. 170 West Tasman Drive San Jose, CA 95134-1706 USA

Tracking of Vehicles by Almost Everyone

Markus Ullmann* † Gerd Nolden,* and Timo Hoss*
* Federal Office for Information Security
D-53133 Bonn, Germany
Email: {markus.ullmann, gerd.nolden, timo.hoss}@bsi.bund.de
† University of Applied Sciences Bonn-Rhine-Sieg
Institute for Security Research
D-53757 Sankt Augustin, Germany
Email: markus.ullmann@h-brs.de

***Abstract*—Increasingly, vehicles will be equipped with information and communication technologies, e.g., wireless communication technologies like IEEE 802.11x, Bluetooth, mobile communication, etc. These communication technologies enable identification and tracking based on identifiers used in communication protocols. Today, the Vehicle Identification Number, and the license plate are regarded as vehicle identifiers. With new communication technologies used in modern vehicles, Secondary Vehicle Identifiers are coming up. This paper analyzes the identification of vehicles based on wireless communication interfaces and presents results of real measurements of vehicular Bluetooth and Wi-Fi interfaces. Moreover, countermeasures are introduced, which reduce the risk of being trackable.**

Keywords–Vehicle Identification; Vehicle Identifier; Wireless Vehicle Interfaces; Privacy; Vehicle Tracking

I. Introduction

Information technology in vehicles has significantly changed during the last 10 years. This is shown by the increasing availability of components for driving assistance: lane keeping support, traffic jam assist, automatic parking assistant, remote parking assistant and so on. This development is a prestage of automatic driving, which is one of the main challenges in automotive engineering at the moment. Besides driving assistance, modern vehicles are equipped with wireless interfaces, e.g., Bluetooth to connect devices (smart phones, tablets, etc.) to the multimedia component (head-unit) of the vehicle. In addition, head-units are more and more capable of establishing a Wi-Fi hot spot to support Internet access for vehicle passengers. Furthermore, the vehicle-2-vehicle communication technology (V2V) based on IEEE 802.11p technology will be deployed in the near future. V2V is one feature of Intelligent Transport Systems (ITS).

Today, only the Vehicle Identification Number (VIN), and the license plate are regarded and used as official vehicle identifiers. This paper analyses vehicle identification capabilities of wireless communication interfaces, called Secondary Vehicle Identifiers, which can be used for vehicle identification and tracking. This issue was first published in [1]. Next, results of further measurements of vehicular Bluetooth interfaces and vehicular Wi-Fi hotspots are presented. The communication interfaces are built into the vehicle to support communication services for passengers. We show, however that these services are also available outside the vehicle and can be misused for unauthorized identification and tracking. We only use cheap measurement equipment, e.g., external Bluetooth USB-Sticks (they cost only a few €) and partially open source tools (software components of the Kali Linux distribution for penetration testing), which are publicly available. The smart phone measurement apps applied can be used by everyone with every modern Android compatible device for the identification of vehicles based on Bluetooth. The aim of this paper is to highlight the issue of identification and tracking of vehicles based on Secondary Vehicle Identifiers. Therefore, we have only investigated selected vehicles instead of performing a study with lots of vehicles. Most of the measurements were already performed in November 2016. The Bluetooth tests presented in Section VI-B were conducted in August 2018.

We primarily investigated simple measurements of existing static Secondary Vehicle Identifier, e.g., static MAC IDs. We know that there exist further device identification capabilities as shown in [2] for Wi-Fi components, which we do not study, here. In comparison, we only propose simple countermeasures which avoid an easy tracking of vehicles.

The subsequent sections of this paper are organized as follows: Section II is a description of related work. Subsequently, identifiers for ITS vehicle stations are presented in Section III. Section IV describes wireless technologies implemented in modern vehicles and analyzes identification capabilities. The aim of the tests performed, test equipment used and test vehicles investigated are presented in Section V. Results of real measurements of Bluetooth and Wi-Fi identifier are given in Section VI. In Section VII, the problem of vehicular tracking is addressed. Section VIII depicts only simple countermeasures to avoid an easy tracking of vehicles. Finally, we summarize our results, and mention open research questions.

II. Related Work

A classification of vehicle identifiers which is also applied in this paper is given in [3]. Hwajeng et al. suggested a vehicle identification and tracking system based on optical vehicle plate number recognition [4]. Tracking of devices based on Bluetooth interfaces is already discussed for a lot of applications, e.g., indoor localization [5] or wireless indoor tracking [6]. In [7], an analysis in Jacksonville, Florida, to capture vehicle traffic streams is described. To this end, a set of Bluetooth receivers were installed at the roadside on specific streets to capture the Bluetooth MAC ID (BD_ADDR) of vehicles passing. A quite similar application is still performed in Bonn to analyse and detect mobility pattern of vehicles based on a network of stationary road side Bluetooth sensors [8].

Besides Bluetooth, IEEE 802.11 compliant devices were suggested for real-time location tracking in indoor and outdoor environments [9].

Since November 1st, 2014, vehicles and motorhomes have to be equipped with a Tire Pressure Monitoring System (TPMS) within Europe. These can be subdivided into direct and indirect TPMS. Direct TPMS means that specific physical sensors measure the air pressure of the tires. These sensors communicate wirelessly with the vehicle and transmit an identifier of 28 to 32 bit length. There are different wireless technologies available for 125 kHz, 315 kHz, and 433 MHz. A detection range of up to 40 m for direct TPMS is mentioned in [10].

Apart from the identification of vehicles based on static identifiers used in communication protocols different feature based identification methods are proposed. One approach is the identification of vehicles based on noise features (individual noise spectrum) [11].

Further identification techniques allow wireless devices to be identified by unique characteristics of their analog (radio) circuitry; this type of identification is also referred to as physical-layer device identification. It is possible due to hardware imperfections in the analog circuitry of transmitters introduced during the manufacturing process. A good overview concerning the physical fingerprinting of different wireless communication technologies is given in [12]. The discussion of device tracking based on static identifier of wireless communication interfaces started 15 years ago [13].

