IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 May 2024, accepted 16 May 2024, date of publication 27 May 2024, date of current version 4 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406206

RESEARCH ARTICLE

Power-Efficient Joint Dynamic Resource Allocation in Virtualized Inter-Data Center Elastic Optical Networks

MAHTA AHMADI, MOHAMMAD HADI, AND MOHAMMAD REZA PAKRAVAN, (Member, IEEE)
Department of Electrical Engineering, Sharif University of Technology, Tehran 1458889694, Iran
Corresponding author: Mohammad Hadi (mohammad.hadi@sharif.edu)

ABSTRACT Network virtualization, as a key driver for revenue enhancement, is under deep exploration by the communication society. Elastic optical network is a growing inter-data center communication solution whose flexible nature facilitates commercial implementation of virtualization concepts. Flexibility and virtuality are two main pillars of efficient networking which meet in a virtualized inter-data center elastic optical network and their interwoven management can significantly reduce the power consumption of data center networks, as a significant energy-consuming contributor. An efficient resource allocation is required to harvest the expected gain of a virtualized network. We propose a power-efficient joint dynamic resource allocation scheme, which optimally allocates computation and communication resources in an inter-data center elastic optical network. The scheme incorporates the benefits of conventional offline and online resource allocation schemes and reduces the power consumption by 19% compared to its traditional counterparts while keeping the delay and service outage below their desired thresholds. The proposed scheme not only supports traffic streams with distinguished service requirements but also exploits the involved service differentiation to further improve power efficiency. Strikingly, in a typical scenario where delay-sensitive and delay-tolerant services co-exist, the reduction in power consumption can be enhanced up to 50%. The efficiency is further improved for a higher number of delay-tolerant services with more relaxed delay requirement. The proposed scheme provides an intelligent compromise between service outage and service requirements under tough networking conditions such that the service requirements of the delay-sensitive services are maintained for the least number of dropped bits in the delay-tolerant services.

INDEX TERMS Elastic optical networks, network virtualization, quality of service, resource allocation, stochastic Lyapunov optimization.

I. INTRODUCTION

Data centers (DCs) and transport networks (TNs), as two significant contributors to the global energy consumption, account for the majority of ICT power consumption [1], approximately 80%, which is predicted to capture 20% of global electricity demand by 2030 [2]. Given the increasing prominence of DCs as significant energy consumers globally, cloud providers are motivated to reduce their energy consumption to address economic and environmental concerns, such as high power consumption bills and government taxes on greenhouse gas emissions [3].

The associate editor coordinating the review of this manuscript and approving it for publication was San-Liang Lee.

Network function virtualization (NFV) has revolutionized the execution of tasks formerly reliant on specialized hardware, offering compelling advantages including cost-effectiveness, flexibility, and simplified management. Virtualized network functions (VNFs) have become effortlessly deployable, scalable, and orchestratable through software. The advent of NFV has provided various optimization opportunities whose effective exploitation can improve power efficiency and reduce implementation and maintenance costs [4], [5]. Optical networks, as the main TN solution for connecting DCs, have been expanded to serve inter- and intra-DC connection requests with heavy and bursty traffic streams. Elastic optical network (EON) is a lucrative implementation of optical TNs which enables power-efficient usage of

network resources [6]. Flexible and online reconfiguration has made EONs an appropriate platform for inter-connection of VNFs in a service chain (SC) processed distributively on several disperse DCs [6], [7].

NFV and EON technologies offer tangible improvements in both capital expenditure (CAPEX) and operational expenditure (OPEX) for cloud service providers and network operators. Following the deployment of equipment and the allocation of CAPEX, the adoption of an optimal resource allocation (RA) strategy, leveraging the inherent flexibility of NFV and EON solutions, can effectively reduce OPEX [4], [7]. Within this context, optimizing energy consumption serves the dual purpose of reducing OPEX and addressing environmental concerns [3]. In order to maximize power efficiency, the implementation of an effective RA approach is necessary.

Static RA, which has been deeply investigated in the literature [8], leads to inefficient over-provisioning and is not applicable to dynamic networks. On the other hand, the proposed dynamic RA schemes solely consider the short-term behavior of the network without any attention to the long-term behaviors, which leads to unnecessary reconfigurations and network instability [9]. Further, they are usually developed according to a simplified model of the TN, which cannot fully describe the involved physical constraints [10]. Lyapunov drift theory is an interesting tool which allows to concurrently handle short- and long-term network behaviors [9]. Moreover, the Lyapunov tool enables decomposition of a complex dynamic RA problem into a sequence of interconnected simple tractable sub-problems [11]. Furthermore, the Lyapunov tool considers a more realistic traffic model with complete dynamism meaning that the traffic value can vary across the time rather than remaining fixed as conventionally assumed in the literature [11]. This dynamic nature holds practical relevance in real-world scenarios. For instance, as highlighted in [12], the internet of things (IoT) often exhibits user data flow with stochastic characteristics attributed to the unpredictable demand arrivals and the diverse range of the involved applications. Although the accurate description of the traffic model for such data flow has been quite challenging [12], [13], the capabilities of the Lyapunov tool allows to elaborately cope with the modeling issues. Considering its successful applications for dynamic RA in EON and elastic optical fronthaul [10], [14], [15] and also in geo-distributed DCs with renewable resources [16], the Lyapunov tool seems to be a potential tool for joint dynamic allocation of communication and computation resources in a virtualized inter-DC EON, as demonstrated in this paper.

In a real networking scenario, a diverse set of quality of service (QoS) requirements should be supported. Particularly, delay, as a key QoS factor, plays a key role in service differentiation in modern networks. As an example, a strict latency below millisecond to several milliseconds is forced for delay-sensitive services such as web service and instant communication while Google frequently handles a significant volume of tasks that can endure delays from a few minutes to even several hours, such as scientific computation and data backup or more specifically, analyzing click-throughs and processing logs of websites, which constitute delay-tolerant services [16], [17]. Despite of its importance, delay-based service differentiation is not investigated deeply in the literature, mainly due to the involved complexity [18], [19]. Fortunately, the Lyapunov tool allows to takes the delay effectively into account and provide a QoS-differentiated service provisioning mechanism for the diverse service requirements coexisted within a common network.

VNFs of a requested SC can be distributed among disperse DCs, entailing inter-DC networking, or be confined inside a single DC, necessitating intra-DC networking [5]. Processing a VNF SC inside a DC can significantly decrease VNF deployment cost although it may negatively affect the network stability due to unbalanced utilization of computation resources [7]. Further, confined processing facilitates handling of interconnected VNF SCs with possible feedback dependencies. In this letter, we assume that each VNF SC is entirely processed within a nominated DC, similar to [6] and [20], and use the Lyapunov tool to guarantee the stability of the involved inter-DC network. Such an assumption can even support distributed processing of a VNF SC since each SC can be decomposed into a sequence of VNFs, each processed independently in a nominated DC. The inherent dynamicity of the Lyapunov tool allows to adaptively reassign the nominated DC of a VNF without stringent QoS requirements over time. This novel approach inserts QoS commitments to the RA process and improves it considerably, particularly, when a shortage condition of computation resources occurs. Fortunately, the Lyapunov tool is general enough to simultaneously consider the DCs and its underlying TN and develop a synergic platform for joint management of the computation and communication resources. This allows a comprehensive description of the optical TN and its associative constraints, such as spectrum and delay limitations, to be incorporated into the RA and makes the results more realistic compared to the previous works on DC management without any detailed attention to the underlying TN and important QoS parameters such as delay.

Simulation results demonstrate that the proposed power-efficient joint dynamic RA scheme improves the power efficiency substantially, up to 19%, compared to the conventional schemes. In a typical mixed networking scenario where delay-sensitive and delay-tolerant services co-exist, the improvement jumps to 50% compared to a counterpart scenario without any attention to different QoS requirements. The proposed schemes offers controllable trade offs between delay, service outage, and power efficiency, as illustrated by the numerical results. These trade offs facilitate network managements and provide meaningful insights for the network operator to better diagnose networking bottlenecks.

