Scaling and Offloading Optimization in Pre-CORD and Post-CORD Multi-Access Edge Computing

Widhi Yahya, Eiji Oki, *Fellow, IEEE*, Ying-Dar Lin, *Fellow, IEEE*, and Yuan-Cheng Lai

***Abstract*—In 5G networks, multi-access edge computing (MEC) can be embedded into an access network (AN-MEC) and a core network (CN-MEC), which composes a two-tier MEC architecture for better scalability. In pre-Central Office Re-architected as a Data center (pre-CORD), AN-MECs are connected to a single but distant CN-MEC through Central Offices (COs). Disaggregation and virtualization of 5G network functions push CN-MEC into COs, which is known as post-CORD. Post-CORD has more CN-MECs closer to User Equipments than pre-CORD. In this work, we propose a scalable two-tier, multi-site, multi-server MEC architecture for pre-CORD and post-CORD. To adjust capacity and traffic allocation in such a distributed two-tier architecture, we integrate scaling and offloading with the objective of minimizing total capacity cost subject to the latency satisfaction percentage constraints, and solve the problem by Latency Aware Two-Phase Iterative Optimization (LA-TPIO). The results show that post-CORD with ten CN-MEC sites requires 30% less capacity than pre-CORD in satisfying 95% of URLLC traffic. Post-CORD utilizes about 48-77% less AN-MEC capacity than pre-CORD because post-CORD's aggregated but close-enough CN-MEC sites are ideal for serving URLLC traffic. Under heavy hotspot traffic, post-CORD's vertical and horizontal offloading percentages are 72% and 28%, respectively, while pre-CORD's are 99% and 1%, which means post-CORD introduces more horizontal offloading because it has links between not only AN-MEC sites but also CN-MEC sites to accommodate hotspot traffic.**

***Index Terms*—Scalable MEC, offloading, scaling, pre-CORD, post-CORD, optimization**

I. Introduction

5G mobile networks are an improvement on 4G in terms of its peak data rate, which is an order of magnitude faster than current LTE networks, short round trip latency (in units of ms), a high number of connected devices, and energy efficiency. 5G networks fulfill the requirements of human-to-human (H2H) and device-to-device (D2D) communications. This technology brings a massive number of devices with various communication models and services to the Internet. In 5G, the various communication uses can be categorized into three essential services: Enhanced Mobile Broadband (eMBB), Ultra-Reliable and Low Latency Communications (URLLC), and Massive Machine Type Communications (mMTC). eMBB is a common 5G broadband service that supports stable connections with high peak data rates such as video streaming, social media, and file transfers. The URRLC supports low-latency communication with a small data payload, and it also needs reliable data transmission in some cases [1] [2]. The mMTC supports an extensive number of devices (IoT characteristics) with intermittently active and small data payloads.

5G network architecture separates control and data plane functions. This separation enables the development of virtualized control and data planes of a 5G network, which can be executed in Multi-Access Edge Computing (MEC) servers. MEC servers can be located in every corner of a 5G network to provide computational resources for hosting infrastructure functions and a service-provider applications. The MEC deployment in 5G network transforms the network as communication and computational infrastructure. MEC servers in 5G networks are expected to reduce Internet backbone traffic and provide computation capacity approximate User Equipments (UEs) to accommodate low latency services [3][4]. Furthermore, integration between cloud and edge computing, called federated computing, can also be carried out given various computing needs.

The authors of [5] and [6] developed MEC prototypes without taking into account scalability. Since arrival traffic rates can dramatically increase and overload a single MEC server, a scalable MEC architecture which can adjust its capacity, is needed. Scalable MEC has been investigated in [7]–[22] with no consideration for hotspot traffic. Hotspot traffic arises in sporting events or a music concerts and can overload an MEC site. In this paper, a two-tier scalable MEC architecture with multi-site and multi-server is proposed. Two-tier refers to the access network MEC (AN-MEC) and core network MEC (CN-MEC) sites. AN-MEC is collocated with the base station in the access network of a cellular system. CN-MEC is an MEC site that is located at the core network of a cellular system. Two-tier is considered because it has more places for placing resources than a single tier AN-MEC for accomodating high arrival traffic, including hotspot traffic.

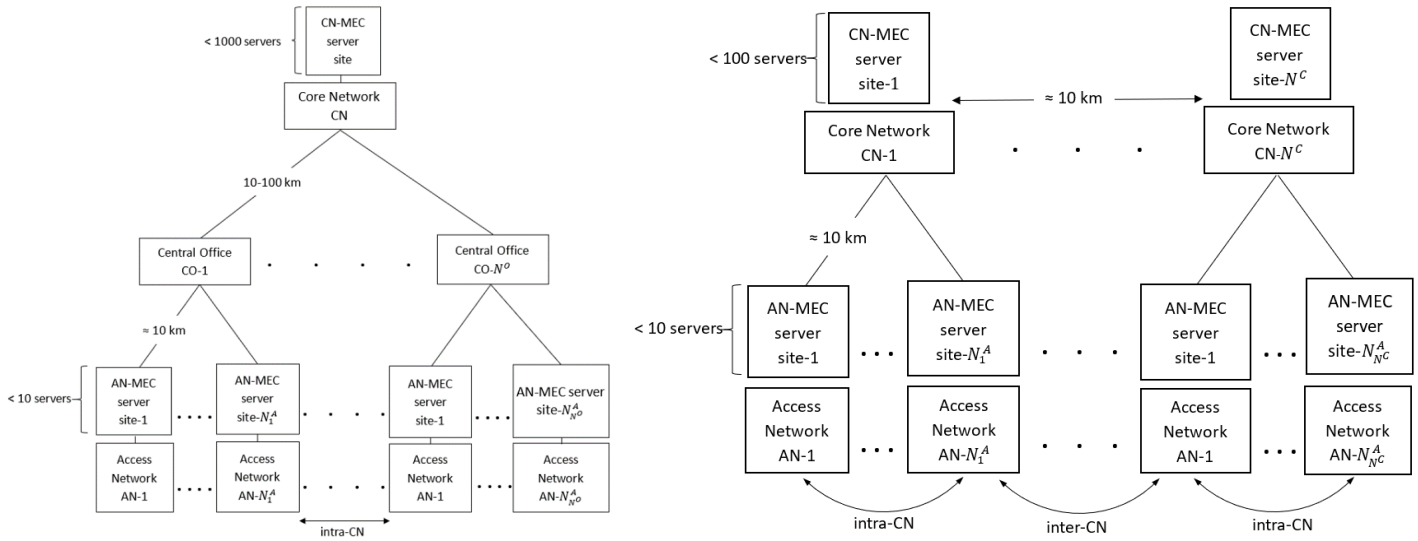
In two-tier MEC architecture, multiple MEC servers can be placed in an access network (AN) using ETSI's bump-the-wire scenario, and be attached to the Packet Data Network Gateway (P-GW) local breakout of the core network (CN). In past years, a CN has been located far from, often hundreds of kilometers from some base stations. A CN is connected to tens of Central Offices (CO), which consists of standard switching equipment and connects to some base stations, within a wide geographical area such as

W. Yahya and Y.-D. Lin are with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan (e-mail: wyahya@cs.nctu.edu.tw; ydlin@cs.nctu.edu.tw).

E. Oki is with the Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: oki@i.kyoto-u.ac.jp).

Y.-C. Lai is with the Department of Information Management, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: laiyc@cs.ntust.edu.tw).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management



(a) Pre-CORD architecture

(b) Post-CORD architecture

Fig. 1: Two-tier MEC architectures.

a province. A country may have one or more CNs, depending on its size, to cover its entire territory. As shown in Fig. 1a, this architecture is known as pre-CORD (Central Office Re-architected as Datacenter) architecture. In a 5G network, because of core network dis-aggregation, the user plane function (UPF) can be placed in the CO, which enables packet steering for redirecting traffic from UEs to MEC applications. This change manifests Central Office Re-architected as Datacenter (CORD) [23], as shown in Fig. 1b, that changes the COs role from communication infrastructure to computing infrastructure. CORD or post-CORD architecture provides more capacity and better latency than pre-CORD because it has MEC servers in some COs which are only tens of kilometers from a base station.

Two-tier MEC architecture with multi-site and multi-server provides a large amount of computing capacity (servers) in a 5G network. However, activating high capacity all the time can increase operational expenditure (OpEx). The management plane function takes place to optimize provider income by minimizing system capacity, while still satisfying latency percentage constraint. A latency percentage constraint describes the minimum amount of traffic that satisfies the latency constraint. The management plane function scales the capacity up when incoming traffic increases, and scales the capacity down when idle capacity exists. This mechanism is usually carried out automatically and is known as autoscaling.

This study addresses the management plane problem in two-tier MEC architecture by integrating the scaling and offloading algorithms in a Latency Aware Two-Phase Iterative Optimization (LA-TPIO). The superiority of LA-TPIO compared to previous algorithms [8]–[22], because it specifically addresses the minimization of resources into control and management planes which adjust offloading ratios and capacity iteratively. The offloading mechanism is carried out by UEs to minimize UE's cost [24] or by an overloaded MEC server/site to minimize latency which is discussed in this paper. In such an offloading mechanism, we extend the TPIO [13] to perform both vertical and horizontal offloading which, considers the latency of the destination site's latency. The offloading mechanism becomes a short-term solution for satisfying latency percentage constraints of hotspot traffic associated with currently allocated capacity. The scaling mechanism is the long-term solution for satisfying arrival traffic by adjusting the MEC system capacity.

Given arrival traffic rates, the system determines the capacity and derives the horizontal and vertical offloading ratios. The objective is to minimize system capacity while satisfying the latency satisfaction percentage constraint. To broaden our knowledge, we also investigated: (1) Capacity allocation of Pre-CORD vs. post-CORD for serving some traffic scenarios; (2) The effect of latency satisfaction percentage threshold; (3) The effect of the latency constraint (4) Uniform vs. non-uniform traffic and (5) LA-TPIO performance comparison.

The remainder of this paper is organized as follows. Section II reviews previous works in scaling and offloading in MEC. Section III defines the system architectures and problem formulation. The solution algorithm, LA-TPIO, is described in Section IV. Section V presents the simulation and results, and Section VI concludes the work.

II. Related Work

Scalability is the property of a system to accommodate a growing number of requests by dynamically allocating capacity. Cloud is an example of a scalable system with huge capacity in a centralized location [12][16]. A centralized location is a long distance from some UEs and not ideal for tight latency services. MEC extends the cloud

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

TABLE I: Scaling and offloading in multi-access edge computing.