In [14], the privacy principles of Bluetooth low energy (BLE) are described and analyzed. It is shown that the privacy mechanisms in BLE are only applicable in connection mode but not during advertising. Moreover, privacy enhancements for advertising are proposed. BLE is widely applied for the connection of fitness trackers to smart phones. Though privacy is an important issue [15] shows that most of the analyzed devices do not implement the privacy mechanisms of the standard or, if they do, implement them in a wrong manner.

Mathy Vanhoef [2] et al. highlight the general difficulty of implementing anti-tracking solution for wireless devices. In particular, they analyzed proprietary Wi-Fi MAC randomization algorithms implemented in iOS (starting from iOS 8), Android (starting from Android 6.0), Linux (starting from Kernel 3.18) and Windows 10. They analyzed that probe requests included in their frame body under the form of Information Elements (IEs), also called tagged parameters, or tags (e.g., ordered lists of tag numbers, extended capabilities, etc.) can be misused for tracking. Besides that sequential frame numbers or predictable scrambling seeds can be used for device identification and device tracking [2].

III. ITS Vehicle Identifier

In this paper, we categorize the available identifiers of vehicles into two classes. Primary vehicle identifiers represent those identifiers which will be typically considered today, e.g., the Vehicle Identification Number (VIN). Secondary Vehicle Identifiers come up with new information technology used in modern vehicles.

A. Primary Vehicle Identifier

To date, every vehicle is identifiable based on its unique VIN. In some areas, the VIN is integrated as human readable information in the windscreen of vehicles.

Besides the VIN, vehicles are marked with a license plate, which is already used for identification.

With the deployment of V2V technology vehicles will be equipped with a long term ECC key pair and an appropriate certificate [16] [17]. This certificate will become an additional primary vehicle identifier in future.

B. Secondary Vehicle Identifier

Modern vehicles are equipped with multi-media components (head-unit), which are able to establish communications with electronic devices of drivers or passengers. Typically, wireless communication technologies, e.g., Bluetooth, are used for that purpose.

A Bluetooth multi-media device emits a static 48 bit Media Access Control address, named MAC ID. The MAC ID is composed of two parts: the first half is assigned to the manufacturer of the device, and the second half is assigned to the specific device. In addition, each Bluetooth device emits a “User-friendly-name” which is typically alterable. Bluetooth devices operate in the ISM band (2.4 to 2.485 GHz).

Moreover, vehicle head-units allow any Wi-Fi ready laptop, tablet or mobile phone to access the internet within the vehicle while travelling if the head-unit has mobile communication capabilities. But head-units configured as access points need a unique Service Set Identifier (SSID) or network name to connect devices. In addition, each head-unit needs a unique MAC address.

If vehicles are equipped with mobile communication capabilities an International Mobile Subscriber Identity (IMSI) is required. This is a unique ID to identify a mobile device within the network. In addition, a SIM card with a dedicated mobile phone number is needed for mobile communication.

In [12], physical fingerprinting of wireless transmitters is investigated. Here, a complete feature set for physical fingerprinting of a transmitter is a secondary vehicle identifier. Vehicle identifiers mentioned so far are sufficient for identification all the time. Furthermore, vehicle identifiers with a limited validity period, e.g., pseudonymous certificates (termed authorization tickets by ETSI) exist. Pseudonymous certificates come up with the V2V technology.

Initially, Secondary Vehicle Identifier have no formal character in contrast to a license plate or VIN. But it is technically very easy to capture Bluetooth and Wi-Fi identifiers of a vehicle as shown in Section VI. So, attackers can misuse them for their purposes.

IV. Wireless Technologies

In this section, wireless technologies, which are applied in vehicles are described. In addition an analysis concerning identification capabilities based on wireless communication technologies is given. We only address local wireless communication technologies, which are quite easy to detect and omit mobile communications according the Global System for Mobile Communications (GSM) or the Long Term Evolution (LTE).

A. Bluetooth

Bluetooth is specified by the Bluetooth special interest group. The information mentioned here is based on the Bluetooth Specification version 5.0 [18].

The concept behind Bluetooth is to provide a universal short-range wireless communication capability using the 2.4 GHz Industrial Scientific Medicine (ISM) bands, available globally for unlicensed low-power uses.

There are two forms of Bluetooth wireless technology systems: Bluetooth Basic Rate (BR) and Bluetooth Low Energy (BLE). During our measurements we detected only Bluetooth (BR) compliant devices in the head-sets of the vehicles investigated.

Both systems include device discovery, connection establishment and connection mechanisms. The Basic Rate system includes optional Enhanced Data Rate (EDR), Alternate Media Access Control (MAC) and Physical (PHY) layer extensions. The Basic Rate system offers synchronous and asynchronous connections with data rates of 721.2 kb/s for Basic Rate, 2.1 Mb/s for Enhanced Data Rate and high speed operation up to 54 Mb/s with the 802.11 AMP. The BLE system includes features designed to enable products that require lower power consumption, lower complexity and lower cost than BR/EDR. The BLE system is also designed for use cases and applications with lower data rates and has lower duty cycles.

1) Bluetooth (BR) Technology: Bluetooth provides support for three application areas using short-range wireless connectivity:

- Data and voice access points: Bluetooth facilitates real-time voice and data transmissions by providing effortless wireless connection of portable and stationary communications devices
- Cable replacement: Bluetooth eliminates the need for numerous, often proprietary cable attachments for connection of practically any kind of communication devices. The range of each radio depends on the output power (up to 100 m)
- Ad hoc networking: A device equipped with a Bluetooth radio can establish an instant connection to another Bluetooth radio as soon as it comes into range

In vehicles, Bluetooth is used for connecting a smart phone to the:

- Hands-free phone system
- Vehicular head-unit to use the loudspeaker of the head-unit to output music from the smart phone

The Bluetooth architecture is divided into different layers. It starts with the Radio Frequency (RF) Layer, also termed physical layer (PHY). To be resilient to disturbances a frequency hopping spread spectrum (FHSS) is used. Three classes of transceivers are available with different output power. Power class 1: 100 mW, power class 2: 2,5 mW and power class 3: 1 mW.

