# Capacity Scaling of Cellular Networks: Impact of Bandwidth, Infrastructure Density and Number of Antennas

Felipe Gómez-Cuba, *Member, IEEE*, Elza Erkip, *Fellow, IEEE*, Sundeep Rangan, *Fellow, IEEE*, and Francisco Javier González-Castaño

*Abstract*—The availability of very wide spectrum in millimeter wave bands combined with large antenna arrays and ultra-dense networks raises two basic questions: What is the true value of overly abundant degrees of freedom and how can networks be designed to fully exploit them? This paper determines the capacity scaling of large cellular networks as a function of bandwidth, area, number of antennas, and base station density. It is found that the network capacity has a fundamental bandwidth scaling limit, beyond which the network becomes power-limited. An infrastructure multi-hop protocol achieves the optimal network capacity scaling for all network parameters. In contrast, current protocols that use only single-hop direct transmissions cannot achieve the capacity scaling in wideband regimes except in the special case when the density of base stations is taken to impractical extremes. This finding suggests that multi-hop communication will be important to fully realize the potential of next-generation cellular networks. Dedicated relays, if sufficiently dense, can also perform this task, relieving user nodes from the battery drain of cooperation. On the other hand, more sophisticated strategies such as hierarchical cooperation, that are essential for achieving capacity scaling in *ad hoc* networks, are unnecessary in the cellular context.

*Index Terms*—Wideband regime, capacity scaling laws, cellular networks.

## I. Introduction

To meet the tremendous growth in demand for cellular wireless data, three new design approaches are widely-considered

for the evolution of next-generation systems [3]:

- Vast spectrum available at very high frequencies, esp. the millimeter wave [4]–[8];
- Massive Multiple-input Multiple-output (MIMO) for increased spatial multiplexing [9], [10];
- Ultra dense deployments of small pico- and femtocells [11], [12].

Together, these technologies offer the potential of orders of magnitude increases in capacity, and, if successful, may fundamentally change the basic constraints that dictate network design today. This possibility leads to two basic questions: What is the fundamental capacity offered by these technologies and how can networks be best designed to fully leverage their potential?

From an information theoretic perspective, millimeter wave transmissions, massive MIMO and ultra-dense deployments are all, in essence, various ways to increase the fundamental degrees of freedom of the network which are controlled by bandwidth, number of antennas and infrastructure density respectively. This paper attempts to characterize the capacity scaling of cellular networks as a function of the scaling of these dimensions. Our analysis follows along the lines of the classic result of Gupta and Kumar [13], but applied to cellular networks rather than *ad hoc* networks with or without infrastructure. Specifically, we consider a large cellular network with $n$ mobile nodes, where the key parameters such as bandwidth, number of antennas, area, and base stations (BSs), all scale as functions of $n$. In addition, traffic in this network travels between each one of the BSs and the nodes in its cell, in separate uplink and downlink phases.

Our main results determine the capacity scaling by finding identically-scaling lower and upper bounds on the throughput. The upper bound is a series of cut-set bounds in which one transmitter is cut from the rest of the network, and all the nodes and BSs in the other side of the cut cooperate perfectly, forming a virtual point-to-point MIMO system where all devices contribute to receive power and all interference is perfectly canceled. The capacity scaling achieving lower bound is found by considering a simple infrastructure multi-hop (IMH) protocol where transmissions are relayed to/from the closest BSs via mobile nodes within the same cell.

We also study the capacity scaling of two additional protocols: The first one is the infrastructure single-hop (ISH) protocol, where transmissions are sent directly between the

BS and each node within its cell, and which is the dominant paradigm in current cellular networks. The second one is the infrastructure relay multi-hop (IRH), modeled after existing two-layer network architectures, where IMH is used for wireless backauling of additional access points called relay nodes (RN), while user nodes only communicate directly with a single nearest access point using ISH to prevent multi-hop implementation difficulties due to mobility, reticence to cooperation, and backwards compatibility.

Our analysis yields several important and in some cases surprising findings:

- *Bandwidth scaling limit:* There is a "critical bandwidth scaling" that defines a maximum useful bandwidth for the whole network. Below the critical point, the capacity scales with the bandwidth, whereas if bandwidth grows faster than its critical limit the capacity becomes power-limited and additional bandwidth growth no longer improves the capacity scaling. Power and bandwidth limited regimes are well-understood for point-to-point channels, and our results provide a generalization to cellular networks.

- *Benefits of increased cell density:* The network capacity always grows with the BS density, whereas the benefits of increased bandwidth or number of BS antennas have a limit. This is valid as long as nodes are sufficiently separated to experience far-field propagation.

- *Interference alignment is not necessary:* Our upper bound implicitly avoids inter-cell interference, whereas our lower bound IMH simply treats interference as noise. Since both have the same scaling, we can conclude that interference-alignment schemes, despite providing significant gains in a non-asymptotic regime [14], do not alter the capacity scaling significantly. On the other hand, our analysis does not discard that BS cooperation, achieved for example by a wired backhaul, could improve the capacity scaling over the non-cooperative BS model either with or without interference canceling. We leave this analysis for future work.

- *Multi-hop is optimal, and outperforms single-hop communication:* The IMH protocol achieves the optimal capacity scaling in all regimes. ISH is optimal at small bandwidth scaling but performs strictly worse than IMH in regimes with wide bandwidths or large numbers of antennas. The reason is that ISH employs longer transmission distances and becomes power-limited earlier than IMH as bandwidth scaling is increased. This suggests that, even though in today's networks capacity is bandwidth-limited and direct transmissions between the mobile nodes and the BS are efficient, in future networks with much larger bandwidths, multi-hop communication may be necessary to fully achieve the network capacity.

- *Hierarchical Cooperation is not necessary in cellular systems:* Optimality of IMH implies that Hierarchical Cooperation (HC) cannot improve the rate scaling achieved with IMH, as opposed to dense ad-hoc networks, where multi-hop is optimal only in some regimes and HC is necessary to achieve the optimal throughput scaling otherwise [15].

- *Wireless backhauling may be optimal, but RN density is critical:* IRH performance depends on the RN density. In the best case scenario with RNs as dense as the user nodes, IRH rate scales as IMH and can be regarded as a practical strategy to achieve capacity scaling while avoiding mobility issues. But if the RN density is lower, the performance of IRH is suboptimal in the power-limited regimes with high bandwidth scaling, and may not offer any gains over ISH in the bandwidth-limited regime with low bandwidth scaling.

- *Applicability to fading and non-coherent communications:* The main results in this paper are obtained under a deterministic path loss channel model with full rank and additive Gaussian noise. However, very similar network scaling laws can be readily argued for the case of the ergodic capacity in frequency selective fading channels with channel state information (CSI) at all network nodes. For the case of non-coherent fading channels –without CSI– there are very few existing results even for capacity of point to point channels [16], [17]. Our results show a behavior similar to those in [18] and [19], which show that in a point-to-point non-coherent wideband channel there is a critical bandwidth occupancy, so that capacity is power limited when the bandwidth exceeds this critical value, and the critical bandwidth threshold grows with the receiver power. However, these results are only for point-to-point channels, and only qualitatively similar to the operating regimes of our network capacity scaling laws. Capacity results for multi-user non-coherent channels are limited, and the scaling exponent for the regime transitions may be different in non-coherent channels than in AWGN and coherent channels even for multiple-antenna point-to-point channels [20].

