Networks Attack Detection on 5G Networks using Data Mining Techniques • Pages 24 - 39 • 1 January 2024

Document type
Book Chapter

Source type
Book

ISBN
978-104001648-0, 978-103253018-5

DOI
10.1201/9781003470281-2

View more

Dynamic deployment and traffic scheduling of UPF in 5G networks

Maaliw, Renato R.[a]; Byeon, Haewon[b]; Kumar, Ravi[c]; Vidyapeeth, D.Y. Patil[d]

Save all to author list

[a] College of Engineering, Southern Luzon State University, Lucban, Quezon, Philippines
[b] Department of Digital Anti-Aging Healthcare, Inje University, Gimhae, South Korea
[c] Department of Electronics and Communication Engineering, Jaypee University of Engineering and Technology, Guna, India
[d] Dr. D. Y. Patil School of Science & Technology, Tathawade, Pune, India

Full text options Export

Abstract

SciVal Topics

Abstract

To achieve dynamic deployment of the user plane in the 5G core network, a multi-stage planning technique based on Benders decomposition is provided for the deployment and traffic scheduling of the User Plane Function (UPF) in an edge network context. This is done so that 5G networks can handle the time-varying data traffic load and provide the low-latency services that users want. User plane data latency, UPF implementation cost, and edge server energy consumption will all be reduced as the influence of deployment options on latency is considered. UPF deployment and traffic management strategies are designed using a tiered approach to planning. The model is decomposed into the primary problem of UPF deployment and various secondary subproblems of traffic scheduling using the Benders decomposition algorithm. We repeatedly address the main problem and its subproblems in order to optimise UPF deployment and traffic scheduling. Results from simulations validate the suggested method's rapid convergence to the correct answer. When compared to the sequential solution approach and a heuristic algorithm based on the Markov Decision Process (MDP), the proposed method yields savings of 10.4 and 5.1 percentage points, respectively. © 2024 selection and editorial matter, Sagar Dhanraj Pande and Aditya Khamparia; individual chapters, the contributors.

SciVal Topics

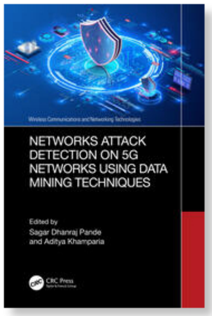
References (22)

View in search results format >

All Export Print E-mail Save to PDF Create bibliography

1 Kang, J., Kang, J., Simeone, O.
On the Trade-Off between Computational Load and Reliability for Network Function Virtualization
(2017) *IEEE Communications Letters*, 21 (8), art. no. 7911307, pp. 1767-1770. Cited 33 times.
doi: 10.1109/LCOMM.2017.2698040
View at Publisher

2 Gil Herrera, J., Botero, J.F.
Tabu Search for Service Function Chain Composition in NFV (Open Access)
(2021) *IEEE Latin America Transactions*, 19 (1), art. no. 9423822, pp. 17-25. Cited 4 times.
http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=9907
doi: 10.1109/TLA.2021.9423822



PREVIEW BOOK

Table of Contents | Book Description

1st Edition

Networks Attack Detection on 5G Networks using Data Mining Techniques

Edited By Sagar Dhanraj Pande, Aditya Khamparia

Hardback £89.99 | eBook £44.09

ISBN 9781032530185
216 Pages 73 B/W Illustrations
Published April 23, 2024 by CRC Press

Description

Artificial intelligence (AI) and its applications have risen to prominence as one of the most active study areas in recent years. In recent years, a rising number of AI applications have been applied in a variety of areas. Agriculture, transportation, medicine, and health are all being transformed by AI technology. The Internet of Things (IoT) market is thriving, having a significant impact on a wide variety of industries and applications, including e-health care, smart cities, smart transportation, and industrial engineering. Recent breakthroughs in artificial intelligence and machine learning techniques have reshaped various aspects of artificial vision, considerably improving the state of the art for artificial vision systems across a broad range of high-level tasks. As a result, several innovations and studies are being conducted to improve the performance and productivity of IoT devices across multiple industries using machine learning and artificial intelligence. Security is a primary consideration when analyzing the next generation communication network due to the rapid advancement of technology. Additionally, data analytics, deep intelligence, deep learning, cloud computing, and intelligent solutions are being employed in medical, agricultural, industrial, and health care systems that are based on the Internet of Things. This book will look at cutting-edge Network Attacks and Security solutions that employ intelligent data processing and Machine Learning (ML) methods.

This book:

- Covers emerging technologies of network attacks and management aspects
- Presents artificial intelligence techniques for networks and resource optimization, and toward network automation, and security
- Showcases recent industrial and technological aspects of next-generation networks
- Illustrates artificial intelligence techniques to mitigate cyber-attacks, authentication, and authorization challenges
- Explains smart, and real-time monitoring services, multimedia, cloud computing, and information processing methodologies in 5G networks
- It is primarily for senior undergraduates, graduate students and academic researchers in the fields of electrical engineering, electronics and communication engineering, computer engineering, and information technology

Table of Contents

Dr.Chaitanya Kulkarni, Dr Zekriia Djabeur Mohamed Seifeddine, Manvitha Gali, Dr. Sheshang Degadwala

Chapter 5: Enhancing Reliability and Security of Power Monitoring Systems in the Era of 5G Networks
Dr. Jitendra Kumar Chaudhary, Dr. Saurabh Kumar, SIVASANKARI S A, Vibhor Kumar Vishnoi

Chapter 6: Passive Voice in 5G Mobile Edge Computing: Optimizing Energy Efficiency and Resource Utilization
Arijeet chandra Sen, Dr Ihtiram Raza Khan, Ajeet Kumar Srivastava, Dr. Sujeet More

Chapter 7: Exchange Matching Algorithm for Low-Complexity Traffic Scheduling for 5G Fronthaul Networks
V. Mahalakshmi, Dr Rajan John, Betty Elezebeth Samuel, Dr. M. Ramkumar Raja

Chapter 8: Attack Path Discovery in Dynamic Network Environments for Automated Penetration Testing over 5G Networks
Dr. Pramod Kumar, Rajiv Sharma

Chapter 9: Enhancing Electric Vehicle Charging Efficiency in Urban Areas with 5G Network Integration and Network Attack Mitigation
Dr. Chintureena Thingom , Arijeet Chandra Sen, Dr. Amit Jain, Dr. Krishan Dutt

