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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

Energy-Efficient Joint Wireless Charging and Computation Offloading In MEC Systems

Rafia Malik and Mai Vu
Department of Electrical and Computer Engineering, Tufts University, MA, USA
Email: rafia.malik@tufts.edu, mai.vu@tufts.edu

***Abstract*—Edge networks offer a promising solution for satisfying the increasing energy and computation needs of user devices with new services including augmented and virtual reality. A mutil-access edge computing (MEC) system with collocated MEC servers and base-stations/access points (BS/AP) has the ability to support multiple users for both data computation and wireless charging. We propose an integrated solution for wireless charging with computation offloading to satisfy the largest feasible proportion of requested wireless charging while keeping the total energy consumption at the minimum, subject to the MEC-AP transmit power and latency constraints. We design a novel nested algorithm to optimally solve the non-convex problem in order to jointly perform data partitioning, time allocation, transmit power control and design the optimal energy beamforming for wireless charging. The proposed resource allocation scheme offers minimal energy consumption compared to other schemes while also delivering a higher amount of wirelessly transferred charge to the users. The results also show that compared to other solutions, the energy charging beams for minimum consumption have a wider main lobe, smaller side lobes, with an absence of the back lobe. Even with data offloading, the proposed solution shows significant charging performance, comparable to the case of charging alone, hence showing the effectiveness of performing partial offloading jointly with wireless charging.**

***Index Terms*—Edge computing, MEC, wireless power transfer, energy efficient network, optimization**

I. Introduction

Multi-access Edge Computing (MEC) is a promising technology which can provide cloud-computing capabilities within the radio access network in close vicinity to mobile subscribers [1]. Computation offloading can be beneficial, for instance, in video surveillance cameras offloading to the edge, or IoT devices or applications like AR/VR offloading their computation intensive tasks to the MEC servers. These servers can be co-located with radio base stations connected via backhaul to the internet core which is connected to the centralized cloud [2]. By moving the computing features to the edge, MEC can offer a distributed and decentralized service environment characterized by proximity, low latency, and high rate access [3] [4]. Currently, ETSI industry specification group is the only international standard available for MEC in the technology field, however, the 3rd Generation Partnership Project (3GPP) has started to include MEC in the 5G network standardization [5].

Radio frequency (RF) energy harvesting has lately garnered significant interest for communication systems with the prospect of far-field wireless power transfer which can enable energy-constrained devices to replenish their charge levels without physical connections. Energy harvesting is a promising technology which has shown some initial commercial deployments [6] [7]. There is also active research in RF power transfer ranging from signal design to maximize energy harvesting potential [8] to application centric research for using UAVs for wireless charging [9]. The potential and promise of wireless energy transfer technologies can come into realization as envisioned for beyond 5G and 6G systems [10].

A. Related Work

The availability of Ultra-High-Definition portable consumer devices and AR/VR applications fuels the growth of mobile video traffic, however, the limited battery lifetime of these devices poses a hindrance to the deployment of such power-hungry designs and computation intensive features [11]. To this end, the synergy between edge computing and wireless power transfer has the potential to provide battery sustainability and to alleviate the computation load. Dense deployments of multiple base-stations with co-located MEC servers [2] in close proximity to connected users can warrant the practicality of wireless charging and offer high access rates and computation capabilities. Such scenarios with edge computing and wireless power transfer are indeed envisioned for future 6G systems which will support AI-based applications and will embrace new radio access interfaces such as THz communications and intelligent surfaces [12].

Prior works have considered wireless charging in MEC systems under different implementations, for instance, wireless charging in cooperation assisted edge computing [13], UAV-enabled mobile edge computing [14] and MEC based heterogeneous networks [15]. Wireless power transfer has been considered in MEC networks for *self-sustained* devices, which rely on wireless charging as their sole power source, in relay-aided edge systems [13], single user [16] and multiple user systems [17]. Different from the concept of self-sustained devices which typically have low power requirements and/or low receiver sensitivity, *on-request* wireless charging can be more widely applicable where user devices use wireless charging to replenish their batteries. In the case of cellular networks, providing charging may can be a billable service assuming that the ground users have knowledge of their battery state, and can inform the MEC-AP about their battery level for requesting recharge in cases where their battery is critically low.

Computation offloading to the edge has been studied under two data models; binary offloading where the task is com-

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

pletely offloaded to the MEC for computation, or kept entirely at the user end for local computation; and partial offloading where the task can be disintegrated such that some of it is offloaded to the MEC and the remaining is computed locally. Modern mobile applications are composed of numerous procedures, for example, an AR/VR application can have multiple computation components such as video rendering, mapping and tracking, object recognition, etc, which allows implementing fine-grained (partial) computation offloading. Most prior works consider the traditional binary offloading scheme [18]–[21], and only recently, partial offloading has been considered for the problem of AP's energy minimization subject to users' latency requirement [17] [22] [31]. While partial offloading is more complicated to implement in comparison, however, it is more realistic with possible implementation in practical edge computing systems and has immense benefits in terms of energy consumption as shown in [22] [31].

Previous works have considered energy consumption in MEC systems, focusing on energy minimization either at the AP [14], [17] or at the user end [11], [16], [23]. In multi-user MEC systems, sequential offloading schemes, such as Time Division Multiple Access (TDMA) are typically assumed where different users offload their computation intensive tasks in a round-robin fashion [16], [17], [23], [24]. Under such sequential schemes, the time for offloading is significantly dominant compared to the computation time at the MEC or the time taken for downloading the computed results. Therefore, in such systems it is common to not optimize for downloading and/or computation time and only schedule and/or optimize the offloading time [16], [17], [23], [24].

B. This Work

In this work, we propose an integrated model which combines both computation offloading and wireless charging. Such models have versatile applicability to different use-cases. Examples include (i) AR/VR applications where complex processing tasks may be offloaded to the edge network for sharing and accessing different context information available in the network, (ii) human-machine interfaces used in smart factories where computation offloading prevents head-mounted AR/VR gear from becoming too warm and uncomfortable to wear [25], (iii) online gaming or training service data between two 5G connected devices [26], (iv) real-time map rendering for autonomous vehicular applications [22], and (v) professional low-latency periodic audio transport services for Audio-Video (AV) production applications, music festivals etc. [27]. Support for such AR/VR use-cases, smart factories and low latency AV production applications for music festivals, all are part of the 3GPP technical specifications and technical reports for 5G cellular networks.

RF wireless charging in these scenarios can help with the growing energy demands for such use-cases and is already on the horizon [6] [7]. Traditional wireless charging, or induction charging, requires large surface area contact to enable inductive charge transfer between the magnetic coils within both devices (charging device and the device being charged) and hence cannot work with devices such as AR/VR headsets which are curvy and not conducive to induction charging. Instead, over-the-air wireless charging which employs smart-lensing technology to focus energy beams for power transfer can offer wireless power solution for multiple applications and devices [28], [29]. Because of this impending need for RF wireless charging, smartphone companies have also initiated collaboration with such technology providers for RF power transfer [30]. RF wireless charging is especially suitable for integration in an edge network because of the proximity between edge servers and user devices and is envisioned as a native support for 5G MEC systems [32] [33].

While edge networks are envisioned in 5G and beyond systems to support both computation offloading and RF wireless charging, little work has examined both services jointly in the same system. We consider a multi-cell multi-user network scenario where access points equipped with massive MIMO antenna arrays and with co-located mobile edge computing servers offer both computation offloading and wireless charging. This is the first work to consider a joint optimization between computation offloading and wireless charging in order to minimize the total energy consumption, while satisfying a strict latency constraint on computation offloading and charging the largest amount as feasible. This joint formulation is different from others such as the sequential formulation for an opportunistic wireless charging scheme designed to maximize the received energy by wireless charging at the user after computation offloading in [31], or the self-sustainable model where the MEC server wirelessly charges end-devices as an incentive for them to subsequently offload computing tasks to the MEC in [34]. These other works do not aim at minimizing the transmitted energy consumption for wireless charging and hence lead to higher energy consumption. In this paper, we formulate a joint problem for resource allocation of data computation, communication and wireless charging resources with the aim of minimizing the overall system's transmitted energy consumption.

Major Contributions

1) This is the first work to consider an integrated system-level problem for joint resource allocation of data transmission, computation and wireless charging resources with shared constraints on latency and power. We formulate a novel problem to minimize a weighted sum of the energy consumption at all users in each cell and at the MEC server, considering the interference from other cells. For computation offloading, we consider partial offloading instead of traditional binary offloading. For wireless charging, we optimize multiple energy beams instead of using the single strongest sub-band for power transfer.
2) We design a novel nested algorithm to solve the non-convex integrated energy minimization problem. Our nested algorithm architecture includes an outer latency-aware algorithm which solves for data partitioning, and an inner two-step primal-dual algorithm which jointly solves for the optimal time allocation and beamforming matrix for

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

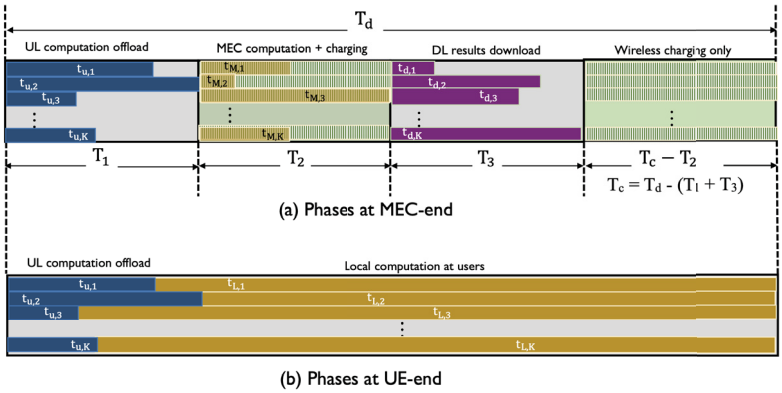

Figure 1: System functions at the MEC and user end that take place within the latency constraint T_d. The user operation contains 2 phases (bottom figure): data offloading and local computation. The MEC operation contains 4 phases (top figure), where wireless charging occurs during two different phases: (i) concurrently with MEC computation in phase II (always); and (ii) during the excess time in phase IV (only if MEC computation offloading finishes before latency constraint).

wireless charging. The proposed algorithm is proved to achieve the optimal solution of the original non-convex problem. Our algorithms also efficiently exploit the joint problem structure to achieve a complexity of $\mathcal{O}(K^3) + \mathcal{O}(K\log(K)/\epsilon^2) + \mathcal{O}((N^2+NK)/\epsilon^2)$, where K is the number of users per MEC-AP and N is the number of antennas at each AP.

3) Results using our algorithm show that our proposed integrated solution warrants higher energy efficiency: it delivers substantially more wireless charge to the users at a significantly lower transmitted energy consumption compared to isotropic and equal power wireless charging schemes. Even with data offloading, our joint algorithm shows that the amount of charged energy is significant and comparable with the case of charging alone, showing the effectiveness of performing partial offloading jointly with wireless charging. Our proposed joint resource allocation algorithm also shows faster convergence compared to sequential wireless charging schemes.

Notation: $\boldsymbol{X}$ and $\boldsymbol{x}$ denote a matrix and vector respectively, $\nabla^2 f(x)$ is the Hessian matrix, and $\nabla^2 f(x)^{-1}$ denotes its inverse. For an arbitrary size matrix, $\boldsymbol{Y}$, $\boldsymbol{Y}^*$ denotes the Hermitian transpose, and $\textbf{diag}(y_1, ..., y_N)$ is an $N \times N$ diagonal matrix with diagonal elements $y_1, ..., y_N$. $\boldsymbol{I}$ is an identity matrix, and $\mathbf{0}, \mathbf{1}$ are all zeros and all ones vector respectively. The standard circularly symmetric complex Gaussian distribution is denoted by $\mathcal{CN}(\mathbf{0}, \boldsymbol{I})$, with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{I}$. $\mathbb{C}^{k\times l}$ and $\mathbb{R}^{k\times l}$ denote the space of $k \times l$ matrices with complex and real entries, respectively.

II. System Model

We consider a system where $L \geq 1$ Access Points (APs), each co-located with an MEC Server, are deployed over a targeted zone/area, for instance in a sports stadium or an exhibition hall at a busy conference, serving ground users with computation offloading and power transfer. Each AP is equipped with a massive antenna array with N antennas while the user-devices are equipped with single antennas. These APs wirelessly charge (upon request) ground users in downlink, collect offloaded data from the users in uplink, and deliver computed results to users in downlink [1]. We consider K users requesting wireless charging service and sending data for computation offloading to each MEC-AP.