Paper	Targeted Networks	MEC Arch.		Objective	Constraint	Offloading direction	Approach	# of compared services	Hotspot
		Two-tier	Multi-site						
[8]	Device/edge	X	O	Minimize cost	Latency	Vertical	SCPA	1	X
[9]		X	X	Minimize UE energy consumption	Energy	Vertical	Heuristic	1	X
[10]		X	X	Minimizing energy	Latency	Vertical	Lyapunov optimization	1	X
[11]		X	X	Minimize cost	Latency, capacity	Vertical	DQN	1	X
[12]	Device/edge/cloud	X	O	Minimize edge-cloud traffic	Latency	Vertical	Heuristic (ENORM)	1	X
[13]		X	O	Minimize service capacity of device, edge and cloud	Latency	Vertical	TPIO	2	X
[14]		X	O	Maximize profit	Latency	Vertical	SMBO	1	X
[15]		X	O	Minimize energy	Latency, capacity	Vertical	THOA	1	X
[16]	Edge/cloud	X	X	Minimize energy	Enegy	Vertical	Online Learning	1	X
[17]	Edge	X	X	Maximize utility	Latency	Vertical	Stackelberg game	1	X
[18]		X	X	Minimize cost	Latency	Vertical	CSAO	1	X
[19]		X	O	Minimize cost	Energy	Vertical	MOACO	1	X
[20]		X	O	Minimizing latency and energy	-	Horizontal	DTORA	1	X
[21]		X	X	Minimize active containers	-	Vertical	Fuzzy	1	X
[22]		X	O	Minimize task offloading	Latency,cost	Vertical	MINLP	1	X
Our	Edge(AN-MEC/CN-MEC)	O	O	Minimize capacity	Latency satisfaction percentage	Vertical, Horizontal	LA-TPIO	3	O

by having distributed resources close to the UEs. Some researchers looked at a scalable MEC for serving multiple UEs. The authors of [9]–[11], [16]–[18], [21] considered an MEC site for accomodating UEs in a particular area. To serve UEs in a wide area, the authors of [8], [12]–[15], [19], [20], [22] investigated a one-tier, multi-site MEC system with capacities behind base stations. We propose a two-tier MEC architecture that places the capacities behind base stations and the cellular system's core networks for better scalability. We also investigated a two-tier MEC system for pre-CORD and post-CORD architecture.

Dynamic capacity allocation is required in a scalable MEC system to scale the system up or down depending on the number of incoming requests over a time interval, so as to minimize OpEx. System information such as the CPU, memory load, and server's latency, owned by the orchestrator, are considered for scaling decisions. Such scaling is a part of the management plane, which is a long-term solution to hotspot traffic. Some studies [9]–[11], [14], [15], [18], [20], [22] have integrated scaling with an offloading mechanism. Offloading is a short-term solution for hotspot traffic because it is part of the control plane that runs in the order of seconds. We integrate control and management plane solutions into LA-TPIO for minimizing capacity in pre-CORD and post-CORD architecture.

The objectives of some studies were driven by implementing scaling and offloading in an MEC system. The authors of [9], [10], [15], and [16] considered energy minimization for longer UE's battery life, and [22] maximized a UE's battery life by maximizing the offloaded task. Multi-objective optimization work that considers energy and latency was detailed in [20]. The authors of [8], [11], [18], and [19] minimized MEC costs by including computation and communication costs. Paper [12] proposed a framework for minimizing edge-cloud traffic that can minimize an Internet service provider's communication cost, while [14] considered offloading and resource allocation to maximize the profit. Active resource minimization was proposed in [13] and [21]. Such active capacity minimization can also minimize OpEx of a service provider.

Our objective was to minimize system capacity with latency satisfaction percentage as a constraint. As shown in Table I, most of previous works used fixed delay as a constraint, which is difficult to achieve for all traffic in a real system because traffic arrivals could fluctuate over time. On the other hand, the management plane adjusts resources in the cycle of minutes to hours to minimize the cost. All of the previous proposals did not take into account hotspot traffic that may appear in some of AN-MEC sites. Only [13] considered more than one service, which can be categorized as URLLC and mMTC. Most of the studies offloaded arrival traffic vertically to another MEC site or a cloud system. In this study, we consider hotspot traffic and enabling horizontal offloading between AN-MEC sites and CN-MEC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

sites. We also investigated some services with different latency constraints which categorize into URLLC, mMTC, and eMBB.

III. System Architectures and Problem Formulation

This section describes the pre-CORD and post-CORD system architectures, gives problem statements, problem descriptions, and provides some notations in Table II.

A. System architectures

This paper tackles the management plane problem in two-tier, multi-site and multi-server MEC architecture. Its objective is to minimize system capacity while satisfying latency requirement. The term 'site' is used to represent an (AN or CN) MEC server cluster in an area. Two-tier MEC architecture is implemented in both pre-CORD and post-CORD architectures. In pre-CORD architecture, the CO, which lie between AN and CN, are only communication hubs and do not have any computation capacity. In pre-CORD architecture, a telecom company could have hundreds of ANs, tens of COs, and only one CN to cover a province. Pre-CORD's CN can be quite a distance away from an base station. On the other hand, in post-CORD architecture, the MEC servers can be placed in the CO, which hosts a CN, since the CO is equipped with UPF to carry out packet steering [9]. Thus, post-CORD architecture has much closer CN-MEC servers than pre-CORD architecture.

Pre-CORD architecture, as shown in Fig. 2, has N^{O} COs. each CO connects N^{A} AN-MEC sites to a CN-MEC site. N_i^{A} denotes the N^{A} AN-MEC sites that are connected to the ith CO. Pre-CORD and post-CORD use index i to represent CO and CN-MEC, respectively. In post-CORD architecture, COs are replaced by N^{C} CN-MEC sites, as shown in Fig. 3. To simplify the simulation process, multiple servers in an MEC site are represented as total capacity μ, with units in packets/second. During initialization, both architectures utilize the same server capacity, which is denoted by μ_i^{C} for CN-MEC capacity and $\mu_{i,j}^{\mathrm{A}}$ for AN-MEC capacity. We assume that the capacity links between AN-MEC and CN-MEC are sufficiently large to accommodate a large volume of traffic. A cellular company would implement a terabit optical link [25] in their mid-haul and backhaul networks to accommodate the proliferation of mobile devices [26]. By using a terabit optical link, the backhaul link's latency is four orders of magnitude less than radio access network (RAN) latency and is five orders of magnitude less than the defined latency constraint for URLLC service (1 ms). The RAN and backhaul link latencies are 0.5 ms and 0.01 µs, respectively, to serve 8000 packets, with each packet being 10 Kb. Since the effect of backhaul link latency is very small, so it is negligible. The propagation delay between AN-MEC and CN-MEC in pre-CORD architecture is equal to $D_{i,j}^{\mathrm{AO}} + D_i^{\mathrm{OC}}$ while in post-CORD architecture equal to $D_{i,j}^{\mathrm{AC}}$, because AN-MEC and CN-MEC are connected directly. $D_{i,j}^{\mathrm{AO}}$, D_i^{OC}, and $D_{i,j}^{\mathrm{AC}}$ are calculate by dividing $d_{i,j}^{\mathrm{AO}}$, d_i^{OC}, and $d_{i,j}^{\mathrm{AC}}$ by speed of light.

Traffic is generated and directed to each AN-MEC site at rate $\lambda_{i,j}$. i and j denote the destination of the arrival traffic in the jth AN-MEC site of the ith CO in pre-CORD or the ith CN-MEC in post-CORD. We consider three types of traffic, which are represented as URLLC, eMBB, and mMTC traffic. Each traffic has a latency constraint, which is denoted as L.

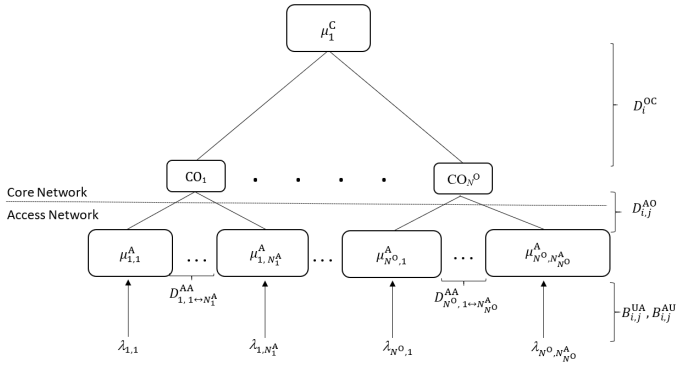


Fig. 2: Parameters used in pre-CORD architecture.

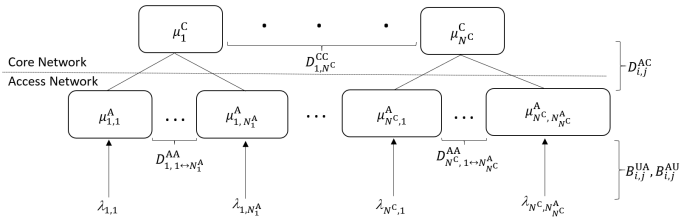


Fig. 3: Parameters used in post-CORD architecture.

B. Problem statement

The given arrival traffic for three kinds of services has different latency requirements, and four kinds of arrival traffic rates which are light uniform, light with some hotspot, heavy uniform, and heavy with some hotspots. Our objective was to minimize the capacity of AN and CN-MEC sites to accommodate that traffic with a latency satisfaction percentage constraint that is higher than a defined threshold T. An offloading mechanism is involved for minimizing latency violation. As shown in Table III, the problem statement of pre-CORD is defined as follows. Input: Traffic $\lambda_{i,j}$; the total number of AN-MEC, CO and CN-MEC sites are denoted by N_i^{A}, N^{O}, N^{C} respectively; link bandwidth: $B_{i,j}^{\mathrm{UA}}$,and $B_{i,j}^{\mathrm{AU}}$; link propagation delay: $D_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}}$, $D_{i,j}^{\mathrm{AO}}$, and D_i^{OC}; latency requirements: L. Output: site's capacity: μ_i^{C}, $\mu_{i,j}^{\mathrm{A}}$; offloading ratio: $\lambda_{i,j}^{\mathrm{A}}$, $\lambda_{i,j_1\rightarrow j_2}^{\mathrm{AA}}$, $\lambda_{i,j}^{\mathrm{C}}$; total of arrival traffic at AN-MEC and CN-MEC site: $\lambda_{i,j}^{\mathrm{AT}}$, λ_i^{CT}. Objective: $\min \mu^{\mathrm{C}} + \sum_{i=1}^{N^{\mathrm{C}}} \sum_{j=1}^{N_i^{\mathrm{A}}} \mu_{i,j}^{\mathrm{A}}$. Constraint: $F(L) \geq T$ where $F(L)$ is CDF of traffic latency.

The problems inherent in pre-CORD and post-CORD architectures have some differences, as shown in Table III. A CO hosts a CN-MEC in post-CORD architecture. The

TABLE II: Notations.

