

Received March 20, 2019, accepted April 13, 2019, date of publication April 25, 2019, date of current version May 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913185

Expanded Combinatorial Designs as Tool to Model Network Slicing in 5G

DANILO GLIGOROSKI, (Member, IEEE), AND KATINA KRALEVSKA^{ID}, (Member, IEEE)

Department of Information Security and Communication Technologies, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Corresponding author: Katina Kralevska (katinak@ntnu.no)

ABSTRACT The network slice management function (NSMF) in 5G has a task to configure the network slice instances and to combine network slice subnet instances from the new-generation radio access network and the core network into an end-to-end network slice instance. In this paper, we propose a mathematical model for network slicing based on combinatorial designs such as Latin squares and rectangles and their conjugated forms. We extend those designs with attributes that offer different levels of abstraction. For one set of attributes, we prove a stability Lemma for the necessary conditions to reach a stationary ergodic stage. We also introduce a definition of utilization ratio function and offer an algorithm for its maximization. Moreover, we provide algorithms that simulate the work of NSMF with randomized or optimized strategies, and we report the results of our implementation, experiments, and simulations for one set of attributes.

INDEX TERMS 5G networks, combinatorial designs, dynamic deployment, Latin squares, Latin rectangles, network slicing, optimal slice selection.

I. INTRODUCTION

In the global market capitalization, 5G technologies are projected to be worth over USD 12.3 trillion by 2035 [1]. Network slicing is seen as the key enabling technology that can bring up to 150% increased revenues for the operators, in comparison with the classical one-big network concept [2]. The idea for network slicing in 5G came from the telecommunication industry alliance NGMN in February 2015 [3] and very shortly afterwards was accepted by 3GPP [4] as an enabling technology that will bring new services and markets.

The role of network slicing is to enable functional and operational diversity on a common network infrastructure [5]. The idea is to create multiple isolated networks, termed Network Slice Instances (NSIs), on a common physical infrastructure where physical and virtual resources of each NSI are customized to satisfy the requirements for a specific communication service. Fig. 1 presents the management phases of a NSI: 1. preparation; 2. commissioning; 3. operation; and 4. decommissioning. The preparation phase includes all steps required before the creation of a NSI (creation and verification of network slice template, evaluation of network slice requirements, capacity planning). The lifecycle of a NSI starts with the second phase. During the commissioning phase, the NSI is created and all resources for the NSI are allocated and instantiated. In the operation phase, the NSI

The associate editor coordinating the review of this manuscript and approving it for publication was Debashis De.

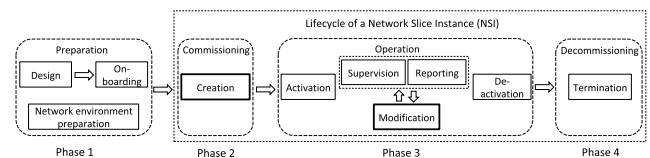


FIGURE 1. Management aspects of network slice instance [6]. We propose a model based on combinatorial designs for the creation and modification steps (represented with thick frames).

supports a communication service. First, the NSI is activated and later performance reporting for KPI monitoring as well as modification and de-activation of the NSI happen. The last phase of NSI lifecycle and NSI management includes termination of the NSI by releasing the dedicated resources and removing the NSI specific configuration from the shared resources. After this phase, the NSI does not exist anymore.

The slicing is performed end-to-end (E2E) [7], [8]. Thus, a NSI contains Network Slice Subnet Instances (NSSIs) in the New-Generation Radio Access Network (AN) and the Core Network (CN), referred to as AN and CN NSSIs in Fig. 2, and the interconnections between them. NSSI is a set of network functions (NFs) which can be physical NFs or virtualized NFs. The 3GPP management system contains information for interconnected NFs such as topology and individual link requirements. Fig. 2 shows that one NSI may support a single (e.g. NSI 1) or multiple communication services (e.g. NSI 3). AN and CN NSSIs can be dedicated

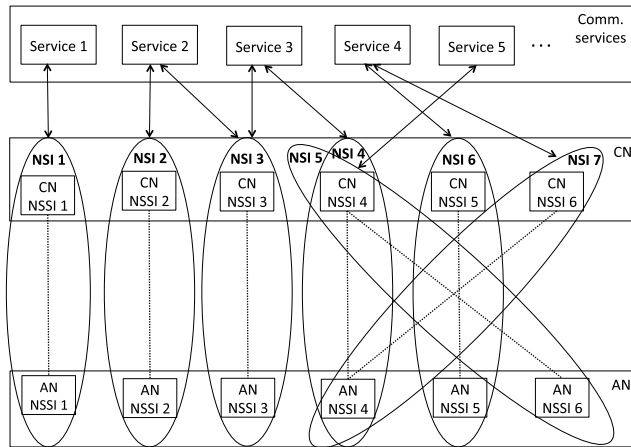


FIGURE 2. Five services supported by seven NSIs. The NSIs contain NFs, belonging to CN and AN NSSIs, and the interconnection information between the NFs.

to one NSI (e.g. CN NSSI 1) or shared by two or more NSIs (e.g. CN NSSI 4).

In general, a demanding tenant issues a communication service request which is translated into a slice request (network functions and infrastructure requirements) for the Mobile Network Operator (MNO). The following management functions manage the NSIs to support communication services: Communication Service Management Function (CSMF), Network Slice Management Function (NSMF) and Network Slice Subnet Management Function (NSSMF). CSMF receives the communication service related requirements by the tenant and converts them into network slice related requirements which are sent to NSMF. NSMF manages and orchestrates the NSI. It configures the NSIs and knows which NSSIs are associated with each NSI (cross-domain management and orchestration (M&O)). One NSSI can be associated with multiple NSIs where NSSMF manages and orchestrates the NSSIs. The network slice is instantiated and configured by NSMF where NSMF manages the interactions among the slice instances in terms of resources and features sharing (cross-slice M&O). For instance, in Fig. 2, both AN NSSI 1 in the access part and CN NSSI 1 in the core part first have to be defined and instantiated. Then NS 1 is instantiated by combining these two NSSIs.