For computation offloading at each MEC, we consider the simple *data-partition model*, where the task-input bits are bit-wise independent and can therefore be arbitrarily divided into different groups to be executed by different entities [35]. We consider the case of partial offloading, such that for the ith user, the u_i computation bits are partitioned into q_i and s_i bits, where q_i bits are computed locally and s_i bits are offloaded to the MEC server. Assuming that such partition at the user-terminal does not incur additional computation bits, then $u_i = q_i + s_i$.

Energy at the user terminal is consumed for two tasks; 1) for local computation which depends on the CPU frequency used, and 2) for transmitting the data for computation offloading to the serving MEC-AP in the uplink which depends on the transmission time and power. Energy at the MEC server is consumed for three tasks; 1) for data computation of offloaded tasks, 2) for transmitting the results of computed data to its users in the downlink, and 3) for wireless charging in the downlink to the users requesting energy. Consider the case where wireless charging is requested jointly with computation offloading. Given a latency constraint of T_d, the time span for data offloading, computation at both the users and the MEC ends, wireless charging, and delivery of computed results to the user should not exceed T_d. For our formulation, we consider only one type of service, for example AR/VR applications. It is out of the current scope of our paper to consider multiple types of services which would require multiple values for T_d for different types of QoS when there are different types of services. Our assumption is in-line with the concept of 5G network slicing where each network slice is an isolated end-to-end network tailored to fulfill diverse requirements requested by a particular application [36]. As an extension however, if we are to consider multiple values for T_d, we can assume sub-groups among users e.g K_1, K_2, where K_1 users adhere to T_{d1} and K_2 users adhere to T_{d2}.

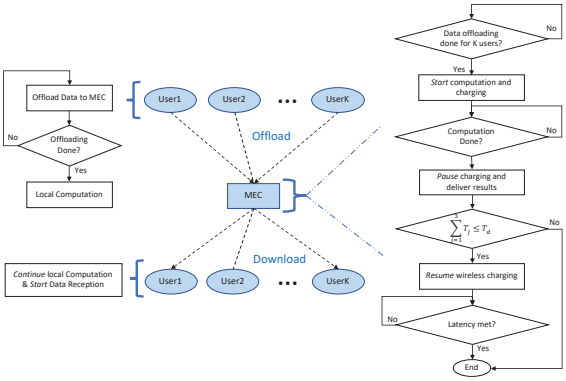

Figure 2: Flow diagram for the tasks performed at the MEC and each of the K users requesting computation offloading and wireless charging.

From the MEC-AP's perspective, the time duration for data offloading from all users to the MEC is denoted by T_1, the time for wireless charging is denoted as T_c, the computation for offloaded data at the MEC spans duration T_2, and the transmission of processed results occupies time T_3, such that $\sum_{j=1}^{3} T_j \leq T_d$. Note that wireless charging can happen concurrently with data computation at the MEC, and can continue after the results have been delivered to the users in downlink. From the user's side, the time taken for data offloading by the i^{th} user, $t_{u,i}$, and the time taken for local computation of any remaining data for this user, $t_{L,i}$ should meet the latency constraint, such that $t_{u,i} + t_{L,i} \leq T_d$. Figure 1 shows these system functions from both the users' and the MEC's perspective.

We assume that computation at the MEC is synchronous, such that the computation only starts once data from all users has been offloaded. While it is possible to perform fine-scale timing optimization by letting the MEC start computing immediately as soon as it receives a user's data, the gain from such a fine-grain optimization is expected to be non-substantial since the MEC time and energy consumptions for computation are both relatively small compared to those for wireless transmissions [22]. Such fine-grain optimization would also significantly increase the formulation and algorithm complexity. As such, in our model we assume that the MEC starts computation after receiving data from all users, where all users upload their data simultaneously. This model can be applied to both the cases of continuous data and of burst data offloading, where bursts are usually short enough such that multiple bursts can fit within the duration T_1 [25]. Figure 2 shows the workflow of offloading, computing and wireless charging processes at the user and the MEC.

A. Wireless Charging

In each cell, we consider K user terminals requesting wireless charging from the MEC-AP, where the i^{th} user requests e_i mJ of energy. To cater for the energy requests from the multiple users, the massive-MIMO enabled MEC-AP employs transmit energy beamforming in downlink to simultaneously charge multiple users.

Let $\boldsymbol{x_q}$ denote the energy bearing signal from the AP to the user-terminal, $\boldsymbol{W_q} \triangleq \mathbb{E}\left[\|\boldsymbol{x_q}\|^2\right]$ be the transmit covariance matrix, and $P_c = \text{tr}(\boldsymbol{W_q})$ be the power transmitted from the AP for wireless charging, in short, the charging power. Then the received (charged) power at the i^{th} user is given as

$$P_{h,i} = \xi_i \mathbb{E}\left[\left|\boldsymbol{h_i^*} \boldsymbol{x_q}\right|^2\right] = \xi_i \text{tr}(\boldsymbol{h_i^*} \boldsymbol{W_q} \boldsymbol{h_i}) \tag{1}$$

where $0 \leq \xi_i \leq 1$ is the energy conversion efficiency from Radio Frequency (RF) to Direct Current (DC), $\boldsymbol{h_i} \in \mathbb{C}^{N \times 1}$ is the channel from the AP to the i^{th} user. We assume a linear energy harvesting model since the received power per user is assumed to be constant over a single time block duration T_d, under strict latency constraints of the problem setting. However, to account for the difference in received power at each user location, each i^{th} user has its own energy conversion efficiency ξ_i based on the received power in the current time block. We define T_c as the time duration for wireless charging, where $T_c = T - (T_1 + T_3)$ and includes the time consumed by the computation phase, over which power is transferred to the users alongside computation at the users and at the MEC server. The energy consumed at the MEC server for power transfer, in short the charging energy, is given by

$$E_c = T_c \text{tr}(\boldsymbol{W_q}) \tag{2}$$

For the i^{th} user requesting e_i mJ of energy, the received (charged) energy, $E_{h,i}$, is given as

$$E_{h,i} = P_{h,i} T_c = \xi_i T_c \text{tr}(\boldsymbol{h_i^*} \boldsymbol{W_q} \boldsymbol{h_i}) \geq \alpha_i e_i \tag{3}$$

Here $0 \leq \alpha_i \leq 1 \ \forall i \in [1, K]$ is defined as an *energy ratio* auxiliary variable to ensure that the received energy is proportional to the requested amount such that only a portion of the requested energy may be charged if it is unfeasible for the AP to satisfy the user's energy request completely due to poor channel conditions or high energy request(s) by a single or few users. The ratio variable α_i therefore serves several reasons: to avoid over charging users' batteries, to conserve energy spending since charging is a billable service, and to ensure that no single user gets an unfairly large amount of the charged energy at the expense of others.

B. Data Transmission

1) Offloading Data in Uplink: In a given time slot, K single-antenna user terminals simultaneously offload to the N antenna AP. We consider $N \gg K$ such that the throughput becomes independent of the small-scale fading with channel hardening [37]. We define $\beta_i \triangleq S_\sigma d_i^{-\gamma}$ as the large scale fading between the i^{th} user and the AP, assuming it to be the same for all AP antennas (independent of N), where S_σ denotes log-normal shadowing with standard deviation σ dB, d_i is the distance from the i^{th} user to the AP, and γ is the path loss exponent. The very large signal vector dimension at a massive MIMO AP enables the use of linear detectors such as maximum ratio combining (MRC), in which case the uplink net achievable transmission rate for the i^{th} user in the l^{th} cell, $r_{u,i}$, is given as [38]

$$r_{u,i} = \nu \log_2\left(1 + \frac{\text{SINR}_{li}^{ul}}{\Gamma_1}\right),\ \text{SINR}_{li}^{ul} = \frac{N\gamma_{li}^l p_{li}}{\sigma_{1,li}^2} \tag{4}$$

where $\Gamma_1 \geq 1$ accounts for the capacity gap due to practical coding schemes, γ_{li} is the mean-square channel estimate, and p_{li} is the transmit power of the i^{th} user in the l^{th} cell. The constant ν represents the portion of transmission symbols spent on data transfer in the coherence interval τ_c. The interference and noise power, $\sigma_{1,li}^2$, includes the receiver noise power, plus interference caused by channel estimation errors due to pilot contamination, and inter-cell interference. Denote the *home cell* as the l^{th} cell, and denote cells using the same pilots as the home cell, called *contaminating cells*, by the set $\mathcal{P}_l$, where $l \in \mathcal{P}_l$. Then the interference plus noise power term is given as

$$\sigma_{1,li}^2 = \sigma_r^2 + \sum_{q\in\mathcal{P}_l}\sum_{i=1}^{K}\beta_{qi}^l p_{qi} + \sum_{q\notin\mathcal{P}_l}\sum_{i=1}^{K}\beta_{qi}^l p_{qi} + N\sum_{q\in\mathcal{P}_l\setminus l}\gamma_{qi}^l p_{qi} \tag{5}$$

where σ_r^2 is the receiver noise variance, the second term represents interference from contaminating cells, the third term is inter-cell interference, and the last term is interference due to the mean-square channel estimates from contaminating cells excluding the home cell and is also called the coherent interference [38].

The energy consumed for offloading the i^{th} user's data is given by $E_{OFF,i} = p_i t_{u,i}$, where p_i is the transmit power and $t_{u,i}$ is the transmission time for the i^{th} user. Let B denote the channel bandwidth, then $t_{u,i} = \frac{s_i}{Br_{u,i}}$. All users offload their computation bits simultaneously, and the total energy and time overhead for simultaneous data offloading is given as

$$E_{OFF} = \sum_{i=1}^{K}\frac{p_i s_i}{Br_{u,i}},\ T_1 = \max_{i\in[1,K]} t_{u,i}. \tag{6}$$

2) Downloading Results in Downlink: For the i^{th} user in the l^{th} cell, the downlink transmission rate with maximum ratio linear precoding at the MEC-AP is given as [38]

$$r_{d,i} = \log_2\left(1 + \frac{\text{SINR}_{li}^{dl}}{\Gamma_2}\right),\ \text{SINR}_{li}^{dl} = \frac{NP\gamma_{li}^l\eta_{lk}}{\sigma_{2,li}^2} \tag{7}$$

where $\Gamma_2 \geq 1$ is the capacity gap similar to (4), and the interference and noise power term is

$$\sigma_{2,li}^2 = \sigma_i^2 + P\sum_{q\in\mathcal{P}_l}\sum_{i=1}^{K}\beta_{qi}^l\eta_{qi} + P\sum_{q\notin\mathcal{P}_l}\sum_{i=1}^{K}\beta_{qi}^l\eta_{qi} + NP\sum_{q\in\mathcal{P}\setminus l}\gamma_{qi}^q\eta_{qi} \tag{8}$$

where σ_i^2 is the noise at the i^{th} user terminal in the l^{th} cell, $\{\eta_{li}\} \in [0,1]$ are the power coefficients satisfying $\sum_{i=1}^{K}\eta_{li} \leq 1$ for all l, and P is the AP's average transmit power. Similar to the uplink transmission, the second term in (8) is pilot contamination, the third term is inter-cell interference which manifests as uncorrelated noise in the home cell, and the last term is coherent interference resulting from mean-square channel estimation errors. Since there is no pilot transmission in this phase, the effective downlink transmission rate is equal to the data rate [22].

The transmission time for delivering the i^{th} user's computation results can be written in terms of the downlink rate in (7) as $t_{d,i} = \frac{\tilde{s}_i}{Br_{d,i}}$. Here $\tilde{s}_i$ denotes the number of information bits generated after processing s_i offloaded bits of the i^{th} user, and is assumed to be proportional to s_i, that is $\tilde{s}_i = \mu s_i$. The AP simultaneously transmits computed results for all users, and the total energy and time overhead for results downloading are then given as

$$E_{DL} = \sum_{i=1}^{K}\frac{P\eta_i\mu s_i}{Br_{d,i}},\ T_3 = \max_{i\in[1,K]} t_{d,i}. \tag{9}$$

C. Data Computation

1) Local computation at the users: The time for computation depends on the amount of data to be computed and the CPU cycle frequency. The energy consumption and the processing time for local computation at the i^{th} user is given as [35]

$$E_{LC} = \sum_{i=1}^{K}\kappa_i c_i(u_i - s_i)f_{u,i}^2,\quad t_{L,i} = \frac{c_i(u_i - s_i)}{f_{u,i}} \tag{10}$$

where κ_i is the effective switched capacitance, $f_{u,i}$ denotes the average CPU frequency, c_i denotes the CPU cycle information, and $q_i = u_i - s_i$ is the total number of bits required to be locally computed at i^{th} user respectively. The users' local computation time can also extend to Phase III while the MEC is sending computed results back to users. This fact is considered later in the problem formulations.