## A. Relation to Prior Work

The seminal work by Gupta and Kumar [13] showed that the feasible rate in a dense *ad-hoc* network scales as $R(n) \propto \Theta(\frac{1}{\sqrt{n}})$, where $n$ is the number of nodes.[1] Ozgur, Lévêque and Tse introduced HC and showed that it achieves linear scaling (i.e. $R(n) = \Theta(1)$) for dense *ad-hoc* networks [22]. Franceschetti, Migliore and Minero described physical constraints which pose an ultimate limitation leading to $R(n) \leq \Theta(\frac{\log(n)^2}{\sqrt{n}})$ [23]. The results of [22] and [23] differ in the channel model, where [22] considers random i.i.d. phases between any pair of nodes, and [23] considers that as $n$ grows, the inter-node distances become smaller than the wavelength and channel phases are determined by spatial characteristics. Ozgur *et al.* [15] argue that linear scaling may still be achievable in a transitory regime where $n$ is very high but finite, such that nodes separations are larger than the carrier wavelength and channels can still be modeled by i.i.d. random phases. Otherwise, if $n$ is so high that inter-node distances are shorter than the wavelength, channel degrees of freedom

---

[1] We use the standard $f(n) = O(g(n))$, $f(n) = \Omega(g(n))$ and $f(n) = \Theta(g(n))$ notations [21] to respectively represent that at sufficiently high $n$ function $f(n)$ becomes less than or equal than $g(n)$, greater than or equal to $g(n)$, and identical to $g(n)$ up to a constant factor.

scaling is spatially limited as in [23]. The connection between capacity scaling results with an i.i.d. random phase model and with a physical spatially-limited channel phase model is further analyzed in [24]–[26], which formalize the unification of [22] and [23]. Ozgur *et al.* [15] also replaced the traditional separate analysis of dense and extended networks with a generalized analysis of *operating regimes*, defining the user density scaling and determining its threshold for which the operating regime changes from dense-like to extended-like networks. More recently, the practicality of hierarchical cooperation to achieve linear scaling was put into question in [27]. There have been extensions of scaling laws of ad-hoc networks introducing cooperation, mobility, broadcast, infrastructure or large bandwidth. See [28] for a comprehensive review.

Most literature on scaling laws follows ad-hoc network models, which are not adequate representations of a cellular network, even in the case of results like [29]–[31] that have modeled ad-hoc networks with infrastructure support. Our analysis still uses the spatial density model for infrastructure proposed in [29]–[32], but we have taken into account that data in a cellular network is *required* to reach the BS and this may create bottlenecks that limit scaling [33]. In [30] and [31] the analysis characterizes an "ad-hoc network with infrastructure support", where source-destination pairs of user nodes of the same type are formed across the network, and BS infrastructure only assists these user nodes. We consider instead a conventional cellular network traffic model, where user nodes are paired with the *nearest* BS, and there are typically asymmetric downlink and uplink rates with the BS as the ultimate source or destination, respectively. Note that in our multi-hop schemes nodes assist each other by forwarding information corresponding to their primary downlink/uplink exchanges with the nearest BS, however user nodes do not maintain direct traffic flows with each other using device-to-device communications underlaying the primary cellular communications, as recently proposed [3], [34]. Due to this cellular traffic model, our analysis requires novel cut-set bounds and achievable schemes, different than those in [30] and [31]. In addition, [30] and [31] considers a specific physical model for the BS antenna arrays, whereas our analysis is agnostic to the antenna model and the results are expressed instead only as a function of the *effective array dimension*.

The main innovation of our analysis method is evaluating the impact of very large bandwidths in capacity scaling. Most scaling analyses consider a constant finite bandwidth; however in such setups links only become power-limited with distance, not with bandwidth. Another approach consists on a priori letting $W \to \infty$ for each finite value of $n$, and *then* let $n$ grow, as in [35] and [36]; but this does not provide insights on the interaction between bandwidth and power-limited scaling operating regimes. In our model the goal is to find out what happens between these two extremes by letting $W$ and $n$ increase to infinity at the same time with an arbitrary relative exponent

$$\psi := \lim_{n,W \to \infty} \frac{\log W}{\log n}, \quad \Leftrightarrow W = \Theta(n^\psi) \qquad (1)$$

where the two extremes correspond to $\psi = 0$ and $\psi = \infty$. The results in [15] also identify operating regimes depending

on power scaling for the ad-hoc case, which may be interpreted implicitly as a scaling bandwidth. Introducing the bandwidth exponent explicitly allows to analyze the relative value of bandwidth scaling in relation to node and infrastructure density.

More recently, several works have studied the impact of density in cellular wireless systems with models based on stochastic geometry [37]. Although this permits a fine characterization of rate beyond scaling in large networks, the ability to model multi-hop protocols through stochastic geometry is more limited, and both analysis techniques are complementary. For example, in [37] only two hops are possible, whereas in this paper we adopt the generalized multi-hop model with arbitrary number of hops developed in the classic Gupta-Kumar model [13], [15].

In this paper we present scaling results characterized up to the exponents only, ignoring logarithmic variations in scaling. That is, we will not distinguish between scaling functions of the form $\Theta(n^x)$ and $\Theta(n^x \log(n))$. Since $n^\epsilon \geq \Theta(\log(n))$ holds for any $\epsilon \geq 0$ and for sufficiently high $n$, our simplification of scaling notation does not affect the main conclusions in this paper, which mainly consider the values of rate exponents and their comparison at regime transition points. While our simplification is sufficient for the particular conclusions in this paper, we do not claim that logarithmic scaling differences are irrelevant in other applications. Researchers have committed considerable effort to study logarithmic gaps in other scenarios [32], [38].

### B. Paper Organization

This paper is structured as follows: Section II describes the cellular network scaling and channel model. Section III obtains an upper bound to capacity. Section IV describes the different achievable protocols. Section V describes capacity scaling and its relation to the throughput of each protocol. Section VI contains observations and interpretations of the results. Finally, Section VII concludes the paper.

## II. NETWORK AND CHANNEL MODELS

### A. Network Model

We consider a sequence of cellular wireless networks indexed by $n$, where $n$ is the number of single-antenna user nodes randomly and uniformly distributed in an area $A$. The network is supported by $m$ BSs, each with $\ell$ effective antennas (see below), and communication takes place over bandwidth $W$. The BSs do not have the ability to perform cooperative transmission/reception through backhaul. In the IRH protocol defined below, we also have $k > m$ single-antenna fixed RNs that communicate with the BSs through a wireless backhaul.

Table I defines the scaling relation between $n$ and the different network parameters. Here $W_0$, $A_0$, $m_0$, $l_0$, $k_0$ are fixed constants. The exponents of the number of BSs and BS antennas are taken from [30]. The constraint $\beta + \gamma \leq 1$ ensures that the number of infrastructure antennas per node does not grow without bounds. The scaling of the network area is as proposed by [15] to model a continuum of operating regimes

TABLE I
SCALING EXPONENTS OF NETWORK PARAMETERS

| Exponent | Range | Parameter (vs. number of nodes $n$) |
|---|---|---|
| $\psi$ | $[0, \infty)$ | Bandwidth $W = W_0 n^\psi$ |
| $v$ | $[0, 1]$ | Area $A = A_0 n^v$ |
| $\beta$ | $[0, 1]$ | No. of BSs $m = m_0 n^\beta$ |
| $\gamma$ | $[0, 1-\beta]$ | No. of BS antenna array effective dimensions $\ell = \ell_0 n^\gamma$ |
| $\rho$ | $[\beta, 1]$ | No. of RNs $k = k_0 n^\rho$ |

between *dense* ($v = 0$) and *extended* ($v = 1$) networks. We introduce the bandwidth scaling exponent $\psi$ as shown in (1). We also introduce the scaling exponent $\rho \geq \beta$ of the number of RNs for the IRH protocol defined below.

Note that $\ell$ is the number of *effective* antenna dimensions, that is the maximum number of independent spatial dimensions over which a BS can communicate. In a rich scattering environments $\ell$ is equal to the number of physical antennas, whereas if scattering is sparse such that some physical antennas are correlated, $\ell$ represents the number of independent propagation paths that the array can exploit. By focusing on a given number of effective dimensions, our analysis can be applied, with appropriate values of $\ell$, to many antenna array architectures and even sparse propagation models in the literature, such as [30] and [39]–[41]. Hereafter, we use the term "number of antennas" to simply refer to the *effective array dimensions* $\ell$ and represent by $\ell_t$ and $\ell_r$ the effective number of transmit and receive array dimensions.

Note that in [30] and [31] it is assumed that $\tilde{\ell} = n^{\tilde{\gamma}}$ physical antennas are uniformly and randomly located in an area $\Theta(n^{v-\beta})$. Certain physical characteristics of the model in [30] and [31] lead to a constraint in the number of exploited transmit dimensions that scales with the *perimeter* of the array, $\Theta(n^{\frac{v-\beta}{2}})$. Therefore the effective number of transmit dimensions in [30], [31] would be $\gamma = \min(\tilde{\gamma}, \frac{v-\beta}{2})$. In addition, [30], [31] imposes the equality requirement $\beta + \tilde{\gamma} = 1$, which we have relaxed as $\beta + \gamma \leq 1$ to account for any physical array model with $\gamma \leq \tilde{\gamma}$, including but not limited to the model in [30] and [31].

