SDN-Based Architecture for Providing Reliable Internet of Things Connectivity in 5G Systems

Luis Tello-Oquendo, Ian F Akyildiz, Shih-Chun Lin, Vicent Pla

SDN-Based Architecture for Providing Reliable Internet of Things Connectivity in 5G Systems

Luis Tello-Oquendo[*†], **Ian F. Akyildiz**[*], **Shih-Chun Lin**[‡], and **Vicent Pla**[†]

[*]Broadband Wireless Networking Laboratory, School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332, USA
[†]Instituto ITACA. Universitat Politècnica de València, Valencia 46022, Spain
[‡]Department of Electrical and Computer Engineering, North Carolina State University Raleigh, NC 27606, USA

***Abstract*—Sheer number of devices in Internet of Things (IoTs) fundamentally challenge the ubiquitous information transmissions through the backbone networks, such as cellular systems. The heterogeneity of IoT devices and the hardware-based, inflexible cellular architectures impose even greater challenges to enable efficient communication. To address these challenges, this paper introduces the so-called SoftAir architecture on wireless software-defined networking and proposes software-defined gateways (SD-GWs) that jointly optimize cross-layer communication functionalities between heterogeneous IoT devices and cellular systems. First, the SoftAir architecture is proposed to support a unified software-defined platform for quality-of-service aware IoT systems and software-defined radio access networks (SD-RANs) with millimeter-wave transmissions. Next, the SD-GWs are designed in SoftAir to explore the interactions between two-types of networks (i.e., IoTs and SD-RANs) and enable cross-layer solutions that simultaneously achieve optimal energy savings and throughput gain in IoTs and maximum sum-rates in SD-RANs. Simulation results validate that our SoftAir solutions surpass classical IoT schemes by jointly optimizing communication functionalities for both IoTs and SD-RANs and bring significant system synergies for reliable 5G IoT communication.**

I. Introduction

The evolution of communication networks, devices, and applications has drawn a new technological age in which everything is connected. Internet of Things (IoTs) has extended the scope of wireless communication services from interpersonal communications to smart interconnection between things and between people and things, allowing wireless communication technologies to penetrate into broader industries and fields. The number of connected IoT devices is expected to increase between 10- and 100-fold beyond 2020 [1]; these devices will range from devices with limited resources that require only intermittent connectivity for reporting (e.g., sensors) to devices that require always-on connectivity for monitoring and/or tracking (e.g., security cameras, transport fleet). IoT connectivity drives the development of 5G cellular systems where a combination of advances such as air interface design, millimeter-wave (mmWave), software-defined networking (SDN), signaling optimization, and intelligent clustering and relaying techniques can all contribute to enable efficient communication and support such hyper-connectivity [2].

Traditional network architectures will not be able to handle both the number of devices and the volume of data they will be draining into the network. Moreover, current IoT solutions rely on low-power wide area (LPWA) networks [3], which complement traditional cellular and short-range wireless technologies in addressing IoT applications. Several technologies, such as LoRa, NB-IoT, SIGFOX, have been developed and designed solely for applications with very limited demands on throughput, reliability, or quality-of-service (QoS). However, without a central regulation among these LPWA technologies, existing IoT solutions cannot support highly diverse QoS requirements from increasing 5G IoT applications. Due to currently fixed and hardware-based infrastructure, no existing work has considered the joint architectural design of IoT networks and software-defined radio access networks (SD-RANs), and the provision of reliable and efficient upstream/downstream IoT transmissions. Another challenge is to efficiently manage the load of traffic and the network resources in the 5G era, to avoid a possible collapse of the network, and to allow the coexistence of different services with different QoS requirements in a scalable and efficient manner.

In this paper, to adequately address the above challenges in 5G IoT, we introduce a new architecture proposed for wireless software-defined networks, the so-called SoftAir [4]; then, software-defined gateways (SD-GWs) are designed in the SoftAir SD-RAN for providing reliable connectivity and services to IoT applications. Our solution overcomes the limitations of existing commercial wireless networks that are inherently hardware-based and rely on closed and inflexible architectural designs by offering five core properties: programmability, cooperativeness, virtualizability, openness, and visibility. These five properties provide functionalities that are essential to enable 5G wireless communication networks and support emerging IoT applications and services. We consider a likely IoT scenario based on several wireless sensor networks (WSNs) that provide IoT services through the SoftAir system. The proposed SD-GW in the SoftAir architecture will manage the sporadic communications from a myriad of the heterogeneous IoT devices and provide local offloading. It aggregates the data from IoT devices clustered geographically and provides them with access to the Internet using SoftAir. It implements the physical, link, and network layers and can support multiple wireless interfaces. For instance, in downstream transmissions, it can connect with the IoT devices using IEEE 802.15.4 or near-field communications (NFC), and in upstream transmissions, it can use the SoftAir protocol stack to communicate with the remote radio heads (RRHs).

Specifically, we develop a cross layer framework and propose a joint optimization of protocols crossing different layers from the IoTs to the SD-RAN according to the devices' QoS requirements and system constraints. Thus, we provide a solution for various performance requirements of applications to handle the heterogeneity of IoT devices. The optimization framework is performed at the SD-GW where a local IoT controller resides; it explores the interactions of these two-types of

This work was supported by the US National Science Foundation (NSF) under Grant No. 1547353, in part by the Ministry of Economy and Competitiveness of Spain under Grants TIN2013-47272-C2-1-R and TEC2015-71932-REDT.

ISBN 978-3-903176-05-8

networks and enables cross-layer solutions to simultaneously achieve optimal energy savings and throughput gain in IoTs, as well as maximum sum-rate in SD-RANs. Furthermore, with the introduction of IPv6, the vast increase in the number of connected devices is properly addressed and the SD-GW can be used to send IoT data to other devices connected to the Internet. The main contributions of this paper are summarized as follows

- We present the SoftAir architecture to provide IoTs connectivity that exploits the emerging features in wireless communications. A study that explores the interactions of communication functionalities for an IoT scenario based on WSNs is provided.
- We design a heterogeneous cross-layer solution for the SD-GW aiming to fulfill a predefined level of QoS, efficient energy consumption, high system performance, and reliable connectivity.
- We develop an optimization framework that achieves optimal energy savings and throughput gain concurrently in WSNs while maximizing the SD-GW rate coverage with mmWave RRHs coordination in SoftAir.

To the best of our knowledge, this work is the first to provide a unified cross-layer optimization framework for IoT communication within 5G systems. The rest of the paper is organized as follows. Section II describes the SoftAir architecture for IoT communications. Section III presents the heterogeneous cross-layer optimization design in SD-GWs that integrates IoTs and SD-RAN of SoftAir. Section IV gives the performance evaluation, and Section V concludes the paper.

II. SoftAir Architecture for 5G IoTs

SoftAir [4] is a unified software-defined platform for 5G systems with network management tools and customized applications of service providers or virtual network operators. It would enable IoT applications to access the data and control the devices without the knowledge of the underlying infrastructure. SoftAir follows a distributed RAN architecture composed of three main parts: (i) the centralized base band servers (BBS) pool, which connects to the core network via backhaul links and consists of software-defined base stations (SD-BSs) from real-time virtualization technology for software-implemented baseband units (e.g., digital processing tasks); (ii) RRHs plus antennas, which are remotely controlled by SD-BSs and serve SD-GWs' transmissions; and (iii) low-latency high-bandwidth fronthaul links (fiber or microwave) using the common public radio interface (CPRI) for an accurate, high-resolution synchronization among RRHs.