Chapter 10: Next-Generation Intrusion Detection System for 5G Networks with Enhanced Security Using Updated Datasets
Komal Shyamsundar Jakotiya, Vishal Shirsath, Raj Gaurav Mishra

Editor(s)

Shipping Options

Book Series

This book is included in the following book series:

- Wireless Communications and Networking Technologies

Related Subjects

Artificial Intelligence | Electrical & Electronic Engineering | Communications & Information Processing | Communications System Design | Digital & Wireless Communication | Telecommunications | Computer & Software Engineering | Power Electronics | Power Engineering | Machine Learning | Computer Science | Engineering & Technology | Electrical Engineering Communications | Systems & Control Engineering | Communications System Design | Digital & Wireless Communications | Power Engineering

Back To Top

Networks Attack Detection on 5G Networks using Data Mining Techniques

Edited by
Sagar Dhanraj Pande
Aditya Khamparia

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Chapter 2

Dynamic deployment and traffic scheduling of UPF in 5G networks

Renato R. Maaliw III
Southern Luzon State University, Lucban, Quezon, Philippines

Haewon Byeon
Inje University, Gimhae, Republic of Korea

Ravi Kumar
Jaypee University of Engineering and Technology, Guna, India

Mukesh Soni
University Centre for Research & Development Chandigarh University, Mohali, India

2.1 INTRODUCTION

The proliferation of mobile data traffic and Internet-connected gadgets in the 5G era has forced a revision of the tried-and-true design of the backbone of wireless networks [1]. The control plane and user plane, as well as the software and hardware, are connected in this design [2]. To overcome these restrictions, new core network architecture has been introduced as part of the 3GPP's 5G technology specifications. By isolating the control plane from the user plane and introducing network function virtualisation and software-defined networking technologies into core network deployment, this design increases network flexibility and scalability [3, 4]. Traditional centralised core network user plane deployment, on the other hand, results in high round-trip latency since it cannot meet the low-latency requirements of 5G services.

Leveraging upcoming technologies like mobile edge computing and fog computing gives a possible answer to these problems. UPF deployment on edge servers nearer users allows for much lower end-to-end latency [5]. Putting in place UPF is a standard example of a challenge while deploying a VNF. Dynamic VNF deployment needs have been largely ignored, while the static deployment of VNFs has received a lot of attention in recent years [6–9]. Joint optimisation techniques for VNF migration costs and network energy usage [10] have been proposed in certain research. These methods, however, ignore the effect that deployment choices have on delay. While other studies have attempted to solve the dynamic VNF deployment problem under

 DOI: 10.1201/9781003470281-2

time-varying traffic conditions by breaking it down into smaller problems [11], the solution accuracy of these approaches is subpar. To deal with fluctuating needs for service function chain resources, deep deterministic policy gradient-based VNF migration algorithms have been developed [12], although training on these algorithms is time-consuming. Additionally, some research has focused on the monetary losses associated with moving VNF instances, and optimisation models for moving VNF instances have been developed, taking energy consumption and reconfiguration costs into account using heuristic algorithms based on the MDP [13].

Energy economy and latency must be considered when deploying UPF because they are two of the most crucial KPIs for 5G networks [14]. UPF network components should be installed as close to the user equipment (UE) as possible to reduce user plane data latency, and UPF network parts should be integrated into as few edge servers as possible to minimise energy usage. The cost of UPF deployment must also be taken into account. Costs related to decisions on deployment made during one time slot may have an impact on those made during the next, and vice versa. The current approaches to time slot optimisation ignore the influence of deployment decisions on delays, which increases the cost of UPF installation.

The overarching goals of this research are to minimise the total energy consumption of edge servers, lessen the expense of UPF deployment, and lessen the impact of operational latency on user plane data. We do this by developing a staged approach for rolling out the user plane UPF and traffic scheduling in the network's core. The model is broken down into its component parts using the Benders decomposition algorithm, with the UPF deployment challenge serving as the major focus. We find the best strategies for UPF deployment and traffic scheduling by iterative, alternating solution of the main problem and subproblems.

Due to the exponential development of mobile data traffic and the rising number of Internet-connected devices, the traditional architecture of wireless network infrastructure must be re-evaluated in the 5G era. The coupling of the control plane and user plane, as well as the software and hardware, limits both flexibility and scalability. As part of the 5G technology specifications, the 3GPP has designed a new central network architecture that incorporates software-defined networking and network function virtualisation. Traditional, centralised deployment of the user plane of the core network results in significant round-trip latency and falls short of the low-latency requirements of 5G services.

This study aids in the design of effective UPF deployment techniques that boost the functionality of 5G networks by taking into account energy usage, latency, and deployment costs. To overcome the difficulties of real-time UPF deployment, a new multi-stage planning model is proposed, which employs the Benders decomposition technique. The results of this research give important information for network operators and researchers working to improve 5G networks through optimised UPF deployment and traffic scheduling.

This research focuses on utilising cutting-edge technology to improve the deployment of the user plane function (UPF) in 5G networks in order to address the aforementioned difficulties. The criteria for dynamic VNF deployment have largely been ignored, despite the fact that static VNF deployment has received a lot of attention. Decomposition-based methods for dynamic VNF deployment in time-varying traffic conditions do not provide appropriate solution accuracy, and existing optimisation approaches do not take the impact of deployment choices on latency into account. Deep deterministic policy gradient-based VNF migration algorithms nevertheless require time-consuming training procedures despite being effective.

Given these limitations, the goal of this study is to lower the amount of energy that edge servers use, the cost of deploying UPF, and the effects of operational latency on user plane data over the whole network operational cycle. In order to do this, a new multi-stage planning method is made for UPF implementation and traffic scheduling in the core network. The Benders decomposition algorithm is used to break the model into a major UPF deployment problem and a number of traffic scheduling subproblems. By solving the main problem and its subproblems one at a time, you can get to the best UPF deployment and traffic scheduling methods.

The findings of this study aid in the development of efficient UPF deployment methods that improve the performance of 5G networks while taking energy efficiency, latency, and deployment costs into account. The problems with real-time UPF deployment are addressed by the multi-stage planning approach that is being suggested, which makes use of the Benders decomposition method. For network operators and researchers trying to optimise UPF deployment and traffic scheduling to boost the performance of 5G networks, the study's conclusions are extremely insightful.

