Proceedings of CECNet 2022
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220568

Cost-Efficient Service Function Chaining with CoMP Zero-Forcing Beamforming in Mobile Edge Networks

Yuan GAO [a],Hai FANG [a], and Kan WANG [b,1]
[a] *Xi'an Institute of Space Radio Technology, Xi'an 710100, China*
[b] *Xi'an University of Technology, Xi'an 710048, China*

Abstract. As two promising paradigms in next generation cellular systems, service function chaining (SFC) and mobile edge computing (MEC) have attracted great interests, and would bring more delay-sensitive services to users in proximity. Nevertheless, owing to time-varying channel conditions and finite server resources, the SFC deployment in edge networks is nontrivial. In this work, leveraging both the coordinated multiple points (CoMP)-based zero-forcing beamforming and ℓ_p $(0 < p < 1)$ norm-based successive convex approximation (SCA) methods, we investigate the SFC deployment in the edge. First, under the constraints of processing and link capacity, transmission power and service function ordering, we build a mixed-integer nonlinear programming (MINLP)-based cost optimization problem, to minimize both the flow and power cost. Then, using the dirty paper-based CoMP zero-forcing beamforming method, the interference among SFSs is canceled, and the original problem is recast as a interference-free one. Next, the ℓ_p $(0 < p < 1)$ norm-based SCA method works to produce a series of convex subproblem, the iterative solution of which is proved to converge to optimal solution of original one at a linear convergence rate. Finally, numerical results are used to validate proposed method, showing that the wireless resource management deserves special interests in the SFC deployment in edge networks.

Keywords. service function chain, mobile edge computing, successive convex approximation, interference cancellation

1. Introduction

Next generation cellular systems are expected to experience the architecture innovation with network function virtualization (NFV), whereby service functions can be virtually instantiated in commodity servers, rather than deployed in specialized hardwares [1]. The cellular network operator can thus on-demand expand the architecture flexibly by supplementing more commodity servers and then instantiating service functions as software [2]. Besides, virtual network functions (VNFs) can be cascaded in a logical sequence through which the packet traverses [2]. Hence, such service function chaining (SFC) facilitates the flexible deployment of diversified services [3].

[1]Corresponding Author: Yuan Gao, Xi'an Institute of Space Radio Technology, Xi'an 710100, China; E-mail: gaoy199034@126.com.

There have been many contributions made to the SFC or VNF deployment in traditional networks, e.g., large-scale networks [4], central cloud networks [5], and content delivery networks [6]. Meanwhile, as another one viable paradigm, mobile edge computing (MEC) is utilized to push caching and computation resources close to devices, without obtaining resources from the central cloud [1]. Integrating MEC with SFC would yield benefits. That is, by deploying SFCs in the edge, users can obtain services nearby, thereby promoting the computation-intensive or delay-sensitive services. Further, edge servers are directly programmable with NFV, and provide richer VNF combinations to build more complicated services [1,3].

Yet, deploying SFC in the edge poses more challenges in resource-constrained edge environment. First, the computing, caching, and networking resources in edge servers are typically insufficient than those of traditional core and cloud networks, and thus the resource allocation needs to be more regulated to save cost. Second, the fast fading of wireless channels in the edge asks for interference cancellation among SFCs [7]. Notice that, in the SFC of edge networks, the cached content needs to be visited before other service functions and thus caching dominates the SFC [8], while baseband unit (BBU) BBU provides signal processing ability, sends packets to users with remote radio heads (RRHs), and thus terminates the SFC [9].

Although there have been some outstanding works in the SFC deployment of edge networks, the channel conditions as well as interference cancellation schemes are not fully investigated [3,5,6]. Thus, assisted by recent advents in Gaussian codebook-based successive dirty paper encoding [10,11], we try to design an interference-free service function chaining method in the edge. First, we investigate the coordinated multiple points (CoMP) transmission-based zero-forcing method to cancel the interference across SFCs, thereby building a cost-efficient SFC deployment problem under the constraints of processing and link capacity. Then, ℓ_p $(0 < p < 1)$ norm penalty is added to the mixed-integer nonlinear programming (MINLP)-based original problem, to build one equivalent counterpart with same optimal solutions. Finally, the successive convex approximation (SCA) method is utilized to solve the counterpart, approaching zero penalty at a linear convergence rate. The contribution can be summarized as follows:

- We investigate the cost-efficient SFC deployment problem in edge networks, by leveraging the CoMP-based zero-forcing beamforming method for interference-free SFCs. In the CoMP system, the effective zero-forcing beamforming matrices are designed to cancel the interference across SFCs, and then the cost is divided into edge server operating and wireless transmission power cost, under the constraints of processing and link capacity.
- To solve the MINLP-based SFC deployment, we then use the ℓ_p $(0 < p < 1)$ norm-based SCA method to recast the problem as a series of convex subproblem. What is important, the equivalence on optimal solutions between original problem and its relaxed counterpart is proved, showing that the SCA converges to optimum with zero penalty at a linear rate.
- Finally, we evaluate the proposed method via simulation results, showing that the cost minimization benefits from not only the SFC routing across inter-connected edge servers, but also careful wireless power regulation. Meanwhile, proposed method shows its vitality with the increasing of SFC and service function number.