Extended from our preliminary study in [4], Fig. 1 depicts an example of the SoftAir-based architecture for 5G IoTs. It consists of three domains: sensing, network, and application. The sensing domain enables *things* to interact and communicate with themselves and with the communication infrastructure; it realizes the data collection of physical targets employing technologies such as WSNs, RFID, ZigBee or NFC. The network domain builds on SoftAir; it aims to transfer

Notations: Throughout this paper, boldface lower and upper case symbols represent vectors and matrices, respectively; $\mathbf{I}_x$ denotes an x by x identity matrix; $\mathbb{C}^{x,y}$ denotes the set of $x \times y$ complex matrices. The trace, transpose, and Hermitian transpose operators are denoted by $\mathrm{tr}(\cdot)$, $(\cdot)^{\mathsf{T}}$, and $(\cdot)^{\mathsf{H}}$, respectively. We use $\mathcal{CN}(\mathbf{X}, \mathbf{Y})$ to denote the circular symmetric complex Gaussian distribution with mean matrix $\mathbf{X}$ and covariance matrix $\mathbf{Y}$; the distribution of a uniform random variable (r.v.) is denoted by $\mathcal{U}(\cdot)$, the distribution of a normal r.v. with mean x and variance σ^2 is denoted by $\mathcal{N}(x, \sigma)$, and $\sim$ stands for "distributed as". Expectation is denoted by $\mathbb{E}[\cdot]$, variance is denoted by $\mathbb{V}[\cdot]$. $\|\mathbf{x}\|$ denotes the Euclidean norm of complex vector $\mathbf{x}$, and $|z|$ denotes the magnitude of a complex number z. The indicator function is denoted by $\mathbb{I}[x]$; it returns 1 when x is true, and 0 otherwise.

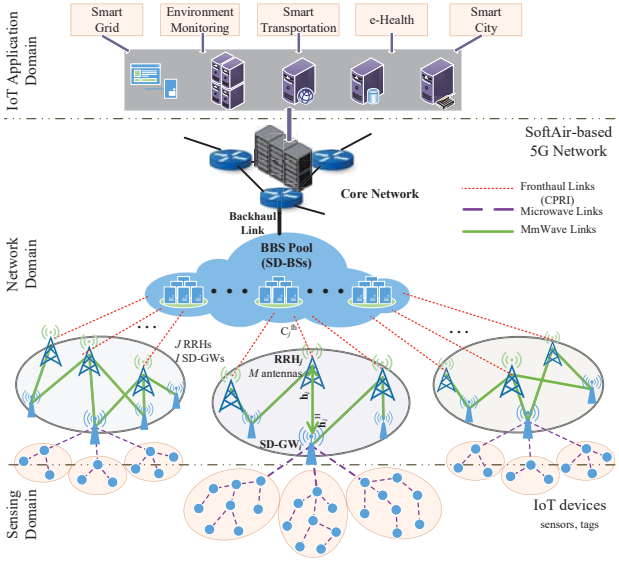


Figure 1. SoftAir [4] network architecture for 5G IoT communication.

the data collected from the sensing domain to the remote destination in the application domain. Finally, the application domain is responsible for data processing and the provision of a wide variety of applications and services.

It is worth noting that a relevant architectural component is the SD-GW, that lies between the sensing and network domain. Besides alleviating the high traffic bursts imposed by sporadic communications from a myriad of heterogeneous IoT devices, the SD-GW aggregates the data from the IoT devices clustered geographically forming several WSNs and provides them access to the Internet using 5G wireless. In that sense, the SD-GW comprises two interfaces: northbound and southbound. The former communicates with the SoftAir system, whereas the latter interconnects the devices inside the cluster, i.e., the SD-GW implements protocols that support co-existence of diverse wireless interfaces, such as intelligent management of interference and distributed management of channel allocation/medium access.

A. System Model

Considering the SoftAir domain of the above architecture, the system consists of a set $\mathcal{I} = \{1, \ldots, I\}$ of SD-GWs that provide connectivity to several WSNs clustered geographically. Inside the WSN, the set $\mathcal{K} = \{1, \ldots, K\}$ of nodes communicate with neighboring devices by 6LoWPAN; we assume that the ith SD-GW acts as cluster head of the set $\mathcal{K}$ of nodes, and relays the data it receives to the SD-RAN. Then, the set $\mathcal{I}$ of SD-GWs is served by a set $\mathcal{J} = \{1, \ldots, J\}$ of associated RRHs. All the RRHs are connected to the BBS pool $\mathcal{B}$ via fronthaul links, where the jth fronthaul link between the $j \in \mathcal{J}$ RRH and the $b \in \mathcal{B}$ BBS has a predetermined capacity C_j^{fh}. Note that by using low-latency high-bandwidth fronthaul links, the software-defined architecture implements an accurate, high-resolution synchronization among RRHs and enables flexible and tangible RRH coordinations. The associations between the RRHs and SD-GWs can be determined based on the distance or channel gain from RRHs to each SD-GW. The RRHs are equipped with an array of M antennas and communicate with the single antenna SD-GWs through mmWave links. One RRH can serve a number of SD-GWs; the jth RRH that is assigned to serve the ith SD-GW will receive the SD-GW's processed base band signal from the BBS pool.

Then, the RRH converts and transmits the corresponding RF signal using a suitable designed pre-coding vector as detailed in Section III-B.

In the SoftAir domain shown in Fig. 1, the SD-GW incorporates a local controller, and has the key role of being a concentrator of several sensor nodes for both control and user planes. It possesses the necessary knowledge to orchestrate the sensors such as network topology, link qualities, and application requirements. Besides performing conversions to communicate between different standards, the SD-GW performs the optimization framework, and can make decisions such as the choice of network parameters and protocols. The network application can then be modified by simply changing the forwarding rules at the local controller, which then propagates the changes to sensors.

III. Heterogeneous Cross-Layer Solution for Software-Defined Gateway

In the following, we develop a cross-layer optimization framework that integrates the sensing domain and the SD-RAN of the SoftAir system, allowing coordination, interaction, and joint optimization of protocols crossing different layers. We elaborate the communication functionalities for both the sensing and SD-RAN domain in Section III-A and Section III-B, respectively. Then, a centralized optimization framework to jointly control the parameters is formulated in Section III-C to ultimately reach an optimum configuration according to an application-dependent objective function. Finally, the protocol operation at the SD-GW is detailed in Section III-D.

A. IoT & WSN Network

In this section, we describe the parameters and communication functionalities at the physical layer (channel, modulation), link layer (channel coding, MAC), and network layer (addressing, routing) for the nodes in the sensing domain.

1) Physical layer functionalities: At the physical layer, the nodes follow the frequency spectrum allocation according to the IEEE 802.15.4 standard [5]; they might have different maximum transmission power and can select different modulation schemes. We use the log-normal channel model, which has been experimentally shown to model the low power communication in WSN accurately [6]. In this model, the total path-loss in dB is given by $l^{\text{WSN}}(d_i)[\text{dB}] = l^{\text{WSN}}(d_0) + 10\bar{n}\log_{10}(d_i/d_0) + \eta$ for $d_i \geq d_0$, where d_i is the transmitter-receiver distance; d_0 is a reference distance; $\bar{n}$ is the path-loss exponent for a particular frequency band or environment; $\eta \sim \mathcal{N}(0,\sigma)$ is the large-scale shadow factor in dB; and $l^{\text{WSN}}(d_0) = 10\log_{10}((4\pi f d_0)/\text{c})^2$ is the path-loss at a reference distance, $d_0 = 1\,\text{m}$, for a given center frequency, $f \in \{800, 900, 2400\}\,\text{MHz}$; $\text{c} = 3\times10^8\,\text{ms}^{-1}$ is the speed of light; and $\bar{n} = 2$. The signal-to-noise ratio (SNR) at a distance d_i in the receiver, $\omega(d_i)$, is given by $\omega(d_i)[\text{dB}] = P_k^{\text{tx}} - l^{\text{WSN}}(d_i) - P^{\text{noise}}$, where P_k^{tx} [dBm] is the output power of the transmitter, and P^{noise} [dBm] denotes the total noise power at the receiver.

