Base Station ON-OFF Switching in 5G Wireless Networks: Approaches and Challenges

Mingjie Feng, Shiwen Mao, and Tao Jiang

Abstract

To achieve the expected 1000× data rates under the exponential growth of traffic demand, a large number of BSs or APs will be deployed in 5G wireless systems to support high data rate services and to provide seamless coverage. Although such BSs are expected to be small-scale with lower power, the aggregated energy consumption of all BSs would be remarkable, resulting in increased environmental and economic concerns. In existing cellular networks, turning off the underutilized BSs is an efficient approach to conserve energy while preserving the QoS of mobile users. However, in 5G systems with new physical layer techniques and highly heterogeneous network architecture, new challenges arise in the design of BS ON-OFF switching strategies. In this article, we begin with a discussion of the inherent technical challenges of BS ON-OFF switching. We then provide a comprehensive review of recent advances on switching mechanisms in different application scenarios. Finally, we present open research problems and conclude the article.

Introduction

With the growing popularity and versatility of mobile devices, the age of the mobile Internet has come. According to a report from comScore Inc., mobile devices accounted for 55 percent of Internet usage in the United States in January 2014, which is the first time that mobile apps overtook the personal computer (PC) for Internet usage in the United States. As a result, the requirements for data-intensive wireless services are boosted at an unprecedented rate, posing great pressure on current wireless systems. To satisfy such demand, the fifth generation (5G) wireless system is under intensive development, and it is expected to provide ubiquitous Internet access with 1000× data rate. Compared to previous generations of mobile communication systems, the 5G system not only involves innovations in physical layer techniques, but also introduces new network architectures and application scenarios. In particular, a major trend in 5G networks is the deployment of a large number of small-scale base stations (BSs) or access points (APs), also known as *network densification*. For example, in the millimeter-wave (mmWave) network, small cell network, distributed antennas system, and femtocaching-enabled network, BSs are deployed close to users to reduce propagation loss, to improve signal-to-noise ratio (SNR), and to reduce service delay. However, these benefits come at a price: the massive BS deployment significantly increases the total energy consumption of wireless systems. For a typical LTE microcell with a cell size of 100 m and bandwidth of 5 MHz, the power consumption ranges from 25 to 40 W depending on the traffic load. To achieve the coverage of a 1500 m macrocell, more than 200 microcells need to be deployed, and the aggregated power consumption is comparable to a typical LTE macrocell BS with a cell size of 1500 m, which has a power consumption of 900 W. The increased energy consumption not only increases the cost of wireless operators, but also generates more greenhouse gas emissions. Thus, energy saving has become an important design objective of wireless systems in recent years. Meanwhile, energy saving needs to be achieved without sacrificing the quality of service (QoS) of users. As the 5G system is expected to provide 1000× data rates, energy efficiency (EE), typically measured by bits per Joule, also needs to be increased by 1000× if the total energy consumption is to be maintained at its original level.

BS ON-OFF switching (also known as BS sleep control) is considered to be an efficient approach for both energy saving and EE improvement. As the traffic pattern fluctuates over both time and space, underutilized BSs can be dynamically turned off to save energy [1]. In 2009, China Mobile began to apply BS sleep control, and the estimated reduction of energy consumption is 36 million kWh/year since that time. Due to such great potential, considerable efforts have been devoted to the design of BS ON-OFF switching strategies in different network scenarios. However, as the 5G system is an integration of different techniques with a highly heterogeneous network architecture [2], the design of BS ON-OFF switching faces special challenges in 5G systems, which can be summarized as follows.

Interoperability with New Techniques: With new techniques in 5G systems, additional constraints and impacts emerge. It is necessary to adjust the BS ON-OFF switching strategy accord-

Mingjie Feng and Shiwen Mao are with Auburn University.

Tao Jiang is with Huazhong University of Science and Technology.

Digital Object Identifier: 10.1109/MWC.2017.1600353

 1536-1284/17/$25.00 © 2017 IEEE

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

ingly. For example, when device-to-device (D2D) communication is available for user equipments (UEs), a UE can still send/receive data to/from a BS with the help of another UE even when it is not within the coverage of the BS. In this case, some BSs can be turned off without worrying about the resulting coverage holes. As another example, when 5G systems operate on unlicensed bands that are currently used by other systems (e.g., WiFi), the BS ON-OFF pattern may be accommodated to maintain the QoS of WiFi users.

Applications in New Network Architectures: In the new 5G network architecture, the functionality and properties of BSs are significantly changed. For instance, in the cloud radio access network (C-RAN), users are connected to remote radio units (RRUs) while the data processing is performed by centralized baseband units (BBUs). Since a BBU operates with much higher power than an RRU, turning off underutilized BBUs is desirable for energy saving. However, when a BBU is switched off, all RRUs linked to the BBU cannot provide service to users. To prevent outage of users, either these RRUs should be connected to a new BBU or the users are handed over to other available RRUs, resulting in a complicated scheduling problem. As another example, when a BS in an mmWave network is turned off, its coverage can hardly be compensated by neighboring BSs due to high propagation loss and blockage by obstacles. Hence, the QoS of users cannot be guaranteed with the existing BS ON-OFF scheduling, so clearly, new approaches are required to address these issues.

Large Number of BSs: To satisfy the sizable aggregated data rate requirements in hotspots, the BSs may be densely deployed with a certain level of redundancy. Thus, turning off a BS increases the traffic loads of multiple nearby BSs. The ON-OFF states and QoS provisioning between different BSs could be highly interdependent. With numerous low-power BSs, a significant amount of BSs need to be switched off to achieve energy saving. This results in a combinatorial problem with a large number of variables, which is generally difficult to solve with standard techniques. In addition, BS ON-OFF switching also impacts the interference pattern if these BSs operate on the same spectrum band. When coordinated multipoint (CoMP) transmission is available, each user can be served by multiple BSs, and the problem becomes even more complicated. A large number of BSs also causes scalability issues, which is a key factor that affects the feasibility and effectiveness of scheduling algorithms.