In spite of the vast number of articles devoted to network slicing, it comes as a surprise that there are still no general precise mathematical models for network slicing and building such models is a challenging task as suggested in [1], [9]. Moreover, even the taxonomy used by different standardization organizations (for example 3GPP and IETF) is not agreed, although they are addressing the same slicing scenarios. For example what is referred as “hard slicing” by IETF, is referred as non-shared network slice subnet instance by 3GPP (see Definition 1 and Definition 2 below). Similarly, “soft slicing” by IETF (Definition 3) corresponds to “shared constituent of network slice instance” (Definition 4) by 3GPP.

Definition 1 (IETF [10]): Hard slicing refers to the provision of resources in such a way that they are dedicated to a specific network slice instance.

Definition 2 (3GPP [11]): A NSSI that is dedicated to one NSI and is not shared as a constituent by two or more NSSI(s) is called a *non-shared NSSI*.

Definition 3 (IETF [10]): Soft slicing refers to the provision of resources in such a way that whilst the slices are separated such that they cannot statically interfere with each other, they can interact dynamically, which means they may compete for some particular resource at some specific time.

Definition 4 (3GPP [11]): A NSSI may be shared by two or more NSIs, this is called a *shared constituent of NSI*. A NF may be shared by two or more NSSI(s), in which case it is called a *shared constituent of NSSI*.

A. RELATED WORK

The idea for network slicing originates from the areas of Cloud Computing [12], Software Defined Networking (SDN) [13], Network Function Virtualization (NFV) [14] and Information-Centric Networking (ICN) [15]. One of the major research problems is the resource allocation across slices. Several works address the resource allocation in the radio access network or cross-domain on VNF level. We mention here some of the most prominent mathematical models developed for network slicing.

Reference [16] presents a mathematical model to construct network slice requests and to map them on the network infrastructure. The mapping process is performed on VNF level where it first places the VNFs to the nodes in the network and later it selects that paths between the VNFs and chains them. With the aim to maximize the long-term network utility, reference [17] uses a genetic algorithm to serve slice requests.

Network slicing brings new business models and interactions between the infrastructure providers, the tenants and the customers. This opens many directions for optimizations. The algorithm for admission and allocation of network slices requests in [18] maximizes the infrastructure provider’s revenue and ensures that the service guarantees provided to tenants are satisfied.

B. OUR CONTRIBUTION

In this paper, we offer one mathematical model for the Network Slice Management Function (NSMF) based on combinatorial designs and their algebraic properties. We see our contribution as one step closer to a general, precise and scalable mathematical model for network slicing. In particular, our mathematical model addresses the tasks of the NSMF in the creation and modification sub-phases of the NSI lifecycle (phases 2 and 3 in Fig. 1). The model uses combinatorial objects known as Latin squares (or Latin rectangles) to describe communication services and the NSSIs. Combinatorial designs [19] have been used for a long time in communications, networking and cryptography. The authors in [20] listed thirteen application areas of combinatorial designs, and in this paper we extend the list with one more application,

i.e., configuration of network slices in 5G. The mathematical properties of our model guarantee conflict resolution for services defined over network slices that compete for resources in CN and AN, as long as the configuration and modification of NSI and NSSI are performed within our model.

The next contribution of this paper is from an optimization point of view. We introduce the notion of utilization ratio function, with aims to describe the functional dependencies between the number of used network resources and the waiting time for establishing the network slice. We present two strategies for the work of NSMF, a non-optimized first-come-first-serve strategy and an optimal strategy, where the optimization objectives are: 1. to maximize the utilization of the network components; and 2. to decrease the average delay time from slice request to slice activation.

Finally, we show some simulation results. The optimal strategy achieved by maximizing the utilization ratio function, provides more than twice better performance in terms of the both objectives compared to the non-optimized strategy.

The rest of this paper is organized as follows. In Section II, we give examples of modeling network slicing with combinatorial designs. In Section III, we develop general and extended combinatorial designs model for cross-domain end-to-end network slicing that includes both hard and soft slicing. In Section IV, we instantiate our general model with several concrete attributes and present algorithms for simulation and optimization of a NSMF for that model. Section V concludes the paper.

II. EXAMPLES OF CROSS-DOMAIN NETWORK SLICING

Fig. 2 shows five services that are provided on the same infrastructure. The resources in the access network part, such as bandwidth, computing and storage, are represented with 6 AN NSSIs, whereas the resources in the core network part are represented with 6 CN NSSIs. AN and CN NSSIs can be associated with one or multiple NSI(s).

TABLE 1. A rectangular scheme, with services as rows, AN NSSIs as columns, and CN NSSIs as table entries, representing the E2E slicing described in Fig. 2.

		AN NSSIs					
		a_1	a_2	a_3	a_4	a_5	a_6
SERVICES	s_1	c_1					
	s_2		c_2	c_3			
	s_3			c_3	c_4		
	s_4				c_6	c_5	
	s_5						c_4

Let us denote the set of 5 services by $S = \{s_1, \dots, s_5\}$, the set of 6 AN NSSIs by $A = \{a_1, \dots, a_6\}$ and the set of 6 CN NSSIs by $C = \{c_1, \dots, c_6\}$. For this concrete example, we can represent the service/NSI/NSSI mapping as a 5×6 rectangular scheme given in Table 1. The services are modeled as rows, and the columns represent the network slice subnet instances of the access network part.

We fill in the rectangular scheme with elements from the set C . For instance, AN NSSI 6 with CN NSSI 4 form an end-to-end slice (NSI 5) for service 5. We model this in

the rectangular scheme by putting c_4 in the row s_5 and the column a_6 . For service 4 there are two scheduled subnet slices in the access network: a_4 is combined with the 6-th core network subnet slice c_6 and a_5 that is combined with the 5-th core network slice c_5 . We model this by placing c_6 in row s_4 and column a_4 , and by placing c_5 in row s_4 and column a_5 .

Note that this configuration is for time slot t . The mapping scheme might change at time slot $t + \Delta t$.

When we apply dedicated resource allocation, then neither the same AN NSSI nor CN NSSI can be scheduled for more than one NSI, i.e., one service. In terms of the rectangular scheme in Table 1 that means that no c_i appears more than once in any column. In other words, a bundle of dedicated resources is allocated.

On the other hand, we can see that we have two c_3 in the 3-rd column a_3 and in the rows s_2 and s_3 . That means that service 2 and service 3 share the 3-rd access slice c_3 . This is a situation when we have shared resources, i.e., soft slicing where the users compete for the resources.