Category	Notations	Meaning	Attribute
Topology	N^{O}	Number of CO	Input
	N^{C}	Number of CN-MEC	Input
	N_i^{A}	Number of AN-MEC sites of ith CO/CN-MEC	Input
	$d^{\mathrm{AA}}_{i,j_1\leftrightarrow j_2}$	Distance between j_1th AN-MEC and j_2th AN-MEC of ith CO/CN-MEC	Input
	$d^{\mathrm{CC}}_{i_1\leftrightarrow i_2}$	Distance between i_1th CN-MEC and i_2th CN-MEC	Input
	$d^{\mathrm{AO}}_{i,j}$	Distance between jth AN-MEC and ith CO ith	Input
	d^{OC}_{i}	Distance between ith CO and CN-MEC	Input
	$d^{\mathrm{AC}}_{i,j}$	Distance between jth AN-MEC and ith CN-MEC	Input
Capacity	$B^{\mathrm{UA}}_{i,j}$	Link bandwidth from UE to jth AN-MEC of ith CO/CN-MEC (uplink)	Input
	$B^{\mathrm{AU}}_{i,j}$	Link bandwidth from jth AN-MEC of ith CO/CN-MEC to UE (downlink)	Input
	μ_i^{C}	Allocated capacity at ith CN-MEC	Output
	$\mu^{\mathrm{A}}_{i,j}$	Allocated AN-MEC capacity at jth AN-MEC of ith CO/CN-MEC	Output
Delay	$D^{\mathrm{AA}}_{i,j_1\leftrightarrow j_2}$	Propagation delay between j_1th AN-MEC and j_2th AN-MEC of ith CO/CN-MEC	Input
	$D^{\mathrm{CC}}_{i_1\leftrightarrow i_2}$	Propagation delay between i_1th CN-MEC and i_2th CN-MEC	Input
	$D^{\mathrm{AO}}_{i,j}$	Propagation delay between jth AN-MEC and ith CO	Input
	D^{OC}_{i}	Propagation delay between ith CO and CN-MEC	Input
	$D^{\mathrm{AC}}_{i,j}$	Propagation delay between jth AN-MEC and ith CN-MEC	Input
	t	Latency experienced by traffic on a node	Variable
	L	Latency constraint	Input
	T	Latency satisfaction percentage constraint	Input
Traffic	$\lambda_{i,j}$	Arrival traffic rate at jth AN-MEC site of ith CO/CN-MEC	Input
	$\lambda^{\mathrm{AT}}_{i,j}$	Total of arrival traffic rate at jth AN-MEC site of ith CO/CN-MEC	Output
	$\lambda^{\mathrm{CT}}_{i}$	Total of arrival traffic rate at ith CN-MEC	Output
Offloading	$\lambda^{\mathrm{A}}_{i,j}$	Offloaded traffic at jth AN-MEC of ith CO/CN-MEC	Output
	$\lambda^{\mathrm{AA}}_{i,j_1\rightarrow j_2}$	Offloaded traffic from j_1th to j_2th AN-MEC	Output
	$\lambda^{\mathrm{CC}}_{i_1\rightarrow i_2,j}$	Offloaded traffic from i_1th to i_2th CN-MEC	Output
	$\lambda^{\mathrm{C}}_{i,j}$	Offloaded traffic from jth AN-MEC to ith CN-MEC	Output

AN-MEC is then directly connected to the CN-MEC in post-CORD. $D^{\mathrm{AO}}_{i,j}$, and D^{OC}_{i} are replaced with $D^{\mathrm{AC}}_{i,j}$. Since the CN-MEC is developed in the CO, this architecture has more than one CN-MEC site, and offloading can take place between CN-MECs and introduces a horizontal offloading ratio in CN sites, $\lambda^{\mathrm{CC}}_{i_1\rightarrow i_2,j}$. Post-CORD's objective is $\min \sum_{i=1}^{N^{\mathrm{C}}} \mu_i^{\mathrm{C}} + \sum_{i=1}^{N^{\mathrm{C}}} \sum_{j=1}^{N_i^{\mathrm{A}}} \mu^{\mathrm{A}}_{i,j}$, and its constraint is the same as in pre-CORD architecture.

The problem mentioned above is solved by implementing the LA-TPIO algorithm, which is a management and control plane integrations. LA-TPIO adjusts the capacity at every MEC site based on its latency satisfaction percentage. After adjusting every MEC site's capacity, the algorithm determines the vertical and horizontal offloading ratio for minimizing the delay violation at each site.

C. Capacity scaling and offloading

In this study, the two-tier multi-server pre-CORD and post-CORD architecture integrate offloading and scaling onto LA-TPIO to minimize system capacity. Some kinds of arrival traffic rates are generated at all AN-MECs in both architectures. The traffic is categorized as uniform or non-uniform. Non-uniform traffic is traffic with a hotspot at some of AN-MEC sites. The generated traffic represent URLLC, eMBB, and mMTC services, which have different latency requirements.

When receiving the traffic, as a short-term solution, the offloading module determines how much and where the traffic should be processed for maximizing the latency satisfaction percentage. In pre-CORD, each AN-MEC site's traffic can be offloaded horizontally to a thousand AN-MEC sites or vertically to a CN-MEC site. A centralized CN-MEC site with high capacity is better than thousands of distributed AN-MEC sites with small capacity. However, because of its centralized location, a CN-MEC site is distant from some AN-MEC sites, which results in its propagation latency violating the latency constraint. Different from pre-CORD, post-CORD's CN-MEC sites are proximate to some AN-MEC sites. Although post-CORD's CN-MEC sites are more widely distributed than the pre-CORD's, the post-CORD's CN-MEC sites are implemented in a smaller number than the number of AN-MEC sites. The trade-off between distributed capacity and distance is investigated in this work.

Scaling is a long-term solution for accommodating arrival traffic. In scaling up MEC sites, their capacity should not be greater than current arrival traffic and still provides a latency satisfaction percentage greater than or equal to the threshold. The scaling down function is executed if the current system capacity is unnecessarily greater than the traffic arrival rate. In scaling down the system, the amount of remaining system capacity must be adequate for arrival traffic and should not over-satisfying the latency satisfaction percentage.

IV. Latency Aware Two-Phase Iterative Optimization

LA-TPIO was chosen because it specifically addresses scaling and offloading problems into management and control planes, respectively. In the first iteration, LA-TPIO adjusts the offloading ratio to minimize the latency constraint violation. Adjusting the offloading ratios can accommodate hotspot traffic by distributing it to some available capacity at AN and CN-MEC sites. LA-TPIO not only determines the vertical but also the horizontal offloading ratio. Because horizontal traffic offloading can cross hundreds of MEC sites, in selecting the destination MEC site the propagation delay from the source must be less than the latency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

TABLE III: Comparison table of pre-CORD and post-CORD's problem statement.

Architecture	Inputs	Outputs	Objective	Constraint
Pre-CORD	$\lambda_{i,j}, N^{\mathrm{C}}, N^{\mathrm{O}}, N_i^{\mathrm{A}}, B_{i,j}^{\mathrm{UA}}, B_{i,j}^{\mathrm{AU}}, D_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}}, D_{i,j}^{\mathrm{AO}}, D_i^{\mathrm{OC}}, L, T$	$\mu^{\mathrm{C}}, \mu_{i,j}^{\mathrm{A}}, \lambda_{i,j}^{\mathrm{A}}, \lambda_{i,j_1\rightarrow j_2}^{\mathrm{AA}}, \lambda_{i,j}^{\mathrm{C}}, \lambda_{i,j}^{\mathrm{AT}}, \lambda^{\mathrm{CT}}$	$\min \mu^{\mathrm{C}} + \sum_{i=1}^{N^{\mathrm{C}}} \sum_{j=1}^{N_i^{\mathrm{A}}} \mu_{i,j}^{\mathrm{A}}$	$F(L) \geq T$
Post-CORD	$\lambda_{i,j}, N^{\mathrm{C}}, N_i^{\mathrm{A}}, B_{i,j}^{\mathrm{UA}}, B_{i,j}^{\mathrm{AU}}, D_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}}, D_{i,j}^{\mathrm{AC}}, D_{i_1\leftrightarrow i_2}^{\mathrm{CC}}, L, T$	$\mu_i^{\mathrm{C}}, \mu_{i,j}^{\mathrm{A}}, \lambda_{i,j}^{\mathrm{A}}, \lambda_{i,j_1\rightarrow j_2}^{\mathrm{AA}}, \lambda_{i_1\rightarrow i_2,j}^{\mathrm{CC}}, \lambda_{i,j}^{\mathrm{C}}, \lambda_{i,j}^{\mathrm{AT}}, \lambda_i^{\mathrm{CT}}$	$\min \sum_{i=1}^{N^{\mathrm{C}}} \mu_i^{\mathrm{C}} + \sum_{i=1}^{N^{\mathrm{C}}} \sum_{j=1}^{N_i^{\mathrm{A}}} \mu_{i,j}^{\mathrm{A}}$	$F(L) \geq T$

constraint. The second iteration determines how much capacity is added and removed at each MEC site. To determine where to offload and how much capacity needs to be added and removed, the latency formulation and the latency satisfaction percentage thresholds need to be defined. The idea is not only to minimize the amount of active capacity but also to satisfy latency satisfaction percentage constraints.

A. Latency distribution

This study applies a Poisson distribution to model the incoming light and heavy traffic with rate $\lambda_{i,j}$, the traffic being directed to the jth AN-MEC of the ith CO/CN-MEC. In light and heavy with hotspot scenarios, hotspot traffic is generated to some AN-MECs with a double arrival rate. Service time follows an exponential distribution with rate $\frac{1}{\mu}$. By considering M/M/1 queuing model [27], the probability density function (PDF) of latency, t, and the cumulative distribution function (CDF) are expressed as:

$$f(t) = (\mu - \lambda)e^{-(\mu-\lambda)\times t}, \tag{1}$$

and

$$F(L) = P(t \leq L) = \begin{cases} 1 - e^{-(\mu-\lambda)\times L} & , L \geq 0 \\ 0 & , L < 0 \end{cases}. \tag{2}$$

1) Pre-CORD latency distributions: In pre-CORD architecture, there are several ways of serving incoming traffic. In the first case, arrival traffic is served at an AN-MEC (type A traffic). In the second case, if an AN-MEC is overloaded, the traffic can be offloaded to another AN-MEC (type AA traffic). If all AN-MECs are overloaded in the third case, traffic is offloaded to CN-MEC (type C traffic).

a) Type A traffic: First, we consider that the traffic can be served in AN-MEC, these sites being the first entities that receive the arrival traffic. The CDF of incoming traffic to an AN-MEC is expressed as:

$$\begin{aligned} F_{i,j}^{\mathrm{A}}(L) =& P(t \leq L) = \int_0^{L} \int_0^{L-t_1} f_{i,j}^{\mathrm{UA}}(t_1) \times f_{i,j}^{\mathrm{A}}(t_2) \\ & \times (1 - e^{-(B_{i,j}^{\mathrm{AU}} - \lambda_{i,j}^{\mathrm{AU}})\times(L-t_1)}) dt_2 dt_1, \end{aligned} \tag{3}$$

where

$$\begin{aligned} f_{i,j}^{\mathrm{UA}}(t) =& (B_{i,j}^{\mathrm{UA}} - \lambda_{i,j}^{\mathrm{UA}})e^{-(B_{i,j}^{\mathrm{UA}} - \lambda_{i,j}^{\mathrm{UA}})\times t}, \\ f_{i,j}^{\mathrm{A}}(t) =& (\mu_{i,j}^{\mathrm{A}} - \lambda_{i,j}^{\mathrm{AT}})e^{-(\mu_{i,j}^{\mathrm{A}} - \lambda_{i,j}^{\mathrm{AT}})\times t}. \end{aligned}$$