To harvest the benefits of BS ON-OFF switching with these challenges, it is necessary to investigate the technical aspects of switching mechanisms and analyze the challenges in the new 5G system context, where different techniques and network architectures are integrated. This article aims to identify the key challenges of BS ON-OFF switching and provide insights into its applications in 5G systems. We first analyze the technical aspects and challenges of the ON-OFF switching operation, followed by a review of the recent advances in different wireless systems. Then we discuss open research problems in some emerging application scenarios and present potential solutions.

Technical Aspects and Challenges of BS ON-OFF Switching

Energy Consumption Model

BS energy consumption models have evolved from simple and approximated models to more sophisticated ones, depending on the BS type and application scenario. A BS consumes a certain amount of energy to maintain its normal operation, including energy for its circuits, cooling system, and so on. Since such static parts constitute a dominant proportion of the total BS energy consumption, binary models are used to approximate the energy consumption in ON-OFF states [3–5]. To capture the impact of traffic load, the dynamic load-dependent part is also considered. For example, the dynamic part can be proportional to the number of users being served [6] or to the number of active antennas [7].

The BS power in the sleep mode is regarded as zero in some works, since it is small compared to the power when the BS is fully active. However, to ensure that the BS can be activated, it should not be completely powered off; it may still consume a certain amount of power, such as detection power [8]. Such power is not negligible, especially for small-scale BSs without on-site cooling systems. Thus, the power under sleep mode is also studied in recent models, and the sleep mode is further classified into different levels of sleep. In [9], the concept of opportunistic sleep is proposed, in which a BS only sleeps in certain time periods to improve system reliability. In [10], the BS states are further classified into four types, including ON, standby, sleep, and OFF, with power consumption ratios given as 100, 50, 15, and 0 percent, respectively.

Traffic Model and Traffic-Aware Scheduling

The traffic models used in prior works are summarized in Table 1. Among these models, the key factors include user arrival rate, user distribution over space, file/task size, and service rate. Based on these parameters, the stochastic geometry framework was widely used to analyze the theoretical system performance. However, since traffic demands from users are heterogeneous in nature, they cannot be characterized by a single traffic model. For example, some demands are delay-sensitive (e.g., a phone call), some are rate-sensitive (e.g., a file download), some are both delay- and rate-sensitive (e.g., online gaming and video conferencing), while some are neither (e.g., information gathered from a sensor network). Considering that BSs can work in different modes with different service provisioning capabilities, one can adjust the operating modes of BSs according to the traffic type and pattern to further enhance system performance.

Considering that BSs can work in different modes with different service provisioning capabilities, one can adjust the operating modes of BSs according to the traffic type and pattern to further enhance system performance.

We use a simple and highly abstracted indicative example to show the potential of traffic-aware scheduling. Consider a given area with multiple BSs, where each BS has two sleep states, namely *standby* and *deep sleep*. The standby state consumes higher power but wakes up more quickly than the deep sleep state. We assume that BSs in both states wake up periodically, and it takes a certain amount of energy to wake up. Without loss of generality, we consider an example of a

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

	BS type	BS energy/power model	Traffic type	Timescale
[6]	Regular BS	ON: fixed + load-proportional; OFF: 0	PPP-based traffic arrival; log normally distributed file size	Slow: BS ON/OFF switching. Fast: user association.
[3]	Regular BS	Fixed value	Poisson arrival UEs, exponential file size.	Slow: duration in objective function. Fast: update of traffic pattern.
[8]	Regular BS	Active: idle power + load-dependent; Sleep: sleep power + detection power.	Poisson arrival tasks; general distribution service time.	Hysteresis sleep time and wake up period.
[4]	Small cell BS	Fixed value	PPP-based user distribution	Traffic pattern update period
[10]	Small cell BS	Four modes: ON (100%), standby (50%), sleep (15%), and OFF (0%).	PPP-based user distribution	Slow: BS ON/OFF switching. Fast: time required to wake up
[11]	Macrocell BS and small cell BS	MBS: Fixed + load-dependent, SBS: fixed for ON and OFF.	Uniform and non-uniform PPP-based user distribution	BS ON/OFF switching
[5]	Small cell BS	Fixed value	Uniform and non-uniform random user distribution	Slow: BS ON/OFF switching. Fast: user association
[13]	BBU and fronthaul	Proportional to the numbers of active BBUs and fronthaul links.	A given traffic distribution on a daily basis	Period for network planning Period for traffic engineering
[7]	Multi-antenna BS	Four parts: circuit + backhauling + transmission + processing.	A certain amount of users to be served	Instantaneous, average CSI, and BS ON/OFF switching.
[9]	BS with EH	Active: fixed + $\frac{\text{RF power}}{\text{Efficiency}}$. Opportunistic sleep: a fraction of active. Deep sleep: 0.	Poisson arrival UEs; given service rate; user data rate requirements.	Period for metric evaluation
[12]	BS with EH	Active: fixed + $\frac{\text{RF power}}{\text{Efficiency}}$. Sleep: 0.	PPP-based user distribution	Large: user density, harvest rate. Small: user location, battery level.
[14]	BS powered by smart grid	On grid + renewable	Five scenarios with different arrival rates	Long-term: one day. Short-term: one hour.
[15]	Regular BS	Fixed + load-dependent	Not specified	Not specified

TABLE 1. Different aspects of system model regarding BS ON-OFF switching. PPP: Poisson point process.

traffic pattern in which half of the traffic demands are delay-sensitive and the other half are rate-sensitive. To fully utilize the advantages of BSs in both states for both energy saving and delay reduction, it is obvious that half of the BSs should be in standby states to serve the delay-sensitive demands, while the other half should be in deep sleep to serve the rate-sensitive demands. We call such a schedule an adaptive partial scheme and compare it to other two schemes with only one BS sleep state.

In Fig. 1, we obtain different energy-delay pairs by changing the wake up intervals. For example, a standby BS reduces its energy consumption by prolonging its duration of sleep; a deep sleep BS reduces its service delay by waking up more frequently and sleeping for shorter durations. It can be seen in Fig. 1 that the adaptive scheme outperforms both benchmark schemes, indicating that a considerable performance gain can be achieved by adjusting the operation states of BSs according to traffic type.