As a concise encapsulation of the research work's contributions, our research focuses on dynamic RA for VNF requests in inter-DC EON with a primary emphasis on power

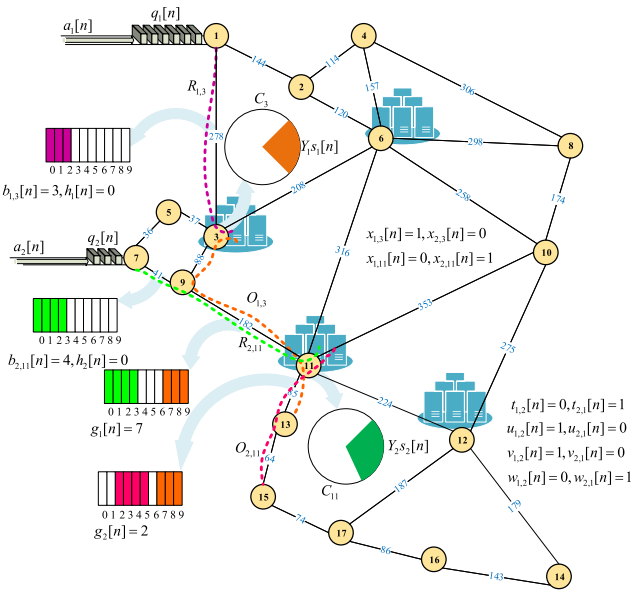

FIGURE 1. System model architecture with the adopted topology of Nobel Germany and two sample requests, one originating from node 1 to node 13 and the other from node 7 to node 15. The VNF of the first and second requests are processed in the nominated DCs 3 and 11 whose available processing capacity of shown by the circles. The uplink and downlink lightpaths of each request and their corresponding assigned spectrum slots are shown in different colors. Satisfaction of the spectrum contiguity, spectrum continuity, non-overlapping, queue stability, and processing capacity constraints are graphically illustrated. Notations in the figure are refined according to the index numbers of the sample connections and the values pertain solely to a single time interval.

efficiency. To accomplish this objective, we leverage the Lyapunov tool, which enables us to incorporate a realistic traffic model characterized by full dynamism, allowing for varying traffic values over time. Within our proposed framework, we carefully consider both delay-sensitive and delay-tolerant services, capitalizing on this opportunity to significantly reduce energy consumption. By thoroughly examining and balancing the trade-offs between delay, service outage, and power efficiency, our work provides valuable insights for making informed decisions in handling diverse QoS requirements and achieving more effective network management. Going forward, in Section II, we provide an overview of the related works. Following that, in Section III, we present the system model and introduce our power-efficient joint dynamic RA scheme designed for a virtualized inter-DC EON. In Section IV, we evaluate the performance of the proposed scheme. Finally, in Section V, we conclude the paper.

II. RELATED WORKS

NFV has attracted substantial attention from academia and industry due to its remarkable flexibility and cost-effectiveness. In recent years, there has been extensive research dedicated to addressing the challenge of RA for VNF requests in inter-DC networks. A set of algorithms is proposed in [20] that manages the admission of requests involving network function services in distributed cloud environments. Reference [13] addresses the challenge of VNF placement in cloud-based DCs by taking into account the dynamic nature of workloads over time. A method utilizing deep reinforcement learning is introduced in [21] for the purpose of achieving optimal placement of VNFs in software-defined and NFV-enabled networks. In [22], an algorithm is presented as a solution for optimizing the deployment of VNF SCs in NFV networks, with an emphasis on low communication delay constraint. Reference [12] investigates a dynamic approach for embedding VNF SCs in IoT networks with deep reinforcement learning. Unfortunately, the physical constraints such as spectrum contiguity, spectrum continuity, and fiber capacity, have not been considered in these studies and consequently, their proposed RA schemes cannot be implemented in real optical networks with network specific physical requirements.

Reference [23] introduces a Genetic Algorithm methodology to optimize the mapping of VNFs and the design of virtual topologies in a metropolitan wavelength division multiplexing (WDM) optical network that is equipped with multi-access edge computing resources. As internet and network technology advance, the limitations of WDM networks become evident in meeting the demands of real-world problems. The fixed grid spectrum allocation and low spectral utilization of WDM networks hinder their ability to address these challenges effectively [18]. Consequently, the WDM-based approach of [23] fails to capture the flexibility of an EON and address its intricate routing and spectrum allocation problem [18]. Reference [24] addresses the problem of optimizing the establishment and task scheduling of VNF SCs in an optical DC interconnection network by two time-efficient algorithms. Despite being implemented in an EON, the constraints and challenges associated with spectrum allocation are not considered. In this work, our focus lies in RA to VNF requests within the EON platform. Notably, our approach encompasses not only DCs and network paths but also the concurrent allocation of spectrum. Moreover, the adopted model serves as a foundation for incorporating practical intricacies of the EON framework in future studies.

In [6], two efficient algorithms are introduced for orchestrating VNF SCs in inter-DC EONs in order to enhance the profit of internet service providers by maximizing the acceptance of user requests. Reference [18] presents a multi-objective optimization model and a few accompanying algorithms for RA in EONs. This model includes routing and spectrum allocation, load balancing across DCs, and efficient deployment of VNFs. Reference [7] explores a solution for dynamic deploying VNF SCs in inter-DC EONs by improving spectrum fragmentation and underutilization. Reference [19] tackles the challenge of dynamic allocation of DC and fiber resources with a hierarchical deep reinforcement learning model, which is built upon a graph neural network framework. Reference [25] introduces a game-theoretic approach to address the dynamic provisioning of both inter-domain and intra-domain VNF requests in edge-cloud EONs. In [26], a proposal is made for the joint optimization of optical path provisioning and VNF placement in virtual content delivery networks, aiming to

enhance efficiency within EONs. The above works do not specifically address the deployment of diverse services with distinguished QoS requirements, especially delay, within the same network while this aspect holds significant importance for the development of novel network management strategies.

In [27], the so-called GreenVoIP framework is introduced, which utilizes NFV and SDN technologies to address energy-efficient RA in virtualized cloud centers providing voice over internet services, aiming to prevent overload and minimize power consumption. Reference [28] presents a provisioning approach for VNF SCs with a specific focus on reducing energy consumption in delay-sensitive applications such as 5G and IoT. A method is proposed in [29] for optimizing VNF placement in substrate networks by utilizing matching theory, with the goal of achieving power efficiency. Reference [30] unveils novel approaches for achieving energy-efficient and interference-aware VNF placement in DCs. These methods leverage techniques such as the first-fit heuristic algorithm and deep reinforcement learning to optimize performance. However, the studies have not used the benefits offered by EON architecture. Reference [31] presents a heuristic algorithm designed to optimize energy and spectrum efficiency when embedding virtual optical networks in EONs. The algorithm aims to provide a balance among energy consumption, spectrum utilization, and acceptance rate. Reference [32] addresses the energy efficient VNF SC deployment problem in inter-DC EON, specifically targeting the optimization of energy consumption. The work proposes an algorithm that leverages reinforcement learning and graph convolutional networks to achieve this goal. To our knowledge, there is a lack of research studies that have specifically addressed the power-efficient allocation of resources to VNF requests in inter-DC EONs, particularly concerning the implementation of services with varying latency requirements. Notably, the mentioned works have overlooked the consideration of delay-tolerant requests. In this study, we not only address the diverse latency requirements but also harness the inherent characteristics of delay-tolerant requests to leverage their potential in optimizing average power consumption.

III. PROBLEM STATEMENT

In this section, we will discuss the system model and then formulate a time-averaged stochastic optimization problem for its power-efficient RA. In order to streamline the optimization problem while maintaining its optimality, we leverage the Lyapunov method. Subsequently, we transform the optimization problem into a practical form that facilitates easier implementation.