2.2 UPF DEPLOYMENT AND TRAFFIC SCHEDULING MULTI-STAGE PLANNING MODEL

Figure 2.1 depicts the edge network's deployment of the 5G core network's user plane. To drastically cut down on user plane latency, the UPF is deployed on edge servers, one of which is placed at each base station (assuming the base station distribution is a Poisson point process). The graphic depicts how the energy optimisation technique consolidates traffic onto a small number of edge servers. By routing connections to closer-by edge servers, latency for the user plane is minimised. This research builds a multi-stage planning model for UPF deployment and traffic scheduling, taking into account the energy consumption and user plane delay of edge servers as well as the cost of redeployment and rescheduling under time-varying load conditions.

2.2.1 Edge network environment

Let I be the set of edge servers, J be the set of edge server resource types, R_{ij} represents the capacity of type j resources on edge server i (where $i \in I$ and j

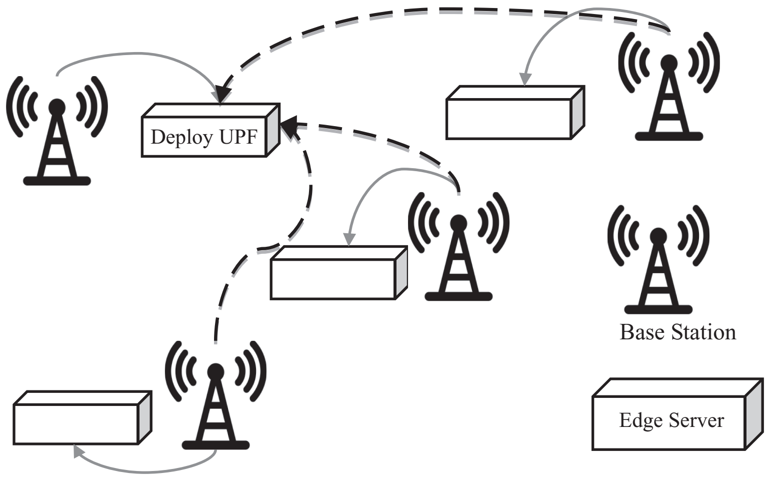


Figure 2.1 Deployment of 5G network.

$\in J$). Let K be the set of UPF virtual network function types, r_{kj} represents the demand for type j resources of UPF instances of type k (where $k \in K$), and u_K^{UPF} represents the processing capacity of UPF instances of type k. Let M be the set of base stations. The data traffic of base stations exhibits a clear day-night pattern and is usually periodic [15]. Let T be the set of time slots for a single period, and Δt be the duration of each time slot. Let u_{mt}^{gNB} represent the data traffic of base station m (where $m \in M$) in time slot t (where $t \in T$).

2.2.2 UPF deployment and traffic scheduling problem

Let $y_{it} \in \{0, 1\}$ determine whether edge server i is enabled in time slot $t \in T$: $y_{it} = 1$ indicates it is enabled, $y_{it} = 0$ indicates it is not enabled. n_{ikt} (where ($n_{ikt} \in N$)) represents the number of UPF instances of type k deployed on edge server i in time slot t. $u_{mit} \geqslant 0$ denotes the data traffic between base station m and edge server i in time slot t. Consider the following three optimisation objectives: edge server energy consumption, UPF deployment cost, and user plane data latency.

2.2.2.1 The edge server energy

The edge server energy consumption adopts a linear power consumption model based on CPU utilisation [16]. The power consumption of edge server i is represented as:

$$P_i = P_i^{idle} + \left(P_i^{peak} - P_i^{idle}\right)u_i^{CPU} \tag{2.1}$$

In the equation: P_i^{idle} and P_i^{peak} represent the idle power consumption and peak power consumption of edge server i, respectively. u_i^{CPU} denotes the CPU utilisation of edge server i. The energy consumption of edge server i in time slot$t$ is represented as:

$$E_t = \sum_{i \in I} y_{it} \left(P_i^{\text{idle}} + \left(P_i^{\text{peak}} - P_i^{\text{idle}} \right) \sum_{k \in K} n_{ikt} r_k^{\text{CPU}} / R_i^{\text{CPU}} \right) \Delta t \tag{2.2}$$

In the equation: R_i^{CPU} represent the CPU capacity of edge server i, and r_k^{CPU} represents the demand of UPF virtual network function of type k for CPU resources.

2.2.2.2 UPF deployment cost

UPF deployment cost includes the energy cost of deploying new UPF instances and the revenue loss caused by deployment delay. Let C_k^{UPF} be the deployment unit cost of UPF instances of type k. Let Δn_{ikt} be the number of newly deployed UPF instances of type k on edge server i in time slot t, represented as:

$$\Delta n_{ikt} = \max\left\{ n_{ikt} - n_{ik(t-1)}, 0 \right\} \tag{2.3}$$

Therefore, the UPF deployment cost in time slot t is represented as:

$$C_t^{\text{dep}} = \sum_{i \in I} \sum_{k \in K} \Delta n_{ikt} C_k^{\text{UPF}} \tag{2.4}$$

2.2.2.3 User plane data latency

User plane latency is mainly composed of four components: processing latency, queuing latency, transmission latency, and propagation latency. Among these, only propagation latency is affected by the geographical location of UPF deployment. Therefore, this study considers only the propagation latency of the user plane. Let δ_{mit} represent the propagation latency between base station m and edge server i. The propagation latency between a base station and an edge server is proportional to the geographical distance between them. The user plane data latency in time slot t is represented as:

$$D_t = \sum_{m \in M} \sum_{i \in I} u_{\text{mit}} \cdot \Delta t \cdot d_{\text{mi}} \tag{2.5}$$