The transmission power and modulation have a direct impact on the bit error rate (BER). Given the link i, the BER Ψ_i is determined as a function of the adopted modulation technique, $mod_i \in \mathcal{M}$, and the SNR, $\omega(d_i)$, as $\Psi_i = \Psi(\omega(d_i), mod_i)$. Note that $\Psi(\cdot)$ is well-known for standard modulations. In the sensing domain, we consider simple modulation schemes following the IEEE 802.15.4 standard [5], such as BPSK and OQPSK, which are suitable for energy-limited WSNs.

2) Link layer functionalities: Concerning the channel coding scheme, we advocate for the use of a hybrid ARQ error control scheme [6], [7] that results from the combination of forward error correction (FEC) codes for poor quality channel conditions (i.e., $\omega(d_i)$ low) as well as the merits of automation repeat request (ARQ) when the channel conditions are good (i.e., $\omega(d_i)$ high). Initially, an uncoded or lightly coded packet is transmitted; if the received packet has more errors than those that can be corrected by the chosen FEC code, a more robust FEC code is chosen. We consider block codes due to their energy efficiency and lower complexity compared to convolutional codes (CCs). For the link i, $cod_i \in \mathcal{C}$ denotes the adopted coding scheme with coding rate RC_i. As far as the BCH$(bl; pl; ce)$ code with rate $RC_i = pl/bl$ is concerned, bl, pl, and ce denote block length, payload length, and the error correcting capability of FEC code in bits, respectively, and $ce < bl$. Given the BER $\Psi_i(\cdot)$, the block error rate, Ψ_i^{block}, becomes $\Psi_i^{\text{block}} = \sum_{j=ce+1}^{bl}\binom{bl}{j}\Psi_i(\cdot)^j(1-\Psi_i(\cdot))^{bl-j}$. Additionally, with $\wp$ bits being the packet length, the packet error rate (PER) Φ_i is calculated as follows

$$\Phi_i = 1 - (1-\Psi_i^{\text{block}})^{\lceil \frac{\wp}{pl} \rceil}, \tag{1}$$

which is approximated as $\lceil \frac{\wp}{pl} \rceil \Psi_i^{\text{block}}$ when Ψ_i^{block} is small. As a result, in each transmission, the initial packet is either coded with a BCH$(128; 106; 3)$ code or not coded to reduce the Φ_i without drastically sacrificing the transmission data rate. If the first transmission fails, i.e., the number of errors is larger than the maximum number of bits that can be corrected, a more robust FEC code is used for the re-transmitted packet [e.g., BCH$(128; 78; 7)$] until the packet is successfully decoded or the maximum number of transmissions (including re-transmissions), $N_i^{\max}$, is reached. Using this hybrid ARQ error control scheme, the overall PER over link i is given by

$$\Phi_i^{\text{Rtx}} = \Upsilon(\Phi_i^{\text{uncoded}}, N_i^{\text{Tx-ub}}, ce), \tag{2}$$

where $\Upsilon(\cdot)$ is a function that relates Φ_i after hybrid ARQ error control scheme, Φ_i^{Rtx}, with the uncoded PER over link i, Φ_i^{uncoded}, which is derived next [see (3)] considering the data storage capacity of nodes; $N_i^{\text{Tx-ub}}$ is the upper-bound for the number of transmissions of a packet with correctly decoding over link i computed as $N_i^{\text{Tx-ub}} = (1-\Phi_i^{\text{uncoded}})^{-1}$. Additionally, we take into account the data storage capacity of the sensor nodes, mem, that is related to the probability of discarding a packet at link i, $\mathbb{P}_i^{\text{pkt-dropout}}$, due to the fact that it can not be queued at the transmitter or at receiver. We define this probability as $\mathbb{P}_i^{\text{pkt-dropout}} = \Gamma(mem_k, F_k)$, where $\Gamma(\cdot)$ is a function that relates the maximum number of packets, mem_k, that can be queued at the transmitter or receiver and the total local traffic (own and relayed), F_k. For instance, assuming Poisson traffic, the transmitter and receiver can be modeled as a single server queue with mem_k buffer size and $(DR_i \cdot RC_i)/\wp$ [pkt/s] service rate, where DR_i [kbps] is the data rate transmission of link i. With these parameters, we determine the uncoded PER over link i, Φ_i^{uncoded}, as follows

$$\Phi_i^{\text{uncoded}} = (1-\mathbb{P}_i^{\text{pkt-dropout}})[1-(1-\Psi_i(\cdot))^{\wp})]. \tag{3}$$

Regarding the MAC functionality, we consider a variation of sleep MAC (SMAC) and carrier sense multiple access with collision avoidance (CSMA/CA) for addressing energy efficiency and scalability. On the one hand, as sensor nodes are likely to be battery powered, we adopt the idea of SMAC in which sensor nodes periodically listen and sleep [8] so that the network lifetime of these nodes is prolonged. On the other

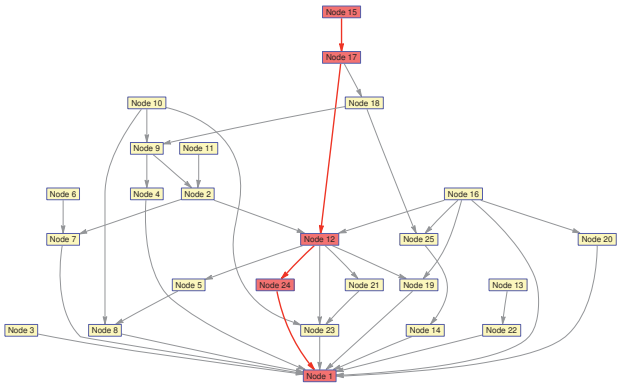


Figure 2. Directed Acyclic Graph (DAG) for a WSN consisting of 25 nodes randomly deployed and the optimal path (red color) from source to destination.

hand, with CSMA/CA, a node attempts to reserve the channel by using request to send/clear to send handshake after it sees the channel idle for an inter-frame space amount of time. If the node fails to reserve the medium, it switches to sleep mode to save energy and waits for the next listening cycle. This medium access method can eliminate the interference drastically if the carrier sensing is properly performed. Note that, if a reservation-based protocol is used, data packet collisions will not occur. Hence, the hybrid of SMAC and CSMA/CA MAC protocol can save energy as well as reduce the interference among the sensor nodes avoiding the degradation of both BER and PER [7]. In our framework, the duration of listen and sleep cycles ($T^{\text{listen}}, T^{\text{sleep}} = 9 \times T^{\text{listen}}$ for a 10% duty cycle) is adaptive to the QoS requirements and they are set the same for all nodes in one cluster. Note that, the longer the sleep duration is, the lower the idle energy consumption, but the longer the end-to-end delay. We consider this duration parameter in the MAC protocol to interplay with physical layer parameters in the proposed cross-layer framework.

3) Network layer functionality: The IoT is expected to have an incredibly high number of *things*, and each of them should be retrievable with a unique IP address. Thus, we advocate for the use of IPv6 addressing in our framework, consistently with 6LoWPAN. A packet with fixed size ($\wp$ bits) is selected and used for all the links throughout a given path; the packet size is computed as $\wp = pl + h + ce$, where pl is the payload (data) length, h is the header length, and ce is the FEC redundancy length. We use the RPL for selecting the multi-hop paths that data packets follow from the source to reach the destination. RPL is a distance vector routing protocol that leaves the process of route selection to an external mechanism called *objective function*. RPL is based on the topological concept of destination oriented directed acyclic graph (DODAG). The DODAG refers to a directed acyclic graph with a single root as shown in Fig. 2; the costs associated with each directional link are derived accordingly to be consistent with the objective function (detailed in Section III-C), the constraints, and the hardware capabilities of the *things* so that the optimal path from the source to destination is provided.