We consider BSs that are placed at fixed distances of each other, dividing the network into regular hexagonal cells around each BS with radius $r_{\text{cell}}$ and with asymptotically (as $n \to \infty$) $\frac{n}{m}$ nodes each, as in Fig. 1. The RNs, when present, are uniformly distributed over cells and placed in a hexagonal layout within each cell. The downlink from the BS to the nodes and the uplink from the nodes to the BS operate independently in alternate time division duplex (TDD) frames. This imposes a $\frac{1}{2}$ penalty in rate but otherwise does not alter the scaling of capacity with $n$. BSs cannot receive in the downlink phase or transmit in uplink, while nodes can do both.

Due to random node placement, the rate achievable by any individual user is a random variable depending on its location and the protocol used. The following definitions are adapted from [13].

*Definition 1:* A downlink (uplink) rate of $R^x_{\text{DL}}(n)$ $(R^x_{\text{UL}}(n))$ bits per second per node is *achieved using protocol x* in a
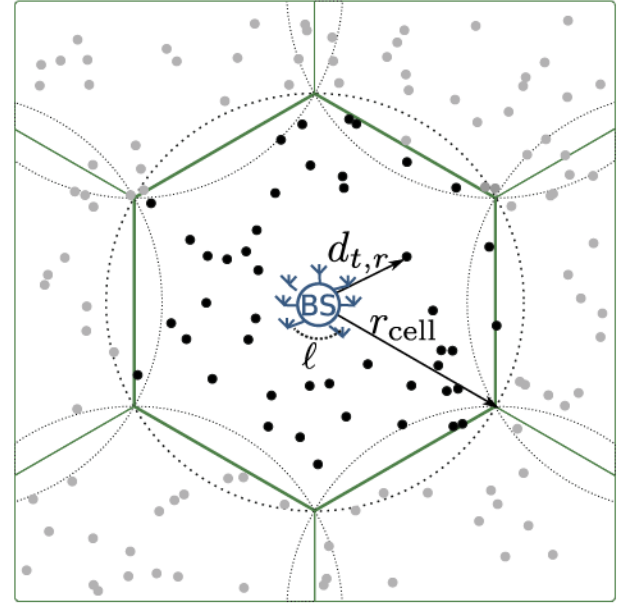


Fig. 1.   One cell of the cellular network.

realization of the cellular network if the protocol can guarantee that all nodes can receive from (transmit to) their assigned BS at least $R^x_{\text{DL}}(n)$ $(R^x_{\text{UL}}(n))$ bits per second.

Note that if we denote by $T^x_{\text{DL}}(n)$ $(T^x_{\text{UL}}(n))$ the sum DL (UL) throughput per BS with protocol $x$, our definition of achievable rate requires that $R^x_{\text{DL}}(n) \leq \frac{m}{n} T^x_{\text{DL}}(n)$.

*Definition 2:* A downlink (uplink) rate of $R_{\text{DL}}(n)$ $(R_{\text{UL}}(n))$ bits per second per node is *feasible* in a realization of the cellular network if there exists a protocol that achieves it. In other words $R_{\text{DL}}(n) = \sup_x R^x_{\text{DL}}(n)$ and $R_{\text{UL}}(n) = \sup_x R^x_{\text{UL}}(n)$.

The definitions above result in random rates depending on the realization of node locations. The definition below is for the largest rate scaling that holds asymptotically with probability 1.

*Definition 3:* The downlink (uplink) per node *throughput capacity scaling* $C_{\text{DL}}(n)$ $(C_{\text{UL}}(n))$ of random cellular network is of the order $\Theta(f(n))$ if there are constants $c_1 < c_2$ such that

$$\lim_{n \to \infty} P\left(R_{\text{DL}}(n) = c_1 f(n)\right) = 1 \qquad (2)$$

$$\lim_{n \to \infty} P\left(R_{\text{DL}}(n) = c_2 f(n)\right) < 1 \qquad (3)$$

We can also define the *achievable rate scaling of protocol x* if in the above definition we replace the feasible rates with achievable rates using protocol $x$.

In this paper we find an upper bound and lower bounds to throughput capacity scaling by studying achievable rate scaling of different protocols. When the two have the same exponent they give the capacity scaling.

*B. Channel Model*

The discrete time received signal observed at a receiver $r$ which can be either a node, BS or RN, is given by

$$\mathbf{y}_r = d_{t,r}^{-\frac{\alpha}{2}} \mathbf{H}_{t,r} \mathbf{x}_t + \sum_{i \in I} d_{i,r}^{-\frac{\alpha}{2}} \mathbf{H}_{i,r} \mathbf{x}_i + \mathbf{z}_r \qquad (4)$$

where $\mathbf{x}_t$ is the signal of the intended transmitter and the set $I$ refers to interfering transmitters active at the same time and over the same frequency band. Furthermore, $d_{j,r}$, $j = t$ or $j \in I$, is the distance between transmitter $j$ and receiver $r$, $\alpha$ is the path loss exponent and $\mathbf{z}_r \sim \mathcal{CN}(0, N_0 \mathbf{I}_{\ell_r})$ is the additive white Gaussian noise.

The channel gain matrix $\mathbf{H} \in \mathbb{C}^{\ell_t \times \ell_r}$ is assumed to be full rank. The full rank assumption is justified by our interpretation of $\ell_t$ and $\ell_r$ as effective antenna dimensions. Each coefficient of the channel matrix has unit gain and an arbitrary phase, $h_{t,r}^{(i,j)} = e^{j\theta_{i,j}}$, and that the channel squared norm satisfies $|\mathbf{H}|^2 = \ell_r \ell_t$. The channel model in (4) is applicable to one symbol transmission with period $T_s = 1/W$ over a frequency-flat channel with power constraint $\mathrm{E}\left[|\mathbf{x}_t|^2\right] \leq \frac{P_t}{W}$ where $P_t$ depends on the type of the transmitter and the fraction of power it dedicates towards $r$. Average transmission power constraints of nodes, BSs and RNs are $P$, $P_{\mathrm{BS}}$ and $P_{\mathrm{RN}}$, respectively.

## III. UPPER BOUND TO CAPACITY SCALING

In order to obtain the upper bound in Theorem 1 below, we develop a series of $m$ cut-set bounds to the downlink sum-rate of the users in each cell, by using the cut that separates the BS -as transmitter- from the rest of the network. Similarly, we develop $n$ cut-set bounds to uplink rate of each user by separating that particular user -as transmitter- from the rest of the network.

*Theorem 1: The downlink throughput capacity scaling of a cellular network is upper bounded*

$$C_{\mathrm{DL}}(n) \leq \Theta\left(n^{\beta+\gamma-1+\min\left(\psi, (1-\nu)\frac{\alpha}{2}\right)}\right) \quad (5)$$

*and the uplink throughput capacity scaling is upper bounded by*

$$C_{\mathrm{UL}}(n \leq \Theta\left(n^{\min\left(\psi, (1-\nu)\frac{\alpha}{2}\right)}\right) \quad (6)$$

*Proof:* We introduce the detailed analysis for downlink. Uplink follows similarly. We first consider the case of no RNs ($k = 0$) and then argue the same bound holds for any $k \leq n$.

We upper bound the sum-rate of the users served by each BS by considering a cut separating that BS from the rest of the network. Each of these $m$ cuts upper bounds the sum downlink rate received by the approximately $n/m$ destination users in one cell. At the receiving side of the cut there is perfect cooperation among $n$ receiver nodes and the remaining $m - 1$ BS transmitters whose transmissions are known to the receivers and can be perfectly canceled. Hence, each cut behaves as a single MIMO channel with array dimensions $\ell_t = \ell$ and $\ell_r = n$.

We represent the distance from each node $r$ to BS $b$ in a diagonal matrix

$$\mathbf{D}_b \triangleq \begin{pmatrix} d_{b,1}^{-\frac{\alpha}{2}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{b,n}^{-\frac{\alpha}{2}} \end{pmatrix},$$

and modify the channel expression (4) to write the signals from all BSs to all nodes in the form

$$\mathbf{y} = \mathbf{D}_t \mathbf{H}_t \mathbf{x}_t + \underbrace{\sum_{t' \neq t} \mathbf{D}_{t'} \mathbf{H}_{t'} \mathbf{x}_{t'}}_{\text{known to all receivers}} + \mathbf{z}, \quad (7)$$

and where $\mathbf{H}_t$ represents the channel matrix between BS $t$ and all receivers.