A. SYSTEM MODEL

Let $\mathbb{Z}_{n_1}^{n_2} = \{n_1, n_1+1, \cdots, n_2\}$ and consider the EON shown in Fig. 1 composed of M nodes connected by optical fiber links according to an arbitrary topology. The allocation of resources is performed for N consecutive discrete time intervals $n \in \mathbb{Z}_1^N$ with each interval having a duration of $\mathcal{T}$. At each time interval n, $a_i[n]$ bits of request $i \in \mathbb{Z}_1^I$ arrive at a queue with backlog $q_i[n]$ and $s_i[n]$ bits are served from the queue. Queues can be virtual and/or actual, where virtual queues are designed to satisfy specific criteria, such as time average constraints [16]. The $s_i[n]$ served bits are then routed over the uplink lightpath $\mathcal{R}_{i,m}$ to a nominated DC, located in node $m \in \mathbb{Z}_1^M$ of the network, to feed a desired VNF and subsequently, they are routed to their destination via the downlink lightpath $\mathcal{O}_{i,m}$. The uplink and downlink lightpaths connecting any source and destination node pair to the DC of node m are precomputed and known with the overall length $\mathcal{L}_{i,m}$. Every link includes $\mathcal{S}$ frequency slots, each having a bandwidth granularity of $\mathcal{W}$. Nominating DC m for the VNF processing of request i, a spectrum band including $b_{i,m}[n]$ contiguous frequency slots are assigned to request i in interval n over the uplink and downlink lightpaths. This batch assignment of the frequency slots ensures that each individual lightpath is allocated a contiguous set of adjacent frequency slots, thereby satisfying the contiguity constraint. The frequency slots assigned to each lightpath remain the same across all the links that the lightpath traverses. Maintaining the same allocated frequency slot index throughout the entire length of the lightpath ensures that the same frequency slots are used, thus satisfying the continuity constraint. In addition to the fulfillment of contiguity and continuity constraints, it is imperative that non-overlapping and guard constraints are also satisfied to avoid spectrum conflict and switching distortion, respectively.

The processing workload of a VNF is $\mathcal{Y}_i$ times of its incoming traffic rate, where the coefficient $\mathcal{Y}_i$ depends in the performance characteristics of the network function [6]. The DC m may be nominated to process a VNF if it has legislative permission and enough processing capacity $\mathcal{C}_m$ to handle the workload of the VNF. Storage resources in DCs play a crucial role; however, they generally do not directly influence the performance metrics of VNFs, such as data processing speed or efficiency of transmission. Consequently, due to the complexity of RA, numerous studies such as [29] and [32] primarily concentrate on computing resources and network bandwidth, the fact followed in this work.

Each lightpath is connected to a pair of bandwidth variable transponders at its begin and end nodes and may pass several bandwidth variable wavelength cross-connects along its path. For each fiber link, it is presumed that there exists a suitable deployment of post, inline, and pre- amplifiers, with their respective gains being equivalent to the switching and fiber losses. Each transponder transmits with a modulation spectral efficiency $\mathcal{K}_i$ in the maximum bandwidth range of $\mathcal{D}_i$ and consumes the power $\mathcal{E}_i + \mathcal{F}_i\mathcal{K}_i$ per each frequency slot [10]. The power consumption of cross-connects and amplifiers are neglected since they are primarily determined by the network topology and are independent of the switched traffic and involved reconfigurable parameters [10]. An active DC requires a constant power $\mathcal{P}_m$ and a varying power which is a linear function of the processing load with the slope

coefficient $\mathcal{Q}_m$ [33]. In the interest of streamlining, the focus of power consumption calculations is directed toward the physical and optical layers. It's worth noting that the proposed platform possesses a broad scope, allowing for the inclusion of considerations in higher layers.

The system model architecture, as depicted in Fig. 1, adopts the topology of Nobel Germany. The DCs are strategically positioned at nodes with the highest degree, nodes 3, 6, 11, and 12. This architecture illustrates two sample requests to demonstrate the system's operation within the time interval n. Two requests are initiated from nodes 1 and 7, targeting nodes 13 and 15, respectively. The process of servicing requests and creating queues can be observed at the source nodes. Their associated VNFs are processed in DCs 3 and 11, respectively. The processing capacity of the DCs, represented by circles, is enough to serve these requests with the necessary processing volume scales in proportion to their served bits, $s_1[n]$ and $s_2[n]$. The uplink and downlink lightpaths, along with their associated spectrum for each request, are visualized using different colors. Spectrum contiguity is demonstrated by allocating contiguous spectrum slots for each request in links, taking into account the fiber bandwidth. Examples of spectrum continuity are lightpaths $\mathcal{O}_{1,3}$ and $\mathcal{R}_{2,11}$, where the assigned spectrum slots remain unchanged across different links of the lightpath. Furthermore, the assigned spectrum slots for different requests do not overlap, and a guard band should be assigned between them. An example of this implementation is the link from 11 to 13, which is common between the requests, and a guard band with a width of 1 frequency slot is assigned between them. Specifically, the figure specifies the values of the variables related to spectrum allocation optimization formulated in the next subsection, ensuring compliance with the necessary constraints.

Table 1 provides a summary of the parameters utilized in the paper.

B. PROBLEM FORMULATION

For a power-efficient network configuration, the nominated DCs and transponders settings must be properly determined to minimize power consumption while queue stability, QoS, and physical constraints are satisfied. Formulation (1) is a stochastic optimization problem whose solution determines the desired power-efficient configuration for the time slots $n \in \mathbb{Z}_i^N$.

$$\min_{\substack{\boldsymbol{b}[n],\boldsymbol{h}[n],\boldsymbol{g}[n],\boldsymbol{z}[n],\boldsymbol{x}[n]\\ ,\boldsymbol{t}[n],\boldsymbol{u}[n],\boldsymbol{v}[n],\boldsymbol{w}[n],n\in\mathbb{Z}_1^N}} \quad \alpha \sum_{i\in\mathbb{Z}_1^I} 4(\mathcal{E}_i + \mathcal{F}_i\mathcal{K}_i) \sum_{m\in\mathbb{Z}_1^M} \bar{b}_{i,m}$$
$$+ \beta \sum_{m\in\mathbb{Z}_1^M} \left(\mathcal{P}_m \bar{z}_m + \mathcal{Q}_m \sum_{i\in\mathbb{Z}_1^I} \mathcal{W}\mathcal{K}_i\mathcal{Y}_i \bar{b}_{i,m}\right) \tag{1a}$$
$$+ \gamma \sum_{i\in\mathbb{Z}_1^I} \sum_{m\in\mathbb{Z}_1^M} \bar{x}_{i,m}\mathcal{L}_{i,m},$$

$$\bar{a}_i \le \bar{s}_i, \quad i \in \mathbb{Z}_1^I, \tag{1b}$$

$$\sum_{m\in\mathbb{Z}_1^M} x_{i,m}[n] \le 1, \quad i \in \mathbb{Z}_1^I, \tag{1c}$$

TABLE 1. Constants and variables along with their corresponding definitions. Vectors, variable, and constants are given in bold, small, and capital format, respectively.