B. 5G Radio Access Network: SoftAir

Following the network model detailed in Section II-A, we formulate the sum-rate optimization in the SoftAir SD-RAN, which jointly optimizes associations between RRHs and SD-GW that use mmWave transmissions, and RRHs' beamforming weights to maximize the SD-GW sum-rate while guaranteeing QoS and system-level constraints. We consider a short frame structure [9], [10] where time is discretized into frames, each frame has duration of T^{frame} symbols. We allocate τ_{ul} symbols for uplink transmission, and τ_{dl} symbols for downlink transmission.

1) Association Scheme: Let $\mathcal{J} = \{1, \ldots, J\}$ and $\mathcal{I} = \{1, \ldots, I\}$ denote the set of RRHs and SD-GWs in the SoftAir system, respectively. Suppose that each SD-GW is served by a specific group of associated RRHs, and a RRH can serve multiple SD-GWs at the same time. To express the association status between RRHs and SD-GWs, we introduce the following binary variables as the indicators. Concretely, RRHs can be active to serve SD-GWs or shutdown to save the energy consumption, let $\{a_j, j \in \mathcal{J}\}$ denotes the activity of RRHs as $a_j = \mathbb{I}[\text{the } j\text{th RRH is in active mode}]$; let $\{g_{ij}, i \in \mathcal{I}, j \in \mathcal{J}\}$ denotes the association between RRHs and SD-GW as $g_{ij} = \mathbb{I}[\text{the } i\text{th SD-GW is served by the } j\text{th RRH}]$; furthermore, to characterize the group (cluster) of serving RRHs, let $\{N_{ij}, i \in \mathcal{I}, j \in \mathcal{J}\}$ be the clustering indicator as $N_{ij} = \mathbb{I}[(i,j) \in \mathcal{L}]$, where $\mathbb{I}[x]$ is the indicator function, $\mathcal{L} = \{(i,j) \mid i \in \mathcal{I}, j \in \mathcal{N}_i\}$ denotes the predetermined set of feasible association, and $\mathcal{N}_i$ denotes the set of near RRHs for the ith SD-GW which can be determined based on the distance or channel gain from RRHs to each SD-GW.

2) Millimeter-Wave Communication: We introduce the link budget for mmWave communication between the ith SD-GW and jth RRH. Particularly we detail the path-loss, l_i, channel vector, $\mathbf{h}_i$, and beamforming gain, G_i^{BF}, for deriving both the achievable uplink and downlink data rates.

Path-Loss: Considering the peculiarities of mmWave propagation, the path-loss for a mmWave communication link i, l_i, can be modeled with three link-states: outage (l_{iO}), LoS (l_{iL}) or NLoS (l_{iN}) [11]. We formulate the path-loss with respect to these three states as follows $l_{iO} = 0$, $l_{iL} = (\alpha_L d_i)^{-\beta_L}$, and $l_{iN} = (\alpha_N d_i)^{-\beta_N}$, where α_L (α_N) can be interpreted as the path-loss of the LoS (NoS) link at 1 [m] distance, and β_L (β_N) denotes the path-loss exponent of the LoS (NLoS) link. From experimental results [11], β_N value (can be up to 4) is normally higher than β_L value (i.e., 2). Then, each link-state is formulated by the channel state probabilities $\mathbb{P}_O$, $\mathbb{P}_L$, and $\mathbb{P}_N$, respectively, as $\mathbb{P}_O = \max(0, 1 - \gamma_O e^{-\delta_O d_i})$; $\mathbb{P}_L = (1 - \mathbb{P}_O)\gamma_L e^{-\delta_L d_i}$; $\mathbb{P}_N = (1 - \mathbb{P}_O)(1 - \gamma_L e^{-\delta_L d_i})$, where d_i denotes the transmitter-receiver distance; the parameters γ_L (γ_O) and δ_L (δ_O) depend on both the propagation scenario and the considered carrier frequency [12]. Thus, the corresponding path-loss component of the channel is modeled as $l_i = \mathbb{I}[U < \mathbb{P}_L(d_i)] l_{iL} + \mathbb{I}[\mathbb{P}_L(d_i) \leq U < (\mathbb{P}_L(d_i) + \mathbb{P}_N(d_i))] l_{iN} + \mathbb{I}[(\mathbb{P}_L(d_i) + \mathbb{P}_N(d_i)) \leq U \leq 1] l_{iO}$, where $U \sim \mathcal{U}[0,1]$ is a uniform random variable. For computing the path-loss model, we use the parameter values at 73 GHz as in [13, Table I].

Channel Vector: Given that the blockage information is not entirely feasible, we exploit the stochastic geometry analysis for modeling the mmWave channel vector [13]. Specifically, we model the channel vector as $\mathbf{h}_i = \sqrt{l_i \boldsymbol{\beta}_i}\, \boldsymbol{\xi}_i \in \mathbb{C}^{M,1}$, where l_i is the large-scale path-loss in power of the mmWave communication link i (which might also include log-normal shadowing), $\boldsymbol{\beta}_i \in \mathbb{C}^{M,M}$ is the co-variance matrix for antenna correlations in small-scale fading, and $\boldsymbol{\xi}_i \in \mathbb{C}^{M,1}$ is a Gaussian vector with the zero-mean circularly symmetric Gaussian noise distribution $\mathcal{CN}(0, \mathbf{I}_M)$ for the fast-fading.

Beamforming: To ensure an acceptable range of the communication in the multi-antenna mmWave transmissions, we introduce the precode vectors, i.e., beamforming weights at the RRHs, where the weight vector $\mathbf{w}_i \in \mathbb{C}^{M,1}$ is the linear downlink beamforming vector at the jth RRH corresponding to the ith SD-GW. The beamforming gain is given as $G_i^{\text{BF}} =$

$\mathbf{w}_i^{\mathsf{H}}\boldsymbol{\beta}_i\mathbf{w}_i$, with $\boldsymbol{\beta}_i$ being the covariance matrix of the channel response vector $\mathbf{h}_i$. In the case where the fading is fully correlated between the antennas, the matched filtering precoding method is exploited as $\boldsymbol{\beta}_i = \mathbf{h}_i^{\mathsf{H}}\mathbf{h}_i$ and $\mathbf{w}_i = \mathbf{h}_i/\|\mathbf{h}_i\|$; therefore, $G_i^{\text{BF}} = \|\mathbf{h}_i\|^2$.

3) Achievable Uplink Rate: Following the above multi-antenna mmWave transmission characterization over a link i, the received base-band signal vector $\mathbf{y} \in \mathbb{C}^{M,1}$ at the BBS at a given instant reads $\mathbf{y}^{\text{ul}} = \sqrt{P^{\text{ul}}}\ \mathbf{H}\mathbf{x}^{\text{ul}} + \eta^{\text{ul}}$, where each element of the received signal vector corresponds to a BBS antenna, $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_I] \in \mathbb{C}^{M,I}$, $\mathbf{h}_i \in \mathbb{C}^{M,1}$ denotes the mmWave channel corresponding to the ith SD-GW, $\mathbf{x} = [x_1 \cdots x_I]^{\mathsf{T}}$ denotes the I×1 vector containing the transmitted signals from all the SD-GWs, P^{ul} is the average transmit power of each SD-GW, and $\eta^{\text{ul}} \sim \mathcal{CN}(0, \sigma)$ is the zero-mean circularly symmetric Gaussian noise with the noise power σ^2.