Using the assumption that channel matrices are full rank and $\ell \leq n/m \leq n$, following standard arguments in [42, eq. (7.10)], it can be shown that $T_{\mathrm{DL}}^t(n)$, the DL sum rate on the cell of BS $t$, can be upper bounded as

$$T_{\mathrm{DL}}^t(n) \leq \max_{\sum_{i=1}^{\ell} P_i \leq P_{\mathrm{BS}}} W \sum_{i=1}^{\ell} \log\left(1 + \frac{P_i \lambda_i^2}{W N_0}\right) \quad (8)$$

where $\lambda_i$ for $i \in [1, \ell]$ are the nonzero, nonnegative singular values of the matrix $\mathbf{D}_t \mathbf{H}_t$.

We know that

$$\lambda_i^2 \leq \sum_{i=1}^{\ell} \lambda_i^2 = \mathrm{tr}\{\mathbf{D}_t \mathbf{H}_t \mathbf{H}_t^H \mathbf{D}_t^H\} \leq \ell \sum_{r=1}^{n} d_{t,r}^{-\alpha}.$$

Concavity of the logarithm suggests that $P_i^* = \frac{P_{\mathrm{BS}}}{\ell}$ maximizes this upper bound. Hence

$$T_{\mathrm{DL}}^t(n) \leq W\ell \log\left(1 + \frac{P_{\mathrm{BS}}}{W N_0} \sum_{r=1}^{n} d_{t,r}^{-\alpha}\right) \quad (9)$$

Notice that if $\lim_{n \to \infty} \frac{P_{\mathrm{BS}} \sum_{r=1}^{n} d_{t,r}^{-\alpha}}{W N_0} = \infty$, then the upper bound in (9) becomes degrees-of-freedom-limited and scales as $\Theta(W\ell)$. Conversely, if $\lim_{n \to \infty} \frac{P_{\mathrm{BS}} \sum_{r=1}^{n} d_{t,r}^{-\alpha}}{W N_0} = 0$, the upper bound is power-limited and scales as $\Theta(\ell P_{\mathrm{BS}} \sum_{r=1}^{n} d_{t,r}^{-\alpha})$.

The sum $\sum_{r=1}^{n} d_{t,r}^{-\alpha}$ can be calculated using the exponential stripping method described in [43]. Consider a series of concentric rings centered at the BS $t$ with inner radius $r_i = n^{\frac{\nu}{2}} e^{\frac{-i}{2}}$ and outer radius $r_{i-1}$. Recall that the user density scales as $n^{1-\nu}$ and network area as $n^{\nu}$, thus the number of nodes contained in each disc is $S_i \leq ne^{1-i}$ with high probability. Using this, we can upper bound the sum over $n$ by summing over the ring. Moreover, the smallest radius that contains one node w.h.p. is $r_s = \Theta(n^{1-\nu})$ so the sum ends at $i \leq \lfloor \log n \rfloor + 1$. For all the outer rings $i \in [1, \lfloor \log n \rfloor]$ we can lower bound distance to the BS by the inner radius $d_{t,i}^{-\alpha} \leq r_{i-1}$. In addition the innermost disk indexed by $i = \lfloor \log n \rfloor + 1$ contains one uniformly-distributed node location, and its distance from the BS scales with $d_{t,i}^{-\alpha} = \theta(r_s) = \Theta(n^{1-\nu})$ with high probability.

$$\sum_{r=1}^{n} d_{t,r}^{-\alpha} \leq \sum_{i=1}^{\lfloor \log n \rfloor + 1} S_i r_i^{-\alpha}$$

$$\leq \left[\sum_{i=1}^{\lfloor \log n \rfloor} ne^{1-i} n^{-\nu \frac{\alpha}{2}} e^{+i\frac{\alpha}{2}}\right] + en^{(1-\nu)\frac{\alpha}{2}}$$

$$\leq n^{-\nu\frac{\alpha}{2}}\left[\log ne^{1+\frac{\alpha}{2}\log(n)}\right] + n^{(1-\nu)\frac{\alpha}{2}}e$$

$$\leq (\log n + 1)n^{(1-\nu)\frac{\alpha}{2}}e \quad (10)$$

where the third inequality is due to $e^{+i\frac{\alpha}{2}} \leq e^{\max(i)\frac{\alpha}{2}}$.

Examining Table I, this leads to $\Theta\left(n^{\gamma+\min(\psi,(1-\nu)\frac{\alpha}{2})}\right)$. Now, by symmetry of the upper bound over all BSs, and by the definition of feasible rate as guaranteed to all users, the throughput capacity of the network is upper bounded by $C_{DL}(n) \leq \frac{m}{n}\min_t T_{DL}^t(n) = \Theta(n^{\beta+\gamma-1+\min(\psi,(1-\nu)\frac{\alpha}{2})})$ completing the proof of Theorem 1 for DL.

Note that this scaling upper bound makes intuitive sense because, with probability 1 as $n \to \infty$, a disc with radius $\Theta(n^{\frac{\nu-1}{2}})$ around a BS contains one receiver, which combined with array gain $n^\gamma$ gives the best-case transfer of power between a single BS and the rest of the network. Also, the degrees of freedom of the cellular network cannot exceed $\Theta(Wm\ell)$.

A similar set of arguments lead to the bound for the uplink. In this case we consider $n$ cuts, each separating one user node from the rest of the network. In this cut, all the BSs and the remaining $n-1$ nodes are on the receiving side of the cut, and their mutual interference is canceled. Due to the fact that the transmitting node has a single antenna (eigenvalue), the degrees of freedom are $\Theta(W)$. The exponential stripping sum (equivalent of (10)) in this case needs to be evaluated over $\sum_{r=1}^{n-1} d_{t,r}^{-\alpha} + \ell \sum_{r=1}^{m} d_{t,r}^{-\alpha} = \Theta(n^{(1-\nu)\frac{\alpha}{2}})$ leading to an upper bound on uplink feasible rate as $\min_t T_{UL}^t(n) = \Theta\left(n^{\min(\psi,(1-\nu)\frac{\alpha}{2})}\right)$

Note that the above scaling laws also apply to the downlink/uplink throughput capacity scaling of a network with $k < n$ RNs. This can be shown evaluating the cut-set bound on an equivalent network with $2n \geq n + k$ user nodes and multiplying the resulting rate per node by 2 which is always greater than $\frac{n+k}{n}$. The capacity scaling exponent does not change when the number of nodes is multiplied by a constant. □

## IV. PROTOCOL MODELS

### A. Infrastructure Single-Hop (ISH)

In the ISH protocol, BSs transmit directly to all nodes in downlink and all nodes transmit directly to the BSs in uplink. The BSs do not cooperate in transmission or reception and the interference signals between different cells are treated as noise. There are $\frac{n}{m}$ nodes uniformly distributed within each cell. The $\ell$ BS antennas are used for Multi-User MIMO (MU-MIMO), implementing a spatial multiplexing scheme to groups of $\ell$ users that allows each BS to transmit or receive $\ell$ signals per bandwidth resource at the same time.

Each BS divides the nodes in its cell in $\frac{n}{m\ell}$ groups of $\ell$ users, and assigns to each group a subchannel with bandwidth $W\ell\frac{m}{n}$. Subchannels are separated using Frequency Division Multiplexing (FDM) in DL and Frequency Division Multiple Access (FDMA) in UL. Within the subchannel of each group, $\ell$ simultaneously transmitted signals coexist using MU-MIMO spatial multiplexing as described in [42, Ch. 10]. The dimensions of the channel matrices always allow this because $\gamma < 1 - \beta$, so there are always more nodes than BS antennas if $n$ is sufficiently large. Also, the multi-user channel matrix, obtained by putting together all point-to-point channel matrices of nodes in the same subchannel and cell, is full rank if nodes are separated at least a quarter of a
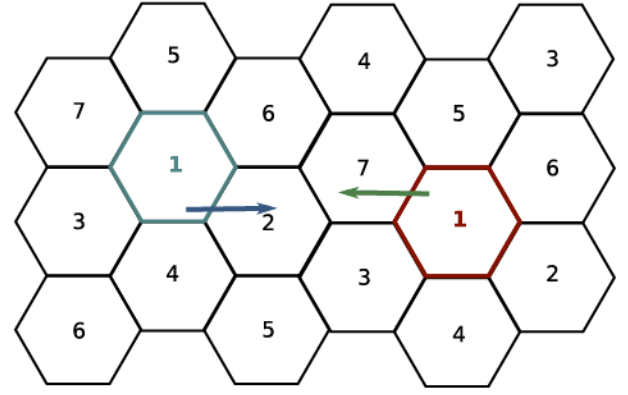


Fig. 2. Time division for multi-hop routing without collisions in a hexagonal lattice. A factor of 1/7 guarantees that two active subcells are always separated by at least two inactive subcells so that transmitters never target the same destination subcell.

wavelength and far-field propagation holds. In downlink, the BS transmits independent signals to each destination with equal power allocation $P_{BS}\frac{m}{n}$.