PRACTICAL AND IMPLEMENTATION CONCERNS

Timescale of Operation: Since the ON-OFF states directly determine the QoS of users, BS ON-OFF switching is always coupled with other design issues, such as user association and traffic offloading. As discussed, BS ON-OFF switching uses both time and energy; it is thus infeasible to perform ON-OFF switching frequently. However, the system states are usually updated at a faster pace, and some technical approaches have to be executed more frequently. For example, due to user mobility, user association is updated more frequently than BS ON-OFF switching. In addition, as the performance of a wireless system largely depends on the channel condition, which is rapidly changing, it is necessary to consider time-averaged channel state information (CSI) instead of instantaneous CSI. The timescale issues of different scenarios are also summarized in Table 1.

How to Acquire System Information and Wake Up when Sleeping: When a BS is turned off, the normal transmission between the BS and UEs is suspended. Hence, the information from nearby UEs, such as CSI and traffic load, cannot be acquired by the BS from the uplink signals. To guarantee the effectiveness of scheduling, the BSs need to be aware of the environment even when they are in the sleep mode so that they can be activated in a timely manner. In [3], waking up a BS is performed with the assistance of neighboring BSs. When a BS is switched off, the neighboring BSs record their own system loads, and use

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

such information as the criterion for BS activation. This way, the neighboring BSs know when and under what conditions a BS should be turned on and can inform the BS to do so.

Alternatively, a BS can wake itself up periodically to detect the environment. It returns to sleep when the criterion for switching on is not satisfied [5, 9, 10]. Under this model, the wake up interval, which determines the wake up frequency, is a key design factor. We show that the wake up interval can be optimized under different sleep modes and traffic patterns with an illustrative example below.

Consider a heterogeneous network (HetNet) with one macrocell BS (MBS) and multiple small cell BSs (SBSs). The MBS is always active to guarantee coverage, while the SBSs can be dynamically switched on or off. The SBSs can operate in two modes: the standby mode and the deep sleep mode. We consider both high- and low-mobility scenarios. In the high-mobility scenario, the number of users in a small cell has a larger variation compared to that in the low-mobility scenario. The system configuration from [5] is used in the simulation. We compare the EE performance under different wake up intervals.

As shown in Fig. 2, a proper value of wake up interval can maximize the EE in all the scenarios. The standby mode outperforms the deep sleep mode when the wake up interval is short, since it can closely capture the dynamics of traffic load. When the wake up interval becomes large, the BSs cannot obtain the load information on time. The standby mode loses its advantage while it consumes more energy, resulting in lower EE compared to deep sleep. As expected, a short wake up interval is preferred in the high mobility scenario.

As another option, the system information can also be obtained with auxiliary devices when a BS is in the sleep mode. In [8], a counting element is used to count the number of accumulated tasks during a sleep period, and a BS wakes up when the number of tasks exceeds a threshold. To address the problem of obtaining CSI from nearby UEs, a possible approach is to deploy a radio receiver at a BS to capture the pilot or uplink UE signals and forward the obtained CSI to the network controller. Then the network controller can activate the BS when necessary, for example, when a large number of users with good channel conditions are nearby.

Trade-off with Other Performance Metrics

The energy savings achieved by switching off BSs come at a price. From the perspective of users, some have to be handed over to another BS and receive a degraded QoS. In addition, the BSs in sleep mode may not be turned on in a timely way, resulting in an extra delay perceived by users. To characterize the trade-off between energy saving and user QoS, a common approach is to add QoS requirements into the constraints of problem formulation [4, 8]. Another approach is to consider the actual relationship between energy consumption and system performance. EE, defined as bits per Joule, is a common performance metric utilized to balance the trade-off between data rate and power consumption [5, 7, 10]. In [6], a cost function is defined to balance energy consumption and delay, which is defined as a weighted sum of energy consumption and delay.

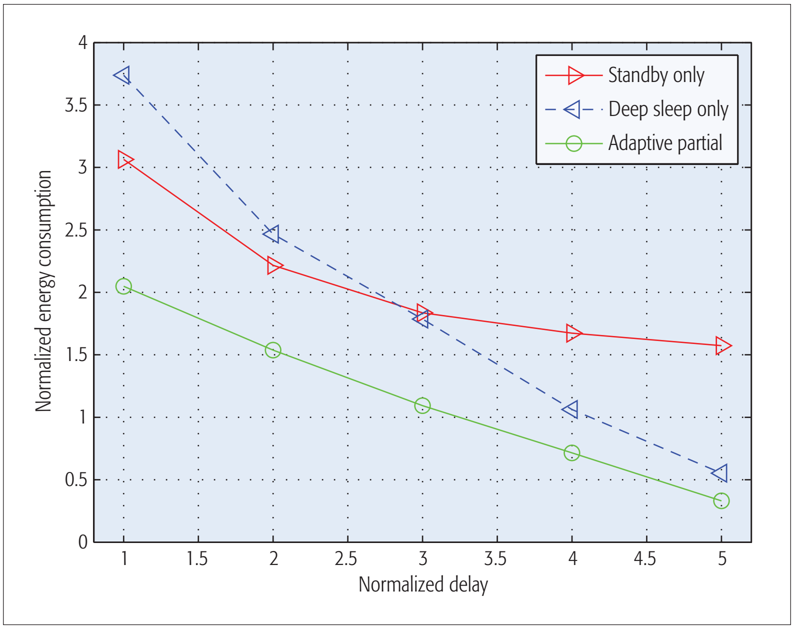

Figure 1. Energy-delay trade-off under different sleep modes with heterogeneous traffic requirements