$f_{i,j}^{\mathrm{UA}}$ and $f_{i,j}^{\mathrm{A}}$ represent the PDF of an AN-MEC's uplink and AN-MEC latency, respectively. t_1 and t_2 are the time that is spent in those entities. The server's capacity at jth AN-MEC site of ith CO/CN-MEC site is denoted by $\mu_{i,j}^{\mathrm{A}}$. Type A traffic experiences latency in access network's uplink ($f_{i,j}^{\mathrm{UA}}$), an AN-MEC site ($f_{i,j}^{\mathrm{A}}$) and downlink, AU, which is represented as residual latency among them. Total arrival traffic rate at jth AN-MEC of ith CO/CN-MEC ($\lambda_{i,j}^{\mathrm{AT}}$) is equal to $\lambda_{i,j}^{\mathrm{A}} + \sum_{x\in\{1,2,\ldots,N_i^{\mathrm{A}}\}\setminus\{j\}} \lambda_{i,x\rightarrow j}^{\mathrm{AA}}$, which are the total of vertical and horizontal offloading traffic.

b) Type AA traffic: In the second case, arrival traffic is offloaded to another AN-MEC. The traffic travels through the UA link and then goes through the link between AN-MECs, called a mid-haul network, with propagation delay $D_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}}$. The queuing delay in a mid-haul is neglected because it is an optical fiber network with a huge bandwidth capacity. Since the defined latency constraint is L, it contains the node and link latencies. We can remove the link latency, and derive the latency constraint of the node $L_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}} = L - D_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}}$. The CDF of traffic arriving at AN-MEC is represented as:

$$\begin{aligned} F_{i,j_1\rightarrow j_2}^{\mathrm{AA}}(L) =& P(t \leq L) = \int_0^{L_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}}} \int_0^{L_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}} - t_1} f_{i,j_1}^{\mathrm{UA}}(t_1) \\ & \times f_{i,j_2}^{\mathrm{A}}(t_2) \\ & \times (1 - e^{-(B_{i,j_1}^{\mathrm{AU}} - \lambda_{i,j_1}^{\mathrm{AU}})\times(L_{i,j_1\leftrightarrow j_2}^{\mathrm{AA}} - t_1)}) dt_2 dt_1, \end{aligned} \tag{4}$$

where

$$\begin{aligned} f_{i,j_1}^{\mathrm{UA}}(t) =& (B_{i,j_1}^{\mathrm{UA}} - \lambda_{i,j_1}^{\mathrm{UA}})e^{-(B_{i,j_1}^{\mathrm{UA}} - \lambda_{i,j_1}^{\mathrm{UA}})\times t}, \\ f_{i,j_2}^{\mathrm{A}}(t) =& (\mu_{i,j_2}^{\mathrm{A}} - \lambda_{i,j_2}^{\mathrm{AT}})e^{-(\mu_{i,j_2}^{\mathrm{A}} - \lambda_{i,j_2}^{\mathrm{AT}})\times t}. \end{aligned}$$

The differences in PDF of delay in type A and type AA traffic lie in the propagation delay that affects the PDF. For example, type AA traffic can be served in an AN-MEC that is located one or some hops away from the source AN-MEC. Total arrival traffic rate at j_2th AN-MEC site is vertical and horizontal offloaded traffic directed to j_2th and calculated as $\lambda_{i,j_2}^{\mathrm{AT}} = \lambda_{i,j_2}^{\mathrm{A}} + \sum_{x\in\{1,2,\ldots,N_i^{\mathrm{A}}\}\setminus\{j_2\}} \lambda_{i,x\rightarrow j_2}^{\mathrm{AA}}$.

c) Type C traffic: When the AN-MEC load is near its capacity, some of the incoming traffic can be offloaded to the CN-MEC. The offloaded traffic goes through AO and OC links. Since the L contains the node and link latencies, we can get the CN-MEC's node delay by $L^{\mathrm{C}} = L - D_{i,j}^{\mathrm{AO}} - D_i^{\mathrm{OC}}$. Total arrival traffic which is served at ith CN-MEC site (λ^{CT}), is a sum of offloaded traffic from all jth AN-MEC sites of all ith CO to a CN-MEC site, $\lambda^{\mathrm{CT}} = \sum_{i\in\{1,2,\ldots,N^{\mathrm{O}}\}} \sum_{j\in\{1,2,\ldots,N_i^{\mathrm{A}}\}} \lambda_{i,j}^{\mathrm{C}}$. We can derive the CDF of type C traffic as:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

$$F^{\mathrm{C}}_{i,j}(L) = P(t \le L) = \int_0^{L^{\mathrm{C}}} \int_0^{L^{\mathrm{C}}-t_1} f^{\mathrm{UA}}_{i,j}(t_1) \times f^{\mathrm{C}}(t_2) \times (1 - e^{-(B^{\mathrm{AU}}_{i,j} - \lambda^{\mathrm{AU}}_{i,j}) \times (L^{\mathrm{C}} - t_1)}) dt_2 dt_1, \quad (5)$$

where

$$f^{\mathrm{UA}}_{i,j}(t) = (B^{\mathrm{UA}}_{i,j} - \lambda^{\mathrm{UA}}_{i,j}) e^{-(B^{\mathrm{UA}}_{i,j} - \lambda^{\mathrm{UA}}_{i,j}) \times t},$$
$$f^{\mathrm{C}}(t) = (\mu^{\mathrm{C}} - \lambda^{\mathrm{CT}}) e^{-(\mu^{\mathrm{C}} - \lambda^{\mathrm{CT}}) \times t}.$$

The CDF latency of arrival traffic in an AN-MEC site of pre-CORD can be derived from (3)-(5) and is represented as:

$$F_{i,j}(L) = \frac{F^{\mathrm{A}}_{i,j}(L)\lambda^{\mathrm{A}}_{i,j} + F^{\mathrm{AA}}_{i,j_1 \to j_2}(L)\lambda^{\mathrm{AA}}_{i,j_1 \to j_2} + F^{\mathrm{C}}_{i,j}(L)\lambda^{\mathrm{C}}_{i,j}}{\lambda_{i,j}}. \quad (6)$$

2) Post-CORD latency distributions: This section covers the latency distributions, which are different from those in pre-CORD architecture. In post-CORD, the CO is equipped with MEC servers, with a capacity of μ^{C}_i, and changes its role to become a CN-MEC site to expand its computing capacity. Since the traffic from the AN-MEC sites can be directly connected to the CN-MEC sites, the latency constraint, L, is only affected by a link propagation delay between AN-MEC and CN-MEC, $D^{\mathrm{AC}}_{i,j}$. The node latency constraint at a CN-MEC site, L^{C}, can be derived by removing the propagation delay from the total latency constraint, $L - D^{\mathrm{AC}}_{i,j}$. Post-CORD architecture also has some CN-MEC sites that enable offloading traffic between them if types A, AA, and AC traffic are not satisfied. Total arrival traffic rate at ith CN-MEC site is a sum of horizontal and vertical offloaded traffic which is calculated as, $\lambda^{\mathrm{CT}}_i = \sum_{j \in \{1,2,\ldots,N^{\mathrm{A}}_i\}} \lambda^{\mathrm{C}}_{i,j} + \sum_{x \in \{1,2,\ldots,N^{\mathrm{C}}\} \setminus \{i\}} \sum_{j \in \{1,2,\ldots,N^{\mathrm{A}}_x\}} \lambda^{\mathrm{CC}}_{x \to i,j}$. We can then carry out CN-MEC to CN-MEC offloading as type CC traffic with the latency CDF, and is given as:

$$F^{\mathrm{CC}}_{i_1 \to i_2,j}(L) = P(t \le L) = \int_0^{L^{\mathrm{CC}}_{i_1 \leftrightarrow i_2}} \int_0^{L^{\mathrm{CC}}_{i_1 \leftrightarrow i_2} - t_1} f^{\mathrm{UA}}_{i_1,j}(t_1) \times f^{\mathrm{C}}_{i_2}(t_2) \times (1 - e^{-(B^{\mathrm{AU}}_{i_1,j} - \lambda^{\mathrm{AU}}_{i_1,j}) \times (L^{\mathrm{CC}}_{i_1 \leftrightarrow i_2} - t_1)}) dt_2 dt_1,$$

where

$$f^{\mathrm{UA}}_{i_1,j}(t) = (B^{\mathrm{UA}}_{i_1,j} - \lambda^{\mathrm{UA}}_{i_1,j}) e^{-(B^{\mathrm{UA}}_{i_1,j} - \lambda^{\mathrm{UA}}_{i_1,j}) \times t},$$
$$f^{\mathrm{C}}_{i_2}(t) = (\mu^{\mathrm{C}}_{i_2} - \lambda^{\mathrm{CT}}_{i_2}) e^{-(\mu^{\mathrm{C}}_{i_2} - \lambda^{\mathrm{CT}}_{i_2}) \times t}. \quad (7)$$

Type CC traffic not only experiences propagation delays between AN-MEC and CN-MEC, $D^{\mathrm{AC}}_{i_1,j}$, but also propagation delay between many CN-MECs, $D^{\mathrm{CC}}_{i_1 \leftrightarrow i_2}$. Node latency, $L^{\mathrm{CC}}_{i_1 \leftrightarrow i_2}$ is then obtained as $L^{\mathrm{CC}}_{i_1 \leftrightarrow i_2} = L - D^{\mathrm{AC}}_{i_1,j} - D^{\mathrm{CC}}_{i_1 \leftrightarrow i_2}$. From (3), (4), (5), and (7), the CDF of latency for arrival traffic in each AN-MEC site of the post-CORD can be expressed as:

$$F_{i,j}(L) = \frac{F^{\mathrm{A}}_{i,j}(L)\lambda^{\mathrm{A}}_{i,j} + F^{\mathrm{AA}}_{i,j_1 \to j_2}(L)\lambda^{\mathrm{AA}}_{i,j_1 \to j_2}}{\lambda_{i,j}} + \frac{F^{\mathrm{C}}_{i,j}(L)\lambda^{\mathrm{C}}_{i,j} + F^{\mathrm{CC}}_{i_1 \to i_2,j}(L)\lambda^{\mathrm{CC}}_{i_1 \to i_2,j}}{\lambda_{i,j}}. \quad (8)$$

The total of $F_{i,j}(L)$ for pre-CORD and post-CORD can be expressed as:

$$F(L) = \frac{\sum_i \sum_j F_{i,j}(L)\lambda_{i,j}}{\sum_i \sum_j \lambda_{i,j}}. \quad (9)$$

B. The LA-TPIO workflow

1) Latency-aware two-phase iterative optimization: The above-mentioned management plane problem is solved by integrating the scaling and offloading algorithms in an LA-TPIO approach, which is shown in Fig. 4. This integration results in a short-term solution, by offloading, and a long-term solution, by scaling, for handling some traffic patterns with defined latency satisfaction percentage constraints. LA-TPIO works for both pre-CORD and post-CORD architectures, the differences being in input, depending on the location and the number of CN-MEC.

In the first phase of LA-TPIO, the offloading algorithm determines the destination of the traffic for maximizing latency satisfaction percentage. If the offloading function meets its iteration limit and the latency satisfaction percentage is lower than the defined threshold in the current system capacity, the algorithm enters the second phase to scale up the system's capacity. If the latency satisfaction percentage is higher than the defined threshold, the algorithm enters the second phase to minimize system capacity.