The rest is organized as follows. The network model as well as problem formulation are presented in Section 2. The zero-forcing beamforming is used in Section 3, and SCA

method is presented in Section 4. Simulation results are described to show the effectiveness of proposed method in Section 5, followed by conclusions in Section 6.

2. Network Model and Problem Formulation

For clarity, all mathematical notations throughout the paper are listed in Table 1.

Table 1. Mathematical Notations

$(\cdot)^\top$	Matrix or vector transpose
$(\cdot)^H$	Matrix or vector conjugate transpose
Upper-case bold letters	Matrices
Lower-case bold letters	Vectors
$\mathrm{diag}\{\cdot\}$	Diagonalization
$\succeq$ or $(\preceq)$	Componentwise inequality between vectors Matrix inequality between symmetric matrices
$\mathrm{Tr}(\cdot)$	Sum of diagonal elements

Consider one MEC network composed of several MEC clouds, each of which includes multiple inter-connected edge servers and a group of RRHs [12].Take one MEC cloud as a instance, and each edge server $n \in \mathcal{N} = \{1,2,\cdots,N\}$ in the MEC cloud can provide diversified VNFs. In particular, the edge server is inter-linked mutually via X2+ links to facilitate the interaction, and meanwhile empowers one RRH via vBBU. Thus, also let n be the index of RRH attached to server n interchangeably. Further, denote $\mathcal{M} = \{1,2,\cdots,M\}$ as the user set, and suppose that each user asks for one service at a time. Each service flow is endowed with one SFC, namely, any packet of the flow has to traverse through the sequenced service functions scattered in the MEC cloud, before the wireless transmission from the RRHs to user $m \in \mathcal{M}$. Denote such SFC associated with user m as $\mathcal{F}(m) = (f_1^m \to \cdots \to f_l^m \to \cdots \to f_{\ell_m}^m)$.

2.1. Signal Model and Data Rate

Each coefficient of user-RRH pair (m,n)'s channel coefficient matrix $\mathbf{H}_{m,n} \in \mathbb{C}^{L_n \times L_m}$ is drawn from the circularly symmetric complex Gaussian (CSCG) distribution. The received signal at each user m becomes as

$$\mathbf{u}_m = \mathbf{H}_m^H \mathbf{W}_m \mathbf{s}_m + \sum_{k<m} \mathbf{H}_m^H \mathbf{W}_k \mathbf{s}_k + \sum_{k>m} \mathbf{H}_m^H \mathbf{W}_k \mathbf{s}_k + \mathbf{n}_m, \tag{1}$$

where $\mathbf{H}_m = [\mathbf{H}_{m,1}; \mathbf{H}_{m,2}; \cdots; \mathbf{H}_{m,N}]^H \in \mathbb{C}^{L \times L_m}$ is the channel matrix between RRHs and user m with $L = \sum_{n=1}^N L_n$, $\mathbf{W}_m \in \mathbb{C}^{L \times d_m}$ denotes the beamforming matrix from RRHs to user m with $d_m \le L_m$ as the data stream number, $\mathbf{s}_m \in \mathbb{C}^{d_m}$ is the sampled Gaussian random codebook with zero mean and covariance $\mathbf{I}_{d_m}$ [10], and $\mathbf{n}_m$ acts as the additive Gaussian while noise with covariance $\mathbf{I}_{L_m}$. Then, via the Gaussian codebook-based successive dirty paper encoding, the second term in right-hand side (RHS) of (1) is removed [10,11], and data rate R_m turns into

$$R_m = \log_2 \frac{\left|\mathbf{I}_{L_m} + \sum_{k \geq m} \mathbf{H}_m^H \mathbf{W}_k \mathbf{W}_k^H \mathbf{H}_m\right|}{\left|\mathbf{I}_{L_m} + \sum_{k > m} \mathbf{H}_m^H \mathbf{W}_k \mathbf{W}_k^H \mathbf{H}_m\right|}, \tag{2}$$

which must be no less than its data rate requirement $R_{m,\text{th}}$, i.e.,

$$R_m \geq R_{m,\text{th}}. \tag{3}$$

2.2. Capacity Constraints

Assume that each virtual machine (VM) is capable of offering one VNF instance [13], and let $\mathcal{N}_f$ denote the set of servers that could afford VNF f. To make each service function only get served on one node at a time, we have

$$\sum_{n \in \mathcal{N}_{f_l^m}} y^m_{f_l^m,n} = 1, \forall f_l^m \in \mathcal{F}(m), \forall m, \tag{4}$$

where $y^m_{f_l^m,n} \in \{0,1\}$ indicates whether or not that the l-th service function for user m is served on n. Then, define $x_{f,n} \in \{0,1\}$ to describe if server n opens VNF f, and we have

$$y^m_{f,n} \leq x_{f,n}, \forall f, \forall m, \forall n, \tag{5}$$

implying that only when $x_{f,n} = 1$ holds that can $y^m_{f,n}$ take 1; otherwise, $y^m_{f,n} = 0$ holds.

