applied sciences

MDPI

Article

Offline Joint Network and Computational Resource Allocation for Energy-Efficient 5G and beyond Networks

Marios Gatzianas [1,2,*], Agapi Mesodiakaki [1,2,*], George Kalfas [1,2], Nikos Pleros [1,2], Francesca Moscatelli [3], Giada Landi [3] and Nicola Ciulli [3] and Leonardo Lossi [3]

[1] Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; gkalfas@csd.auth.gr (G.K.); npleros@csd.auth.gr (N.P.)
[2] Center for Interdisciplinary Research and Innovation, 57001 Thessaloniki, Greece
[3] Nextworks, 56122 Pisa, Italy; f.moscatelli@nextworks.it (F.M.); g.landi@nextworks.it (G.L.); n.ciulli@nextworks.it (N.C.); l.lossi@studenti.unipi.it (L.L.)
[*] Correspondence: mgkatzia@csd.auth.gr (M.G.); amesodia@csd.auth.gr (A.M.)

Abstract: In order to cope with the ever-increasing traffic demands and stringent latency constraints, next generation, i.e., sixth generation (6G) networks, are expected to leverage Network Function Virtualization (NFV) as an enabler for enhanced network flexibility. In such a setup, in addition to the traditional problems of user association and traffic routing, Virtual Network Function (VNF) placement needs to be jointly considered. To that end, in this paper, we focus on the joint network and computational resource allocation, targeting low network power consumption while satisfying the Service Function Chain (SFC), throughput, and delay requirements. Unlike the State-of-the-Art (SoA), we also take into account the Access Network (AN), while formulating the problem as a general Mixed Integer Linear Program (MILP). Due to the high complexity of the proposed optimal solution, we also propose a low-complexity energy-efficient resource allocation algorithm, which was shown to significantly outperform the SoA, by achieving up to 78% of the optimal energy efficiency with up to 742 times lower complexity. Finally, we describe an Orchestration Framework for the automated orchestration of vertical-driven services in Network Slices and describe how it encompasses the proposed algorithm towards optimized provisioning of heterogeneous computation and network resources across multiple network segments.

Keywords: multi-access edge computing; virtual network function; service function chaining; mixed integer linear program; network orchestration

check for updates

Citation: Gatzianas, M.; Mesodiakaki, A.; Kalfas, G.; Pleros, N.; Moscatelli, F.; Landi, G.; Ciulli, N.; Lossi, L. Offline Network and Computational Resource Allocation for Energy-Efficient beyond 5G Networks. *Appl. Sci.* **2021**, *11*, 10547. https://doi.org/10.3390/app112210547

Academic Editor: Alberto Gatto

Received: 9 October 2021
Accepted: 5 November 2021
Published: 9 November 2021

1. Introduction

Beyond 5G (B5G) and 6G networks are envisioned to meet a plethora of service requirements supporting high resource and technology heterogeneity, while adopted architectural paradigms, such as Centralized-Radio Access Network (C-RAN) and Network Function Virtualization (NFV), offer the necessary high flexibility and configurability to the network. The increasing use of NFV, in particular, to "softwarize" functions and applications [1–3] (i.e., decouple the function logic from the underlying hardware running the actual code) in numerous Open System Interconnection (OSI) layers has transformed the majority of "traditional" network functions (e.g., Network Address Translation (NAT), firewall, load balancing, and Intrusion Detection/Prevention System (IDPS)) into Virtual Network Functions (VNFs), which can run on generic computational infrastructure that can be dynamically deployed. Although we use the generic term "Central Processing Unit (CPU) resources" to refer to this computational infrastructure, our model and ensuing analysis can capture any type of computational device, including Graphics Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) and other computational acceleration platforms.

The NFV-supporting physical computational infrastructure contains a number of nodes, including traffic-forwarding switches and computational nodes of different capabilities that can host VNFs in the form of Virtual Machines (VMs), containers, or unikernels. Although traditional cellular architectures were mainly equipped with cloud servers, located at distant locations offering high computational power at low cost, the need for supporting ultra-low latency B5G services has motivated Multi-Access Edge Computing (MEC) [4]. To this end, MEC nodes offer computational capabilities very close to the user, often being attached to Base Stations (BSs), and are able to achieve ultra-low latency at the expense of high cost, thus leading to a nontrivial cost/distance tradeoff that needs to be quantitatively examined.

The above flexible deployment introduces a new exploitable "degree of freedom" regarding the potential location of the deployed VNFs (referred to as the "VNF placement problem" [5,6]) but also introduces new challenges to traditional user association and traffic routing problems, since these problems are strongly coupled with the location of the nodes running the VNFs. Furthermore, the VNFs themselves typically operate in synergy with each other to form Service Function Chains (SFCs), i.e., ordered sequences of VNFs which process packets in an End-to-End (E2E) manner, within the operator's network, and according to specific rules matching a given service request. This ordering adds a crucial constraint to the routing problem, as the SFC is considered to successfully meet the service request **only** when the selected routed path passes through the VNFs comprising the specific SFC in exactly the order specified; for example, if an SFC is described by the ordered sequence NAT→Firewall→NAT, then all packets belonging to this SFC must be processed by only these three VNFs in the specified order.

The above discussion motivates the need to jointly study VNF placement alongside computational and communication resource allocation in a mobile network, while satisfying the SFC, throughput, and delay requirements [7–9]. In addition, B5G is expected to include a variety of different access and transport technologies with distinct characteristics that should be jointly studied. Specifically, apart from fiber links interconnecting the physical nodes, wireless links may also be deployed as an X-haul transport solution, to offer high flexibility close to the Access Network (AN) [10]. Millimeter wave (mmWave) constitutes a very promising candidate to serve this purpose, due to its high bandwidth availability and antenna gains that are able to compensate for the higher path loss in this band. Moreover, for the AN, comprised of gNodeBs (gNBs) densely overlaid with Small Cells (SCs), 5G-New Radio (5G-NR) proposes the use of multiple frequencies (including mmWave). Hence, a holistic network resource planning study should jointly consider: (1) all types of technologies, e.g., 5G-NR, mmWave, fiber, along with their benefits and constraints, and (2) the allocation of all different resource types (i.e., communication, computational, and storage).

It is also crucial to consider the whole network path from the traffic source to the destined User Equipment (UE), so as to satisfy the service latency constraint and offer true E2E optimality, within the operator's network boundaries. Furthermore, achieving high energy efficiency is of utmost importance not only to limit the network operator's operational costs (thus, increasing its revenue) but also to decrease the Information and Communications Technology (ICT) carbon footprint, leading to eco-friendly B5G networks. Hence, we focus on energy-efficient network resource planning solutions to jointly solve the user association, VNF placement (SFC chaining), and traffic routing problem in the highly heterogeneous B5G networks.

In addition, although the discussion so far has focused on (energy-optimal) establishment of the data plane to satisfy the requested services, there is a strong need for an accompanying flexible management and control plane that supports the actual dynamic provisioning and/or configuration of resources in B5G network infrastructures. Network Slicing is considered as a key enabling concept for achieving a high degree of automation in the provisioning of services in B5G networks, allowing, at the same time, resource sharing within the operator's network and the fulfilment of performance requirements depending on the service type (i.e., enhanced Mobile Broadband – eMBB, ultra Reliable Low Latency Communication – uRLLC, and massive Machine Type Communication – mMTC). Network

Slicing improves the way network and computational resources are allocated and also offers the possibility of performing runtime optimization that targets Quality of Service (QoS) preservation as well as energy-efficient deployments and the reduction of operational cost.

Combining Network Slicing, Software Defined Networking (SDN), and NFV/MEC orchestration techniques with advanced strategies for resource planning, VNFs of different types, potentially belonging to different network segments (radio, core, and transport network), can be placed and configured in an optimal manner, while the SFC connectivity is guaranteed through the establishment of optimized jointly computed network paths.

1.1. Related Work

The joint problem of VNF placement, SFC chaining, and routing in wired networks, mostly targeting cloud environments in the network core, has been widely studied with tools such as Dynamic Programming [11], knapsack algorithms [12], Monte Carlo Tree Search [13], and Benders decomposition [14]. The above works, including the recent ones in [7,15,16], formulate NP-hard Mixed Integer Linear (MILP) and Nonlinear Programs, for which heuristic algorithms are proposed and numerically evaluated. Hence, their essential differences lie in the considered constraints and objectives (i.e., minimum total network power consumption in [15], link utilization, overhead, and server power consumption in [16], and monetary profit in [7]). Models with a distinct MEC/cellular flavor appear in [17,18] (see also survey in [19]); however, [17] does not consider power consumption, while [18] models the link constraints more abstractly than in our paper and uses a different objective (i.e., minimize maximum link utilization).

Regarding Network Slicing, SDN, and NFV/MEC orchestration, many platforms, both commercial and open-source, provide functionalities that target the management and control of technology-specific resources. The Open Source Management and Orchestration (MANO) framework proposed by the European Telecommunications Standards Institute (ETSI), considered the most mature standardized solution for Network Service Lifecycle Management (LCM) [20], also supports Network Slicing but is not currently targeting the management of radio elements. On the other hand, O-RAN, being the most successful of the open Radio Access Network (RAN) initiatives, targets specifically the virtualization and management of radio functions [21]. During Horizon 2020 5G Infrastructure Public Private Partnership (5G-PPP) phase 2, some research initiatives focused on prototyping Network Slicing solutions considering (mostly) computational resources [22] and integrating, in some cases, mechanisms for joint management of the transport [23] and radio network segments [24].

1.2. Research Gap

Most of the work on resource allocation (i.e., VNF placement, SFC chaining and routing) mentioned in Section 1.1 does not consider mobile networks and, even when they do, they ignore the wireless AN segment. The last remark motivates our paper, which also studies joint VNF placement and routing in a MEC/cloud-enabled heterogeneous mobile network consisting of macro BSs and SCs. We explicitly include the AN, as well as potential wireless X-haul links, while accounting for the inherent wireless channel fluctuations. Hence, this paper extends the concrete and detailed communication model of [25] by adding all necessary controls for the computational resources and by modeling the associated delay and capacity constraints. In addition, the extensive technical documentation released by Standards Developing Organizations (SDOs) (such as ETSI) regarding the E2E network efficiency of mobile networks in conjunction with the NFV-supporting infrastructures [26,27], indicate that solutions targeting E2E network energy efficiency are of prime importance.

On the other hand, the multiplicity of technology-specific orchestration platforms presents interoperability challenges, which motivates current research in 5G and B5G networks towards delivery of E2E Network Slicing solutions that support the orchestration and management of lheterogeneous resources in distributed edge to cloud infrastructures and across the different network segments [28], i.e., radio, core, and transport. In addition, although there are solutions that optimize the hardware (optimized design of dedicated solutions) or the job allocation in the computing nodes, they provide local optimization benefits, while overlooking the holistic network optimization. To this end, so far, there has been no complete solution for a unified Orchestration Framework having a holistic view of the entire network (potentially by integrating existing technology-specific platforms such as ETSI MANO and O-RAN) and using it for the management and orchestration of Vertical-driven services within E2E Network Slices provisioned and configured over multi-technology components. The search for such a solution provides additional motivation for our paper.

