applied sciences

MDPI

Article

Energy-Efficient Caching for Mobile Edge Computing in 5G Networks

Zhaohui Luo [1], Minghui LiWang [1], Zhijian Lin [1], Lianfen Huang [1,*], Xiaojiang Du [2] and Mohsen Guizani [3]

[1] Department of Communications Engineering, Xiamen University, Xiamen 361005, China; luozhaohui@stu.xmu.edu.cn (Z.L.); minghuilw@stu.xmu.edu.cn (M.L.W.); linzhijian1234@126.com (Z.L.)

[2] Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA; dux@temple.edu

[3] Department of Electrical and Computer Engineering, University of Idaho, Moscow, ID 83844, USA; mguizani@uidaho.edu

* Correspondence: lfhuang@xmu.edu.cn; Tel.: +86-592-258-0142

Academic Editor: Elli Kartsakli

Received: 31 March 2017; Accepted: 23 May 2017; Published: 27 May 2017

Abstract: Mobile Edge Computing (MEC), which is considered a promising and emerging paradigm to provide caching capabilities in proximity to mobile devices in 5G networks, enables fast, popular content delivery of delay-sensitive applications at the backhaul capacity of limited mobile networks. Most existing studies focus on cache allocation, mechanism design and coding design for caching. However, grid power supply with fixed power uninterruptedly in support of a MEC server (MECS) is costly and even infeasible, especially when the load changes dynamically over time. In this paper, we investigate the energy consumption of the MECS problem in cellular networks. Given the average download latency constraints, we take the MECS's energy consumption, backhaul capacities and content popularity distributions into account and formulate a joint optimization framework to minimize the energy consumption of the system. As a complicated joint optimization problem, we apply a genetic algorithm to solve it. Simulation results show that the proposed solution can effectively determine the near-optimal caching placement to obtain better performance in terms of energy efficiency gains compared with conventional caching placement strategies. In particular, it is shown that the proposed scheme can significantly reduce the joint cost when backhaul capacity is low.

Keywords: edge caching; energy-efficient; mobile edge computing; 5G cellular networks

1. Introduction

Nowadays, with the rapid development of mobile communication technologies and mobile devices, wireless data traffic is experiencing an explosive increase, especially in terms of mobile video streaming, high definition (HD) video and video webcasting [1]. A recent Cisco report estimates that the global mobile data volume will grow nearly ten times in the next five years, and the world's mobile data traffic will reach 30.6 monthly exabytes by 2020 [2]. It has led to significant increases in user latency and imposed a heavy burden on backhaul links connecting local base stations (BSs) to the core network (CN). In addition, the rapid growth of mobile data traffic has been compelling mobile network operators (MNOs) to provide more and more network capacities and meet these pressing traffic demands, which are achieved by extending their network infrastructure and enhancing spectral efficiency. For example, it is still far from enough to satisfy the mobile data traffic demands although the capacity of cellular network which can be immensely increased by deploying a large amount of BSs [3]. In fact, since plenty of the available backhaul networks are of low capacity and often cannot catch up with the rate requirements, backhaul capabilities have been regarded as a bottleneck for mobile cellular networks.

One promising solution to meet the demand is edge caching, which brings video contents closer to the users, reduces data traffic going through the backhaul links, the time required for content delivery, as well as help to smooth the traffic during peak hours. In wireless edge caching, highly sought-after videos are cached in the cellular BSs or wireless access points so that demands from users to the same content can be accommodated easily without duplicate transmissions from remote servers. Specifically, local caching can be more effective when a fraction of requested contents has high popularity.

Recently, Mobile Edge Computing (MEC) [4,5] has been introduced as an emerging paradigm enabling a capillary distribution of cloud storage capabilities to the edge of the cellular radio access network (RAN). In particular, the MECSs are implemented directly at the BSs using generic-computing platforms, which enable context-aware services and caching deployment in close-proximity to the mobile users. As a consequence of this, MECS presents a unique opportunity to not only implement edge caching but also perform caching placement strategy design. With the benefits of avoiding potential network congestion and alleviating the backhaul links burden, caching popular content at MECSs for backhaul capacity-limited mobile networks has emerged as a cost effective solution [6,7]. Recently, a good deal of works have been focused on big data analysis strategies for edge caching [8,9], context-aware caching deployment strategy design [10,11], and decentralized coded caching strategies [12,13]. Nevertheless, the cache allocation mechanism, more specifically, the energy efficiency (EE) cache deployment, has received less attention. When the actual budget is given, the cache size deployed at MECS will not be arbitrarily large. Caching more content requires activating more MECS, which results in more energy consumption. Moreover, providing grid power supply with fixed power uninterruptedly in support of MECS is costly and even infeasible, especially when the load changes dynamically over time. Hence, the cost energy of MECS should be carefully investigated, and the EE of MECS within the 5G cellular network should be optimized. As a result, the interplay between the EE and backhaul capacity is supposed to be intensively studied.