2) Computation of the offloaded data at the MEC server: MEC servers, with high computation capacities, compute the tasks of all users in parallel [39] [35]. The energy and time consumed for computing offloaded bits is given as

$$E_{OC} = \sum_{i=1}^{K}\kappa_m f_{mi}^2 d_m s_i,\quad t_{M,i} = \frac{d_m s_i}{f_{mi}}\ \forall i \in [1,K],$$
$$T_2 = \max\{t_{M,i}\}. \tag{11}$$

where $t_{M,i}$ is the time for computing i^{th} user's offloaded task, s_i is the number of bits offloaded by the i^{th} user to the MEC, d_m is the number of CPU cycles required to compute one bit

Table I: Symbols Table

Symbol	Definition	Variable	Parameter
$\boldsymbol{h_i}$	channel from the AP to the i^{th} user		✓
$\boldsymbol{W_q}$	transmit covariance matrix	✓	
e_i	energy requested (mJ) by i^{th} user		✓
$\boldsymbol{\alpha}$	energy ratio auxilliary variable	✓	
ϵ_i	energy conv. efficiency for i^{th} user		✓
Γ_1/Γ_2	capacity gap for UL/DL		✓
$\sigma^2_{1/2,i}$	interference for UL/DL	✓	
β_i	large scale fading		✓
γ_i	mean-sqaure channel estimate		✓
p_i	i^{th} user's transmit power	✓	
P	AP's average transmit power		✓
B	channel bandwidth		✓
$\boldsymbol{f_{u/m}}$	CPU cycle frequency at user/MEC		✓
u_i	computation bits for i^{th} user		✓
$\boldsymbol{t_u}, \boldsymbol{t_d}$	offload/download time	✓	
$r_{u/d,i}$	UL/DL transmission rate	✓	
η_i	i^{th} user's power coefficient in DL	✓	
$\boldsymbol{s}$	offloaded bits	✓	
$t_{M/L,i}$	i^{th} user's compute time at MEC/user	✓	

at the MEC, f_{mi} is the CPU frequency assigned to the i^{th} user's task, and κ_m is the effective switched capacitance of the MEC server.

For our formulation to follow in Section III, we consider equal frequency allocation for users' tasks, that is $f_{m,i} = f_m \ \forall i$, based on previous results in [22] showing that dynamic frequency allocation has little effect on the system energy consumption since in a typical network setting, the wireless transmission energy consumption is significantly dominant compared to the computation energy consumption. Table I summarizes all the symbols we use in modeling the system.

III. Formulation and Analysis of Joint Energy Minimization

Considering a multi-cell multi-MEC network, we formulate an edge computing problem which explicitly accounts for physical layer parameters including available transmit powers from each user and the MEC, associated massive MIMO data rates with realistic pilot contamination and interference. For simplicity of notation, we assume that all K users which are offloading their computation to the MEC server are also requesting wireless charging.

A. Joint Energy Minimization Problem Formulation

The total energy consumption by all users, based on equations (6) and (10), can be written as

$$E_u = \sum_{i=1}^{K} \left[\frac{t_{u,i}(2^{\frac{s_i}{\nu t_{u,i} B}} - 1)\Gamma_1 \sigma^2_{1,i}}{N\gamma_i} + \kappa_i c_i (u_i - s_i) f^2_{u,i} \right] \quad (12)$$

Similarly, the total energy consumption at the MEC server, based on equations (2), (9) and (11), can be written as $E_m = E_{m1} + E_{m2}$ where

$$E_{m1} = \sum_{i=1}^{K} \left[\frac{t_{d,i}(2^{\frac{\mu s_i}{t_{d,i} B}} - 1)\Gamma_2 \sigma^2_{2,i}}{N\gamma_i} + \kappa_m d_m f^2_{mi} s_i \right] \quad (13)$$

$$E_{m2} = (T_d - T_1 - T_3)\text{tr}(\boldsymbol{W_q}) \quad (14)$$

Here E_{m1} is the energy consumed for computation and transmission and E_{m2} is the energy consumed for wireless charging. In these expressions, using (4) and (7), and by definition of the uplink and downlink transmission rates as $r_{u,i} = \frac{s_i}{\nu t_{u,i} B}$ and $r_{d,i} = \frac{\mu s_i}{t_{d,i} B}$ respectively, we have implicitly replaced the power allocation variables for per-user uplink transmission power (p_{li}) and per-user downlink power (η_{li}) as functions of the time allocation and data partitioning as follows

$$p_{li} = \frac{(2^{\frac{s_i}{\nu t_{u,i} B}} - 1)\Gamma_1 \sigma^2_{1,i}}{N\gamma_i}, \quad \eta_{li} = \frac{(2^{\frac{\mu s_i}{t_{d,i} B}} - 1)\Gamma_2 \sigma^2_{2,i}}{PN\gamma_i} \quad (15)$$

Below we discuss an integrated formulation which jointly optimizes for the wireless charging transmit beamforming matrix, the amount of data offloaded from each user, and the time duration for each phase within a total latency requirement with aim of system level energy minimization. The joint energy minimization problem can be written as

$$(P_{\text{int}}): \min_{\boldsymbol{s},\boldsymbol{t},\boldsymbol{W_q}} E_{\text{total}} = (1-w)E_u + w(E_{m1} + E_{m2}) \quad (16)$$

$$\text{s.t. Eqs. } (12) - (14) \quad \text{(a-c)}$$

$$\sum_{j=1}^{3} (T_j) \leq T_d, \quad \text{(d)}$$

$$\frac{c_i(u_i - s_i)}{f_{u,i}} + t_{u,i} - T_d \leq 0 \ \forall i \in [1, K] \quad \text{(e)}$$

$$t_{u,i} - T_1 \leq 0 \ \forall i \in [1, K], \quad \text{(f)}$$

$$t_{d,i} - T_3 \leq 0 \ \forall i \in [1, K], \quad \text{(g)}$$

$$\frac{d_m s_i}{f_{mi}} - T_2 \leq 0 \ \forall i \in [1, K], \quad \text{(h)}$$

$$T_c = T_d - T_1 - T_3 \quad \text{(i)}$$

$$\text{tr}(\boldsymbol{W_q}) - P \leq 0, \quad \text{(j)}$$

$$\xi_i \text{tr}(\boldsymbol{h_i^* W_q h_i}) T_c - \alpha_i e_i \geq 0 \quad \text{(k)}$$

Here E_{total} is weighted sum of energy consumed at all users (E_u given in (12)) and the MEC (E_{m1} and E_{m2} given in (13) and (14)), with $1-w$ and w as the respective weights. The optimizing variables of this problems are time allocation $\boldsymbol{t} = [t_{u,1}...t_{u,K}, t_{d,1}...t_{d,K}, T_1, T_2, T_3, T_c]$, offloaded data $\boldsymbol{s} = [s_1...s_K]$, and beamforming matrix for wireless charging $\boldsymbol{W_q} \in \mathbb{R}^{N \times N}$. Given parameters of the problems are T_d as the total latency constraint, P as the AP's transmit power, B as the channel bandwidth, Γ_1, Γ_2 as the uplink and downlink capacity gaps, (κ_i, c_i) and (κ_m, d_m) as the switched capacitance and CPU cycle information at the users and the MEC respectively.

Parameter T_d is the total latency constraint, and (d) represents the constraint that both the time consumed for all three phases at the MEC, and the time consumed for offloading $\boldsymbol{t_u}$ and local computation at each user $\boldsymbol{t_L}$ should not exceed T_d. Constraints (f-h) show that the time consumed separately for offloading $\boldsymbol{t_u}$, computation of users' tasks at the MEC $\boldsymbol{t_M}$, and downloading time $\boldsymbol{t_d}$ for each user's results must be less than the maximum allowable time, $\{T_1, T_2, T_3\}$, for that phase as given in {(6),(11), (9)} respectively. Constraint (h) denotes

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

that wireless charging occupies all the time within T_d outside the data transmission operations.

In terms of power constraints, (j) represents the maximum transmission power of the AP. which can be used for wireless charging. Note that wireless charging and data transmission from MEC do not occur at the same time, and the power constraint for data transmission is implicitly included in (15) where the transmit power allocation in downlink is constrained such that $\sum_{i=1}^{K} \eta_{li} \leq 1$. Constraint (k) shows that the amount of received (charged) energy at the i^{th} user is proportional to the amount of requested energy depending on the availability of the system. Here the proportional factor α_i $(0 \leq \alpha_i \leq 1)$ is an auxiliary variable to ensure feasibility of the charging problem for cases when the available time or MEC-AP power for wireless charging cannot satisfy the full requested energy amount. We are interested in the largest values of α_i for which the problem is feasible, as such these α_i values will also be optimized.

P_{int} jointly optimizes for both computation offloading variables and wireless charging beams. To our knowledge, this is the first formulation to consider such a joint optimization. It is different from [31] where we considered two sequential problems, one to minimize the transmitted energy consumption for computation offloading, followed by the other to maximize the energy received by the users through wireless charging. Here the objective for (P_{int}) is to jointly minimize the total transmitted energy consumption for both computation offloading and wireless charging. The inclusion of transmitted energy for wireless charging in the minimizing objective function consequently leads to completely different algorithms and optimization results.

B. Problem Analysis and Decomposition

In this section we analyze the integrated problem (P_{int}) and show that they can be decomposed into simpler problems. The multivariable problem in (16) is a non-linear and non-convex optimization problem. This is due to constraint (16b) in which the term $f_{mi}^2 s_i$ is neither convex nor concave since its Hessian is indefinite with one positive and one negative eigenvalue, making this constraint and consequently problem (P_{int}) non-convex. We can show, however, that the objective function f_0 for (P_{int}) is a convex function of s_i since the second-order derivative for the objective function with respect to s_i is positive for all considered ranges of problem parameters. [22]. Furthermore, if the gradient of $f_0(\cdot)$ with respect to s_i evaluated at $s_i = 0$ satisfies the non-negativity condition

$$\left\{ \frac{(1-w)2^{\frac{s_i}{\nu t_{u,i} B}} \ln 2 \Gamma_1 \sigma_{1,i}^2}{\nu B N \gamma_i} + \frac{w\mu 2^{\frac{\mu s_i}{t_{d,i} B}} \ln 2 \Gamma_2 \sigma_{2,i}^2}{B N \gamma_i} + w\kappa_m d_m f_{m,i}^2 - (1-w)\kappa_i c_i f_{u,i}^2 \right\} \Bigg|_{s_i \to 0} \geq 0, \tag{17}$$

then the total energy in problem (P_{int}) is an increasing function of each s_i. For typical network settings, with multiple APs and users located in a reasonable size target area, because of the dominant energy consumptions for wireless transmissions and MEC computation over local computation, the condition in (17) will hold true [22]. On the other hand, by offloading data to the MEC, the total computation time can be reduced. Therefore, there exists an optimal point, $s_i^\star$ $\forall i \in [1, K]$, which minimizes E_{total} within the latency constraint.

If offloaded data $\boldsymbol{s}$ is fixed, then problem (P_{int}) turns out to be convex in the remaining variables as stated in the following lemma. Lemma 1 lets us decompose the original non-convex problem (P_{int}) into simpler convex subproblems which will be used in the subsequent algorithm design.

Lemma 1. *For a given set of offloaded data $\boldsymbol{s}$, the problem (P_{int}) is convex in the remaining variables $\boldsymbol{t}, \boldsymbol{W_q}$.*

Proof. Proof follows by examining each constraint and showing that with fixed s_i, it is a convex function. Details in Appendix A. □

Since CPU frequencies are not optimizing variables due their negligible impact on the total energy consumption [22], for a given value of the offloaded data s_i, the computation time for the offloaded data T_2 can be pre-determined in closed form directly from constraint (g) in (16) and hence constraint (16g) can be excluded from the problem (P_{int}). Also, at a fixed value of $\boldsymbol{s}$, considering wireless charging as an opportunistic feature in addition to computation offloading, (P_{int}) is also separable in $\boldsymbol{t}$ and $\boldsymbol{W_q}$ as stated in the lemma below.

Lemma 2. *Given that wireless charging is opportunistic, at a fixed value of s_i, problem (P_{int}) is separable in terms of variable $\boldsymbol{t}$ and $\boldsymbol{W_q}$ as follows*

$$(P2) \quad \min_{\boldsymbol{t}} \quad (1-w)E_u + wE_{m1} \quad \textit{s.t.} \quad (16\textit{a,b,d-f}) \tag{18}$$

$$(P3) \quad \min_{\boldsymbol{W_q}} \quad wE_{m2} \quad \textit{s.t.} \quad (16\textit{c,h-j}) \tag{19}$$

This decomposition of (P_{int}) into sub-problems (P2) and (P3) is optimal and retains the optimality of the solutions, where the optimal $T_c^\star$ from (P2) is used as a parameter for (P3).