1.3. Our Contributions

The paper's contributions are summarized as follows:

- We formulate a concrete joint user association, traffic routing, and VNF placement optimization problem targeting at the **overall** E2E network performance optimization, with minimal assumptions, that accounts for both communication and computation resources in **all** segments of a mobile network (i.e., AN, edge, and core) and explicitly account for the AN segment, typically ignored in the literature.
- Due to the NP-hardness of the resulting problem, we also propose a heuristic algorithm, evaluate its performance via simulations and demonstrate its superior performance compared with other State-of-the-Art (SoA) algorithms. The proposed solutions can be applied for internode network optimization, in conjunction with approaches targeting intranode optimization for maximum performance.
- Expanding upon our previous work in [29], we also describe the proposed orchestration solution, which, integrated with the proposed algorithm, enables the automated and optimized provisioning and configuration of heterogeneous computational and network resources across **all** network segments, targeting the orchestration of virtualized services according to the expected performance requirements and the specified SFC.

Our formulation and algorithm can also be employed by a mobile network operator as an offline tool, during the network planning stage, to provide quantitative answers on the power expenditure and computational resources (both of them major components of Operational Expenditure (OPEX) and Capital Expenditure (CAPEX)) required to support a given set of services.

The rest of the paper is structured as follows: Section 2 describes the system model and problem formulation, while the proposed heuristic and the employed Orchestration Platform are presented in Sections 3 and 4, respectively. Our performance evaluation methodology and comparison between the optimal solution and other SoA algorithms is described in Section 5 with the actual results being presented in Section 6. Section 7 concludes the paper. Notation-wise, sets are denoted with calligraphic symbols $\mathcal{V}, \mathcal{F}$ etc. and $\triangleq$ denotes equality by definition.

2. System Model and Problem Statement

We consider the RAN and Core segments of a mobile network and model them as a graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (excluding mobile users) and $\mathcal{E}$ is the set of non-access edges/links among them, as illustrated in Figure 1 (which also shows access links, for completeness). The nodes in $\mathcal{V}$ comprise gNB and/or SCs (hereafter referred to as BSs) in the RAN, as well as switches/routers and other middlebox devices (e.g., load balancers, firewalls, etc.) in the Core. These devices typically operate as VNFs running in virtual instances (e.g., VMs, containers, etc.) utilizing computational resources collocated with network nodes.

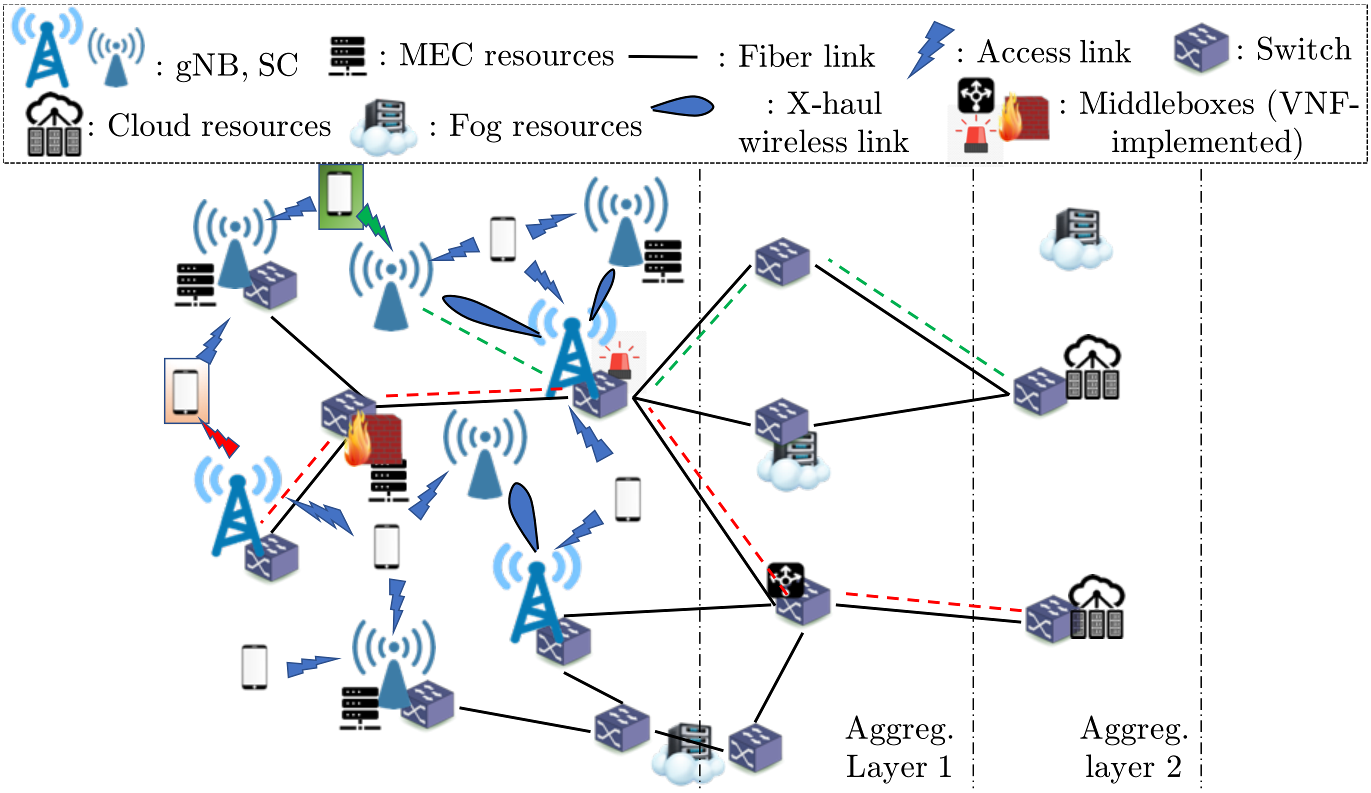

Figure 1. Mobile network of heterogeneous communication and computation nodes along with middlebox functionality offered by deployed VNFs. Dashed color lines indicate 2 illustrative E2E paths selected for 2 highlighted UEs.

Each link $e = (u, v) \in \mathcal{E}$, where $u, v \in \mathcal{V}$, can be wired or wireless (the latter enables wireless X-hauling, typically mmWave), has a communication capacity c_e, and induces a delay δ_e (being the sum of transmission and propagation delays) to all packets traversing it. We partition $\mathcal{E}$ into the sets $\mathcal{E}_{fi}$ and $\mathcal{E}_{wl}$ of wired/fiber and wireless links, respectively, and denote with $\mathcal{V}_{sw} \subseteq \mathcal{V}$ the set of nodes in $\mathcal{V}$ that have at least one incident link in $\mathcal{E}_{fi}$ (i.e., $u \in \mathcal{V}_{sw}$ if there exists $w \in \mathcal{V}$ such that $(u, w) \in \mathcal{E}_{fi}$). We abstractly refer to the nodes in $\mathcal{V}_{sw}$ as "switches" since most fiber links in a mobile Core are typically Point-to-Point links among switches or routers. For modeling reasons, and although all physical links are inherently bidirectional, we explicitly distinguish between links (u, v) and (v, u) in $\mathcal{E}$ (and also for $\mathcal{E}_{fi}, \mathcal{E}_{wl}$). We denote with $h(e) = u$ and $t(e) = v$ the head and tail, respectively, of directed link (u, v).

The UEs are **not** included in $\mathcal{V}$ but rather in $\mathcal{J}$ (we also define $\tilde{\mathcal{V}} \triangleq \mathcal{V} \cup \mathcal{J}$). Each UE $j \in \mathcal{J}$ connects to a BS $a \in \mathcal{A}^{(j)} \subseteq \mathcal{V}$, where the dependence on j captures the typical signal level-based association rules. We define $\mathcal{A} \triangleq \cup_{j \in \mathcal{J}} \mathcal{A}^{(j)}$ as the set of all BSs and $\mathcal{S}^{(a)} \triangleq \left\{ j : a \in \mathcal{A}^{(j)} \right\} \subseteq \mathcal{J}$ as the set of UEs which **may** be served by BS $a \in \mathcal{A}$. We focus on Downlink (DL) and consider the set of DL AN links $\mathcal{E}_{AN} \triangleq \left\{ (a, j) : a \in \mathcal{A},\, j \in \mathcal{S}^{(a)} \right\}$ with $\tilde{\mathcal{E}} \triangleq \mathcal{E} \cup \mathcal{E}_{AN}$. To capture the underlying Physical layer constraints in the air interface, we also assume that each BS $a \in \mathcal{A}$ has a **maximum** number of $\bar{N}_a^{(RB)}$ Resource Blocks (RBs) to allocate to the UEs in $\mathcal{S}^{(a)}$.

Let $\mathcal{V}_c \subseteq \mathcal{V}$ be the set of network nodes which also have *computational* (i.e., CPU) resources able to host one or more VNFs. Additional computational resources such as memory and storage can be similarly handled and are omitted for simplicity and without loss of generality. For each node $y \in \mathcal{V}_c$, we denote with c_y the amount of CPU resources, measured in Giga-Floating Point Operations per Second (GFLOPS). We denote with $\mathcal{F}$ the set of all available VNFs (viewed as **complete software stacks**) that can be deployed and allow for multiple **instances** of a given VNF in the same or different nodes depending on network traffic. Each VNF $f \in \mathcal{F}$ is described by the tuple $f \triangleq (T_f, \pi_f, w_f, \tau_f)$, where T_f is an identifier of the VNF's functionality (e.g., NAT etc.), $\pi_f > 0$ is the data processing capacity of the VNF (in Mbps), $w_f > 0$ is the amount of CPU resources (in GFLOPS) required for the VNF's operation, and $\tau_f > 0$ is the data processing **delay** experienced by an individual data packet as it passes through the VNF. To ensure proper service endpoints, we also introduce the set $\mathcal{F}_{dum} \triangleq \{f_{dum1}, f_{dum2}\}$ of two "dummy" VNFs and include it into $\mathcal{F}$. The VNFs $f \in \mathcal{F}_{dum} \subseteq \mathcal{F}$ are characterized by $\pi_f = \infty$, $w_f = 0$, $\tau_f = 0$, which implies that the "dummy" VNFs work transparently, w.r.t., our model.