Next, let $R_{m,\text{th}}$ be the data rate requirement for user m, let $\mathcal{F}_n$ be the set of VNFs readily embedded in n, and assume that processing one bit expends one processing unit in the VNF. Also, each VNF's being served flow rate must be no greater than its processing capacity $\mu_{f,n}$, i.e.,

$$\sum_m y^m_{f,n} R_{m,\text{th}} \leq \mu_{f,n}, \forall f \in \mathcal{F}_n, \forall n. \tag{6}$$

Meanwhile, the link capacity $\mu_{n,s}$ between n and s also needs to satisfy

$$\sum_m \sum_{f_l^m \in \mathcal{F}(m)} z^m_{f_l^m,n,s} R_{m,\text{th}} \leq \mu_{n,s}, \forall n, s, \tag{7}$$

where variable $z^m_{f_l^m,n,s} \in \{0,1\}$ shows whether or not that f_l^m and f_{l+1}^m are serially offered by server n and s. Note in both (6) and (7) that the fixed data rate $R_{m,\text{th}}$ is set in the SFC routing across edge serves, since the backhaul data rate is typically assumed to be no greater than wireless data rate [14].

Finally, we have

$$z^m_{f_l^m,n,s} \geq y^m_{f_l^m,n} + y^m_{f_{l+1}^m,s} - 1, \forall f_l^m \in \mathcal{F}(m), \forall m, \forall n, s, \tag{8}$$

implying that $z^m_{f_l^m,n,s}$ can take 1 only when both $y^m_{f_l^m,n} = 1$ and $y^m_{f_{l+1}^m,s} = 1$ hold.

2.3. Problem Formulation

It is desired that service functions in one SFC are centralized in as few nodes as possible, to minimize the flow cost. Each edge server is nevertheless with restricted processing capacity, only accommodating finite service functions, and inevitably resulting in the flow routing cost. Further, not only the flow routing across servers, but also the VNF provisioning in edge servers do incur cost. Thus, the power cost includes both the wireless transmission power in the air and service function sustaining power in edge servers.

We aim to minimize the total service flow cost plus power cost in the edge. First, define $\sum_m \mathrm{Tr}\left(\mathbf{W}_m\mathbf{W}_m^H\right)$ as the wireless transmission power. Further, let $P^m_{f^m_{\ell_m},n}$ and $P^m_{n,\mathrm{th}}$ be the vBBU processing and wireless transmission power budget for server n and RRH n to serve user m, separately, and $y^m_{f^m_{\ell_m},n} \in \{0,1\}$ denotes whether or not user m accesses server n's vBBU. The beamforming matrix $\mathbf{W}_m$ should satisfy

$$\mathrm{Tr}\left(\mathbf{A}_n\mathbf{W}_m\mathbf{W}_m^H\right) \le P^m_{f^m_{\ell_m},n} y^m_{f^m_{\ell_m},n} + P^m_{n,\mathrm{th}}, \tag{9}$$

with

$$\mathbf{A}_n = \mathrm{diag}\{\underbrace{0,\cdots,0}_{\sum_{n'=1}^{n-1} L_{n'}},\underbrace{1,\cdots,1}_{L_n},\underbrace{0,\cdots,0}_{\sum_{n'=n+1}^{N} L_{n'}}\}. \tag{10}$$

Next, denote $P_{f,n}$ and $P^m_{f,n}$ as the constant power consumption for server n to sustain VNF f and for n to supply user m with f, respectively, and the total cost becomes as

$$\begin{aligned} C = &\sum_{n,s}\sum_{m}\sum_{f^m_l \in \mathscr{F}(m)} z^m_{f^m_l,n,s} R_{m,\mathrm{th}} \\ &+\eta\left(\sum_n\sum_{f\in\mathscr{F}_n} x_{f,n}P_{f,n} + \sum_n\sum_m\sum_{f^m_l\in\bar{\mathscr{F}}(m)} y^m_{f^m_l,n}P^m_{f^m_l,n} + \sum_m \mathrm{Tr}\left(\mathbf{W}_m\mathbf{W}_m^H\right)\right), \end{aligned} \tag{11}$$

where η is a positive trade-off factor between power and flow cost. In particular, the four different terms respectively denote service flow, supply power to sustain VNFs in servers, power consumption on service function provisioning for users, and wireless transmission power at the RRHs. Note in (11) that, we define that $\sum_m \mathrm{Tr}\left(\mathbf{W}_m\mathbf{W}_m^H\right)$ also comprises the vBBU provisioning power. Thus, $\bar{\mathscr{F}}(m) = \mathscr{F}(m)\backslash\{f^m_{\ell_m}\}$ holds in the third term.

Till now, considering the above link and service instance capacity constraints as well as data rate requirement and transmission power budget, the network cost minimization can be formulated as

$$\begin{aligned} \mathscr{P}_0: \quad &\min_{\substack{[\mathbf{W}_m],\{x_{f,n}\}\\ \{y^m_{f^m_l,n}\},\{z^m_{f^m_l,n,s}\}}} \quad C \\ &\text{s.t.} \quad (3)-(9) \end{aligned} \tag{12}$$

where constraint (8) together with the objective (11) implies that $z^m_{f^m_l,n,s} = 1$ makes sense only when both $y^m_{f^m_l,n}$ and $y^m_{f^m_{l+1},s}$ take 1, and the minimization nature of (12) prevents the optimal solution from taking 1 on $z^m_{f^m_l,n,s}$, when only one of $y^m_{f^m_l,n}$ and $y^m_{f^m_{l+1},s}$ equals 1.