Note that while our equal division of power and bandwidth may be suboptimal we show, as a part of the analysis of ISH, that in a scaling law sense this power allocation suffices to get the best scaling possible with any single-hop protocol.

### B. Infrastructure Multi-Hop (IMH)

In the IMH protocol, each cell is subdivided regularly into smaller regions of area $A_r$ called *routing subcells*, and information is forwarded to/from the BS via multi-hop communication using a node in each routing subcell as a relay as shown in Fig 3(a). For multi-hopping, the routing subcells must have at least one node with high probability, which results in $A_r > \frac{A}{m}\frac{2\log(\frac{n}{m})}{\frac{n}{m}}$ [30]. Routes are defined as successions of transmissions between adjacent subcells, where each hop covers a distance no longer than four subcell radii, $4r_{subcell} \propto \sqrt{A_r}$, given the largest distance between any two points in adjacent subcells. Sub-cells alternate in becoming active using a non-scaling (i.e. constant) time division scheduling by a constant factor to avoid collisions (transmissions to the same destination subcell) and satisfy the half-duplex constraint. For example in a hexagonal tessellation a 1/7 constant can prevent collisions as illustrated in Fig. 2.

All downlink routes start, and all uplink routes end, at the BS in the center of the cell. We call this point the *head* of all routes, where the BS communicates with its closest $\ell$ users only, using the same MU-MIMO we described for ISH with minor adaptations. Since there are $\ell$ nodes and $\ell$ BS antennas, a single channel with bandwidth $W$ without FDMA/FDM is employed, with MU-MIMO spatial multiplexing in exactly $\ell$ spatial dimensions. The channel and rate models for these links are the same as in ISH, with the new bandwidth allocation and a reduced maximum distance between the BS and the destinations scaling as $4r_{subcell} = \Theta(n^{\frac{\nu-1}{2}})$.

The BS serves as head for a total of $\frac{n}{m}$ routes (one per cell user), but only $\ell$ can be spatially multiplexed at the same time, so the routes are time-multiplexed in a round robin fashion in
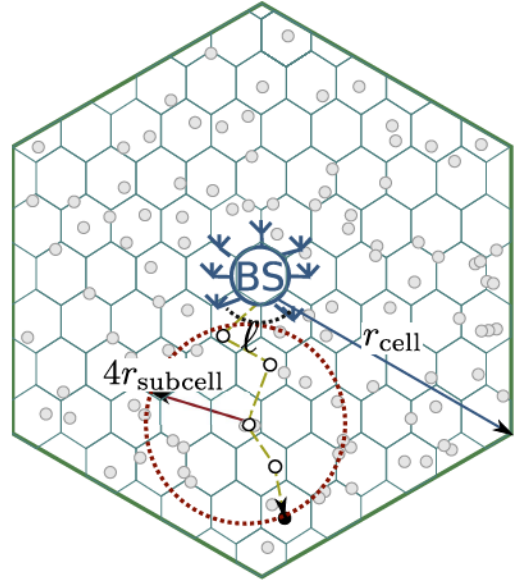
the links between the BS and its neighbors, with each route being served a $\frac{m\ell}{n}$ portion of the time. For the remaining hops on each route, a single node in each routing subcell forwards its received data of a single path to a single node in the next routing subcell in the path. Since each node has a single antenna, there is no MIMO and all the bandwidth and node power are exploited. Again, inter-node distances scale at most as $4r_{\text{subcell}} = \Theta(n^{\frac{\nu-1}{2}})$.

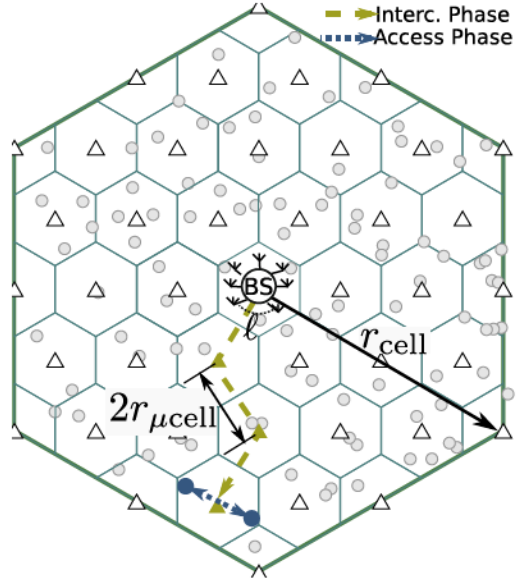## C. Infrastructure Relay-Multi-Hop (IRH)

In the IRH protocol, the network area is divided regularly in a nested double hexagonal grid of $m$ cells and $m + k$ microcells, where nodes in each microcell are served by an Access Point (AP) that is either a BS or a RN. We consider that a controller may decide to use the RNs or not, falling back to the behavior of ISH if the RNs are not sufficiently dense. In downlink, the RNs are exploited if $\rho \geq \beta + \gamma + (\beta - \nu)\frac{\alpha}{2} - \psi$. In uplink, the RNs are exploited if $\rho$ is high enough such that $\min\left(\psi - (\beta + \gamma - \rho)^+, (\rho - \nu)\frac{\alpha}{2}\right) \geq \min\left(\psi, (\beta - \nu)\frac{\alpha}{2} + 1 - \beta\right)$. These thresholds are justified by the analysis of IRH in the next section.

When the above conditions are satisfied and the RNs are utilized, a wireless backhauling connection for all $k/m$ RNs in each cell is provided by their closest BS using IMH. To implement backhauling, time is divided into an *access phase* and a *interconnection phase*, with relative durations $\tau_a \in [0, 1]$ and $1 - \tau_a$.

- In the *access phase*, for a fraction $\tau_a \in [0, 1]$ of the time, in each microcell, all APs exchange data with the user nodes using an ISH protocol. Signals that propagate between different microcells are treated as interference. There are $\frac{n}{m+k}$ nodes within each microcell with high probability. Unlike BSs, RNs do not have $\ell$ antennas and therefore rates in RN microcells create a bottleneck for throughput scaling. APs use FDMA/FDM with a single antenna (no MU-MIMO), allocating transmissions to each user node on orthogonal subchannels with bandwidth $W\frac{m+k}{n}$. A BS transmits with node power allocation $P_{\text{BS}}\frac{m+k}{n}$ and a RN does the same split $P_{\text{RN}}\frac{m+k}{n}$. Note that the capacity scaling is by definition the rate scaling of the worst user. Since microcells where the AP is a single antenna RN are more constrained, we can assume for simplicity that BSs also have a single antenna so that all microcells are represented equally in the analysis. Note also that even if $\rho = \beta$, the numbers $m$ and $k$ may differ by a constant factor and some users in the system would be served by single-antenna APs.
- In the *interconnection phase*, for a fraction $1 - \tau_a$ of the time, BSs exchange data with RNs using an IMH protocol. Each microcell of area $A_r \sim n^{\nu-\rho}$ becomes the *routing subcell* of IMH, and information is forwarded to/from the BS via multi-hop communication using the single RN in each microcell as a relay as shown in Fig 3(b). The BS uses MU-MIMO to transmit or receive up to $\min(\ell, \frac{k}{m})$ routing paths at the same time. Note that unlike IMH protocol, we are no longer guaranteed to have more RNs than transmit antennas.



(a) IMH: Multi-hop is performed by one user node selected in each routing subcell.



(b) IRH: Multi-hop is performed by RNs across microcells during the interconnection phase.

Fig. 3. The two multi-hop protocols in this paper, IMH and IRH.

Each hop covers a distance of exactly two microcell radii, $2r_{\mu\text{cell}}$, as RNs are regularly placed at the centers of their microcells. Microcells alternate in becoming active using a non-scaling (i.e. constant) time or frequency division scheduling to avoid collisions and satisfy the half-duplex constraint.