Proof. The objective function of (P_{int}) in (16) constitutes of distinct components as separate functions of the time allocation variables $t_{u,i}, t_{d,i}$ and the transmit covariance matrix $\boldsymbol{W_q}$, as seen from the expressions for E_u, E_{m1} and E_{m2} in (12), (13) and (14) respectively. Thus P_{int} can be divided into two sub-problems of minimizing the energy consumption for wireless charging (P3) and for computation offloading (P2), where the only variable coupling these two problems is T_c which must satisfy $T_c = T_d - T_1 - T_3$. If we fix the value of T_c, the two problems are then completely separable in terms of the offloading and the wireless charging variables. The question then becomes what is the optimal value for T_c.

Consider the computation offloading problem, the "best" T_c is the optimal $T_c^\star$ resulting from solving problem (P2), since that value of T_c corresponds to the minimum energy consumed for computation offloading. Any change from $T_c^\star$ will result in an increased energy consumption for computation offloading.

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

For wireless charging, since the goal is to minimize the amount of consumed energy while satisfying the largest feasible portion of the request (by picking the largest $\boldsymbol{\alpha}$ feasible), it is of interest for wireless charging to be able to have the largest possible T_c while the largest feasible $\alpha_i < 1 \ \forall i$. Noting also that the energy consumption for wireless charging is a monotonously increasing function of T_c, thus taking $T_c = T_c^\star$, the optimal value from the computation offloading problem (P2), is also optimal for wireless charging while feasible $\boldsymbol{\alpha} < 1$. Any increase of T_c beyond this value $T_c^\star$ will increase the energy consumption for both computation offloading and wireless charging. On the other hand, if $\boldsymbol{\alpha} = 1$ is feasible with $T_c = T_c^\star$, that means the charging phases can fully satisfy the charging requests within a time duration less than $T_c^\star$. Therefore, $T_c^\star$ now acts as an upper bound on the time necessary for wireless charging, and changing T_c slightly to a smaller value from $T_c^\star$ does not change the energy consumption for charging, while increasing the energy consumption for computation offloading. Thus again $T_c^\star$ results in the minimum amount of energy consumption for both computation offloading and wireless charging while satisfying the charing requests.

Thus in all cases of the largest feasible α value, the optimal $T_c^\star$ resulting from solving sub-problem (P2) is optimal for the joint problem (P_{int}). Thus (P_{int}) can be optimally decomposed into two sub-problems (P2) and (P3) where the optimal value $T_c^\star$ from sub-problem (P2) is used as a parameter for sub-problem (P3).

□

C. Wireless Charging With Largest Feasible $\boldsymbol{\alpha}$

The goal of sub-problem (P3) is to design the optimal energy beamforming $\boldsymbol{W_q}$ to minimize the energy consumption for wireless charging (during both computation phase II and the excess latency time), while satisfying the largest portion of the energy requests as feasible. We re-write (P3) with relevant constraints from (16) as follows

$$
\begin{aligned}
(P3): \ \min_{\boldsymbol{W_q}} \quad & T_c \text{tr}(\boldsymbol{W_q}) && (20)\\
\text{s.t.} \quad & \text{tr}(\boldsymbol{W_q}) \le P && (a)\\
& \xi_i \text{tr}(h_i^* \boldsymbol{W_q} h_i) T_c \ge \alpha_i e_i \quad \forall i = 1...K && (b)
\end{aligned}
$$

Here α_i is an auxiliary variable representing a proportion of the requested amount, whereas $0 < \alpha_i \le 1$, with the largest feasible value of α_i will be sought for the optimal solution.

Before analyzing (P3), it is worthwhile noting the difference between this formulation and others which maximize the amount of charged (received) energy $\xi_i \text{tr}(h_i^* \boldsymbol{W_q} h_i) T_c$, such as the one considered in [31]. These two different objectives lead to different constraints where the amount of received energy is upper-bounded in [31] to ensure that it is no more than requested. Here since the goal is to minimize the amount of transmitted energy, the amount of received energy is instead lower-bounded.

$P3$ introduces a best-feasibility approach towards wireless charging, such that the energy delivered to the users is the largest feasible while also minimizing the overall energy consumption. This best-feasibility result is obtained by using the auxiliary variable α_i to ensure that the received energy is at least an α_i portion of the requested energy. The largest values for α_i which are feasible, that is, ensuring problem (P3) have a energy beamforming solution within the power constraint and available time, will be sought as the solution of the problem.

IV. Optimality Conditions and Optimal Solutions

In this section, we analyze the optimality conditions for the two sub-problems established in Section III to derive the optimal time durations for computation offloading and the optimal beam directions for wireless charging as functions of the dual variables. We also derive the solution for the largest feasible value of auxiliary variable $\boldsymbol{\alpha}$, which provides the portion of requested energy that can be charged by the transmit power constraint within one time slot. These optimality conditions are then used in designing a nested algorithm in the next section for solving the original problem P_{int}.

A. Optimal Time Durations For Computation Offloading

Here we present the solution for the optimal time allocation for the computation offloading problem (P2). Since the problem is convex based on Lemma 1, we adopt a primal-dual solution using the Lagrangian duality analysis similar to that proposed in [22, Theorem 1] and derive the optimal solution as given in Theorem 1 below.

Theorem 1. *The offloading and downloading time, $t_{u,i}$ and $t_{d,i}$ respectively, can be obtained as a solution of the form*

$$
x = \frac{cB}{\ln 2}\Big(W_0\Big(\frac{-y}{\sigma^2 e} - \frac{1}{e}\Big) + 1\Big) \tag{21}
$$

where $y = -\frac{\beta_i + \theta_i}{(1-w)}$, $x = x_{1,i} = \frac{1}{t_{u,i}}$, $c = \frac{\nu}{s_i}$, $\sigma^2 = \frac{\Gamma_1 \sigma_{1,i}^2}{N\gamma_i}$ to solve for $t_{u,i}$, and $y = \frac{-\phi_i}{w}$, $x = x_{2,i} = \frac{1}{t_{d,i}}$, $c = 1/\mu s_i$, and $\sigma^2 = \frac{\Gamma_2 \sigma_{2,i}^2}{N\gamma_i}$ to solve for t_d, i. Here θ_i, β_i and ϕ_i are the dual variables associated with the constraints (d), (e) and (g) of problem (P_{int}) in (16) respectively.

Proof. The solution in (21) can be obtained directly by applying KKT conditions on the Lagrangian dual of the problem (P2) or $P_{\text{seq,CO}}$ with respect to $t_{u,i}$ and $t_{d,i}$. Detailed proof can be obtained using an approach similar to that in [22, Theorem 1] and is omitted for brevity. □

B. Optimal Energy Beam Directions and Power

Problem (P3) is a semi-definite programming with linear objective function and linear constraints and hence is convex. We can show that strong duality holds since Slater's condition is satisfied, that is, we can find a strictly feasible point ($\boldsymbol{W_q} = p\boldsymbol{I}_{N\times N}$, $p \le P/N$, $0 \le \alpha_i \le 1 \ \forall i$) in the relative interior of the domain of the problem where the inequality constraints hold with strict inequalities [40].

From the definition of charging time, as $T_c = T_d - T_1 - T_3$, the problem (P3) has an interdependency on the optimization problem (P2). However, based on Lemma 2, since (P2) and (P3) are separable, we can use the optimal time allocation

obtained as a solution of (P2) to find the energy beamforming matrix, $\boldsymbol{W_q}$ in (P3). Theorem 2 below provides the optimal beam directions for wireless charging.

Theorem 2. *Let the eigenvalue decomposition of the optimal energy beamforming matrix be* $\boldsymbol{W_q}^\star = \boldsymbol{U_q}\boldsymbol{\Lambda_q^\star}\boldsymbol{U_q^*}$, *where* $\boldsymbol{U_q} \in \mathbb{R}^{N\times N}$ *defines the directions of energy beams and diagonal* $\boldsymbol{\Lambda_q^\star}$ *is the beam power allocation matrix. Then the optimal directions for energy beams are* $\boldsymbol{U_q^\star} = \boldsymbol{U_B}$, *where* $\boldsymbol{U_B}$ *is obtained from the eigenvalue decomposition of* $\boldsymbol{B} = \boldsymbol{U_B}\boldsymbol{\Lambda_B}\boldsymbol{U_B^*}$, *such that* $\lambda_{B,1} \leq \ldots \leq \lambda_{B,N}$, *where*

$$\boldsymbol{B} = (T_c + \lambda_5)\boldsymbol{I} - \xi_i T_c \sum_{i=1}^{K} \psi_i \alpha_i \boldsymbol{h_i}\boldsymbol{h_i^*}$$

Here λ_5 *and* ψ_i *are the dual variables associated with constraint (20a) and the* i^{th} *constraint in (20b) respectively.*

Proof. See Appendix B. □

Theorem 2 provides the optimal directions of the energy beams for the beamforming matrix, $\boldsymbol{W_q}$. What is left now is to obtain the optimal power allocation across the energy beams, that is, the eigenvalues of the transmit covariance matrix for wireless charging. To this end, we substitute the optimal beam directions from Theorem 2 into (P3) and re-write the formulation in terms of the beam power allocation only as (P4) below. Beam power allocation, $\boldsymbol{\lambda_q}$, can then be obtained as a solution to a Linear Programming (LP) problem given in Theorem 3 below.

Theorem 3. *The optimal beam power allocation is derived by solving the LP problem below*

$$(P4): \min_{\boldsymbol{\lambda_q}} \quad \sum_{i=1}^{K} \lambda_{q,i} \tag{22}$$

$$s.t. \quad \sum_{i=1}^{K} \lambda_{q,i} \leq P, \tag{a}$$

$$\lambda_{q,1} \geq ... \geq \lambda_{q,K} \geq 0 \tag{b}$$

$$\boldsymbol{A}\boldsymbol{\lambda_q} \geq \text{diag}(\boldsymbol{\alpha})\boldsymbol{b} \tag{c}$$

where $\boldsymbol{\lambda_q} = [\lambda_{q,1}, ..., \lambda_{q,K}]^T$, $\boldsymbol{A} \in \mathbb{R}^{K\times K} = [\boldsymbol{a_1^*}...\boldsymbol{a_K^*}]$, $\boldsymbol{a_i}^* = \text{diag}(\boldsymbol{q_i}\boldsymbol{q_i^*})$, $\boldsymbol{q_i^*} = \boldsymbol{h_i^*}\boldsymbol{U_B}$ *and* $\boldsymbol{b} \in \mathbb{R}^{K\times 1} = [\pi_1...\pi_K]$, $\pi_i = \frac{e_i}{\xi_i T_c}$ $\forall i = 1...K$.

Proof. See Appendix B. □

C. Largest Feasible Charged Ratio $\boldsymbol{\alpha}$

The inclusion of the auxiliary α_i variables in the original problem (P_{int}) in (16) for feasibility and consequently finding the largest feasible values of α_i for wireless charging is a novel feature of this formulation. As such, $\boldsymbol{\alpha}$ appears in problem (P4) as an auxiliary variable to ensure that the amount of charged energy is feasible within the transmit power constraint. While different values of $\boldsymbol{\alpha}$ will result in different power allocation, we are interested in the largest $\boldsymbol{\alpha} \in [\boldsymbol{0}, \boldsymbol{1}]$ that makes (P4) feasible, so that the amount of received energy is largest while minimizing the transmit power.

Solving for the largest feasible $\boldsymbol{\alpha}$ usually requires establishing a sequence of feasibility problems, where we increase the value of $\boldsymbol{\alpha}$ in each subsequent problem until the problem just becomes infeasible. In (P4), however, we are able to exploit the problem structure to solve for the largest feasible $\boldsymbol{\alpha}$ in closed form, thus requiring no separate algorithms for finding $\boldsymbol{\alpha}$. The following lemma provides the optimal value of $\boldsymbol{\alpha}$.