With the objective of minimising the total edge server energy consumption, UPF deployment cost, and user plane data latency over the operating cycle, the UPF deployment and traffic scheduling problem can be formulated as follows:

$$\min \sum_{t \in T} \left(\beta_e \cdot E_t + C_t^{\text{dep}} + \beta_d \cdot D_t \right) \tag{2.6a}$$

$$\&\text{s.t} \sum_{k \in K} n_{ikt} \cdot r_{kj} \leq y_{it} \cdot R_{ij};\ t \in T,\ i \in I,\ j \in J \tag{2.6b}$$

$$\& \sum_{m \in M} u_{\text{mit}} \leq \sum_{k \in K} n_{ikt} \cdot u_k^{\text{UPF}};\ t \in T,\ i \in I \tag{2.6c}$$

$$\& u_{mt}^{\text{gNB}} = \sum_{i \in I} u_{\text{mit}};\ t \in T,\ m \in M \tag{2.6d}$$

$$\Delta n_{ikt} \geq n_{ikt} - n_{ik(t-1)};\ t \in T,\ i \in I,\ k \in K \tag{2.6e}$$

$$\& y_{it} \in \{0,1\};\ t \in T,\ i \in I \tag{2.6f}$$

$$\& n_{ikt} \in \mathrm{N};\ t \in T,\ i \in I,\ k \in K \tag{2.6g}$$

$$\& u_{\text{mit}} \geq 0;\ t \in T,\ m \in M,\ i \in I \tag{2.6h}$$

$$\& \Delta n_{ikt} \geq 0;\ t \in T;\ i \in I;\ k \in K \tag{2.6i}$$

Equation (2.6b) represents the resource capacity constraint of the edge servers, ensuring that the resource allocation for each edge server does not exceed its resource capacity. Equation (2.6c) represents the processing capacity constraint of the edge servers, ensuring that the data traffic scheduled to each edge server does not exceed its processing capacity. Equation (2.6d) represents the flow conservation constraint, indicating that the data traffic from base station m is equal to the sum of the data traffic scheduled to all edge servers. Equations (2.6e) and (2.6i) are derived from the definition of auxiliary variables, where λ and φ are the cost factors for energy consumption and unit data latency, respectively, used to convert energy consumption and data latency into monetary costs. It can be observed that Equation (2.6) represents a mixed-integer linear programming problem.

Theorem 1

The UPF deployment and traffic scheduling problem described in Equation (2.6) is an NP-hard problem.

Proof: We will prove this by constructing a special case of the problem using a reduction method [17] and showing that this special case is a known NP-hard problem.

Let $C_k^{\mathrm{UPF}} = 0$ and $\delta = 0$, which allows us to decompose the problem into time slots.

Let $\beta_d - 0$ be set to such values that the traffic scheduling becomes irrelevant.

Consider a restriction where only a single edge server and a single resource type are considered. Thus, Equation (2.6a–2.6h) transforms into an integer minimum knapsack problem.

Since the integer minimum knapsack problem is known to be NP-hard [18], it follows that Equation (2.6) is also NP-hard.

2.3 TRAFFIC SCHEDULING ALGORITHM BASED ON BENDERS DECOMPOSITION

The Benders decomposition approach is used to resolve the UPF deployment and traffic scheduling problems. Equation (2.6) is divided into several smaller problems for traffic scheduling and one primary problem for UPF deployment in order to effectively reduce the computational complexity of the issue.

Theorem 2

The deployment and scheduling of traffic in the user plane of the 5G core network is a mixed-integer linear programming problem with complicated variables [19].

Proof: From Equation (2.6e), it can be observed that the deployment cost for each time slot depends not only on the deployment decisions for the current time slot but also on the deployment decisions for the previous time slot. The decision variable δ hinders the decomposition of Equation (2.6) into time slots. If the variable δ is fixed to a given value, Equation (2.6) can be decomposed into a series of subproblems for traffic scheduling. Therefore, the variable δ is a complex variable in Equation (2.6).

Equation (2.6) is resolved using the Benders decomposition procedure in accordance with Theorem 2. Iterative algorithms include the Benders decomposition algorithm. To determine the lower and upper bounds of Equation (2.6), the main problem and the subproblem are each solved independently in each iteration. When the lower and higher boundaries are sufficiently near, the process stops.

2.3.1 Traffic scheduling subproblem

By fixing the complex variable δ to a given value, Equation (2.6) can be decomposed into a series of traffic scheduling subproblems based on time slots. In the iteration, the traffic scheduling subproblem for time slot is given by:

$$\min D_t \tag{2.7a}$$

$$\text{s.t.}(6\text{c}).(6\text{d}).(6\text{h}) \tag{2.7b}$$

$$\&n_{ikt} = n^{*}_{iktv}; i \in I, k \in K \tag{2.7c}$$

The optimisation objective is to minimise the user plane data latency for time slot. Equation (2.7b) represents the constraints related to traffic scheduling and it is fixed for the given value which is obtained from solving the main problem in the current iteration. In Equation (2.7), the variable is treated as a continuous variable. It can be observed that the traffic scheduling subproblem is a continuous linear programming problem.

Consider the user plane data latency for time slot obtained from solving the subproblem in Equation (2.7). The total operational cost for the user plane in the core network over the running period is represented as:

$$z_v^{\text{ub}} = \sum_{t \in T} \left(\beta_e \cdot E^{*}_{tv} + C_{tv}^{\text{dep}*} + \beta_d \cdot D^{*}_{tv} \right)$$

The energy use of the edge servers during the time slot determined from solving the primary issue in the current iteration is represented in the equation E^{*}_{tv}. After resolving the key issue in the current iteration, $C_{tv}^{\text{dep}*}$ is the deployment cost of UPF for time slot. An upper limit on Equation (2.6)'s ideal value is given by z_v^{ub}. The Benders optimal cut in the main problem is produced using λ^{*}_{iktv}, the best value of the dual variable connected to Equation (2.7c).

2.3.2 The problem of UPF deployment

In iteration v, the UPF deployment master problem is

$$\min \sum_{t \in T} \left(\beta_e \cdot E_t + C_t^{\text{dep}} + \beta_d \cdot \alpha_t \right) \tag{2.8a}$$

$$\&\text{s.t.Mode}(6\text{b}).(6\text{e}) \sim (6\text{g}).(6\text{i}), t \in T \tag{2.8b}$$

$$\alpha_t \geq D_{tj}^* + \sum_{i \in I} \sum_{k \in K} \lambda_{iktj}^* \left(n_{ikt} - n_{iktj}^* \right) t \in T, j = 1, \cdots, v-1 \tag{2.8c}$$

$$\alpha_t \geq \alpha_t^{\text{down}}; t \in T \tag{2.8d}$$

The objective function in Equation (2.8a) corresponds to Equation (2.6a), which represents the estimated value of the user plane data latency in time slot. Equation (2.8b) represents the constraints related to UPF deployment. Equation (2.8c) represents the Benders optimal cut set generated in previous iterations. Equation (2.8d) sets the lower bound, which accelerates the convergence of the Benders decomposition algorithm. It can be observed that the main problem is an integer linear programming problem.