Note that the access phase and interconnection phase may have different scalings, which determines the optimal time allocation $\tau_a$ and the overall rate scaling of the IRH protocol.

We assume that RNs have one or a fixed number of antennas that do not scale with $n$. This is a realistic model since, in the near future, it is likely that BSs will still have many more antennas than nodes or RNs. However, it is not difficult to extend the results in this paper to the case where the number of RN antennas also scale with n.

## V. CAPACITY SCALING AND RATE SCALING OF PROTOCOLS

### A. IMH Achieves Downlink Capacity Scaling

Our main result is the characterization of the scaling of *throughput capacity* for cellular wireless networks, which is limited by the upper bound on Section III and as we show below achieved by the IMH protocol.

*Theorem 2: For the IMH protocol, downlink rate per node scales as*

$$R_{\mathrm{DL}}^{\mathrm{IMH}}(n) = \Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(1-v)\frac{\alpha}{2}\right)}\right) \quad (11)$$

*and uplink rate per node scales as*

$$R_{\mathrm{UL}}^{\mathrm{IMH}}(n) = \Theta(n^{\beta+\gamma-1+\min\left(\psi,(1-v)\frac{\alpha}{2}\right)}) \quad (12)$$

*Proof:* Appendices A and B. □

Combining upper bound in Theorem 1 and the achievable scaling in Thorem 2, we obtain the following.

*Theorem 3: The downlink per node throughout capacity scales as*

$$C_{\mathrm{DL}}(n) = \Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(1-v)\frac{\alpha}{2}\right)}\right) \quad (13)$$

*and for $\beta + \gamma = 1$, the uplink per node throughut capacity scales as*

$$C_{\mathrm{DL}}(n) = \Theta\left(n^{\min\left(\psi,(1-v)\frac{\alpha}{2}\right)}\right) \quad (14)$$

*Corollary 1: IMH is optimal for downlink, that is it achieves throughput capacity scaling in downlink. For uplink, IMH achieves rate scaling within a gap no larger than $n^{1-\beta-\gamma}$ to capacity, and is optimal for $\beta + \gamma = 1$*

*Remark 1: In our model we have $\beta + \gamma \le 1$. When scattering is rich and the scaling of the number of independent transmits dimensions $\gamma$ equals the number of physical antennas the condition $\beta+\gamma = 1$ corresponds to the total number of infrastructure investment scaling as the number of users. This, for example, is within the realm of ultra-dense networks [44]. In this case IMH is optimal (in terms of throughput capacity scaling) in the uplink as well.*

Next, we obtain the achievable rate scalings of the other protocols introduced in Sec. IV.

### B. ISH is Suboptimal

The ISH protocol is representative of the dominant communication mode in current cellular networks, consisting of direct transmissions between BS and nodes. Our analysis shows that single-hop protocols can not fully exploit large bandwidths, and therefore cellular architectures must adopt multi-hop in future generations if large bandwidths are to be utilized optimally.

*Theorem 4: For the ISH protocol, downlink rate per node scales as*

$$R_{DL}^{\mathrm{ISH}}(n) = \Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(\beta-v)\frac{\alpha}{2}\right)}\right) \quad (15)$$

*and uplink rate scales as*

$$R_{UL}^{\mathrm{ISH}}(n) = \Theta\left(n^{\beta+\gamma-1+\min\left(\psi,(\beta-v)\frac{\alpha}{2}+(1-\beta)\right)}\right) \quad (16)$$

*Proof:* Appendix A. □

*Remark 2: For $\beta = 1$ ISH has the same rate scaling as IMH in all regimes and is optimal. In downlink, for all $\beta < 1$ we have $(\beta - v)\frac{\alpha}{2} < (1 - v)\frac{\alpha}{2}$ and ISH achieves a rate scaling worse than IMH when the bandwidth scaling is $\psi \ge (\beta - v)\frac{\alpha}{2}$. Similarly, in uplink, for all $\beta < 1$ we have $(\beta - v)\frac{\alpha}{2} + (1 - \beta) < (1 - v)\frac{\alpha}{2}$ and ISH performs worse than IMH for $\psi \ge (\beta - v)\frac{\alpha}{2} + (1 - \beta)$.*

### C. IRH Performance Depends Critically on RN Density

There may be practical issues related with multi-hop implementation through mobile users as in IMH, and the use of static dedicated RNs in this protocol provides a reasonable middle-ground. The gap between IRH and IMH gets smaller and can be closed as RN density increases.

*Theorem 5: For the IRH protocol, if $\rho \ge \beta + \gamma + (\beta - v)\frac{\alpha}{2} - \psi$, downlink rate per node scales as*

$$R_{\mathrm{DL}}^{\mathrm{IRH}}(n) = \Theta\left(n^{\min(\beta+\gamma,\rho)-1+\min\left(\psi,(\rho-v)\frac{\alpha}{2}\right)}\right) \quad (17)$$

*and if $\min\left(\psi - (\beta + \gamma - \rho)^+, (\rho - v)\frac{\alpha}{2}\right) \ge \min\left(\psi, (\beta - v)\frac{\alpha}{2} + 1 - \beta\right)$, uplink rate per node scales as*

$$R_{\mathrm{UL}}^{\mathrm{IRH}}(n) = \Theta\left(n^{\min(\beta+\gamma,\rho)-1+\min\left(\psi,(\rho-v)\frac{\alpha}{2}+(\beta+\gamma-\rho)^+\right)}\right), \quad (18)$$

*otherwise rates scale as in ISH.*

*Proof:* Appendix C. □

*Remark 3: The IRH controller always uses the RNs in downlink if $\rho > \beta + \gamma$, and in uplink if $\rho \ge \beta + \max(\gamma, (1 - \beta)\frac{2}{\alpha})$.*

*Remark 4: If $\rho = 1$ the rate scaling with IRH matches the rate scaling with IMH and therefore IRH is optimal in downlink. If $\rho = \beta + \gamma = 1$ IRH is optimal in uplink as well.*

*Remark 5: However if $\rho = 1$ and $\beta + \gamma < 1$ IRH does not meet the uplink upper bound rate scaling. This means that the amount of wired-backhauled infrastructure is a limiting factor even for networks with very high density of wireless-backhauled infrastructure. Finally, if $\rho < 1$, IRH rate scaling is dominated by that of IMH.*

### D. Illustration of the results

Figures 4(a) and 4(b) illustrate the scaling exponents of the upper bound, IMH, ISH and IRH protocols, for the downlink and uplink cases respectively. The horizontal axes represent the exponent of bandwidth, $\psi$, which together with the exponent of number of independent array dimensions, $\gamma$, represents the scaling of the degrees of freedom of BS transmission. The vertical axes show the exponent of the feasible per node rate $\log(R(n))$.

In downlink, the upper bound behaves exactly like IMH, and hence the capacity scaling of the cellular network is fully characterized and IMH is optimal, whereas in uplink IMH cannot achieve the upper bound if $\gamma + \beta < 1$, with the gap between IMH and the upper bound scaling as $n^{1-\gamma-\beta}$. These results generalize the particular case from [30] where the bandwidh does not scale ($\psi = 0$) and the antenna array at the BS follows a physical model with $\tilde{\ell} = n^{\tilde{\gamma}}$ physical antennas
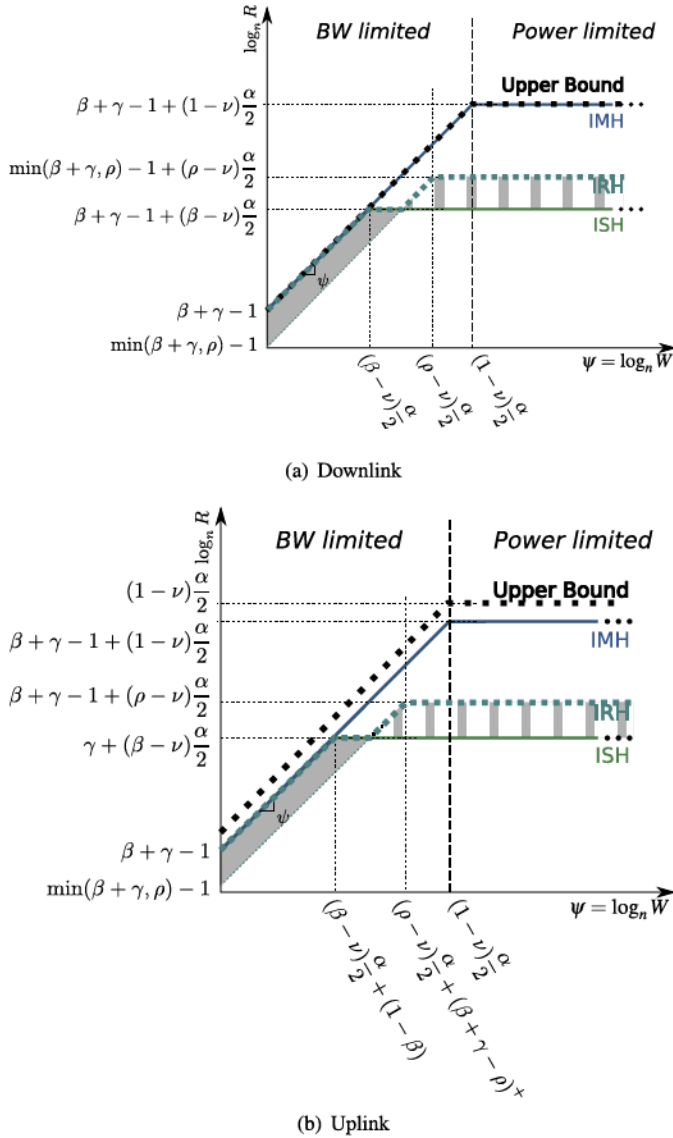
(a) Downlink