Bluetooth (BR) uses 79 frequency channels, spaced 1 MHz apart. Channel n uses (where n is in the range 0 - 78) a carrier frequency of 2402+n MHz. Each frequency channel is divided into 1600 time slots per second; each slot is 625 μs long. Each data packet may use between 1 and 5 slots and is transmitted on a different frequency channel, following a pseudo-random hopping sequence determined by the device address of the master device.

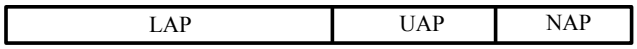


Figure 1. Structure of a Bluetooth Device Address

At first, Bluetooth devices have to establish a connection, termed pairing, to exchange data. This procedure is initiated by the host device based on the inquiry process. During this process Bluetooth devices respond with inquiry reply messages including BD_ADDR and clock rate (CLK), etc. During the pairing process the jump sequence for sharing the channels is calculated by the master device and synchronized with the slave devices.

There exists a range of Bluetooth Specification versions from Bluetooth 1.0a (published 1999) to Bluetooth 5.0 (published 2016).

2) Identification Capabilities: A Bluetooth multi-media device emits a static 48 bit MAC identifier (BD_ADDR). The MAC ID is composed of three parts: Lower Address Part (LAP), Upper Address Part (UAP), and Nonsignificant Address Part (NAP). NAP (16 bit) and UAP (8 bit) are assigned to the manufacturer of the device, and LAP (24 bit) is assigned to the specific device.

In addition, each Bluetooth device emits a "User-friendly-name" which is typically alterable. BD_ADDR and the "User-friendly-name" are the primary identifiers. In addition, the data set of a Bluetooth device: CLK, Bluetooth device profile, and the Host Controller Interface (HCI) can be used for identification purposes (Table I), too.

3) Bluetooth Low Energy: BLE is a low-power wireless technology for short-range control and monitoring applications. It operates in the 2.4 GHz ISM band as well. It uses 40 radio channels. 3 channels are primarily used for advertising. For BLE only one packet format is specified in the link layer. It consists of:
Preamble | Access Code | PDU | CRC.
The access code includes the 48 bit device address.

There are only two PDU formats in BLE, one for advertising packets and one for data packets.

The standard distinguishes between public and random device addresses. A public and a random device address are both 48 bits in length. To avoid tracking of a device, random device addresses should be used. But random device addresses can only be applied for data packets in a connection mode not for advertising packets.

The random device address may belong to either of the following two sub-types:

- Static address
- Private address

The term "Static address" means that the device initializes its static address to a new value after each power cycle. Private addresses are changed during operation at a fixed frequency.

As long as the Bluetooth device is not powered down and up, the static address has not changed and sniffed bluetooth advertising packets of one Bluetooth device can be linked. For privacy reasons private random addresses should thus be used.

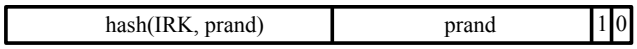


Figure 2. Structure of a resolvable private Bluetooth device address

A private address may belong to either of the following two sub-types:

- Non-resolvable private address
- Resolvable private address

Resolvable private addresses have the positive side effect that already connected devices can be identified later on though the device address has changed in the meantime. Therefore, a specific device, namely, Identity Resolving Key IRK is needed, which is transmitted from the Bluetooth device to the Bluetooth component in the vehicle after a first paring procedure. The IRK is linked to an identity at the Bluetooth host. Figure 2 depicts the structure of a 48 bit resolvable private Bluetooth address. It consists of three different parts:

- bit mask '10' indicating a random resolvable private address
- 22 bit random value *prand* and
- 24 bit hash value $hash(IRK, prand)$

If a Bluetooth host receives a data packet with resolvable private address it calculates for all known IRK_i $hash'(IRK_i, prand)$ and compares this value with the current value of $hash(IRK, prand)$.

$$hash'(IRK_i, prand) \stackrel{?}{=} hash(IRK, prand) \qquad (1)$$

If this equation holds for one IKS the component is identified and pairing can again be established.

B. Wireless Local Area Network (Wi-Fi)

Primary, Wi-Fi is based on the communication standards which was made for cable based Local Area Networks (LAN), IEEE 802.11x.

1) Technology: Briefly spoken, Wi-Fi devices support two different modes:

- Ad hoc mode, termed independent BSS (IBSS): Wi-Fi devices communicate peer-to-peer. During the communication data pakets are sent to all devices of the network but discarded by the devices if the destination address does not fit
- Access point mode, termed Basic Service Set (BSS): All Wi-Fi devices are connected with the access point (hot spot)

Head-units of modern vehicles provide Wi-Fi hot spots. So any Wi-Fi ready laptop, tablet or mobile phone is able to access the internet within the vehicle while travelling if the head-unit has mobile communication facilities (GSM, LTE).

Different Wi-Fi Standards exist: IEEE 802.11b / g / a / n / ac. They differ in the frequency band used (2,4 GHz and/or 5 GHz), and communication speed (1 Mbit/s . . . 6,96 Gbit/s). The frequency band is split into channels (2,4 GHz: 13 channels with a bandwidth of 20 or 40 MHz, hence 5 channels are needed to establish a network). In the 5 GHz Wi-Fi frequency band channels have a bandwidth of 20, 40, 80 or 160 MHz.

TABLE I. TECHNOLOGY SPECIFIC IDENTIFICATION FEATURES

Technology	First Level Features	Second Level Features
Bluetooth	MAC ID (BD_ADDR) "friendly name"	CLK, Bluetooth device profil Host Controller Interface
IEEE 802.11 X (Wi-Fi)	MAC ID (BSSID) "SSID"	Information in Beacon Frames

One type of the management frames in IEEE 802.11 based Wi-Fis is a beacon frame. Beacon frames are transmitted periodically to announce the presence of a wireless LAN and contain information about the network. Beacon frames are transmitted by the access point in an infrastructure Basic Service Set (BSS). In IBSS networks beacon generation is distributed among the stations.