Remark that $\mathscr{P}_0$ is actually one $\mathscr{NP}$-hard problem, since it is easy to prove that $\mathscr{P}_0$ reduces to the uncapacitated facility location (UFL) one (which is $\mathscr{NP}$-hard) [8].

3. CoMP-Assisted Beamforming to Cancel Interference

We try to utilize the zero-forcing beamforming in [10,11] at RRHs to cancel the interference on user m from other users $m+1$ to M. First, stack the channel matrices for all users as $[\mathbf{H}_m]_{m=1}^{M} \in \mathbb{C}^{L\times\sum_{m=1}^{M} L_m}$, which is column full rank, and its QR decomposition is

$$[\mathbf{H}_m]_{m=1}^{M} = [\mathbf{Q}_m]_{m=1}^{M+1} \begin{bmatrix} \mathbf{R}_{1,1}, & \cdots & \mathbf{R}_{1,M} \\ & \ddots & \vdots \\ & & \mathbf{R}_{M,M} \\ \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}, \tag{13}$$

where $[\mathbf{Q}_m]_{m=1}^{M+1} \in \mathbb{C}^{L\times L}$ is the unitary matrix, and $\mathbf{R}_{m,m} \in \mathbb{C}^{L_m\times L_m}$ is the full-rank upper triangular matrix. From (13), we have $\mathbf{H}_m = \sum_{k=1}^{m} \mathbf{Q}_k \mathbf{R}_{k,m}$.

Thus, define $\mathbf{S}_m \in \mathbb{C}^{\sum_{k=m}^{M+1} L_k\times d_m}$ as the weight matrix, and the zero-forcing beamforming matrix for user m can be designed as $\mathbf{W}_m = [\mathbf{Q}_k]_{k=m}^{M+1} \mathbf{S}_m$, giving rise to

$$\mathbf{H}_m^H \mathbf{W}_k = \mathbf{0}, \forall k > m, \tag{14}$$

whereby the third term in (1) is also canceled.

Note that $\mathbf{H}_m^H \mathbf{W}_m = \left[\mathbf{R}_{m,m}^H, \mathbf{0}\right] \mathbf{S}_m$, and thus only the first L_m rows of $\mathbf{S}_m$ could affect the data rate. The beamforming matrix of user m can reduce to $\mathbf{W}_m = \mathbf{Q}_m \bar{\mathbf{S}}_m$, with $\bar{\mathbf{S}}_m \in \mathbb{C}^{L_m\times d_m}$ and $\mathbf{S}_m = \left[\bar{\mathbf{S}}_m; *\right]$.

Afterwards, we define the positive semi-definite beamforming matrix as $\mathbf{\Sigma}_m = \mathbf{R}_{m,m}^H \bar{\mathbf{S}}_m \bar{\mathbf{S}}_m^H \mathbf{R}_{m,m}$, and the data rate (2) becomes as

$$R_m = \log_2 |\mathbf{I}_{L_m} + \mathbf{\Sigma}_m| \geq R_{m,\text{th}}, \tag{15}$$

Notice in (15) that, the interference among SFSs is canceled now, due to (14).

Till now, (9) can be recast as

$$\text{Tr}\left(\mathbf{A}_n \mathbf{W}_m \mathbf{W}_m^H\right) = \text{Tr}\left(\underbrace{\mathbf{R}_{m,m}^{-1} \mathbf{Q}_m^H \mathbf{A}_n \mathbf{Q}_m \mathbf{R}_{m,m}^{-H} \mathbf{\Sigma}_m}_{\bar{\mathbf{A}}_{m,n}} \right) \leq P^m_{f^m_{\ell_m},n} y^m_{f^m_{\ell_m},n} + P^m_{n,\text{th}}. \tag{16}$$

Finally, $\mathscr{P}_0$ is equivalently recast as

$$\begin{aligned} \mathscr{P}_1: \quad & \min_{\substack{[\mathbf{\Sigma}_m], \{x_{f,n}\} \\ \{y^m_{f^m_l,n}\}, \{z^m_{f^m_l,n,s}\}}} \quad C \\ & \text{s.t.} \quad (4)-(8), (15), (16) \end{aligned} \tag{17}$$

4. Norm Penalty-based Iterative Solving

To solve $\mathscr{P}_1$, we would like to relax $\{x_{f,n},\}$ $\{y^m_{f^m_l,n}\}$ and $\{z^m_{f^m_l,n,s}\}$. Yet, $\mathscr{P}_1$ is generally not equal to its relaxed counterpart, since the optimal solution in relaxed $\mathscr{P}_1$ cannot be insured to be binary. Thus, besides relaxation, we also add some ℓ_p $(0 < p < 1)$ norm penalty terms to the objective, thereby forcing the relaxed variables binary.