Recently, the issue of energy efficiency has received a lot of attention in the MEC system [14]. In [15], user association and power allocation in millimeter-wave-based ultra-dense networks is considered with attention to load balance constraints, energy harvesting by base stations, user quality of service requirements, energy efficiency, and cross-tier interference limits. Literature [16] investigates the power control and sensing time optimization problem in a cognitive small cell network, where the mitigation of cross-tier interference, imperfect hybrid spectrum sensing, and energy efficiency are considered. As one of the most popular and efficient energy saving schemes [17,18], BS sleeping has been proposed and widely studied to realize substantial energy saving in cellular networks [19–22]. However, integrating MEC with BSs significantly complicates the energy saving issue due to the fact that BSs now provide not only radio access services but also caching services. Furthermore, since caching resources on MECS are limited, downloading some content from the CN is inevitable. As a result, energy consumption couples the caching capacity and MECSs' sleeping decisions over time. It has been observed that the content popularity and caching capacity are two main factors affecting the MECSs' sleeping decisions. Literature [23] discussed the caching deployment problem with a given wireless transmission rate, and it also made an assumption that three factors of the backhaul transmission rate, MECSs' storage capacity and system energy consumption are fixed. However, in practical mobile networks, base stations should consider different wireless channel states and conditions as well as different types of backhaul links and system power. Thus, it is necessary and crucial for caching deployment and active MECS to consider the above three factors [7,11]. As a consequence, how to design an optimal solution to minimize energy cost while guaranteeing high user's quality of experience (QoE) is a challenging issue.

In this paper, we study the joint optimization of average download latency and average energy consumption in cellular networks with MEC integration in order to maximize the QoE for users while keeping the energy consumption of the system as low as possible. The main technical contributions of this work can be summarized as follows:

- We make a trade-off between system average download latency (SADL) and system average energy consumption (SAEC) by developing an effective caching placement strategy. Our algorithm achieves a close-to-minimum delay cost to SADL compared to the delay optimal algorithm, while minimize SAEC.
- We indicate the influence of the content popularity distribution satisfying the required number of active MECS, as well as the influence of the backhaul capacities on the required cost of system. Numerical results show that the proposed joint cost optimal (JCO) algorithm outperforms the conventional caching placement strategies achieve a significant performance improvement and effectively reduces system energy consumption.

The remainder of this paper is organized as follows. In Section 2, we first describe the system model and problem optimization. Then, we derive and propose the MECS allocation algorithm for joint cost delay and power. Simulation results are shown and discussed in Section 3. Finally, conclusion is drawn in Section 4.

2. System Model and Problem Formulation

In this section, we introduce the system model and explain the considered network architecture. In the next section, we formally introduce the optimization problem.

2.1. System Model

As a major deployment method of the MEC, we consider an edge system consisting of a BS and multiple MECSs from the set $\mathcal{M} = \{1, \cdots, M\}$, which are physically co-located and share the same power supply in the cell site. Also, the MECSs serve the content requests submitted by N mobile users (MUs). N MUs are uniformly distributed within the scope of coverage radius of BS denoted by R. The MECSs store the contents which can be downloaded by the MUs in the coverage areas of the BS. Through separated backhaul links, base stations are connected to the core network, which stores the whole content library. The storage capacity denoted by $C_Z(\text{bits})$ of each MECS is limited to handle a large set of contents in total. This architecture is depicted in Figure 1.

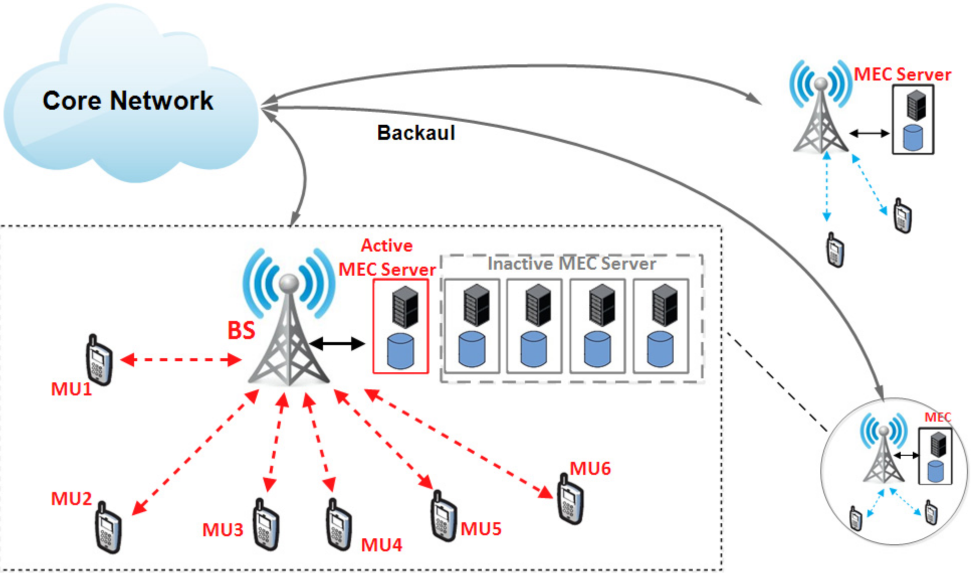


Figure 1. Architecture of Mobile Edge Computing (MEC) servers' caching system.

2.1.1. Wireless Transmission and Content Caching Model

The wireless transmission model can be simplified as follows. The transmission rate is set to be C_{MEC} (Mbps) if the MU can download the interested content from the MECS. Otherwise, a MU can only take the content from the core network (CN) by backhaul link at the rate of C_{Back} Mbps. Also, the backhaul capacity of BS is denoted as C_{bmax}(Mbps) which is generally limited in dense cellular network scenarios. We further consider that $C_{Back} < C_{MEC}$. In the model, we focus on a multi-user orthogonal frequency division multiple access (OFDMA) system in mobile networks, where each channel in the system is orthogonal to the others—in other words, no interference among MUs [24]. For simplicity, we assume that channel gains have the same distribution and small-scale fast fading will average out. Hence the wireless transmission rate of one MU depends on its available bandwidth and signal-to-noise ratio (SNR) [7]. In the cell, with available bandwidth W for each MU and a given BS transmission power P_{trans}, the wireless transmission rate of a MU is given by the Shannon's theorem:

$$C = W\left(\log_2\left(1 + \frac{P_{trans}\beta R^{-\epsilon}}{W\delta^2}\right)\right) \tag{1}$$

$$\sum_{MU} C_{Back} \leq C_{bmax} \tag{2}$$

where δ^2 is the noise power, ϵ is the pathloss exponent and β is the pathloss constant.