2) Identification Capabilities: Primary identifiers are:

- Basic Service Set ID (BSSID) or MAC address of the Wi-Fi device and
- SSID (primary name associated with an 802.11 wireless local area network with a maximum length of 32 characters)

In addition, information in Wi-Fi beacon frames could be used for identification, too (Table I).

3) Random Device Address: A common specification on Wi-Fi MAC address randomization does not yet exist. However, proprietary Wi-Fi MAC randomization algorithms are implemented in iOS (starting from iOS 8), Android (starting from Android 6.0), Linux (starting from Kernel 3.18) and Windows 10. Unfortunately, these mechanisms are only available if the Wi-Fi card and driver support it.

V. MEASUREMENTS

In this section, the test cases performed and the test equipment used are described.

A. Aim of the Measurements

By means of the measurements, we investigate vehicular Bluetooth as well as Wi-Fi communication capabilities especially for identification purposes outside the vehicle. Therefore, the following measurements, divided into test cases, are performed:

- Test case 1: Radiation characteristics
- Test case 2: Signal strength
- Test case 3: Activity of the transmitter
- Test case 4: Detection of Secondary Vehicle Identifier in stand still mode of the vehicle
- Test case 5: Detection of Secondary Vehicle Identifier in driving mode of the vehicle

B. Test Vehicles

The following vehicles were investigated during the measurements:

- Skoda Octavia III (two different models equipped with Bluetooth chips from Qisda Corporation or Alps Electronics Co. LTD are investigated): Only used for Bluetooth measurements
- VW Passat B8: Only used for Bluetooth measurements

- Opel Astra 2016 incl. OnStar: Only used for Wi-Fi measurements
- Opel Insignia Innovation 2016 incl. OnStar: Only used for Wi-Fi measurements

C. Test Equipment

1) Bluetooth Test Equipment for the Tests in Section VI-A:

- Notebook
  - ThinkPad X201 with Kali Linux (64 Bit, version 2016.2), BTScanner version 2.0, and Kismet version 2016-07-R1
  - Ubertooth One (firmware git-579f25) with Ubertooth-Specan-Ui, and Ubertooth-Rx version 201-10-R1 [19]
  - Standard antenna, LogPer antenna, and directional antenna WIFI-LINK WAVEGUIDE Antenna PN: WCA-2450-12, frequency range 2,4 - 2,5 GHz, 12 dBi
- Smart phone
  - Samsung Galaxy S6, Android 6.0.1, Bluetooth-Scanner app version 1.1.3 (from Google Playstore)

2) Bluetooth Test Equipment for the Tests in Section VI-B:

- Notebook
  - Lenovo ThinkPad T400 with Kali Linux (64 Bit, version 2018.2), BTScanner version 2.1-6
  - Ubertooth One (firmware-version: 2018-06-R1, API 1.03) with Ubertooth-Specan-Ui, and Ubertooth-Rx [19] and the following antennas are used: standard antenna, Ettus VERT2450 antenna, and WIFI-LINK WAVEGUIDE antenna PN: WCA-2450-12, 2,4-2,5 GHz, 12 dBi
  - USB Bluetooth stick: AVM BlueFritz! USB v2.0
- Smart phone
  - Sony Smartphone Xperia Z5 Compact

3) Wi-Fi Test Equipment:

- Notebook
  - Notebook Lenovo ThinkPad T400, Ubuntu 16.04 LTS and LinSSID version 2.7
  - USB-Wi-Fi-device: TP-Link TL-WN722N with standard antenna and directional antenna WIFI-LINK WAVEGUIDE Antenna PN: WCA-2450-12, 2,4-2,5 GHz, 12 dBi
- Smart Phone
  - Huawei P8 lite 2017, Wifi-Analyzer App (from Google Playstore)
  - Samsung S7, Wifi-Analyzer App (from Google Playstore)

VI. Measurements and Results

In this section the test results of the performed tests are described. In Section VI-A an Octavia III head-unit with a Bluetooth chip of the Qisda Corporation is examined, whereas in Section VI-B an Octavia III head-unit with a Bluetooth chip of Alps Electronic Co. LTD is considered.

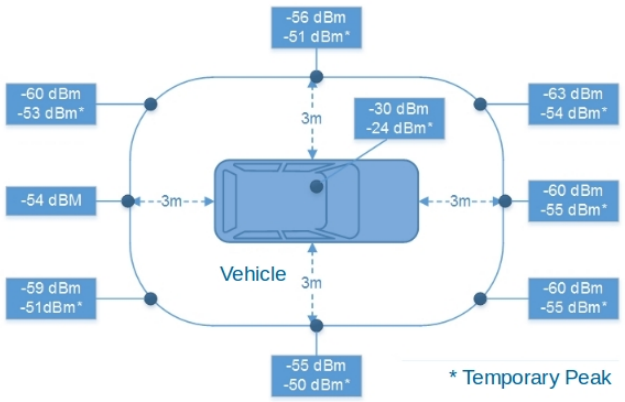


Figure 3. Radiation characteristic of the Octavia III Bluetooth device

TABLE II. SIGNAL STRENGTH OF THE OCTAVIA BLUETOOTH DEVICE

Distance	Standard Antenna	LogPer Antenna	Directional Antenna
3 m	-50 dBm	-56 dBm	-47 dBm
6 m	53 dBm	-60 dBm	-51 dBm
9 m	-63 dBm	-63 dBm	-54 dBm
12 m	-67 dBm	-65 dBm	-56 dBm
15 m	-71 dBm	-68 dBm	-60 dBm
18 m	-75 dBm	-69 dBm	-63 dBm
21 m	-78 dBm	-72 dBm	-65 dBm
30 m		-75 dBm	-68 dBm

A. Bluetooth Measurements for the Octavia III equipped with a Bluetooth chip of the Qisda Corporation (and partly Passat)

1) Test Case 1: As test equipment, a Lenovo ThinkPad X201, with Ubertooth One, Ubertooth-Specan-Ui and standard antenna is used. Measurements are performed at one position inside and 8 positions outside the vehicle. The positions and results are plotted in Figure 3. As we expected, the highest signal strength of -30 dBm has been detected inside the vehicle. But outside the vehicle, a strong signal strengh has also been measured.

