SoftAir: A software defined networking architecture for 5G wireless systems

Ian F. Akyildiz, Pu Wang, Shih-Chun Lin

ABSTRACT

One of the main building blocks and major challenges for 5G cellular systems is the design of flexible network architectures which can be realized by the software defined networking paradigm. Existing commercial cellular systems rely on closed and inflexible hardware-based architectures both at the radio frontend and in the core network. These problems significantly delay the adoption and deployment of new standards, impose significant challenges in implementing and innovation of new techniques to maximize the network capacity and accordingly the coverage, and prevent provisioning of truly-differentiated services which are able to adapt to growing and uneven and highly variable traffic patterns. In this paper, a new software-defined architecture, called SoftAir, for next generation (5G) wireless systems, is introduced. Specifically, the novel ideas of network function cloudification and network virtualization are exploited to provide a scalable, flexible and resilient network architecture. Moreover, the essential enabling technologies to support and manage the proposed architecture are discussed in details, including fine-grained base station decomposition, seamless incorporation of OpenFlow, mobility-aware control traffic balancing, resource-efficient network virtualization, and distributed and collaborative traffic classification. Furthermore, the major benefits of SoftAir architecture with its enabling technologies are showcased by introducing software-defined traffic engineering solutions. The challenging issues for realizing SoftAir are also discussed in details.

1. Introduction

Existing commercial wireless networks are inherently hardware-based and rely on closed and inflexible architectural designs. Such inflexible hardware-based architectures typically lead to a 10-year cycle for a new generation of wireless networks to be standardized and deployed, impose significant challenges into adopting new wireless networking technologies to maximize the network capacity and coverage, and prevent the provision of truly-differentiated services able to adapt to increasingly growing, uneven, and highly variable traffic patterns. In particular, for 5G cellular system requirements, the ultra high capacity should have 1000-fold capacity/km² compared to LTE, the user-plane latency should be less than 1 ms over the radio access network, and the ultra high data rates should provide 100-fold increase in user-experienced throughput (targeting 1 Gbps experienced user throughput everywhere). The challenges faced by the current network architectures cannot be solved without a radical paradigm shift in the design of next-generation wireless networks. Hence, in this paper, we propose the utilization of Software-Defined Networking (SDN) concept for next generation (5G) wireless networks, introduce a new architecture for wireless software-defined networks, called SoftAir, and
present solutions and challenges for related research in this domain.

SDN has been recently introduced primarily for data center networks and for the next-generation Internet [6,25]. The main ideas are (i) to separate the data plane from the control plane, and (ii) to introduce novel network control functionalities based on an abstract representation of the network. In current instantiations of this idea, these are realized by (i) removing control decisions from the hardware, e.g., switches, (ii) enabling the hardware to be programmable through an open and standardized interface, e.g., OpenFlow [37], and (iii) using a network controller to define the behavior and operation of the network forwarding infrastructure. SDN makes it easier to introduce and deploy new applications and services than the classical hardware-dependent standards. So far, the majority of SDN developments has concentrated on wired networks [6]. In this paper, we propose the utilization of SDNs for next generation (5G) wireless networks and present a new architecture for wireless SDNs, called SoftAir, and the solutions and challenges for related research in this domain. In our proposed SoftAir architecture, the control plane consists of network management and optimization tools and is implemented on the network servers. The data plane consists of software-defined base stations (SD-BSs) in the radio access network (RAN) and software-defined switches (SD-switches) in the cellular core network. Their control logic, e.g., physical/MAC/network functions, are implemented in software on general purpose computers and remote data centers.

Our proposed SoftAir architecture offers five core properties: (i) programmability, i.e., SDN nodes (e.g., SD-BSs and SD-switches) can be reprogrammed on-the-fly by dynamically associating with different network resources and networking algorithms; (ii) cooperativeness, i.e., SDN nodes can be implemented and aggregated at data centers for joint control and optimization to enhance the global network performance; (iii) virtualizability, i.e., multiple virtual wireless networks can be created on a single SoftAir, each of which operates under its own independent network protocols with network resources allocated based on demand; (iv) openness, i.e., data plane elements (i.e., BSs and switches), regardless of the underlying forwarding technologies and vendors, have unified data/control interfaces, e.g., CPRI and OpenFlow [20,38], thus significantly simplifying the data plane monitoring and management; and (v) visibility, i.e., centralized controllers have a global view of the network status collected from BSs and switches.

The above five properties provide functionalities that are essential to enable 5G wireless communication networks to possess the following promising features:

- **Evolvability and adaptiveness**: Because of the inherent separation between data plane and control plane, in SoftAir, both hardware forwarding infrastructure and software networking algorithms can easily, continuously and independently be upgraded quickly, which allows to timely adopt emerging radio technologies (e.g., millimeter wave (mm Wave), full-duplex, massive MIMO, and TeraHertz [5,10,29]) in the hardware infrastructure, while deploying novel traffic engineering, network management, and network optimization solutions at controllers. Moreover, the programmable data plane allows controllers to dynamically allocate network resources and adopt new networking solutions, according to the highly variable traffic patterns, unexpected network failures, and diverse quality of service (QoS) requirements of traffic flows.

- **Infrastructure-as-a-service**: Emerging network services, such as machine-to-machine communications, smart grid applications, mobile virtual network operators (MVNOs), and over-the-top content services (e.g., Netflix video streaming), require highly differentiated networking capabilities to be integrated and deployed over the same network infrastructure. The network virtualizability of SoftAir allows the wireless hardware infrastructure to be offered as a service rather than as a physical asset. Specifically, in SoftAir, the service providers are provided with the ability to control, optimize, and customize the underlying infrastructure without owning it and without interfering with the operations and performance of other service providers, thus leading to more cost-efficient operations and enhanced QoS. Moreover, thanks to the programmable data plane, the network resources, e.g., spectrum, can be dynamically shared among the service providers, e.g., MVNOs.

- **Maximal spectral efficiency**: In SoftAir, SD-BSs (e.g., both macro and small-cell base stations) can be implemented and aggregated at a server or a data center. There, they can easily share control information, mobile data and channel state information (CSI) associated with different active users in the system. Therefore, with SoftAir, it is much easier to implement algorithms to mitigate or exploit inter-cell interference towards universal frequency reuse, i.e., achieving frequency reuse factor 1 in the entire network.

- **Convergence of heterogeneous networks**: The constant influx of new and amended wireless standards (e.g., small cells, WiFi, WiMAX, LTE, LTE-A, and super-WiFi) has created a rich but chaotic wireless environment in which multiple standards are competing and coexisting. Moreover, the fundamentally different features between wireless RAN and wired core network prevent the deployment of simple and unified network planning and control. By utilizing open and technology-independent interfaces, SoftAir can enable smooth transition and unified management among different wireless standards and between wireless RANs and wired core networks.

- **Low carbon footprints**: The high energy-efficiency of SoftAir relies on its software-defined data plane, where the processing capacity of the SD-BSs can be dynamically scaled according to the uneven network traffic patterns in such a way that the number of idle BSs, which consume almost the same amount of energy as the active ones, is reduced. Moreover, with the centralized implementation of SD-BSs at data centers, the number of physical BS sites can be significantly reduced. Thus, air conditioning and other onsite power-hungry equipment can be considerably reduced.

The rest of the paper is organized as follows. Section II provides the related work. Section III introduces the architecture design of SoftAir. Section IV summarizes the essential management tools for SoftAir. Section V presents the
software-defined traffic engineering solutions enabled by SoftAir. Section VI concludes this paper.

2. Related work

In literature, the software-defined architectures are well-studied in wired networks. For example, in data center networks and campus Local Area Networks (LANs) [12,37,54], these architectures mainly support centralized and adaptive management of flow tables at switches and routers. Furthermore, considering wired network virtualization, cloud computing and computer virtualization have maintained strong foothold for the past few years. In particular, the virtualization of routers and switches has occurred with Virtual Private Networks (VPNs) over service provider networks common in Wide Area Networks (WANs) and Metropolitan Area Network (MANs) and with virtual LANs in enterprise networks. It is achieved by logically partitioning a physical network into virtual networks that share the physical routers/switches/crossconnects, physical links, and bandwidth on each link. The utilization of the physical resources needs to be carefully managed to maintain the QoS and security needs of the users of each virtual network. Therefore, SDN’s effectiveness and great potential for 5G data networking come with many new technical challenges, which need to be addressed by the new research advances.