Lemma 3. *The largest energy ratio auxiliary variable* $\boldsymbol{\alpha}$ *which ensures that (P3) and (P4) stay feasible is obtained as*

$$\boldsymbol{\alpha} = \begin{cases} \boldsymbol{1}, & \sum_{i=1}^{K} \lambda_{q,i}^{(0)} <= P \\ \text{diag}(\boldsymbol{b})^{-1}\boldsymbol{A}\boldsymbol{\lambda}_q^{(0)} \frac{P}{\sum_{i=1}^{K} \lambda_{q,i}^{(0)}}, & \textit{otherwise} \end{cases} \tag{23}$$

where $\boldsymbol{\lambda}_q^{(0)}$ *are the optimal values of* $\lambda_{q,i}$ *obtained from (P4) when setting* $\alpha_i = 1$ *and sum power constraint* P *is removed.*

Proof. Since the optimal solution of the LP in (P4) is linear in the constraint P, we can solve this problem without loss of optimality by first setting $\boldsymbol{\alpha} = \boldsymbol{1}$ and removing the power constraint, then solve for the resulting LP. If the sum of solved $\lambda_{q,i}^{(0)}$ is more than P, then $\boldsymbol{\alpha}$ will be the scaling vector to bring this sum to be equal to P while still satisfying constraint (22c) with equality, and all optimal values $\lambda_{q,i}^\star$ will be scaled by α_i. Otherwise α_i stays as 1 and $\lambda_{q,i}^\star$ stays unchanged as $\lambda_{q,i}^{(0)}$. The largest energy ratio $\boldsymbol{\alpha}$ is hence obtained as in (23). □

V. A Nested Algorithm

While problem (P_{int}) is not convex in all the optimizing variables, Lemma 1 shows that by fixing the offloaded bits $\boldsymbol{s}$, the problem is convex in all the remaining optimizing variables with a convex objective function and a convex feasible set. This suggests an iterative procedure where we can fix the offloaded bits $\boldsymbol{s}$ and solve for the rest of the variables, then adjust $\boldsymbol{s}$ and repeat until convergence is achieved. As long as the gradient condition (17) holds in typical network settings, the total energy consumption is increasing with $\boldsymbol{s}$ and the optimization in terms of $\boldsymbol{s}$ can be achieved using a descent algorithm with an added criterion for the latency.

When fixing $\boldsymbol{s}$ and solving for the rest of the variables including the time allocation $\boldsymbol{t}$ and transmit covariance matrix $\boldsymbol{W_q}$, instead of using a convex solver which is unable to exploit the problem structure and hence can be inefficient, we make use of Lemma 2 and optimally divide this convex problem further into two sub-problems: problem $(P2)$ to find the optimal time allocation $\boldsymbol{t}^\star$, and $(P3)$ to solve for the transmit covariance matrix $\boldsymbol{W_q^\star}$.

Note that the decompositions into sub-problems still maintain the optimality of the solution for the original joint problem. We next propose an optimal and customized nested algorithm which includes an outer algorithm to determine $\boldsymbol{s}^\star$ and an inner two-step algorithm to solve for $\boldsymbol{t}^\star$ and $\boldsymbol{W_q^\star}$ to efficiently reach the solution for problem (P_{int}).

A. Nested Algorithm Architecture

Based on Lemma 1, the algorithm for solving (P_{int}) is designed to have a nested architecture with an outer and an

inner loop, in which the outer loop solves for s_i decrementally while the inner loop solves for the remaining variables at a fixed value of s_i. Specifically, the nested algorithms work as follows.

1) Outer Latency-Aware Descent Algorithm for s_i: We first initialize the offloaded bits $\boldsymbol{s}$ and the dual variables in the outer algorithm. At the current value of s, the inner algorithm is executed, for which we use a primal-dual approach employing a subgradient method. At convergence where the stopping criterion for the dual problem is satisfied, the inner algorithm returns the control to the outer algorithm. Based on the newly updated primal solution from the inner algorithm, we proceed to updating $\boldsymbol{s}$ by some Δs_i for each user for the next iteration of the outer algorithm, using a latency aware descent algorithm [40]. Similar to [22], the latency aware descent algorithm is based on the standard Newton's method with a novel modification to the classical stopping criterion to account for the latency constraint [22]. The latency based stopping criterion is given as

$$T_{\text{total}} = \max\ (t_{u,i} + t_{L,i}, \sum_{j=1}^{3} T_j) \leq T_d \tag{24}$$

The outer algorithm works as follows. We initialize $0 < s_{i,0} < u_i$, input the simulation parameters, and update the step or search direction $\Delta \boldsymbol{s}$. We then execute the inner algorithm for finding the optimal time and frequency allocation for the given value of $\boldsymbol{s}$. Next we proceed to the sequential update of s_i. We use backtracking line search to find the step-length at the k^{th} iteration as the vector $\boldsymbol{t}^{(k)}$, with $t_i^{(k)}$ as the step-length for the i^{th} user, and update the offloaded bits for the next iteration as $s_i^{(k+1)} = s_i^{(k)} + t_i \Delta s_i$. We then check the stopping criteria for convergence of the outer algorithm. In this step, we introduce a novel modification to the classical stopping criterion for descent methods, which is necessary to arrive at the optimal solution for the original problem (P_{int}) as shown in the next proposition. Note that for the implicit constraint on computation bits, $0 \leq s_i \leq u_i\ \forall i$, the upper bound automatically holds since we start the latency-aware Newton's method with a feasible point, with an initial s_i that is smaller than u_i, and then keep decreasing until zero or until the latency condition is met. The lower bound condition is checked in the algorithm such that in the update step, the value for s_i is positive, if not then it is set to zero.

2) Inner Algorithms for Other Variables: For each iteration of the outer algorithm, we solve for the inner optimization problems (P2) and (P3) in sequence. For a given value of $\boldsymbol{s}$, we solve (P2) to obtain time allocation, and calculate the charging time $T_c = T_d - T_1 - T_3$ which is used by problem (P3) to find the optimal energy beamforming matrix, as discussed in detail in the next subsection. These steps for the nested optimization are repeated until a minimum point for the weighted total energy consumption is reached where all the constraints in the original problem (P_{int}) are satisfied. At each iteration of (P3), we solve the LP problem (P4) to find the optimal beam power allocation using a standard convex solver. The algorithm flow is depicted in Figure 3 and the steps for solving (P_{int}) are given in Alg. 1.

The nested algorithm proposed here is different from the algorithm in [31]. The formulation in (P_{int}) is a joint optimization of all variables, where the solution for energy beamforming is a part of the inner algorithm solving subproblem (P3). On the other hand, the formulation in [31] constructs two sequential problems to solve for $\boldsymbol{W_q}$ and the optimal $\boldsymbol{s}$ and $\boldsymbol{t}$ separately, where the transmit covariance matrix $\boldsymbol{W_q}$ is solved as an independent problem after obtaining the optimal $\boldsymbol{s}$ and $\boldsymbol{t}$. As a result, here the energy beamforming matrix $\boldsymbol{W_q}$ is updated at every iteration along with the current values of variables s and t, instead of being updated separately only after establishing the optimal values of $\boldsymbol{s}$ and $\boldsymbol{t}$ as in [31]. The resulting optimal $\boldsymbol{W_q}^\star$ is also very different from [31] in both beam directions and beam power allocation, as will be illustrated in our numerical results section.

B. Inner Primal-Dual Algorithms

For the inner algorithm to solve for $\boldsymbol{W_q}$ and $\boldsymbol{t}$, we design two primal-dual algorithms where the primal variable values are obtained as closed form functions of the dual variables, and the dual variables are found by solving the dual problem using a sub-gradient methods. The Lagrangian of problem (P2) is given in (25). The dual-function for the convex optimization problem (P2) can be defined as

$$g_{P2}(\lambda_1, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \inf_{\boldsymbol{t}} \mathcal{L}_{P2}(\boldsymbol{t}, \lambda_1, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}) \tag{26}$$

and the dual-problem is defined as

$$\begin{aligned} \text{P2-dual:}\ \max\ & g_{P2}(\lambda_1, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}) \\ \text{s.t. } & \lambda_1 \geq 0, \beta_i, \theta_i, \phi_i \geq 0\ \forall i = 1...K \end{aligned} \tag{27}$$

where λ_1, $\boldsymbol{\beta}, \boldsymbol{\theta}$, and $\boldsymbol{\phi}$ are the dual variables associated with constraints (c-f) in (16), respectively.

Similarly, the dual-function for the convex optimization problem (P3) is obtained as

$$g_{P3}(\boldsymbol{\psi}, \lambda_5) = \min_{\boldsymbol{W_q}} \mathcal{L}_{P3}(\boldsymbol{W_q}, \boldsymbol{\psi}, \lambda_5) \tag{28}$$

$$\begin{aligned} \mathcal{L}_{P2} = \lambda_1 (\sum_{j=1}^{3} T_j - T_d) + \sum_{i=1}^{K_u} \frac{t_{u,i} (2^{\frac{s_i}{\nu t_{u,i} B}} - 1) \Gamma_1 \sigma_{1,i}^2}{N \gamma_i} + \sum_{i=1}^{K_u} \kappa_i c_i (u_i - s_i) f_{u,i}^2 + \sum_{i=1}^{K_u} \frac{t_{d,i} (2^{\frac{\mu s_i}{t_{d,i} B}} - 1) \Gamma_2 \sigma_{2,i}^2}{N \gamma_i} \\ + \sum_{i=1}^{K_u} \kappa_m d_m f_{mi}^2 s_i + \sum_{i=1}^{K} \beta_i (t_{u,i} - T_1) + \sum_{i=1}^{K} \theta_i \left(\frac{c_i q_i}{f_{u,i}} + t_{u,i} - T_d \right) + \sum_{i=1}^{K} \phi_i (t_{d,i} - T_3) \end{aligned} \tag{25}$$

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

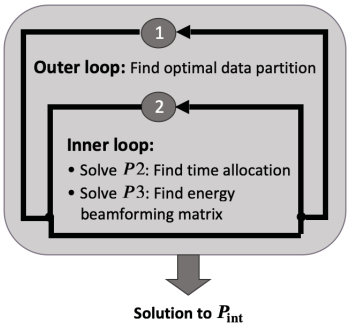

Figure 3: Nested algorithm architecture for P_{int}: Each outer loop advances the data partitioning using a latency-aware descent algorithm, whereas each inner loop jointly solves for the optimal time allocation and the optimal energy beamforming matrix using a primal-dual algorithm. All iterations satisfy the latency and power constraints.

where the Lagrangian is given in (33). The dual-problem for (P3) is then given as

$$\text{P3-dual: } \max \quad g_{P3}(\boldsymbol{\psi}, \lambda_5) \\ \text{s.t. } \lambda_5 \geq 0, \psi_i \geq 0 \text{ for } i = 1...K \tag{29}$$

Using closed form expressions for primal variables in terms of dual-variables as in Theorems 1-2, the dual functions above are functions of only the dual-variables. These dual problems can then be solved using the subgradient method [41].

The subgradient terms with respect to all dual variables of the original problem (P2) and (P3) are as given in (30a-d-f).

We use the subgradient method, for the inner algorithms, to solve the constrained convex optimization problems (P2) and (P3) [41]. Since the dual problems in (29, 27) are maximization problems for the respective dual functions, the inner algorithms find the subgradient for the negative dual functions ($-g_{P2}, -g_{P3}$). The primal variables are updated, at each iteration, based on Theorems 1-3.

The dual variables vector x is updated as

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \beta_k \boldsymbol{g}^{(k)} \tag{31}$$

where β_k is the k^{th} step-size, and $\boldsymbol{g}^{(k)}$ is the subgradient vector at the k^{th} iteration evaluated using the sub-gradient expressions in (30a-d-f). In our proposed algorithm, we use $\beta_k = 1/\sqrt{k}$. For this non-summable diminishing step size, the algorithm is guaranteed to converge to the optimal value as $k \to \infty$ with a theoretical iteration complexity of $\mathcal{O}(1/\epsilon^2)$ [41] [42]. At each iteration of the inner algorithm, the best point for the dual functions is retained since the subgradient method is not a descent method. These primal-dual update steps are repeated until the desired level of precision, ϵ_2, is reached for the stopping criterion.

Algorithm 1 Solution for (P_{int})

Given: Distances d_i $\forall i$. Channel $\boldsymbol{H} = \boldsymbol{G^T}$. Precision, ϵ_1, ϵ_2, Data u_i, Latency T_d. Initialize: s_i

Begin Outer Algorithm for P_{int}

Given a starting point $\boldsymbol{s}$, **Repeat**

1) Initialize dual variables, $\lambda_1, \lambda_5, \beta_i, \theta_i, \phi_i, \psi_i \forall i$ and compute $\Delta \boldsymbol{s}$ using Newton's method, where
$$\Delta \boldsymbol{s} := -\nabla^2 f_0(\boldsymbol{s})^{-1} \nabla f_0(\boldsymbol{s})$$
and $f_0(.)$ is the objective function in (18) .
2) **Offloading Sub-Algorithm for (P2)**
   - Calculate $t_{u,i}$ and $t_{d,i}$, using (21). Then $T_1^\star = \max t_{u,i}^\star$ and $T_3^\star = \max t_{d,i}^\star$.
   - Update p_i and η_i using (15) and calculate $\sigma_{1,i}^2$ and $\sigma_{2,i}^2$.