Let $\mathcal{C}$ be the set of SFCs, where SFC $\rho \in \mathcal{C}$ is described by the **ordered** sequence $\rho \triangleq \langle f_{dum1}, f_1^{(\rho)}, \dots, f_{N_\rho}^{(\rho)}, f_{dum2} \rangle$, where N_ρ is the number of non-dummy VNFs in ρ and $f_i^{(\rho)} \in \mathcal{F} \setminus \mathcal{F}_{dum}$. The traffic of ρ is **properly served only if** it passes through the VNFs in ρ **exactly matching the specified order**. We also write $f \rightsquigarrow \rho$ to state that VNF f is contained in ρ and define $\mathcal{R}_\rho \triangleq \{f \in \mathcal{F} : f \rightsquigarrow \rho\}$. An equivalent description for ρ is via a **directed** graph $\mathcal{G}^{(\rho)}(\mathcal{V}^{(\rho)}, \mathcal{E}^{(\rho)})$ with the virtual node set $\mathcal{V}^{(\rho)} \triangleq \{f_{dum1}, f_1^{(\rho)}, f_2^{(\rho)}, \dots, f_{N_\rho}^{(\rho)}, f_{dum2}\}$ (i.e., each VNF $f \rightsquigarrow \rho$ is a node of $\mathcal{V}^{(\rho)}$) and the virtual edge set $\mathcal{E}^{(\rho)} = \left\{ \left(f_{dum1}, f_1^{(\rho)}\right), \left(f_i^{(\rho)}, f_{i+1}^{(\rho)}\right)_{1 \le i \le N_\rho - 1} \left(f_{N_\rho}^{(\rho)}, f_{dum2}\right) \right\}$ describing the relative order of the VNFs (see bottom part of Figure 2 for an example of such a virtual graph). For any virtual edge $e' \in \mathcal{E}^{(\rho)}$, we denote with $h(e'), t(e') \in \mathcal{F}$ the respective VNFs at the head and tail of e'.

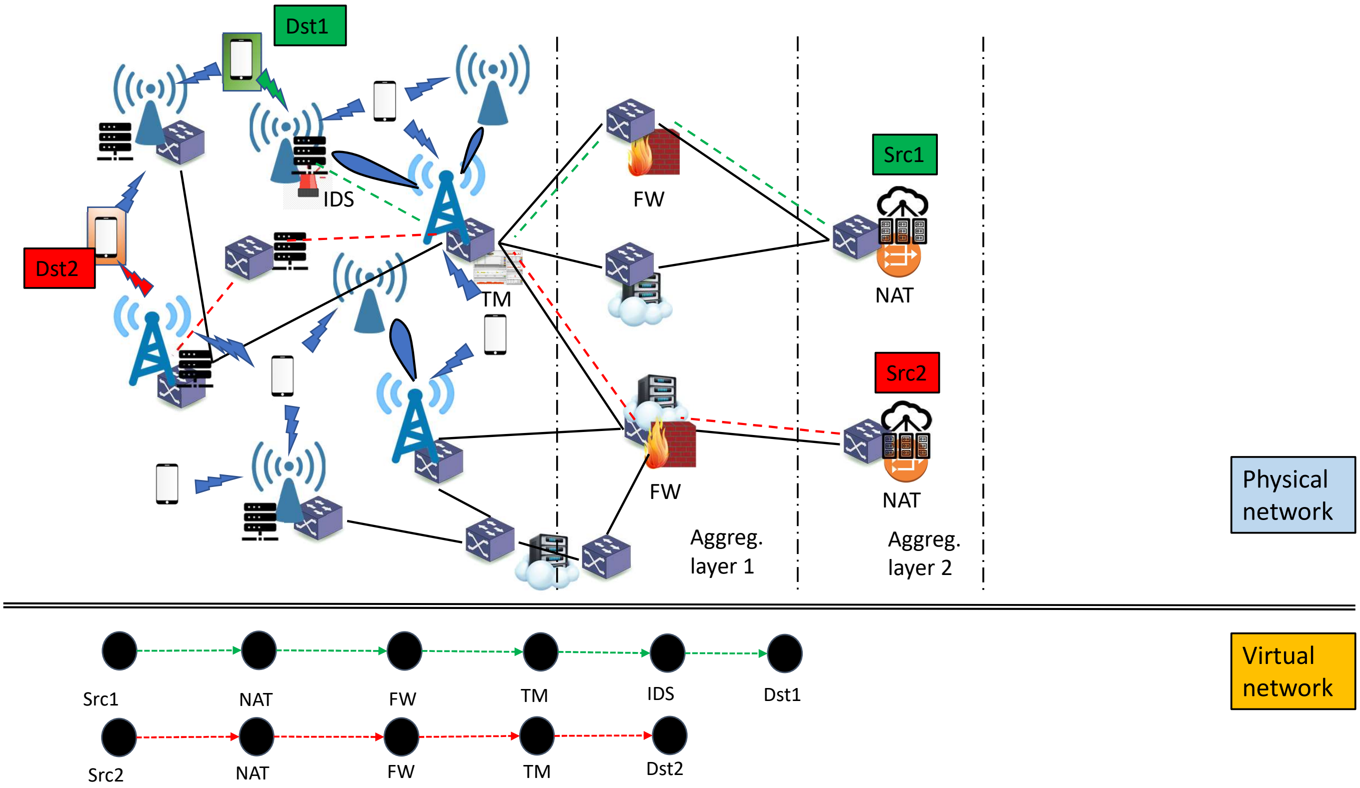

Figure 2. Embedding of virtual network (bottom part) into the physical network (top part). For illustration, 2 distinct SFCs are shown (in green and red) corresponding to 2 different UE requests, and each SFC is described by its own virtual network. $\mathcal{G}^{(\rho_1)}$ (green color) contains 6 virtual nodes and 5 virtual edges, whereas $\mathcal{G}^{(\rho_2)}$ contains 5 virtual nodes and 4 virtual edges. The virtual nodes will be mapped to actual physical nodes, while the virtual edges will be mapped to actual paths in the physical network.

Each UE $j \in \mathcal{J}$ requests a service type $q_j \in \mathcal{Q}$, with $q_j \triangleq (s_{q_j}, r_{q_j}, \delta_{q_j}, \rho_{q_j})$, where $s_{q_j} \in \mathcal{V}$ is the "source" node of the service (the "destination" node of service q_j is UE j), $r_{q_j} > 0$ is the E2E required throughput, $\delta_{q_j} > 0$ is the E2E maximum allowed latency, and $\rho_{q_j} \in \mathcal{C}$ is the required SFC. We explicitly allow for sharing a VNF instance among two (or more) different SFCs, provided the VNF can meet the requirements imposed by the aggregate traffic of the services sharing this VNF. Furthermore, for each UE j, we use our knowledge (or estimates) of the Signal to Interference plus Noise Ratio (SINR) $\sigma_{a,j}$ of link $(a, j) \in \mathcal{E}_{AN}$ and the requested service rate r_{q_j} to compute the number of RBs $N_{a,j}^{(RB)}$ needed to achieve this rate on link (a, j). Assuming frequency-flat slow fading [25], we consider the simple case of uniform BS power allocation among the RBs, so that each RB is assigned a power of $p_a^{(RB)}$.

We can now succinctly state the problem to be solved as follows: *for a set of service requests $\mathcal{Q}$ generated by a set of UEs $\mathcal{J}$, we seek to jointly determine the location of the VNFs that must be deployed to **properly serve** the requested SFCs, as well as the E2E routing path for each request, so that the total system power consumption is minimized (equivalently, the energy efficiency in bits/Joule over all requests is maximized)*. Computation of the selected routing path also includes determination of the BS to which each UE attaches to.

2.1. Power Consumption Model and Problem Formulation

Unless otherwise stated, we consistently use the following indices ranging over the respective sets: $j \in \mathcal{J}$, $a \in \mathcal{A}$, $q \in \mathcal{Q}$, $y \in \mathcal{V}_c$, $\tilde{y} \in \mathcal{V}_c \cup \mathcal{J}$, $f \in \mathcal{F}$, $q \in \mathcal{Q}$, $u, v, w \in \mathcal{V}$, $m, n \in \mathcal{V}_{sw}$. We introduce the decision variables $x_{j,a}$ as the Boolean indicator of whether UE j attaches to BS a, and $\phi_{\tilde{y},f,q}$ as the Boolean indicator of whether VNF f requested by service q is deployed on node $\tilde{y}$. Furthermore, $\theta_e^{e',q}$ (or its alias $\theta_{u,v}^{e',q}$) is the Boolean indicator variable of whether the directed **physical** link $e = (u, v) \in \mathcal{E}$ belongs to the **physical** path in $\mathcal{G}$ onto which the **virtual** edge $e' \in \mathcal{E}^{(\rho_q)}$ is mapped for SFC ρ_q. Finally, $N_{f,\tilde{y}}$ is the number of **instances** of VNF f deployed on node $\tilde{y}$.

Towards superior energy-saving performance, we employ resources, devices, and links **only** when needed. Specifically, collocated CPU resources at node y are employed only when y **actually** runs deployed VNFs, as captured by the Boolean indicator variable ξ_y. Hence, the power consumed by CPU processing at node y is given by

$$P_y^{(CPU)} = P_y^{(CPU,i)} \xi_y + \left(P_y^{(CPU,m)} - P_y^{(CPU,i)} \right) \cdot U_y, \tag{1}$$

where $P_y^{(CPU,m)}$, $P_y^{(CPU,i)}$ are the maximum and idle power of the CPU deployed at y and $U_y \triangleq \sum_{f \in \mathcal{F}} \frac{N_{f,y} w_f}{c_v}$ is the CPU load factor at y [15].

Similarly, for a fiber link $(n, m) \in \mathcal{E}_{fi}$, the Boolean variable $z_{n,m}$ indicates whether the link **actually** carries traffic, in **either** link direction, for any request. To examine whether link $(n, m) \in \mathcal{E}_{fi}$ carries any traffic in the **specific** direction from n to m, we introduce the Boolean variable $w_{n,m}$, which implies that it must hold

$$\begin{aligned} z_{m,n} &\geq w_{m,n} \\ z_{m,n} &\geq w_{n,m} \end{aligned} \quad \forall\, (n, m) \in \mathcal{E}_{fi}. \tag{2}$$

We denote with ψ_n the Boolean variable of whether the switch of node $n \in \mathcal{V}_{sw}$ is **actually** used. There exist certain consistency relations between these variables as shown in (3) below, where $C_1 > 0$ is a sufficiently large constant.

$$\begin{aligned} &\sum_{q \in \mathcal{Q}} \sum_{e' \in \mathcal{E}^{(\rho_q)}} \theta_e^{e',q} \leq C_1 w_e,\ \forall\, e \in \mathcal{E}_{fi}, \\ &\sum_{e \in \mathcal{E}_{fi}: h(e) = n} w_e + \sum_{e \in \mathcal{E}_{fi}: t(e) = n} w_e \leq C_1 \psi_n,\ \forall\, n \in \mathcal{V}_{sw}. \end{aligned} \tag{3}$$

The above relations capture the fact that a node is active (i.e., $\psi_n = 1$) only if it has active incident links carrying traffic and, similarly, a link is active only if it is carrying traffic for one of the requested SFCs.

The total power consumed by the switch in node n is

$$P_n^{(sw)} = P_{idle}^{(sw)} \psi_n + P_{port} \sum_{m \in \mathcal{V}_{sw}:(n,m)\in\mathcal{E}_{fi}} z_{n,m}, \tag{4}$$

where $P_{idle}^{(sw)}$ denotes the switch idle power and the second term accounts for the **active** fiber links of the switch, with P_{port} being the power consumed by each active port [15].