Let $\mathbf{A}$ be the M×I linear detection matrix (which depends on the channel matrix $\mathbf{H}$) used by the BBS $b \in \mathcal{B}$ to separate the received signal into user streams. The BBS processes its received signal vector and obtains the estimated channel matrix (assuming no estimation errors) by multiplying the detection matrix with the Hermitian-transpose of the linear receiver as $\tilde{\mathbf{y}}^{\text{ul}} = \mathbf{A}^{\mathsf{H}}\mathbf{y}^{\text{ul}} = \mathbf{A}^{\mathsf{H}}\mathbf{H}\mathbf{x} + \mathbf{A}^{\mathsf{H}}\eta^{\text{ul}}$. The ith element of $\tilde{\mathbf{y}}^{\text{ul}}$ can be written as $\tilde{y}_i^{\text{ul}} = \sqrt{P_i^{\text{ul}}}\mathbf{a}_i^{\mathsf{H}}\mathbf{H}\mathbf{x} + \mathbf{a}_i^{\mathsf{H}}\eta^{\text{ul}}$, where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$. By the elements multiplication, we further get $\tilde{y}_j^{\text{ul}} = \sqrt{P_i^{\text{ul}}}\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_i x_i + \sum_{k=1,k\neq i}^{I}\sqrt{P_k^{\text{ul}}}\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_i x_k + \mathbf{a}_i^{\mathsf{H}}\eta^{\text{ul}}$, where x_i denotes the ith element of $\mathbf{x}$ and $\mathbf{h}_i$ is the ith column of $\mathbf{H}$. Then, the signal-to-interference-plus-noise ratio (SINR) achieved by the ith SD-GW, γ_i^{ul}, is

$$\gamma_i^{\text{ul}} = P_i^{\text{ul}}|\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_i|^2/(\textstyle\sum_{k=1,k\neq i}^{I} P_k^{\text{ul}}|\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_k|^2 + \|\mathbf{a}_i\|^2\sigma^2). \quad (4)$$

Assuming an ergodic channel [14], the achievable uplink rate of the ith SD-GW is given by $R_i^{\text{ul}} = B\log_2(1+\gamma_i^{\text{ul}})$, where B denotes the wireless transmission bandwidth. We define the uplink sum rate [bits/s/Hz] per cell considering the associations between RRHs and SD-GWs as follows

$$C^{\text{ul}} = \textstyle\sum_{i=1}^{I} g_{ij}N_{ij}R_i^{\text{ul}}, \forall j \in \mathcal{J}. \quad (5)$$

4) Achievable Downlink Rate: The received base band signal $y^{\text{dl}} \in \mathbb{C}$ at the ith SD-GW is given as $y^{\text{dl}} = \sqrt{P_j^{\text{dl}}}\mathbf{h}_i^{\mathsf{H}}\,\mathbf{s} + \eta^{\text{dl}}$, where $\mathbf{s} \in \mathbb{C}^{M,1}$ is the signal vector intended for the ith SD-GW with P_j^{dl} average power; $\eta^{\text{dl}} \sim \mathcal{CN}(0, \sigma^2)$ is the receiver noise. We assume channel reciprocity, i.e., the downlink channel $\mathbf{h}_i^{\mathsf{H}}$ is the Hermitian transpose of the uplink channel $\mathbf{h}_i$. The transmit vector $\mathbf{s}$ is given as $\mathbf{s} = \sqrt{\upsilon}\sum_{i=1}^{I}\mathbf{w}_i x_i^{\text{dl}} = \sqrt{\upsilon}\,\mathbf{W}\,\mathbf{x}^{\text{dl}}$, where $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_I] \in \mathbb{C}^{M,I}$ is a pre-coding matrix (i.e. the network beamforming design) and $\mathbf{x}^{\text{dl}} = [x_1 \cdots x_I]^{\mathsf{T}} \in \mathbb{C}^{I,1}$ contains the data symbols for the ith SD-GW. The parameter υ normalizes the average transmit power per RRH to $\mathbb{E}[\frac{P_j^{\text{dl}}}{I}\mathbf{s}^{\mathsf{H}}\mathbf{s}] = P_j^{\text{dl}}$, i.e., $\upsilon = \left(\mathbb{E}\left[\frac{1}{I}\text{tr}(\mathbf{W}\mathbf{W}^{\mathsf{H}})\right]\right)^{-1}$.

The associated SINR achieved by the ith SD-GW, γ_i^{dl}, is

$$\gamma_i^{\text{dl}} = \upsilon|\mathbf{h}_i^{\mathsf{H}}\mathbf{w}_i|^2/\left(\textstyle\sum_{k=1,k\neq i}^{I}\upsilon\,|\mathbf{h}_i^{\mathsf{H}}\mathbf{w}_k|^2 + \sigma^2\right). \quad (6)$$

Since the SD-GWs do not have any channel estimate, we provide an ergodic achievable rate based on the techniques developed in [14, Theorem 1] as $R_i^{\text{dl}} = B_i(1-\kappa)\log_2(1+\gamma_i^{\text{dl}})$, where B_i is the bandwidth allocated to the ith SD-GW, κ accounts for the spectral efficiency loss due to signaling at RRH. The downlink sum rate [bits/s/Hz] per cell considering the associations between RRHs and SD-GWs is

$$C^{\text{dl}} = \textstyle\sum_{i=1}^{I} g_{ij}N_{ij}R_i^{\text{dl}}, \forall j \in \mathcal{J}. \quad (7)$$

C. Optimization Framework

In the following, we elaborate the optimization framework for the sensing and SD-RAN domain in Section III-C1 and Section III-C2, respectively.

1) IoT Sensing Domain: The IoT should provide differentiated services for applications with different QoS requirements, ranging from error-limited applications or minimum energy consumption applications to highly-delay-sensitive applications or any combination of them. Hence, we consider a multi-objective optimization problem which can simultaneously optimize multiple conflicting end-to-end objectives such as PER (Φ^{e2e}), delay (T^{e2e}), and energy consumption (E^{e2e}), subject to certain constraints.

We construct a single aggregate objective function which is defined by the weighted linear combination of each objective; we use w_{PER}, w_{E}, and w_{T} as the three weights for the end-to-end PER, energy consumption, and time delay objectives, respectively. As these three objectives differ in the units in which they are measured as well as their order of magnitude, we normalize each term and optimize their deviations with respect to some predefined utopia values (unattainable minimum values which are used to provide the non-dimensional objective functions and can be computed offline [7]). Therefore, the overall objective function for WSN communication becomes

$$\text{minimize} \quad w_{\text{PER}}|\tfrac{\Phi^{\text{e2e}}}{\Phi^{\text{opt}}} - 1| + w_{\text{E}}|\tfrac{E^{\text{e2e}}}{E^{\text{opt}}} - 1| + w_{\text{T}}|\tfrac{T^{\text{e2e}}}{T^{\text{opt}}} - 1|, \quad (8)$$

where $w_{\text{PER}} + w_{\text{E}} + w_{\text{T}} = 1$; Φ^{opt}, E^{opt}, T^{opt} are the end-to-end PER, energy consumption, and delay utopia values for normalizing purposes, respectively. Note that (8) may target at different degrees of QoS requirements for various IoT applications by adapting the specific weight value (w_{PER}, w_{E}, or w_{T}) according to the application.

Statistical QoS Guarantee: The higher transmission reliability associated with lower PER is crucial for almost all types of WSN. Also, having a bounded delay is especially important for real-time monitoring and applications with timing constraints. Aiming to support the distributed functionalities among sensors, in the following we form the per-node based constraints (i.e., for transmissions upon link i) of link reliability, delay, and energy.

Given the tolerable maximum end-to-end PER, Φ^{TH}, the corresponding reliability constraint is

$$\Phi^{\text{e2e}} = \left(1 - (1 - \Phi_i^{\text{Rtx}})^{N^{\text{hops}}}\right) \leq \Phi^{\text{TH}}, \quad (9)$$

where $(1 - \Phi_i^{\text{Rtx}})^{N^{\text{hops}}}$ represents the PER of multi-hop transmission; Φ_i^{Rtx} [see (2)] is the PER over link i with hybrid ARQ error control, and N^{hops} is the number of traversed hops for an incoming packet to node k.