(b) Uplink

Fig. 4. Scaling exponents for the upper bound on capacity and feasible rate with ISH, IRH and IMH. Donwlink capacity is fully characterized, whereas uplink capacity lies in the small gap between the upper bound and the IMH protocol.

that experience a constraint on the number of independent transmit dimensions given by $\gamma = \min(\tilde{\gamma}, \frac{1-\beta}{2})$.

In ISH transmissions have to cover longer distances, from the BS to the cell edge, resulting in an earlier transition into a power-limited regime and a lower utility of increasing the bandwidth compared to IMH. In UL this is partially compensated by the fact that the BS receives the total power transmitted by $n/m$ nodes. However, since for a $\beta \leq 1$, we have $(\beta - \nu)\frac{\alpha}{2} + (1 - \beta) \leq (1 - \nu)\frac{\alpha}{2}$, and ISH rate scaling is dominated by that of IMH.

The analysis of IRH in Theorem 5 shows that both in downlink and uplink we can identify a minimum density of RNs such that, beyond this density, IRH outperforms ISH.

- A first scenario where the IRH controller **must not** use the RNs is identified in the switching conditions on Theorem 5. We have highlighted the region where RNs

must not be used with a solid gray shadowed area in the figures. This gap region exists only if $\rho \leq \beta + \gamma$, whereas if $\rho > \beta + \gamma$ the rate scaling with bandwidth exponent $\psi \leq (\beta - \nu)\frac{\alpha}{2}$ is unchanged regardless of whether the controller use the RNs or not.

- A second scenario where the IRH controller **must** use the RNs is identified by comparing the rates of IRH and ISH at $\psi > (\beta - \nu)\frac{\alpha}{2} + (\beta + \gamma - \rho)^+$. We have highlighted this gap with a striped gray shadowed area in the figures. The received power defines a bottleneck when $\psi > (\beta - \nu)\frac{\alpha}{2} + (\beta + \gamma - \rho)^+$ in downlink and $\psi > (\beta - \nu)\frac{\alpha}{2} + (1 + \beta) + (\beta + \gamma - \rho)^+$ in uplink. RNs introduce new bandwidth-limited and power-limited areas to the rate exponents, and allow IRH to outperform ISH. In downlink this gap always exists for any $\rho \geq \beta$, and RNs always increase power-limited rates, because the receiver power depends only on the distance and pathloss exponents, which are improved by RNs to $(\rho - \nu)\frac{\alpha}{2} > (\beta - \nu)\frac{\alpha}{2}$. In uplink, however, if $\rho$ is too low, the striped gray region denoted in the figure collapses and RNs do not improve the power-limited rates of ISH. This occurs because the power bottleneck depends also on the number of uplink transmitters per receiver, and RNs only improve the received power if $(\rho - \nu)\frac{\alpha}{2} + (\beta + \gamma - \rho)^+ > (\beta - \nu)\frac{\alpha}{2} + (1 - \beta)$.

## VI. DISCUSSION AND OBSERVATIONS

### A. The Limitations of ISH are Fundamental to Any Single-Hop Protocols

The ISH protocol defined in Section IV-A assumes equal power allocation in downlink and suboptimal linear MU-MIMO processing. It is not difficult to show that a more general version of single hop would not improve the rate scaling beyond what was obtained in Theorem 4. Uplink and downlink rates in a general cellular network restricted to single-hop communications can be upper bounded by considering broadcast and multiple access channel results [42], respectively. For example, our downlink analysis, which is asymptotic in number of nodes, can be identified as an asymptotic high-SNR broadcast channel when $\psi \leq (\nu - \beta)\frac{\alpha}{2}$. The degrees of freedom region of the broadcast channel in the high-SNR regime is known (see for example [42]) and the worst user performance does not exceed –in terms of scaling exponent– what we achieve with ISH. Similar arguments can be made regarding the low-SNR capacity of the broadcast channel with $\psi \geq (\nu - \beta)\frac{\alpha}{2}$, and the multiple access channel. As a result, no other single-hop protocol throughput scales better than that of ISH, and the differences between ISH and IMH in our achievable schemes arise from the differences between the single-hop and multi-hop architectures, not from our simplifications in ISH.

### B. Operating Regimes of Large Cellular Networks

For point to point channels operating with power $P$, bandwidth $W$, it is well known that there are two operating regimes: When $\frac{P}{WN_0} \ll 1$, the capacity, given by $C(W) = W \log(1 + \frac{P}{WN_0})$. behaves as $\Theta(\frac{P}{N_0})$, and we say that it is *power*

(a) Network bandwidth-limited regime type I.   (b) Network bandwidth-limited regime type II.   (c) Network power-limited regime.
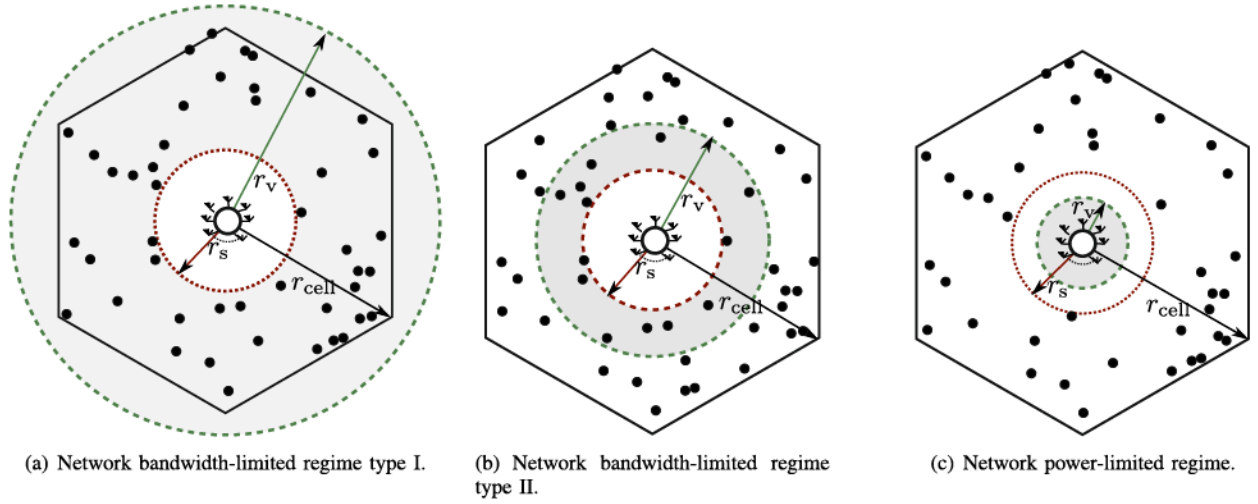
Fig. 5.   Capacity scaling regimes. In the *network bandwidth-limited regime type I* all nodes lie sufficiently close to the BS to receive direct high-SNR transmissions. In the *network bandwidth-limited regime type II* capacity is bandwidth-limited, but most nodes lie far from the BS and cooperation among nodes is imperative. In the *network power-limited* regime capacity is power-limited.