Type	Notation	Definition
Indices	$n \in \mathbb{Z}_1^N$	Time interval index
	$m, m' \in \mathbb{Z}_1^M$	Network node index
	$i, i' \in \mathbb{Z}_1^I$	Request index
Input parameters	N	Number of time intervals
	M	Number of network nodes
	I	Number of requests
	$\mathcal{T}$ [s]	Time interval duration
	$\mathcal{W}$ [Hz]	Frequency slot bandwidth
	$\mathcal{S}$	Number of frequency slots
	$\mathcal{B}$	Number of guard frequency slots
	$\boldsymbol{\mathcal{K}} = (\mathcal{K}_i)$ $[\frac{\text{bps}}{\text{Hz}}]$	Modulation spectral efficiency
	$\boldsymbol{\mathcal{D}} = (\mathcal{D}_i)$ [Hz]	Maximum transponder bandwidth
	α	Transponder power penalty coefficient
	β	DC power penalty coefficient
	γ	Lightpath length penalty coefficient
	λ	Lyapunov penalty coefficient
	$\boldsymbol{\mathcal{E}} = (\mathcal{E}_i)$ [W]	Transponder power bias
	$\boldsymbol{\mathcal{F}} = (\mathcal{F}_i)$ $[\frac{\text{W.Hz}}{\text{bps}}]$	Transponder power slope
	$\boldsymbol{\mathcal{P}} = (\mathcal{P}_m)$ [W]	DC idle power
	$\boldsymbol{\mathcal{Q}} = (\mathcal{Q}_m)$ $[\frac{\text{W}}{\text{bps}}]$	DC load power
	$\boldsymbol{\mathcal{C}} = (\mathcal{C}_m)$ [bps]	DC processing capacity
	$\boldsymbol{\mathcal{X}} = (\mathcal{X}_{i,m})$	DC processing permission
	$\boldsymbol{a}[n] = (a_i[n])$ [bit]	Arrived bits
	$\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_i)$	Required processing coefficient
	$\boldsymbol{\mathcal{A}} = (\mathcal{A}_i)$ [bit]	Queue length limit
	$\boldsymbol{\mathcal{L}} = (\mathcal{L}_{i,m})$ [m]	Overall length of lightpath
	$\boldsymbol{\mathcal{R}} = (\mathcal{R}_{i,m})$	Uplink lightpath
	$\boldsymbol{\mathcal{O}} = (\mathcal{O}_{i,m})$	Downlink lightpath
Output parameters	$\boldsymbol{b}[n] = (b_{i,m}[n])$	Assigned frequency slots
	$\boldsymbol{q}[n] = (q_i[n])$ [bit]	Queue length
	$\boldsymbol{h}[n] = (h_i[n])$	Uplink start frequency slot
	$\boldsymbol{g}[n] = (g_i[n])$	Downlink start frequency slot
	$\boldsymbol{z}[n] = (z_m[n])$	DC activation indicator
	$\boldsymbol{x}[n] = (x_{i,m}[n])$	VNF deployment indicator
	$\boldsymbol{s}[n] = (s_i[n])$ [bit]	Served bits
	$\boldsymbol{t}[n] = (t_{i,i'}[n])$	Uplink spectrum locator
	$\boldsymbol{u}[n] = (u_{i,i'}[n])$	Uplink-downlink spectrum locator
	$\boldsymbol{v}[n] = (v_{i',i}[n])$	Downlink-uplink spectrum locator
	$\boldsymbol{w}[n] = (w_{i,i'}[n])$	Downlink spectrum locator

$$x_{i,m}[n] \le \mathcal{X}_{i,m}, \quad i \in \mathbb{Z}_1^I, m \in \mathbb{Z}_1^M, \tag{1d}$$

$$\sum_{i\in\mathbb{Z}_1^I} \mathcal{W}\mathcal{K}_i\mathcal{Y}_i b_{i,m}[n] \le \mathcal{C}_m, \quad m \in \mathbb{Z}_1^M, \tag{1e}$$

$$\sum_{i\in\mathbb{Z}_1^I} x_{i,m}[n] \le I z_m[n], \quad m \in \mathbb{Z}_1^M, \tag{1f}$$

$$h_i[n] + \sum_{m\in\mathbb{Z}_1^M} b_{i,m}[n] \le \mathcal{S}, \quad i \in \mathbb{Z}_1^I, \tag{1g}$$

$$g_i[n] + \sum_{m\in\mathbb{Z}_1^M} b_{i,m}[n] \le \mathcal{S}, i \in \mathbb{Z}_1^I, \tag{1h}$$

$$\mathcal{W} b_{i,m}[n] \le \min\{\mathcal{S}\mathcal{W}, \mathcal{D}_i\} x_{i,m}[n], \quad i \in \mathbb{Z}_1^I, m \in \mathbb{Z}_1^M, \tag{1i}$$

$$t_{i,i'}[n] + t_{i',i}[n] = 1, \quad i, i' \in \mathbb{Z}_1^I : i \ne i', \tag{1j}$$

$$u_{i,i'}[n] + v_{i',i}[n] = 1, \quad i, i' \in \mathbb{Z}_1^I : i \ne i', \tag{1k}$$

$$w_{i,i'}[n] + w_{i',i}[n] = 1, \quad i, i' \in \mathbb{Z}_1^I : i \ne i', \tag{1l}$$

$$h_i[n] + b_{i,m}[n] + \mathcal{B} \leq h_{i'}[n] + (\mathcal{S} + \mathcal{B}) (3 - t_{i,i'}[n] - x_{i,m}[n] - x_{i',m'}[n]), \quad i, i' \in \mathbb{Z}_1^I, m, m' \in \mathbb{Z}_1^M : i \neq i', \mathcal{R}_{i,m} \cap \mathcal{R}_{i',m'} \neq \varnothing, \quad (1m)$$

$$h_i[n] + b_{i,m}[n] + \mathcal{B} \leq g_{i'}[n] + (\mathcal{S} + \mathcal{B}) (3 - u_{i,i'}[n] - x_{i,m}[n] - x_{i',m'}[n]), \quad i, i' \in \mathbb{Z}_1^I, m, m' \in \mathbb{Z}_1^M : i \neq i', \mathcal{R}_{i,m} \cap \mathcal{O}_{i',m'} \neq \varnothing, \quad (1n)$$

$$g_i[n] + b_{i,m}[n] + \mathcal{B} \leq h_{i'}[n] + (\mathcal{S} + \mathcal{B}) (3 - v_{i,i'}[n] - x_{i,m}[n] - x_{i',m'}[n]), \quad i, i' \in \mathbb{Z}_1^I, m, m' \in \mathbb{Z}_1^M : i \neq i', \mathcal{O}_{i,m} \cap \mathcal{R}_{i',m'} \neq \varnothing, \quad (1o)$$

$$g_i[n] + b_{i,m}[n] + \mathcal{B} \leq g_{i'}[n] + (\mathcal{S} + \mathcal{B}) (3 - w_{i,i'}[n] - x_{i,m}[n] - x_{i',m'}[n]), \quad i, i' \in \mathbb{Z}_1^I, m, m' \in \mathbb{Z}_1^M : i \neq i', \mathcal{O}_{i,m} \cap \mathcal{O}_{i',m'} \neq \varnothing, \quad (1p)$$

The objective function (1a) is a weighted mixture of the average power consumption and lightpath length, where $\bar{b}_{i,m}$, $\bar{z}_m$, and $\bar{x}_{i,m}$ denote time-averaged values of the corresponding variables $b_{i,m}[n]$, $z_m[n]$, and $x_{i,m}[n]$ over the intervals $n \in \mathbb{Z}_i^N$, respectively. The first and second terms reflect the total power consumption of the transponders and DCs. Request i has 4 transmit and receive transponders in its uplink and downlink lighpaths, each consuming a power of $(\mathcal{E}_i + \mathcal{F}_i\mathcal{K}_i)\sum_{m\in\mathbb{Z}_1^M} \bar{b}_{i,m}$, where the integer variable $b_{i,m}[n]$ shows the number of frequency slots assigned to request i if its VNF is processed in DC m. An active DC designated by the binary variable $z_m[n]$ consumes the power $\mathcal{P}_m\bar{z}_m + \mathcal{Q}_m \sum_{i\in\mathbb{Z}_1^I} \mathcal{W}\mathcal{K}_i\mathcal{Y}_i\bar{b}_{i,m}$, where $\sum_{i\in\mathbb{Z}_1^I} \mathcal{W}\mathcal{K}_i\mathcal{Y}_i\bar{b}_{i,m}$ is the total input traffic rate to the DC. Clearly, in full load processing conditions, the DC consumed the peak power of $\mathcal{P}_m + \mathcal{Q}_m\mathcal{C}_m$. The third term in the objective function (1a) stands for the path length, where the binary variable $x_{i,m}[n]$ indicates the deployment of the VNF of request i on the DC of node m. Specifically, minimizing this term would be beneficial in improving time-averaged delay and facilitating network load balancing. The weighing factors α, β, and γ adjust the importance of the three terms, and their values are determined based on the contextual requirements of the optimization. For instance, in an eco-friendly scenario, $\beta \gg \gamma, \alpha$ to force reducing the carbon footprint caused by DCs.