Regarding the energy consumption, let E_k denote the energy consumed on the kth node, it is defined by the product of the packet size and the energy required for one bit as $E_k = \wp(2E_{\text{elec}}^{\text{bit}} + P_k^{\text{tx}}/F_k)$, where $E_{\text{elec}}^{\text{bit}} = E_{\text{elec}}^{\text{bit-Tx}} = E_{\text{elec}}^{\text{bit-Rx}}$ in Joule/bit is the distance-independent energy to transmit one bit; $E_{\text{elec}}^{\text{bit-Tx}}$ is the energy per bit needed by the transmitter electronics, and $E_{\text{elec}}^{\text{bit-Rx}}$ is the energy per bit utilized by the receiver electronics; P_k^{tx} and F_k are the transmission power and the total local traffic at the kth node, respectively. Restricted by the constraint E^{TH}, the overall energy consumption over the entire path is computed by

$$E^{\text{e2e}} = \textstyle\sum_{k=1}^{N^{\text{hops}}} E_k \leq E^{\text{TH}}. \quad (10)$$

Finally, restricted by the maximum end-to-end delay T^{TH}, the statistical delay guarantee is modeled as the probability

that a packet is delivered under the deadline should be at least φ as follows

$$\mathbb{P}(T^{\text{e2e}} \leq T^{\text{TH}}) \geq \varphi. \tag{11}$$

The end-to-end delay, T^{e2e}, is calculated as $T^{\text{e2e}} = \sum_{i=1}^{N^{\text{hops}}} (T_i^{\text{queuing}} + T_i)$, where T_i^{queuing} is the queuing delay at link i and T_i is the delay at link i excluding the queuing delay. T_i is composed of the time for handshake $T_i^{\text{handshake}}$, time for data transmission T_i^{data}, timeout delay T_i^{timeout}, time for acknowledgment T_i^{ack}, sleep time T^{sleep}, and the signal processing time T_i^{DSP}, and is calculated as $T_i \leq (T_i^{\text{handshake}} + T_i^{\text{data}} + T_i^{\text{timeout}})(N_i^{\text{Tx-ub}} - 1) + (T_i^{\text{handshake}} + T_i^{\text{data}} + T_i^{\text{ack}}) + T^{\text{sleep}} + T_i^{\text{DSP}}$. Note that the queuing delay is determined by many factors such as current traffic, other nodes' behavior or hardware status. Therefore, the overall end-to-end delay is modeled, by applying the central limit theorem, as a Gaussian random variable $T^{\text{e2e}} \sim \mathcal{N}\left(\sum_{i=1}^{N^{\text{hops}}} \left(T_i + \mathbb{E}[T_i^{\text{queuing}}]\right), \left(\sum_{i=1}^{N^{\text{hops}}} \mathbb{V}[(T_i^{\text{queuing}})]\right)^{1/2}\right)$. Then, the end-to-end delay constraint (11) is transformed into

$$\sum_{i=1}^{N^{\text{hops}}} \left(T_i + \mathbb{E}[T_i^{\text{queuing}}]\right) + \phi^{-1}(\varphi)\left(\sum_{i=1}^{N^{\text{hops}}} \mathbb{V}[(T_i^{\text{queuing}})]\right)^{1/2} \leq T^{\text{TH}}, \tag{12}$$

where $\phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. Finally, the end-to-end throughput, G^{e2e}, is inversely proportional to the end-to-end delay as $G^{\text{e2e}} = pl/T^{\text{e2e}}$.

2) SoftAir SD-RAN Domain: As IoT applications demand services with different rate requirements, we formulate these requirements in terms of SINR coverage and achieved sum-rate per cell at the SD-RAN. Given ϑ as the minimum tolerable SINR over a link i, the SINR constraints of SD-GWs can be formulated as

$$\gamma_i \geq \vartheta, \forall i \in \mathcal{I}, \tag{13}$$

where γ_i is computed by either (4) or (6) in case of uplink or downlink transmission, respectively. From the association scheme, we can obtain the equality $a_j = 1 - \prod_{j=1}^{J}(1 - g_{ij}N_{ij}), \forall j \in J$ and the following sets of association constraints between RRHs and SD-GW:

$$a_j \geq g_{ij}N_{ij}, \forall i \in \mathcal{I}, j \in \mathcal{J}; \tag{14}$$

$$\sum_{j=1}^{J} g_{ij}N_{ij} \geq 1, \forall i \in \mathcal{I}, \tag{15}$$

where (14) implies that a RRH is in active mode if it is associated with at least one SD-GW whereas (15) ensures that each SD-GW is served by at least one RRH. On the other hand, given the pre-coding vector at the jth RRH for the ith SD-GW, the transmitter power used by this RRH to serve the ith SD-GW is $\mathbf{w}_i^{\mathsf{H}}\mathbf{w}_i$ [15]. Let $P_j^{\text{r-max}}$ denote the maximum power of the jth RRH, we impose the constraints on RRHs' downlink beamforming weights as follows

$$\sum_{i=1}^{I} \mathbf{w}_i^{\mathsf{H}}\mathbf{w}_i \leq a_j P_j^{\text{r-max}}, \forall j \in \mathcal{J}; \tag{16}$$

$$\mathbf{w}_i^{\mathsf{H}}\mathbf{w}_i \leq g_{ij}N_{ij}P_j^{\text{r-max}}, \forall i \in \mathcal{I}, j \in \mathcal{J}, \tag{17}$$

where (16) limits the total transmit power of RRHs and (17) ensures that the transmit power from the jth RRH to the ith SD-GW is set to zero if there is no association between them. Furthermore, by only allowing the links in $\mathcal{L}$ (see Section III-B1) we set the beamforming weights of mmWave communication links as

$$\mathbf{w}_i^{\mathsf{H}}\mathbf{w}_i = 0 \text{ if } N_{ij} = 0, \forall i \in \mathcal{I}, j \in \mathcal{J}, \tag{18}$$

so that we reduce all possible links between J RRHs and I SD-GWs to $|\mathcal{L}|$ links (given that $|\mathcal{L}| \ll JI$), which in turns dramatically shrinks the possible solution sets of precoding vectors for lower computation complexity [13], [15]. Additionally, the per-fronthaul capacity constraints (neglecting the fronthaul capacity consumption for transferring compressed beamforming vector) are formulated as follows

$$C \leq C_j^{\text{fh}}, \forall j \in \mathcal{J}, \tag{19}$$

where C is computed by (5) in uplink transmission or by (7) in downlink transmission. This indicates that the total data rate transmitted at the jth RRH should be less or equal to the rate forwarded by the jth fronthaul link.

We aim to maximize the total achievable uplink/downlink data rates at SD-GWs; the overall objective function for the SD-RAN becomes

$$\text{maximize} \quad C = \sum_{i=1}^{I} R_i, \tag{20}$$

where R_i depends on the communication direction: uplink (see Section III-B3), downlink (see Section III-B4).