   Repeat
   - Find dual function in (26) using Theorem 1
   - Find subgradients in (30a-d)
   - Update dual variables using the subgradient method

   Until subgradients converge with ϵ_2 as in (32)
3) **Charging Sub-Algorithm for (P3)**
   - Calculate time for wireless charging
   $$T_c^\star = T_d - T_1^\star - T_3^\star$$
   - Find $\boldsymbol{\lambda_q}^\star$ by solving the LP in (P4).
   - Set $\boldsymbol{W_q}^\star = \boldsymbol{U_B} \boldsymbol{\Lambda_q^\star} \boldsymbol{U_B^*}$, where $\boldsymbol{\Lambda_q^\star} = \textbf{diag}(\boldsymbol{\lambda_q}^\star)$

   Repeat
   - Find dual function in (28) using Theorem 2
   - Find subgradients in (30e-f)
   - Update dual variables using the subgradient method

   Until subgradients converge with ϵ_2 as in (32)
   - Obtain $\boldsymbol{\alpha}$ as in (23)
4) *Line search and Update*. $s_i := s_i + t_i \Delta s_i$.
   - If any $s_i < 0$, set $s_i = 0$

Until stopping criterion for Newton's method is satisfied: $\lambda^2/2 < \epsilon_1$, where $\lambda := -\nabla f_0(\boldsymbol{s})^T \Delta \boldsymbol{s}$, or latency constraint in (24) is met.

End Outer Algorithm for P_{int}

In the subgradient method, the key quantity is not the function value but rather the Euclidean distance to the optimal set [41]. Therefore, for our implementation we employed a stopping criterion as

$$\|\boldsymbol{g}^{(k+1)} - \boldsymbol{g}^{(k)}\|_2 \leq \epsilon_2 \tag{32}$$

such that the iterations stop when the relative change is less than ϵ_2. This is a classical stopping criterion similar to the one

$$\nabla_{\lambda_1} \mathcal{L} = \sum_{j=1}^{3} T_j - T_{\text{delay}}, \quad \nabla_{\beta_i} \mathcal{L} = t_{u,i} - T_1, \quad \nabla_{\phi_i} \mathcal{L} = t_{d,i} - T_3 \quad \nabla_{\theta_i} \mathcal{L} = \frac{c_i q_i}{f_{u,i}} + t_{u,i} - T_d, \tag{30a-d}$$

$$\nabla_{\psi_i} \mathcal{L} = \alpha_i e_i - \xi_i \text{tr}\left(\boldsymbol{h_i^* W_q h_i}\right) T_c \quad \nabla_{\lambda_5} \mathcal{L} = \text{tr}(\boldsymbol{W_q}) - P \tag{30e-f}$$

proposed for the rapidly convergent iterative method in [43]. The steps for the integrated algorithm are shown in Alg. 1.

C. Complexity Analysis

We proceed to analyze the complexity of the proposed nested Algorithm 1. This algorithm consists of four mains steps as numerically labeled in Algorithm 1. Except for Step 4 which is a simple line search and update, we will discuss the computational complexity per iteration in each other step, as well as the number of iterations, or iteration complexity, required in each step. These analyses will let us compute the overall order of complexity.

1) Computational Complexity: First, consider the outer algorithm based on Newton's method. Step 1 in the outer algorithm is a calculation of Newton's search direction. For K users and $\boldsymbol{s} \in \mathbb{R}^{K\times 1}$, the computation cost for each Newton search direction requires $\mathcal{O}(K^3)$ flops [44]. In Step 4, the backtracking line search requires $\mathcal{O}(K)$ flops per inner backtracking step. The novel latency-aware stopping criterion in (24) is a max operation over K users, with complexity $\mathcal{O}(K)$ [22]. Putting these together, the main computation cost in each outer algorithm's iteration, excluding the inner algorithm steps, is therefore $\mathcal{O}(K^3)$.

Next, consider Step 2 and Step 3 in Algorithm 1, which correspond to the inner primal-dual algorithms for (P2) and (P3) respectively. In Step 2, we use the subgradient method for computation offloading resource allocation in (P2), for which the Lambert function evaluation for the primal variables $\boldsymbol{t_u}, \boldsymbol{t_d}$ is more computationally dominant, since it requires Halley's iteration to invert $x\exp(x)$, using a first-order asymptotic approximation as the initial estimate. Halley's iteration method is a higher-order generalization of Newton's method which requires analytical and numerical computation of higher-order derivatives of the function, such that using FFT multiplication, it has a complexity of $\mathcal{O}(\log(K))$ [45]. For $2K$ primal variables, $(t_{u,i}, t_{d,i} \forall i \in [1, K])$, the complexity for this step is of the order $\mathcal{O}(2K\log(K))$. The chosen stopping criterion for both (P2) and (P3) is a norm calculation which requires requires $\mathcal{O}(2(3K+1))$ and $\mathcal{O}(2(K+1))$ flops based on the size of the subgradient vector $\boldsymbol{g}^{(k)}$ for (P2) and (P3), respectively. The total computation complexity for each iteration in the primal-dual algorithm for (P2) in Step 2 is then $\mathcal{O}(K\log(K))$.

In Step 3, the charging sub-algorithm for $(P3)$, finding the optimal beam directions for an N-antenna massive MIMO array requires performing the SVD of $\boldsymbol{C} \in \mathbb{C}^{N\times N}$ with a computation cost of $\mathcal{O}(N^2)$. At each iteration of this inner algorithm for (P3), we also solve a linear programming problem $(P4)$ with computational complexity $\mathcal{O}(3K+1)$ [46]. Finally, $\boldsymbol{W_q} = \boldsymbol{U_B}\boldsymbol{\Lambda_q^\star}\boldsymbol{U_B^*}$ is obtained through matrix multiplication. Taking into account the K non-zero-elements of $\boldsymbol{\Lambda_q}$ by writing $\boldsymbol{W_q} = \sum_{i=1}^{K} \lambda_{q,i}\boldsymbol{u_i}\boldsymbol{u_i}^*$, the complexity for this multiplication operation is of the order $\mathcal{O}(NK)$. Thus the total computation complexity for each iteration of the primal-dual algorithm for (P3) in Step 3 is $\mathcal{O}(N^2 + NK)$.

2) Iteration Complexity: The convergence of the outer algorithm depends on Newton's method which has a linear start and then hits the quadratic convergence after a certain number of iterations which depends on the starting point [40]. In our latency-aware descent outer algorithm, since we add an additional stopping criterion based on the latency, the algorithm may stop earlier than the standard implementation. The latency-aware Newton's method may not hit the quadratic convergence if the latency constraint is met before that.

The sub-algorithm for (P2) in Step2 and sub-algorithm for (P3) in Step3 are both based on the subgradient method where we use the non-summable diminishing step size for which the number of iterations required to reach convergence is of the order $\mathcal{O}(1/\epsilon^2)$ [41] [42]. The sub-algorithm for (P2) in Step2 also includes the Halley's iterative method for computing $\boldsymbol{t_u}, \boldsymbol{t_d}$ which uses linear-over-linear approximation and has a cubic rate of convergence, $\mathcal{O}(\log_3(n))$, for n-bit accuracy [45]. The sub-algorithm for (P3) in Step3 includes a linear-programming step which is solvable in polynomial time [46].

3) Total Complexity: Putting together the analyses above, we can compute the complexity of each sub-algorithm alone, and then put them together to compute the complexity of the overall nested algorithm. The complexity for the sub-algorithm for (P2) in Step 2 is equal to the product between its iteration complexity and computation complexity, which gives $\mathcal{O}(c_1 K\log(K)/\epsilon^2)$, where $c_1 = \mathcal{O}(\log_3(n))$ is the number of iterations for Halley's method which does not grow with K or N. Similarly, the complexity for the sub-algorithm for (P3) in Step 3 is $\mathcal{O}((N^2 + NK)/\epsilon^2)$.

The total complexity of the nested algorithm can then be computed as

$$
\begin{aligned}
&\text{Total complexity of nested Algorithm 1}\\
&= \text{iteration complexity of outer algorithm}\\
&\times \big(\text{computation complexity for Step 1}\\
&+ \text{iteration complexity} \times \text{computation complexity for Step 2}\\
&+ \text{iteration complexity} \times \text{computation complexity for Step 3}\big)\\
&= c\left(\mathcal{O}(K^3) + \mathcal{O}(c_1 K\log(K)/\epsilon^2) + \mathcal{O}((N^2 + NK)/\epsilon^2)\right)\\
&= \left(\mathcal{O}(K^3) + \mathcal{O}(K\log(K)/\epsilon^2) + \mathcal{O}((N^2 + NK)/\epsilon^2)\right)
\end{aligned}
$$

where c is the number of latency-aware Newton iterations for the outer algorithm which does not grow with problem size. The final complexity expression is a function of both the number of antennas N and number of users per cell K. In a typical network scenario, we often have $N \gg K$, which leads to the complexity dominated by and growing quadratically with N. The number of users K, however, also plays an important role since often N is fixed in a given network but K can change. In the numerical results section, we analyze the complexity of each sub-algorithm in terms of K.

VI. Numerical Results

In this section, we evaluate the solution of energy minimization problem (P_{int}) with respect to energy and time consumption, the partition of bits offloaded to the MEC for computation and the received energy via wireless charging. For simulations, we consider an exhibition room setting within

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

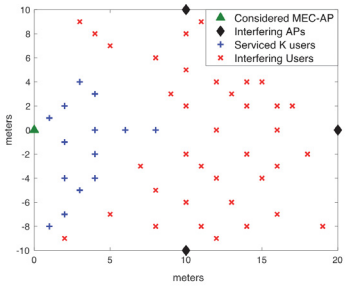

Figure 4: Simulation system layout: A 4-cell network with 4 MEC-AP, each serving 10 randomly located users within an area of 20m $\times$ 20m.

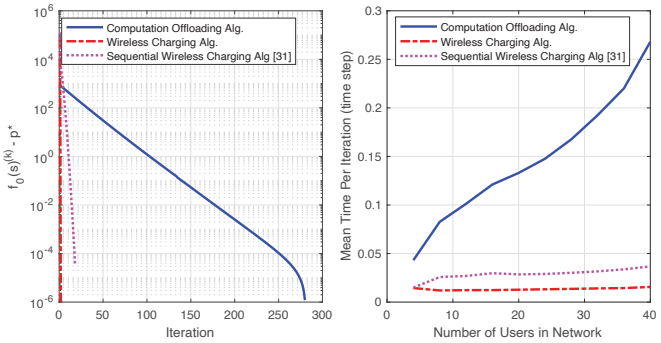

Figure 5: Alg. 1 convergence vs. (a) number of iterations, and (b) number of users in the network, for both computation offloading (sub-problem P2, $\epsilon_1 = 10^{-5}, \epsilon_2 = 10^{-3}$) and wireless charging (sub-problem P3, $\epsilon_2 = 10^{-6}$), in comparison with sequential wireless charging in [31].

an area of 20m $\times$ 20m with 4 APs each with $N = 100$ antennas and 40 active users randomly located with $K = 10$ users per AP's coverage area as shown in Figure 4. Note that the total number of users on the ground can be much larger, but these are the number of active UEs requesting offloading and wireless charging services at each time. For simulations, $w = 10^{-3}$, $T_d = 20$ms (for AR/VR applications [47]), $B = 5$MHz, $\tau_c = BT_d$, $\Gamma_1 = \Gamma_2 = 1.25$, $\mu = 2$, $\kappa_i = 0.5$pF, $\kappa_m = 5$pF, $c_i = 1000$, $d_m = 500$, $\gamma = 2.2$, $\sigma = 2.7$dB, $\sigma_r^2 = -127$dBm, $\sigma_k^2 = -122$dBm, $f_{u,i} = f_u = 1800$ MHz $\forall i$. Each MEC processor has 24 cores with maximum frequency of 3.4GHz, and we use $f_{m,i} = f_m = (24 \times 3400)/K$ MHz $\forall i$. Transmit power available at user and AP is 23 dBm and 46 dBm respectively. To calculate the interference and noise power ($\sigma_{1,i}^2, \sigma_{2,i}^2$) which include massive MIMO pilot contamination and intercell interference, we assume that user terminals transmit at their maximum power, that is $p_{qi} = 23$dBm, and the interfering APs use equal power allocation in the downlink, that is $\eta_{qi} = \frac{1}{K}$ $\forall i$. Numerical results are averaged independent channel realizations of $\mathbf{H}$ and $\mathbf{G}$ for 1000 spatial realizations (randomly generated user locations).