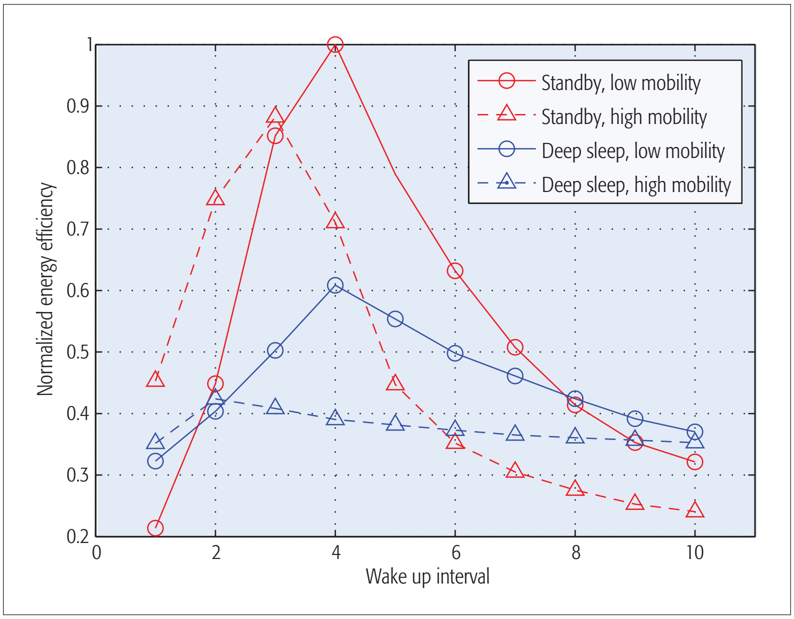

Figure 2. EE vs. wake up interval under different sleep modes and levels of mobility.

When a BS is turned off, the traffic load of neighboring BSs would increase, which potentially degrades the QoS of users served by these BSs. Thus, it is also necessary to consider the trade-off among different BSs. In a two-tier HetNet with a large number of SBSs turned off, the MBS may be overloaded, resulting in delay or even congestion. In [11], a constraint on the number of users served by the MBS is enforced to guarantee the QoS of its users.

Low-Complexity Algorithms

As the BS ON-OFF scheduling problem is combinatorial in nature, it is generally NP-hard and cannot be solved with standard techniques. It is more challenging when BS ON-OFF is jointly scheduled

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

	Objective	Constraint	Approach	Solution
[6]	Minimize weighted sum of energy and delay	N.A.	BS ON-OFF switching and user association.	Decompose into two subproblems; greedy ON-OFF scheduling.
[3]	Minimize energy over a period	BS traffic load	BS ON-OFF switching	Step-by-step process with evaluation on the network impact of turning ON-OFF.
[8]	Minimize average power	Average delay	Optimize hysteresis sleep time and wake up period	Simple bisection search based on exploiting the special structure
[4]	Minimize the ratio of turned off small cells	UE outage probability	BS ON-OFF scheduling and spectrum allocation	Solutions obtained when equalities hold for outage constraints
[10]	Maximize EE	Coverage probability and delay	BS sleep mode design	Maximize a quasi-convex lower bound; an iterative scheme for solution.
[11]	Minimize total BS power	Macrocell BS power	Active/sleeping schedule	Step-by-step process by evaluating the gain/loss of operation
[5]	Maximize EE	BS traffic load	BS ON-OFF switching and user association	Decompose into two subproblems; centralized and distributed solution
[13]	Minimize the number of active BBUs	Bandwidth; cell and user processing capability.	Time-wavelength allocation; user-fronthaul-BBU mapping	Load balancing for overloaded BBU; activate a BBU if still overloaded
[7]	Maximize EE	Data rate, power, and precoding	BS and antenna switching; power allocation	Turn OFF the antenna/BS such that total power is minimized afterwards
[9]	Minimize on-grid power	Blocking probability	BS ON-OFF switching; resource allocation; renewable energy allocation	Two-stage dynamic programming. First: BS ON-OFF; second: resource allocation.
[12]	Minimize on-grid power	Outage probability and resource availability	BS ON-OFF switching and offloading strategy	SBSs with positive power saving gain are activated first
[14]	Minimize power cost	Battery storage; supply and consumption balance	Dynamic power usage of different sources	Formulate and solve a stochastic programming
[15]	Each operator minimizes its cost	Rewards and penalties of the game	Provide incentive for infrastructure sharing by formulating a game	Analyze the outcome of the game

TABLE 2. Problem formulations and solution algorithms.

for trade-offs among multiple performance metrics, as the formulated problem may be mixed integer programming with multiple sets of variables. To decouple BS ON-OFF switching with other tasks, it is natural to decompose the original problem into several subproblems and iteratively solve them [5, 6, 9]. With regard to BS ON-OFF switching, step-by-step greedy algorithms are developed [6, 3, 7, 11, 12]. By evaluating the system performance gains under different operations, the BS that obtains the largest gain is selected in each step. The optimal ON-OFF states can also be obtained by fully utilizing the special properties of the problem [4, 8]. In [4], using a comprehensive performance analysis, the optimal BS ON-OFF states are found to correspond to the case that equalities hold for the outage constraints.

Another possible approach for a low-complexity solution is to transform the original problem into a more solvable form. In [10], to deal with the non-convex objective function, a lower bound that is quasi-convex is derived and serves as the new objective function. The solution obtained by maximizing the lower bound is shown to be near-optimal. Designing distributed schemes is another way to achieve low-complexity solutions. In [9], a bidding game between users and BSs is formulated, and the initial outcome of the game determines a user association strategy. When user associations are determined, each BS makes its own decision regarding ON-OFF switching depending on whether it is profitable to serve these users. As a result, BSs with low traffic loads would choose to turn off. The system EE can thus be improved. Table 2 provides a summary of different problem formulations and solution algorithms presented in the literature.

RECENT ADVANCES IN EMERGING WIRELESS NETWORKS

RENEWABLE ENERGY EMPOWERED NETWORK

The use of renewable energy can minimize on-grid energy consumption as well as reduce operating expenditures. This can be realized by equipping BSs with energy harvesting (EH) devices that transform natural energy (e.g., solar and wind energy) into electricity to power the BSs. However, as the generation of renewable energy depends on uncontrollable environmental conditions, the supply of renewable energy could be highly unstable. In addition, it is very likely that such a supply does not match the fluctuating traffic demands in both temporal and spatial domains. Thus, renewable energy should be properly stored and allocated for efficient utilization.