2) Test Case 2: As test equipment a Lenovo ThinkPad X201, with Ubertooth One, Ubertooth-Specan-Ui and different antennas is used: Standard antenna, LogPer antenna and directional antenna WIFI-LINK WAVEGUIDE. The test results are presented in Table II. With all antennas the Bluetooth signal can always be detected, within a distance of 21 m.

3) Test Case 3: The Bluetooth module of the head-unit starts with scanning of Bluetooth devices which were already paired in the past and are registered in the pairing list of the head-unit after starting the ignition. Scanning is switched off after the deactivation of the ignition and removal of the key.

4) Test Case 4: First, a Samsung Galaxy S6 with the Bluetooth scanner app is utilised as test equipment. Figure 4 presents the test setting. The following information about the Bluetooth device of the head-unit can be captured with the test equipment mentioned:

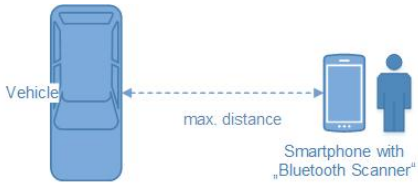


Figure 4. Test arrangement for the detection of Secondary Vehicle Identifier in stand still mode

Listing 1. BD_ADDR and "friendly name" of the head-unit of a Skoda Octavia

```
Skoda_TF
00:17:CA:D9:6B:77 (-65 dBm)
AUDIO_VIDEO_HANDSFREE
Scan Cycle 199 (20.11.16 15:01)
```

SSID “Skoda_TF”, BSSID “00:17:CA:D9:6B:77”, the service “AUDIO_VIDEO_HANDSFREE” and the “Scan Cycle 199 (20.11.16 15:01)” with date were captured. This information is readable up to a distance of 24 m (signal strength at this distance: -83 dBm) (it has to be mentioned that the owner of the Skoda Octavia III has already altered its SSID. “Skoda_TF” is not the factory setting).

The following information is captured from the Bluetooth device of the head-unit of the Passat up to a distance of 12 m (signal strength at this distance: -84 dBm):

Listing 2. BD_ADDR and "friendly name" of the head-unit of a VW Passat

```
VW BT 2058
A8:54:B2:FE:30:35 (-79 dBm)
AUDIO_VIDEO_HIFI_AUDIO
Scan Cycle 25 (02.11.16 13:15)
```

From a privacy perspective it is remarkable, that the name of the car manufacturer is part of the SSID and that the number part “2058” of the SSID is chosen from the VIN of the Passat.

Next, Lenovo ThinkPad X201, Ubertooth One with Ubertooth-Rx are used as test equipment to perform the same test case. The subsequent information can be captured if the test equipment is switched on and a Samsung Galaxy S6 is connected to the Octavia III head-unit:

Listing 3. Galaxy S6 connected to the Octavia head-unit

```
systime=1479652524 ch=39 LAP=d96b77 err=0
clkn=100728 clk_offset=1540 s=-35 n=-55 ...
systime=1479652571 ch=39 LAP=68dae3 err=0
clkn=250437 clk_offset=5596 s=-21 n=-55 ...
systime=1479652571 ch=39 LAP=68dae3 err=0
clkn=251217 clk_offset=5613 s=-16 n=-55 ...
```

This information can be captured up to 18 m with the standard antenna and up to 42 with the directional antenna.

5) Test Case 5: Using the test equipment Samsung Galaxy S6 with the Bluetooth scanner app, the subsequent information can be captured up to a speed of 30 km/h. Figure 5 shows the test case.

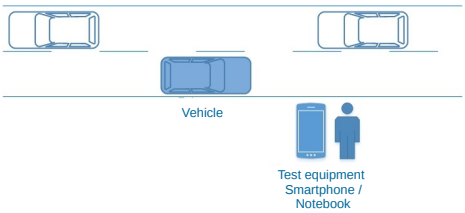


Figure 5. Test arrangement for the detection of secondary vehicle identifier in driving mode

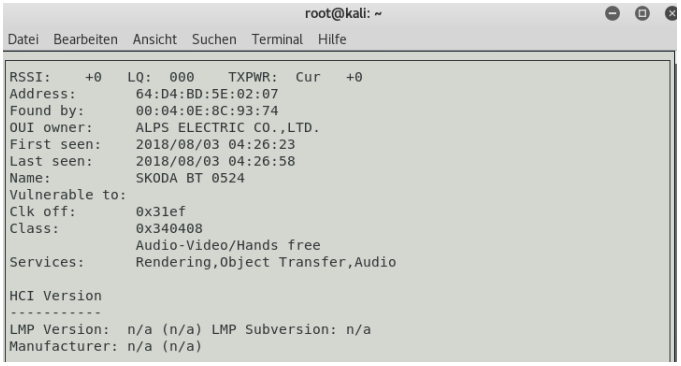


Figure 6. Screenshot of the BTScanner during the measurement

Listing 4. SSID and BSSID in driving mode

```
Skoda_TF
00:17:CA:D9:6B:77 (-65 dBm)
AUDIO_VIDEO_HANDSFREE
Scan Cycle 199 (20.11.16 15:01)
```

B. Bluetooth Measurements for the Octavia III equipped with a Bluetooth chip of Alps Electronic Co. LTD

1) Test Case 4: As test equipment a Lenovo ThinkPad T400 with Kali Linux Version 2018.2, an USB Bluetooth stick AVM BlueFritz! USB 2.0 and a BTScanner version 2.1-6 were used to sniff information of the Bluetooth head-unit of the Octavia in stand still mode and with enabled ignition. Figure 6 presents the captured information. This information (BD_ADDR, “friendly name”) can be captured up to a distance of 67 m between vehicle and measurement device. The test arrangement is shown in Figure 4.