The time-averaged constraint (1b) guarantees queue stability by ensuring that the average served bits of each queue is equal to or greater than its corresponding average arrived bits [15], where $\bar{a}_i$ is the time-averaged of the arrived bits of request i while $\bar{s}_i$ is the time-averaged of the served bits

$$s_i[n] = \sum_{m\in\mathbb{Z}_1^M} \mathcal{T}\mathcal{W}\mathcal{K}_i b_{i,m}[n], \quad i \in \mathbb{Z}_1^I. \quad (2)$$

With constraint (1c), only a single DC processes each VNF. Particularly, if the VNF i cannot be processed in node m due to lack of DC or any legislative limitation, the binary constant $\mathcal{X}_{i,m} = 0$ excludes the impossible processing by constraint (1d). Note that legislative limitations can arise from physical factors such as geographical distance or security restrictions imposed by the client. A DC with the processing capacity $\mathcal{C}_m$ affords processing its assigned VNFs by constraint (1e). A DC is active and consumes power if it serves at least one VNF subject to constraint (1f).

In our work, $b_{i,m}[n]$ frequency slots are assigned to request i in batches instead of individually. This batch assignment approach guarantees the allocation of adjacent spectrum slots to each request, effectively satisfying the contiguity constraint. The spectrum continuity constraint in optical networks imposes that the assigned spectrum band must occupy the same portion of the spectrum without any alterations across all the links comprising the a lightpath. In our problem, in DC, data traverses from the optical domain to the electrical domain, undergoes processing, and is subsequently transmitted back in the optical domain. Consequently, it becomes necessary to ensure spectrum continuity independently for both the uplink and downlink lightpaths. Actually, rearrangement of the spectrum band in the downlink lightpath offers increased flexibility for the network RA and results in a lower blocking rate compared to scenarios, where both uplink and downlink lightpaths possess identical spectrum configurations. Constraint (1g) and (1h) locate the assigned spectrum bands of the uplink and downlink lightpaths inside the fiber working spectrum, where $h_i[n]$ and $g_i[n]$ denote the first frequency slot of the spectrum bands assigned to the uplink and downlink lightpaths, respectively. Constraint (1i) restrains the spectrum assignment beyond the fiber bandwidth $\mathcal{SW}$ as well as maximum bandwidth $\mathcal{D}_i$ of the transponder. The index of frequency slots assigned to each lightpath remains fixed across all its links. The constancy of the allocated frequency slot index across all links of a lightpath ensures that the same spectrum slots are utilized over all the links of a lightpath, satisfying the continuity constraint.

There are four auxiliary binary variables in constraints (1j)-(1l) that determine the relative location of the assigned spectrum band of any two requests in their uplink and downlink lightpaths. $t_{i,i'}[n]$ equals 1 if $h_i[n] \leq h_{i'}[n]$, and 0 otherwise. $u_{i,i'}[n]$, $v_{i,i'}[n]$ and $w_{i,i'}[n]$ are defined similar to $t_{i,i'}[n]$ to express $h_i[n] \leq g_{i'}[n]$, $g_i[n] \leq h_{i'}[n]$ and $g_i[n] \leq g_{i'}[n]$, respectively. Constraints (1m)-(1p) assures the non-overlapping and guard constraints and are only activated when there is at least one shared link between lighpaths. They are designed to aim at precluding the sharing of spectrum band between two requests over an optical link, while simultaneously allocating a guard band with $\mathcal{B}$ frequency slots between any two adjacent spectrum bands. Actually, the index of the start and end frequency slots of the lightpaths is selected based on the relative location of the assigned spectrum band of the lightpaths, in such a way that the guard band is also placed between them. The considered guard band facilitates

non-ideal filtering in the transponders and cross-connects. Further, the consideration of spectrum non-overlapping is contingent upon the sharing of lightpaths between two requests, indicating the presence of at least one common link between them.

C. PRACTICAL ALGORITHM

The optimization problem (1) exhibits a purely theoretical framework, aiming to achieve simultaneous and continuous optimization for all time intervals $n \in \mathbb{Z}_1^N$. However, practical implementation is hindered by the requirement of the future arrival information $\boldsymbol{a}[n]$, which is unattainable. In dynamic RA, the arrival information $\boldsymbol{a}[n]$ of the current and previous time slots is available. Even when the formulation is restricted to a subset of time intervals with available arrival information, the resulting complexity render it unsuitable for practical usage. Consequently, a causal and accurate approximation of the formulation is pursued using the Lyapunov method. The Lyapunov method is characterized by a series of interconnected, typically linear, optimization steps, linked by a recursive updating expression. This iterative approach can approximate the optimal solution for the stochastic optimization problems formulated in (1) [11]. Adopting the Lyapunov method, a near-optimal solution of the optimization (1) is effectively derived by solving a linear optimization with the objective function

$$\begin{aligned}\min_{\substack{\boldsymbol{b}[n],\boldsymbol{h}[n],\boldsymbol{g}[n],\boldsymbol{z}[n],\boldsymbol{x}[n]\\ \boldsymbol{t}[n],\boldsymbol{u}[n],\boldsymbol{v}[n],\boldsymbol{w}[n]}} \quad & \alpha \sum_{i\in\mathbb{Z}_1^I} 4(\mathcal{E}_i + \mathcal{F}_i\mathcal{K}_i) \sum_{m\in\mathbb{Z}_1^M} b_{i,m}[n] \\ & + \beta \sum_{m\in\mathbb{Z}_1^M} \left(\mathcal{P}_m z_m[n] + \mathcal{Q}_m \sum_{i\in\mathbb{Z}_1^I} \mathcal{W}\mathcal{K}_i\mathcal{Y}_i b_{i,m}[n]\right) \\ & + \gamma \sum_{i\in\mathbb{Z}_1^I}\sum_{m\in\mathbb{Z}_1^M} x_{i,m}[n]\mathcal{L}_{i,m} + \lambda \sum_{i\in\mathbb{Z}_1^I} q_i[n](a_i[n] - s_i[n])\end{aligned} \tag{3}$$

subject to the constraint (1c)-(1p) in each time interval n. Two consecutive optimizations of the intervals n and $n+1$ are related by the queue length update equation

$$q_i[n+1] = \max\{q_i[n] + a_i[n] - s_i[n], 0\}, \quad i \in \mathbb{Z}_1^I \tag{4}$$

with the initial queue length $q_i[0] = 0$, where $s_i[n]$ is given by (2). Note that queue lengths $q_i[n]$ and arrival bits $a_i[n]$ are known parameters in the linear optimization of each interval n. Furthermore, to expedite optimization process, the solution of the linear optimization in time interval n is utilized as a starting point for the optimization in the next time interval $n+1$.

As demonstrated in [9], the last term in (3), called service drift, provides an adjustable compromise, controlled by the Lyapunov penalty coefficient λ, between time-averaged minimization of the objective terms in (1a) and satisfaction of the queue stability constraint in (1b). It enforces the servicing of heavily backlogged queues, thereby rendering the inclusion of constraint (1b) unnecessary in the linear optimization of each step. However, while this term holds the queue lengths below a finite value, it may not be suitable for delay-sensitive requests that have strict requirements on backlog length. Certainly, by increasing λ, it becomes possible to impose stricter restrictions on the backlog length. However, this approach entails employing the same RA strategy for all requests, thereby disregarding the differentiation in QoS requirements they may have. To address this particular issue, we add the constraint

$$q_i[n] + a_i[n] - s_i[n] \le \mathcal{A}_i, \quad i \in \mathbb{Z}_1^I, \tag{5}$$

to the linear optimization solved in each time interval. Constraint (5) holds the backlog of queue i below the certain threshold $\mathcal{A}_i$ in each time interval n to expedite serving a delay-sensitive request. Clearly, $\mathcal{A}_i = \infty$ deactivates the constraint for delay-tolerant requests.