A. Algorithm Convergence

Figure 5 shows, on the left, the convergence of the two algorithms solving optimization sub-problems $(P2)$ and $(P3)$ with $u_i = u = 10$kbits, $e_i = e = 1$J $\forall i$. On the right, the mean time per iteration is plotted against the number of users K. Note that both $(P2)$ and $(P3)$ use the subgradient method with an iteration complexity of $\mathcal{O}(1/\epsilon^2)$, however $(P3)$ converges in fewer iterations compared to $(P2)$ since $(P3)$ uses a linear-programming step which is solvable in polynomial time, and the primal-dual steps in the sub-algorithm for (P3) converge significantly faster compared to those in (P2) which use the iterative Halley's method for each time allocation variable $(t_{u,i}, t_{d,i} \forall i)$. We observe, however, that even though the sub-algorithm for (P2) takes more iterations to converge, it eventually hits the quadratic convergence region where the difference between the current objective function value and the optimal value drops off dramatically with each additional iteration. This observation agrees with our convergence analysis of the outer algorithm based on Newton's method.

Figure 5 (right) shows that the run time per iteration for the sub-algorithm for (P3), or *charging sub-algorithm*, increases linearly with the number of users K, in agreement with the computation complexity analysis result of $\mathcal{O}(NK + N^2)$. On the other hand, the mean time per iteration for subalgorithm (P2) increases super-linearly with the number of users in the network. This result also agrees with the earlier complexity analysis which shows the computation complexity for (P2) sub-algorithm as $\mathcal{O}(K \log(K))$.

We also compare the proposed *charging sub-algorithm* for the joint energy minimization to the sequential energy maximizing opportunistic charging scheme in [31], showing significantly faster convergence for the proposed joint charging scheme as seen in Figure 5. For our implementation on a personal computer, the time unit in Figure 5 (right) is a second, however for faster machines, such as MEC servers, with the high-performance CPUs and parallel processing, this time-step may be significantly smaller. Notwithstanding the complexity, the performance of these algorithms can also be used as a benchmark for joint computation offloading and wireless charging on MEC systems.

Figure 6 (left) shows the total runtime for the sub-algorithm for (P2) and the sub-algorithm for (P3). These results are also in agreement with those in Figure 5 in that the total runtime for computation offloading is significantly higher than for wireless charging. For wireless charging (P3), we see that the total run time increases linearly with the increase in number of users. For computation offloading (P2) however, we observe an interesting result that the total runtime actually decreases as the number of users increases beyond a certain threshold. This phenomenon can be explained by a decrease in the number of iterations as the number of user increases. According to Figure 6 (right), as the number of users increases, the average amount of data offloaded to the MEC also increases, which leads to a longer mean time per iteration. However, the number of outer iterations in the latency-aware algorithm actually decreases, leading to an overall faster convergence time.

B. Comparison of Wireless Charging Schemes

Figure 7 shows a comparison of the proposed integrated and sequential wireless charging schemes with two other schemes: (i) isotropic scheme where $\boldsymbol{W_q} = \frac{P}{N}\boldsymbol{I}$ and equal charging power P/N is allocated across all N antennas of

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

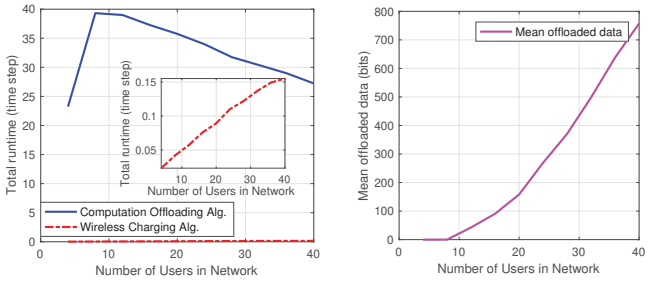

Figure 6: Total runtime versus the number of users for each algorithm on computation offloading and wireless charging (left) and the mean amount of offloaded data to the MEC-AP with increasing network size (right).

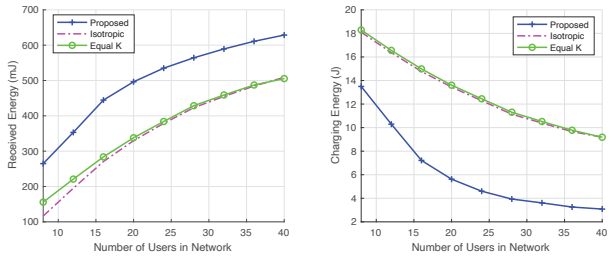

Figure 7: Comparison of the proposed energy beamforming scheme for wireless charging with isotropic wireless charging, and directed K-beam charging with equal power allocation (equal K): (left) Total received wirelessly charged energy; (right) Total transmitted energy consumption for wireless charging.

the AP, and (ii) equal K with directional charging using the beamforming directions proposed in Theorem 2, but with equal power allocation P/K across K energy beams. For fairness of comparison, we use power scaling for the other two schemes such that the users receive energy at most equal to the requested amounts similar to the proposed scheme. Since wireless charging is proposed as a billable service for future networks, this is also a necessary design consideration from the service providers' and consumers' perspective.

Figure 7 shows the received energy on the left and the transmitted energy on the right. As illustrated in this figure, the sum received energy for the proposed scheme is significantly larger than the other two schemes. The wireless charging performance for the isotropic and beamforming with equal power allocation scheme are similar. However, for smaller networks the equal power allocation scheme with directed power transfer does offer some improvement over the isotropic scheme in terms of the received energy. The proposed integrated charging energy minimization scheme consumes the lowest charging energy overall and offers significantly better received energy performance than both the isotropic and equal power schemes.

C. Energy Charging Beams

Figure 8 shows the beam radiation pattern for the proposed joint energy minimization scheme and the opportunistic max-received wireless charging scheme proposed in [31] respectively, under the same channel conditions. For the proposed energy minimization scheme, we see that the nulls are not as deep which allows for increased charging energy levels to users in low coverage areas. This "*null-fill*" property is a

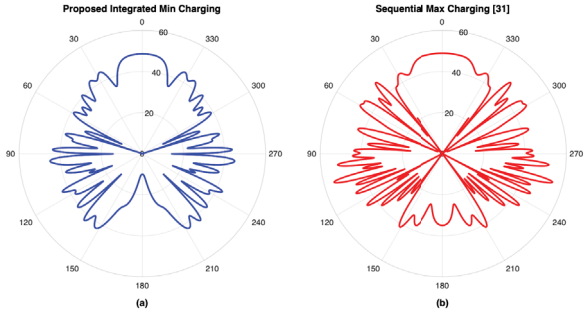

Figure 8: A typical wireless charging beam pattern for simultaneously charging multiple UEs from an MEC-AP, where shown is the strongest beam out of 10 beams for this channel realization: (a) The proposed joint energy minimization scheme, (b) Sequential energy maximization opportunistic charging scheme in [31]. System setting: 10 UEs simultaneously receiving wireless charging from this MEC-AP, with 30 interfering UEs in other cells, locations of UEs and MEC-APs are as given in Figure 4. The beam patterns show that the proposed joint wireless charging scheme consumes a much smaller amount of energy by having lower intensity beams and no backscatter beams.

common design feature to alter the energy distribution for the various antenna elements in the array [48]. In both schemes, however, users may receive wireless charging not only from the main beam but also from the side lobes which can be an important consideration for wireless charging. One significant difference among the two schemes is the reduction in number of side lobes and elimination of back lobes for the proposed scheme, which curbs energy losses and enhances the objective of energy minimization.

Another interesting finding presented in the plot (bottom) in Figure 9 is the optimal number of energy beams for $K = 10$ users per cell. For the isotropic wireless charging, there are always $N > K$ energy beams. For the case of K beams with equal power allocation, the number of beams is equal to the number of users in the cell. While multiple energy beams may be necessary for a multi-user system as also previously discussed in [49], the optimal number of energy beams for the integrated charging energy minimization scheme is usually less than the number of users as seen in Figure 9. Since each energy beam can contribute as additional RF charging sources for neighboring users, the transmit beamforming can be intelligently designed as proposed to limit the number of energy beams which can prevent energy losses caused by transmitting energy in numerous directions and hence also contribute to energy minimization.

D. Charging Efficiency

Figure 10 shows the average percentage of received energy at the users end compared to the requested energy. Here the requested energy is assumed to be 66mJ/ms for each user, however, for a lower value of the requested energy, the percentage received energy would be higher. The figure shows a comparison for the directed equal power and isotropic schemes to the proposed scheme under two operating modes, the charging only mode where connected users request wireless charging but do not require computation at the edge, and the *data and charging* mode where connected users request both wireless charging as well as computation offloading. The average charged percentage is seen to decrease with an

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/JSTSP.2021.3098963, IEEE Journal of Selected Topics in Signal Processing

$$\mathcal{L}_{P3} = T_c \text{tr}(\boldsymbol{W_q}) + \lambda_5 \left(\text{tr}(\boldsymbol{W_q}) - P\right) - \xi_i T_c \text{tr}\left(\left(\sum_{i=1}^{K} \psi_i h_i h_i^*\right) \boldsymbol{W_q}\right) + \sum_{i=1}^{K} \psi_i \alpha_i e_i$$

$$= \text{tr}\left(\left[(T_c + \lambda_5)\boldsymbol{I} - \xi_i T_c \sum_{i=1}^{K} \psi_i \boldsymbol{h_i h_i^*}\right] \boldsymbol{W_q}\right) + \sum_{i=1}^{K} \psi_i \alpha_i e_i - \lambda_5 P = \text{tr}\left(\boldsymbol{B W_q}\right) + \sum_{i=1}^{K} \psi_i \alpha_i e_i - \lambda_5 P \tag{33}$$

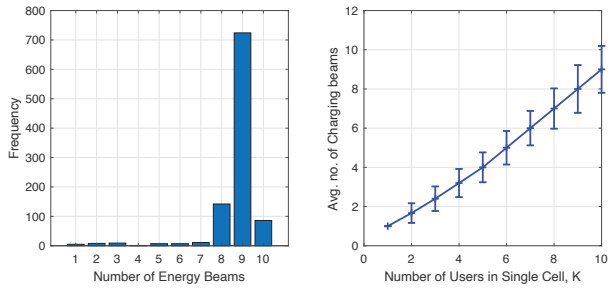

Figure 9: Distribution of the number of charging energy beams for the proposed joint scheme for K = 10 users over 1000 spatial realizations. System setting: 10 UEs receiving wireless charging from this MEC-AP, with 30 interfering UEs in other cells, locations of UEs and MEC-APs are as given in Figure 4.

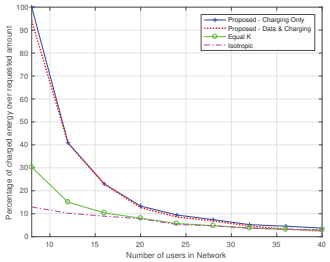

Figure 10: Per-user charged energy as percentage of requested energy with each user requesting 66mJ, 4 MEC-APs and system layout as shown in Figure 4 averaged over 1000 spatial realizations. While the amount of charged energy per-user reduces as the number of users in the network increases, the total amount of energy received via wireless charging increases with more users as shown in Figure 7.

increase in the network size as expected. While the directed beams equal charging scheme shows some improvement over the isotropic scheme for small network sizes, both schemes generally have less than 50% efficiency compared to the proposed scheme. Further, as illustrated, even under the joint data and charging mode, where the MEC-AP simultaneously optimizes the resources required for both computation offloading as well as wireless charging, the decrease in charging efficiency is negligible compared to the charging only mode.

VII. Conclusion

We considered a latency constrained multi-cell multi-user wireless system with collocated MEC-AP servers providing computation offloading and wireless charging services to its connected users. We formulated a novel system-level problem to minimize the total transmit energy consumption while ensuring the largest amount of received energy as feasible. We design an efficient nested algorithm by an optimal division into convex subproblems to solve for data partitioning, time allocation and transmit energy beamforming matrices. Our algorithm demonstrates that even with significant amounts of data to be computed, the network can deliver decent amounts of charged energy to the users. The MEC-AP wireless charging beams for minimizing the overall energy consumption also have no back lobe and have significantly more power concentrated in the main lobe, hence delivering a more efficient and effective energy transfer. These results validate the potential of wireless charging in concurrent with computation offloading from edge networks and can provide a performance benchmark for practical implementations.