Another way of modeling the network slicing architecture is the rows to represent the core subnet slices, the columns to represent the access subnet slices and services are the entries in the table, as it is presented in Table 2. In the case when we want to have exclusivity, for instance one NSI for low latency and ultra reliable service, then we allocate a specific subnet slice only to one service, i.e., the services are placed in the table exactly only once in each row and column. We will elaborate this later with one theorem.

TABLE 2. A rectangular scheme, with CN NSSIs as rows, AN NSSIs as columns, and services as table entries, representing the E2E slicing described in Fig. 2.

		AN NSSIs					
		a_1	a_2	a_3	a_4	a_5	a_6
CN NSSIs	c_1	s_1					
	c_2		s_2				
	c_3			s_2, s_3			
	c_4				s_3		s_5
	c_5					s_4	
	c_6				s_4		

TABLE 3. A rectangular scheme, with service as rows, CN NSSIs as columns, and AN NSSIs as table entries, representing the E2E slicing described in Fig. 2.

		CN					
		c_1	c_2	c_3	c_4	c_5	c_6
SERVICES	s_1	a_1					
	s_2		a_2	a_3			
	s_3			a_3	a_4		
	s_4					a_5	a_4
	s_5				a_6		

Finally, for a completeness, we present the third rectangular scheme (conjugate to the previous two), with services as rows, CN NSSIs as columns, and AN NSSIs as table entries in Table 3.

III. COMBINATORIAL MODEL OF NETWORK SLICING

We start with some basic definitions about Latin squares and related combinatorial structures.

Definition 5: A Latin square of order n is an $n \times n$ array in which each cell contains a single symbol from a n -set S , such that each symbol occurs exactly once in each row and exactly once in each column.

Definition 6: A $k \times n$ Latin rectangle is an $k \times n$ array (where $k \leq n$) in which each cell contains a single symbol from a n -set S , such that each symbol occurs exactly once in each row and at most once in each column.

Definition 7: A partial Latin square (rectangle) is a square (rectangular) array L with cells that are either empty or contain exactly one symbol such that no symbol occurs more than once in any row or column.

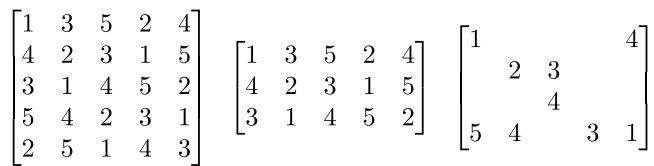


FIGURE 3. A 5×5 Latin square, a 3×5 Latin rectangle and a partial 4×5 Latin rectangle.

In Fig. 3 we show an example of a 5×5 Latin Square, a derived 3×5 Latin rectangle and a derived partial 4×5 Latin rectangle.

Definition 8: Let L be a $n \times n$ Latin square on symbol set E_3 , with rows indexed by the elements of a n -set E_1 and columns indexed by the elements of a n -set E_2 . Let us define a set of triplets $T = \{(x_1, x_2, x_3) : L(x_1, x_2) = x_3\}$. Let $\{a, b, c\} = \{1, 2, 3\}$. The (a, b, c) -conjugate of L , $L_{(a,b,c)}$, has rows indexed by E_a , columns by E_b , and symbols by E_c , and is defined by $L_{(a,b,c)}(x_a, x_b) = x_c$ for each $(x_1, x_2, x_3) \in T$.

Instead of using some general symbol sets E_1, E_2 and E_3 in Definition 8, and in the rest of this paper let us use the set of services $E_1 \equiv S = \{s_1, \dots, s_{n_s}\}$, the set of AN NSSIs $E_2 \equiv A = \{a_1, \dots, a_{n_a}\}$ and the set of CN NSSIs $E_3 \equiv C = \{c_1, \dots, c_{n_c}\}$. In this context, we write (S, A, C) -conjugate instead of $(1, 2, 3)$ -conjugate, (S, C, A) -conjugate instead of $(1, 3, 2)$ -conjugate and (C, A, S) -conjugate instead of $(3, 2, 1)$ -conjugate.

In the light of our introduced mathematical formalism that uses the combinatorial objects of Latin squares and rectangles, instead of the descriptive Definition 1 for hard slicing and its equivalent Definition 2 for dedicated (non-shared) slice subnet instances we offer another definition for hard network slicing in the core and access parts.

Definition 9 (Hard Core Network Slicing): Hard network slicing of C is a set of triplets $T_{hard,C} = \{(s_i, a_j, c_k) : s_i \in S, a_j \in A, c_k \in C\}$, such that for any two triplets $(s_{i_1}, a_{j_1}, c_{k_1}), (s_{i_2}, a_{j_2}, c_{k_2}) \in T_{hard,C}$ it holds:

$$\begin{cases} \text{if } s_{i_1} = s_{i_2} \text{ then } a_{j_1} \neq a_{j_2} \text{ and } c_{k_1} \neq c_{k_2}, \\ \text{if } a_{j_1} = a_{j_2} \text{ then } s_{i_1} \neq s_{i_2} \text{ and } c_{k_1} \neq c_{k_2}. \end{cases} \quad (1)$$

Definition 10 (Hard Access Network Slicing): Hard network slicing of A is a set of triplets $T_{hard,A} = \{(s_i, a_j, c_k) : s_i \in S, a_j \in A, c_k \in C\}$, such that for any two triplets

$(s_{i_1}, a_{j_1}, c_{k_1}), (s_{i_2}, a_{j_2}, c_{k_2}) \in T_{hard,A}$ it holds:

$$\begin{cases} \text{if } s_{i_1} = s_{i_2} \text{ then } a_{j_1} \neq a_{j_2} \text{ and } c_{k_1} \neq c_{k_2}, \\ \text{if } c_{k_1} = c_{k_2} \text{ then } s_{i_1} \neq s_{i_2} \text{ and } a_{j_1} \neq a_{j_2}. \end{cases} \quad (2)$$

Theorem 1: $T_{hard,C} = \{(s_i, a_j, c_k) : s_i \in S, a_j \in A, c_k \in C\}$ is hard network slicing, if and only if there exists a partial (S', A', C') -conjugate Latin rectangle where $S' \subseteq S, A' \subseteq A$ and $C' \subseteq C$.