The power expenditure model for the mmWave links in $\mathcal{E}_{wl}$ follows [25] (see Equations (7)–(10) therein); specifically, the power consumed on link $e \in \mathcal{E}_{wl}$ is given by

$$P_e^{(mmW)} = N_{RF}^{(mmW)} \left(\chi_e P_e^{(mmW,i)} + \Delta_e^{(mmW)} F(\ell_e) \right), \tag{5}$$

where $N_{RF}^{(mmW)}$ is the number of Radio Frequency (RF) chains in the link, $P_e^{(mmW,i)}$ is the idle power of the link's transmitter, $\ell_e \triangleq \sum_{q\in\mathcal{Q}} \sum_{e'\in\mathcal{E}^{(\rho_q)}} \theta_e^{e',q} / b_e$ is a load-dependent variable (where b_e is the utilized bandwidth of link e), $\Delta_e^{(mmW)}$ is a slope parameter depending on the power electronics used in the link's transmitter, and $F(\cdot)$ is a **piecewise-linear** function accounting for the nonlinear dependence between achieved throughput and power expenditure. Finally, χ_e is a Boolean indicator variable for whether link e is **actually** used to serve any traffic; if not, the link's transceiver is turned off.

The power expenditure for the AN links in $\mathcal{E}_{AN}$ is similarly modeled as follows: the power expended by a gNB BS $a \in \mathcal{A}$ is

$$P_a^{(gNB)} = N_{RF}^{(gNB)} \cdot \left(\mu_a P_a^{(gNB,i)} + \Delta_a^{(gNB)} \sum_{j\in\mathcal{S}^{(a)}} x_{j,a} p_a^{(RB)} N_{a,j}^{(RB)} \right), \tag{6}$$

where μ_a is the Boolean indicator for whether BS a actually serves any UEs, and $N_{RF}^{(gNB)}$, $P_a^{(gNB,i)}$, and $\Delta_a^{(gNB)}$ have the same semantics as in (5). For an SC BS, it similarly holds

$$P_a^{(SC)} = N_{RF}^{(SC)} \cdot \left(\mu_a P_a^{(SC,i)} + \Delta_a^{(SC)} \sum_{j\in\mathcal{S}^{(a)}} x_{j,a} p_a^{(RB)} N_{a,j}^{(RB)} \right). \tag{7}$$

We use the clever trick of [25] to convert the activation/power saving constraints into a set of linear constraints by introducing auxiliary Boolean variables χ_e, for $e \in \mathcal{E}_{wl}$, and ν_a, for $a \in \mathcal{A}$

$$\begin{aligned} &\chi_e + C_2\sigma_e \geq 1, \ \forall\, e \in \mathcal{E}_{wl}, \\ &1 - C_2\chi_e \leq \sum_{q\in\mathcal{Q}} \sum_{e'\in\mathcal{E}^{(\rho)q}} \theta_e^{e',q} \leq C_2(1-\sigma_e), \ \forall\, e \in \mathcal{E}_{wl}, \\ &1 - C_1\nu_a \leq \sum_{j\in\mathcal{J}} x_{j,a} \leq C_2(1-\nu_a), \ \mu_a + C_1\nu_a \geq 1, \ \forall a \in \mathcal{A}. \end{aligned} \tag{8}$$

In addition to the basic self-consistency conditions

$$\begin{aligned}
&\phi_{s_{q_j},dum1,q_j} = \phi_{j,dum2,q_j} = 1, \quad \forall\, j \in \mathcal{J}, \\
&\phi_{\tilde{y},f_{dum1},q_j} = 0, \quad \forall\, j \in \mathcal{J},\ \forall\, \tilde{y} \in \mathcal{V}_c \setminus \{s_{q_j}\}, \\
&\phi_{\tilde{y},f_{dum2},q_j} = 0, \quad \forall\, j \in \mathcal{J},\ \forall\, \tilde{y} \in \mathcal{V}_c \setminus \{j\},
\end{aligned} \tag{9}$$

and

$$\begin{aligned}
&\phi_{j,f,q} = 0, \quad \forall\, j \in \mathcal{J},\ \forall\, q \in \mathcal{Q},\ \forall\, f \in \mathcal{F} \setminus \{f_{dum2}\}, \\
&\phi_{\tilde{y},f,q} = 0, \quad \forall\, q \in \mathcal{Q},\ \forall\, f \in \mathcal{F} \setminus \mathcal{R}_{\rho_q},\ \forall\, \tilde{y} \in \mathcal{V}_c \cup \mathcal{J}, \\
&\theta_{a,j}^{\dot{e}'_{q_j},q_j} = x_{j,a}, \quad \forall\, j \in \mathcal{J},\ \forall\, a \in \mathcal{A},
\end{aligned} \tag{10}$$

following from the definitions, where $\dot{e}'_{q_j}$ is the "last" virtual edge in $\mathcal{E}^{(\rho_q)}$, i.e., $t\left(\dot{e}'_{q_j}\right) = 1$, we also impose the constraints:

$$\begin{aligned}
&\sum_{y \in \mathcal{V}_c} \phi_{y,f,q} = 1, \quad \forall\, q \in \mathcal{Q},\ \forall\, f \in \mathcal{R}_q \setminus \mathcal{F}_{dum}, \\
&\sum_{a \in \mathcal{A}^{(j)}} x_{j,a} = 1, \quad \forall\, j \in \mathcal{J}, \\
&x_{j,b} = 0, \quad \forall\, b \notin \mathcal{A}^{(j)},
\end{aligned} \tag{11}$$

which require that each requested VNF must be deployed at a node and that each UE must properly attach to exactly one of its allowable BSs (i.e., those BSs which can allocate a sufficient number of Resource Blocks to serve its requested traffic).

The constraint

$$\begin{aligned}
&\sum_{q \in \mathcal{Q}: f \rightsquigarrow \rho_q} \phi_{y,f,q} r_q \le N_{f,y} \pi_f, \quad \forall\, f \in \mathcal{F},\ \forall\, y \in \mathcal{V}_c, \\
&\sum_{f \in \mathcal{F}} N_{f,y} w_f \le c_y, \quad \forall\, y \in \mathcal{V}_c, \\
&\sum_{f \in \mathcal{F} \setminus \mathcal{F}_{dum}} N_{f,y} \le C_1 \xi_y, \quad \forall\, y \in \mathcal{V}_c,
\end{aligned} \tag{12}$$

ensures that the number of deployed VNF instances on a node is sufficient to meet the data processing requirements for the incoming traffic and does not exceed the amount of available CPU resources, while ensuring that the computational node is deployed only when needed.

The link and BS capacity constraints are captured as follows:

$$\begin{aligned}
&\sum_{q \in \mathcal{Q}} \sum_{e' \in \mathcal{E}^{(\rho_q)}} \theta_e^{e',q} r_q \le c_e,\ \forall\, e \in \tilde{\mathcal{E}}, \\
&\sum_{j \in \mathcal{S}^{(a)}} N_{a,j}^{(RB)} \le \bar{N}_a^{(RB)},\ \forall\, a \in \mathcal{A}
\end{aligned} \tag{13}$$

which restricts the total amount of traffic flowing through any link and the total number of Resource Blocks allocated by any BS, while

$$\sum_{v:(u,v)\in\tilde{\mathcal{E}}} \theta_{u,v}^{e',q} - \sum_{w:(w,u)\in\tilde{\mathcal{E}}} \theta_{w,u}^{e',q} = \phi_{u,h(e'),q} - \phi_{u,t(e'),q}, \qquad \forall\, q \in \mathcal{Q},\ \forall\, e' \in \mathcal{E}^{(\rho_q)},\ \forall\, u \in \tilde{\mathcal{V}}, \tag{14}$$

is a flow conservation and routing condition, which ensures that the packets of each requested SFC meet the corresponding VNFs in the correct order as they are routed through the selected path. The above equation essentially ensures the "proper" embedding of the virtual graph $\mathcal{G}^{(\rho_q)}$, for each request $q \in \mathcal{Q}$, into the physical graph (see Figure 2). Finally, E2E delay constraint is captured by

$$\sum_{\substack{y\in\mathcal{V}_c,\\ f\in\mathcal{R}_{q_j}}} \phi_{y,f,q_j}\tau_f + \sum_{\substack{e\in\mathcal{E},\\ e'\in\mathcal{E}^{(\rho_{q_j})}}} \theta_e^{e',q_j}\delta_e + \sum_{a\in\mathcal{A}^{(j)}} x_{j,a}\delta_{(a,j)} \le \delta_{q_j},\ \forall\, j \in \mathcal{J}. \tag{15}$$

Hence, the total power expenditure in the network is

$$P_{total} \triangleq \sum_{n\in\mathcal{S}_{sw}} P_n^{(sw)} + \sum_{y\in\mathcal{V}_c} P_y^{(CPU)} + \sum_{e\in\mathcal{E}_{wl}} P_e^{(mmW)} + \sum_{a\in\mathcal{A}} P_a^{(gNB/SC)} \tag{16}$$

and we formulate our problem as the following NP-hard MILP

$$\begin{aligned} &\text{minimize} \quad P_{total}, \\ &\text{s.t.} \quad (2),\ (3),\ (8)\text{–}(15), \end{aligned} \tag{17}$$

where the control variables in (17) are $N_{f,v}$, ξ_y, $z_{m,n}$, $w_{m,n}$, ψ_n, $\theta_e^{e',q}$, χ_e, σ_e, μ_a, ν_a, $\phi_{\tilde{y},f,q}$, $x_{j,a}$, $N_{a,j}^{(RB)}$ with index semantics as previously described. The MILP property of (17) follows from the simple observation (by visual inspection) that **all** of the above control variables appear as linear terms in the constraints (2), (3), (8)–(15) and as linear (in (1), (4), (6), (7)) or piecewise linear terms (in (5)) in the components of (16) comprising the objective function of (17), combined with the fact that the control variables take only integer or Boolean values. Note that a piecewise linear objective function can be readily converted into a purely linear form by introducing auxiliary variables (see Section 4.3.1 of [30]). Due to the high complexity of solving (17), we next propose a low-complexity heuristic algorithm and use (17) as a yardstick against which the heuristic's performance is evaluated. For the reader's convenience, we have collected all introduced notation into Table 1 at the end of the paper.

Table 1. List of Symbols and Notations.