Fig. 4 shows that the algorithm starts by measuring the CDF of all MEC sites' traffic latency, $F(L)$. If $F(L)$ is less than T, then the algorithm enters the first phase. The algorithm adjusts the offloading ratio to satisfy the latency percentage constraint in this phase. Index g is used to limit looping with the maximum number G in the first phase, in case that the system cannot satisfy the latency percentage constraint by utilizing the current capacity. g is incremented by one for each loop. The algorithm enters the second phase if $F(L)$ satisfies the T or $g > G$. The algorithm adjusts the system capacity in phase two. The adjustment is bounded by the latency percentage constraint. After adjusting the system capacity, the algorithm re-enters the first phase. These two phases are repeated until there are very small changes in the system capacity or the incremented looping index h is greater than H, which indicates that no better solution can be found.

2) First phase, adjusting the offloading ratio: In this phase, the system adjusts the offloading ratios of all AN-MEC sites in two-tier MEC architecture. Suppose an AN-MEC site is overloaded in pre-CORD architecture, traffic can be offloaded to another AN-MEC (horizontal offloading) or be directed to CN-MEC (vertical offloading). The latency distribution in each site is derived from (3), (4), and (5). Since post-CORD has more than one CN-MEC site, it initiates horizontal offloading between CN-MEC sites whose latency distribution is derived from (7). In both architectures, if an MEC-site latency does not satisfy the latency percentage constraint, the traffic is then shifted to another MEC site.

Fig. 5 and 6 show the offloading algorithm flows. The algorithm computes $F_{i,j}$ for each jth AN-MEC site of ith

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

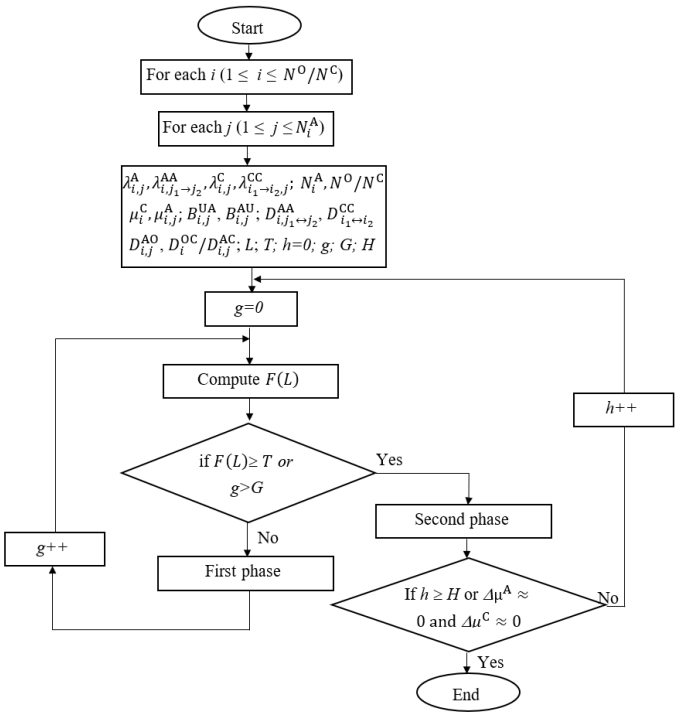


Fig. 4: Two-phase iterative optimization.

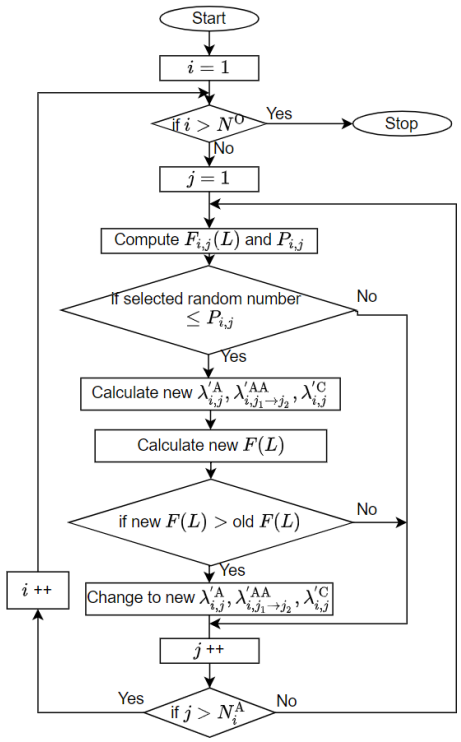


Fig. 5: Traffic allocation algorithm for pre-CORD.

CO/CN-MEC. It also computes the latency violation probability of a site, $P_{i,j}$. We derive $P_{i,j}$ as follows:

$$P_{i,j} = \max\left(1 - \frac{F_{i,j}(L)}{T}, 0\right). \tag{10}$$

The traffic at a site has the probability of being shifted if $F_{i,j} < T$, which means that the traffic shifting probability is greater than 0. We then choose a random number between 0 and 1: if the random number is less than $P_{i,j}$ of a traffic, traffic is then shifted from one traffic type to another traffic type. For example, AN-MEC site's traffic, $\lambda_{i,j}^A$, which has a high shifting probability, is shifted to another CN-MEC or AN-MEC site, one with the highest latency satisfaction percentage. In determining how much traffic is shifted from the AN-MEC site to another, we uses $F_{i,j}^A(L) = T$ to calculate a new traffic distribution ($\lambda_{i,j}^{'A}$, $\lambda_{i,j_1\to j_2}^{'AA}$, $\lambda_{i,j}^{'C}$) for pre-CORD and ($\lambda_{i,j}^{'A}$, $\lambda_{i,j_1\to j_2}^{'AA}$, $\lambda_{i,j}^{'C}$, $\lambda_{i_1\to i_2,j}^{'CC}$) for post-CORD. After obtaining the new traffic distribution, $F^{'}(L)$ is computed. If $F^{'}(L) > F(L)$, then the new traffic distribution is retained. Otherwise the old traffic distribution is retained.

3) Second phase, adjusting capacity: The scaling algorithm is bonded with the latency satisfaction percentage to adjust the capacity of each AN-MEC, $\mu_{i,j}^A$, and CN-MEC site, μ_i^C. As shown in Fig. 7, the new $\mu_{i,j}^{'A}$ and $\mu_i^{'C}$ are calculated to satisfy the latency satisfaction percentage which is derived from (6) and (8) for pre-CORD and the post-CORD architecture, respectively. First, the algorithm calculates $\mu_{i,j}^{'A}$ for each AN-MEC, while still using current μ_i^C. Second, the algorithm calculates $\mu_i^{'C}$ for each CN-MEC, while still using current $\mu_{i,j}^A$. $\Delta\mu^C$ and $\Delta\mu^A$ are calculated as $\sum_{i=1}^{N^C}\mu_i^C - \sum_{i=1}^{N^C}\mu_i^{'C}$ and $\sum_{i=1}^{N^C}\sum_{j=1}^{N_i^A}\mu_{i,j}^A - \sum_{i=1}^{N^C}\sum_{j=1}^{N_i^A}\mu_{i,j}^{'A}$,

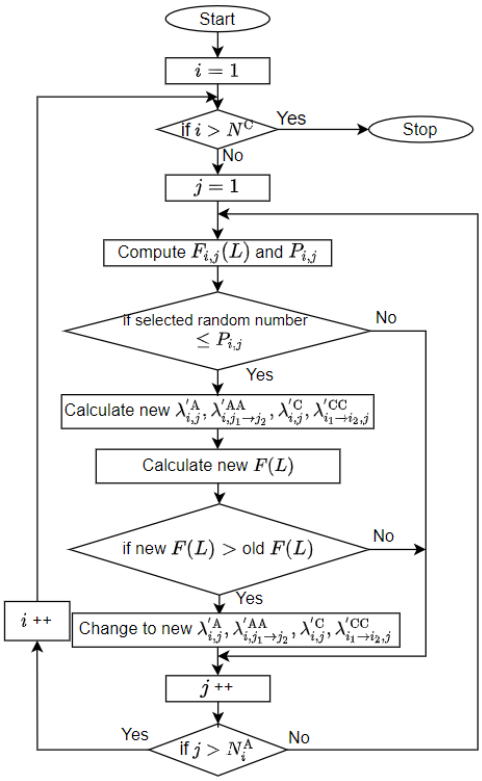


Fig. 6: Traffic allocation algorithm for post-CORD.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

respectively. Then, the algorithm calculates $\Delta\mu^{total} = \max(\Delta\mu^{A}, \Delta\mu^{C})$ in scaling up or scaling down the system to minimize its capacity. Before the changes, it calculates total capacity, μ^{total}, which is equal to $\sum_{i=1}^{N^{C}} \mu_i^{C} + \sum_{i=1}^{N^{C}} \sum_{j=1}^{N_i^{A}} \mu_{i,j}^{A}$.

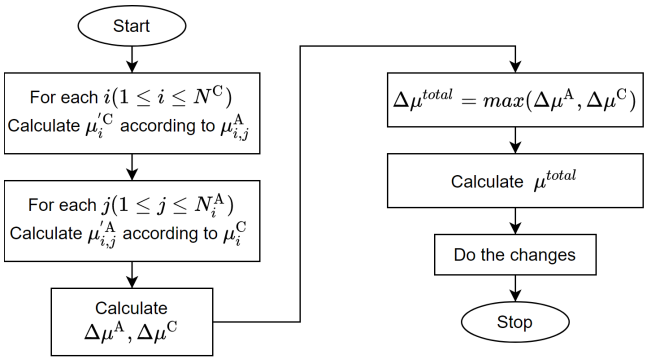


Fig. 7: Scaling algorithm for each MEC site.

4) Time complexity of LA-TPIO: LA-TPIO adjusts the offloading ratio in the first phase and scales the capacity in the second phase. As shown in Fig. 5 and 6, offloading ratio adjustment depends on N^{O}/N^{C} CO/CN-MEC sites and each of CO/CN-MEC site is connected to N_i^{A} AN-MEC sites. In the worst-case, where current capacity cannot satisfy the arrival traffic, the offloading adjustment iterations are limited by G. So, the time complexity of the first phase is $O(G \times N^{C} \times N^{A})$ where N^{A} is the maximum of all N_i^{A}. Because N^{O} of pre-CORD is equal to N^{C} of post-CORD, we can consider only one of them to get the time complexity.

The capacity scaling, shown in Fig. 7, is carried out for every AN-MEC and CN-MEC site. (9) is used to calculate how much capacity needs to be added or removed. The second phase's time complexity can be represented as $O(N^{C} \times (1 + N^{A}))$. As shown in Fig. 4, both phases have iterations limited by H. So, the worst case time complexity is $O(H \times ((N^{C} \times N^{A} \times G) + (N^{C} \times (1 + N^{A}))) = O(H \times G \times N^{C} \times N^{A})$.

V. Simulation and Results

The simulation was conducted to investigate scaling and offloading in pre-CORD and post-CORD architectures, which use LA-TPIO for minimizing the allocated capacity. Four types of traffic arrival rates and three types of services were generated. We also investigated capacity allocation in pre-CORD vs. post-CORD for satisfying different latency satisfaction percentages, the effect of latency satisfaction percentage, the effect of latency constraint, uniform vs. hotspot traffic, and LA-TPIO performance.