Proof: If we are given hard network slicing $T_{hard,C}$, then we can build an array L as in Table 1, where the row indexing is by s_i elements in $T_{hard,C}$ that forms a subset $S' \subseteq S$, column indexing is by a_j elements in $T_{hard,C}$ that forms a subset $A' \subseteq A$, and entries by c_k elements in $T_{hard,C}$ that form a subset $C' \subseteq C$. Due to Equation (1) in Definition 9 it follows that the cells in L are either empty or contain exactly one symbol, and no symbol occurs more than once in any row or column. Thus, the array obtained from T_{hard} is a partial Latin rectangle.

Let L be a partial (S, A, C) -conjugate Latin rectangle. Then we can build a set of triplets $T_{hard,C} = \{(s_i, a_j, c_k) : s_i \in S, a_j \in A, c_k \in C\}$, from the non-blank cells in L such that Equation (1) holds. ■

Definition 9, Definition 10 and Theorem 1 address the modeling of the hard core slicing with the (S, A, C) -conjugate. However, in practice we have network slices with components that are of mixed nature: sometimes a network slice has both core network and access network components as hard components, but sometimes one or both of those components are shared. That situation is best modeled with the (C, A, S) -conjugate rectangles, as shown in the next Theorem.

Theorem 2: Let all network slices are represented as a set of triplets $T = \{(c_i, a_j, s_k) : c_i \in C, a_j \in A, s_k \in S\}$, where $i \in \{1, \dots, n_c\}, j \in \{1, \dots, n_a\}$ and $k \in \{1, \dots, n_s\}$. Then, there is a rectangular array $\mathcal{R}_{n_c \times n_a}$ of type (C, A, S) and size $n_c \times n_a$ and there are values $1 \leq n_1 \leq n_c$ and $1 \leq n_2 \leq n_a$ such that the array is partitioned in four rectangular sub-arrays

$$\mathcal{R}_{n_c \times n_a} = \begin{array}{|c|c|c|} \hline & & A \\ \hline C & \mathcal{R}_{1,1} & \mathcal{R}_{1,2} \\ \hline & \mathcal{R}_{2,1} & \mathcal{R}_{2,2} \\ \hline \end{array} \quad (3)$$

where $\mathcal{R}_{1,1} \equiv \mathcal{R}_{n_1 \times n_2}, \mathcal{R}_{1,2} \equiv \mathcal{R}_{n_1 \times (n_a - n_2)}, \mathcal{R}_{2,1} \equiv \mathcal{R}_{(n_c - n_1) \times n_2}, \mathcal{R}_{2,2} \equiv \mathcal{R}_{(n_c - n_1) \times (n_a - n_2)}$, and the following holds:

- 1) every row and every column in $\mathcal{R}_{1,1}$ have at most one non-empty cell;
- 2) every row in $\mathcal{R}_{1,2}$ has at most one non-empty cell, but its columns can have none, one or several non-empty cells;
- 3) every column in $\mathcal{R}_{2,1}$ has at most one non-empty cell, but its rows can have none, one or several non-empty cells;
- 4) every column and every row in $\mathcal{R}_{2,2}$ can have none, one or several non-empty cells.

TABLE 4. A rectangular scheme equivalent to Table 2.

	a_1	a_2	a_5	a_3	a_4	a_6
c_1	s_1					
c_2		s_2				
c_5			s_4			
c_3				s_2, s_3		
c_4					s_3	s_5
c_6					s_4	

Proof: Let us reorder the elements of C as follows: $C_{hard} = \{c_1, \dots, c_{n_1}\}$ are components from the core network part that can be used only as dedicated, i.e., hard slicing, $C_{soft} = \{c_{n_1+1}, \dots, c_{n_c}\}$ are components that can be shared among NSIs. Then it is clear that $C = C_{hard} \cup C_{soft}$ is represented as a disjunctive union of dedicated and shared core network components. Let us apply the same reordering for the components in the access part, i.e., let us represent $A = A_{hard} \cup A_{soft}$ where $A_{hard} = \{a_1, \dots, a_{n_2}\}$ and $A_{soft} = \{a_{n_2+1}, \dots, a_{n_a}\}$. With this reordering for every slice $(c_i, a_j, s_k) \in T$ it holds:

$$\begin{cases} s_k \in \mathcal{R}_{1,1} \text{ if } 1 \leq i \leq n_1 \text{ and } 1 \leq j \leq n_2, \\ s_k \in \mathcal{R}_{1,2} \text{ if } 1 \leq i \leq n_1 \text{ and } n_2 + 1 \leq j \leq n_a, \\ s_k \in \mathcal{R}_{2,1} \text{ if } n_1 + 1 \leq i \leq n_c \text{ and } 1 \leq j \leq n_2, \\ s_k \in \mathcal{R}_{2,2} \text{ if } n_1 + 1 \leq i \leq n_c \text{ and } n_2 + 1 \leq j \leq n_a. \end{cases}$$

Thus, for $s_k \in \mathcal{R}_{1,1}$ we can apply both conditions (1) and (2) from Definitions 9 and 10, and claim 1 from Theorem 2 will follow. To see the validity of the claim 2 for $s_k \in \mathcal{R}_{1,2}$ we need only to apply the condition (1). Similarly, for the validity of the claim 3 and $s_k \in \mathcal{R}_{2,1}$ we need only to apply the condition (2). Then, the correctness of the remaining final claim 4 when $s_k \in \mathcal{R}_{2,2}$ follows. ■

Example 1: Let us represent the network slicing case presented in Fig. 2 and Table 2 as a table following Theorem 2.

Definition 11: We say that a network slice is represented in an *extended* (C, A, S) -conjugate form if it is given as a tuple $(c, a, s, attr_1, \dots, attr_l)$ where $c \in C, a \in A, s \in S$ and $attr_v$ are some additional attributes that are considered as important features of the slice.

IV. SIMULATION OF NSMF WITH SEVERAL OPTIMIZATION OBJECTIVES

Equipped with Theorem 2 and Definition 11 we can implement and simulate any realistic scenario for NSMF. We assume that requests for resources in the AN and CN parts for implementing slices with different requirements arrive according to Poisson distribution with arrival rate λ in each time unit. NSMF checks if the pool of resources can support the creation of the slice. If not, then the request is re-queued for the next time unit. Upon acceptance, the NSMF creates a new NSI and allocates a corresponding resource bundle (NSSI AN and NSSI CN) to the new NSI. We consider dynamic deployment where slices have life time of ν time units distributed with exponential distribution, and the resources allocated to the slices will be released and added back to the resource pool when the slice is deactivated.