For notational simplicity, let $\mathbf{x} = \{x_{f,n}\}$, $\mathbf{y} = \{y^m_{f^m_l,n}\}$ and $\mathbf{z} = \{z^m_{f^m_l,n,s}\}$ hereinafter, and then (4) can be recast as

$$\left\|\mathbf{y}^m_{f^m_l}\right\|_1 = 1, \forall f^m_l, \forall m, \tag{18}$$

with $\mathbf{y}^{m,t}_{f^m_l} := \left[y^{m,t}_{f^m_l,1}, y^m_{f^m_l,2}, \cdots, y^m_{f^m_l,|\mathscr{N}_{f^m_l}|}\right]^\top$. Then, one instance problem is as follows:

$$\begin{aligned} \min_{\mathbf{y}^m_{f^m_l}} \quad & \left\|\mathbf{y}^m_{f^m_l} + \delta\mathbf{1}\right\|_p^p = \sum_{n\in\mathscr{N}_{f^m_l}} \left(y^{m,t}_{f^m_l,n} + \delta\right)^p \\ \text{s.t.} \quad & \left\|\mathbf{y}^m_{f^m_l}\right\|_1 = 1, \\ & \mathbf{0} \preceq \mathbf{y}^m_{f^m_l} \preceq \mathbf{1}, \end{aligned} \tag{19}$$

where δ takes one arbitrarily small positive value [15]. (19)'s optimal point is precisely 0-1 valued, i.e., only one server from $\mathscr{N}_{f^m_l}$ can afford f^m_l for user m, and thus (19)'s optimal value is $c_{\delta,f^m_l} = \left(|\mathscr{N}_{f^m_l}| - 1\right)\delta^p + (1+\delta)^p$.

Next, in addition to relaxation, we also add one L_p penalty term (with regulating parameter σ) onto $\mathscr{P}_1$'s objective, reformulating $\mathscr{P}_1$ as

$$\begin{aligned} \mathscr{P}_{1-L_p}: \min_{[\mathbf{\Sigma}_m],\mathbf{x},\mathbf{y},\mathbf{z}} \quad & C + \sigma \underbrace{\sum_m \sum_{f^m_l\in\mathscr{F}(m)} \left(\left\|\mathbf{y}^m_{f^m_l} + \delta\mathbf{1}\right\|_p^p - c_{\delta,f^m_l}\right)}_{P_\delta(\mathbf{y})} \\ \text{s.t.} \quad & (4)-(8), (15), (16) \\ & \left\|\mathbf{y}^m_{f^m_l}\right\|_1 = 1, \mathbf{0} \preceq \mathbf{y}^m_{f^m_l} \preceq \mathbf{1}, \mathbf{0} \preceq \mathbf{z}^m_{f^m_l} \preceq \mathbf{1} \end{aligned} \tag{20}$$

with $\mathbf{z}^m_{f^m_l} = \left[z^m_{f^m_l,n,s}\right]_{n=1,s=1}^{|\mathscr{N}_{f^m_l}|,|\mathscr{N}_{f^m_l}|}$ and σ as a penalty parameter.

The asymptotic optimality of $\mathscr{P}_{1-L_p}$ to $\mathscr{P}_1$ with increasing σ is already established in [15, Theorem 3], and we next prove its linear convergence rate to zero penalty by proposing one assumption.

Assumption 1. *C is L_C-Lipschitz continuous and the maximum ℓ_2 norm distance between two solutions in $\mathscr{P}_{1-L_p}$ is bounded by a constant R.*

Under Assumption 1, we investigate the convergence rate of $\mathscr{P}_{1-L_p}$ with the increasing $\{\sigma_v\}$, where v is the iteration index. Denote $P_\delta(\mathbf{y})$ in the v-th iteration as $P_\delta(\mathbf{y}^v)$.

Theorem 1. *Assume Assumption 1 is satisfied. When the penalty parameter is updated as* $\sigma_{v+1} = \tau\sigma_v$ *with* $\tau > 1$ *and* $\sigma_0 = 1$*, the penalty term sequence* $\{P_\delta(\mathbf{y}^v)\}$ *satisfies*

$$P_\delta(\mathbf{y}^v) \leq \frac{L_C R}{\tau^v}. \tag{21}$$

Proof. From the Appendix (proof of Theorem 3) in [15], we have

$$\sigma_v(P_\delta(\mathbf{y}^v) - P_\delta(\mathbf{y}^{v+1})) \leq C_{v+1} - C_v, \tag{22}$$

where C_v denotes the value of C in (20) at the v-th iteration for simplicity. Next, set $\sigma_{v+1} = \tau\sigma_v$, $\sigma_0 = 1$ and $\tau > 1$, and the following holds as

$$\tau^v(P_\delta(\mathbf{y}^v) - 0) = \tau^v P_\delta(\mathbf{y}^v) \leq C^* - C^v, \tag{23}$$

since the penalty term vanishes at optimal solutions, and C^* is the optimal value in $\mathscr{P}_1$ with binary $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