Statistical QoS Guarantee: To ensure low transmission delay, the size of each packet, $\wp$ bits, is small enough such that it can be transmitted within one uplink phase; the transmission time interval (TTI) is the same as the frame duration, T^{frame}, and $T^{\text{frame}} \ll T^{\text{TH}}$. Thus, the uplink (downlink) transmission can be finished within the duration of τ^{ul} (τ^{dl}). Furthermore, the expected queuing delay for the packets at the SD-GW should be bounded as

$$\mathbb{E}[T_i^{\text{queuing}}] \leq T^{\text{TH}} - T^{\text{frame}}. \tag{21}$$

Then, to guarantee the stringent QoS requirements for all SD-GW $i \in \mathcal{I}$, our cross-layer design satisfies that: (i) the probability that the queuing delay is larger than $(T^{\text{TH}} - T^{\text{frame}})$ is smaller than a predefined violation probability D^{TH}, i.e., $\mathbb{P}(T_i^{\text{queuing}} > (T^{\text{TH}} - T^{\text{frame}})) < D^{\text{TH}}$; (ii) with finite BCH codes, the transmission of each packet is finished within one frame with a small error probability, i.e., $\Phi_i^{\text{e2e}} \leq \Phi^{\text{TH}}$; (iii) to guarantee the end-to-end delay and its reliability with finite transmit power, the packet dropout probability, $\mathbb{P}_i^{\text{pkt-dropout}}$ is smaller than a predefined violation probability Q^{TH}; (iv) the probability that the SINR coverage is smaller than ϑ is smaller than a predefined violation probability ϑ^{TH}, i.e., $\mathbb{P}(\gamma_i < \vartheta) < \vartheta^{\text{TH}}$. Finally, the end-to-end system reliability is controlled by

$$1 - (1 - D^{\text{TH}})(1 - \Phi^{\text{TH}})(1 - Q^{\text{TH}})(1 - \vartheta^{\text{TH}}) \leq \Omega, \tag{22}$$

where Ω dictates the overall reliability requirement. The entire formulation of the joint cross-layer optimization for SD-GWs is summarized in Table I.

D. Protocol Operation

The SD-GWs possess the necessary knowledge (e.g., network topology, link qualities) to orchestrate sensors at the southbound interface and the application requirements at the northbound interface. Therefore, they are able to receive IoT data traffic from the sensing devices and forward this traffic to the SoftAir SD-RAN. Depending on the communication direction, each SD-GW will either perform protocol conversions in such a way it can forward the data to the SoftAir system with the maximum achievable rate or forward the data to the WSN meeting the application QoS requirements by performing the optimization framework. The SD-GW first builds the hierarchical topology (DODAG) that specifies a route from each node to itself using a DODAG Information Object (DIO) message. Once a node receives a DIO, it can first, calculate its rank, then, choose a set of parent nodes

Table I
HETEROGENEOUS CROSS-LAYER OPTIMIZATION FRAMEWORK

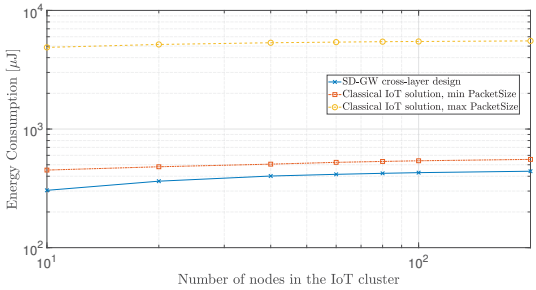
Inputs:	
Sensor domain:	Φ^{opt}, E^{opt}, T^{opt}, Φ^{TH}, E^{TH}, T^{TH}, D^{TH}, Q^{TH}, h, N_k^{max}, mem_k, $\forall k \in \mathcal{K}$
Cellular domain:	$P_j^{\text{r-max}}$, C_j^{fh}, ϑ^{TH}, Ω, T^{frame}, $\forall j \in \mathcal{J}$
Compute (offline):	w_{PER}, w_{E}, w_{T}, $N_k^{\text{Tx-ub}}$, F_k, $\forall k \in \mathcal{K}$
Find:	
Sensor domain:	P_k^{tx}, mod_i, cod_i, $\wp$, T^{listen}, $\forall k \in \mathcal{K}, i \in \mathcal{I}$
Cellular domain:	P_i^{ul}, P_j^{dl}, a_j, g_{ij}, $\mathbf{w}_i$, $\forall i \in \mathcal{I}, j \in \mathcal{J}$
Objectives:	
minimize	$w_{\text{PER}}\lvert\frac{\Phi^{\text{e2e}}}{\Phi^{\text{opt}}} - 1\rvert + w_{\text{E}}\lvert\frac{E^{\text{e2e}}}{E^{\text{opt}}} - 1\rvert + w_{\text{T}}\lvert\frac{T^{\text{e2e}}}{T^{\text{opt}}} - 1\rvert$ (8)
maximize	$C = \sum_{i=1}^{I} R_i$. (20)
Subject to:	
Packet error rate constraints:	(9), (3), (2).
Energy consumption constraint:	(10).
Delay constraints:	(21), (12).
SINR constraints:	(13); (4) uplink, (6) downlink.
Association constraints:	(15), (14).
Beamforming weights constraints:	(18), (17), (16).
Per-fronthaul capacity constraint:	(19).
System reliability constraint:	(22).

(candidate nodes where data can be forwarded) and finally, send a new DIO message to inform other neighbors. The rank is an integer that increases linearly from the SD-GW and identifies the position of a node about the SD-GW and other nodes in the network. The parents must have a rank equal or lower than the node. Each node has a default path (i.e., preferred parent) but maintains a list of parents for resilience purposes, overhead reduction in case of link degradation, or increasing performance. Hence, immediately after the reception of a DIO message, a node has an optimal path towards the SD-GW. The optimal path is set according to the optimization framework detailed in Section III-C. The SD-GW translates the application requirements into network QoS requirements and constructs the optimal architecture finding the optimal communication parameters on both the sensor and the SD-RAN domains and forwarding the data through the correspondent interface. The SD-GWs determine the routing paths at the local level, and at the intercluster level, cluster coordinators, who are elected by the transmission algorithm, facilitate the communication.

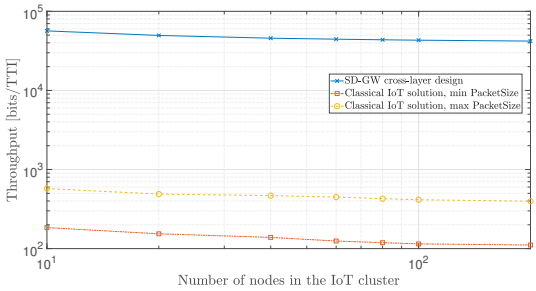
IV. PERFORMANCE EVALUATION

In this section, we present simulation results to assess and compare the performance of the proposed cross-layer design detailed in Section III with that of conventional layered protocol solutions, i.e., individual communication functionalities that do not share information and operate in separate layers. In all experiments, each point represents the average value of 10^5 samples. The overall reliability requirement is $\Omega = 10^{-5}$.

Following the system model (see Section II-A), we design a set $\mathcal{J}$ of $J = 12$ associated RRHs in the SD-RAN, each RRH is equipped with $M = 4$ antennas; the coverage area of every RRH has a radius of 200 m. The channel vectors are generated according the mmWave communication characterization detailed in Section III-B2, where the three-state path-loss model with log-normal shadowing is considered; the carrier frequency is set at 73 GHz. The transmit power constraint for each RRH is $P_j^{\text{r-max}} = 45$ dBm. Moreover, we assume that all the RRHs possess the same fronthaul capacity, i.e., $C_j^{\text{fh}} = 6$ bps/Hz, since 64 QAM is the highest constellation supported by the network, and thus the maximum spectrum efficiency per data stream is 6 bps/Hz. The bandwidth of the wireless link is $B = 500$ MHz. In the sensing domain,



(a)



(b)

Figure 3. IoT performance metrics vs. number of nodes in the network for the proposed design and the classical IoT solution. (a) Energy Consumption. (b) Throughput.

the set $\mathcal{J}$ of associated RRHs serve a set $\mathcal{I}$ of clustered sensor nodes. The ith cluster has one SD-GW as the cluster head. The maximum transmission power of each SD-GW is set at 23 dBm and thermal noise power is assumed to be -101 dBm/Hz. The sensor nodes inside the cluster are randomly deployed; the concerned coverage area of each sensor node has a radius of 50 m. Each node uploads packets with rate 0.02 packets/TTI; the TTI has set to 0.1 ms. We compare the results when the QoS requirements are focused on end-to-end delay minimization and energy consumption minimization while the PER is constrained to be below $\Phi^{\text{TH}} = 10^{-6}$. The packet sizes that we consider in the simulations are $\wp = \{20, 40, 100, 133\}$ bytes.