Some previous studies have shown that in practical networks, the request probability of content can be fitted with some popularity distributions. In the proposed work, we assume that the MUs request content (i.e., videos, files, news, etc.) from a library $\mathcal{F} = \{1, \cdots F\}$, where each content-$f$ in this library has a same size of $L(f)$ bits and different popularity. The probability content-$f(f = 1 \cdots F)$ being requested is denoted as $P_{\mathcal{F}}(f)$, i.e., $\sum_{f=1}^{F} P_{\mathcal{F}}(f) = 1$. As a matter of fact, the popularity of requested contents follows the Zipf's distribution [23–25], which can be expressed as:

$$P_{\mathcal{F}}(f) \; = \frac{\Omega}{f^{\alpha}} \tag{3}$$

where

$$\Omega \; = \left(\sum_{i=1}^{F} \frac{1}{i^{\alpha}}\right)^{-1}$$

The parameter α in equation (3) describes the steepness of the distribution. Like the distribution of contents in the web proxies and the traffic dynamics of cellular devices, this kind of power law is used to characterize many real world phenomena [23]. The higher α value corresponds to a steeper distribution, and indicates that a fraction of the content is more popular than the rest of the catalog (i.e., users have very similar interests). For another, lower values describe more uniform behavior almost as popular as the content (i.e., users have different interests) The parameter α can take different values depending on the MUs' behaviour and MECS deployment strategies (i.e., campus, enterprise, urban, and rural environments), and its practical value in our experimental setup will be given in the subsequent sections.

Without the loss of generality, the contents of the library are sorted in line with the descending order of popularity, in which content-*1* indicates the content with the highest downloading probability. As a consequence, MUs are considered to make independent downloading requests based on $P_{\mathcal{F}}(f)$. Caching the most popular contents will be regarded as the optimal caching strategy for MECS, and hence the caching hit ratio (Q) of A_c MECSs can be written as:

$$Q = 1 - \frac{\sum_{f=A_c*C_s}^{F} \frac{1}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{1}{f^{\alpha}}} \tag{4}$$

where $C_s \triangleq \frac{C_Z}{L(f)}$ is defined as the number of the caching contents of MECS , the number of active MECSs is A_c.

$$1-Q=\frac{\sum_{f=A_c*C_s}^{F}\frac{1}{f^{\alpha}}}{\sum_{f=1}^{F}\frac{1}{f^{\alpha}}}$$

Proof. The proof is presented in Appendix A.

2.1.2. Users Model

In the model, we assume that the number of MUs, denoted as $U(t)$, which requests interested contents from BS, follows the Poisson process with parameter λ [25,26]. Thus, it yields:

$$P(U(t)=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t} \tag{5}$$

2.1.3. Delay Cost Model

For MUs downloading the target content workload, we mainly consider the transmission delay as the delay cost due to the limitations of the backhaul capacity of the BS. It costs less in terms of transmission delay for a MU downloading the target content from the MECS to a MU (denoted as T_{MEC}),as compared to that from the CN to a MU (denoted as T_{Back}). In order to quantify the delay performance of services without restricting our model to any particular metric, we utilize an average response time to represent the delay cost:

$$T_{Aver}=\eta\Lambda\beta^T \tag{6}$$

$$\eta=[T_{MEC},T_{Back}] \tag{7}$$

where $T_{MEC}=\frac{L}{C_{MEC}}$, $T_{Back}=\frac{L}{C_B}$ represents the duration of one MU finishing a download on interested content from the MECS and finishing a download on interested content from the CN, respectively.

$$C_B=\begin{cases} C_{Back} \quad , if\ K_{Back}\le B(t) \\ \frac{C_{bmax}}{K_{Back}}, if\ K_{Back}>B(t) \end{cases} \tag{8}$$

where K_{Back}^n is defined as the number of MUs which finishing a download on interested content from the CN. The maximal number of MUs that can download interested content at the same time by the backhaul network is defined as $B(t)$. If $K_{Back}^n\le B(t)$, then each user's transmission rate of backhaul is C_{Back}, else if $K_{Back}^n>B(t)$, each user's transmission rate of backhaul is $C_B=\frac{C_{bmax}}{K_{Back}^n}$, $B(t)=\frac{C_{bmax}}{C_{Back}}$.

$$\Lambda=\begin{bmatrix} K_{MEC}^1 & K_{MEC}^2 & \cdots & K_{MEC}^{U(t)} \\ K_{Back}^1 & K_{Back}^2 & \ldots & K_{Back}^{U(t)} \end{bmatrix} \tag{9}$$

where K_{MEC}^n is defined as the number of MUs which finishing a download of interested content from the MECS. The average number of MUs requesting interested contents from the BS is $E[U(t)]=K_{MEC}^n+K_{Back}^n$. Since $U(t)$ yields to the Poisson process, $E[U(t)]=E\left[\frac{(\lambda t)^k}{k!}e^{-\lambda t}\right]=\lambda t.$

$$\beta=[P(1),P(2),\ldots,P(N)] \tag{10}$$

where $P(n)=\begin{pmatrix} U \\ K_{MEC}^n \end{pmatrix}H^{K_{MEC}^n}(1-H)^{K_{Back}^n}$ represents the probability that K_{MEC}^n users hit the cache simultaneously.