By choosing different types of attributes we have opportunity to model different objectives (one or several) of the NSMF such as:

- 1) to maximize the utilization of the network components;
- 2) to decrease the average delay time from slice request to slice activation;
- 3) to decrease the number of rejected slice requests;
- 4) to maximize network operator revenue;
- 5) to maximize the number of slices with high throughput.

In this section we give simulation results of an implementation of NSMF for simple network slicing described with the following attributes:

a high level abstraction of a network slice instance

$$(c, a, s, t_s, t_w) \tag{4}$$

where t_s is the remaining life time of the slice and t_w is the time passed from the slice request until the slice was activated. By default $t_w = 1$ when the request is issued. A full description of all components necessary for implementation of the NSFM is given in Table 5.

Note: With the attribute list described in expression (4) we work with a NSMF model where all hard resources and all soft resources from the core network and the access network are picked from a pool of resources. The NSMF in this model has a higher level of abstraction and it does not take into account the specific capacity of the requested resources. Still, as we will show further in this work, even with this very abstracted model, we can infer important conclusions about the functionality of the network slicing concept and NSMF. Nevertheless, our combinatorial model of network slicing can describe more detailed variants of NSMF. For example,

a network slice with quantitative resources

$$(c, a, s, t_s, t_w, r_c, r_a) \tag{5}$$

where t_s is the remaining life time of the slice, t_w is the time passed from the slice request until the time the slice was activated, r_c is the quantitative value requested from the core network and r_a is the quantitative value requested from the access network.

We now give the algorithm that simulates the work of NSMF with network slices described with the expression (4) and a scenario where rejected requests are added in the waiting queue to be considered for scheduling in the next time unit. Those rejected requests will compete for the network resources with the newly arrived requests.

In Algorithm 1 we use several sub-functions that we comment here. In Step 5 the variable N_{req} gets a random value according to a Poisson distribution with a parameter λ . In Step 6 the function $GetRequests[N_{req}, \mu, p_c, p_a]$ returns a set of initial requests Req , according to the parameters N_{req}, μ, p_c, p_a as they are described in Table 5.

In Step 8 there is a call to a procedure that rearranges the active list of requests $Req \leftarrow HeuristicRearrangement[Req]$. That rearrangement can return just the original list of requests if we do not have developed any optimization strategy,

TABLE 5. A list of all components used by NSMF for network slices given in a form of an extended (C, A, S) -conjugate as described by the attributes in expression (4).

Notation	Meaning	Relations/functions	Comment
(c, a, s, t_s, t_w)	Network slice	$c \in C, a \in A, s \in S,$ t_s - remaining life time of the slice, t_w - waiting time before slice was activated.	Initial value of t_s is a random variable with exponential distribution and average value of μ time units. In the simulation, t_s is decreased by 1 in every time unit.
μ	An average life time of a network slice expressed in number of time units and modeled with an exponential distribution	$P(t_s \leq x) = \begin{cases} 1 - e^{-\frac{x}{\mu}} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$ is the probability that the value of t_s is less or equal to some value x i.e. its cumulative distribution function.	Expected value of t_s is $E[t_s] = \mu$.
C_{hard}	Set of hard slice core network components	$C_{hard} = \{c_1, \dots, c_{n_1}\}, C_{hard} = n_1$	An important parameter for the set C_{hard} is the number of its elements n_1 .
C_{soft}	Set of shared core network components	$C_{soft} = \{c_{n_1+1}, \dots, c_{n_c}\}, C_{soft} = n_c - n_1$	If we denote the total number of core network components with n_c then the number of elements in C_{soft} is given as $n_c - n_1$.
C	Set of all core network components	$C = C_{hard} \cup C_{soft}, C = n_c$	Total number of core network components is n_c .
A_{hard}	Set of hard slice access network components	$A_{hard} = \{a_1, \dots, a_{n_2}\}, A_{hard} = n_2$	An important parameter for the set A_{hard} is the number of its elements n_2 .
A_{soft}	Set of shared access network components	$A_{soft} = \{a_{n_2+1}, \dots, a_{n_a}\}, A_{soft} = n_a - n_2$	If we denote the total number of access network components with n_a then the number of elements in A_{soft} is given as $n_a - n_2$.
A	Set of all access network components	$A = A_{hard} \cup A_{soft}, A = n_a$	Total number of access network components is n_a .
S	Set of all established network slices.	$S = \{s_1, \dots, s_{n_s}\}, S = n_s$	Number of active network slices in one particular moment t is n_s . Notice, that in the next time moment $t + 1$ the number n_s might change.
$req = (c, a, s, t_s, 1)$	Initial request for a network slice	$c \in C, a \in A, s \in S$	If NSMF decides that there are no resources for this request, t_w is increased by 1, and the request is put back to the waiting queue for the next time unit.
p_c	For a requested slice $req = (c, a, s, t_s, 0)$ the probability that $c \in C_{hard}$	$P(c \in C_{hard}) = p_c$ $P(c \in C_{soft}) = 1 - p_c$	Since dedicated resources are more expensive to use, the probability for requests that will ask for dedicated core network components is usually $p_c < 0.5$.
p_a	For a requested slice $req = (c, a, s, t_s, 0)$ the probability that $a \in A_{hard}$	$P(a \in A_{hard}) = p_a$ $P(a \in A_{soft}) = 1 - p_a$	The probability for requests that will ask for dedicated access network components is usually $p_a < 0.5$.
N_{req}	Number of received requests for network slice in certain time moment t	$Req = \{req_1, \dots, req_{N_{req}}\}, Req = N_{req}$	N_{req} is a random variable with Poisson distribution and average value of λ .
λ	An average number of received requests for network slices in certain time moment t , modeled with a Poisson distribution	$P(N_{req} = k) = \frac{e^{-\lambda} \lambda^k}{k!}$.	Expected value of N_{req} is $E[N_{req}] = \lambda$.

or it can perform some heuristics in order to achieve better results with the next subroutine $Dispatch[req, C, A]$ called in Step 10. Based on the rearrangement described in Theorem 2 we have developed one very simple but effective heuristics described in Algorithm 2. The idea can be briefly described as: give priorities to requests that belong to the rectangles $\mathcal{R}_{1,1}$, then $\mathcal{R}_{1,2}$ and $\mathcal{R}_{2,1}$ and finally to the rectangle $\mathcal{R}_{2,2}$. Within the subsets of requests in these rectangles, give priority to the requests that will finish sooner rather than later (that is sorting in ascending order in Steps 16 - 19).