A. Parameters settings

Table IV lists the parameters used to simulate both pre-CORD and post-CORD architectures. There are some differences between them related to the topology and delay settings. Pre-CORD was allocated 1000 AN-MEC sites that connected to one CN through 100 COs. The distances between the AN-MEC sites and COs, $d_{i,j}^{AO}$, ranged from 5 to 10 km, and the distances between COs and a CN, d_i^{OC}, ranged from 10 to 100 km. The area was assumed to be 40,000 km^2 with 200 km length. Since the CN-MECs were placed at COs, post-CORD had one or more CN-MEC sites. The simulation used 100 CN-MEC sites. Each CN-MEC site covered 10 AN-MEC sites. In a 40,000 km^2, the distances between AN-MEC and CN-MEC, $d_{i,j}^{AC}$, sites in post-CORD ranged from 5 to 10 km. Post-CORD also established connections between CN-MEC sites with distance, d^{CC}, ranging from 20 to 40 km.

AN-MEC site's initial capacity was half of the arrival traffic rate. The CN-MEC site's initial capacity was the arrival traffic rate multiplied by the number of connected AN-MEC sites. Initially, all arrival traffic was directed to AN-MEC sites. This initial configuration triggered capacity scaling and traffic offloading at the AN-MEC and CN-MEC sites.

Three kinds of services, represented by URLLC, mMTC, and eMBB with latency requirements ranging from 1 to 100 ms, were considered in this simulation. Rather than considering only three different latency constraints, 1 ms for URLLC, 4 ms for eMBB and 100 ms for mMTC, this investigation extended 25 ms [28] and 50 ms [29] latency constraints for mMTC. That traffic was generated to each AN-MEC site with a rate of 2000 packets/second and 4000 packets/second to represent light and heavy arrival traffic rates generated by UEs. The link between UE and AN-MEC had an asymmetric link capacity [30]. In non-uniform traffic, the hotspot traffic rate was a double uniform traffic rate. Hotspot traffic was generated to half of the AN-MEC sites. Some latency satisfaction percentage thresholds were considered, which ranged from 60% to 95%.

B. Results

1) Pre-CORD vs. post-CORD: Fig. 8 shows the comparative results between pre-CORD and post-CORD with URLLC traffic. The objective of post-CORD provided a closer CN-MEC site than the pre-CORD's CN-MEC, which was ideal for the traffic with tight latency constraints such as URLLC. The results show that post-CORD required less capacity than pre-CORD, because some pre-CORD's capacities were located 100 km away from UEs, contributing to additional propagation delays. As the delay consists of a link and node delay, this propagation delay shoved the node delay requirements, which was why the pre-CORD sites allocated more capacity for satisfying the node's delay requirements. Post-CORD required 5-30% less capacity than pre-CORD. The most significant result was obtained when using ten CN-MEC sites. Scattering a resource into some areas in queuing systems produced poor capacity allocation. For example, one MEC site with a capacity of 30,000 packets/sec has 99.99% latency satisfaction in serving traffic with a rate equal to 10,000 packets/second. However, when that one MEC site was scattered into ten MEC sites with 3,000 packets/second capacity for serving 1,000 packets/second traffic, the MEC system only satisfied 72.93% traffic. Nevertheless, a small number of CN-MEC

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

TABLE IV: Parameters settings

Category	Notations	Pre-CORD	Post-CORD
Topology	N^{O}	100	-
	N^{C}	1	100
	N_i^{A}	10	10
	$d_{i,j_1 \leftrightarrow j_2}^{AA}$	$5\,km \leq d_{i,j_1 \leftrightarrow j_2}^{AA} \leq 10\,km$	$5\,km \leq d_{i,j_1 \leftrightarrow j_2}^{AA} \leq 10\,km$
	$d_{i_1 \leftrightarrow i_2}^{CC}$	-	$20\,km \leq d_{i_1 \leftrightarrow i_2}^{CC} \leq 40\,km$
	$d_{i,j}^{AO}$	$5\,km \leq d_{i,j}^{AO} \leq 10\,km$	-
	d_i^{OC}	$10\,km \leq d_i^{OC} \leq 100\,km$	-
	$d_{i,j}^{AC}$	-	$5\,km \leq d_{i,j}^{AC} \leq 10km$
Capacity	μ^{C}	$\mu_{i,j}^{A} \times N_i^{A}$	$\mu_{i,j}^{A} \times N_i^{A}$
	$\mu_{i,j}^{A}$	$\frac{\lambda_{i,j}}{2}$	$\frac{\lambda_{i,j}}{2}$
	$B_{i,j}^{UA}$	100 Mbps	100 Mbps
	$B_{i,j}^{AU}$	200 Mbps	200 Mbps
Delay	$D_{i,j_1 \leftrightarrow j_2}^{AA}$	$\frac{d_{i,j_1 \leftrightarrow j_2}^{AA}}{3\times10^5 km/s}$	$\frac{d_{i,j_1 \leftrightarrow j_2}^{AA}}{3\times10^5 km/s}$
	$D_{i_1 \leftrightarrow i_2}^{CC}$	-	$\frac{d_{i_1 \leftrightarrow i_2}^{CC}}{3\times10^5 km/s}$
	$D_{i,j}^{AO}$	$\frac{d_{i,j}^{AO}}{3\times10^5 km/s}$	-
	D_i^{OC}	$\frac{d_i^{OC}}{3\times10^5 km/s}$	-
	$D_{i,j}^{AC}$	-	$\frac{d_{i,j}^{AC}}{3\times10^5 km/s}$
	L	1, 4, 25, 50, and 100 ms	1, 4, 25, 50, and 100 ms
	T	60%, 70%, 80%, 90%, and 95%	60%, 70%, 80%, 90%, and 95%
Traffic	Light $\lambda_{i,j}$	2000 packets/s	2000 packets/s
	Heavy $\lambda_{i,j}$	4000 packets/s	4000 packets/s
	Hotspot $\lambda_{i,j}$	2 x $\lambda_{i,j}$	2 x $\lambda_{i,j}$
	Packet size of URLLC, mMTC, and eMBB	1, 5, 10 Kb	1, 5, 10 Kb [31]

sites decreased CN-MEC coverage and increased the distance between AN-MEC sites and a CN-MEC site. Fig. 9 shows that the best number of CN-MEC sites was ten because CN-MEC could cover all AN-MEC sites with ideal distances for URLLC. When the number of CN-MEC sites was reduced to four, the distance between some CN-MEC and AN-MEC sites increased, and some distances exceeded the required distance for URLLC. Fig. 10 shows the effect of different traffic rates of URLLC traffic at both architectures. The relation between allocated capacity and arrival traffic rate was linear.

Fig. 8 shows that pre-CORD utilized more AN-MEC sites than post-CORD because some of the AN-MEC sites were located far away from the CN-MEC site; AN-MEC sites had a capacity of 70% to 77%. In contrast, post-CORD utilized 17% to 44% of an AN-MEC site's capacity. Having more capacity in an AN-MEC site is more expensive than in a CN-MEC site because AN-MEC sites are widely distributed in some areas and need space, electrical capacity, and a cooling system for only a small number of servers. By comparison, CN-MEC sites are usually placed in COs, which are more centralized and already have space, electrical capacity, and a cooling system for networking equipment. Compared to the total arrival traffic, post-CORD had better system utilization than pre-CORD, ranging from 73% to 87% while pre-CORD ranged from 65% to 85%.

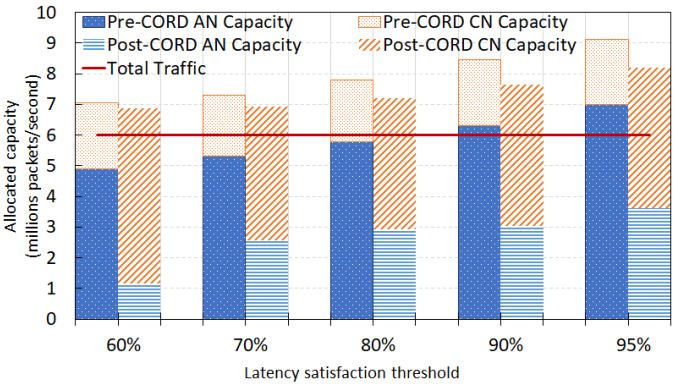


Fig. 8: pre-CORD vs. post-CORD.

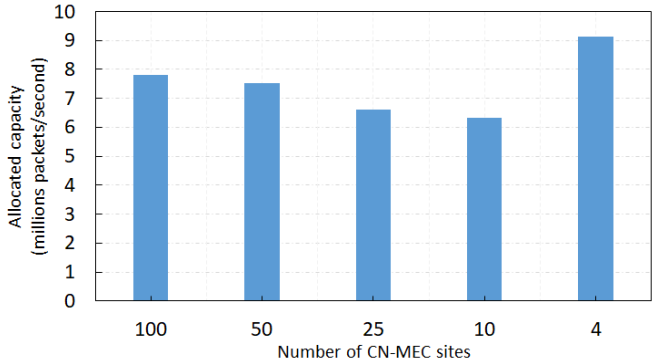


Fig. 9: Effect of number of CN-MEC sites in post-CORD.

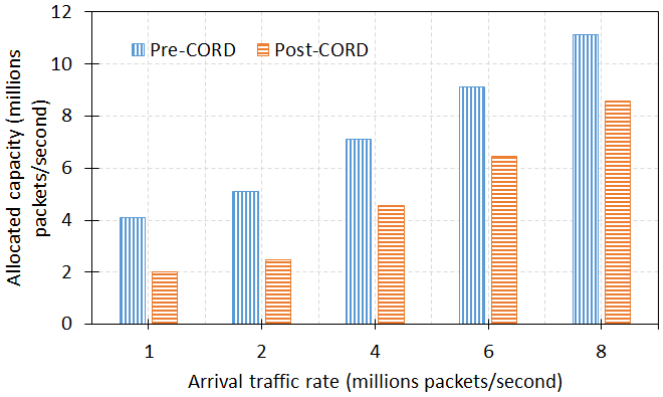


Fig. 10: Effect of arrival traffic rates.

2) Effect of the latency satisfaction percentage threshold: Various latency satisfaction percentage thresholds were also considered in the investigation. Fig. 11 shows a comparison of pre-CORD and post-CORD for accommodating URLLC traffic with different latency satisfaction percentage thresholds. Higher traffic satisfaction percentages needed greater allocated capacity. Post-CORD needed up to 22% less capacity than pre-CORD in serving some traffic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

scenarios, with latency satisfaction percentage thresholds ranging from 60% to 95%. The most significant result was obtained in serving traffic with latency satisfaction threshold, which was set at 95% because the system required to allocate more capacity to satisfy more arrival traffic. Since post-CORD capacities were closer to the UEs and had lower propagation delays, it left more time for node processing time and resulted in fewer resources being allocated than to pre-CORD.