As such, the formulation of average response can be written as:

$$T_{Aver} = \sum_{n=0}^{U(t)} \begin{pmatrix} U(t) \\ n \end{pmatrix} Q^n (1-Q)^{U(t)-n} T_{MEC} + \sum_{n=0}^{U(t)} \begin{pmatrix} U(t) \\ n \end{pmatrix} Q^n (1-Q)^{U(t)-n} T_{Back} \tag{11}$$

The Equation (11) can be rewritten as:

$$T_{Aver} = \sum_{n=0}^{U(t)} \begin{pmatrix} U(t) \\ n \end{pmatrix} Q^n (1-Q)^{U(t)-n} \left(\frac{nL}{C_{MEC}} + \frac{(U(t)-n)L}{\min\{C_{Back}, C_{Max}\}} \right) \tag{12}$$

Proof. The proof is presented in Appendix B.

2.1.4. Power Consumption Model

The power model can be assumed as follows. The total power consumption P_{Sys} demand of the 5G cellular networks consists of operational power P_{Op}, transmission power P_{trans} , and MECSs power P_{MEC}.

$$P_{Sys} = P_{Op} + P_{trans} + P_{MEC} \tag{13}$$

The operational power is load-independent, consisting of the baseband processor, the converter, the cooling system, etc. Therefore, for the BS in time slot t:

$$P_{Op} = P_O \mu \tag{14}$$

where P_O is a constant which describes system power consumption, μ is a synchronous workload coefficient.

Transmission occurs on wireless links between the MU and the BS, as well as the backhaul link between the BS and CN. Usually the wireless transmission power consumption dominates, so that we consider only the wireless portion. We assume that the small-scale fast fading will average out since the considered time slot is relatively long. Hence, we focus on pathloss effects. We can approximate the pathloss effect by considering the maximum coverage radius (R) of the BS, in order to keep the maximum achievable transmission rate of all users which are under the coverage of the BS larger than C_o [27,28]. Given the transmission power P_{trans}, the maximum achievable transmission rate is given by the Shannon's theorem,

$$C_O = W\left(\log_2\left(1 + \frac{P_{trans}\beta R^{-\epsilon}}{W\delta^2}\right)\right) \tag{15}$$

where W is the channel bandwidth, δ^2 is the noise power, ϵ is the pathloss exponent and β is the pathloss constant. We suppose the noise-limited is set by assuming that the BS operates on orthogonal channels [28–30].

Similarly, we consider each transmission must meet a target rate C_o to satisfy a transmission delay requirement, the transmission power should satisfy:

$$P_{trans} = \frac{\left(2^{C_o W^{-1}} - 1\right) W\delta^2}{\beta R^{-\epsilon}} \tag{16}$$

The MECS system power at the edge server is load-dependent. Let

$$P_{MEC} = P_{Server}\rho^T \tag{17}$$

where $\rho = [a_1, a_{2,\ldots,}a_N]$, let $a_N \in \{0,1\}$ represent the active (1)/inactive (0) decision for the MECSs.

2.2. Problem Formulation

2.2.1. Delay Optimization

Firstly, we want to quantify the delay performance of services without restricting our model to any particular metric (for the delay-optimal algorithm). Thus, the minimum delay cost can be computed as:

$$P1: \underset{P}{arg\ min}\ T_{Aver} \tag{18}$$

$$\text{s.t. } C_b \leq C_{bmax}$$

$$\sum\nolimits_{f=1}^{C_s} L(f) \leq C_Z$$

The optimization problem is a linear programming problem with a computational complexity of $O(F * K * C_s)$ and can be solved by using a conventional solver, i.e., MATLAB.

2.2.2. Joint Delay Cost with Power Consumption Optimization

In this subsection, we will give the joint optimization for time delay and energy cost. Due to the different impact of time and energy cost, we introduce a weight factor, denoted as ω, which indicates the emphasis on either time or energy cost. Thus, minimizing the time and energy cost of system can be specified as the following problem:

$$P2: \ arg \min_{K} \left(T_{Aver} - T_{opt}\right)^2 + \omega * P_{Sys} \tag{19}$$

$$\text{s.t. } C_b \leq C_{bmax}$$

$$\sum\nolimits_{f=1}^{K*C_s} f \leq F$$

$$K = |\rho|_1$$

The joint optimization problem is hard to solve, so we decouple $\mathcal{P}2$ problem in two stages as follows:

Stage 1: We consider the system power allocation is fixed, i.e., P_{MEC} is already known. We find that the optimization of $P2$ can be decoupled into K sub-problems with regard to the number of active MECSs, and the series of problems $G\{K\}$, in order to find the minimum delay cost for each sub-problem K, which follows the solution for the $P1$.

Stage 2: In this stage, we focus on minimizing the energy cost in the JCO algorithm.

Algorithm 1. Joint Cost Optimal (JCO) algorithm

```
Set $T_{opt} = 0$, $P_{MEC} = 0$, and $T_{Aver} = 0$
whlie
Calculate $T_{opt}$, $P_{MEC}$, $T_{Aver}$, $P_{Sys}$, then set $P2 = P2'$
    If (6) is satisfied for active a new MECS
    If (8) is satisfied for inactive a new MECS
    If $\left(T_{Aver} - T_{opt}\right) > \theta$ then
        $G\{K\} = G\{K+1\}$ ,and use genetic algorithm minimum $P1$
    else then
    obtain $P'_{Sys}$, $T'_{Aver}$
    update $P2'$ ,and if $\frac{P2'}{T'_{Aver}} > \delta$ then reset $G\{K\}$
End If
Until $P2' - P2 < \varphi$, end whlie
Output $T_{Aver}$,$P_{Sys}$,$K$, $P2$
```

3. Numerical Results

In this section, we present numerical results to verify the effect of energy efficiency scheme and illustrate the impact of various MEC network parameters. In simulations, we use the MEC network parameters as shown in Table 1.