Symbol	Interpretation	Notation for Element (In Case of Sets)	Symbol	Interpretation	Notation for Element (In Case of Sets)
Input parameters					
$\mathcal{V}$	Set of network nodes (**excluding** mobile users)	u, v	$\mathcal{V}_{sw}$	Set of nodes equipped with switch	u, v
$\mathcal{E}$	Set of network links (**excluding** access links)	$e = (u, v)$	$\mathcal{J}$	Set of UEs	j
$\mathcal{E}_{fi}, \mathcal{E}_{wl}$	Set of fiber (resp. wireless) **non-access** links	(u, v)	$\mathcal{E}_{AN}$	Set of **access** links	(a, j)
c_e	Communication capacity of link e	NA	δ_e	Delay (i.e., transmission + propagation) induced by link e	NA
$\mathcal{A}^{(j)}$	Set of BSs that UE j can connect to	a	$\mathcal{S}^{(a)}$	Set of UEs that can be served by BS a	j
$\mathcal{V}_c$	Set of network nodes equipped with computational resources	y	c_y	Amount of computational resources available at node y	NA
$\mathcal{F}$	Set of available VNFs	$f = (T_f, \pi_f, w_f, \tau_f)$ where T_f: VNF id, π_f: VNF data processing capacity, w_f: amount of CPU resources required by VNF, τ_f: delay induced by VNF processing			
$\mathcal{C}$	Set of available SFCs	$\rho = \langle f_{dum1}, f_1^{(\rho)}, \ldots, f_{N_\rho}^{(\rho)}, f_{dum2} \rangle$			
$\mathcal{R}_\rho$	Set of VNFs comprising SFC ρ	f	$\mathcal{G}^{(\rho)}$	Virtual directed graph describing SFC via virtual node set $\mathcal{V}^{(\rho)}$ and virtual edge set $\mathcal{E}^{(\rho)}$	NA
$\mathcal{Q}$	Set of requests by UEs	$q = (s_q, r_q, \delta_q, \rho_q)$ where s: source node of service request, r_q: E2E requested rate, δ_q: E2E requested latency, ρ_q: requested SFC			
$N_{a,j}^{(RB)}$	Number of RBs needed to be assigned by BS a to UE j to meet its requested rate				
$P_{idle}^{(sw)}$	Switch idle power	NA	$P_e^{(mmW,i)}$	Idle power of transmitter in mmWave link e	NA

Table 1. *Cont.*

Symbol	Interpretation	Notation for Element (In Case of Sets)	Symbol	Interpretation	Notation for Element (In Case of Sets)
Input parameters					
$P_a^{(gNB,i)}$	Idle power of gNB a	NA	$P_a^{(SC,i)}$	idle power of SC a	NA
$N_{RF}^{(mmW)}$	Number of RF chains used in transmitter of mmWave link	NA	$N_{RF}^{(gNB)}$, $N_{RF}^{(SC)}$	Number of RF chains used by gNB, SC	NA
$\Delta_e^{(mmW)}$	Slope parameter of transmitter in mmWave link e	NA	$\Delta_a^{(gNB)}$, $\Delta_a^{(SC)}$	Slope parameter of transmitter in gNB, SC a	NA
Decision variables					
$x_{j,a}$	Boolean indicator of whether UE j actually connects to BS a		$\phi_{\tilde{y},f,q}$	Boolean indicator of whether VNF f requested by service q is deployed on node $\tilde{y}$	
$\theta_e^{e',q}$, $\theta_{u,v}^{e',q}$	Boolean indicator of whether physical link $e = (u, v)$ belongs to the physical path onto which the virtual link $e' \in \mathcal{E}^{(\rho_q)}$ is mapped for SFC rho_q		$N_{f,\tilde{y}}$	Number of instances of VNF f deployed on node $\tilde{y}$	
$z_{n,m}$	Boolean indicator of whether fiber link (n, m) actually carries traffic in **either** of its directions		w_e, $w_{n,m}$	Boolean indicator of whether fiber link $e = (n, m)$ actually carries traffic from n to m	
ξ_y	Boolean indicator of whether the computational resources of node y are used		ψ_n	Boolean indicator of whether the switch at node n is actually used to forward traffic	
μ_a	Boolean indicator for whether BS a serves any UEs		χ_e	Boolean indicator of whether wireless link e is actually used to serve any traffic	
ν_a	Auxiliary variable for μ_a [25]		σ_e	Auxiliary variable for χ_e [25]	

3. Proposed Energy-Efficient Vnf Placement, Traffic Routing, and User Association (Hero)

Our proposed Heuristic for Energy-efficient VNF placement, traffic Routing, and user assOciation (HERO) aims to maximize the network energy efficiency while ensuring low UE blocking probability. HERO consists of two stages, as shown in Figure 3. In the first stage, the traffic path is selected (i.e., user association and routing is performed), while in the second stage VNF placement takes place, ensuring correct VNF ordering.

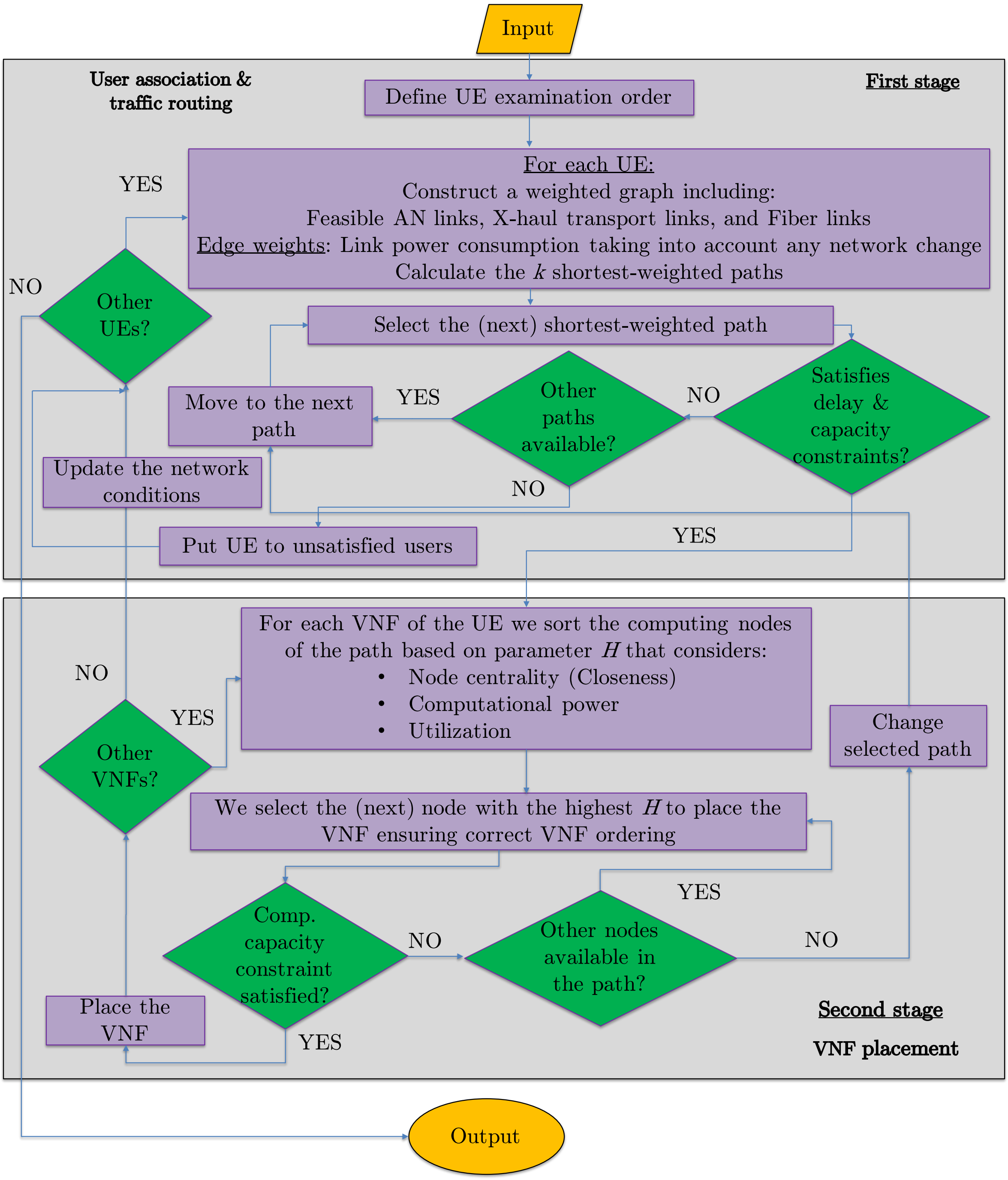
Input
User association & traffic routing
First stage
Define UE examination order
YES
For each UE:
Construct a weighted graph including:
Feasible AN links, X-haul transport links, and Fiber links
Edge weights: Link power consumption taking into account any network change
Calculate the k shortest-weighted paths
NO
Other UEs?
Select the (next) shortest-weighted path
Move to the next path
YES
Other paths available?
NO
Satisfies delay & capacity constraints?
Update the network conditions
NO
Put UE to unsatisfied users
YES
For each VNF of the UE we sort the computing nodes of the path based on parameter H that considers:
Node centrality (Closeness)
Computational power
Utilization
NO
YES
Other VNFs?
Change selected path
We select the (next) node with the highest H to place the VNF ensuring correct VNF ordering
YES
Comp. capacity constraint satisfied?
NO
Other nodes available in the path?
NO
Place the VNF
YES
Second stage
VNF placement
Output

Figure 3. Flowchart of the proposed energy-efficient VNF placement, traffic routing, and user association algorithm (HERO).

Initially, to ensure a high UE acceptance ratio, the UEs are sorted based on their service demands, giving priority to the UEs with the most delay-intolerant services. For UEs with the same delay requirements, priority is given to the UEs with higher rate demands. Then, for each UE, a weighted graph is constructed from the service traffic source to the UE with all feasible links and their respective power consumption acting as weights. HERO calculates the k shortest-weighted paths and starts with the first, as long as it satisfies the delay and link capacity constraints, taking into account the decisions for the already examined users. Otherwise, the next path is selected until either a path that satisfies all constraints is found or there are no other paths. In the latter case, the UE is blocked and HERO proceeds with the next as long as there is one.

After a valid path is found for the current UE, HERO proceeds to the second stage, where the UE VNFs are being placed. To that end, for each VNF, following the order of the UE SFC, a list is constructed with all the available computational nodes based on a parameter, denoted by H. This parameter is equal to the sum of the normalized node centrality (closeness), the normalized node computational capabilities (c_y), and the node CPU utilization. The latter is equal to (a) 1 when the studied VNF can be placed in the examined node without initiating a new VNF instance, (b) 0.1 when there is enough computational capacity to host the studied VNF in the examined node, but a new VNF instance is required, and (c) 0 otherwise. Subsequently, the node with the highest H for the selected VNF is selected, as long as it has sufficient computational resources to host it. Otherwise, the node with the next highest H is selected, until either the VNF is placed or there is no other node to examine in the selected path. In the latter case, the algorithm returns to stage 1 and the next path out of the k calculated is examined. The process is repeated for the new path until either all VNFs of the UE are placed or there is no other path to study and the UE is blocked. In the case where all UE VNFs are placed, the network conditions are updated, and the algorithm proceeds to the next UE. The aforementioned steps are repeated until all UEs are examined.

4. Interaction with Orchestration Framework

The proposed Orchestration Framework, whose high-level architecture is depicted in Figure 4, is based on the Vertical Slicer prototype in [31]. The Orchestration Framework aims at integrating and coordinating technology-specific platforms (e.g., ETSI Open Source MANO, SDN Controllers, O-RAN, etc.) to achieve the automated orchestration of E2E Network Slices, including the joint provisioning and configuration of heterogeneous resources, while considering the full chains of virtual functions associated with both mobile connectivity and application logic. The Vertical Slicer follows a service-driven approach: the high-level requirements provided by the Service Provider (e.g., number of expected UEs, Ultra-High Definition (UHD) streaming type) are dynamically mapped into performance requirements at the infrastructure layer in order to determine the characteristics of the E2E Network Slice in terms of mobile connectivity (e.g., downlink throughput) and the actual composition of the overall SFC (i.e., including network and application virtual functions) along with the dimension of its included components.