Recently, few SDN architectures [9,14,18,22,31,42,48,53] are exploited in wireless networks. OpenRoads [53] is mainly targeted at WiFi networks with little support for cellular networks. OpenRadio [9] proposes a novel programmable wireless data plane that provides modular programming capability for the entire wireless stack. However, OpenRadio does not provide any network controller that take advantage of such programmable data plane. CellSDN [31] aims to achieve a centralized control plane for cellular core networks. ADRENAline [14] provides an industrial solution of SD-CN for 5G cellular systems with optical OFDM. However, these architectures neither consider the scalability issue of SD-CN nor the incorporation of CN with RAN. On the other hand, SoftTRAN [22] attempts to restructure the control plane of RAN in a software-defined manner. An emerging distributed RANs, called Cloud-RAN [18], proposes SD-RAN architecture that connects SD-RANs to virtual BS pool through fibers and provides centralized control solution upon the BS pool. However, its coarse-grained BS decoupling limits the scalability and evolvability of such distributed RANs due to the excessive (and redundant) I-Q transmissions. DOCOMO [42] and SK Telecom [48] also provide their own industrial SD-RAN solution for 5G systems, respectively. Similar to Cloud-RAN, these solutions adopt coarse-grained decoupling and thus, bring bottlenecks to fronthaul fiber networks. Different from our proposed SoftAir architecture, all of the above architectures lack a coherent framework that integrates cellular core network and radio access network in a software-defined manner. What is more important, these solutions lack underlying concrete algorithms that leverage the promising properties of software-defined wireless network architecture. In addition, network virtualization, serving as a key building block to enable the sharing of (mobile) carrier networks for enhanced spatio-temporal spectrum utilization [19], is omitted by all existing software-defined architectures.

3. SoftAir architecture design

As shown in Fig. 1, the architecture of SoftAir consists of a data plane and a control plane. The data plane is an open, programmable, and virtualizable network forwarding infrastructure, which consists of software-defined radio access network (SD-RAN) and software-defined core network (SD-CN). The SD-RAN consists of a set of SD-BSs, while the SD-CN is composed of a collection of SD-switches. The control plane mainly consists of two components: (1) network management tools, and (2) customized applications of service providers or virtual network operators. In the following, we present the scalable SoftAir architecture in detail, explain the network virtualization, and introduce three essential management tools, namely mobility-aware control traffic balancing, resource-efficient network virtualization, and distributed and collaborative traffic classifier.

3.1. Scalable network function cloudification/network function virtualization (NFV)

The network function cloudification, also known as network function virtualization (NFV), decouples network functions from the underlying hardware and centralizing/clouding them at network servers. The function cloudification makes the network architecture highly flexible as the network can be reconfigured quickly and adaptively. Indeed, the main advantage of the wired OpenFlow-based SDNs relies on the network-layer function cloudification, which decouples the routing function from the hardware switches and centralizes it at a network controller through an open network interface, namely OpenFlow. Despite its great advantages, such cloud-based network architecture also imposes challenges on the network scalability.

(1) Scalable SD-CN design: To provide the cellular core network with high flexibility, our proposed SoftAir adopts SD-switches to form the SD-CNs as shown in Figs. 1 and 2. By SD-CNs, the customized SDN applications, e.g., mobility management, QoS routing, and billing policies; as well as global management tools, e.g., traffic classification and network virtualization, can be designed, deployed and updated on the network controller to fit the specific and ever-changing needs. The practicality of SD-CN can be envisioned because of the successful field deployment of SDNs, e.g., B4 from Google [25], SWAN from Microsoft [23], and ADMCF from Huawei [2]. The recent deployments of software-defined WANs, e.g., B4 from Google [25], has successfully demonstrated the promising performance of SDNs by boosting the average link utilization from 30–40% to over 70%.

The scalability of SD-CN can be greatly improved by leveraging high-performance controllers and optimized network management schemes. For example, with the current SDN technology, one single controller can achieve 12 million requests per second processing speed for the control messages between the switches and the controller [6]. Such high processing capacity has been further enhanced by the adoption of controller clusters and advances in multi-threading technologies [6]. More importantly, our recent research on the scalability of SDNs [32] has shown that even in large scale SDNs with practical in-band control channels, the control message forwarding delay between the controllers and
the switches can be minimized by our proposed control traffic balancing scheme, which employs emerging parallel optimization theories, e.g., Alternating Direction Method of Multipliers (ADMM) [11], to achieve fast and reliable control message forwarding.

To further increase the scalability of SD-CN, we can adopt a mobility-aware and proactive control traffic balancing scheme, which minimizes the control message forwarding delay by taking into account the unique mobile feature of SD-RANs (formally defined in the following subsection), where the control traffic oriented from SD-RAN are dynamic but follows certain spatial and temporal patterns. Accordingly, the mobility-aware control traffic balancing solution can employs a hybrid approach. On the one hand, the average control message forwarding delay will be minimized through semistatic control message forwarding rules by incorporating the historical mobility pattern of the mobile users. On the other hand, real-time control traffic forwarding rules, empowered by the fast-converging optimization solutions [32], will be enforced to address the dramatic and unexpected control traffic changes in both SD-CN and SD-RAN.

(2) Scalable SD-RAN design: to further enhance the network flexibility, we simultaneously realize the physical-, MAC-, and network-layer function cloudification for RAN, thus forming the SD-RAN. As shown in Fig. 1, the proposed SD-RAN follows a distributed RAN architecture. Here, each SD-BS is split into hardware-only radio head(s) and software-implemented baseband units - these two components are not necessarily co-located. In particular Remote Radio Heads (RRHs) are connected to the baseband units on baseband servers (BBS) through fronthaul network (fiber or microwave) using standardized interfaces, such as CPRI.

![Fig. 1. Overall architecture of SoftAir.](image1)

![Fig. 2. Function cloudification of SoftAir.](image2)
(Common Public Radio Interface) or OBSAI (Open Base Station Architecture Initiative) interface. The baseband or PHY/MAC cloudification through distributed RAN architecture has two main features. First, its decoupling of baseband processing from the radio hardware allows the independent evolution of radio technology at RRHs and baseband processing solutions at BBS. Second, centralizing baseband processing at data centers can facilitate network-wide cooperative processing among different base stations, while at the same time contribute to cost reduction. Thus, the distributed RAN architecture has recently received significant attention from both industry and academia. Cloud-RAN [18] is an example. The applicability and practicality of distributed RAN is enabled by the fast evolution of software-defined radio architectures. Significant efforts have been devoted to the standardization of open and technology-independent communication interfaces to connect RRHs and BBS. In particular, the standardized interface CPRI, jointly developed by Ericsson, Huawei, NEC, Alcatel Lucent and Nokia Siemens, can enable high-speed (up to 10 Gbit/s, low bit error rate \(10^{-12}\)), and long-distance (up to 40 miles) data exchange between RRHs and BBS, while providing the high-resolution synchronization.

Existing distributed RAN architectures such as Cloud-RAN mainly focus on the high-performance computing of baseband processing functions (mostly for physical layer operations) at remote servers or data centers. It faces two fundamental limitations. First, they cannot achieve scalable PHY/MAC-layer cloudification. Second, they do not support network-layer cloudification as the SD-CN. To this end, our proposed SD-RAN offers significantly enhanced scalability, evolvability, and cooperativeness through fine-grained base station decomposition.

Fine-grained base station decomposition: The existing distributed RAN adopts a coarse-grained function splitting between RRHs and BBS, where the entire baseband processing along with MAC layer operations are located at the BBS, while RRHs only implement the radio frontend. Through this approach, the digital I–Q samples must be transported between BBS and RRHs, which inevitably demands extremely high data rates on the transport network (i.e., fronthaul network). More specifically, the data rate requirement \(R_{IQ}\) for transporting I–Q samples of a waveform depends on the bandwidth \(W\), the oversampling ratio \(r > 1\), the quantification resolution \(Q_s\) of the analog signal, the duplex mode \(D_m\), and the number of antennas \(A\) at the radio head \((R_{IQ}(B,A,D_m) = 2 \times W \times r \times Q_s \times D_m \times A)^1\) For example, to transport a 20 MHz LTE waveform over CPRI between BBS and a half-duplex RRH with eight antennae a transmission rate of 7.36 Gbps is required. Such linear scaling of data rate requirement with the radio bandwidth, the number of antennas, and the duplex mode contradicts with the current trend of wireless technologies moving towards massive MIMO, full-duplex transceivers, mm-waves, and Terahertz band, thus significantly degrading the scalability and evolvability of the current distributed RANs.