Next, the sniff distance of an existing Bluetooth communication between a paired smartphone (Sony Smartphone Xperia Z7 Compact) and the head-unit was investigated. As test equipment Lenovo ThinkPad T400 with Kali Linux Version 2018.2 and Ubertooth One with Ubertooth-Rx was applied. Table III presents the maximum detection distance with different antennas for a successful receiving of the BD_ADDR of the head-unit.

2) Test Case 5: Figure 5 depicts the test arrangement. As test equipment a Lenovo ThinkPad T400 with Kali Linux Version 2018.2, an USB Bluetooth stick AVM BlueFritz! USB 2.0 and BTScanner version 2.1-6 were used to detect the BD_ADDR of the head-unit of the Octavia III in driving mode.

TABLE III. DETECTION DISTANCE OF A PAIRED COMMUNICATION

Antenna	Detection Distance
Standard antenna	73 m
VERT2450	89 m
LogPer Antenna	112 m

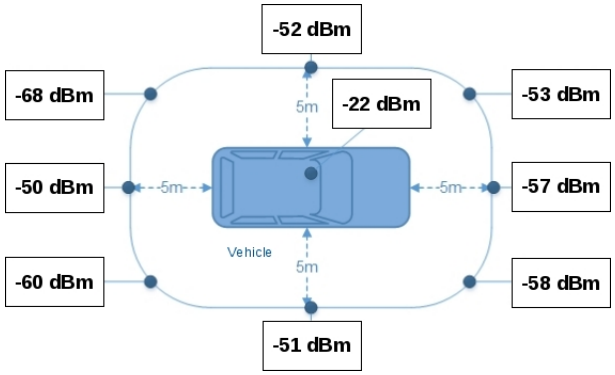


Figure 7. Radiation characteristic of the Opel Insignia Wi-Fi device

The test equipment was located at a distance of 10 m from the street to monitor the driving Octavia III. Up to a speed of 50 km/h we could identify the BD_ADDR of the head-unit. We stopped the investigation at this point. 50 km/h is the speed-limit inside cities in Europe.

C. Wi-Fi Measurements for the Opel Insignia (partly Opel Astra)

1) Test Case 1: As test equipment a Lenovo ThinkPad T400, TP-Link TL-WN722N with standard antenna, and LinSSID is used. The signal strength of the Wi-Fi access point (Wi-Fi-AP) has been measured at 8 fixed points outside and at 1 point inside the vehicle. The positions are equal to the Bluetooth test case. But in contrast to the Bluetooth measurement, the distance between the vehicle and the measurement tool is 5 m. The results for the Opel Insignia are plotted in Figure 7. As we expected, the highest signal strength of -22 dBm has been detected inside the vehicle. But outside the vehicle, a strong signal strengh has also been measured.

2) Test Case 2: As test equipment a Lenovo ThinkPad T400, TP-Link TL-WN722N with standard antenna, and LinSSID on the one hand and Samsung S7, and Wifi-Analyzer on the other hand are used. Using the TP-Link TL-WN722 and the Samsung S7 the signal strength is measured in increasing distance from the vehicle, in the direction of the right front door. The results are plotted in Figure 8. Only small differences in signal strength can be detected between an active connection and a non connection of a client to the Wi-Fi-AP of the Opel Insignia. The measurement sensitivity of the smart phone is about 10 dBm lower for distances greater 10 m in contrast to the measurements with the TP-Link. With both measurement devices the signal of the Wi-Fi-AP can always be detected, within a distance of 60 m.

3) Test Case 3: As test equipment a Lenovo ThinkPad T400, TP-Link TL-WN722N with standard antenna, and LinSSID is used.

General Motors and Opel provide vehicle online connectivity based on the OnStar service. Only if the OnStar service is

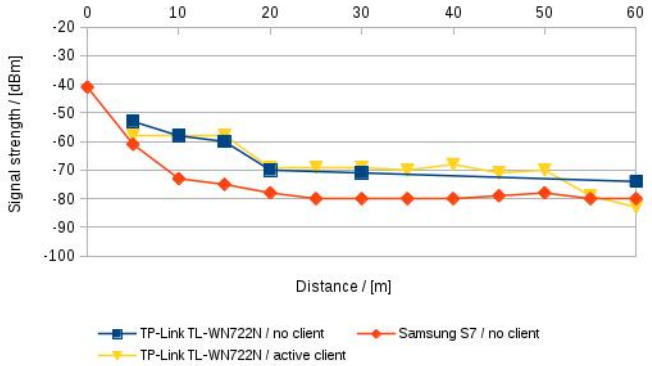


Figure 8. Radiation characteristic of the Opel Insignia Wi-Fi device

TABLE IV. SIGNAL STRENGTH OF THE ASTRA WI-FI DEVICE IN STAND STILL MODE

Distance	Signal strength Huawei P8 lite 2017	Signal strength TP-Link TL-WN722N
216 m	-82 dBm	-81 dBm
424 m	no signal	-91 dBm

enabled the Wi-Fi-AP of the Opel Insignia can be switched on. The Wi-Fi transmitter is activated when the ignition is started and deactivated when the key is removed from the ignition lock. Enabling or disabling the Wi-Fi-AP is not possible for the driver, using only the configuration menu implemented in the vehicle (disabling is possible with an appropriate smartphone app).

4) Test Case 4: As test equipment a Lenovo ThinkPad T400, TP-Link TL-WN722N with standard antenna, and LinSSID on the one hand and Samsung S7, and Wifi-Analyzer on the other hand are used. Figure 4 presents the test setting. In stand still mode the following Secondary Vehicle Identifier and additional information has been measured for the Wi-Fi device of the Opel Insignia, for all distances up to 60m with both test equipments.