If $Dispatch[]$ subroutine returns that there are resources both in the core network and in the access network, then

that request is activated by adding it in Step 12 to the list of active slices, and the list of core network and access network resources is updated in Steps 13 and 14. If $Dispatch[]$ subroutine returns that there are no available resources then the waiting time for the request is increased by one, and the rejected request is added to the set of rejected requests.

Steps 20 to 30 update the state of the active slices by reducing by 1 all their t_s values. If the slice has a value t_s that is still positive, it will continue to be active for the next time unit. Otherwise, the slice is deactivated and its resources are released and are added back in the pool of available resources (Steps 26 and 27).

Algorithm 1 Simulation of NSMF With Dynamic Deployment and Re-queuing of Rejected Requests

```

1: ActiveSlices ← ∅
2: RejReq ← ∅
3: ns ← 0
4: for t = 1 to TimeSimulation do
5:   Nreq ← Poisson(λ)
6:   Req ← GetRequests[Nreq, μ, pc, pa] ∪ RejReq
7:   RejReq ← ∅
8:   Req ← HeuristicRearrangement[Req]
9:   for req = (c, a, s, ts, tw) ∈ Req do
10:    (FoundC, FoundA) ← Dispatch[req, C, A]
11:    if FoundC AND FoundA then
12:      ActiveSlices ← ActiveSlices ∪ {req}
13:      C ← C \ {req.c}
14:      A ← A \ {req.a}
15:    else
16:      req.tw ← req.tw + 1
17:      RejReq ← RejReq ∪ {req}
18:    end if
19:  end for
20:  NewActive ← ∅
21:  for req = (c, a, s, ts, tw) ∈ ActiveSlices do
22:    req.ts ← req.ts - 1
23:    if req.ts > 0 then
24:      NewActive ← NewActive ∪ {req}
25:    else
26:      C ← C ∪ {req.c}
27:      A ← A ∪ {req.a}
28:    end if
29:  end for
30:  ActiveSlices ← NewActive
31: end for

```

Lemma 1: The necessary conditions for Algorithm 1 to reach a stationary ergodic stage are the following:

$$p_c < \frac{n_1}{n_c}, \tag{6}$$

$$p_a < \frac{n_2}{n_a}, \tag{7}$$

$$\mu\lambda < \min\{n_c, n_a\}. \tag{8}$$

Proof: (sketch) The proof is by assuming that any of the given inequalities is not true and by showing in that case Algorithm 1 produces an ever increasing list of requests Req. For example, let us assume that $p_c \geq \frac{n_1}{n_c}$. This means that in average, there will be more requests asking for hard core network components than there are available, thus rejecting those requests, i.e., producing longer and longer requests lists Req.

A similar reasoning is if we suppose that $\mu\lambda \geq \min\{n_c, n_a\}$. This means that there will be a situation when the number of requests times the number of time units necessary to finish the activity of those requests will surpass the minimum number of available resources either in the core

Algorithm 2 HeuristicRearrangement[Req]

```

1: Req1,1 ← ∅, Req1,2 ← ∅, Req2,1 ← ∅, Req2,2 ← ∅
2: for req = (c, a, s, ts, tw) ∈ Req do
3:   if c ∈ Chard AND a ∈ Ahard then
4:     Req1,1 ← Req1,1 ∪ req
5:   end if
6:   if c ∈ Chard AND a ∈ Asoft then
7:     Req1,2 ← Req1,2 ∪ req
8:   end if
9:   if c ∈ Csoft AND a ∈ Ahard then
10:    Req2,1 ← Req2,1 ∪ req
11:   end if
12:   if c ∈ Csoft AND a ∈ Asoft then
13:    Req2,2 ← Req2,2 ∪ req
14:   end if
15: end for
16: Req1,1 ← SortAscending[Req1,1, ts]
17: Req1,2 ← SortAscending[Req1,2, ts]
18: Req2,1 ← SortAscending[Req2,1, ts]
19: Req2,2 ← SortAscending[Req2,2, ts]
20: Req ← Req1,1 || Req1,2 || Req2,1 || Req2,2
21: Return Req

```

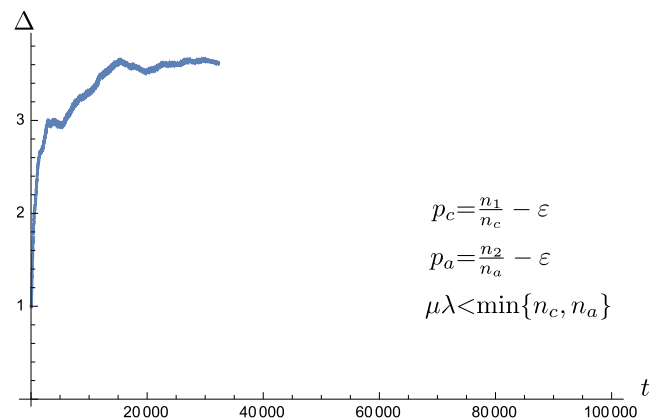


FIGURE 4. An average activation delay simulating the work of NSMF for 100,000 time units. The average is taken over 10 experiments. After a transitioning phase of about 15,000 time units, the process becomes stationary ergodic and the average delay Δ is around 3.5.

part or in the access part. In that case the rejected requests will be added to the queue of requests, thus contributing for ever-increasing length of the list of requests Req. ■

We have an initial implementation of Algorithms 1 and 2 in Mathematica, and next we show several experimental results that confirm the claims of Lemma 1, especially the effects of compliance vs non-compliance with the conditions (6), (7) and (8).