To address this challenge, SoftAir adopts a new fine-grained base station decomposition architecture by leaving partial baseband processing at the RRH (e.g., modem), while implementing the remaining baseband functions, e.g., MIMO coding, source coding, and MAC, at the BBS, as shown in Fig. 2. Such function splitting is convenient because CPRI, which is not only defined for I–Q sample transport, can still be adopted without designing new interfaces and can lead to considerably reduced data rate requirements between BBS and RRHs. In addition, our decomposition still preserves sufficient flexibility offered by the distributed RAN architecture. More specifically, by this new decomposition method, the data rate requirement \(R_t\) of transporting the information bits conveyed by a waveform can be flexibly adjusted by moving the partial baseband functionalities to the BBS. In summary, with reduced data rate requirements, our SD-RAN offers excellent scalability, cooperative gain, and evolvability to next-generation wireless networks by allowing the aggregation of a large number of technology-evolving RRHs at BBS through the diverse, cost-efficient, CPRI-supported fronthaul network topologies, and over different fronthaul mediums.

Seamless incorporation of OpenFlow: Different from the existing distributed RAN [18], our proposed SD-RAN, as shown in Fig. 2, incorporates OpenFlow into the software-defined Base Station (SD-BS). In our SD-RAN design, we implement an OpenFlow interface for each SD-BS by utilizing Open vSwitch (OVS) [43], which is an OpenFlow-capable software switch that can easily be realized in BBS. With OVS, each SD-BS will be able to interpret, exchange, and respond to the OpenFlow protocol messages. Equipping SD-BSs with OpenFlow capabilities provides a unified interface to control and manage base stations with different wireless standards, thus leading to a multi-technology converged RAN that allows smooth transitions among different radio access technologies. For example, with OpenFlow-capable SD-BSs, seamless vertical mobility can be easily achieved, which allows mobile users to transparently roam among BSSs with different wireless standards, without experiencing disruptions in network services. This can be done by simply re-routing the traffic flows through the core network over the best paths to different BSs via the technology-independent OpenFlow interface enabled on both core network switches and BSs. Moreover, adopting the common OpenFlow interface on both SD-switches and SD-BSs promises the transparent interconnections between SD-CNs and SD-RANs and allows the unified management of the entire SoftAir.

3.2. Network virtualization capability

The network virtualization enables multiple isolated virtual networks, e.g., M2M, smart grid, over-the-top service provider, cellular provider, to share the same physical network infrastructure, as shown in Fig. 3. That means it focuses on slicing network resources for multiple virtual networks so that they can simultaneously share the same physical network architecture. More specifically, each virtual network can adopt its customized PHY/MAC/NET layer protocols, without interrupting the operations and performance of other virtual networks. These virtual networks can also be deployed on demand and dynamically allocated. Our proposed SoftAir with network virtualization capability enables a wide range of emerging applications, e.g. (1) allowing MVNOs to adopt different wireless standards (i.e., small

---

1 The factor of 2 captures both I-plane and Q-plane digital samples.
Fig. 3. Network virtualization of SoftAir.

cells, HetNets, LTE, WiMAX, or WiFi), (2) enabling active RAN sharing, which could allow the operators to save up to $60 Billion in a period of five years, through the significant reduction in equipment investments in low traffic areas [1], (3) promising virtualization-enabled QoS routing by simultaneously satisfying the strict end-to-end performance requirements (e.g., delay, jitter, and throughput) of different network services that generate fundamentally different traffic flows [44], and (4) accelerating technology innovation by allocating isolated wireless resources to deploy and test innovative technologies on the operational networks in large-scale real-life scenarios.

To realize these isolated virtual networks, SoftAir implements three functions as shown in Fig. 3: network hypervisor for high-level virtualization as well as wireless hypervisor and switch hypervisor for low-level virtualization.

- **The network hypervisor** is a high-level resource management framework, which adaptively allocates nonconflicting multi-dimensional network resources to service providers or virtual network operators.

- **The wireless hypervisor** is a low-level resource scheduler that enforces or executes the resource management policies determined by network hypervisor by employing a variety of wireless resource dimensioning schemes, e.g., OFDMA or wireless scheduling, so that isolation among virtual networks are guaranteed.

- **The switch hypervisor** focuses on bandwidth management in a single SD-switch. In particular, bandwidth provisioning on switches offers bandwidth assurance for the designated virtual network. In SoftAir, we adopt a well-known and open-source switch hypervisor, namely FlowVisor [47].

By utilizing the proposed wireless hypervisor at SD-BSs along with the Flowvisor at SD-switches, SoftAir will have the capability to enable the end-to-end network virtualization traversing both SD-RAN and SD-CN, thus realizing a truly multi-service converged network infrastructure.

3.3. SoftAir management tools

Cloud orchestration aims to automate the configuration, coordination and management of software and software interactions in the cloud environment. To support cloud orchestration in SoftAir, to enable the promising features and to maximize the overall performance of SoftAir, as shown in Fig. 2, three essential and general management tools need be developed: (1) mobility-aware control traffic balancing, (2) resource-efficient network virtualization, and (3) distributed and collaborative traffic classifier.

- **Mobility-aware control traffic balancing:** To promise on-line and adaptive traffic engineering in SoftAir, the control messages should be forwarded from SD-switches or SD-BSs to the controller(s) in a fast and robust manner. While out-band control channel is cost-prohibitive, the usage of in-band control channel imposes a great challenge into timely and reliable transmissions of control traffic. Moreover, due to mobile users, the unique mobile traffic features in SD-RANs should be examined for control traffic balancing. More specifically, considering users’ mobility, a nonlinear optimization framework is proposed with an objective to find the optimal control traffic forwarding paths for each SD-switch and SD-BS to the controller in such a way the average control traffic delay in the whole network is minimized.

- **Resource-efficient network virtualization:** Network virtualization capacity is essential to support Infrastructure-as-a-Service, thus enabling a wide range of emerging applications as mentioned above. Since wireless network resources are limited, resource-efficient wireless network virtualization solutions are highly desirable (Section 4.2). Specifically, the utilization-optimal network hypervisor is
favorable, which can maximize the global resource utilization, while guaranteeing the respective data rate requirements demanded by the virtual networks. Moreover, the throughput-efficient wireless hypervisor is demanding, which aims to achieve the isolation of multiple virtual networks at each SD-BS, while ensuring the efficient resource utilization for enhanced throughput.

• **Distributed and collaborative traffic classification:** The proposed SoftAir has the great capability to significantly improve the spectrum efficiency. However, without integrating advanced traffic engineering tools, improved spectrum efficiency does not lead to satisfied QoS for mobile users because of the highly bursty mobile data traffic emitted from diverse network applications with differentiated QoS demands [50]. To this end, we propose a distributed and collaborative traffic classifier at SD-BS, which collaborates with a global traffic learner at the network controller to achieve fast, fine-grained and accurate traffic classification (Section 4.3). The objective of this distributed traffic classifier is to identify the application, the QoS requirement, and the stochastic features associated with each traffic flow. With such traffic flow information, highly sophisticated and adaptive traffic engineering solutions can be adopted at both the BS and the network level (Section 6).

4. SoftAir management tools

4.1. Mobility-aware control traffic balancing

Rather than exploiting costly out-band control due to a separate control channel, the in-band mode is favored and adopted gradually in practical SDN implementation. In particular, for SoftAir in Fig. 4, each SD-switch or SD-BS (i.e., BBS) needs to send the control purpose traffic, such as the route setup requests for new flows and real-time network congestion status, to the SDN controller. Based on the continuously received control messages, the controller optimizes the best routes for data flows according to dynamically changing traffic patterns and flow QoS requirements and sets up the forwarding tables of SD-switch or SD-BS along the optimal path via certain secure protocols (e.g., Openflow), thus enabling highly efficient data transmissions and superior link utilization. While the effectiveness and scalability of SDN highly depend on the timely delivery of control messages from switches or BSs to the controller, control traffic balancing faces several challenges including: (i) the need of fast balancing algorithm that provides the required system QoS as well as guarantees link capacity in a timely manner; (ii) the decision of optimal controller locations that should be driven by the proposed balancing algorithm; and (iii) the unique mobile features from SD-RAN should be reflected into the balancing design for accurate traffic modeling and better system performance.