Table 1. List of main simulation parameters.

Parameters	Value
Number of MECS	M_{server} = 15∼20
Number of MU	MU = 10∼50 [25]
Radius of the MEC range	R = 100 m
Number of alternative contents in total	F = 1000
Storage capacity of one MECS	C_z = 3∼10 GB
Size of each content	L= 50∼100 Mb [23]
Maximum transmission rate of MEC to MU	C_{MEC} = 2 Mbps
Maximum transmission rate of CN to MU	C_{Back} = 1 Mbps
The backhaul capacity of BS	C_{bmax} = 15∼30 Mbps
Noise power	$\delta^2 = -102$ dbm [7,33]
BS transmit rate requirement	C_o = 2 Mbps
Contents Request Pattern	α = 0.56~1.16 [7,10,31,32]
Power of system	P_{Op} = 800 W [35]
Power of one MECS	P_{Server} = 200 W
Transmit power of BS	P_{trans} = 20 W
Transmit bandwidth W	10 MHz
Path-loss exponent α	4 [10]
Trade-off weight	ω = 0.5 [34]

3.1. Impact of the Backhaul Capacities

We analyze the impact of the backhaul capacities on the algorithms' performance in Figure 2, meanwhile illustrating the average power consumption and average delay cost. As expected, increasing backhaul capacities not only reduces average power consumption but also decreases average delay cost. We observe that if the number of active MECSs equals to 0, all contents will be downloaded with the rate C_o from the CN, so $T_{Aver} = \frac{U(t)L}{C_b}$. If the number of active MECSs equals to ∞, the MECSs are enough to ensure that each content can be downloaded from the MECS, so the transmission rate is always C_{MEC}, and thus $T_{Aver} = \frac{U(t)L}{C_{MEC}}$. On the other hand, for a low value of backhaul capacities increasing the MECS's power consumption, we can reduce the average delay cost effectively when the system is in overloaded conditions, as the MECS can serve more requests. However, if excessively active MECSs cache a lot of unpopular content, it brings about a bad performance in effectively reducing average delay cost. Figure 2 explicitly demonstrates the trade-off between the average delay cost and the average power consumption for given backhaul capacity.

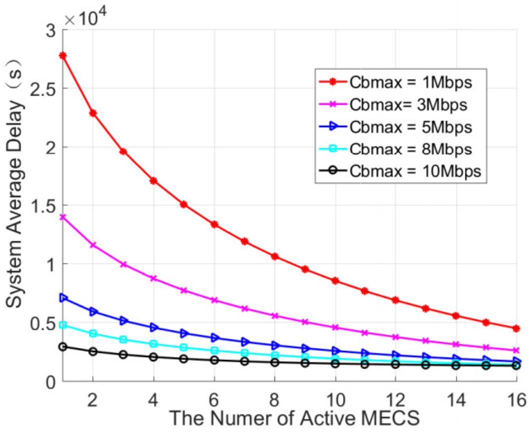


Figure 2. System joint performance versus backhaul capacity.

3.2. Impact of the Content's Popularity Pattern

We analyze the impact of the content's popularity pattern on the algorithms' performance in Figure 3. Namely, we vary the shape parameter α of the contents popularity with the value 0.46 to 1.16. As expected, with increase of α, the active MECSs decrease in the scheme, and the energy efficiency improves as the popularity distribution gets steeper. When α is high, the vast majority of user requests refer to a small number of contents. Clearly, caching the above contents provides significant benefits to the provider. To conclude, when the content popularity distribution is highly concentrated (i.e., $\alpha > 1$), our algorithm achieves at least 1.5 times higher performance than $\alpha < 1$.

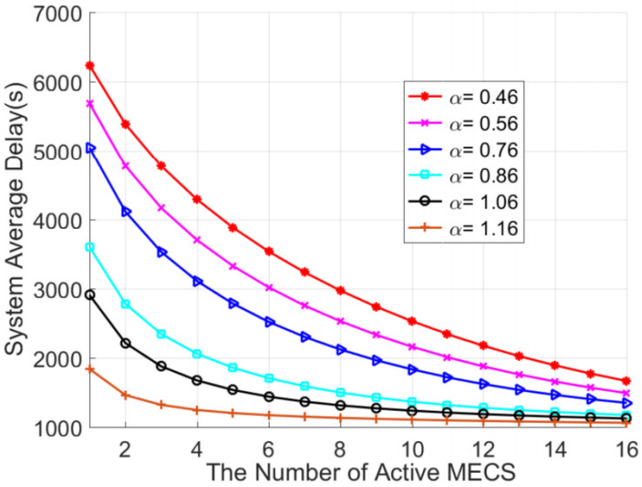


Figure 3. System performance versus popularity shape parameter.

3.3. Impact of the Joint System Cost

Figure 4 shows the joint performance of the delay optimal, energy optimal and joint optimal algorithms. In Figure 4, we observe that the proposed algorithm outperforms the other two schemes in terms of the system cost. As expected, increasing backhaul capacities not only reduces the average power consumption but also decreases the average delay cost. Therefore, with the increase of backhaul capacities, the joint cost rapidly decreases.