In terms of E2E Network Slice modelling, the Vertical Slicer implements a Network Slice Template where each SFC is split into three different Network Slice Subnets which specify, respectively, the deployment characteristics for the RAN, the core, and the Vertical service. The proposed modelling enhances the 3rd Generation Partnership Project (3GPP) Network Slice Network Resource Model [32] by integrating the Vertical services' components. In addition, the Vertical Slicer also builds an overview of the needed connectivity services to be established over the transport network to fulfil the SFC.

In the orchestration and provisioning of the needed resources, the Orchestration Framework encompasses the resource allocation algorithm proposed in this paper. Specifically, when the Orchestration Framework receives a request for instantiating a service, the automated translation mechanism implemented at the Vertical Slicer determines the performance requirements and the SFC to be orchestrated. Then, the Vertical Slicer triggers a Placement Service, which calls the proposed algorithm, and provides as inputs both the service requirements and the SFC. The Placement Service takes decisions about network paths to be established at the data-plane and the placement of the different components in the SFC. The result of the computation is returned to the Vertical Slicer, which proceeds with the orchestration and configuration of the needed resources, in particular, by triggering and coordinating the target technology-specific platforms: NFV/MEC Orchestrators for VNFs' LCM and SDN Controllers for the management and control of the data-plane. Once the service and the corresponding E2E Network Slice are properly instantiated and configured, the Vertical Slicer continues coordinating the overall LCM, interacting as needed with the technology-specific platforms for executing orchestration procedures.

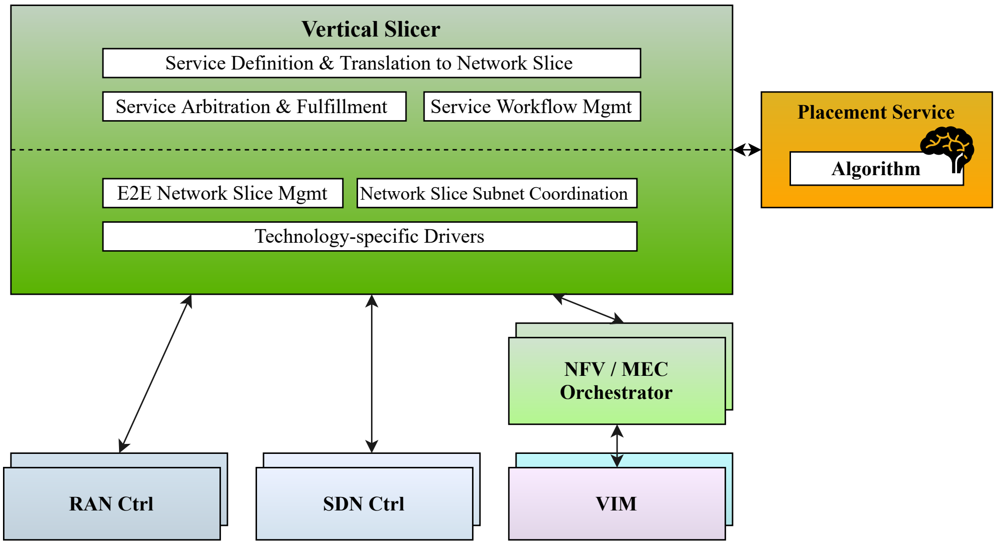
Vertical Slicer
Service Definition & Translation to Network Slice
Service Arbitration & Fulfillment
Service Workflow Mgmt
E2E Network Slice Mgmt
Network Slice Subnet Coordination
Technology-specific Drivers
Placement Service
Algorithm
NFV / MEC Orchestrator
RAN Ctrl
SDN Ctrl
VIM

Figure 4. Orchestration Framework high-level architecture.

5. Evaluation Methodology and Simulation Scenarios

In our work, the heuristics were developed in MATLAB, while IBM CPLEX [33] was used to compute the optimal solution of the MILP described in Section 2.1. The simulation scenarios considered a gNB sector area of 500 m radius overlaid with two SC clusters [25], as shown in Figure 5. Each cluster contained four possible SCs, which were randomly and uniformly placed in an 100-m-radius from the cluster centers. The minimum allowable distances are given in [25]. A subset of BSs, namely one SC (randomly selected) per cluster and the gNB, was assumed to have fiber access to the aggregation network. mmWave X-haul links could be deployed among BSs as long as their distance was lower than 200 m. The aggregation network consisted of two layers each one comprising four possible node positions, as shown in Figure 5. The aggregation layer nodes were connected among them and with the BSs via fiber so that there was no disconnected node. Hotspot UE traffic was also assumed [25].

We considered five different SFCs containing ordered combinations of the following VNFs: NAT (Network Address Translation), FW (Firewall), TM (Traffic Monitor), WOC (WAN Optimization Controller), IDPS (Intrusion Detection Prevention System), and VOC (Video Optimization Controller). For each SFC, the corresponding ordered VNF combination, data rate demand (uniformly distributed in the provided set), E2E delay requirement and share of the total SFC requests, are shown in Table 2. The data processing capacity as well as the GFLOPS requirement of each VNF type are given in Table 3. For a given number of UEs, we ran 10 different scenarios, with five different UE distribution snapshots each.

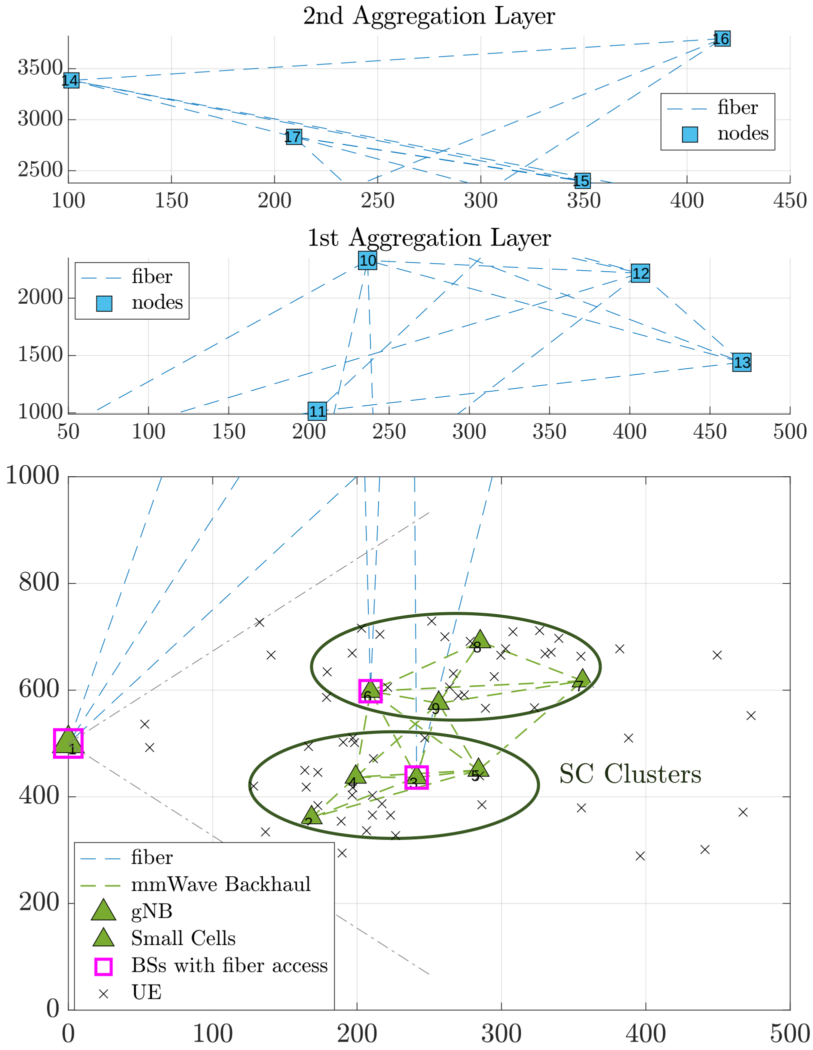

Figure 5. Simulation scenario example topology.

Table 2. SFC details [15].

Type	VNF Ordering	Throughput (Mbps)	Delay (ms)	Share (%)
Web	NAT→FW→TM→WOC→IDPS	[0.6–1]	500	20
VoIP	NAT→FW→TM→FW→NAT	[0.404–0.64]	100	20
Streaming	NAT→FW→TM→VOC→IDPS	[5–24]	100	39
Gaming	NAT→FW→VOC→WOC→IDPS	[0.24–0.5]	60	6
Ultra RT AI/ML	NAT→NAT	[15–25]	1	15

Table 3. VNF details [15].

Type	NAT	FW	TM	VOC	WOC	IDPS
Process Capacity (Mbps)	500	400	200	578	300	600
GFLOPS Requirement	110	440	55	110	110	440

We assumed 100 RBs allocated per gNB or SC (corresponding to $\mu = 0$ in 5G numerology). The operating frequency of the gNB and SCs was 2 GHz, assuming orthogonal channels between the gNB and the SCs. However, the SCs of different clusters could interfere with each other. The mmWave X-haul links operated at 60 GHz, with 200 MHz channel bandwidth. For the AN and the mmWave links, we employed the link budget equation and related parameter values of [25]. The number of CPU cores was equal to 8, 24, and 48 for each node at the MEC, 1st, and 2nd Aggregation layers, respectively. Parameter $P_y^{(CPU,m)}$ was selected randomly from the set $\{55, 70\}$ for the MEC nodes, from $\{150, 220\}$ for the 1st Aggregation Layer and from $\{200, 278\}$ for the 2nd Aggregation Layer nodes, while the idle power was assumed to be equal to 10% of the assigned maximum power values. Each fiber link had a capacity of 10 Gbps, while the rest of the simulation parameters are summarized in Table 4.

Table 4. Simulation Parameters [15,25,34].

Parameter	Value	Parameter	Value	Parameter	Value
$P_{idle}^{(sw)}$	315 W	P_{port}	7 W	Packet length	1.5 KB
$N_{RF}^{(gNB)}$	8	$N_{RF}^{(SC)}$	4	$N_{RF}^{(mmW)}$	64
$\Delta_a^{(gNB)}$	4.7	$\Delta_a^{(SC)}$	4	$\Delta_e^{(mmW)}$	100
$P_a^{(gNB,i)}$	130 W	$P_a^{(SC,i)}$	6.8 W	$P_e^{(mmW,i)}$	3.9 W

The optimal solution and the proposed heuristic (HERO) were compared with the following SoA algorithms [15]:

- Holu: This algorithm first performed the VNF placement based on the node centrality (closeness) and the CPU utilization of computing nodes and then decided upon traffic routing targeting the minimization of power consumption, while satisfying the E2E service delay constraint.
- BCSP: It considered node centrality (betweeness) for the VNF placement and the shortest-path in terms of delay for routing, while meeting the E2E service delay constraint.

Given that both reference algorithms did not handle user association, we employed the default user association criterion for both, i.e., the UEs were connected to the BSs based on the highest received signal-to-interference-plus-noise ratio (SINR), and consequently, lowest number of required RBs. In addition, for a fair comparison, their UE examination order was selected to be the same as HERO.