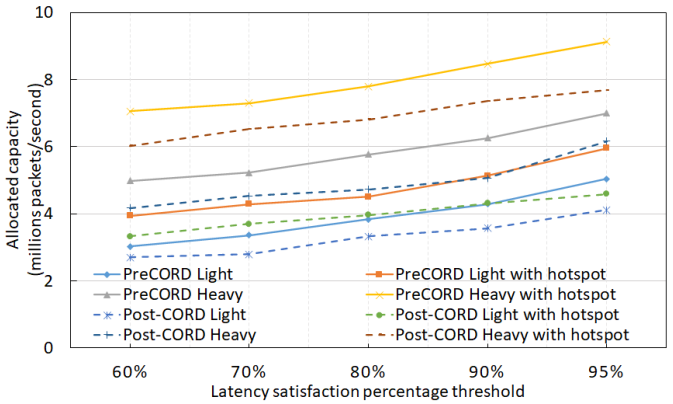
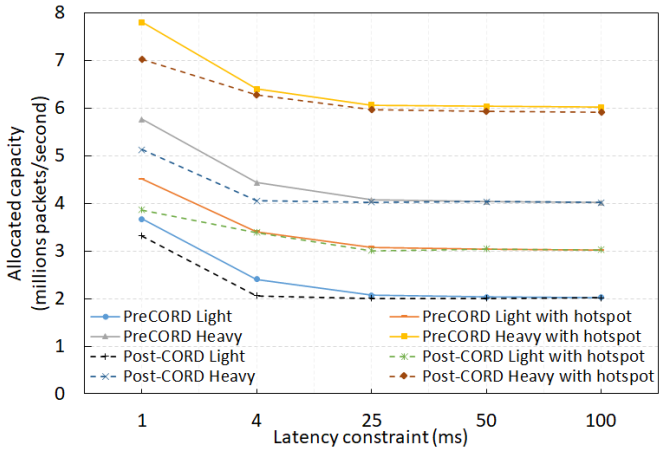
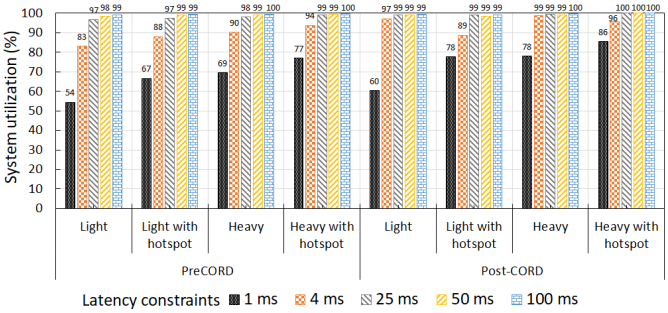


Fig. 11: Effect of latency satisfaction percentage threshold.

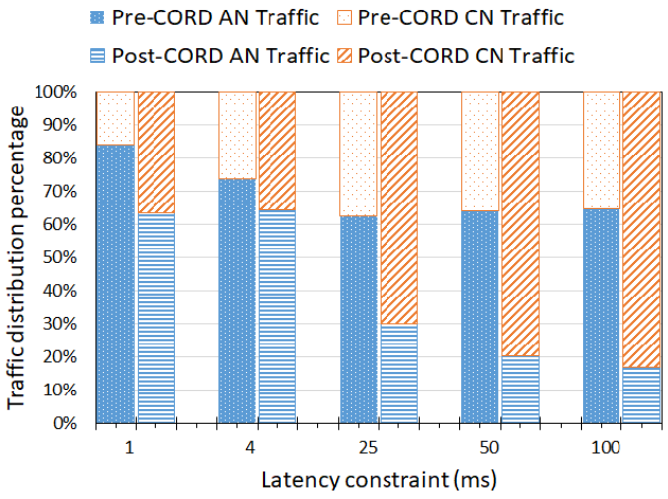
3) Effect of latency constraint: Different traffic from different services were generated into pre-CORD and post-CORD with an 80% latency satisfaction threshold. Fig. 12a shows that the service with the tightest latency constraint consumed the most capacity. In this case, the tightest latency constraint was generated by URLLC traffic. Tight latency constraint services utilized more AN-MEC sites than loose latency constrain services, as shown in Fig. 12c. A system utilization comparison between pre-CORD and post-CORD is shown in Fig. 12b, which shows the ratio of allocated resources and arrival traffic rate for serving some services. Greater system utilization indicates greater system efficiency in traffic handling. The URLLC and eMBB with 1 ms and 4 ms latency constraints, respectively, were services with less system utilization. This means that the allocated capacity must be greater than the incoming traffic in order to satisfy 80% traffic. Pre-CORD system utilization ranged from 54% to 77% and 83% to 94% for serving URLLC and eMBB, respectively. Post-CORD had higher system utilization than pre-CORD, which ranged from 60% to 86% and 89% to 97% for serving URLLC and eMBB traffic, respectively. The greater system utilization of post-CORD indicated that post-CORD needed about 8% to 14% less capacity, than pre-CORD in serving URLLC and eMBB traffic, because post-CORD has more capacity that is ideal for serving tight delay constraints. Because of large distances between pre-CORD's CN-MEC and some UEs, pre-CORD allocated more capacity to accommodate tighter node delays after being shoved by longer propagation delays. The longer propagation delay between AN-MEC and CN-MEC also caused pre-CORD to has more AN traffic than post-CORD, shown in Fig 12c. In serving services with latency constraints longer than 4 ms, both pre-CORD and post-CORD allocated nearly the same amount of capacity, and the system utilization was ranging from 97% to 100%. This showed that the pre-CORD's CN-MECs, located a long distance from some UEs, were also ideal for providing that kind of service, as we assumed that the back-haul and mid-haul networks were using optical fiber links where queuing can be neglected.



(a) Allocated capacity.



(b) System utilization.



(c) Traffic distribution.

Fig. 12: Effect of different latency constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSM.2021.3101862, IEEE Transactions on Network and Service Management

4) Uniform vs. non-uniform traffic: This section evaluates the effect of uniform and non-uniform (hotspot) traffic. Fig. 13 shows that hotspot traffic triggered more horizontal offloading than uniform traffic in post-CORD. Post-CORD resulted in more horizontal offloading than pre-CORD because horizontal offloading appeared not only between AN-MEC sites but also CN-MEC sites. Fig. 13 shows that the required link capacity for mid-haul was 28% of total traffic for post-CORD. Pre-CORD did not show much horizontal offloading in this simulation because it seems that the LA-TPIO kept the traffic of an overloaded AN-MEC site at the same site rather than distributed it to another site, and waiting for the scaling process to adjust capacity.

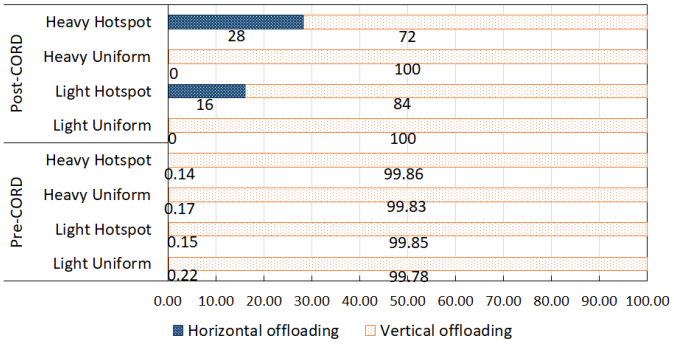
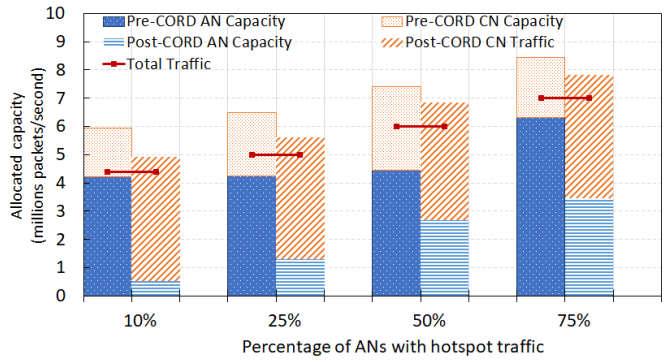
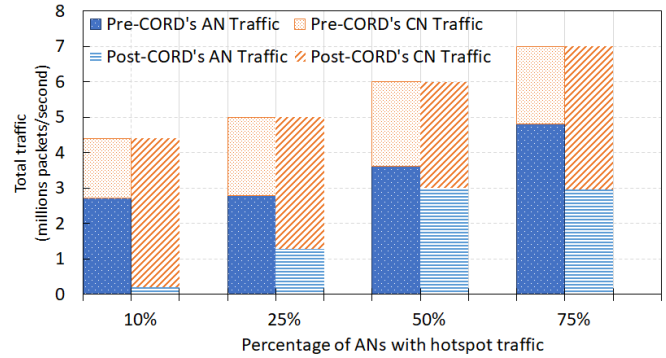


Fig. 13: Horizontal vs. vertical offloading.



(a) Allocated capacity.



(b) Traffic distribution.

Fig. 14: Effect of hotspot density.

Effect of hotspot density was also investigated. Hotspot URLLC traffic with 80% latency satisfaction constraint was generated in 10%, 25%, 50%, and 75% of ANs. Fig. 14a shows that total traffic and allocated capacity were increasing linearly with a number of ANs with hotspot. Because pre-CORD and post-CORD architectures used the same LA-TPIO algorithm to distribute traffic and to allocate capacity, the hotspot density affected only 2% to 4% in the allocated capacity comparison. Post-CORD with the least hotspot density has the least allocated capacity. Fig. 14b shows post-CORD tended to use CN-MEC while pre-CORD tended to use AN-MEC first to server arrival traffic because of long propagation delay to pre-CORD's CN-MEC. The density of the hotspot AN increased the pre-CORD's AN-MEC traffic ratio from 56% to 68%. In post-CORD, low-density hotspot traffic, which was generated at 10% and 25% of AN-MECs, occupied only 5% and 25% AN-MECs, and occupied nearly 50% of AN-MECs when the density of hotspot increases from 50% to 75%.

5) LA-TPIO performance comparison: Fig. 15 shows a comparison of the algorithms which was carried out to allocate capacity in high arrival traffic scenarios with URLLC traffic. Even in some satisfaction percentage threshold settings, simulated annealing (SA) and brute-force outperformed LA-TPIO in term of minimum capacity, the difference being no more than 0.9%. Brute-force and SA are time-consuming algorithms, with a complexity that is higher than LA-TPIO. The LA-TPIO needed 5 to 15 minutes to converge on 1000 AN-MEC sites, while SA and brute-force needed more than 30 minutes to converge.

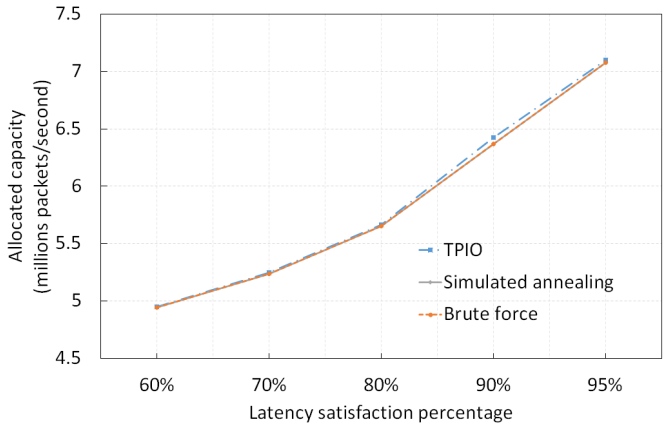


Fig. 15: TPIO vs. Simulted annealing vs. Brute force.