To address these challenges, inspired by our previous work [32], SoftAir provides a novel mobility-aware control traffic balancing that decides the optimal controller locations and optimal forwarding paths of control flows with respect to the mobile traffic. First, through the unique features of SoftAir, users’ mobility pattern as well as the corresponding traffic model are well characterized by the derivation of mobile traffic distribution. Moreover, while the placement of large numbers of controllers is a NP hard problem, the feasible placement set can be illustrated via the development of approximation algorithms. Next, regarding the modeled traffic distribution and a feasible solution of controller placement, a nonlinear optimization framework for control traffic balancing is formulated to find the optimal forwarding paths from SD-switch and SD-BS to the controller in such a way the average control traffic delay in the whole network is minimized. Following the proposed framework, several well-known fast algorithms, e.g. ADMM [11], might be exploited to provide the optimal solution within few milliseconds. Finally, the obtained delay results are further feedback to the placement decision, and trigger the adaptive control for better controller placement with the fulfillment of timely control message delivery in SoftAir.

(1) Research challenges:

• **Mobile feature extraction of SD-RAN:** to provide mobility-aware design, users’ movement should be characterized by suitable mobility models such as random walk, random waypoint, etc. Moreover, users’ transmission behaviors from their applications should also be identified, e.g., Poisson or Pareto traffic arrival. Based on these info, the mobile traffic distribution can be well established to capture the unique mobile feature of traffic in SD-RAN. The derived traffic distribution thus serves as an accurate mobile traffic model and will be used in the control traffic balancing design.

• **Optimal placement of controller clusters:** multicontrollers placement problem will decide the minimum required number of controllers, the controllers’ locations, and control domain assignments between SD-switch (SD-BS) and controllers. In addition to the planar placement of multiple controllers, more complicated placement problem can be considered including controller clustering and multi-threading technologies. In particular, several controllers can form a cluster and several clusters can follow the hierarchical management structure. Moreover, multi-threading methods can be further adopted to enhance the task managements among many controllers, realizing a centralized control plane.

• **Timely control traffic balancing:** the most important requirement in the design of control traffic balancing is its computation time. More specifically, while the
considered network size is often huge with many traffic flows, the corresponding parameters in the optimization framework of balancing problem are always tremendous. Therefore, how to design a fast algorithm such that it provide a solution with the desired accuracy as well as a fast convergent rate will be the focus. Moreover, this timely balancing solution should also be developed base on the above mentioned results of mobile features and optimal placement.

4.2. Resource-efficient network virtualization

The network virtualization layer of SoftAir is designed to create a set of virtual (or logic) networks on the shared network infrastructure. The virtual networks can be dedicated (i) to different network services/applications so that each service/application can be treated with customized and independent resource provisioning algorithms, (ii) to different network operators so that multiple operators can dynamically share the same network infrastructure along with the associated spectrum and infrastructure sharing, and (iii) to facilitate the cooperation and coexistence of different technologies, e.g., multi-radio access technologies (Multi-RATs).

To realize these isolated virtual networks, two functions are proposed as shown in Fig. 3: the network hypervisor for high-level virtualization, and the wireless hypervisor and switch hypervisor for low-level virtualization.

(1) Network hypervisor: It focuses on high-level resource management as shown in Fig. 5, which determines how to distribute non-conflicting network resource blocks among virtual network operators based on their demands. A utilization-optimal network hypervisor is proposed to maximize the global resource utilization, while guaranteeing the data rate requirements demanded by each virtual operators. In SoftAir, network resources include (1) wireless spectrum resources consisting of time slots and frequency channels, (2) wireless infrastructure resources, including SD-BSs and maximum power of each SD-BS, and (3) radio access technology options. Given a set of virtual operators, the data rate required by a virtual network can be formulated, which means that within the coverage area of each SD-BS, the virtual network needs to offer a certain average data rate for its users with certain spatial distribution and density. Then, at each SD-BS, we can assign the wireless resource blocks (RBs) to the virtual network. In particularly, for each channel, there is a corresponding weight parameter, which defines the percentage of time the virtual operator can exclusively use the channel. Moreover, we can further allocate the power and radio access technology at the SD-BS for the virtual network. With the allocated resource blocks, power, and radio access technology, the average data rate that this SD-BS can offer to the virtual network is thus obtained, where users follow certain spatial distribution. This average data rate relates to the effective SINR at an arbitrary user of the SD-BS and the distance vectors from each user to each RRH within the considered SD-BS.

(2) Wireless hypervisor: It is a low-level resource scheduler as shown in Fig. 6 that enforces or executes the resource management policies determined by the network hypervisor by employing a variety of wireless resource dimensioning schemes, e.g., OFDMA or wireless scheduling, so that 100% isolation among virtual networks is guaranteed. As a consequence, each virtual network can implement and adopt its own and customized NET/MAC/PHY layer protocols. Besides providing 100% isolation between the virtual networks, the wireless hypervisor has to ensure efficient utilization of the limited spectrum resources, which exhibit inherent channel
and user diversity. In the literature, very few solutions have been proposed to design wireless hypervisors so far [3, 52], which have two fundamental limitations. First, they are not flow-based and user-centric, while SDN’s traffic engineering control is flow-based. Second, they are normally not queue-length based solutions, and therefore, according to classical queueing theory, are generally not throughput-efficient.

(3) Switch hypervisor: Enabled by OpenFlow protocol, switch hypervisor (or switch fabric) focuses on bandwidth partitioning in a single SD-switch. In particular, bandwidth provisioning on switches aims to offer predefined bandwidth for any specific traffic flow or virtual network. An well-known OpenFlow(-based) virtualization, called FlowVisor, is an layer responsible for the isolation among slices of the virtualized infrastructure and employs leaky-bucket scheme for bandwidth provisioning, which however, is not a work-conserving policy and thus can not guarantee full bandwidth utilization. To solve this issue, the recent work [26] exploits GPS models upon Combined Input Cross-point Queued (CICQ) switches, which ensures full bandwidth utilization while providing accurate bandwidth provision.

(4) Research challenges:

• Utilization-optimal network hypervisor: To facilitate network hypervisor, the distance vectors that depend on the user spatial distribution of the virtual network are first investigated based on the random graph theory [40]. The utilization-optimal network hypervisor then aims at finding the optimal resource allocation for each virtual network, in terms of resource blocks, power levels, and radio access technologies so that the data rate requested by each virtual network at every SD-BS is guaranteed, and the total data rate of all the virtual networks is maximized. This problem is a nonseparable non-quadratic problem, which is NP hard. Also, a huge parameter set brings a multidimensional data decision problem. To encounter these difficulties, three algorithms can be implemented, including dimensional reduction via kernel principal component analysis (PCA) [46], convex optimization for big data [15], and randomized matrix algorithm [36], which have been known to be effective for big data processing. Moreover, the performance and the complexity of the above three algorithms will be compared to obtain the optimal solving strategy.

• Throughput-efficient and flow-based wireless hypervisor: To enable flow-based wireless hypervisor, the generalized processor sharing solution (GPS) can be applied so that each virtual network or network flow has the exclusive usage of wireless resources (e.g., channels) for a certain percentage of time. Since the conventional GPS enables each network virtual operator to apply the customized MAC and Physical layer algorithms (e.g., the one proposed in section V-B) independent of other virtual network operators. However, the conventional GPS does not provide throughput efficiency in reallocating unused network resources (e.g., channels) when some virtual networks are temporally idle i.e., having all flow queues empty. This is because the weight calculation of conventional GPS is insensitive to the channel and queue-length conditions. Inspired by the throughput-optimal property of maximum weight scheduling, the queue-length based GPS (Q-GPS) can be developed, which assigns weights proportional to the product of queue length and channel rate. Towards this, the theoretical performance bounds of Q-GPS in terms of fairness and throughput enhancements need to be investigated.

4.3. Distributed and collaborative traffic classifier

The third enabling tool that will be developed is a distributed traffic classifier (Fig. 7). Traffic classification is the task of associating network traffic flows with the application that generated it, or of categorizing network traffic flows into different QoS classes (e.g., interactive, bulk data transfer,
streaming, and best effort). It is an essential function to enable differentiated resource provisioning and service-based pricing in future broadband mobile systems. Compared with wired networks, traffic classification in wireless networks faces new and additional challenges, including: (i) As mobile applications keep evolving, fine-grained and adaptive traffic classification is required at BSs for efficient uplink resource allocations for highly dynamic and interactive mobile traffic, e.g., VoIP and mobile gaming; and (ii) because of the high volume of high-speed and delay-sensitive traffic generated by emerging mobile multimedia applications, it is desirable to adopt fast and light-weight traffic classification methods that still achieve high classification accuracy.