6. Simulation Results and Discussion

In Figures 6 and 7, we show the normalized w.r.t. the Optimal solution of (17) energy efficiency (in bits/Joule) and computational time (in logarithmic scale), respectively, for all algorithms and different number of UEs. As can be seen, HERO provided a very good tradeoff between energy efficiency and complexity compared to the other approaches, achieving up to 78% of the Optimal value, with up to 742 times lower complexity. All algorithms had a 100% user acceptance ratio in all cases, except for BCSP which achieved 96% for N = 20, 93% for N = 40, and 91% for N = 40. This is due to the fact that in BCSP the CPU utilization of the computing nodes was not taken into account, resulting in less efficient VNF placement, which under higher traffic load could lead to a few UEs being blocked. The inefficiency of BCSP's VNF placement is also shown in Figure 8, where the power break down of all algorithms for different number of UEs per gNB sector area is presented. As shown, inefficient BCSP VNF placement led to a higher number of deployed computational nodes, and thus, higher power consumption.

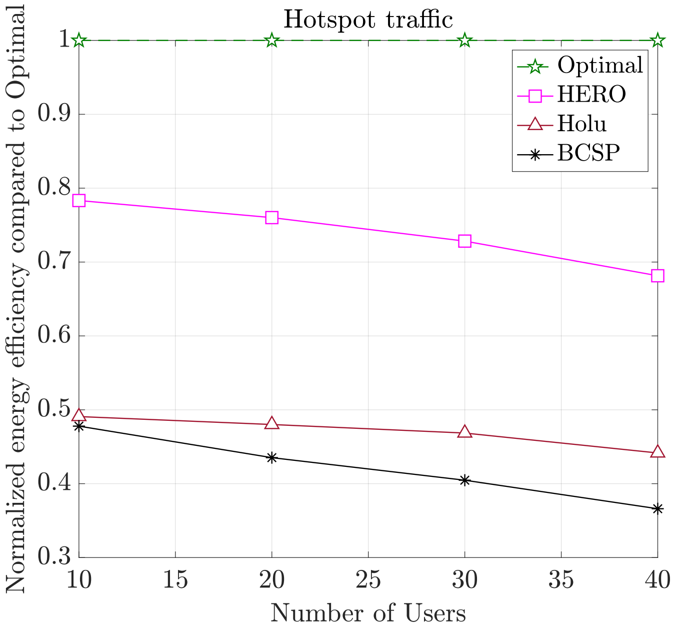

Figure 6. Energy efficiency (bits/Joule) of all algorithms for different numbers of UEs per gNB area.

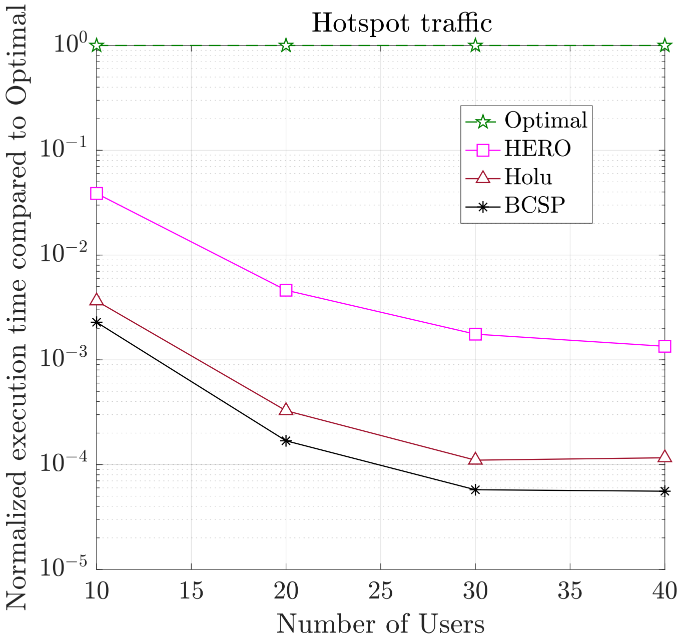

Figure 7. Execution time (s) in logarithmic scale of all algorithms for different numbers of UEs per gNB area.

Compared to Holu and BCSP, HERO achieved up to 60% and 86% higher energy efficiency, respectively, while keeping the complexity low, as shown in Figure 7. This is due to the fact that HERO additionally considered user association as part of the optimization problem leading to higher flexibility at the expense of a little higher complexity. On the other hand, in both Holu and BCSP, the serving BSs were already decided (based on the best SINR criterion) and then the optimal VNF placement and traffic routing from the UE traffic source to its serving BS were performed. This is also demonstrated in Figure 8, where the Optimal and HERO did not deploy the gNB in any case but used SCs as serving BSs (in contrast to Holu and BCSP), hence, leading to much lower power consumption. We also observed that the power consumption of the Optimal and HERO were scaling better than the SoA with increasing load (HERO still achieved 68% of the Optimal energy efficiency value when $N = 40$).

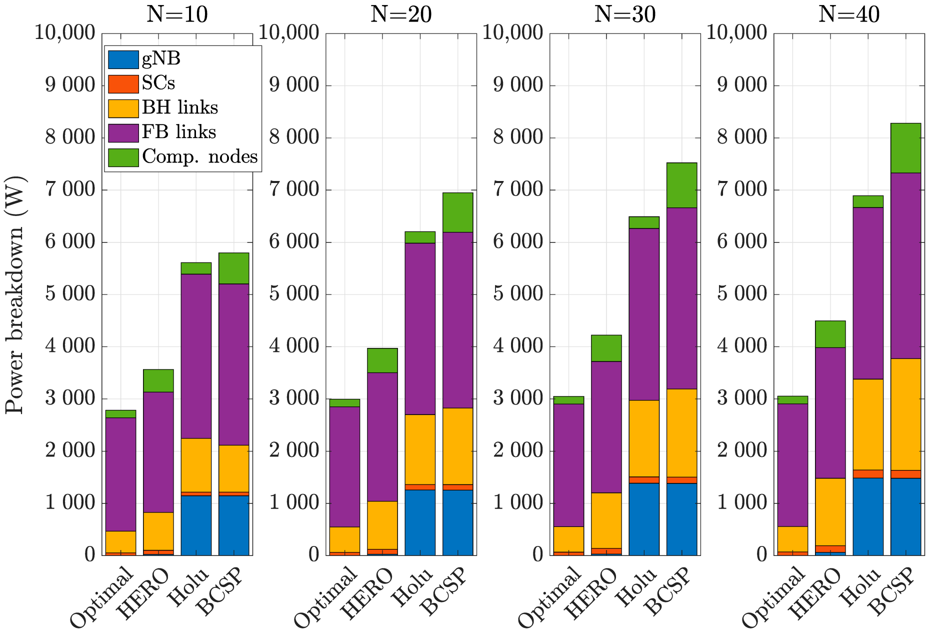

Figure 8. Power (W) breakdown of all algorithms for low traffic (N = 10) and high traffic (N = 40).

To further prove the improved performance of the proposed solutions (Optimal and HERO), we show in Figure 9, the percentage of active nodes and links for all algorithms for different number of UEs. As can be observed, Holu and BCSP, which did not optimize user association, kept the gNB always active, thus resulting in much higher power consumption, as shown in Figure 8. In addition, in Holu and BCSP, the number of active SCs, increased at a much higher rate with increasing load compared to the proposed approaches, which proves the scalability in terms of energy efficiency of the proposed solutions unlike the SoA. Compared to Holu, HERO activated fewer BH and FB links, and consequently fewer switches at the expense, however, of a higher number of active computational nodes. This is due to the fact that, in the considered scenario, which, however, was an accurate representation of actual networks of this type, the fiber links and switches had much higher energy impact ($P^{(sw)}_{idle}$=315 W) compared to the computational nodes, so that more efficient user association and routing had a higher impact on the overall performance compared to the VNF placement. As a result, Holu, which gave higher priority to VNF placement at the expense of traffic routing efficiency, while not taking user association into account at all, resulted, as already shown, in poorer overall performance. It is worth noting, however, that the proposed solutions are independent of the power consumption values, since the latter constitute the algorithms's input, and thus, different values could lead to different results always targeting high energy efficiency. As a final remark, the performance gains of the proposed algorithms justify the motivation of our work that user association, VNF placement, and traffic routing should be jointly considered to guarantee true optimal E2E network performance.

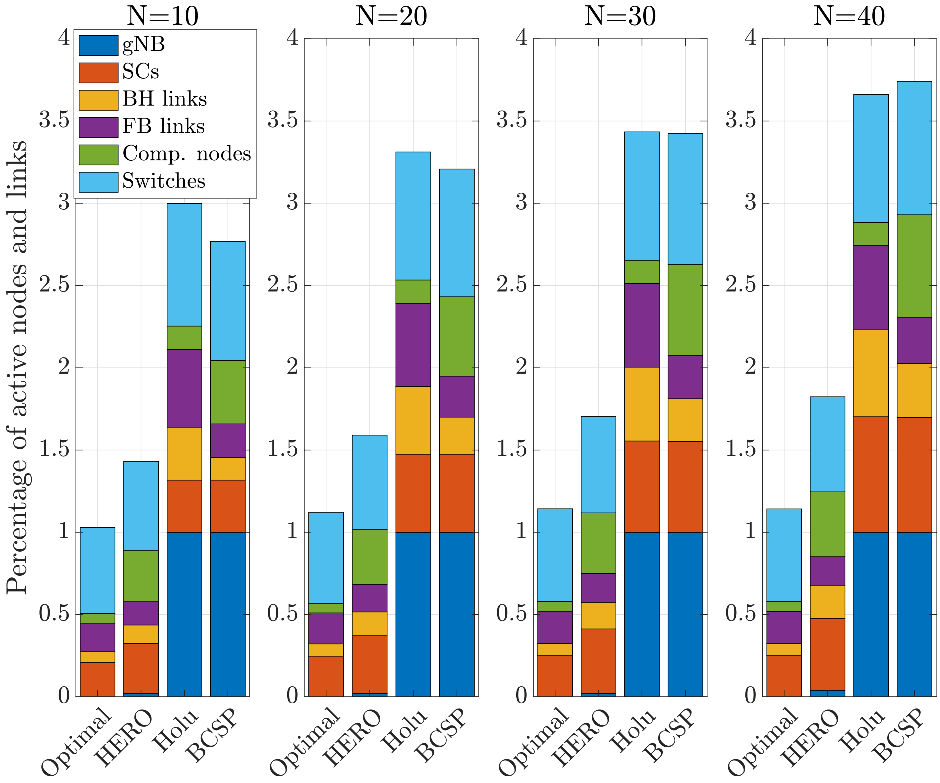

Figure 9. Percentage of active nodes and links of all algorithms for different numbers of UEs per gNB area.

7. Conclusions

In this paper, we studied the joint VNF placement, user association, and traffic routing in B5G networks targeting energy efficiency maximization, while ensuring a high UE acceptance ratio. We modeled the aforementioned problem as a MILP with minimal assumptions, which captured all characteristics of the employed technologies, resource, and service types as well as their constraints and power consumption. To tackle the prohibitive complexity of the studied problem, we proposed HERO, an energy-efficient resource planning heuristic, which was shown to significantly outperform the SoA, while achieving up to 78% of the optimal value, with up to 742 times lower complexity. Finally, we described the Orchestration Framework that is responsible for the dynamic and automated orchestration of Vertical-driven services in tailored E2E Network Slices, and explained how such a framework is assisted by the proposed algorithms, which jointly compute the optimal allocation of both network and compute resources across the radio, core, and transport network segments.