IV. EVALUATION

Within this section, we conduct a comprehensive evaluation of performance of the proposed method. The first subsection introduces the simulation setup and scenarios, while the second subsection presents and analyzes the obtained results.

A. SIMULATION SETUP

We have adopted the Nobel Germany topology for the performance validation, as delineated in Fig. 1, where the first 4 high-degree nodes are equipped with DC facilities. The numbers on the links indicate their length in km. The constants are set to $\mathcal{T}$= 5 s, $\mathcal{W}$ = 6.25 GHz, $\mathcal{S}$= 640, and $\mathcal{B}$ = 1 [14]. Assuming that similar transponders and DCs are deployed across the network, $\mathcal{E}_i$ = 75.6 W, $\mathcal{F}_i$ = 18.75 W.Hz/bps, $\mathcal{K}_i$ = 2 bps/Hz, $\mathcal{D}_i$ = 50 GHz, $\mathcal{P}_m$ = 56 kW, $\mathcal{Q}_m$ = 90.36 W/bps, and $\mathcal{C}_m$=1.4 Tbps [10], [15], [33]. I = 25 requests with random source and destination pairs are considered and served for $N = 200$ time intervals. The arrival bits $a_i[n]$ are assumed to be independent and identically distributed across all time slots, following a log-normal distribution characterized by a variable standard deviation σ, that is half the average $\mu = 2\sigma$. We assume a uniform distribution over [0.5,1] for $\mathcal{Y}_i$. Moreover, the Lyapunov penalty coefficient λ is set to 10^{-4} while the weighting factors α, β and γ are valued as 0.9, 0.9, and 0.1, respectively, since energy consumption is deemed more important for the simulation analysis. To solve the RA problem, we employ the combined capabilities of MATLAB and Gurobi [34].

We consider two versions of the proposed scheme and compare them to three static RA benchmark schemes. In static RA, the network configuration is planned to accommodate the anticipated peak of network traffic, and it remains unchanged even during fluctuations in network traffic. The anticipated amount of traffic peak in various static scenarios can vary based on the adopted policies. In a radical policy with a lower amount of anticipated traffic peak, there is an increased likelihood of encountering times with incoming traffic exceeding the anticipated peak and any excess traffic beyond the planned traffic peak remains unserved. To address this issue, considering more conservative static scenarios

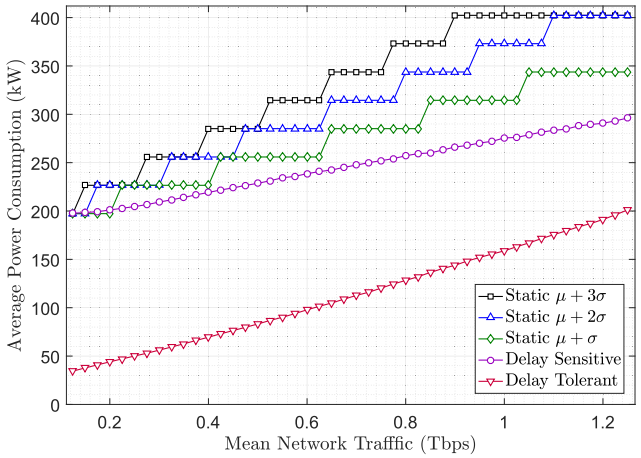

FIGURE 2. Average power consumption versus mean network traffic for the proposed and benchmark schemes.

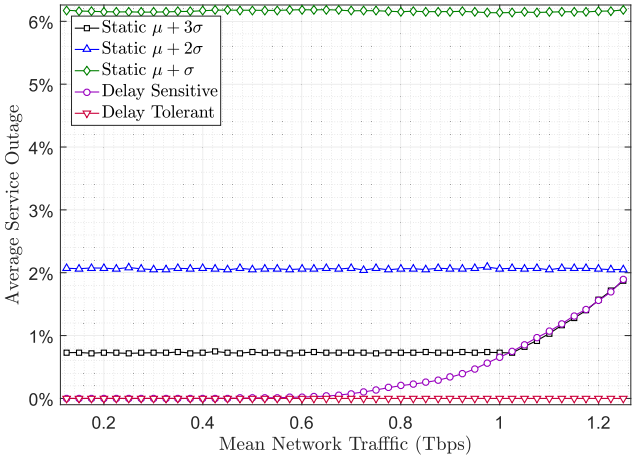

FIGURE 3. Average service outage versus mean network traffic for the proposed and benchmark schemes.

with higher amount of anticipated traffic peak can be a potential solution. However, there is an increased likelihood of allocating more resources than necessary. Consequently, this can lead to a waste of resources and an increase in energy consumption. Note that no queuing is assumed for the static schemes and therefore, the static schemes do not distinguish requests according to their QoS requirements to defer service to the delay-tolerant requests to future times in order to optimize energy consumption. In the first version of proposed scheme, called delay-tolerant scheme, resources are allocated without any limitations on the queue length, i.e. $\mathcal{A}_i = \infty$, while in the second version, named delay-sensitive scheme, the queue length is limited to $\mathcal{A}_i = 0.01\mu$. The static $\mu + 3\sigma$ benchmark scheme allocates resources for the static traffic peak $\mu + 3\sigma$, which is more conservative approach than the two other $\mu + 2\sigma$ and $\mu + \sigma$ benchmark schemes. As mentioned before, a more conservative RA strategy establishes the prerequisites for servicing during intervals of high traffic. However, this enhanced servicing capability comes at the expense of increased power consumption and the inefficient utilization of resources. If the RA of a scheme fails, the least number of bits are dropped such that the RA becomes feasible.

B. NUMERICAL RESULTS

In this subsection, we present and discuss the numerical results, shedding light on the effectiveness and efficiency of our proposed approach. Fig. 2 shows average power consumption versus mean network traffic for the proposed and benchmark schemes. The average power consumption is determined by calculating the average of the power consumption of DCs and transponders across all considered time intervals. The average power consumption in the static schemes is a discrete step-wise ascending function of the mean network traffic, where the discontinuities occurs when the fixed configuration does no longer afford the mean network traffic. For the proposed scheme, the average power consumption is a smooth linear-wise ascending function of the mean network traffic with a higher vertical intercept for the delay-sensitive version than that of the delay-tolerant version. The smoothness originates from the adaptive behaviour of the Lyapunov, where the network is reconfigured to adaptively follow the traffic dynamism. As the mean network traffic exceeds 1 Tbps, the average power consumption of the static $\mu + 3\sigma$ scheme remains unchanged, since failing to satisfy the network physical constraints and therefore, dropping part of the network traffic. The static schemes encounter these limitations at lower traffic levels compared to dynamic schemes, as they allocate more resources than necessary for the requests.

In the delay-tolerant version, the traffic transmission can be postponed by accumulating the bits in the queues until a power-efficient situation occurs resulting in a less power consumption at the cost of longer delay. The delay-sensitive version holds the average queuing delay around 0.8 ms. Obviously, the proposed scheme reduces average power consumption by 19% for delay-sensitive version and 52% for delay-tolerant version, when compared to the static $\mu + \sigma$ scheme at the mean network traffic of 1.05 Tbps. Furthermore, it achieves an average power consumption reduction of 31% and 59% for delay-sensitive and delay-tolerant versions, respectively, when compared to the static $\mu + 3\sigma$ scheme at the same mean network traffic.