To address these challenges, SoftAir adopts a distributed and collaborative traffic classification system [24], which will leverage the unique features of SoftAir to achieve fast, fine-grained and accurate traffic QoS classification i.e., inferring which QoS class a traffic flow belongs to. Our proposed system will take advantages of the centralized and computationally powerful network controller to jointly exploit the advantages of deep packet inspection (DPI) [39] and semi-supervised machine learning [34]. The proposed system will consist of two components: (i) the local traffic classifiers in SD-BSs at the network edge and (ii) the global traffic learner at the network controller located at the core network. The local traffic classifier will be responsible for identifying the QoS class of a traffic flow through a mapping function. The mapping function is simply a function that takes a few features of the traffic flow, e.g., the average packet interarrival time, Hurst parameter and port number, as the inputs and gives the QoS class of the traffic flow as the output. The global traffic learner at the controller is responsible for learning, building and refining the mapping function based on the historical traffic information. The proposed traffic classification system has two advantages. First, the light-weight and fast traffic classifier can be easily implemented as a mapping function and seamlessly integrated with the SD-BS, as shown in Fig. 2. Second, the accuracy and the adaptability of such low-complexity traffic classifier can be guaranteed by the network controller, which can utilize the global view of the network flows to build the accurate mapping functions through time-consuming computationally intensive DPI and machine learning. In particular, our system employs a publish-subscribe event system (i.e. client-server model) to let the controller proactively send the updated mapping functions to the local traffic classifier.

(1) Local traffic classifier: In our design, the local traffic classifier will be implemented as a set of mini-classifiers connected in a parallel structure. Each mini-classifier is responsible for checking one feature of the traffic flow, e.g., port number [30]. The mapping function of the local classifier defines the combination pattern of the traffic classifiers. In this way, the local classifier will be dynamically refined or rebuilt by the global traffic learner at the network controller.

(2) Global traffic learner: The global traffic learner will first define the QoS classes by associating the reference or training applications with each QoS class. The reference applications of the QoS class can be defined arbitrarily as needed and updated as time proceeds. For example, the reference applications in the streaming video QoS class can include some popular and representative applications, e.g., Netflix and Youtube. Then, based on a set of historical traffic traces, e.g., training data, the global traffic learner will apply highly accurate but time-consuming DPI or stochastic packet inspection (SPI) to label or detect a portion of the traffic flows that are generated from the reference applications in each QoS class. As a consequence, each QoS class will be naturally associated with a set of features (feature vector) that are related to the labeled flows. For example, one feature of the streaming video QoS class will be Hurst parameter, which is a stochastic feature of the labeled or detected Netflix flows and used to characterize the long-range dependent nature of variable-bit-rate video traffic [21]. Next, semi-supervised machine learning will be applied, which utilizes the feature vector to group the statistically similar unlabeled and labeled flows into the same cluster, i.e., the same reference application. In this way, the mapping function between the feature vectors and the corresponding QoS classes will be trained, refined, and established through a large and adjustable volume of real-time traffic traces collected at the network controller.

(3) Research challenges:

- Complexity reduction of the mapping function: To further reduce the complexity of the mapping function, it is necessary to apply and study the dimensionality reduction approach [16], which is the technique of combining non-linearly the features in order to represent the feature vector in a lower-dimensional space. Reducing the dimension of the feature vector can directly reduce the number of mini-classifiers implemented at the SD-BSs, thus further enhancing the efficiency of local traffic classifiers.

- Optimal feature set selection: The efficiency of the machine learning based traffic classification depends on the suitable selection of a set of features (e.g., port numbers, packet interarrival time, and Hurst parameter) regarding the traffic flows. Therefore, it is of great importance to choose a suitable set of features, which can characterize the most dominant behaviors of the traffic flows.

- Adaptive feature vector training: Obtaining the accurate statistical features of the traffic flows can help to increase the classification accuracy. This, however, may cause high control overhead by collecting a large volume of traffic statistical data from the SD-switches and SD-BSs. Therefore, the optimal tradeoff between classification accuracy and control overheads has to be investigated. Accordingly, the adaptive feature vector training need be developed to cope with current network congestion conditions, while satisfying the desired classification accuracy.

5. Software-defined traffic engineering for SoftAir

In this section, we propose new traffic engineering solutions designed to leverage the full potential of the SoftAir architecture. Specifically, BS clustering, collaborative scheduling and rerouting solutions proposed in this section are directly supported by the enabling tools discussed in Section 3. In particular, the distributed traffic classification solutions in Section 3.3 provide fine-grained, accurate, and fast QoS classification for incoming traffic. By leveraging this feature, differentiated treatments, e.g., clustering, scheduling, and rerouting solutions proposed in this section, can be provided to different network applications, service providers,
and virtual operators. Moreover, the mobility-aware control traffic balancing algorithm in Section 3.1 makes collaborative gain through a large number of RRHs possible, and enables both dynamic BS clustering and collaborative scheduling. The throughput-optimal collaborative scheduling algorithm to be proposed in this section can achieve its theoretical performance limits, since the largest possible network capacity region is provided by dynamic SD-BS formation. Last but not the least, the network virtualization algorithms in Section 3.2 enable hybrid routing algorithms, e.g., QoS-guaranteed handover rerouting, that can simultaneously offer the best QoS performance for fundamentally different traffic flows.

5.1. Capacity-optimal dynamic base station formation

In next-generation wireless networks, cooperative multiuser MIMO (MU-MIMO), also known as network MU-MIMO, is deemed as a key candidate to exploit the inter-cell interference and improve the system throughput [4,6]. This approach also addresses interference limitations of the original MU-MIMO solutions [27]. Network MU-MIMO is realized by BS coordination or clustering, where the transmitting antennas of a cluster of BSs cooperatively act as a single antenna array. Despite its great potential, the realization of network MU-MIMO in current wireless network architectures is facing fundamental challenges because of limitations in fine-grained synchronization among BSs, low-cost and low-delay channel/user information sharing, and the “cluster edge” effect, where the users at the edge of a cluster suffer from severe out-of-cluster interference.

In SoftAir, our novel SD-RAN design in Fig. 2 promises three major advantages to address these challenges: (1) the SD-BSs, implemented in software and co-located at the BBS, can be dynamically formed by associating with any clusters of RRHs to mitigate the ‘cluster edge’ effect, while providing cost- and delay-free CSI exchange. (2) The SD-BS is able to provide centralized, accurate, and high-resolution synchronization among the RRHs using standardized interfaces such as CPRI [20]. (3) Different from the emerging distributed RANs, e.g., Cloud-RAN, our SD-RAN employs an innovative fine-grained BS decomposition, which significantly reduces the bandwidth burden on the fronthaul networks, thus enabling highly scalable and flexible cooperation among a large number of RRHs. In summary, our proposed SoftAir architecture can allow next-generation wireless networks to approach the economic and theoretically-optimal linear scaling of network capacity by aggressively deploying a large number of low-cost RRHs and facilitating diverse cooperation modes among them. Ideally, by forming a single giant SD-BS that incorporates all RRHs for full cooperation, the inter-cell interference in SoftAir can be mitigated, enabling an optimal and universal frequency reuse factor of one. However, realizing such a fully cooperative system needs to face the realistic network conditions. An immediate new problem arises in SoftAir: how much cooperation gain can be achieved by dynamically forming the SD-BSs.

**Dynamic software-defined base station formation:** We introduce a generic clustering optimization framework for dynamic SD-BS formation with the objective to maximize the spectral efficiency, while considering the specific cooperation costs in SoftAir and extend this for high density RRH deployments. More specifically, as shown in Fig. 8, the proposed framework aims at finding optimal associations between SD-BSs and RRHs, according to the dynamic user distribution in such a way that the network capacity is maximized by jointly mitigating the impact of the fronthaul delay, the inter SD-BS interference, and the channel estimation cost. To formulate the dynamic SD-BS formation problem, we consider a cellular network with a set of RRHs. Here, the objective is to maximize the overall capacity by forming several clusters or SD-BSs, while considering the cooperation cost induced by the fronthaul network. In particular, given the distances between SD-BSs and RRHs, the fronthaul delays can be easily obtained as the distances divide the speed of light. Then, the cluster-specific maximum fronthaul delay is given as the
maximum delay in a considered cluster. In order to achieve the maximum capacity, the RRHs of an SD-BS need to transmit the precoded data at the right phase to minimize symbol timing offsets at the mobile users. It means that at each time slot, this cluster delay is needed at most to synchronize the transmissions of the RRHs, which leads to a channel utilization ratio of the corresponding SD-BS as the ratio of the slot duration to the total time duration, i.e., slot duration pluses the cluster delay. Moreover, considering a set of mobile users served by this specific SD-BS, the signal-to-(out-of-cluster)-interference-plus-noise ratio (SINR) received by a mobile user is a function of channel gain (including distance decay, shadowing, building penetration losses, and antenna patterns) from each RRH to the mobile user, the noise spectral density, and the transmitted bandwidth. Therefore the observation at the user can be formulated including three terms: the noise, the out-of-cluster interferences, and the aggregated strength from each RRH in the cluster, where fading channel coefficients are assumed mutually independent random variables with unit-variance. Note that since the Gaussian distribution correctly models thermal noise and the interference is made up of a large number of independent components, the aggregate interference plus noise is approximately Gaussian with unit variance. Also the out-of-cluster interference (and thus the received signal quality) now scales not only with the power levels of out-of-cluster RRHs, but with other parameters such as cell size or noise variance.