Although the proposed formulation and heuristic algorithm can be extended to also handle service requests dynamically arriving in time (i.e., they can be used to implement an **online** joint computational and communications resource allocation policy), the computational complexity of the heuristic may still be too high to properly adapt to the small time scale of request arrivals. To this end, in the future, we plan to explore learning-based techniques (inspired from approximate reinforcement learning) to construct low-complexity online energy-efficient resource allocation algorithms.

Author Contributions: Conceptualization, M.G. and A.M.; architecture design, F.M. and G.L.; methodology, A.M. and M.G.; prototype implementation, L.L.; validation, A.M. and M.G.; formal analysis, M.G.; resources, G.K.; writing—original draft preparation, M.G. and A.M.; writing—review and editing, G.K., N.P. and F.M.; visualization, A.M.; supervision, N.P. and N.C.; project administration, G.K. and F.M.; funding acquisition, G.K. and N.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by H2020 5G-COMPLETE (GA 871900).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

3GPP	3rd Generation Partnership Project
5G	Fifth Generation
5G-NR	5G-New Radio
5G-PPP	5G Infrastructure Public Private Partnership
6G	Sixth Generation
B5G	Beyond Fifth Generation
BS	Base Station
AN	Access Network
CAPEX	Capital Expenditure
CPU	Central Processing Unit
C-RAN	Centralized-Radio Access Network
DL	Downlink
E2E	End-to-end
eMBB	Enhanced Mobile Broadband
ETSI	European Telecommunications Standards Institute
FPGA	Field Programmable Gate Array
FW	Firewall
gNB	gNodeB
GPU	Graphics Processing Unit
ICT	Information and Communications Technology
IDPS	Intrusion Detection/Prevention System
LCM	Lifecycle Management
MANO	Management and Orchestration
MEC	Multi-Access Edge Computing
MILP	Mixed Integer Linear Program
mMTC	Massive Machine Type Communication
NAT	Network Address Translation
NFV	Network Function Virtualization
OPEX	Operational Expenditure
OSI	Open System Interconnection
QoS	Quality of Service
RAN	Radio Access Network
RB	Resource Block
RF	Radio Frequency
SINR	Signal to Interference plus Noise Ratio
State-of-the-Art	SoA
SC	Small Cell
SDN	Software Defined Networking
SDO	Standards Developing Organization
SFC	Service Function Chain
TM	Traffic Monitor
uRLLC	Ultra Reliable Low Latency Communication
UE	User Equipment

UHD	Ultra-High Definition
VM	Virtual Machine
VNF	Virtual Network Function
VOC	Video Optimization Controller
WOC	WAN Optimization Controller

References

1. Ali, J.; Roh, B.H.; Lee, S. QoS improvement with an optimum controller selection for software-defined networks. *PLoS ONE* **2019**, *14*, e0217631. [CrossRef] [PubMed]
2. Ali, J.; Roh, B.H. Quality of Service improvement with optimal software-defined networking controller and control plane clustering. *Comput. Mater. Contin.* **2021**, *67*, 849–875. [CrossRef]
3. Barakabitze, A.; Ahmad, A.; Mijumbi, R.; Hines, A. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Comput. Netw.* **2020**, *167*, 106984. [CrossRef]
4. Abbas, N.; Zhang, Y.; Taherkordi, A.; Skeie, T. Mobile Edge Computing: A survey. *IEEE Internet Things J.* **2018**, *5*, 450–465. [CrossRef]
5. Tang, L.; Yang, H.; Ma, R.; Hu, L.; Wang, W.; Chen, Q. Queue-aware dynamic placement of virtual network functions in 5G access network. *IEEE Access* **2018**, *6*, 44291–44305. [CrossRef]
6. Laghrissi, A.; Taleb, T. A survey on the placement of virtual resources and Virtual Network Functions. *IEEE Commun. Surv. Tutor.* **2019**, *1*, 1409–1434. [CrossRef]
7. Liu, J.; Lu, W.; Zhou, F.; Lu, P.; Zhu, Z. On dynamic Service Function Chain deployment and readjustment. *IEEE Trans. Netw. Serv. Manag.* **2017**, *14*, 543–553. [CrossRef]
8. Dong, L.; da Fonseca, N.; Zhu, Z. Application-Driven provisioning of Service Function Chains over heterogeneous NFV platforms. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 3037–3048. [CrossRef]
9. Cao, H.; Du, J.; Zhao, H.; Luo, D.; Kumar, N.; Yang, L.; Yu, F. Resource-Ability assisted Service Function Chain embedding and scheduling for 6G networks With virtualization. *IEEE Trans. Veh. Technol.* **2021**, *70*, 3846–3859. [CrossRef]
10. Cudak, M.; Ghosh, A.; Ghosh, A.; Andrews, J. Integrated access and backhaul: A key enabler for 5G millimeter-wave deployments. *IEEE Commun. Mag.* **2021**, *59*, 88–94. [CrossRef]
11. Ghribi, C.; Mechtri, M.; Zeghlache, D. A dynamic programming algorithm for joint VNF placement and chaining. In Proceedings of the 2016 ACM Workshop on Cloud-Assisted Networking, Irvine, CA, USA, 12 December 2016.
12. Ma, Y.; Liang, W.; Huang, M.; Xu, W.; Guo, S. Virtual Network Function service provisioning in MEC via trading off the usages between computing and communication resources. (early access). *IEEE Trans. Cloud Comput.* **2020**. [CrossRef]
13. Soualah, O.; Mechtri, M.; Ghribi, C.; Zeghlache, D. Energy efficient algorithm for VNF placement and chaining. In Proceedings of the 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, Spain, 14–17 May 2017.
14. Ayoubi, S.; Sebbah, S.; Assi, C. A logic-based Benders decomposition approach for the VNF assignment problem. *IEEE Trans. Cloud Comput.* **2019**, *7*, 894–906. [CrossRef]
15. Varasteh, A.; Madiwalar, B.; van Bemten, A.; Kellerer, W.; Mas–Machuca, C. Holu: Power-aware and delay-constrained VNF placement and chaining. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 1524–1539. [CrossRef]
16. Tajiki, M.; Salsano, S.; Chiaraviglio, L.; Shojafar, M.; Akbari, B. Joint energy efficient and QoS-aware path allocation and VNF placement for Service Function Chaining. *IEEE Trans. Netw. Serv. Manag.* **2019**, *16*, 374–388. [CrossRef]
17. Dietrich, D.; Papagianni, C.; Papadimitriou, P.; Baras, J. Network function placement on virtualized cellular cores. In Proceedings of the 2017 9th International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, 4–8 January 2017.
18. Yang, S.; Li, F.; Trajanovski, S.; Chen, X.; Wang, Y.; Fu, X. Delay-aware Virtual Network Function placement and routing in edge clouds. *IEEE Trans. Mobile Comput.* **2021**, *20*, 445–459. [CrossRef]
19. Spinelli, F.; Mancuso, V. Toward enabled industrial verticals in 5G: A survey on MEC-based approaches to provisioning and flexibility. *IEEE Commun. Surv. Tutor.* **2021**, *23*, 596–630. [CrossRef]
20. Open Source MANO. 2021. Available online: https://osm.etsi.org/ (accessed on 7 October 2021).
21. O-RAN Alliance. 2021. Available online: https://www.o-ran.org/ (accessed on 7 October 2021).
22. Casetti, C.; Chiasserini, C.; Martín-Pérez, J.; Molner, N.; Deiss, T. The Vertical Slicer: Verticals' Entry Point to 5G Networks. In Proceedings of the 27th European Conference on Networks and Communications (EuCNC 2018), Ljubljana, Slovenia, 18–21 June 2018.
23. Perez, M.G.; Perez, G.M.; Giardina, P.; Bernini, G.; Neves, P.; Alcaraz-Calero, J.; Wang, Q.; Koutsopoulos, K. Self-Organizing Capabilities in 5G Networks: NFV & SDN Coordination in a Complex Use Case. In Proceedings of the 27th European Conference on Networks and Communications (EuCNC 2018), Ljubljana, Slovenia, 18–21 June 2018.
24. Khalili, H.; Papageorgiou, A.; Siddiqui, M.; Meixner, C.C.; Carrozzo, G.; Nejabati, R.; Simeonidou, D. Network Slicing-aware NFV Orchestration for 5G Service Platforms. In Proceedings of the 28th European Conference on Networks and Communications (EuCNC 2019), Valencia, Spain, 18–21 June 2019.

25. Mesodiakaki, A.; Zola, E.; Santos, R.; Kassler, A. Optimal user association, backhaul routing and switching off in 5G heterogeneous networks with mesh millimeter wave backhaul links. *Ad Hoc Netw.* **2018**, *78*, 99–114. [CrossRef]
26. ETSI ES 203 228: Environmental Engineering (EE): Assessment of Mobile Network Energy Efficiency, November 2020. v.1.3.1. Available online: https://www.etsi.org/deliver/etsi_es/203200_203299/203228/01.03.01_60/es_203228v010301p.pdf (accessed on 7 October 2021).
27. ETSI EN 303 471: Energy Efficiency Measurement Methodology and Metrics for Network Function Virtualisation (NFV), January 2019. v.1.1.1. Available online: https://www.etsi.org/deliver/etsi_en/303400_303499/303471/01.01.01_60/en_303471v010101p.pdf (accessed on 7 October 2021).
28. Li, X.; Garcia-Saavedra, A.; Costa-Perez, X.; Bernardos, C.; Guimaraes, C.; Antevski, K.; Mangues-Bafalluy, J.; Baranda, J.; Zeydan, E.; Corujo, D.; et al. 5Growth: An End-to-End Service Platform for Automated Deployment and Management of Vertical Services over 5G Networks. *IEEE Commun. Mag.* **2012**, *59*, 84–89. [CrossRef]
29. Gatzianas, M.; Mesodiakaki, A.; Kalfas, G.; Pleros, N. Energy-efficient joint computational and network resource planning in Beyond 5G networks. *Proc. IEEE Globecom.* **2021**, to appear.
30. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
31. Vertical Slicer. Available online: https://github.com/nextworks-it/slicer (accessed on 7 October 2021).
32. TR 28.541: Management and Orchestration; 5G Network Resource Model (NRM); Stage 2 and Stage 3 (Rel. 17), 2020. v. 17.0.0. Available online: https://www.3gpp.org/ftp//Specs/archive/28_series/28.541/28541-h01.zip (accessed on 7 October 2021).
33. ILOG CPLEX Optimization Studio. 2020. Available online: https://www.ibm.com/products/ilog-cplex-optimization-studio (accessed on 7 October 2021).
34. Heddeghem, W.V.; Idzikowski, F.; Vereecken, W.; Colle, D.; Pickavet, M.; Demeester, P. Power consumption modeling in optical multilayer networks. *Photon. Netw. Commun.* **2012**, *24*, 86–102. [CrossRef]