Listing 5. SSID and BSSID of an Opel Insignia head-unit

```
SSID: WiFi Hotspot 1760
BSSID: C4:49:BB:21:91:DE
Frequency: 2437 MHz; 2448-2426 = 22 MHz
Channel: 6
Misc.: WPA2-PSK-CCMP+TKIP, ESS,
       MITSUMI ELECTRIC Co.,LTD
```

Next, we determine the maximum detection distance for the Secondary Vehicle Identifiers. As test equipment a HP notebook, TP-Link TL-WN722N with standard antenna, and a LinSSID on the one hand and a Huawei P8 lite 2017 with a Wifi-Analyzer on the other hand are used. The results are shown in Table IV for the Wi-Fi device of the Opel Astra. If a signal has been detected, then the SSID and the BSSID can always be extracted. The smart phone detected a signal up to 216 m, the USB-Wi-Fi-device up to 424 m.

5) Test Case 5: As test equipment a HP notebook, TP-Link TL-WN722N with standard antenna, and a LinSSID on the one hand and a Huawei P8 lite 2017 with Wifi-Analyzer

TABLE V. SIGNAL STRENGTH OF THE ASTRA WI-FI DEVICE IN DRIVING MODE

Speed	Maximum signal strength Huawei P8 lite 2017	Maximum signal strength TP-Link TL-WN722N
50 km/h	-60 dBm	-55 dBm
100 km/h	-71 dBm	-50 dBm

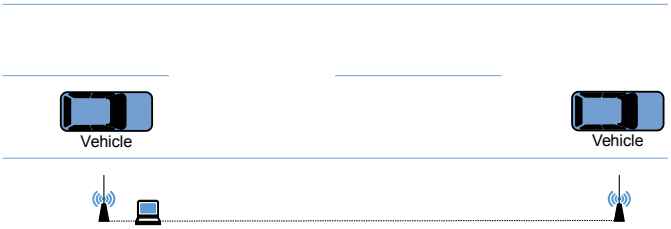


Figure 9. Attack scenario: Tracking of vehicles

app on the other hand are used. The notebook with USB - Wi-Fi device and the smart phones operate 1 m above the floor beside the roadway. Figure 5 shows the general test case. The results for the Wi-Fi device of the Opel Astra are presented in Table V. The maximum signal strength has been detected by the USB-Wi-Fi-device. The measured signal strengths with the TP-Link for 50 and 100 km/h are surprising. We assume that this issue is caused by the moving vehicle and the sample rate of the measurement devices of about 1 Hz (vehicle moves 13,9 m/s at 50 km/h and 27,8 m/s at 100 km/h).

VII. Tracking of Vehicles

We have shown that the identification and tracking of vehicles based on Secondary Vehicle Identifier can be performed with very cheap technical measurement equipment. This capability can be misused for tracking of vehicles. A technical measurement infrastructure to perform such kind of tracking is shown in Figure 9. Here, we consider only adversaries who passively sniff the communication.

To monitor vehicle motions in a specific geographic region a dense net of road side stations operating as sniffer would be needed. Due to current privacy regulations such an infrastructure for tracking of individual vehicles can be precluded in Western Europe [20]. It seems more realistic to be identified by scattered receivers of crucial neighbours monitoring vehicle motions in a street.

VIII. Countermeasures

A. Technology Independent Measures

In principle, wireless communication technology enables the identification and tracking of vehicles. One basic requirement to avoid privacy violations based on wireless interfaces or communication technologies is to avoid static identifiers in the whole communication stack. For example, communication technology for the vehicle-2-vehicle communication technology applies this rule [16].

Identification and tracking are completely excluded if the wireless communication components are powered down. But in vehicles Bluetooth (BR) is used for connecting a smart phone to the:

- Hands-free phone system
- Vehicular head-unit to use the loudspeaker of the head-unit to output the music from the smart phone

If drivers use this Bluetooth capabilities they will not quit the usage due to possible privacy risk.

The tracking problem can be reduced, however, if the antennas are located inside the vehicle and the field strength of the Bluetooth and Wi-Fi transmitters are limited. Especially during connection mode the field strength can be reduced to a necessary range to retain the data communication.

B. Technology Dependent Measures

1) Bluetooth: An alternative to Bluetooth (BR) to avoid simple tracking is the usage of Bluetooth Low Energy. BLE has specific privacy features, which are briefly described in Section IV-A3. In particular, private Bluetooth addresses should be used. This feature avoids the tracking of Bluetooth devices in connection mode. But the issue of tracking is still valid if BLE components are in advertising mode.

Obviously, private addresses have to be used for the Bluetooth interface in the head-unit als well as for the connected Bluetooth components.

2) Wi-Fi: A common standard for MAC ID randomization for Wi-Fi components is still missing. There are proprietary implementations for operating systems mentioned in Section IV-B3. The mechanism implemented in Windows 10 [21], [2]. Random MAC IDs are possible with Windows 10 if hardware and driver supports this issue. Interesting is that Windows 10 does not only use random device addresses during probe requests. It also employs a random address when connecting to a network. Further detailed investigations are needed to suggest adequate solutions for Wi-Fi MAC ID randomization for vehicles which are compliant with the Wi-Fi standard.

IX. Conclusion and Future Work

As shown in Section VI, it is technically very easy to capture Secondary Vehicle Identifiers based on wireless interfaces of vehicles, especially Bluetooth and Wi-Fi (even with low cost equipment as shown in this paper). Although, these interfaces are designed to connect the devices of passengers, vehicle identifiers can be detected far away from the vehicle (424 m for Wi-Fi with a TP-Link device) and high vehicle speed of up to 100 km/h. This enables the misuse of vehicle identifiers for the tracking of vehicles. At least, MAC ID randomization is needed for Bluetooth and Wi-Fi interfaces in vehicular head-units. BLE already supports random device addresses. For Wi-Fi only proprietary solutions are available. Altogether further investigations are needed to propose random MAC ID solutions which can be broadly applied in vehicular devices.