In Fig. 4 and Fig. 5 we give the results of performing 10 simulations with the following parameters: $n_1 = 50, n_c = 350, p_c = 0.99 \frac{n_1}{n_c} = 0.141429, n_2 = 100, n_a = 500, p_a = 0.99 \frac{n_2}{n_a} = 0.198, \lambda = 10, \mu = \lfloor 0.99 \frac{\min\{n_c, n_a\}}{\mu} \rfloor = 34$. The simulation was performed for 100,000 time units. As we can see in Fig. 4, there is a transition period of about 15,000 time

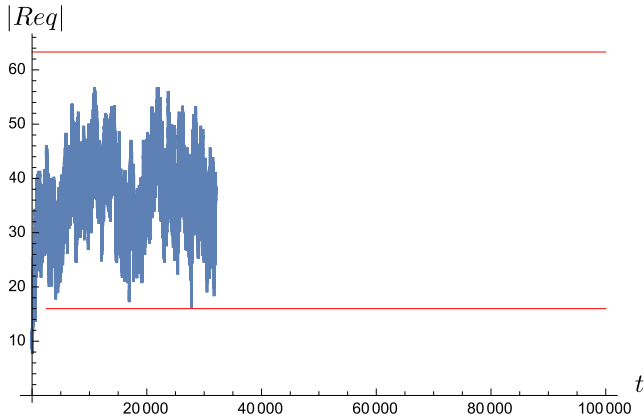


FIGURE 5. An average request queue size simulating the work of NSMF for 100,000 time units. The average is taken over 10 experiments. The size of the requests queue $|Req|$ is stationary ergodic and varies between 16 and 63.

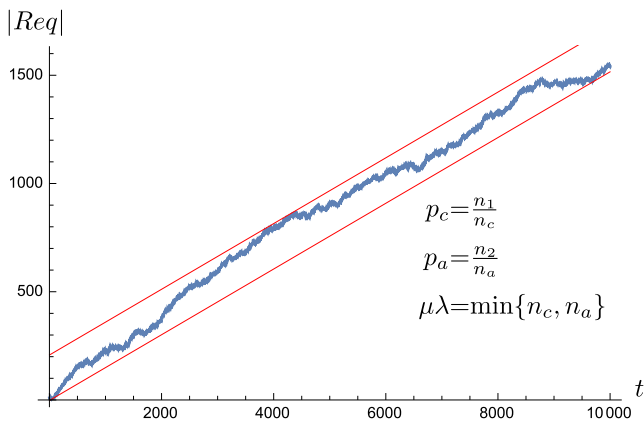


FIGURE 6. An average request queue size simulating the work of NSMF for 10,000 time units. The average is taken over 10 experiments. By having parameters that are upper bounds in Lemma 1, the functioning of the NSMF is not a stable process since the size of the requests queue $|Req|$ is always increasing.

units until the process becomes stationary ergodic with an average delay Δ around 3.5 time units. In Fig. 5 we show the corresponding queue size for the same simulation. The size of the queue $|Req|$ is stationary ergodic and varies between 16 and 63 requests.

In Fig. 6 we show the results of 10 simulations with the values that are the upper bounds in Lemma 1, i.e., $p_c = \frac{n_1}{n_c} = 0.142857$, $p_a = \frac{n_2}{n_a} = 0.2$ and $\mu\lambda = \min\{n_c, n_a\} = 35$. As we can see the size of the requests queue is always increasing as times goes on, indicating that the parameters chosen by the network operator are not sustainable in this model. This simulation analysis indicates also that the network operator should either increase the pool size for the access and core network resources in order to avoid the strict equality or some rejection policy should be introduced.

As mentioned before, we can seek for several optimization objectives within one model of NSMF. Here we give results from simulation of optimized and non-optimized NSMF given by the NSI expression (4), where the optimization is performed in Step 8 of the Algorithm 1, and where the

optimization heuristics is given in Algorithm 2. The optimization objectives are: 1. to maximize the utilization of the network components; and 2. to decrease the average delay time from slice request to slice activation. We argue that by these objectives, indirectly we are achieving also the objective to maximize the network operator revenue.

Definition 12: Let $U(t), t = 1, \dots$, be a function that denotes the number of network slice resources scheduled by the NSMF at time t . An average utilization $V_{[T_1, T_2]}$ of the network components for the NSMF model given by the NSI expression (4), for the time period $[T_1, T_2]$ is defined as

$$V_{[T_1, T_2]} = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2} U(t). \quad (9)$$

Without a proof we state here the following Corollary.

Corollary 1: For NSMF model given by the NSI expression (4), for any time interval $[T_1, T_2]$, $V_{[T_1, T_2]}$ is upper bounded by $\min(m_c, n_a)$, i.e.,

$$V_{[T_1, T_2]} \leq \min(m_c, n_a). \quad (10)$$

Seeking for optimization strategies that will increase the average utilization of the network components is a desired goal, but it is not the most rational optimization objective because it excludes the delay time between the request and the service delivery. Thus, it is much better to set another optimization objective which we define with the next definition.

Definition 13: Let $U(t), t = 1, \dots$, be a function that denotes the number of network slice resources scheduled by the NSMF at time t , and let $\Delta(t), t = 1, \dots$, be a function that denotes the average delay units that network slice requests issued at time t should wait until their activation. An utilization ratio function $W(t)$ is defined as:

$$W(t) = \frac{U(t)}{\Delta(t)}. \quad (11)$$

An average utilization ratio $W_{[T_1, T_2]}$ for the time period $[T_1, T_2]$ is defined as

$$W_{[T_1, T_2]} = \frac{1}{T_2 - T_1} \sum_{t=T_1}^{T_2} W(t). \quad (12)$$

Our objective is to define optimization strategies that maximize $W_{[T_1, T_2]}$ for any time interval $[T_1, T_2]$.

In Fig. 7 we show comparison between two utilization ratio functions where the first one is obtained without any optimization heuristics, i.e., the requests are processed as they come in a first-come-first-serve manner (the blue curve), and the second one is obtained by the Algorithm 2 (the orange curve). For the non-optimized version we get $W_{[4000, 10000]} = 42.248$ that means in every moment the ratio between the number of used resources and the waiting time is 42.248. On the other hand, the optimal strategy gives us $W_{[4000, 10000]} = 98.865$ which is more than double than the non-optimized strategy.