From Equations (2.6c) and (2.6d), it is known that the variable must satisfy the constraints:

$$\sum_{i \in I} \sum_{k \in K} n_{ikt} u_k^{\text{UPF}} \geq \sum_{m \in M} u_{mt}^{\text{gNB}}, t \in T \tag{2.9}$$

Because Equation (2.9) is not a component of the main issue, obtaining the values of the complex variables generated from the main problem solution may make the subproblem infeasible. In such cases, a feasibility cut is applied to the fundamental issue, increasing the computing cost of the Benders decomposition technique significantly. Equation (2.9) is attached to the core issue to guarantee that the subproblem is always doable. When a feasibility cut is removed from the fundamental issue, the computing cost of the Benders decomposition procedure is greatly decreased.

Let z_v^{lb} be the optimal solution to the main problem at the *v*th iteration, which functions as a lower bound for the optimal solution of Equation (2.6). Let θ represent the optimal solution to the main problem in the *v*th iteration, which will serve as the fixed value for the complex variables in the current iteration of the subproblem.

2.3.3 Benders decomposition algorithm

As shown in Figure 2.2, the specific steps to solve the UPF deployment and traffic scheduling problem using the Benders decomposition algorithm are as follows:

(1) Initialise parameters: Set the upper bound of the objective function of the UPF deployment and traffic scheduling problem, $z^{\text{ub}} = \infty$,, and set the lower bound, $z^{\text{lb}} = -\infty$. Set the Benders optimal cut set in the UPF deployment master problem to empty.

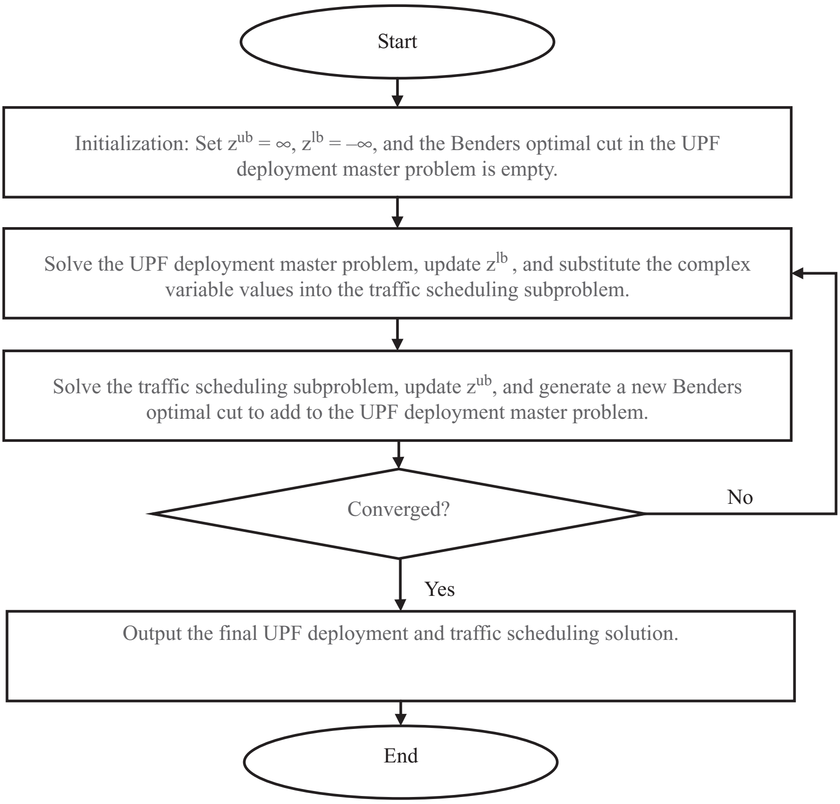


Figure 2.2 Flow chart based on Benders decomposition approach.

(2) Solve the UPF deployment master problem: Obtain z_v^{1b} and n_{iktv}^* as the optimal solution. Update the lower bound of the objective function of the UPF deployment and traffic scheduling problem to $z^{lb} = z_v^{1b}$ and replace n_{iktv}^* in the traffic scheduling subproblem.

(3) Solve the traffic scheduling subproblem: Obtain D_{tv}^* and λ_{iktv}^* as the optimal solution. Update the upper bound of the objective function of the UPF deployment and traffic scheduling problem.

$$z^{ub} = \min\left\{ z^{ub}, z_v^{ub} = \sum_{t \in T} \left(\beta_e \cdot E_{tv}^* + C_{tv}^{dep*} + \beta_d \cdot D_{tv}^* \right) \right\} \tag{2.10}$$

(4) Check if the convergence condition is satisfied. Let ε be the predetermined error tolerance. If $(z^{ub} - z^{lb})/z^{lb} \le \varepsilon$, then terminate the iteration. Otherwise, generate the Benders optimal cut.

$$\alpha_t \geq D_{tv}^* + \sum_{i \in I} \sum_{k \in K} \lambda_{iktv}^* \left(n_{ikt} - n_{iktv}^* \right), t \in T \tag{2.11}$$

And add Formula (2.11) to the main problem of UPF deployment, return to step 2).

2.3.4 Complexity analysis

The proposed algorithm in this research separates the traffic scheduling problem into $|T|$ subproblems, one of which is the UPF deployment master problem. Assuming that each edge server can host a maximum of $n_{\max}$ UPF instances, the UPF deployment master problem is an integer programming problem in which just a single UPF type is taken into account. The UPF deployment master problem has a complexity of $O(2^{|I||T|}(n_{\max} + 1)^{2|I||T|})$. Linear programming problem of complexity $O((|M||I|)^3)$ describes the traffic scheduling subproblem. Each iteration has a complexity of $O(2^{|I||T|}(n_{\max} + 1)^{2|I||T|} + |T|(|M||I|)^3)$ in terms of computation. Where L is the number of iterations needed for the algorithm to converge, the overall computational complexity is written as $O(L(2^{|I||T|}(n_{\max} + 1)^{2|||T|} + |T|(|M||I|)^3))$. Our experiments show that only a few iterations are needed to reach convergence. $O(|T|2^{|I|}(n_{\max} + 1)^{2|I|}(|M||I|)^3)$ is the computational complexity of the direct solution of Equation (2.6). Therefore, the computing complexity of Equation (2.6) is much reduced by the proposed technique. The suggested algorithm is evaluated in a simulation experiment alongside a staged solution approach and an MDP-based heuristic algorithm. The staged solution approach has an $O(|T|2^{|I|}(n_{\max} + 1)^{2|I|}(|M||I|)^3)$ computational complexity. The computational complexity of the MDP-based heuristic algorithm is $O(|T|2^{|I|}(n_{\max} + 1)^{2|I|}(|M||I|)^3)$ for determining the action set via the staged solution method, and $O(L' |T|^3(|M||I|)^3)$, for determining the optimal policy via policy iteration, where L' is the number of iterations required for algorithm convergence. $O(|T|2^{|I|}(n_{\max} + 1)^{2|I|}(|M||I|)^3 + L' |T|^3(|M||I|)^3)$.