To further examine practical aspects of SD-BS formation, we consider explicit channel estimation followed by coherent detection of payload data. Specifically, pilot-assisted channel estimation is exploited as follows. For a specific cluster, the share of symbols reserved for pilots for users and for payload data consist the entire transmissions. The pilot transmissions should be orthogonally multiplexed from each of the transmissions. Thus, each receiver’s estimation of the channel coefficients corresponding to in-cluster RRH transmitters shares the pilot symbols. Moreover, the fading channel coefficients now consist two uncorrelated terms: the channel estimations and the corresponding estimation errors. Therefore, we can formulate the minimum mean square error and rewrite the observation for each user. The mobile users customarily utilize the channel estimates as if they were correct, in which case the estimation errors play the role of additional Gaussian noise. Accordingly, the effective SINR at the user upon payload data detection is then given, and the average spectral efficiency (bits/Hz/user) of the cluster is derived, which depends on the type of cooperation among the users in the cluster.

(1) Network massive MU-MIMO (Cooperation mode): The SD-BS clustering problem can be formulated as a capacity maximization problem based on the entire average spectral efficiency from the given clustering assignment constrained such that each RRH is covered by one cluster set. Through the proposed Dynamic Software-Defined Base Station Formation, multiuser multiplexing gains of network MU-MIMO can be leveraged to improve global spectral efficiency and network capacity; by optimally managing the inter-cluster interference, the fronthaul network delay, and cluster edge effects. First, a fast algorithm will be investigated to solve the proposed BS formation problem through Fully Polynomial Time Approximation Schemes (FPTAS) [28] or ADMM [11]. Such a fast algorithm is of significant importance to ensure the adaptability of the SD-BS formation to the dynamic user distribution. Then, the superior spectrum efficiency will be demonstrated of SoftAir by advancing the concept of network MU-MIMO to network massive MIMO as shown in Fig. 9, which aims at enhancing the network capacity by forming virtual directional antenna via the very large antenna array offered by the network MIMO. In particular, through the cooperation among RRHs via the BBS, the pilot contamination in massive MIMO can be completely eliminated. This means in the asymptotic regime, the interference free systems can be realized from some simple pre-coding schemes as the number of cooperated RRHs (i.e., the size of antenna array) goes to infinity.

(2) Millimeter wave usage (Coordination mode): Comparing to rough 600 MHz available spectrum in conventional microwave cellular systems, millimeter wave (mmWave) provides enormous amounts of spectrum at frequencies ranging from 3 to 300 GHz. Such a plentiful resource in mmWave leads many researches with diverse backgrounds, from hardware component to system processing algorithm designs, to study mmWave transmissions in different aspects. The major difference between mmWave and microwave frequencies, also being the most challenging problem, is the sensitivity to blockages. In particular, the path loss exponent for non-line-of-sight propagation is double than the one for line-of-sight case, making the channel modeling very complicated. This challenge becomes even more pronounced when the channel changes rapidly due to mobility and high Doppler shifts at mmWave frequencies. To deal with the mentioned difficulty, in SoftAir architecture, SD-BSs provide an effective solution through the operations with RRHs as shown in Fig. 10. That is, SD-BSs simply enable the adaptive antenna arrays for net array gain and interference mitigation via the coordinates of RRHs’ orientations and functionalities. It thus supports the ubiquitously good channel qualities in the entire geographic area. Toward this, two essential functionalities should be involved to realize this ubiquitous coverage as: (1) the adaptive hosting scheme among SD-BSs (i.e., the changing of the supporting RRHs) and (2) the coordination of omnidirectional microwave and directional mmWave transmissions. In
particular, when users move around a certain area, the host SD-BS as well as the corresponding RRHs’ mmWave operations should be also changed accordingly to ensure the constant good receptions at users. Furthermore, once the mmWave transmissions cannot support the required signal qualities, the coordination mechanism takes place for joint mmWave and microwave transmissions to greatly increase the spectrum efficiency.

3) Research challenges:

- Coordinate virtual directional antennas: After solving network MU-MIMO problem through the suggested solutions, the network massive MU-MIMO is addressed as follows. While massive MIMO utilizes the uplink transmission to estimate the downlink channel qualities, it suffers the well-known pilot contamination problem. Same situation happens when network massive MU-MIMO is considered. Moreover, in addition to deal with the conventional contamination problem, the cooperation among SD-BSs (and RRHs) should also be determined at the same time. These considerations bring out the coordination problem of virtual directional antennas. In particular, the dynamic SD-BS formation solutions with hybrid network MIMO modes should be supported, which optimally configure the cooperation strategies of RRHs to maximize the joint gain of performing the multiuser multiplexing at the center of the RRH cluster and the beamforming at the edge of the RRH cluster.

- Ubiquitous millimeter wave covering: To enable the ubiquitous covering whenever the user with mmWave transmission moves, the entire optimal planning of transmission techniques can be formulated. Specifically, while there are two options, i.e., adaptive SD-BSs and coordination transmissions, for each user in any specific location, the optimization objective can be set as the minimum signaling overheads. And it should subject to the required data rates, which will be the function of SD-BSs’ set, RRHs’ operations, and the available combinations of joint microwave and mmWave transmissions. As this planning problem might be NP-complete (NPC), the randomized approximation algorithm can be applied to obtain a workable solution within few rounds, i.e., very less time for the calculations. Then, based on this solution as the staring point, several fast optimal searching algorithms can be further applied to approach the suboptimal solution for the original complicated problem.

5.2. Throughput-optimal collaborative resource provisioning

As shown in Section V-A, each SD-BS can be optimally formed so that fine-grained coordination among a cluster of RRHs is enabled. Such promising feature along with the distributed traffic classifier on each SD-BS, for the first time, allows to perform throughput-optimal joint resource provisioning on a SD-BS. Throughput-optimal collaborative resource provisioning in Fig. 11, based on throughput-optimal scheduling [33], aims to optimally and collaboratively distribute network resources of a cluster of RRHs among their
users based on the statistical properties of their traffic flows so that (i) the predefined QoS of each flow can be guaranteed and (ii) the overall throughput of the SD-BS is maximized.

Consider \( N \) users served by a SD-BS, which is associated with a set \( \mathcal{M} \) of RRHs with each RRH having \( d \) antennas. Consequently, the SD-BS has a set \( \mathcal{L} \) of antennas, where \( |\mathcal{L}| = M \times d \), and each antenna can operate on a set \( C \) of channels. The SD-BS assigns a queue \( q_n \) to each user \( n \). Let \( \lambda_n(t) \) denote the traffic flow arriving at queue \( n \) at time slot \( t \) with the traffic rate \( \lambda_n = E[\lambda_n(t)] \). At each time slot \( t \), the SD-BS assigns several antennas to each user. If antenna \( l \) is assigned to user \( n \) at time slot \( t \), the SD-BS allocates transmission power \( P_{nl}^c \) with modulation scheme \( M_{nl}^c \) for each channel \( c \in C \) on this antenna \( l \). The objective of the throughput-optimal QoS provisioning problem is to find the optimal antenna configuration \( \{S_{nl}(t)\}_{l,c} \), transmission power \( \{P_{nl}^c\}_{l,c} \) and modulation scheme \( \{M_{nl}^c\}_{l,c} \) for all users \( n \in N \) with an objective to maximize the total SD-BS throughput RBS, while at the same time satisfying that (1) the average data rate of user \( n \) is larger than its incoming traffic rate, i.e., \( R_n > \lambda_n \) and (2) the probability that the queuing delay \( D_n(t) \) of user \( n \) is larger than a delay threshold \( T_n \) is smaller than a predefined violation value \( \delta_n \), i.e., \( Pr(D_n(t) > T_n) < \delta_n \). The throughput-optimal QoS provisioning problem can be formulated as a stochastic optimization problem.