Next, from Assumption 1, we further have

$$C^* - C_v \leq L_C R. \tag{24}$$

Till now, by substituting (24) into (23), we obtain

$$P_\delta(\mathbf{y}^v) \leq \frac{L_C R}{\tau^v}. \tag{25}$$

□

From (21), it follows that $\{P_\delta(\mathbf{y}^v)\}$ approaches 0 at a linear convergence rate. Till now, although the asymptotic optimality of $\mathscr{P}_{1-L_p}$ to $\mathscr{P}_1$ is presented, $\mathscr{P}_{1-L_p}$ is non-convex. We then have to leverage the SCA method [15], i.e., obtain the first-order Taylor approximation of penalty. As such, $P_\delta(\mathbf{y})$ can be upper bounded by $P_\delta(\mathbf{y}) \leq P_\delta(\mathbf{y}^v) + \nabla_{\mathbf{y}} P_\delta(\mathbf{y}^v)^\top(\mathbf{y} - \mathbf{y}^v)$, where the optimal point in the preceding SCA iteration is taken as $\mathbf{y}^v$, while $\nabla_{\mathbf{y}} P_\delta(\mathbf{y}^v)$ is the derivative of $P_\delta(\mathbf{y})$ on $\mathbf{y}^v$.

Finally, in the $(v+1)$-th SCA iteration, $\mathscr{P}_{1-L_p}$ is recast as one convex problem as

$$\begin{aligned} \mathscr{P}_{1-S}: \min_{[\mathbf{\Sigma}_{m,t}],\mathbf{y},\mathbf{z}} \quad & C_t + \sigma_{v+1}\nabla_{\mathbf{y}} P_\delta(\mathbf{y}^v)^\top \mathbf{y} \\ \text{s.t.} \quad & (4)-(8),(15),(16) \\ & \left\|\mathbf{y}_{f_l^m}^m\right\|_1 = 1, \mathbf{0} \preceq \mathbf{y}_{f_l^m}^m \preceq \mathbf{1}, \mathbf{0} \preceq \mathbf{z}_{f_l^m}^m \preceq \mathbf{1} \end{aligned} \tag{26}$$

which is readily solved via lots of algorithms, e.g., the Lagrangian dual [16].

Remark that, $\mathscr{P}_{1-S}$ is indeed a one-slot SFC deployment problem. The dynamic SFC deployment in time-varying scenarios have also received lots of interests, via either deep reinforcement learning [17,18] or graph neural networks [19,20], which is yet beyond the scope of this work, and would be left in future work.

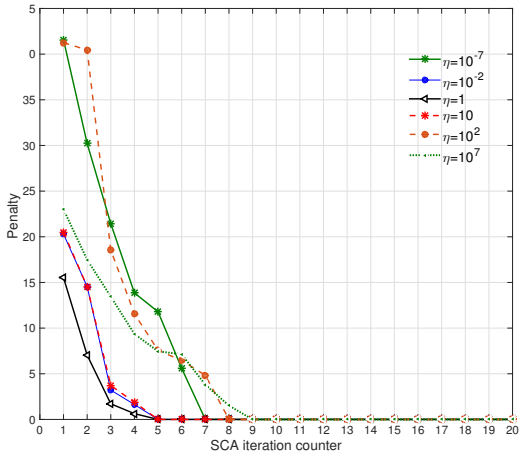

Figure 1. Convergence performance of penalty-based SCA iteration. (7 edge server and 6 SFCs exist in the system, and each SFC includes 4 service functions.)

5. Simulation Results

To simulate edge networks, the random network model [21] is used to create connections between any two edge servers n and s, with the connecting probability $\Pr(n,s) = \beta \exp\left(\frac{-d(n,s)}{\alpha R_d}\right)$, where $d(n,s)$ is the ℓ_2 norm distance, R_d denotes the maximum distance among all (n,s)-pairs, and both $\alpha \in [0,1]$ and $\beta \in [0,1]$ are control parameters. 7 edge servers with 8 antennas and 6 users with 3 antennas are distributed in a rectangular coordinate grid 300 m $\times$ 300 m. The fast fading is Rayleigh distributed, the log-normal shadowing fading per user is 8 dB, noise power spectrum density is -174 dBm/Hz, the path loss is $\mathrm{PL}(\mathrm{dB}) = 32.45 + 10\log_{10}(d(m))$, and the maximum transmission power per RRH is 46 dBm.

Five different types services coexist, and their data rate requirements are separately 0.5 Mbps, 1 Mbps, 4 Mbps, 5 Mbps, and 10 Mbps. A total of seven different VNFs exist in edge networks, and each server have randomly accommodated three of them. The processing capacity of each VNF ranges from 5 Mbps to 15 Mbps, the link capacity between any two linked edge servers ranges from 10 Mbps and 50 Mbps, while both $P_{f,n}$ and $P^m_{f_l^m,n}$ in (11) range from 1 W to 4 W.

For performance comparison, three other benchmarks are presented as random SFC routing + CoMP zero-forcing beamforming (abbreviated as random zero-forcing), optimal SFC routing + unicast beamforming (abbreviated as optimal unicast), and random SFC routing + unicast beamforming (abbreviated as random unicast).

5.1. Numerical Results

Fig. 1 shows the convergence behavior of SCA iteration under different settings of η. We set error to zero duality as 0.01. All settings converge to zero penalty within 10 iterations, thus verifying the linear convergence rate of proposed ℓ_p $(0 < p < 1)$ penalty-based SCA iteration.