2.4 SIMULATION EXPERIMENT

The effectiveness of the UPF deployment and traffic scheduling algorithm in the user plane of 5G core network based on Benders decomposition is verified by simulation experiments.

2.4.1 Experimental setup

In a 4.0 km × 4.0 km area, the positions of base stations are assumed to follow a Poisson point process [20] with a density of 1 station/km^2. Each base station is equipped with one edge server for deploying 5G core network user-plane UPF elements. Each edge server has 48 CPU cores (considering only CPU as the bottleneck resource) with a peak power consumption of 1,000 W. The idle power ratio (the ratio of idle power consumption to peak power consumption) is set to 0.4. Only one type of UPF is considered, which requires 1 CPU core and has a throughput of 0.1 Gb/s. The backhaul link between the base station and the edge server is established using millimetre-wave communication, and the propagation delay is proportional to the geographical distance [21]. The traffic of the base stations exhibits periodic variations with a cycle of 1 day. Therefore, we set $\Delta t = 24$ hours, and we take $\Delta = 1$ hour. The traffic of the base stations is randomly selected from the set {0.5, 1.0, 1.5, 2.0, 2.5} Gb/s with a uniform distribution.

2.4.2 Experimental results

We examine the convergence of the proposed algorithm. Figures 2.3 and 2.4 illustrate the variations of the upper and lower bounds of the UPF deployment and traffic scheduling problem with the number of iterations. In the figure, C represents the total cost, and η denotes the iteration number. The energy price is set to 1 yuan/(kW·h), the cost of unit data latency is 5 Unit/(Gb·ms), the unit price for UPF instance deployment is 0.02 Unit, and the error tolerance is set to 0.01. It can be observed that as the number of iterations increases, the upper and lower bounds gradually converge. The relative error between the upper and lower bounds approaches 0. After 29 iterations, the relative error drops below the error tolerance.

The proposed algorithm was compared to the sequential solution method and the MDP (MDP) heuristic algorithm [13]. The stepwise optimisation approach optimises each time period in sequence without considering the delayed effect of deployment choices. The MDP-based heuristic approach represents the original issue as a MDP and solves it using the policy iteration process, although its accuracy is restricted by the action sets used. The total cost of the three techniques for varied UPF instance deployment costs (cUPF) is shown in Figure 2.4. The suggested method has the lowest overall cost, and its advantage rises as the cUPF becomes larger. as cUPF = 0.06

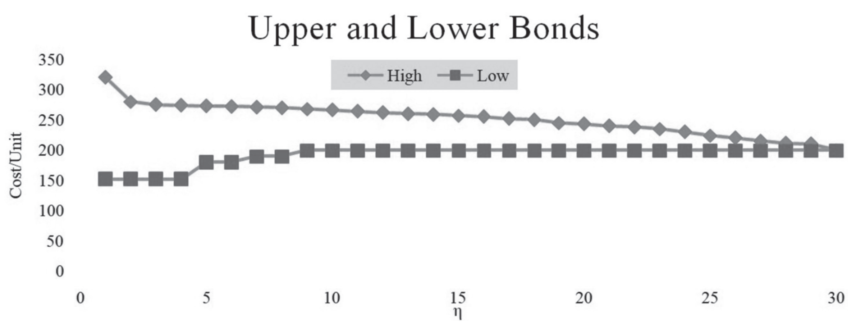


Figure 2.3 Cost estimation using iterations.

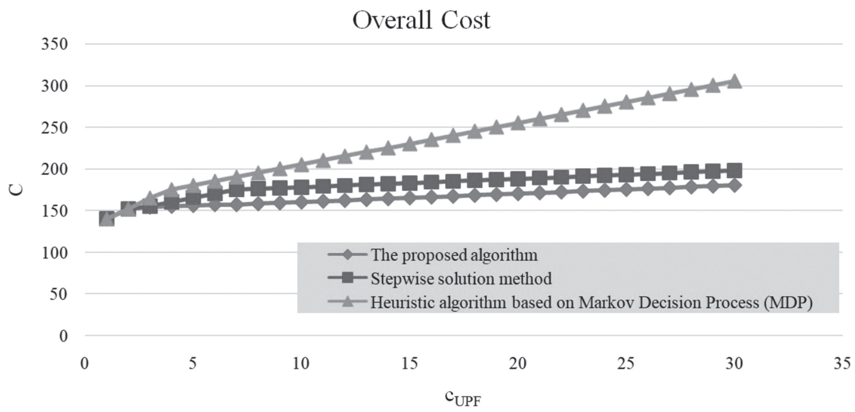


Figure 2.4 Comparative cost analysis (a) average number of servers used, (b) average user plane delay, (c) average number of servers used, (d) average user plane delay.

yuan, the suggested approach cuts overall operating expenses by 10.4% and 5.1%, respectively, as compared to the sequential solution technique and the MDP-based heuristic algorithm.

The effects of the inactive power coefficient and the latency cost on the algorithm's performance were analysed. Figure 2.5(a–d) depicts the average number of edge servers utilised per time interval and the average user-facing latency for various parameter values. Us represents the average number of server requests, while Du represents the average user-facing latency. As the idle power coefficient () increases or decreases, it is observed that the UPF instances are consolidated onto fewer edge servers, resulting in a smaller average number of edge servers used per time slot but a higher average user-facing latency. Inversely, eitherdecreasingor increasing, the UPF instances are deployed to a larger average number of edge servers per time interval, resulting in a smaller average user-facing latency.