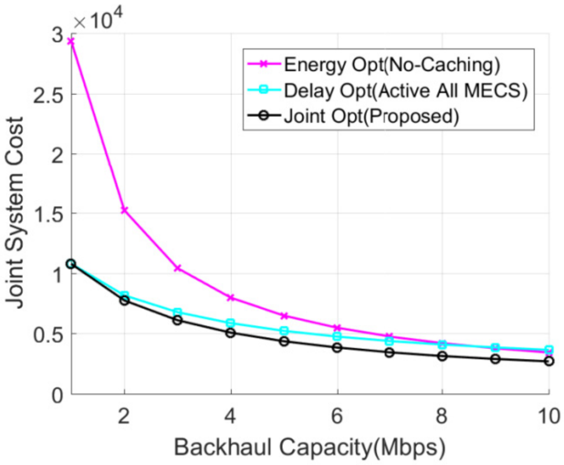
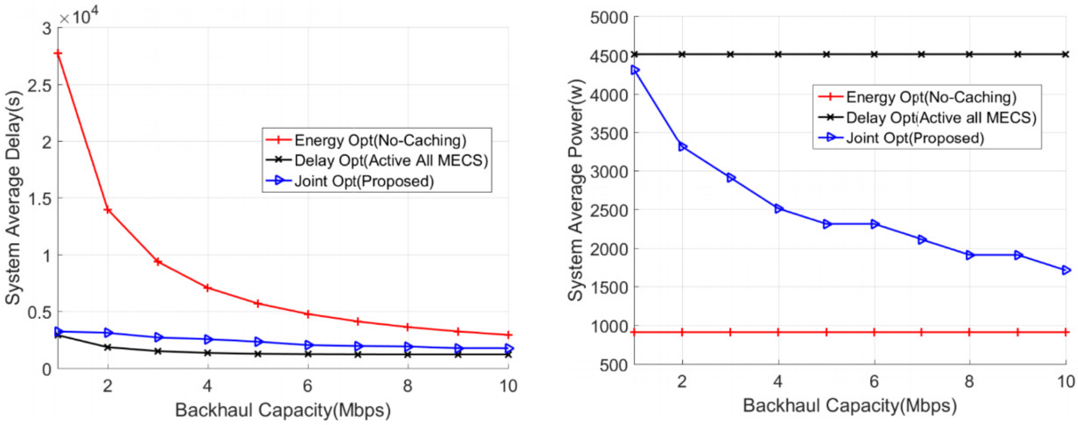


Figure 4. Joint system cost versus backhaul capacities.

Figure 5 shows the delay and energy performance of the delay optimal, energy optimal and joint optimal algorithms. In Figure 5a, the delay optimal algorithm achieves the best delay performance at the maximum cost of energy, as shown in Figure 5b. Otherwise, the energy optimal algorithm achieves the minimum cost of energy at the worst delay performance. Obviously, both methods have very low energy efficiency. In Figure 5a, the performance of the proposed joint optimal is in close proximity to the delay optimal, which achieves the theoretical delay performance but much better than the energy optimal. In Figure 5b, although the performance of the proposed method causes a slightly larger delay than the delay optimal (increasing minimum delay by around 20%), energy consumption can be reduced up to about 63%.



(**a**) (**b**)

Figure 5. (**a**) Delay performance; (**b**) Energy performance.

4. Conclusions

In this paper, we have focused on energy-efficient strategies for MECS in a backhaul capacity-limited cellular network for minimizing the power consumption while satisfying a computation delay cost constraint. For the motivation, taking the backhaul capacity, contents popularities and the number of users into account, a constraint expression that can trade off the system energy consumption and average delay has been derived, and it can illustrate the impacts of different MECS network parameters. The numerical results indicate that our method can reduce the energy consumption by about 63% with the trade-off in delay efficiency (increasing the minimum delay by around 20%), perform very close to the optimal solution, and much better than the worst-case scenario, i.e., the approximation bound.

Acknowledgments: The work presented in this paper was partially supported by National High Technology Research and Development Program(Grant number 2015AA01A707), 2014 National Natural Science Foundation of China (Grant number 61371081), 2015 National Natural Young Science Foundation of China (Grant number 61401381).

Author Contributions: All the authors contributed to the conception, design and performance of the experiment, the analysis of the data and the writing of the paper. Zhaohui Luo and Lianfen Huang initiated and discussed the research problem; Zhaohui Luo conceived and designed the scheme; Zhaohui Luo, Minghui LiWang and Zhijian Lin performed the experiments and made the figures. Zhaohui Luo and Xiaojiang Du and Mohsen Guizani analyzed the data; Zhaohui Luo, Minghui LiWang, Lianfen Huang, Zhijian Lin, Xiaojiang Du and Mohsen Guizani wrote the paper.

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

Abbreviations

The following abbreviations are used in this manuscript:

MEC	Mobile Edge Computing
MECS	Mobile Edge Computing Server
MNOs	Mobile network operators
MCS	Mobile communication system
MUs	Mobile Users
HD	High Definition
EE	Energy efficiency
BS	Base station
OFDMA	Orthogonal Frequency Division Multiple Access
SNR	Signal-to-Noise Ratio
SADL	System Average Download Latency
SAEC	System Average Energy Consumption
CN	Core Network

Appendix A

From Equations (4), Q can be expressed as the following:

$$P_{\mathcal{F}}(f) = \frac{\Omega}{f^{\alpha}} = \frac{\left(\sum_{i=1}^{F} \frac{1}{i^{\alpha}}\right)^{-1}}{f^{\alpha}}$$

when the number of active one MECS , the hit ratio can be written as:

$$Q = \frac{\sum_{f=1}^{C_s} \frac{\Omega}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{\Omega}{f^{\alpha}}} = \frac{\sum_{f=1}^{C_s} \frac{\left(\sum_{i=1}^{F} \frac{1}{i^{\alpha}}\right)^{-1}}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{\left(\sum_{i=1}^{F} \frac{1}{i^{\alpha}}\right)^{-1}}{f^{\alpha}}} = \frac{\sum_{f=1}^{C_s} \frac{1}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{1}{f^{\alpha}}}$$

hence when the number of active MECS is A_c

$$Q = \frac{\sum_{f=1}^{A_c * C_s} \frac{1}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{1}{f^{\alpha}}} = \frac{\sum_{f=1}^{F} \frac{1}{f^{\alpha}} - \sum_{A_c * C_s}^{F} \frac{1}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{1}{f^{\alpha}}} = 1 - \frac{\sum_{A_c * C_s}^{F} \frac{1}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{1}{f^{\alpha}}}$$

Hence $1 - Q$ can be written as:

$$1 - Q = \frac{\sum_{A_c * C_s}^{F} \frac{1}{f^{\alpha}}}{\sum_{f=1}^{F} \frac{1}{f^{\alpha}}}$$

This completes the proof.