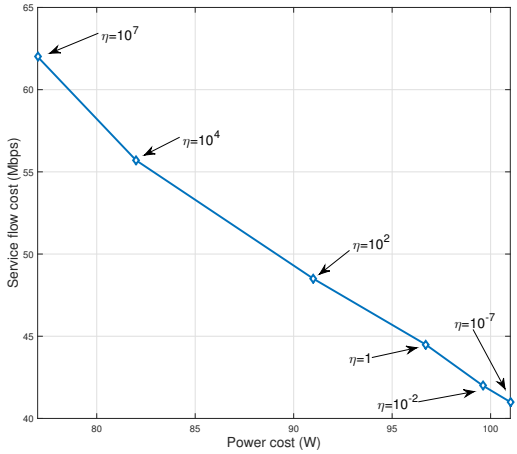

Figure 2. Tradeoff between power and service flow cost. (7 edge server and 6 SFCs exist in the system, and each SFC includes 4 service functions.)

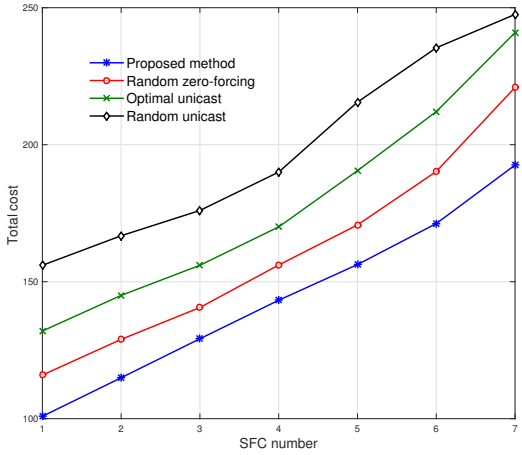

Figure 3. The impact of SFC number on total cost. (7 servers exist in the system, each SFC includes 4 service functions, and $\eta = 10^2$.)

Fig. 2 shows the tradeoff curve between power and service flow cost, under different η in (11). When $\eta = 10^7$, proposed method almost puts all emphases on power cost, and thus $\mathscr{P}_0$ reduces to the power minimization problem. The flow cost reaches above 60 Mbps, since no emphases are put on it. Conversely, when $\eta = 10^{-7}$, almost all emphases are imposed on service flow cost, $\mathscr{P}_0$ reduces to one service flow minimization problem, and thus power cost reaches above 100 W. Fig. 2 can provide empirical values of η for practical SFC deployment in next generation cellular systems.

Fig. 3 compares different methods regarding SFC number, in terms of total cost. As

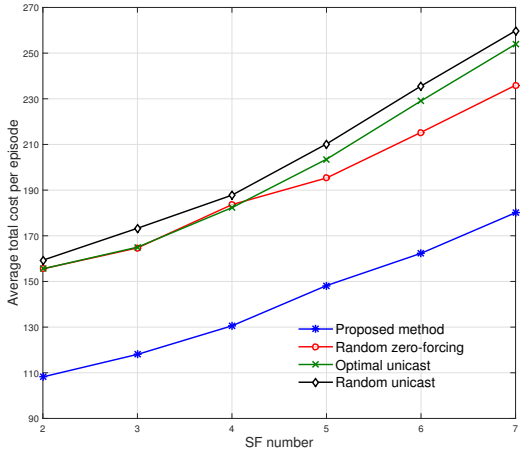

Figure 4. The impact of service function number on total cost. (7 servers exist in the system, each SFC includes 4 service functions, and $\eta = 10^2$.)

shown in Fig. 3, proposed method obtains the least system cost, due to its advantage in both SFC routing and wireless transmission exploitation. In addition, random zero-forcing gets less cost than optimal unicast, since the setting $\eta = 10^3$ gives more emphases on power cost, and random zero-forcing obtains less power cost via CoMP-based beamforming. More especially, proposed method respectively surpasses other three benchmarks on average by 38.1%, 55.3%, and 70.1%.

Fig. 4 shows the impact of service function number per SFC. As shown in Fig. 4, proposed method always has the least cost, due to its advantage in both optimal routing and optimal beamforming. In addition, random zero-forcing surpasses optimal unicast, since $\eta = 10^2$ setting would incur less power cost for the former. More especially, proposed method respectively surpasses other three benchmarks on average by 25.2%, 27.5%, and 30.1%.

6. Conclusion and Future Work

We investigated the cost-efficient SFC deployment in the mobile edge networks. First, both the flow routing and power sustaining costs were incorporated, and a SFC deployment problem was formulated to minimize the total cost, under the constraints of wireless interference, processing and link capacity, transmission power, as well as service function ordering. Then, to cancel the interference among SFCs, a dirty paper-based zero-forcing beamforming technique was used for interference-free SFC deployment. Next, to solve the MINLP-based original problem, the original one was relaxed, and the ℓ_p $(0 < p < 1)$ norm penalty term was added onto the objective to make the optimal solutions of relaxed one also binary. Finally, simulation results verified the convergence behavior of proposed SCA method, as well as the importance of transmission power allocation in the SFC deployment in edge networks. In future work, we will study the dynamic SFC deployment in time-varying scenarios, using reinforcement learning.