The solution of the throughput-optimal collaborative resource provisioning problem depends on the stochastic properties of the traffic flows. Based on the burstiness level, network traffic can be classified into two categories: heavy-tailed (HT) flows and light-tailed (LT) flows. Specifically, a traffic flow is heavy-tailed if its traffic arrivals follow heavy-tailed (HT) flows and light-tailed (LT) flows. Specifically, a traffic flow is heavy-tailed if its traffic arrivals follow a heavy-tailed distribution. Otherwise, the traffic flow is light-tailed. Specifically, a random variable \( X \) follows heavy-tail distribution if its tail distribution decreases slower than exponentially. Recent significant empirical results collected from commercially operated wireless networks [35,45] show that the mobile data traffic on WiFi networks and cellular networks is heavy-tailed, thus extremely bursty nature. The observed heavy-tailed behavior in mobile traffic is largely departing from the conventional light-tailed traffic assumptions in wireless network research, e.g., Poisson and Markovian traffic. However, this phenomenon is highly consistent with the current trend that the web and video traffic, which is inherently heavy tailed, is dominating wireless networks. Because of the heavy-tailed nature of mobile traffic, solving the throughput-optimal collaborative scheduling problem is very challenging. Under the conventional light-tailed traffic assumption, this problem can be solved by applying large deviation theory [41]. However, large deviation theory relies on the assumption that the exponential expectation of the traffic arrivals is finite. This assumption holds for light-tailed traffic, but not for heavy-tailed traffic – all heavy-tail distributed random variables have infinite exponential expectation.

To address this challenge, the distributed traffic classifier of SoftAir can be leveraged. More specifically, with such traffic classifier, the SD-BS not only has the capacity to identify heavy-tailed traffic flows but can also “extract” their stochastic attributes, e.g., the tail index or burstiness level of traffic flows. With such information, the burstiness-aware resource provisioning algorithms [49,50,51] can be applied to solve the throughput-optimal QoS provisioning problem. Such algorithms make provisioning decisions based on the queueing delay of user raised to the \( J_3 \)-th power, where \( J_3 \) is determined by (i) the burstiness or the tail index of the traffic arrivals \( An(t) \) and (ii) the QoS parameters of each flow, including the delay threshold and the delay violation value.

1. **Research challenges:**

- **Low-complexity optimal resource provisioning:** In SoftAir, each SD-BS can manage a large number of RRHs and each RRH can be equipped with tens or hundreds of antennas (e.g., massive MIMO). Due to its programmable feature, SD-BS can be also associated with a large pool of spectrum bands ranging from TV channels to mmWaves. As a result, it is demanding to develop large-scale convex or non-convex optimization algorithms for resource provisioning at SD-BSs, e.g., weight design for large antenna arrays of mmWaves, precoding for network/massive MIMO, and collaborative scheduling over a massive number of RRHs, while maintaining low computational complexity and fast convergence rate.

- **Virtualization-enabled Multi-RAT QoS provisioning:** Because of the network virtualization capability of SoftAir, multiple virtual networks can be created on the same SD-BS, each of which is dedicated for a different application, which requires totally different radio access technology along with different QoS requirements. For example, machine-to-machine type applications may rely on the contention based radio access technology, e.g., IEEE 802.15.4 with ISM bands to deliver low data rate information. The multimedia streaming applications may utilize reservation based radio access, e.g., OFDMA with cellular bands to meet the strict delay and throughput requirements. As a result, Multi-RAT QoS provisioning schemes need to be developed for the SD-BS. In this case, each virtual network is adopting suitable resource provisioning schemes optimized for its associated radio access technology, while the dynamic resource distributions among the virtual networks can be realized to cope with the traffic dynamics of the virtual networks.

5.3. **Software-defined mobility management framework**

Mobility management contains two components: location management and handoff management [8]. Location management enables the system to track the locations of mobile users between consecutive communications. Location management consists of registration and paging. Registration is performed by a mobile user to update its current location with the system when leaving a registration area, where the registration area is a group of cells within which the user is allowed to move without updating its location. Paging is performed by the system to search for a mobile user by sending poll messages to the cells close to the last reported location of the mobile user at the arrival of an incoming call. In addition, handoff management is the process by which users keep their connections active when they move from one BS to another. Handoff management consists of BS association and handoff rerouting. BS association is a procedure by which a mobile user is disconnected from the current serving BS and
associated with a new one. Handoff rerouting is performed to establish the new routing paths for the active connections of a mobile user when its associated BS changes.

Mobility management [8] faces significant new challenges in the 5G wireless systems, including:

- **Minimum signaling overhead**: To improve resource efficiency and guarantee QoS, signaling messages and delay should be minimized when (i) performing registration and paging to track mobile users and (ii) when establishing new routing paths during handoffs. However, as wireless networks are becoming increasingly heterogeneous, conventional and closed network architectures inevitably lead to increased signaling costs because of the deployment of interworking or interoperating functions to accommodate roaming between networks with different radio access technologies. Moreover, the impact of such high signaling overheads is even exacerbated since wireless networks are moving towards denser BS deployment with frequent initiations of handoff procedures and where the system may have to page a large number of BSs to locate a mobile user.

- **QoS guarantees during handoff rerouting**: Handoff management needs to simultaneously support best-effort traffic and real-time traffic, e.g., video streaming and VoIP, that requires strict QoS guarantees, e.g., delay, jitter, bandwidth, thus demanding the establishment of QoS-optimal rerouting path when handoffs occur. However, in current wireless networks, heuristic rerouting solutions (e.g., anchor-, chain-, and multicasting-based rerouting [8]), are conventionally adopted, which, however, can cause high routing complexity, long or looped routing paths, and excessive bandwidth consumption.

To address these challenges, a new software-defined mobility management framework, as shown in Fig. 12, is considered for 5G wireless networks, which leverages the multi-service multi-technology multi-network converged SoftAir to offer low-cost, QoS-guaranteed, and seamless mobility services. More specifically, the proposed framework consists of two components: minimum-cost location management and QoS-guaranteed handoff rerouting.

1. **Minimum-cost location management via Virtual registration areas**: In SoftAir, each SD-BS is a multi technology converged platform, i.e., through the OpenFlow interface homogeneous control can be applied to the SD-BSs independently of their underlying radio access technologies. Such features can enable unified and seamless registration and paging in the heterogeneous networks, thus avoiding the need for deploying inter-networking functions that can induce high signaling costs. Moreover, by exploiting the multi-technology convergence feature of SD-BSs, we can further reduce the signaling cost and increase the paging efficiency by proposing the concept of virtual registration area (VRA), which is a group of cells with different access technologies, within which the mobile user does not need to update or register its location with the cellular BSs. The key advantage of the virtual registration area is to exploit the asymmetric features of different radio access technologies. On one hand, the frequent location registration with macro-cell cellular BSs will reduce the spectrum efficiency of the already limited cellular bands. On the other hand, updating locations with the high data rate WiFi or small-cell BSs will not evidently degrade the spectrum efficiency, while enhancing the paging accuracy by providing additional location updates of the mobile users. In SoftAir, we can develop an optimal VRA design, designed to select the optimal set of heterogeneous cells for VRAs in such a way that the cellular registration signaling cost is minimized with the maximum paging accuracy enhancement.

The VRA design can exploit the innovative road-traffic based cell grouping technique in [13], by incorporating the asymmetric and heterogeneous features of next-generation wireless networks. More specifically, in the optimal VRA design, the intercell traffic prediction scheme is applied to determine the expected intercell movements of mobiles by examining the road typologies (highways, carriageways,
footpaths, etc.). These predictions are used by the road-traffic based cell grouping scheme to group cells into VRAs. In particular, each cellular cell pair (i.e., a pair of neighboring cellular macro-cells) will have two types of road traffic: cellular road-traffic and WiFi road-traffic. The former is the expected amount of mobile movements between the two cellular BSs in a cellular cell pair. The latter is the expected amount of mobile movements between the WiFi cells that reside in the cellular cell pair. With the optimal VRA design, the cellular cell pair with higher cellular road-traffic and lower WiFi road-traffic will be assigned to the same VRA. There are two advantages of such design. First, by grouping the cellular cells with high cellular road-traffic into the same VRA, we can decrease the inter-VRA cellular road-traffic (i.e., road traffic across the boundaries of two VRAs). This can considerably reduce registration overheads in cellular system since only inter-VRA cellular road-traffic will trigger registration signaling between cellular BSs and mobile users. Second, by grouping the cellular cells that have low WiFi road-traffic into the same VRA, we can increase the inter-VRA WiFi road-traffic, which can improve the location registration frequencies over high-bandwidth WiFi access points, thus allowing cost-efficient paging process in the future.