VIII. Appendix

A. Appendix A - Proof for Lemma 1

Consider problem (P_{int}) in (16) at fixed values of s_i. The objective function is affine and convex. Convexity in $\boldsymbol{t}$ can be established similar to [22, Lemma 1]. Constraint (b) contains a function of the form $f_1 = (T_d - T_1 - T_3)\text{tr}(\boldsymbol{W_q})$, with affine term $T_d\text{tr}(\boldsymbol{W_q})$. Considering $\tilde{f}_1 = -T_1\text{tr}(\boldsymbol{W_q})$, to check for joint convexity in T_1 and $\boldsymbol{W_q}$, the Hessian of $\tilde{f}_1$ is the block matrix, $\nabla^2 \tilde{f}_1 = [\boldsymbol{0}_{N\times N} \quad -\boldsymbol{I}_{N\times N}; -\boldsymbol{I}_{N\times N} \quad \boldsymbol{0}_{N\times N}]$, with repeated eigenvalues ± 1 and therefore doesn't show convexity. However, the sublevel sets $\{(T_1 \in \mathbb{R}^+, \boldsymbol{W_q} \in \mathbb{R}^{N\times N}), -T_1\text{tr}(\boldsymbol{W_q}) \leq \alpha\}$ are jointly convex in T_1 and $\boldsymbol{W_q}$ in the domain of the function, $(T_1 \geq 0, \boldsymbol{W_q} \succcurlyeq \boldsymbol{0})$, therefore the function $\tilde{f}_1$ is quasiconvex [40, Example 3.31]. Therefore constraint (16b) is a sum of convex and quasiconvex functions with convex sets and sublevel sets respectively. Similarly, constraint (i) also has convex sublevel sets with a quasiconvex function of the form $-T_c\text{tr}(\boldsymbol{h_i^* W_q h_i})$. Constraints (j) is the linear trace of $\boldsymbol{W_q}$ and hence is convex.

Based on the above, the objective is convex and all constraints are either convex or have convex sub-level sets in the remaining variables. Thus the problem is convex at a given s_i.

B. Appendix B - Proof for Theorem 2 and 3

1) Proof for Theorem 2: To minimize the Lagrangian for problem (P3) as given in (33) to obtain the dual function in (28), we only need to consider the term involving $\boldsymbol{W_q}$

$$\min_{\boldsymbol{W_q}} \quad \text{tr}\left(\boldsymbol{B W_q}\right) \tag{34}$$

By applying an inequality relating the trace of a matrix product to the sum of eigenvalue products [50, Ch. 9, H.1.h.], $\text{tr}(\boldsymbol{B W_q})$ is minimized by choosing $\boldsymbol{U_q} = \boldsymbol{U_B}$ such that

$$\text{tr}(\boldsymbol{B W_q}) = \sum_{i=1}^{N} \lambda_{B,i} \cdot \lambda_{q,i} \tag{35}$$

where the eigenvalues of $\boldsymbol{W_q}$ are in descending order $\lambda_{q,1} \geq \lambda_{q,2} \geq \ldots \geq \lambda_{q,N}$ and those of matrix $\boldsymbol{B}$ are in ascending

order such that $\lambda_{B,1} \leq \lambda_{B,2} \leq \ldots \leq \lambda_{B,N}$ and the eigenvectors $\boldsymbol{U_B}$ are obtained based on this order of the corresponding eigenvalues in $\boldsymbol{\Lambda_B} = \text{diag}(\boldsymbol{\lambda_B})$. Since the eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{W_q}$ are in reverse order to each other, the sum of their eigenvalue products yields the minimum value for $\text{tr}(\boldsymbol{BW_q})$ in (35).

2) Proof for Theorem 3: In the eigenvalue decomposition of $\boldsymbol{W_q^\star}$ as $\boldsymbol{W_q} = \boldsymbol{U_q \Lambda_q U_q^*}$, the diagonal matrix $\boldsymbol{\Lambda_q} \in \mathbb{R}^{N \times N}$ has power allocated across K diagonal elements and the remaining eigenvalues for the $N-K$ beams is set to zero. Based on Theorem (2), equation (20b) can be rewritten as

$$\text{tr}(\boldsymbol{h}_i^* \boldsymbol{U_q \Lambda_q U_q^* h}_i) = \pi_i \tag{36}$$

where $\pi_i = \frac{e_i}{\xi_i T_c}$ $\forall i = 1...K$. We define the row vector, $\boldsymbol{q}_i^* = \boldsymbol{h}_i^* \boldsymbol{U_q} = \boldsymbol{h}_i^* \boldsymbol{U_B}$. Then

$$\text{tr}(\boldsymbol{q_i^* \Lambda_q q_i}) = \pi_i \tag{37}$$

Define row vector $\boldsymbol{a_i}^* = \text{diag}(\boldsymbol{q_i q_i^*})$ for $i \in [1, K]$, $\boldsymbol{A} \in \mathbb{R}^{K \times K} = [\boldsymbol{a_1^*}...\boldsymbol{a_K^*}]$, and vector $\boldsymbol{b} \in \mathbb{R}^{K \times 1} = [\pi_1 ... \pi_K]$. This results in constraint (22c) in (P4). The ordering of $\boldsymbol{\lambda_q}$ needs to be in reverse from $\boldsymbol{\lambda_B}$, that is, in descending order, so as to minimize (33) as in (35), which gives us (22b) in (P4).

REFERENCES

[1] ETSI, "Mobile-Edge Computing - Introductory Technical White Paper," Huawei, IBM, Intel, Nokia, DOCOMO, Vodafone, Tech. Rep., 2014.

[2] ——, "ETSI White Paper No. 28: MEC in 5G networks," ETSI, Tech. Rep., 2018.

[3] M. S. Elbamby, C. Perfecto, M. Bennis, and K. Doppler, "Toward Low-Latency and Ultra-Reliable VR," *IEEE Network*, March 2018.

[4] E. Cau, M. Corici, P. Bellavista, L. Foschini, G. Carella, A. Edmonds, and T. M. Bohnert, "Efficient Exploitation of Mobile Edge Computing for Virtualized 5G in EPC Architectures," in *2016 4th IEEE International Conference on MobileCloud*, March 2016.

[5] D. Sabella, et al., "Edge Computing: from standard to actual infrastructure deployment and software development," White Paper, Intel Corporation, Tech. Rep., 2019, revision 1.0.

[6] Ossia, "Cota: Real Wireless Power," http://www.ossia.com/cota/.

[7] Energous, "Far Field Wattup Transmitter," "http://energous.com".

[8] J. Kim, B. Clerckx, and P. D. Mitcheson, "Signal and System Design for WPT : Prototype, Experiment and Validation," 2019.

[9] L. Xie, J. Xu, and R. Zhang, "Throughput Maximization for UAV-Enabled Wireless Powered Communication Networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1690–1703, 2019.

[10] I. F. Akyildiz, A. Kak, and S. Nie, "6g and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, 2020.

[11] S. Sardellitti, G. Scutari, and S. Barbarossa, "Joint Optimization of Radio and Computational Resources for Multicell Mobile-Edge Computing," *IEEE Trans. on Sig. and Info. Processing over Networks*, June 2015.

[12] K. B. Letaief, W. Chen, Y. Shi, J. Zhang, and Y. A. Zhang, "The roadmap to 6g: AI empowered wireless networks," *IEEE Comms Mag*, 2019.

[13] X. Hu, K. Wong, and K. Yang, "Wireless Powered Cooperation-Assisted Mobile Edge Computing," *IEEE Trans on Wir. Comms*, April 2018.

[14] F. Zhou, Y. Wu, H. Sun, and Z. Chu, "UAV-Enabled Mobile Edge Computing: Offloading Optimization and Trajectory Design," in *2018 IEEE ICC*, May 2018.

[15] Y. Zhao, V. C. M. Leung, H. Gao, Z. Chen, and H. Ji, "Uplink Resource Allocation in Mobile Edge Computing-Based Heterogeneous Networks with Multi-Band RF Energy Harvesting," in *2018 IEEE ICC*, May 2018.

[16] C. You, K. Huang, and H. Chae, "Energy Efficient Mobile Cloud Computing Powered by WET," *IEEE JSAC*, May 2016.

[17] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint Offloading and Computing Optimization in Wireless Powered Mobile-Edge Computing Systems," *IEEE Trans on Wir Comms*, March 2018.

[18] H. Guo, J. Liu, and J. Zhang, "Efficient Computation Offloading for Multi-Access Edge Computing in 5G HetNets," in *IEEE ICC*, May 2018.

[19] K. Zhang, Y. Mao, S. Leng, Q. Zhao, L. Li, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Energy-Efficient Offloading for MEC in 5G Het Nets," *IEEE Access*, 2016.

[20] X. Zhang, Y. Mao, J. Zhang, and K. B. Letaief, "Multi-objective resource allocation for mobile edge computing systems," in *IEEE PIMRC*, 2017.

[21] X. Lyu, H. Tian, C. Sengul, and P. Zhang, "Multiuser Joint Task Offloading and Resource Optimization in Proximate Clouds," *IEEE Trans on Veh. Tech*, April 2017.

[22] R. Malik and M. Vu, "Energy-efficient computation offloading in delay-constrained massive MIMO enabled edge network using data partitioning," *IEEE Trans on Wir. Comms*, vol. 19, 2020.

[23] C. You, K. Huang, H. Chae, and B. Kim, "Energy-Efficient Resource Allocation for Mobile-Edge Computation Offloading," *IEEE Trans on Wir Comms*, March 2017.

[24] S. Bi and Y. J. Zhang, "Computation Rate Maximization for Wireless Powered Mobile-Edge Computing With Binary Computation Offloading," *IEEE Trans on Wir Comms*, June 2018.

[25] 3GPP, "Service requirements for cyber-physical control applications in vertical domains; Stage 1," 3GPP, Technical Specification (TS) 22.104.

[26] 3GPP, "Service requirements for the 5G system; Stage 1," TS 22.261.

[27] 3GPP, "Service requirements for video, imaging and audio for professional applications (VIAPA); Stage 1, (TS) 22.263, 2020, version 17.1.0.

[28] Humavox, "Wireless Charging in the Age of Connectivity," "http://www.humavox.com/".

[29] Guru, "Guru Wireless - Delivering an Ecosystem of Wireless Power"

[30] Energous, "Energous Corporation Announces Collaboration with vivo Global," "https://ir.energous.com/press-releases/detail/630/energous-corporation-announces-collaboration-with-vivo", accessed: 2021-04-11.

[31] R. Malik and M. Vu, "On-Request Wireless Charging and Partial Computation Offloading In Multi-Access Edge Computing Systems," *IEEE Trans on Wireless Comms*, 2021.

[32] ETSI, "Harmonizing standards for edge computing - A synergized architecture leveraging ETSI ISG MEC and 3GPP specifications," 2020.

[33] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of MEC in 5g and beyond: Fundamentals, tech. integration, and state-of-the-art," *IEEE Access*, 2020.

[34] F. Wang, J. Xu, and S. Cui, "Optimal energy allocation and task offloading policy for wireless powered MEC systems," *IEEE Trans on Wir Comms*, 2020.

[35] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A Survey on Mobile Edge Computing: The Communication Perspective," *IEEE Comms Svys Tuts*, Fourthquarter 2017.

[36] 3GPP, "Study on enhancement of Radio Access Network (RAN) slicing for NR,", Technical Report (TR) 38.832, 2020, version 1.0.0.

[37] H. Q. Ngo and E. G. Larsson, "No Downlink Pilots Are Needed in TDD Massive MIMO," *IEEE TWC*, May 2017.

[38] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[39] T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, and D. Sabella, "On MEC: A Survey of the Emerging 5G Network Edge Cloud Architecture and Orchestration," *IEEE Comms Svys Tuts*, thirdquarter 2017.

[40] S. Boyd and L. Vandenberghe, *Convex optimization*. CUP, 2004.

[41] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient Methods," *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2003.

[42] R. Tibshirani, "Subgradient Method," *lecture notes 10-725/36-725: Convex Optimization, Spring 2015*, 2015.

[43] D. Powell and J. Macdonald, "A rapidly convergent iterative method for the solution of the generalised nonlinear least squares problem," *The Computer Journal*, 1972.

[44] R. Tibshirani, "Newton method," *Notes for Convex Optimization: Machine Learning 10-725, UC Berkeley*, 2019.

[45] G. Dahlquist and Å. Björck, *Numerical methods in scientific computing, volume I*. SIAM, 2008.

[46] N. Megiddo, "On the complexity of linear programming," in *Advances in economic theory: Fifth world congress*. CUP.

[47] D. Robbins, C. Cholas, M. Brennan, and K. Critchley, "Augmented and Virtual Reality for Service Providers," Intel Corp, Tech. Rep., 2017.

[48] B. Lindmark, "Analysis of pattern null-fill in linear arrays,"*EuCAP*, 2013.

[49] Y. Zeng, B. Clerckx, and R. Zhang, "Communications and signals design for wireless power transmission," *IEEE Trans on Comms*, 2017.

[50] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and its Applications*. Springer, 1979, vol. 143.