In [9], with statistical information on both

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

renewable energy and traffic patterns, renewable energy allocation, BS ON-OFF scheduling, and resource allocation are jointly considered to minimize the on-grid energy consumption, subject to the average blocking probability of users. The original problem is transformed into an unconstrained problem, and a two-stage dynamic programming algorithm with low complexity is proposed to obtain a near-optimal solution. The HetNet case is considered in [12], where SBSs are equipped with EH devices. The switching mechanism of the SBSs and traffic loading from the MBS are considered to minimize the on-grid power usage while satisfying constraints of outage probability and resource availability. Based on theoretical analysis on outage probability, a two-stage SBS activation scheme is proposed. In the first stage, the energy saving gain obtained by offloading MBS traffic to each SBS is obtained. In the second stage, SBSs with positive energy saving gain are activated first.

Massive MIMO HetNet

A massive multiple-input multiple-output (MIMO) HetNet employs a large number of antennas at the MBS [5]. The channel estimation overhead is a major concern due to the large dimensions of the channel matrix. When the traffic load of the MBS is increased, more symbols must be used as pilots in each frame. As a result, a smaller proportion of symbols can be used for data transmission, and the average throughput of users is reduced. Thus, in a massive MIMO HetNet, the trade-off between the traffic load of the MBS and that of the SBSs needs to be considered in the switching mechanism design. In [5], joint BS ON-OFF scheduling and user association are gauged in order to maximize the EE of a massive MIMO HetNet, subject to traffic load constraints of all BSs. Due to the convexity and a special property, the optimal user association strategy under a given BS set can be achieved using a series of Lagrangian dual methods. Then ON-OFF states can be optimized with a subgradient method using the optimal Lagrangian multipliers derived in the user association subproblem. A distributed scheme based on user bidding is also proposed, which has been discussed previously.

Cloud Radio Access Network

In C-RAN, the functionality of a BS is separated: data processing is performed by centralized BBUs and wireless signal transmission/reception is performed by geographically distributed RRUs. The ON-OFF switching of RRUs is relatively simple, and some existing methods based on the traditional BS model can be directly applied. However, since the RRUs serve only as transmitters/receivers with low power consumption, the energy saving gained from turning off RRUs alone would be limited. While it is highly appealing to investigate the ON-OFF switching for BBUs, turning off a BBU impacts all its serving RRUs and users connected to the RRUs. Hence, the ON-OFF states of BBUs are coupled with BBU-RRU and RRU-user mappings, resulting in a challenging scheduling problem. Such a problem is referred to as virtual BS formulation in [13], and a joint time-wavelength allocation for optical fronthaul, user-RRU-BBU mapping, and BBU ON-OFF scheduling scheme is proposed to minimize the number of active BBUs. The idea of the solution algorithm is to locate the overload BBUs and then apply load balancing. If any BBU is still overloaded, a new BBU has to be activated. This process can guarantee that all constraints are satisfied, and the number of active BBUs is kept as small as possible.

Cooperative Communication

CoMP transmission/reception, which allows a user to be served by multiple cooperating BSs, is an effective approach to enhance spectral efficiency and link reliability, but at the price of increased overhead. Within this context, detailed physical layer signal analysis is required to study the impact of BS ON-OFF switching. In [7], power allocation and ON-OFF switching for both the BS and antenna are considered to maximize the EE of a CoMP system. A low-complexity iterative greedy approach is proposed, in which the idea is to switch the antenna/BS that brings the largest reduction of power consumption until any constraint is violated.

Smart Grid Empowered Wireless System

The smart grid is a new paradigm of power systems that enables highly efficient use of energy, especially renewable energy. Thus, using the smart grid to power wireless systems is a promising approach to reduce both the operation costs and greenhouse gas emissions. For example, the BS of a wireless system can be powered by a combination of a renewable power source and an electrical grid. However, this approach is challenged by a set uncertainties, including renewable power generation, power price, and wireless traffic load. In [14], an adaptive demand-side power management scheme is proposed to make intelligent decisions about the power usage between renewable power and on-grid power. Such a problem is formulated as a stochastic programming problem with the objective of minimizing the cost of power consumption, and the optimal solution is derived by solving an equivalent linear programming problem.

The smart grid is a new paradigm of power systems that enables highly efficient use of energy, especially renewable energy. Thus, using the smart grid to power wireless systems is a promising approach to reduce both the operation costs and greenhouse gas emissions.

Infrastructure Sharing among Different Operators

While most existing works consider BS ON-OFF switching from the perspective of a single wireless operator, infrastructure sharing among different operators has the potential to achieve better system performance. As the BSs of multiple operators coexist in a cell, an operator can switch off its BS and migrate its traffic to the active BSs of another operator covering the same area. To enable such a process, it is necessary to design an incentive mechanism to motivate the cooperation of operators. In [15], a distributed game of multiple operators is formulated to provide incentive for infrastructure sharing, and each operator can make a rational decision on its switching strategy.

Challenges and Open Problems

5G Application Scenarios

D2D Communications: With the help of D2D communications, a BS can be turned off even if this leads to coverage holes, resulting in more energy saving. However, this comes at a price of increased energy consumption by UEs. Thus, it