(2) QoS-Guaranteed handoff rerouting: In SoftAir, hand-off rerouting happens when a user is moving from one SD-BS to another SD-BS. Since the SD-BS is associated a large number of RRHs covering a broad geographic area, the handoff management, including BS association and handoff rerouting, will not be triggered frequently, thus leading to considerably reduced signaling overhead. On the other hand, due to the strict QoS requirements of real-time traffic, the handoff rerouting scheme needs a more sophisticated design to incorporate delay, jitter, and bandwidth requirements. In particular, our key objectives for optimal handover re-routing of real-time traffic are to achieve per-flow QoS guarantees, seamless mobility and network-wide performance guarantees. Using the SoftAir architecture, per-flow and network-wide information at the central controller can be utilized to achieve fine-grained per-flow QoS and network-wide performance guarantees. The control and data plane separation allows us to simultaneously setup forwarding rules on all switches instead of hop-by-hop signaling to enable seamless mobility. To further guarantee seamless mobility, we propose a proactive re-routing approach where new routes are computed even before the users arrive at a location under the new serving SD-BS using mobility prediction, which tells the highest possible cell a mobile user can move to. This can be realized by exploiting our predictive user mobility profile framework [7], which has been demonstrated to have superior prediction accuracy.

Formally speaking, several variables are first given to the rerouting problem as the set of flows rerouting request, the set containing all links in the SoftAir, the link capacity and the occupied data rates by the existing flows, and the end-to-end delay and the average data rate. The objective of the QoS-guaranteed handover rerouting problem is to find the optimal flow allocation so that the corresponding rerouting paths satisfy the QoS requirements and provide the maximum network utility at the same time. In particular, the objective utility function can be formulated as the the total residue link resources with respect to allocated resources. Moreover, considering network-wide performance guarantees, the per-flow requirements of delay, jitter, and bandwidth should be satisfied. In addition to the above performance guarantees, the average data rates from the existing flows and the rerouting flows should be further maintained to go below the average link capacity.

(3) Research challenges:

- **Optimal VRA design**: To minimize the registration cost while maintaining desired paging accuracy, the optimal VRA design should be developed in such a way that the average registration frequency with cellular BSs is minimized and the average registration frequency with WiFi BSs is larger than a predefined threshold for enhanced paging efficiency. The optimal VRA design can be modeled as an integer programming problem, which is generally NP-hard. Thus, low-complexity approximation algorithms need to be developed. Finally, the performance of the optimal VRA design should be evaluated for different distributions of the road traffic and the base stations.

- **QoS-aware rerouting design**: As mentioned, the considered QoS guarantees should include delay, jitter, and band-width. Specifically, the delay constraint is provided as the probability that the queueing delay of the flow is larger than a delay threshold is smaller than a predefined violation value. The jitter requirement is examined as the delay variation should be less than a predefined value. For bandwidth requirement, the average data rate of the flow should be larger than its incoming traffic rate. Then, to solve such a stochastic optimization rerouting problem, the linear traffic transformation [32] can be utilized first to map the end-to-end traffic flows into per-link traffic. Then, queueing network theory [17] and stochastic network calculus can be adopted to derive the queueing backlog and delay as well as the QoS performance bounds.

6. Conclusion

In this paper, we propose SoftAir as a new paradigm towards next-generation wireless networks. SoftAir provides high flexible architecture, which can accelerate the innovations for both hardware forwarding infrastructure and software networking algorithms through control and data plane separation, enable the efficient and adaptive sharing of network resources through network virtualization, achieve maximum spectrum efficiency through cloud-based collaborative baseband processing, encourage the convergence of heterogeneous networks through open and technology-independent interfaces, and enhance energy efficiency through the dynamic scaling of computing capacity of the SD-BSs. To realize the promising properties of SoftAir, the essential management tools, including mobility-aware control traffic balancing, resource-efficient network virtualization, and distributed and collaborative traffic classifier, are introduced. Moreover, the novel software-defined traffic engineering solutions, including dynamic BS formation, throughput-optimal collaborative resource provisioning and software-defined mobility management, are presented. In addition, the research challenges of realizing SoftAir are also summarized. As being investigating these challenges, we will be able to provide the complete solution suite in near future.
References


[17] I. Norros, Large deviations and overflow probabilities for the general single-server queue, with application, Queueing Syst. 16 (1994) 387–396.


[34] I. Norros, Large deviations and overflow probabilities for the general single-server queue, with application, Queueing Syst. 16 (1994) 387–396.


[38] SK Telecom, SK telecom 5G white paper: SK telecom’s view on 5G vision, architecture, technology, and spectrum, Technical report, 2014 [online], Available: http://www.sktelecom.co.kr (last accessed 22.05.15).


Ian F. Akyildiz (M’86–SM’89–F’96) received the B.S., M.S., and Ph.D. degrees in Computer Engineering from the University of Erlangen-Nuremberg, Germany, in 1978, 1981 and 1984, respectively. Currently, he is the Ken Byers Chair Professor in Telecommunications with the School of Electrical and Computer Engineering, Georgia Institute of Technology (Georgia Tech), Atlanta, GA USA; the Director of the Broadband Wireless Networking (BWN) Laboratory and the Chair of the Telecommunication Group at Georgia Tech. Since 2013, he is a FiDiPro Professor (Finland Distinguished Professor Program (FiDiPro) supported by the Academy of Science) with the Department of Electronics and Communications Engineering, at Tampere University of Technology, Finland. Since 2008, he is also an honorary professor with the School of Electrical Engineering at Universitat Politcnica de Catalunya (UPC) in Barcelona, Catalonia, Spain, and the founding director of NCC (Nano Communications Center). Since 2008, he is also an honorary professor with the School of Electrical Engineering at Universitat Politcnica de Catalunya (UPC) in Barcelona, Catalonia, Spain, and the founding director of N3Cat (NaNoNetworking Center in Catalonia). Since 2011, he is a Consulting Chair Professor at the Department of Information Technology, King Abdulaziz University (KAU) in Jedda, Saudi Arabia. He is the Editor-in-Chief of Computer Networks (Elsevier) Journal, and the founding Editor-in-Chief of the Ad Hoc Networks (Elsevier) Journal, the Physical Communication (Elsevier) Journal and the Nano Communication Networks (Elsevier) Journal. He is an IEEE Fellow (1996) and an ACM Fellow (1997). He received numerous awards from IEEE and ACM. His h-index is 88 and the total number of citations is above 69K due to Google scholar as of March 2015. His current research interests are in wireless sensor networks in challenged environments, 5G Cellular systems, nanonetworks, Terahertz Band, and software defined networks.

Pu Wang (M’05) received the B.S. degree in electrical engineering from the Beijing Institute of Technology, Beijing, China, in 2003, the M.Eng. degree in computer engineering from the Memorial University of Newfoundland, St. Johns, NL, Canada, in 2008, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, in 2013. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS. Dr. Wang received the BWN Lab Researcher of the Year award in 2012, Georgia Institute of Technology. He received the TPC top ranked paper award of IEEE DySPAN 2011. He was also named Fellow of the School of Graduate Studies, 2008, Memorial University of Newfoundland. His research interests include wireless sensor networks, cognitive radio networks, software defined networks, nanonetworks, multimedia communications, wireless communications in challenged environment, and cyber-physical systems.

Shih-Chun Lin (S’08) received the B.S. degree in electrical engineering and the M.S. degree in communication engineering from National Taiwan University in 2008 and 2010, respectively. He is a Graduate Research Assistant in the Broadband Wireless Networking Laboratory (BWN Lab), School of Electrical and Computer Engineering, Georgia Institute of Technology. Currently, he is working toward the PhD degree in electrical and computer engineering under the supervision of Prof. Ian F. Akyildiz. His research interests include wireless underground sensor networks, software defined networking, large machine-to-machine communication, cognitive radio networks, and statistical scheduling in wireless systems.