References

[1] Cabrera JA, Fitzek FH, Hanisch S, Itting SA, Zhang J, Zimmermann S, Strufe T, Simsek M, Fetzer CW. Intelligent networks. In: Fitzek FH, Li SC, Speidel S, Strufe T, Simsek M, Reisslein M, editors. Tactile Internet. San Diego, CA: Academic Press; 2021. p. 131–149.

[2] Cui L, Tso FP, Jia W. Federated service chaining: Architecture and challenges. IEEE Communications Magazine. 2020 Mar; 58(3): 47–53.

[3] Song S, Lee C, Cho H, Lim G, Chung J. Clustered virtualized network functions resource allocation based on context-aware grouping in 5G edge networks. IEEE Transactions on Mobile Computing. 2020 May; 19(5): 1072–1083.

[4] Luo Z, Wu C, Li Z, Zhou W. Scaling geo-distributed network function chains: A prediction and learning framework. IEEE Journal On Selected Areas In Communications. 2019 Aug; 37(8): 1838–1850.

[5] Alhussein O, Do PT, Ye Q, Li J, Shi W, Zhuang W, Shen X, Li X, Rao J. A virtual network customization framework for multicast services in NFV-enabled core networks. IEEE Journal On Selected Areas In Communications. 2020 Apr; 38(6): 1025–1039.

[6] Dieye M, Ahvar S, Sahoo J, Ahvar E, Glitho R, Elbiaze H, Crespi N. CPVNF: Cost-efficient proactive VNF placement and chaining for value-added services in content delivery networks. IEEE Transactions on Network and Service Management. 2018 Jun; 15(2): 774–786.

[7] Li J, Wang R, Wang K. Service function chaining in industrial Internet of Things with edge intelligence: A natural actor-critic approach. IEEE Transactions on Industrial Informatics. 2022, to appear, online. doi:10.1109/TII.2022.3177415.

[8] Zheng G, Tsiopoulos A, Friderikos V. Optimal VNF chains management for proactive caching. IEEE Transactions on Wireless Communications. 2018 Oct; 17(10): 6735–6748.

[9] Harutyunyan D, Shahriar N, Boutaba R, Riggio R. Latency and mobility–aware service function chain placement in 5G networks. IEEE Transactions on Mobile Computing. 2022 May; 21(5): 1697–1709.

[10] Dong Y, Zhang H, Li J, Yu FR, Guo S, Leung VCM. An online zero-forcing precoder for weighted sum-rate maximization in green CoMP systems. IEEE Transactions on Wireless Communications. 2022 Sept; 21(9): 7566–7581.

[11] Lu HF. Optimal sum rate-fairness tradeoff for MIMO downlink communications employing successive zero forcing dirty paper coding. IEEE Communications Letters. 2021 Mar; 25(3): 783–787.

[12] Zhou Z, Wu Q, Chen X. Online orchestration of cross-edge service function chaining for cost-efficient edge computing. IEEE Journal On Selected Areas In Communications. 2019 Aug; 37(8): 1866–1880.

[13] Pu L, Jiao L, Chen X, Wang L, Xie Q, Xu J. Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks. IEEE Journal On Selected Areas In Communications. 2018 Aug; 36(8): 1751–1767.

[14] Tao M, Chen E, Zhou H, Yu W. Content-centric sparse multicast beamforming for cache-enabled cloud RAN. IEEE Transactions on Wireless Communications. 2016 Sept; 15(9): 6118–6131.

[15] Zhang N, Liu YF, Farmanbar H, Chang TH, Hong M, Luo ZQ. Network slicing for service-oriented networks under resource constraints. IEEE Journal On Selected Areas In Communications. 2017 Nov; 35(11): 2512–2521.

[16] Boyd S, Vandenberghe L. Convex optimization. New York: Cambridge University Press; 2004. 716 p.

[17] Qi S, Li S, Lin S, Saidi MZ, Chen K. Energy-efficient VNF deployment for graph-structured SFC based on graph neural network and constrained deep reinforcement learning. In Proceedings of 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS); 2021 Sept 08-10; Tainan, Taiwan, China: IEEE Press; p. 348-353.

[18] Wang T, Fan Q, Li X, Zhang X, Xiong Q, Fu S, Gao M. DRL-SFCP: Adaptive service function chains placement with deep reinforcement learning. In Proceedings of International Conference on Communications (ICC); 2021 June 14-23; Montreal, Canada: IEEE Press; p. 1-6.

[19] Jiang W. Graph-based Deep Learning for Communication Networks: A Survey. Computer Communications. 2022 Mar; 185:40-54.

[20] Heo DN, Lange S, Kim HG, Choi H. Graph neural network based service function chaining for automatic network control. In Proceedings of 21st Asia-Pacific Network Operations and Management Symposium (APNOMS); 2020 Sept 22-25; Daegu, South Korea: IEEE Press; p. 7-12.

[21] Newman M. The structure and function of networks. Computer Physics Communications. 2022 Aug; 147(1): 40–45.