Fig. 3 reports average service outage versus mean network traffic. The average service outage is defined as the ratio of the number of dropped bits required for a feasible RA to the number of arrived bits in a duration of time. Put differently, the average service outage can be calculated as the discrepancy between the number of arrived bits and the number of serviced bits divided by the number of arrived bits during the considered time intervals. The inability to achieve a feasible RA can arise from insufficient computation resources, inadequate communication resources, or a combination of both factors. In such cases, a potential remedy

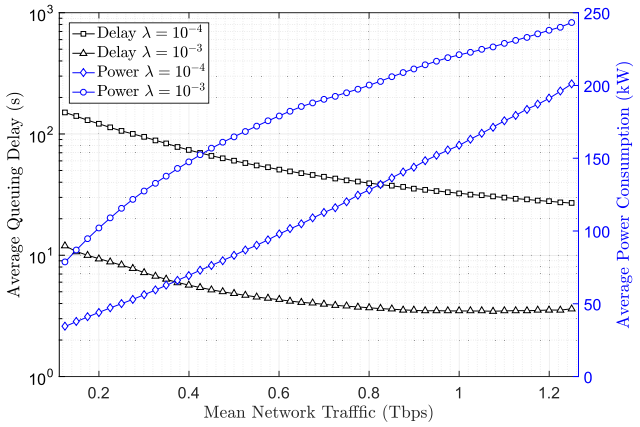

FIGURE 4. Average power consumption and average queuing delay versus mean network traffic for different Lyapunov penalty coefficients in delay-tolerant scheme.

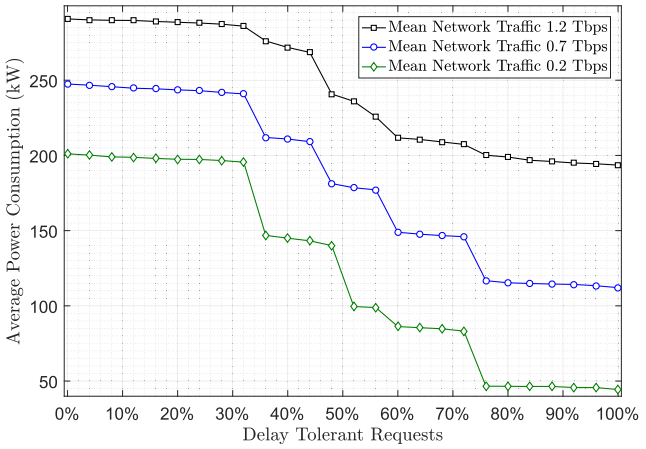

FIGURE 5. Average power consumption versus the percentage of delay-tolerant requests in the network for different mean network traffic.

involves discarding a portion of the received bits, thereby enabling the servicing of the requests while addressing the physical constraints. Here, the necessity for a joint RA approach becomes evident, as allocating DCs independently from routing and spectrum allocation can result in solution failure, even if sufficient resources are available. In other words, joint RA reduces service outage. The service outage in the static schemes remains almost constant as traffic increases no more than a specified value. The constant outage is due to the sudden rare peaks in the traffic which cannot be supported by the fixed configuration of the network. The service outage grows suddenly for a specified network traffic after the physical constraints are no longer satisfied for the main portion of the traffic. In the more conservative $\mu+3\sigma$ scheme, the resources are over-provisioned more so the sudden traffic peaks are handled with a lower service outage; however, the network capacity shortage appears faster at a lower network traffic due to high over-provisioning. The service outage is zero for the proposed schemes while the mean network traffic is less than 0.7 Tbps, where the delay-sensitive version imposes a negligible average service outage to keep the delay low. The delay-tolerant version provides zero average service outage even for high mean network traffic conditions since it can smooth sudden traffic peaks in the queues and serve them gradually over time.

In accordance with Little's law [11], the average number of customers in a stable system over an extended period is equivalent to the average arrival rate during that period multiplied by the average customer residence time. Consequently, the average queuing delay can be computed by dividing the average queue length by the average arrival rate during the considered time intervals multiplied by the duration of a time interval. Fig. 4 reports average power consumption and average queuing delay versus mean network traffic for different Lyapunov penalty coefficients, $\lambda = 10^{-4}$ and 10^{-3}. Observing the scenario where $\lambda = 10^{-3}$, it becomes evident that the average queuing delay has experienced a substantial reduction. However, this reduction has come at the cost of an increase in the average power consumption. Indeed, lower values for λ corresponds to a more pronounced improvement in the optimization objective, albeit at the expense of increased queue length and subsequent delays. Conversely, a higher value of λ can help reduce delay, but it compromises the optimality of power consumption. It is worth noting that the requirements for delay vary among network requests, contingent upon the application and the specific QoS criteria they demand. In actuality, a higher value of λ entails imposing greater power consumption across the network in an attempt to minimize delay for all requests, which is not desirable due to the fact that not all requests possess a sensitivity to delays. Consequently, by selecting a small value of λ and appropriately determining the queue length limit, it becomes possible to optimize both power consumption and satisfy the requested delay requirements. Subsequently, we will analyze the performance in scenarios where network requests vary in type, and resources are concurrently allocated to both delay-sensitive and delay-tolerant requests.

Fig. 5 presents the average power consumption versus the percentage of delay-tolerant requests in the network for various network traffic values of 0.2, 0.7, and 1.2 Tbps. The power consumption decreases as the percentage of the delay-tolerant requests increases since more bits can be accumulated in the queues and served collectively at a suitable time in some nominated DCs. Obviously, Fig. 5 reveals that how the proposed scheme can exploit QoS differentiation for power efficiency or equivalently, revenue enhancement.

V. CONCLUSION

An innovative dynamic RA scheme is proposed for power-efficient VNF management in an inter-DC EON. The proposed scheme not only locates suitable DCs for processing the submitted VNF requests but also considers the physical network limitations for routing and spectrum assignment in

the underlying EON. Since the proposed RA is formulated as a stochastic Lyapunov optimization, it inherits all the potential benefits of the Lyapunov optimization such as effective handling of time-average, instantaneous, and stability constraints. The scheme provides applicable tradeoffs between power efficiency and QoS commitments such that in a typical network scenario, it can reduce the power consumption by 50%.

The integration of renewable energy sources in DCs holds the potential to mitigate costs and minimize carbon footprint. Inherent uncontrollability and unpredictability of renewable energies, coupled with the high expense of energy storage, makes the integration of the renewable energy sources challenging [16]. However, the proposed RA platform seems capable enough to cope with these challenges. In fact, by prioritizing the service of delay-tolerant requests during periods when inexpensive and environmentally-friendly energy is abundant, DCs can effectively reduce both costs and carbon footprint through the utilization of renewable energy sources. Considering the significance of delay-based service differentiation in addressing this issue, future endeavors can consider modeling of renewable energy sources and their integration within the proposed RA problem. Furthermore, the generality and versatility of the proposed scheme allows more realistic constraints to be included in the RA, the fact that opens new aspects for future extensions of the research work.

REFERENCES

[1] C. Freitag, M. Berners-Lee, K. Widdicks, B. Knowles, G. S. Blair, and A. Friday, "The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations," *Patterns*, vol. 3, no. 8, Aug. 2022, Art. no. 100576.

[2] N. Jones, "The information factories," *Nature*, vol. 561, no. 7722, p. 163, 2018.

[3] A. Tarafdar, S. Sarkar, R. K. Das, and S. Khatua, "Power modeling for energy-efficient resource management in a cloud data center," *J. Grid Comput.*, vol. 21, no. 1, p. 10, Mar. 2023.

[4] J. Sun, Y. Zhang, F. Liu, H. Wang, X. Xu, and Y. Li, "A survey on the placement of virtual network functions," *J. Netw. Comput. Appl.*, vol. 202, Jun. 2022, Art. no. 103361.

[5] H. U. Adoga and D. P. Pezaros, "Network function virtualization and service function chaining frameworks: A comprehensive review of requirements, objectives, implementations, and open research challenges," *Future Internet*, vol. 14, no. 2, p. 59, Feb. 2022.

[6] H. Yu, Z. Chen, G. Sun, X. Du, and M. Guizani, "Profit maximization of online service function chain orchestration in an inter-datacenter elastic optical network," *IEEE Trans. Netw. Service Manag.*, vol. 18, no. 1, pp. 973–985, Mar. 2021.