We first examine the interactions between link layer and routing functionalities via end-to-end energy consumption and throughput performance of one IoT cluster at the sensing domain and compare the results of our design with that of a classical IoT communication solution. Then, we examine both the sum-rate and the achievable rate per SD-GW in the SD-RAN as the number of IoT clusters increases. For this, we consider that each cluster has a fixed number of sensor nodes and one SD-GW as cluster head. A layered protocol architecture is built for the comparison with our proposed design; its configuration is the following:

Classical IoT Solution (sensing domain: IEEE 802.15.4 + RPL; SD-RAN domain: conventional association schemes used in mmWave [12], [13]) In the sensing domain, this protocol configuration follows the frequency spectrum allocation according to the IEEE 802.15.4 standard at 2400 MHz (OQPSK modulation, 250 kbps transmission rate) and Sleep MAC + CSMA/CA for the PHY and link layer, respectively. In the NET layer, this protocol uses RPL and, for a fair comparison, the objective function is similar to that of our design (focused on the minimization of end-to-end delay and energy consumption while the PER is constraint to be below $\Phi^{\text{TH}} = 10^{-6}$). In the SD-RAN, the following association

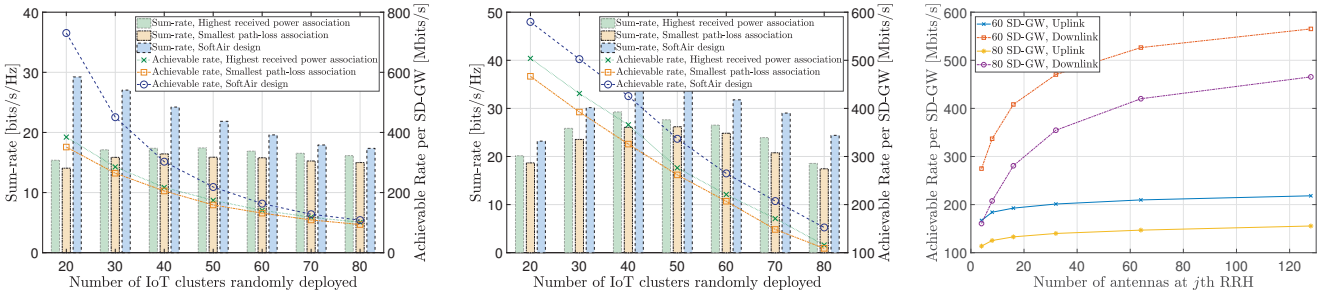


Figure 4. Sum-rate and achievable rate vs. number of SD-GWs deployed for the SoftAir design and conventional association schemes in mmWave; (a) upstream transmissions, (b) downstream transmissions. (c) Impact of increasing the number of antenna elements at RRHs on SD-GWs' achievable rate.

schemes used in conventional mmWave communication are configured: (i) highest received power association and (ii) smallest path-loss association. Although the described layered protocol applies previously proposed functionalities more or less, it only considers its related layers without information sharing but with reasonable assumptions for the other layers.

Fig. 3a and Fig. 3b show, in logarithmic scale, the end-to-end energy consumption [μJ] and throughput [bits/TTI], respectively, as a function of the number of nodes in the cluster. In Fig. 3a we observe that the energy consumption of the SD-GW design is always lower than that of the classical IoT solution; on average, the energy savings of our solution ranges from 22.6% up to 92.5%. Also, we can observe that the energy consumption increases gradually as the number of nodes in the network increases since the higher node density essentially creates more paths for the data transmission. It is evident that using large packet sizes imply the highest levels in energy consumption. Fig. 3b shows the significant improvement of throughput with the SD-GW cross-layer design. The reason is that our solution selects the optimum path from the source to destination and the best parameter configuration for the device such as power, modulation, coding scheme, packet size.

On the other hand, we analyze the achievable rate per SD-GW in the SD-RAN domain, the set $\mathcal{J}$ of associated RRHs serve different densities of IoT clusters; each IoT cluster has 100 nodes and one SD-GW as cluster head. In upstream transmissions, Fig. 4a shows that our design outperforms conventional association schemes used in mmWave transmissions with coordinated multi-point. On average, the spectral efficiency of our design is 40% higher than that of other association schemes and the achieved data-rate is up to 54% higher than that of conventional solutions. Regarding downstream transmissions, Fig. 4b shows that, on average, our design outperforms 26% conventional association schemes in terms of spectral efficiency. Furthermore, the spectral efficiency peaks at 34 bits/s/Hz; then, it slightly declines as the number of SD-GW increases. Although the achievable data rate per SD-GW decreases with the increasing SD-GW density, our architecture can provide high data-rate for each SD-GW by increasing the number of antennas at the RRHs. This fact is shown in Fig. 4c for high densities of SD-GWs (60 and 80); specifically, with an antenna array of 128 elements at the RRHs and 80 SD-GWs deployed in such an area, our design can support each SD-GW with at least 450 [Mbits/s] rate in downlink and 150 [Mbits/s] rate in uplink through mmWave transmissions.

V. Conclusion

We presented a SoftAir architecture for providing IoT communication by exploiting a set of emerging features such as mmWave and SDN. Our solution brings significant system synergies by jointly optimizing functionalities in different communication layers for both IoTs and SD-RANs; SD-GWs are proposed to (i) explore the interactions of two-type networks, (ii) enable cross-layer solutions, and (iii) render efficient energy consumption and throughput in IoT, while maximizing the sum-rate at the SD-RAN for reliable IoT communication. Simulation results validate the superiority of our solutions that provide performance improvements in terms of energy savings, throughput, and spectral efficiency in comparison with conventional IoT solutions. It allows enormous and reliable IoT connectivity with high data rates at 5G SD-RANs.

References

[1] Cisco. (2017, March) Cisco visual networking index (VNI): Global mobile data traffic forecast update, 2016-2021.

[2] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17 – 48, 2016.

[3] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: An overview," *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, 2017.

[4] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1 – 18, 2015.

[5] I. S. Association. IEEE Standard for Low-Rate Wireless Networks. [Online]. Available: http://www.ieee802.org/15/pub/TG4.html

[6] M. C. Vuran and I. F. Akyildiz, "Error control in wireless sensor networks: A cross layer analysis," *IEEE/ACM Transactions on Networking*, vol. 17, no. 4, pp. 1186–1199, Aug 2009.

[7] C. Han, J. M. Jornet, E. Fadel, and I. F. Akyildiz, "A cross-layer communication module for the internet of things," *Computer Networks*, vol. 57, no. 3, pp. 622 – 633, 2013.

[8] W. Ye, J. Heidemann, and D. Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 3, pp. 493–506, June 2004.

[9] E. Lähetkangas, K. Pajukoski, J. Vihriälä, G. Berardinelli, M. Lauridsen, E. Tiirola, and P. Mogensen, "Achieving low latency and energy consumption by 5G TDD mode optimization," in *2014 IEEE International Conference on Communications Workshops (ICC)*, June 2014, pp. 1–6.

[10] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer transmission design for tactile internet," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[11] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[12] M. D. Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5038–5057, Sept 2015.

[13] S.-C. Lin and I. F. Akyildiz, "Dynamic base station formation for solving NLOS problem in 5G millimeter-wave communication," in *IEEE Conference on Computer Commun. (INFOCOM)*, may 2017, pp. 1–9.

[14] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, August 2011.

[15] V. N. Ha, L. B. Le, and N. D. Dao, "Coordinated multipoint transmission design for cloud-rans with limited fronthaul capacity constraints," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sept 2016.