Appendix B

Proof.

Using (16), we can rewrite

$$P(n) = \begin{pmatrix} U \\ K_{MEC}^{n} \end{pmatrix} Q^{K_{MEC}^{n}} (1 - Q)^{K_{Back}^{n}}$$

So the formulation of average response can be written as:

$$T_{Aver} = \sum_{n=0}^{U(t)} \begin{pmatrix} U(t) \\ n \end{pmatrix} Q^{n} (1 - Q)^{U(t) - n} T_{MEC} + \sum_{n=0}^{U(t)} \begin{pmatrix} U(t) \\ n \end{pmatrix} Q^{n} (1 - Q)^{U(t) - n} T_{Back}$$

$$\sum_{n=0}^{U(t)} \begin{pmatrix} U(t) \\ n \end{pmatrix} Q^{n} (1 - Q)^{U(t) - n} T_{Back}$$

$$= \sum\nolimits_{n=0}^{j} \binom{j}{n} Q^n (1-Q)^{j-n} \frac{nL}{C_{Back}} + \sum\nolimits_{n=j}^{U(t)} \binom{U(t)}{n} Q^{U(t)} (1-Q)^{U(t)-n} \frac{(U(t)-n)^2 L}{C_b}$$

Then

$$\sum\nolimits_{n=0}^{U(t)} \binom{U(t)}{n} Q^n (1-Q)^{U(t)-n} T_{Back} = \sum\nolimits_{n=0}^{U(t)} \binom{U(t)}{n} Q^n (1-Q)^{U(t)-n} \frac{(U(t)-n)L}{\min\{C_{Back}, C_{Max}\}}$$

$$T_{Aver} = \sum\nolimits_{n=0}^{U(t)} \binom{U(t)}{n} Q^n (1-Q)^{U(t)-n} \left(\frac{nL}{C_{MEC}} + \frac{(U(t)-n)L}{\min\{C_{Back}, C_{Max}\}} \right)$$

This completes the proof.

References

1. Paschos, G.; Bastug, E.; Land, I.; Caire, G.; Debba, M. Wireless caching: Technical misconceptions and business barriers. *IEEE Commun. Mag.* **2016**, *54*, 16–22. [CrossRef]
2. Cisco CVNI. Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper. Document ID: 1465272001663118. Available online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/ (accessed on 1 June 2016).
3. Li, C.; Zhang, J.; Letaief, K.B. Throughput and energy efficiency analysis of small cell networks with multi-antenna base stations. *IEEE Trans. Wirel. Commun.* **2014**, *13*, 2505–2517.
4. Hu, Y.C.; Patel, M.; Sabella, D.; Sprecher, N.; Young, V. Mobile edge computing-A key technology towards 5G. In *ETSI White Paper*; ETSI: Antipolis, France, 2015.
5. Roman, R.; Lopez, J.; Mambo, M. Mobile edge computing, Fog et al.: A survey and analysis of security threats and challenges. *Future Gener. Comput. Syst.* **2016**. [CrossRef]
6. Baştuğ, E.; Bennis, M.; Kountouris, M.; Debbah, M. Cache-enabled small cell networks: Modeling and tradeoffs. *EURASIP J. Wirel. Commun. Netw.* **2015**, *1*, 1–11.
7. Peng, X.; Zhang, J.; Song, S.H. Cache size allocation in backhaul limited wireless networks. In Proceedings of the IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 23–27 May 2016.
8. Zeydan, E.; Bastug, E.; Bennis, M.; Kader, M.A. Big data caching for networking: Moving from cloud to edge. *IEEE Commun. Mag.* **2016**, *54*, 36–42. [CrossRef]
9. Zheng, K.; Yang, Z.; Zhang, K.; Chatzimisios, P.; Yang, K.; Xiang, W. Big data-driven optimization for mobile networks toward 5G. *IEEE Netw.* **2016**, *30*, 44–51. [CrossRef]
10. Poularakis, K.; Iosifidis, G.; Tassiulas, L. Approximation algorithms for mobile data caching in small cell networks. *IEEE Trans. Commun.* **2014**, *62*, 3665–3677. [CrossRef]
11. Peng, X.; Shen, J.C.; Zhang, J.; Letaief, K.B. Backhaul-aware caching placement for wireless networks. In Proceedings of the IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA, 6–10 December 2015.
12. Maddah-Ali, M.A.; Niesen, U. Decentralized coded caching attains order-optimal memory-rate tradeoff. *IEEE/ACM Trans. Netw.* **2015**, *23*, 1029–1040. [CrossRef]
13. Niesen, U.; Maddah-Ali, M.A. Coded caching with nonuniform demands. *Trans. Inform. Theory.* **2017**, *63*, 1146–1158. [CrossRef]
14. Antonopoulos, A.; Kartsakli, E.; Bousia, A.; Alonso, L.; Verikoukis, C. Energy-efficient infrastructure sharing in multi-operator mobile networks. *IEEE Commun. Mag.* **2015**, *53*, 242–249. [CrossRef]
15. Zhang, H.; Huang, S.; Jiang, C.; Long, K.; Leung, V.C.M.; Poor, H.V. Energy Efficient User Association and Power Allocation in Millimeter Wave Based Ultra Dense Networks with Energy Harvesting Base Stations. *IEEE J. Sel. Areas Commun. arXiv* **2017**, arXiv:1704.07037.
16. Zhang, H.; Nie, Y.; Cheng, J.; Leung, V.C.; Nallanathan, A. Sensing time optimization and power control for energy efficient cognitive small cell with imperfect hybrid spectrum sensing. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 730–743. [CrossRef]
17. Bousia, A.; Kartsakli, E.; Antonopoulos, A.; Alonso, L.; Verikoukis, C. Game-theoretic infrastructure sharing in multioperator cellular networks. *IEEE Trans. Veh. Technol.* **2016**, *65*, 3326–3341. [CrossRef]
18. Datsika, E.; Antonopoulos, A.; Zorba, N.; Verikoukis, C. Green cooperative device-to-device communication: A social-aware perspective. *IEEE Access* **2016**, *4*, 3697–3707. [CrossRef]