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

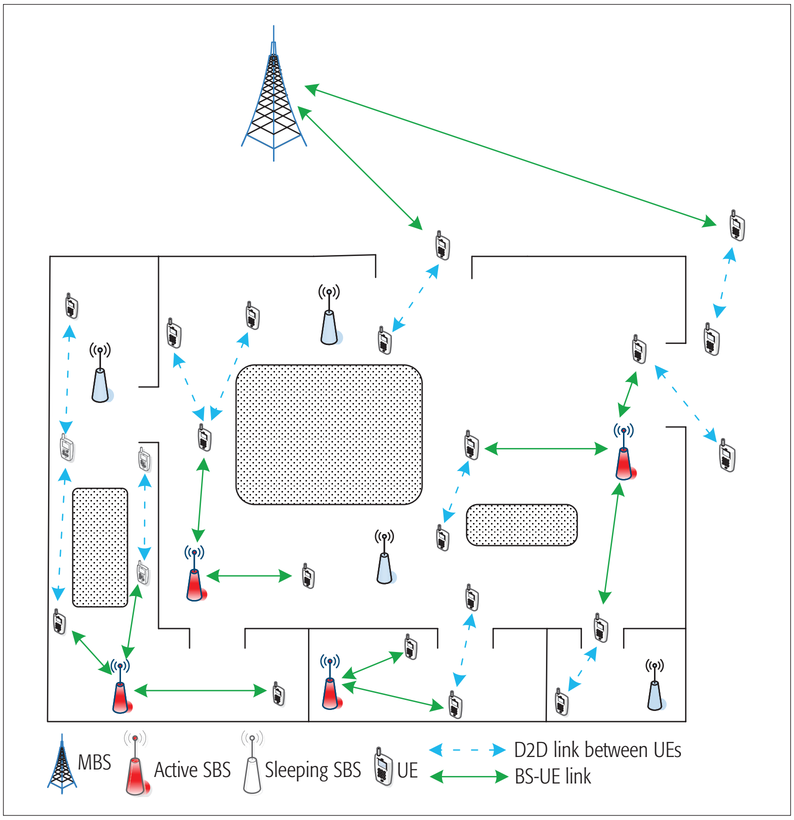

FIGURE 3. An example of BS ON-OFF scheduling in a D2D assisted mmWave network.

is necessary to motivate the participation of UEs and efficiently use their limited power storage. BS ON-OFF switching should be jointly considered with other design issues, such as D2D path selection and resource allocation. For example, suppose UE 1 relays the signal of UE 2 to a BS. To avoid congestion, the data rate between the BS and UE 1 should be no less than the data rate between UE 1 and UE 2 plus the actual data rate of UE 1. With such constraints, BS ON-OFF switching depends on the availability of UEs, link scheduling, and resource allocation. All these factors should be considered in future research.

MmWave Networks: MmWave communications is also a key technique for 5G wireless. Due to the large propagation loss and vulnerability to blockage, the signal transmission in an mmWave network can only rely on line-of-sight transmission or reflection. As a result, when a BS is turned off, its coverage may not easily be compensated by other BSs. A possible approach to deal with this challenge is to optimize the deployment pattern of BSs. With proper setting of BS locations and well designed beamforming schemes, the coverage holes caused by turning off BSs can be minimized. In the case where direction of arrival (DoA) estimation is feasible, the instantaneous positions of UEs can be obtained by each BS. Then a BS can know whether a UE can be served by another BS based on the environment layout, and make decisions regarding its ON-OFF state.

D2D communication is an effective means to combat blockage of signals in an mmWave network. As shown in Fig. 3, D2D communication can significantly enhance the coverage of both indoor and outdoor UEs. Apart from the issues mentioned before, some inherent challenges of mmWave networks need to be taken into account, such as neighbor discovery and efficient beamforming. Moreover, the UEs may not be able to perform "pseudo-wired" beamforming, and thus their transmission signals may still have a large beamwidth. Interference management is necessary to guarantee good link quality.

Heterogeneous Networks with Wireless Backhaul: Wireless backhaul (WB) between an MBS and SBSs in a HetNet has gained growing attention due to its easy implementation and relatively low cost. As an inherent constraint, the data rate of WB between the MBS and an SBS should be no less than the aggregated data rate of all small cell UEs (SUEs) served by the SBS. Hence, without proper configuration, WB may become a bottleneck that limits the system performance. When switching off an SBS, the aggregated data rate requirement of neighboring SBSs would be increased, putting pressure on their WBs. Thus, guaranteeing the data rates of WBs is a key concern when SBS ON-OFF scheduling is enabled. From the MBS perspective, the WB can be regarded as a macrocell UE (MUE) to be served. Hence, the data rates of WBs are also related to the traffic load and scheduling strategy of the MBS, and this is an important factor for system design. The network architecture of a HetNet is shown in the first case of Fig. 4, and it can seen that the wireless backhaul also needs to transmit the information exchange between the MBS and SBSs. Thus, guaranteeing the effectiveness and timeliness of scheduling is another design factor for a HetNet with wireless backhaul.

As massive MIMO is a key technique for 5G, we consider a massive MIMO HetNet as an example, in which the MBS is equipped with a large number of antennas, and the SBSs can be turned off when their load is low. The WBs and MUEs are put into several beamforming groups, and the channel estimation overhead is proportional to the number of active WBs and MUEs. When an SBS is turned off, the UEs originally served by the SBS may be handed over to the MBS or neighboring SBSs depending on the CSI. If the UEs are handed over to the MBS, the average throughput of the active WB and MUE would decrease, as analyzed in [9]. If the UEs are handed over to neighboring SBSs, the WBs of these SBSs need to be allocated with more spectrum resources to maintain the average throughput of UEs served by these SBSs. All these trade-offs need to be fully balanced for the design of an energy-efficient massive MIMO HetNet.

Mobile Edge Computing: The mobile edge computing (MEC) architecture allows content providers to deploy MEC servers at cellular BSs so that the applications and task processing are performed closer to the cellular users. In a HetNet with SBSs equipped with MEC servers, as shown in the second case of Fig. 4, the storage and computation capability of MEC stations are limited compared to the cloud server. Thus, optimizing the ON-OFF switching should consider not only the traffic load, but also the property of each task. For example, a computational intensive task should be directly processed at the cloud server and the nearby SBS should be turned off if it has no other ongoing task. In addition, the processing

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

delay is impacted by both the processing rate of the MEC server and transmission data rate to an SBS. This brings another factor to be considered in user association, which in turn impacts the BS ON-OFF schedule.

Wireless Caching Stations: In a small cell AP with caching capability, the frequently requested data-intensive tasks can be stored at caching stations in advance, resulting in high QoS for users. As the popular contents vary over time and location, it would be desirable to turn off the underutilized caching stations for energy saving. Different from cellular BSs, ON-OFF switching of a caching station impacts all the tasks stored in the station. Thus, a thorough consideration is required. The system performance can be further enhanced if the demand for content at each station is predictable, which can be implemented using statistical information or pattern recognition techniques. As shown in the third case of Fig. 4, the caching stations can gather the user preference information and forward it to the server in the core network. After processing such information with machine-learning-based approaches, the updated caching contents are sent back to the caching stations. The information analysis of user preference can also be executed locally to reduce latency if such capability is available for the caching stations.