VI. Conclusions

A two-tier, multi-site, multi-server MEC applied LA-TPIO for integrating offloading and a scaling mechanism. Although LA-TPIO displayed nearly the same performance as SA and the minimum solution of brute-force, its convergence time was shorter than that of SA and brute-force. A two-tier architecture was evaluated under pre-CORD and post-CORD architectures. Post-CORD required less capacity than pre-CORD in all traffic scenarios. The most significant

difference was obtained when serving the URLLC traffic with a 95% latency satisfaction percentage threshold. Post-CORD with ten CN-MEC sites needed 30% less capacity than pre-CORD. Post-CORD resulted in a trade-off between the number of CN-MEC sites and the distance to CN-MEC sites, which affected the allocation of capacity. In satisfying 95% URRLC traffic, post-CORD had higher system utilization than pre-CORD, ranging from 73% to 87%, whereas pre-CORD was 65% to 85%. Pre-CORD utilized AN-MEC sites first because the CN-MEC location was distant from some UEs. Post-CORD utilized about 48% to 77% less AN-MEC sites capacity than pre-CORD because post-CORD's AN-MEC and CN-MEC sites were ideal for serving URLLC traffic. Post-CORD was more beneficial than pre-CORD because the development of AN-MEC sites at base stations, which was more widely distributed than CO, and was costlier than the development of CN-MEC sites in CO. Given some services with different latency requirements, our results show that tight latency services required greater capacity. In post-CORD, the AN-MEC and CN-MEC sites were ideal for serving ultra-low latency services. Post-CORD required 8% to 14% less capacity compared to pre-CORD in serving URRLC and eMBB with an 80% latency satisfaction percentage threshold. There were not many differences between pre-CORD and post-CORD's allocated capacity for providing services, which had latency constraints of more than 4 ms because the CN-MEC site of pre-CORD could also satisfy the arrival traffic. In uniform vs. non-uniform traffic comparison, the hotspot traffic triggered more horizontal traffic offloading than uniform traffic. The post-CORD utilized 28% mid-haul links and 72% back-haul links. The post-CORD had more horizontal offloading traffic than pre-CORD because it introduced horizontal offloading between not only AN-MEC sites but also CN-MEC sites.

Further work that needs to be considered in the future includes the following: First, AN-MEC and CN-MEC site placement optimization is needed to cover an area representing an actual geographical area. Second, the future model also needs to consider a temporal characteristic such as dynamic user traffic rate, which changes over time at an MEC site. Third, heterogeneous service rates, which introduce a hyper-exponential service rate model, can be taken into consideration. Lastly, LA-TPIO should be considered for a scalable and efficient two-tier MEC system in the actual implementation of the MEC management plane.

ACKNOWLEDGMENT

This work was supported in part by Ministry of Science and Technology, Taiwan, and also in part by Lanner Inc.

REFERENCES

[1] O. Queseth *et al.*, *5GPPP architecture working group: View on 5G architecture (version 2.0, December 2017)*, English. Belgium: European Commission, Dec. 2017.

[2] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[3] F. Giust *et al.*, "MEC deployments in 4G and evolution towards 5G," *ETSI White Paper*, vol. 24, pp. 1–24, 2018.

[4] M. Satyanarayanan, "The emergence of edge computing," *Computer*, vol. 50, no. 1, pp. 30–39, 2017.

[5] C.-Y. Chang, K. Alexandris, N. Nikaein, K. Katsalis, and T. Spyropoulos, "MEC architectural implications for LTE/LTE-A networks," *Proceedings of the Workshop on Mobility in the Evolving Internet Architecture*, pp. 13–18, 2016.

[6] C.-Y. Li, H.-Y. Liu, P.-H. Huang, H.-T. Chien, G.-H. Tu, P.-Y. Hong, and Y.-D. Lin, "Mobile edge computing platform deployment in 4G LTE networks: A middlebox approach," in *USENIX Workshop on Hot Topics in Edge Computing (HotEdge 18)*, 2018.

[7] P. A. Apostolopoulos, E. E. Tsiropoulou, and S. Papavassiliou, "Risk-aware data offloading in multi-server multi-access edge computing environment," *IEEE/ACM Transactions on Networking*, vol. 28, no. 3, pp. 1405–1418, 2020.

[8] B. Yang, W. K. Chai, Z. Xu, K. V. Katsaros, and G. Pavlou, "Cost-efficient NFV-enabled mobile edge-cloud for low latency mobile applications," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 475–488, 2018.

[9] Y. Liao, L. Shou, Q. Yu, Q. Ai, and Q. Liu, "Joint offloading decision and resource allocation for mobile edge computing enabled networks," *Computer Communications*, vol. 154, pp. 361–369, 2020.

[10] Q. Zhang, L. Gui, F. Hou, J. Chen, S. Zhu, and F. Tian, "Dynamic task offloading and resource allocation for mobile-edge computing in dense cloud RAN," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3282–3299, 2020.

[11] S. Nath, Y. Li, J. Wu, and P. Fan, "Multi-user multi-channel computation offloading and resource allocation for mobile edge computing," *IEEE International Conference on Communications*, pp. 1–6, 2020.

[12] N. Wang, B. Varghese, M. Matthaiou, and D. S. Nikolopoulos, "ENORM: A framework for edge node resource management," *IEEE Transactions on Services Computing*, vol. 13, no. 6, pp. 1086–1099, 2017.

[13] Y. D. Lin, Y. C. Lai, J. X. Huang, and H. T. Chien, "Three-tier capacity and traffic allocation for core, edges, and devices for mobile edge computing," *IEEE Transactions on Network and Service Management*, vol. 15, no. 3, pp. 923–933, 2018.

[14] H. Yuan and M. Zhou, "Profit-maximized collaborative computation offloading and resource allocation in distributed cloud and edge computing systems," *IEEE Transactions on Automation Science and Engineering*, pp. 1–11, 2020.

[15] H. Li, H. Xu, C. Zhou, X. Lu, and Z. Han, "Joint optimization strategy of computation offloading and resource allocation in multi-access edge computing

environment," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 10 214–10 226, 2020.

[16] J. Xu, L. Chen, and S. Ren, "Online learning for offloading and autoscaling in energy harvesting mobile edge computing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 361–373, 2017.

[17] S. Guo, X. Hu, G. Dong, W. Li, and X. Qiu, "Mobile edge computing resource allocation: A joint Stackelberg game and matching strategy," *International Journal of Distributed Sensor Networks*, vol. 15, no. 7, 2019.

[18] Q. Li, J. Zhao, and Y. Gong, "Computation offloading and resource allocation for mobile edge computing with multiple access points," *IET Communications*, vol. 13, no. 17, pp. 2668–2677, 2019.

[19] S. Vimal, M. Khari, N. Dey, R. G. Crespo, and Y. H. Robinson, "Enhanced resource allocation in mobile edge computing using reinforcement learning based MOACO algorithm for IIOT," *Computer Communications*, vol. 151, pp. 355–364, 2020.

[20] Z. Tong, X. Deng, F. Ye, S. Basodi, X. Xiao, and Y. Pan, "Adaptive computation offloading and resource allocation strategy in a mobile edge computing environment," *Information Sciences*, vol. 537, pp. 116–131, 2020.

[21] F.-H. Tseng, M.-S. Tsai, C.-W. Tseng, Y.-T. Yang, C.-C. Liu, and L.-D. Chou, "A lightweight autoscaling mechanism for fog computing in industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4529–4537, 2018.

[22] T. X. Tran and D. Pompili, "Joint task offloading and resource allocation for multi-server mobile-edge computing networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 856–868, 2018.

[23] L. Peterson *et al.*, "Central office re-architected as a data center," *IEEE Communications Magazine*, vol. 54, no. 10, pp. 96–101, 2016.

[24] X.-L. Huang, X. Ma, and F. Hu, "Machine learning and intelligent communications," *Mobile Networks and Applications*, vol. 23, no. 1, pp. 68–70, 2018.

[25] J. Hecht, "Ultrafast fibre optics set new speed record," *New Scientist*, vol. 210, p. 24, 2011.

[26] I. Ajewale Alimi, N. Jesus Muga, A. M. Abdalla, C. Pinho, J. Rodriguez, P. Pereira Monteiro, and A. Luis Teixeira, "Towards a converged optical-wireless fronthaul/backhaul solution for 5g networks and beyond," *Optical and Wireless Convergence for 5G Networks*, pp. 1–29, 2019.

[27] P. G. Harrison, "Response time distributions in queueing network models," in *Performance Evaluation of Computer and Communication Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1993, pp. 147–164.

[28] Y. Jeon, S. Kuk, and H. Kim, "Reducing message collisions in sensing-based semi-persistent scheduling (SPS) by using reselection lookaheads in cellular V2X," *Sensors (Switzerland)*, vol. 18, no. 12, 2018.

[29] ETSI, "5G; Service requirements for next generation new services and markets (3GPP TS 22.261 version 15.5.0 Release 15)," vol. 0, p. 53, 2018.

[30] I. Fogg, *Benchmarking the global 5g experience — opensignal*. [Online]. Available: https://www.opensignal.com/2021/04/15/benchmarking-the-global-5g-experience-april-2021 (visited on 07/05/2021).

[31] L. Cominardi, L. M. Contreras, C. J. Bernardos, and I. Berberana, "Understanding qos applicability in 5g transport networks," *International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–5, 2018.

Widhi Yahya received his M.S. degree in Computer Science, National Central University (NCU), Taiwan in 2014. He is a lecturer and a researcher in computer science at University of Brawijaya, Indonesia. He is currently pursuing Ph.D. degree at Electrical Engineering and Computer Science (EECS) International Graduate Program of National Yang Ming Chiao Tung (NYCU). His research interest area are in network programming, software-defined networking, and multi-access edge computing (MEC) optimization.

Eiji Oki is a Professor at Kyoto University, Kyoto, Japan. He was with Nippon Telegraph and Telephone Corporation (NTT) Laboratories, Tokyo, from 1993 to 2008, and The University of Electro-Communications, Tokyo, from 2008 to 2017. From 2000 to 2001, he was a Visiting Scholar at Polytechnic University, Brooklyn, New York. His research interests include routing, switching, protocols, optimization, and traffic engineering in communication and information networks.

Ying-Dar Lin is a Chair Professor of computer science at National Yang Ming Chiao Tung University (NYCU), Taiwan. He received his Ph.D. in computer science from the University of California at Los Angeles (UCLA) in 1993. His research interests include network softwarization, cybersecurity, and wireless communications. His work on multi-hop cellular was the first along this line, and has been cited over 1000 times. He is an IEEE Fellow and IEEE Distinguished Lecturer. He has served or is serving on the editorial boards of several IEEE journals and magazines, and was the Editor-in-Chief of IEEE Communications Surveys and Tutorials (COMST), during 2016–2020.

Yuan-Cheng Lai received his Ph.D. degree from the Department of Computer and Information Science from National Chiao Tung University in 1997. He joined the faculty of the Department of Information Management at National Taiwan University of Science and Technology in August 2001 and has been a professor since February 2008. His research interests include performance analysis, network security, 5G mobile networks and Internet of Things.