2.5 CONCLUSION

This study examined the issue of dynamic UPF (UPF) implementation in peripheral networks for 5G core networks. A multistage planning algorithm for UPF deployment and traffic scheduling using Benders decomposition was developed. We were able to achieve the optimal UPF deployment and traffic scheduling by considering how much energy edge servers consume, how much UPF deployment costs, how long user plane data takes to arrive, and the delayed effects of deployment choices. The Benders decomposition method was used to determine how to solve the UPF deployment and traffic scheduling multi-stage planning model. Simulations demonstrated that the proposed algorithm decreased the amount of energy used by peripheral servers, the cost of deploying UPF, and the time required for user plane

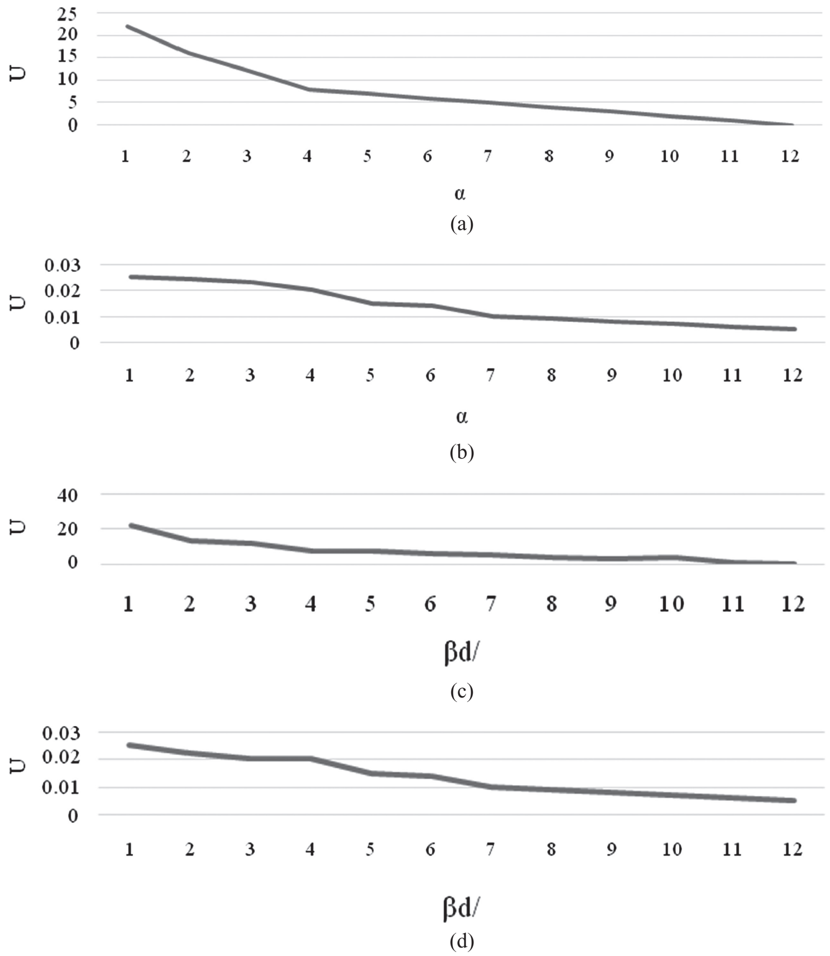


Figure 2.5 Analysis of latency cost. (a) Average number of servers used. (b) Average user plan delay. (c) Average number of servers used. (d) Average user plan delay.

data to arrive. However, it should be noted that the suggested algorithm is still difficult to implement for large problems. Future research will concentrate on determining how to implement user plane functions in large-scale core networks.

REFERENCES

[1] J. Kang, O. Simeone and J. Kang, "On the Trade-Off between Computational Load and Reliability for Network Function Virtualization," in *IEEE Communications Letters*, vol. 21, no. 8, pp. 1767–1770, Aug. 2017, doi: 10.1109/LCOMM.2017.2698040

[2] J. Gil Herrera and J. F. Botero, “Tabu Search for Service Function Chain Composition in NFV,” in *IEEE Latin America Transactions*, vol. 19, no. 01, pp. 17–25, January2021, doi: 10.1109/TLA.2021.9423822
[3] Q. Chen, X. Xu and H. Jiang, “Online Green Communication Scheduling for Sliced Unlicensed Heterogeneous Networks,” in *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10657–10670, Oct. 2021, doi: 10.1109/TVT.2021.3110823
[4] S. Agarwal, F. Malandrino, C. F. Chiasserini and S. De, “VNF Placement and Resource Allocation for the Support of Vertical Services in 5G Networks,” in *IEEE/ACM Transactions on Networking*, vol. 27, no. 1, pp. 433–446, Feb. 2019, doi: 10.1109/TNET.2018.2890631
[5] X. Wang, G. B. Giannakis and A. G. Marques, “A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks,” in *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2410–2431, Dec. 2007, doi: 10.1109/JPROC.2007.907120
[6] K. M. Zuberi and K. G. Shin, “Design and Implementation of Efficient Message Scheduling for Controller Area Network,” in *IEEE Transactions on Computers*, vol. 49, no. 2, pp. 182–188, Feb. 2000, doi: 10.1109/12.833115
[7] R. Casellas, R. Vilalta, R. Martínez and R. Muñoz, “Highly Available SDN Control of Flexi-Grid Networks with Network Function Virtualization-Enabled Replication,” in *Journal of Optical Communications and Networking*, vol. 9, no. 2, pp. A207–A215, Feb. 2017, doi: 10.1364/JOCN.9.00A207
[8] J. Zhanget al., “Aeronautical $Ad~Hoc$ Networking for the Internet-Above-the-Clouds,” in *Proceedings of the IEEE*, vol. 107, no. 5, pp. 868–911, May2019, doi: 10.1109/JPROC.2019.2909694
[9] Z. Wang, J. Zhang and T. Huang, “Determining Delay Bounds for a Chain of Virtual Network Functions Using Network Calculus,” in *IEEE Communications Letters*, vol. 25, no. 8, pp. 2550–2553, Aug. 2021, doi: 10.1109/LCOMM.2021.3065147
[10] C. Bu, X. Wang, M. Huang and K. Li, “SDNFV-Based Dynamic Network Function Deployment: Model and Mechanism,” in *IEEE Communications Letters*, vol. 22, no. 1, pp. 93–96, Jan. 2018, doi: 10.1109/LCOMM.2017.2654443
[11] K. Mebarkia and Z. Zsóka, “Service Traffic Engineering: Avoiding Link Overloads in Service Chains,” in *Journal of Communications and Networks*, vol. 21, no. 1, pp. 69–80, Feb. 2019, doi: 10.1109/JCN.2019.000007
[12] L. Linguaglossaet al., “Survey of Performance Acceleration Techniques for Network Function Virtualization,” in *Proceedings of the IEEE*, vol. 107, no. 4, pp. 746–764, April2019, doi: 10.1109/JPROC.2019.2896848
[13] G. M. Saridiset al., “Lightness: A Function-Virtualizable Software Defined Data Center Network With All-Optical Circuit/Packet Switching,” in *Journal of Lightwave Technology*, vol. 34, no. 7, pp. 1618–1627, 1April1, 2016, doi: 10.1109/JLT.2015.2509476
[14] Al-Shuwaili, O. Simeone, J. Kliewer and P. Popovski, “Coded Network Function Virtualization: Fault Tolerance via In-Network Coding,” in *IEEE Wireless Communications Letters*, vol. 5, no. 6, pp. 644–647, Dec. 2016, doi: 10.1109/LWC.2016.2611510
[15] S. Cherrared, S. Imadali, E. Fabre, G. Gössler and I. G. B. Yahia, “A Survey of Fault Management in Network Virtualization Environments: Challenges and Solutions,” in *IEEE Transactions on Network and Service Management*, vol. 16, no. 4, pp. 1537–1551, Dec. 2019, doi: 10.1109/TNSM.2019.2948420