M2M in IoT: Machine-to-machine (M2M) communication is expected to be widely applied in the paradigm of the Internet of Things (IoT) due to the massive data generated by various types of machines and devices. Similar to D2D, the direct transmission of M2M creates opportunities for APs or BSs to be turned off, as a machine connected to an AP can relay the signals of machines that are out of coverage. However, this requires additional functionality in the relaying machine/device, while most devices in the context of IoT are simple, low-power, low-cost devices that act as sensors and are not always switched on. Hence, the feasibility of exploiting M2M for energy saving depends on the availability of devices, and efficient coordination with low overhead is required.

Although the machines/devices in IoT are only required to be connected to the APs occasionally, the upload of the gathered data and the download of updated settings must be carried out with a certain frequency due to the storage limits of the devices and the required timeliness of the data. This requires the APs to be turned on in a proper manner to balance the trade-off between service provision and energy saving. In the presence of multiple machines with different data sizes and requirements for update frequency, the task schedule can be jointly considered with AP ON-OFF switching to minimize the long-term energy consumption as well as guarantee the QoS.

Operation in Unlicensed Bands: When a 5G wireless network operates on unlicensed bands, the coexistence with other systems that already use the spectrum band is a fundamental challenge. Take WiFi as an example. Due to the differences in medium access control (MAC) protocols, WiFi may get stuck in constant backoff states if the 5G network transmits in its normal pattern. For BS ON-OFF switching, a BS may be switched off during certain periods, not only for the purpose of energy saving, but also for QoS provisioning of WiFi users. When a BS is turned off, the UEs originally served by the BS may be handed over to the WiFi system for better QoS. Consequently, a trade-off between 5G UEs and existing system UEs is an interesting problem that can be addressed in future research.

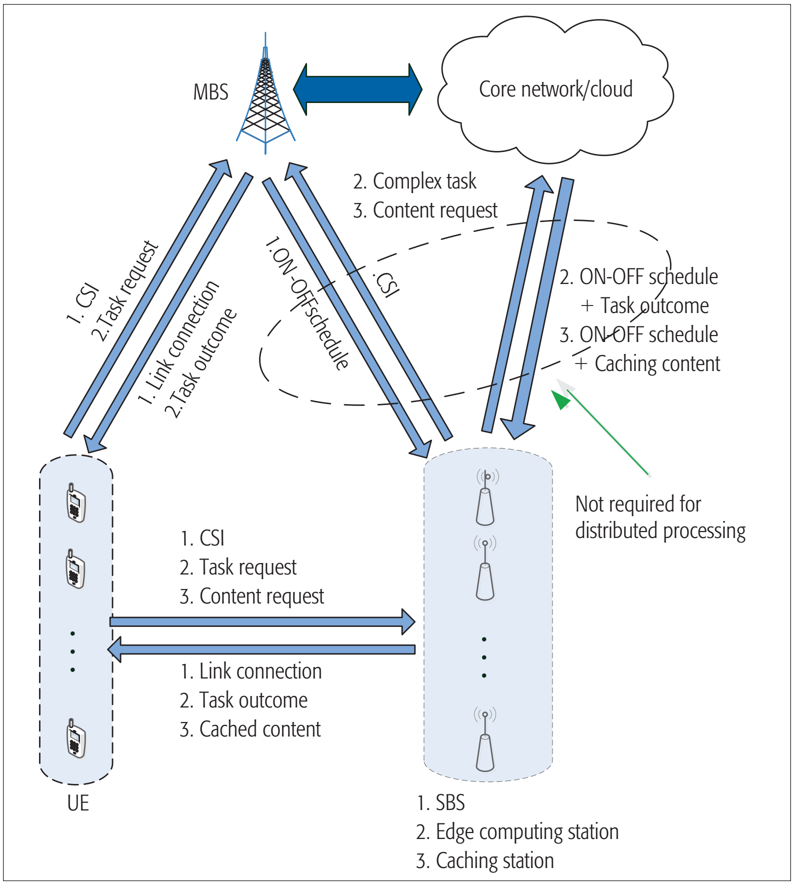

FIGURE 4. Network architecture and information exchange in three heterogeneous application scenarios.

LTE unlicensed (LTE-U) is regarded as a promising application scenario in 5G systems. In an LTE-U system, UEs are allocated with orthogonal time-frequency resource blocks (RBs). Then the ON-OFF schedule should be jointly considered with the user association and resource allocation of each BS, since these factors determine the aggregated interference level of both LTE and WiFi users. Thus, efficient approaches are required to decouple BS ON-OFF switching with other system configurations. Moreover, since LTE adopts centralized control for service provision with information exchange through an X2 interface, the added BS ON-OFF schedule would increase the processing overhead and complexity, making it necessary to design decentralized or computationally efficient schemes to enhance feasibility.

ADDITIONAL TECHNICAL CHALLENGES

Scalability: With the expected massive deployment of BSs in 5G networks, the control overhead could be overwhelming. In particular, if software-defined networking architecture is applied, the network control would be performed in a centralized pattern, resulting in prohibitive complexity. To address this, a wireless network may

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.

The detailed technical aspects needed to implement this approach still require further study. Moreover, the learning-based techniques have the potential to be applied to other approaches, and therefore, to be explored in future research.

be partitioned into multiple parts to guarantee the feasibility of scheduling. As for BS ON-OFF switching, we can divide BSs into different clusters, each with a controller. At the lower level, intra-cluster scheduling can be performed with reasonable overhead as long as the cluster size is relatively small. At the higher level, different controllers can cooperatively adjust their strategies by taking the inter-cluster impacts into consideration. The cluster size provides a trade-off between complexity and system performance, and clustering strategy would be an interesting topic for future research.

Application of Machine Learning Techniques: With the development of hardware, machine learning techniques become feasible for network control and can be used in ON-OFF switching to achieve better performance. As shown in Figs. 1 and 2, the trade-off between energy and delay, and the trade-off between EE and the wake up interval are observed. However, if we can predict the traffic pattern, each BS can wake up at proper time instances such that energy saving is achieved without causing additional delay. Then machine learning techniques can be applied to predict the user distribution, traffic load, and traffic type of each BS. Based on historical data and the movement of users, the approximate number of users arriving at each BS can be estimated, and the expected underutilized BSs can maintain sleep mode. However, the detailed technical aspects needed to implement this approach still require further study. Moreover, the learning-based techniques can also be applied to other aspects such as resource allocation and traffic offloading, and therefore should be explored in future research.