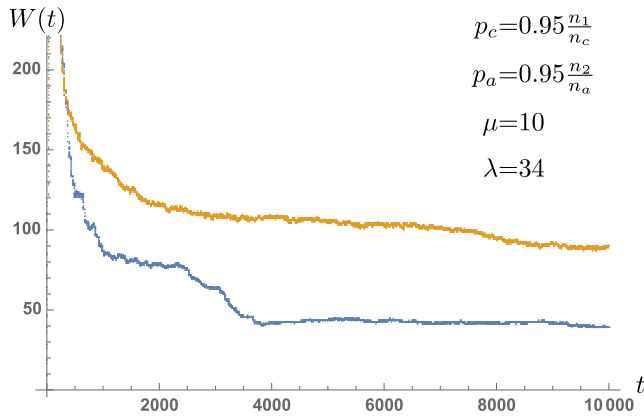


FIGURE 7. Utilization ratio function simulating the work of NSMF for 10,000 time units. The average is taken over 10 experiments. The orange curve is obtained by the optimal strategy in algorithm 2 and the blue curve is for simulation without any optimizations (requests are processed as they arrive).

V. CONCLUSION

We proposed a mathematical model for network slicing based on combinatorial designs such as Latin squares and rectangles and their conjugate forms. These combinatorial designs allow us to model both soft and hard slicing in the access and core parts. Moreover, by the introduction of the extended attribute description our model can offer different levels of abstractions for NSMF that combines cross-domain NSSIs in one end-to-end NSI.

From an optimization point of view, in this work we also introduced the notion of utilization ratio function, with aims to describe the functional dependencies between the number of used network resources and the waiting time for establishing the network slice. Then, we presented two strategies for the work of NSMF, a non-optimized first-come-first-serve strategy and an optimal strategy, where the optimization objectives are: 1. to maximize the utilization of the network components; and 2. to decrease the average delay time from slice request to slice activation. Simulations results presented in this work show that optimal strategy, achieved by maximizing the utilization ratio function, provides more than twice better performances in terms of the both objectives compared to the non-optimized strategy.

REFERENCES

- [1] D. Kim and S. Kim, "Network slicing as enablers for 5G services: State of the art and challenges for mobile industry," in *Telecommunication Systems*. New York, NY, USA: Springer, Oct. 2018, pp. 1–11.
- [2] Ericsson Study, "An economic study of 5G network slicing for IoT service deployment," Ericsson, Stockholm, Sweden, Tech. Rep., 2017.
- [3] N. Alliance, "5G white paper," Next Gener. Mobile Netw., Frankfurt, Germany, White Paper, 2015, pp. 1–125.
- [4] *Study on New Services and Markets Technology Enablers*, document 3GPP TS 22.891 version 1.0.0, Sep. 2015.
- [5] *System Architecture for the 5G System*, document 3GPP TS 23.501 Version 15.2.0 Release 15, 2018.
- [6] NGMN Alliance, "5G white paper," Next Gener. Mobile Netw., Frankfurt, Germany, White Paper, 2015, pp. 1–125.
- [7] Q. Li, G. Wu, A. Papatthanassiou, and U. Mukherjee. (2016). "An end-to-end network slicing framework for 5G wireless communication systems." [Online]. Available: <https://arxiv.org/abs/1608.00572>
- [8] A. Kaloylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Commun. Standards Mag.*, vol. 2, no. 1, pp. 60–65, Mar. 2018.
- [9] "The evolution of network slicing," ABI Res. Intel, White Paper, Nov. 2017.
- [10] L. Geng *et al.*, *Network Slicing Architecture*, document Internet-Draft draft-geng-netslices-architecture-02, Internet Engineering Task Force, 2017.
- [11] *Study on Management and Orchestration of Network Slicing for Next Generation Network*, document 3GPP TR 28.801 version 15.1.0 Release 15, 2018.
- [12] A. Regalado, "Who coined 'cloud computing,'" *Technol. Rev.*, vol. 31, 2011.
- [13] L. Yang, R. Dantu, T. Anderson, and R. Gopal, "Forwarding and control element separation (ForCES) framework," Tech. Rep., 2004.
- [14] ETSI, "'001: Network functions virtualisation (NFV)," Architectural Framework, ETSI, Sophia Antipolis, France, 2013.
- [15] A. Ghodsi, S. Shenker, T. Koponen, A. Singla, B. Raghavan, and J. Wilcox, "Information-centric networking: Seeing the forest for the trees," in *Proc. 10th ACM Workshop Hot Topics Netw.*, Nov. 2011, p. 1.
- [16] W. Guan, X. Wen, L. Wang, Z. Lu, and Y. Shen, "A service-oriented deployment policy of end-to-end network slicing based on complex network theory," *IEEE Access*, vol. 6, pp. 19691–19701, 2018.
- [17] B. Han, J. Lianghai, and H. D. Schotten, "Slice as an evolutionary service: Genetic optimization for inter-slice resource management in 5G networks," *IEEE Access*, vol. 6, p. 33137–33147, 2018.
- [18] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Atlanta, GA, USA, May 2017, pp. 1–9.
- [19] C. J. Colbourn and J. H. Dinitz, *Handbook of Combinatorial Designs* (Discrete Mathematics and Its Applications), 2nd ed. Boca Raton, FL, USA: CRC Press, 2006.
- [20] C. J. Colbourn, J. H. Dinitz, and D. R. Stinson, *Applications of Combinatorial Designs to Communications, Cryptography, and Networking* (London Mathematical Society Lecture Note Series), vol. 187. London, U.K., 1999.



DANILO GLIGOROSKI was born in Skopje, Macedonia. He received the B.S. and M.S. degrees in applied mathematics and the Ph.D. degree in computer science from Ss Cyril and Methodius University in Skopje, in 1997.

From 1997 to 2008, he was an Assistant Professor with the Faculty of Natural Sciences, Skopje University. He has authored over 180 scientific publications. He holds 10 inventions. Since 2008, he has been a Professor of information security and cryptography with the Norwegian University of Science and Technology (NTNU). His main research interests include the applications of various algebraic structures in cryptography, information security, and coding theory.



KATINA KRALEVSKA was born in Skopje, Macedonia, in 1987. She received the B.Sc. degree and the M.Sc. degree in telecommunications from Ss Cyril and Methodius University in Skopje, Macedonia, in 2010 and 2012, respectively, and the Ph.D. degree from the Norwegian University of Science and Technology (NTNU), in 2016.

She was a Postdoctoral Researcher with the Department of Information Security and Communication Technology, NTNU, in 2017, where she became an Associate Professor, in 2018. Since 2019, she has been the Deputy Head of the Department of Information Security and Communication Technology, NTNU. She has authored over 25 scientific publications. She holds over eight inventions. Her research interests include applied erasure coding in networks and distributed storage systems, and 5G and network slicing.

...