In general, vehicle manufacturer avoid to produce country specific vehicles. So we expect that our measurements hold for all instances of the analyzed car models at least in Europe.

In the context of the upcoming V2V communication our results are worrysome concerning the privacy of vehicles and drivers. The V2V communication is a short range communication technology with a communication range of about 800 m in open space.

In the future, every vehicle will periodically broadcast Cooperative Awareness Messages (CAM) with a packet generation rate of 1 up to 10 Hz. A CAM contains a lot of

data about the sending vehicle: current geographic position, speed, driving direction, etc., at a specific time. One privacy requirement is that a receiver can not link a CAM to a specific vehicle. Secondary Vehicle Identifiers can be misused to link captured CAM messages to a specific vehicle [22].

ACKNOWLEDGEMENT

The authors would like to thank our colleagues Christian Berghoff, Tobias Franz and Thomas Strubbe for detailed discussions. Also thanks to the anonymous reviewers for the valuable comments.

REFERENCES

[1] M. Ullmann, T. Franz, and G. Nolden, "Vehicle Identification Based on Secondary Vehicle Identifier - Analysis, and Measurements -," in Proceedings VEHICULAR 2017: The Sixth International Conference on Advances in Vehicular Systems, Technologies and Applications. IARIA, 2017, pp. 32–37.

[2] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, and F. Piessens, "Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms," in Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. ACM, 2016, pp. 413–424.

[3] M. Ullmann, T. Strubbe, and C. Wieschebrink, "Technical Limitations, and Privacy Shortcomings of the Vehicle-to-Vehicle Communication," in Proceedings VEHICULAR 2016: The Fifth International Conference on Advances in Vehicular Systems, Technologies and Applications. IARIA, 2016, pp. 15–20.

[4] H. Lee, D. Kim, D. Kim, and S. Y. Bang, "Real-time automatic vehicle management system using vehicle tracking and car plate number identification," in Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on, vol. 2. IEEE, 2003, pp. II–353.

[5] R. Bruno and F. Delmastro, "Design and Analysis of a Bluetooth-Based Indoor Localization System," 2003, pp. 711–725.

[6] R. Zhou, "Wireless Indoor Tracking System (WITS)," Aktuelle Trends in der Softwareforschung, Tagungsband zum IT Software-Forschungstag. Dpunkt Verlag Heidelberg, Germany, 2006, pp. 163–177.

[7] C. Carpenter, M. Fowler, and T. Adler, "Generating Route-Specific Origin-Destination Tables Using Bluetooth Technology," Transportation Research Record: Journal of the Transportation Research Board, no. 2308, 2012, pp. 96–102.

[8] M. Mueller, D. Schulz, M. Mock, and D. Hecker, "Detecting mobility patterns with stationary bluetooth sensors: A real-world case study," in Proceedings of the 18th AGILE International Conference on Geographic Information Science, 2015.

[9] M. Emery and M. K. Denko, "IEEE 802.11 WLAN Based Real-Time Location Tracking in Indoor and Outdoor Environments," in Electrical and Computer Engineering, 2007. CCECE 2007. Canadian Conference on. IEEE, 2007, pp. 1062–1065.

[10] R. M. Ishtiaq Roufa, H. Mustafaa, S. O. Travis Taylora, W. Xua, M. Gruteserb, W. Trappeb, and I. Seskarb, "Security and privacy vulnerabilities of in-car wireless networks: A tire pressure monitoring system case study," in 19th USENIX Security Symposium, Washington DC, 2010, pp. 11–13.

[11] S. Astapov and A. Riid, "A Multistage Procedure of Mobile Vehicle Acoustic Identification for Single-Sensor Embedded Device," International Journal of Electronics and Telecommunications, vol. 59, no. 2, 2013, pp. 151–160.

[12] B. Danev, D. Zanetti, and S. Capkun, "On Physical-Layer Identification of Wireless Devices," ACM Computing Surveys (CSUR), vol. 45, no. 1, 2012, p. 6.

[13] T. Jiang, H. J. Wang, and Y.-C. Hu, "Preserving location privacy in wireless lans," in Proceedings of the 5th international conference on Mobile systems, applications and services. ACM, 2007, pp. 246–257.

[14] Wang, Ping, "Bluetooth Low Energy-privacy enhancement for advertisement," Master's thesis, Norwegian University of Science and Technology, Departement of Telematics, 2014.

[15] Andrew Hilts, Christopher Parsons, and Jeffrey Knockel, "Every Step You Fake: A Comparative Analysis of Fitness Tracker Privacy and Security," 2016, https://openeffect.ca/reports/Every_Step_You_Fake.pdf access date: July 30, 2018.

[16] ETSI, "ETSI TR 102 893 V1.1.1: Intelligent Transport Systems (ITS); Security; Threat, Vulnerability and Risk Analysis (TVRA); Technical Report," 2010, http://www.etsi.org/, Access Date: June 02, 2017.

[17] ——, "ETSI EN 302 665 V1.1.1: Intelligent Transport Systems (ITS) - Communications Architecture," 2010, http://www.etsi.org/, Access Date: June 02, 2017.

[18] "Bluetooth Core Specification, v 5.0," 2011, https://www.bluetooth.com/specifications/bluetooth-core-specification, access date: July 25, 2018.

[19] Ubertooth Developer, "Ubertooth Bluetooth Sniffer," 2017, https://github.com/greatscottgadgets/ubertooth/, access date: March 24, 2017.

[20] European Commission, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016, https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en access date: August 03, 2018.

[21] C. Huitema, "Experience with mac address randomization in windows 10," in 93th Internet Engineering Task Force Meeting (IETF), 2015.

[22] M. Ullmann, and T. Strubbe, and C. Wieschebrink, "Misuse Capabilities of the V2V Communication to Harm the Privacy of Vehicles and Drivers," in International Journal On Advances in Networks and Services, vol 10 no 12. IARIA, 2017.

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601