[7] A. Khatiri and G. Mirjalily, "A cost-efficient, load-balanced and fragmentation-aware approach for deployment of VNF service chains in elastic optical networks," *Comput. Commun.*, vol. 188, pp. 156–166, Apr. 2022.

[8] J. G. Herrera and J. F. Botero, "Resource allocation in NFV: A comprehensive survey," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 518–532, Sep. 2016.

[9] M. Hadi and E. Agrell, "Iterative configuration in elastic optical networks," in *Proc. Int. Conf. Opt. Netw. Design Model. (ONDM)*, May 2020, pp. 1–3.

[10] M. Hadi and E. Agrell, "Joint power-efficient traffic shaping and service provisioning for metro elastic optical networks," *J. Opt. Commun. Netw.*, vol. 11, no. 12, pp. 578–587, Dec. 2019.

[11] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*, vol. 1. San Rafael, CA, USA: Morgan & Claypool, 2010.

[12] X. Fu, F. R. Yu, J. Wang, Q. Qi, and J. Liao, "Dynamic service function chain embedding for NFV-enabled IoT: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 507–519, Oct. 2019.

[13] D. Li, P. Hong, K. Xue, and J. Pei, "Virtual network function placement considering resource optimization and SFC requests in cloud datacenter," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 7, pp. 1664–1677, Jul. 2018.

[14] M. Hadi, M. R. Pakravan, and E. Agrell, "Dynamic resource allocation in metro elastic optical networks using Lyapunov drift optimization," *J. Opt. Commun. Netw.*, vol. 11, no. 6, pp. 250–259, Jun. 2019.

[15] F. S. Vajd, M. Hadi, C. Bhar, M. R. Pakravan, and E. Agrell, "Dynamic joint functional split and resource allocation optimization in elastic optical fronthaul," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 4, pp. 4505–4515, Dec. 2022.

[16] G. Zhang, S. Zhang, W. Zhang, Z. Shen, and L. Wang, "Distributed energy management for multiple data centers with renewable resources and energy storages," *IEEE Trans. Cloud Comput.*, vol. 10, no. 4, pp. 2469–2480, Oct. 2022.

[17] A. K. Mishra, J. L. Hellerstein, W. Cirne, and C. R. Das, "Towards characterizing cloud backend workloads: Insights from Google compute clusters," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 37, no. 4, pp. 34–41, Mar. 2010.

[18] Y. Wang, Q. Yang, and X. Guo, "A resource and task scheduling based multi-objective optimization model and algorithms in elastic optical networks," *Sensors*, vol. 22, no. 24, p. 9579, Dec. 2022.

[19] B. Li and Z. Zhu, "GNN-based hierarchical deep reinforcement learning for NFV-oriented online resource orchestration in elastic optical DCIs," *J. Lightw. Technol.*, vol. 40, no. 4, pp. 935–946, Feb. 15, 2022.

[20] Y. Ma, W. Liang, Z. Xu, and S. Guo, "Profit maximization for admitting requests with network function services in distributed clouds," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 5, pp. 1143–1157, May 2019.

[21] J. Pei, P. Hong, M. Pan, J. Liu, and J. Zhou, "Optimal VNF placement via deep reinforcement learning in SDN/NFV-enabled networks," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 263–278, Dec. 2019.

[22] G. Sun, G. Zhu, D. Liao, H. Yu, X. Du, and M. Guizani, "Cost-efficient service function chain orchestration for low-latency applications in NFV networks," *IEEE Syst. J.*, vol. 13, no. 4, pp. 3877–3888, Dec. 2019.

[23] L. Ruiz, R. J. D. Barroso, I. De Miguel, N. Merayo, J. C. Aguado, R. De La Rosa, P. Fernandez, R. M. Lorenzo, and E. J. Abril, "Genetic algorithm for holistic VNF-mapping and virtual topology design," *IEEE Access*, vol. 8, pp. 55893–55904, 2020.

[24] Z. Xu and Z. Zhu, "On establishing and task scheduling of data-oriented VNF-SCs in an optical DCI," *J. Opt. Commun. Netw.*, vol. 14, no. 3, pp. 89–99, 2022.

[25] S. Li, B. Li, and Z. Zhu, "On the game-theoretic analysis of dynamic VNF service chaining in edge-cloud EONs," *J. Lightw. Technol.*, vol. 41, no. 10, pp. 2940–2952, May 15, 2023.

[26] T. Miyamura and A. Misawa, "Joint optimization of optical path provisioning and VNF placement in vCDN," *Opt. Switching Netw.*, vol. 49, May 2023, Art. no. 100740.

[27] A. Montazerolghaem, M. H. Yaghmaee, and A. Leon-Garcia, "Green cloud multimedia networking: NFV/SDN based energy-efficient resource allocation," *IEEE Trans. Green Commun. Netw.*, vol. 4, no. 3, pp. 873–889, Sep. 2020.

[28] G. Sun, R. Zhou, J. Sun, H. Yu, and A. V. Vasilakos, "Energy-efficient provisioning for service function chains to support delay-sensitive applications in network function virtualization," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6116–6131, Jul. 2020.

[29] M. Chen, Y. Sun, H. Hu, L. Tang, and B. Fan, "Energy-saving and resource-efficient algorithm for virtual network function placement with network scaling," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 1, pp. 29–40, Mar. 2021.

[30] Y. Mu, L. Wang, and J. Zhao, "Energy-efficient and interference-aware VNF placement with deep reinforcement learning," in *Proc. IFIP Netw. Conf.*, Jun. 2021, pp. 1–9.

[31] W. Wei, H. Gu, A. Pattavina, J. Wang, and Y. Zeng, "Optimizing energy and spectrum efficiency of virtual optical network embedding in elastic optical networks," *Opt. Switching Netw.*, vol. 37, May 2020, Art. no. 100568.
[32] R. Zhu, W. Zhang, P. Wang, J. Chen, J. Wang, and S. Yu, "Energy-efficient graph reinforced vNFC deployment in elastic optical inter-DC networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 2, pp. 1591–1604, Mar. 2024.
[33] L. Nonde, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Energy efficient virtual network embedding for cloud networks," *J. Lightw. Technol.*, vol. 33, no. 9, pp. 1828–1849, May 1, 2015.
[34] *Gurobi Optimizer Reference Manual*. Accessed: Feb. 2024. [Online]. Available: https://www.gurobi.com/documentation/current/refman/index.html

MAHTA AHMADI received the B.Sc. degree in electrical engineering from the University of Isfahan, in 2018, and the M.Sc. degree in communication systems from the Sharif University of Technology, in 2021. Her research interests include optical communication networks and machine learning.

MOHAMMAD HADI received the Ph.D. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2018. He was a Postdoctoral Researcher with the Chalmers University of Technology, in 2019. Since 2021, he has been an Assistant Professor of communication systems with the Sharif University of Technology. His main research interest includes resource allocation in optical and radio networks.

MOHAMMAD REZA PAKRAVAN (Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the University of Tehran, Tehran, Iran, in 1990, and the M.Sc. and Ph.D. degrees in electrical engineering from the University of Ottawa, Canada, in 1992 and 2000, respectively. In 2001, he joined the Department of Electrical Engineering, Sharif University of Technology, where he is currently an Associate Professor. He is also the Director of the Data Networks Research Laboratory and the Networking Group with theAdvanced Communication Research Institute. His research interests include optical communication systems and networks, wireless communication networks, and data networking algorithms and protocols. He has received many awards for his academic and engineering achievements. Some of the notable awards are the IEEE Neal Shepherd Memorial Best Propagation Paper Award from the IEEE Vehicular Technology Society, in 2001; two International Khwarizmi Awards from the Iranian Research Organization for Science and Technology, in 2014 and 2019, and the Entrepreneurship Award of the Year from the IEEE Iran Section, in 2020. He has also been recognized as a top engineer of the year by the Iranian Academy of Science for his outstanding contributions to the development of electrical and computer engineering, in 2018.