QoE-Aware Scheduling: As shown in Table 2, most existing works take into account QoS metrics in system models and problem formulations. However, the actual satisfaction level of a user is determined by multiple QoS factors, such as outage probability, delay, downloading rate, congestion probability, video type and format, and energy consumption of mobile devices. Besides, different users have different evaluations regarding all these aspects. Thus, it is necessary to introduce quality of experience (QoE) into the system design. Moreover, in terms of QoE of users and the energy cost, realistic economic models can be established (e.g., QoE-based user payment). The minimization of the cost of the wireless operator can be a new design objective.

Conclusion

In this article, we aim to identify the challenges of BS ON-OFF switching in 5G wireless networks and provide insights for potential solutions. We analyze the technical aspects of BS ON-OFF switching and present an overview of recent advances in different 5G wireless networks. We conclude this article with a discussion of open problems and outlook.

Acknowledgment

This work is supported in part by the National Science Foundation under Grant CNS-1320664, the Wireless Engineering Research and Education Center (WEREC) at Auburn University, the National Science Foundation for Distinguished Young Scholars of China with Grant number 61325004, and Major State Basic Research Development Program of China (973 Program) with Grant number 2013CB329006.

References

[1] Z. Niu, "TANGO: Traffic-Aware Network Planning and Green Operation," *IEEE Wireless Commun.*, vol. 18, no. 5, Oct. 2011, pp. 25–29.
[2] J. Andrews *et al.*, "What Will 5G Be?," *IEEE JSAC*, vol. 32, no. 6, June 2014, pp. 1065–82.
[3] E. Oh, K. Son and B. Krishnamachari, "Dynamic Base Station Switching-On/Off Strategies for Green Cellular Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, May 2013, pp. 2126–36.
[4] S. Zhang *et al.*, "How Many Small Cells Can Be Turned Off via Vertical Offloading under a Separation Architecture?," *IEEE Trans. Wireless Commun*., vol. 14, no. 10, Oct. 2015, pp. 5440–53.
[5] M. Feng, S. Mao, and T. Jiang, "BOOST: Base Station On-Off Switching Strategy for Energy Efficient Massive MIMO HetNets," *Proc. IEEE INFOCOM 2016*, San Francisco, CA, Apr. 2016, pp. 1395–1403.
[6] K. Son *et al.*, "Base Station Operation and User Association Mechanisms for Energy-Delay Trade-offs in Green Cellular Networks," *IEEE JSAC*, vol. 29, no. 8, Sept. 2011, pp. 1525–36.
[7] Q. Zhang *et al.*, "Energy Efficient Downlink Cooperative Transmission with BS and Antenna Switching Off," *IEEE Trans. Wireless Commun*., vol. 13, no. 9, Sept. 2014, pp. 5183–95.
[8] X. Guo *et al.*, "Delay-Constrained Energy-Optimal Base Station Sleeping Control," *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1073–85.
[9] J. Gong *et al.*, "Base Station Sleeping and Resource Allocation in Renewable Energy Powered Cellular Networks," *IEEE Trans. Commun*., vol. 62, no. 11, Nov. 2014, pp. 3801–13.
[10] C. Liu, B. Natarajan, and H. Xia, "Small Cell Base Station Sleep Strategies for Energy Efficiency," *IEEE Trans. Vehic. Tech*., vol. 65, no. 3, Mar. 2016, pp. 1652–61.
[11] S. Cai *et al*., "Green 5G Heterogeneous Networks through Dynamic Small-Cell Operation," *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1103–15.
[12] S. Zhang *et al*., "Energy-Aware Traffic Offloading for Green Heterogeneous Networks," *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1116–29.
[13] X. Wang *et al*., "Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN," *IEEE JSAC*, vol. 34, no. 5, May 2016, pp. 1130–39.
[14] D. Niyato, X. Lu, and P. Wang, "Adaptive Power Management for Wireless Base Stations in a Smart Grid Environment," *IEEE Wireless Commun.*, vol. 19, no. 6, Dec. 2012, pp. 44–51.
[15] A. Antonopoulos *et al*., "Energy-Efficient Infrastructure Sharing in Multi-Operator Mobile Networks," *IEEE Commun. Mag*., vol. 53, no. 5, May 2015, pp. 242–49.

Biographies

Mingjie Feng [S'15] received his B.E. and M.E. degrees from Huazhong University of Science and Technology, Wuhan, P. R. China in 2010 and 2013, respectively, both in electrical engineering. He is currently a Ph.D. student in the Department of Electrical and Computer Engineering, Auburn University, Alabama. His research interests include cognitive radio networks, femtocell networks, massive MIMO, and full-duplex communication. He is a recipient of a Woltosz Fellowship at Auburn University.

Shiwen Mao [S'99, M'04,SM'09] received his Ph.D. in electrical and computer engineering from Polytechnic University, Brooklyn, New York, in 2004. He is the Samuel Ginn Distinguished Professor and director of the Wireless Engineering Research and Education Center at Auburn University. His research interests include wireless networks and multimedia communications. He is a Distinguished Lecturer of the IEEE Vehicular Technology Society, 2014–2018. He is on the Editorial Boards of *IEEE Transactions on Multimedia*, the *IEEE Internet of Things Journal*, *IEEE Multimedia*, and *ACM GetMobile*, among others.

Tao Jiang [M'06, SM'10] is a Distinguished Professor in the Wuhan National Laboratory for Optoelectronics, School of Electronics Information and Communications, Huazhong University of Science and Technology. He received his Ph.D. degree in information and communication engineering from Huazhong University of Science and Technology in April 2004. He is/was an Associate Editor of *IEEE Transactions on Signal Processing*, *IEEE Communications Surveys & Tutorials*, *IEEE Transactions on Vehicular Technology*, the *IEEE Internet of Things Journal*, and *China Communications*, among others.

Authorized licensed use limited to: Auburn University. Downloaded on April 13,2020 at 02:03:18 UTC from IEEE Xplore. Restrictions apply.