19. Wu, J.; Bao, Y.; Miao, G.; Zhou, S.; Niu, Z. Base-station sleeping control and power matching for energy-delay tradeoffs with bursty traffic. *IEEE Trans. Veh. Technol.* **2016**, *65*, 3657–3675. [CrossRef]
20. Yu, N.; Miao, Y.; Mu, L.; Du, H.; Huang, H.; Jia, X. Minimizing Energy Cost by Dynamic Switching ON/OFF Base Stations in Cellular Networks. *IEEE Trans. Wirel. Commun.* **2016**, *15*, 7457–7469. [CrossRef]
21. Bousia, A.; Kartsakli, E.; Antonopoulos, A.; Alonso, L.; Verikoukis, C. Multiobjective Auction-Based Switching-Off Scheme in Heterogeneous Networks: To Bid or Not to Bid? *IEEE Trans. Veh. Technol.* **2016**, *65*, 9168–9180. [CrossRef]
22. Han, F.; Zhao, S.; Zhang, L.; Wu, J. Survey of strategies for switching off base stations in heterogeneous networks for greener 5G systems. *IEEE Access* **2016**, *4*, 4959–4973. [CrossRef]
23. Baştuğ, E.; Bennis, M.; Zeydan, E.; Kader, M.A.; Karatepe, I.A. Big data meets telcos: A proactive caching perspective. *J. Commun. Netw.* **2015**, *17*, 549–557. [CrossRef]
24. Zhang, K.; Mao, Y.; Leng, S.; Zhao, Q.; Li, L.; Peng, X. Energy-Efficient Offloading for Mobile Edge Computing in 5G Heterogeneous Networks. *IEEE Access* **2016**, *4*, 5896–5907. [CrossRef]
25. Ding, R.; Wang, T.; Song, L.; Han, Z.; Wu, J. Roadside-unit caching in vehicular ad hoc networks for efficient popular content delivery. In Proceedings of the Wireless Communications and Networking Conference (WCNC), New Orleans, LA, USA, 9–12 March 2015; pp. 1207–1212.
26. Hu, Z.; Zheng, Z.; Wang, T.; Song, L. Poster: Roadside Unit Caching Mechanism for Multi-Service Providers. In Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing, Hangzhou, China, 22–25 June 2015; pp. 387–388.
27. Chen, L.; Zhou, S.; Xu, J. Energy Efficient Mobile Edge Computing in Dense Cellular Networks. *arXiv* **2017**, arXiv:1701.07405.
28. Xu, J.; Sun, Y.; Chen, L.; Zhou, S. E2M2: Energy Efficient Mobility Management in Dense Small Cells with Mobile Edge Computing. *arXiv* **2017**, arXiv:1701.07363.
29. Rimal, B.P.; Van, D.P.; Maier, M. Mobile Edge Computing Empowered Fiber-Wireless Access Networks in the 5G Era. *IEEE Commun. Mag.* **2017**, *55*, 192–200. [CrossRef]
30. Mao, Y.; Zhang, J.; Letaief, K.B. Dynamic computation offloading for mobile-edge computing with energy harvesting devices. *IEEE J. Sel. Areas Commun.* **2016**, *34*, 3590–3605. [CrossRef]
31. Liu, J.; Bai, B.; Zhang, J.; Letaief, K.B. Content caching at the wireless network edge: A distributed algorithm via belief propagation. In Proceedings of the IEEE International Conference on Communications (ICC), Kuala Lumpur, Malaysia, 23–27 May 2016.
32. Gill, P.; Arlitt, M.; Li, Z.; Mahanti, A. Youtube traffic characterization: a view from the edge. In Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, CA, USA, 24–26 October 2007; pp. 15–28.
33. Lee, W.; Jung, B.C. Energy-Efficient On-Off Power Control of Femto-Cell Base Stations for Cooperative Cellular Networks. *Appl. Sci.* **2016**, *6*, 356. [CrossRef]
34. Chen, M.; Hao, Y.; Lai, C.F.; Wu, D.; Li, Y.; Hwang, K. Opportunistic task scheduling over co-located clouds in mobile environment. *IEEE Trans. Ser. Comput.* **2016**. [CrossRef]
35. Xu, J.; Ren, S. Online Learning for Offloading and Autoscaling in Renewable-Powered Mobile Edge Computing. *arXiv* **2016**, arXiv:1609.05087.