[16] S. Yang, F. Li, S. Trajanovski, R. Yahyapour and X. Fu, "Recent Advances of Resource Allocation in Network Function Virtualization," in *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 2, pp. 295–314, 1Feb. 2021, doi: 10.1109/TPDS.2020.3017001
[17] Y. Tao, J. Wu, X. Lin and W. Yang, "DRL-Driven Digital Twin Function Virtualization for Adaptive Service Response in 6G Networks," in *IEEE Networking Letters*, vol. 5, no. 2, pp. 125–129, June2023, doi: 10.1109/LNET.2023.3269766
[18] R. Mijumbi, J. Serrat, J.-L. Gorricho, N. Bouten, F. De Turck and R. Boutaba, "Network Function Virtualization: State-of-the-Art and Research Challenges," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 236–262, Firstquarter 2016, doi: 10.1109/COMST.2015.2477041
[19] T. Lin, Z. Zhou, M. Tornatore and B. Mukherjee, "Demand-Aware Network Function Placement," in *Journal of Lightwave Technology*, vol. 34, no. 11, pp. 2590–2600, 1June1, 2016, doi: 10.1109/JLT.2016.2535401
[20] Y. Liu, H. Lu, X. Li and D. Zhao, "An Approach for Service Function Chain Reconfiguration in Network Function Virtualization Architectures," in *IEEE Access*, vol. 7, pp. 147224–147237, 2019, doi: 10.1109/ACCESS.2019.2946648
[21] L. Ge, J. Zhou and Z. Zheng, "Dynamic Hierarchical Caching Resource Allocation for 5G-ICN Slice," in *IEEE Access*, vol. 9, pp. 134972–134983, 2021, doi: 10.1109/ACCESS.2021.3116602
[22] V. Petrovet al., "Dynamic Multi-Connectivity Performance in Ultra-Dense Urban mmWave Deployments," in *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2038–2055, Sept. 2017, doi: 10.1109/JSAC.2017.2720482

Close

"Dynamic Deployment and Traffic Scheduling of UPF in 5G Networks"

Original Submission

Evidence of Review

(Summary of Reviews)

Reviewer Recommendation Term:	Major Revisions

Comments to Author:

Professional Suggestions/Comments on the paper:

1. The introduction outlines the problem of high latency in traditional centralized core network deployment. Clarify and elaborate on this problem statement to set a strong foundation for the rest of the paper.

2. Include a comprehensive literature review that discusses previous works in detail, highlighting the gaps your research aims to fill.

3. Clearly outline the novel contributions of your research and how they differ from existing methods.

4. Provide a more detailed explanation of the methodology, including the steps of the Benders decomposition algorithm, the parameters used, and how the model was tested.

5. Include detailed quantitative results, performance metrics, and comparisons with existing methods to demonstrate the effectiveness of your approach.

6. Include a section that outlines the assumptions and limitations of your study, which can help contextualize the findings and guide future research.

7. Refine the abstract to succinctly capture the essence of the study, and strengthen the conclusion by summarizing key findings, implications, and potential future work.

8. Provide clear definitions and explanations for technical terms and concepts, possibly in a glossary or within the text.

9. If applicable, explain how you introduced noise or anomalies into the data to test the model's robustness. This will help in understanding how the model performs under different scenarios.

10. Add diagrams, flowcharts, and tables to visually represent the core network architecture, the Benders decomposition process, and other key aspects of the methodology.

11. Create a detailed reference list following a standardized citation format and ensure all references are up-to-date and relevant.

Close

FACULTY POSITION RECLASSIFICATION FOR SUCS

(DBM-CHED Joint Circular No. 3, series of 2022)

CERTIFICATION OF PERCENTAGE CONTRIBUTION

(Research Output with Multiple Authors)

Title of Research: Dynamic Deployment and Traffic Scheduling of UPF in 5G Networks
Type of Research Output: Book Chapter (Scopus-Indexed, CRC Press (Taylor & Francis))

Instruction: Supply ALL the names of the authors involved in the publication of Research output and indicate the contribution of each author in percentage. Each author shall sign the conforme column if he/she agrees with the distribution. The conforme should be signed by all the authors in order to be considered. Please prepare separate Certification for each output.

	Name of Authors	Current Affiliations	% Contribution	Conforme *(Sign if you agree with the % distribution)*
1	Renato R. Maaliw III	SLSU	50%	RM3
2	Haewon Byeon	Inje University	20%	Haewon Byeon
3	Ravi Kumar	Jaypee University of Engineering & Technology	15%	Ravi Kumar
4	D.Y. Patil Vidyapeeth	Patil School of Science & Technology	15%	
	** Should have a total of 100%*		100.00%	

Prepared by:

RM3

Renato R. Maaliw III, DIT
(Name and Signature)
Faculty

Certified by:

Nicanor L. Guinto, Ph.D
(Name and Signature)
Director, Office of Research Services