*limited.* Conversely, when $\frac{P}{WN_0} \gg 1$, the capacity behaves as $\Theta(W \log(\frac{P}{N_0}))$, and we say it is *bandwidth limited.*

Our analysis shows that large cellular networks also have two capacity scaling regimes: *Network bandwidth limited regime* and the *network power limited regime*. Furthermore, the network bandwidth limited regime can be categorized into two types depending on whether cooperation among nodes is necessary or not to ensure that network power is not a limitation.

We illustrate these regimes in figure 5. The regimes apply to both downlink and uplink, but we describe them only for downlink for the sake of compactness. We denote the cell radius by $r_{\text{cell}} = \Theta(n^{\frac{\nu - \beta}{2}})$, the distance between two closest nodes by $r_{\text{s}} = \Theta(n^{\frac{\nu - 1}{2}})$, and the longest distance where the BS can transmit or receive without being power limited by $r_{\text{v}} = (W)^{-\frac{1}{\alpha}}$.

*1) Network Bandwidth-Limited Regime Type I:* If $\psi < (\beta - \nu)\frac{\alpha}{2}$, $r_{\text{v}}$ scales faster than $r_{\text{cell}}$, as $n \to \infty$, and it is possible to deliver bandwidth-limited rates $\Theta(n^{\beta + \gamma - 1 + \psi})$ separately to each node in the cell using single hop protocols such as ISH. In this regime there is no requirement for cooperation.

*2) Network Bandwidth-Limited Regime Type II:* For $(\beta - \nu)\frac{\alpha}{2} < \psi \le (1 - \nu)\frac{\alpha}{2}$ network capacity is still bandwidth limited, but single-hop protocols are not. The radius $r_{\text{v}}$ scales slower than $r_{\text{cell}}$ but faster than $r_{\text{s}}$. A few nodes in each cell are sufficiently close to their BS to establish high-SNR direct communications, while a majority of nodes are further away at low-SNR distances. Therefore, in this regime multi-hop is imperative to achieve the bandwidth-limited capacity scaling $\Theta(n^{\beta + \gamma - 1 + \psi})$.

*3) Network Power-Limited Regime:* If $\psi > (1 - \nu)\frac{\alpha}{2}$, $r_{\text{cell}}$ grows faster than $r_{\text{v}}$ and we cannot guarantee that there is at least one user sufficiently close to the BS with high probability. In this regime the SNR in expression (9) is low and the upper bound to capacity is power-limited ($\Theta(n^{\beta + \gamma - 1 + (1 - \nu)\frac{\alpha}{2}})$).

If $\beta = 1$ ISH is optimal, the type II bandwidth-limited regime collapses and the rates of ISH and IMH scale with the same exponent. However, the constraint on the total number of infrastructure units per user $\beta + \gamma \le 1$ means that $\beta = 1$ is incompatible with the exploitation of large antenna degrees of freedom in the BSs.

### C. Relation to the History and Future of Cellular Technologies

Current cellular networks are limited in degrees-of-freedom, namely bandwidth and number antennas. Hence they operate only in the network bandwidth limited regime type I, where user nodes are individually bandwidth limited and single hop protocols are optimal. Therefore, it is not surprising that the gains obtained by the early implementations of relaying in 4G systems have been modest [45].

These early multi-hop implementations correspond to our IRH protocol with low relay density (low $\rho$). In broad terms, multi-hopping uses degrees-of-freedom in exchange of power gain, and therefore is not as advantageous in traditional bandwidth limited cellular networks such as LTE, where degrees of freedom are a precious resource [45]. However, future cellular systems, such as mmWave, will have an abundance of bandwidth and most likely operate in the network power-limited regime, necessitating multi-hopping, either using network nodes or infrastructure RNs, for increased network capacity [46].

Interference studies have shown that mmWave networks operate very close to the threshold between having interference- and noise-limited links [46] (roughly equivalent to network bandwidth- and power-limited rates). The analysis in [46] also showed that a small increase in node density makes the mmWave network transition from noise- to interference-limited. Therefore, the operating regime transitions identified in our analysis have a direct practical impact in the design of future mmWave cellular systems.

## D. IRH Is More Practical Than ISH or IMH for High Density Networks

Both ISH and IRH improve with large invesments in infrastructure, increasing $\beta$ or $\rho$, respectively, but in practice IRH offers advantages over ISH because there are practical limitations to the deployment of wired BSs to increase $\beta$, such as curb excavation rights and cost of fiber-optic for backhaul connection.

IRH also has practical advantages over IMH as a multi-hop implementation, due to the fact that RNs are typically static, connected to the energy grid, and owned by the same network operator. Conversely, user nodes are typically mobile, battery powered and owned by customers. Thus the implementation of IMH would pose greater practical challenges such as support for mobility in the multi-hop protocol, battery efficiency optimization, and behavioral incentives to prevent customers from rejecting cooperation.

## E. Hierarchical Cooperation is Not Necessary for Cellular Capacity Scaling

In an ad-hoc network both direct transmission and multi-hop may be suboptimal in some regimes, and a HC protocol is employed to achieve capacity scaling [15]. This demonstrates the utility of cooperative *virtual antenna arrays*, formed through coordinated joint transmissions by single antenna nodes grouped in clusters. Our analysis of cellular networks shows instead that IMH achieves downlink capacity scaling in all regimes. Therefore node clusters forming virtual antenna arrays are not necessary to achieve capacity scaling in cellular networks.

We leave for future work a scenario in which HC might regain relevance for cellular networks, where we relax the assumption that BSs cannot exchange their messages through a backhaul connection and perform joint transmissions. For cooperative BSs, it is possible that virtual antenna arrays formed by clusters of cooperative devices become necessary at the user side.

## F. Cellular-Specific Traffic Bottlenecks Affect Capacity Scaling

Some existing "ad-hoc networks with infrastructure support" scaling laws analyses [29]–[31], [47]–[49] model infrastructure only as an intermediary to assist the same ad-hoc type communications. These works study rates from some user nodes to others, where the infrastructure is a mere intermediary, and when BSs are so far apart that communicating with them does more harm than good, these works ignore the infrastructure and apply ad-hoc protocols such as HC. The more realistic cellular network analyzed in this paper requires traffic to always flow through BSs even when this causes bottlenecks as illustrated in our analysis.

## VII. Conclusions

In this paper we have obtained the throughput capacity scaling of cellular wireless networks in a model comprising scaling of area, BS density, number of antennas per BS,

total bandwidth, and also, optionally, number of wireless backhauled RNs. We have shown that cellular network capacity scaling exhibits a transition between network bandwidth-limited and network power-limited operating regimes as the bandwidth increases, equivalent to the well known transition in point-to-point links from bandwidth-limited to power-limited capacity. Moreover, we have shown that different protocols can experience protocol-specific suboptimal transitions into power-limited behavior earlier than (i.e. for bandwidth scaling exponent lower than) the transition experienced by the capacity scaling. The transition thresholds are fundamentally related to the typical distance between a transmitter and a receiver for each protocol, suggesting that cooperative multi-hop schemes transmitting between nearest neighbor across the minimal distance have an advantage in networks with wide bandwidths. In fact, our results show multi-hop is optimal for downlink.

Single-hop protocols deal with the longest transmission distances and transition into power limitation the earliest. This means the network bandwidth-limited capacity regime is further divided into two subtypes: type I, where bandwidth is low enough that single-hop protocols are bandwidth-limited and all users can be served *independently* with bandwidth-limited rates; and type II, where the network capacity is bandwidth-limited but single hop protocols are power-limited and cooperation is imperative to serve capacity-achieving rates to all users.

In cellular networks with additional wireless-backhauling RNs the capacity scaling law depends strongly on RN density, so that if the number of RNs is insufficient the network is better off disregarding the RNs altogether and using only the BSs. Conversely, if a sufficiently dense set of RNs is installed, cooperative multi-hop capacity can be achieved. However, a low density of wired-backhauling infrastructure can still be a limitation even if the RN density is high.

Our analysis provides a theoretical framework to explain historical experiences with the implementation of multi-hop and relaying in cellular networks, where the gains have been very modest. Current cellular systems operate in the network bandwidth-limited type I regime, where our analysis predicts that adding RNs brings little advantage and may even decrease rate scaling.

Our analysis is also highly relevant for the design of future cellular networks with increased bandwidth, such as mmWave or carrier aggregation systems. Preliminary studies in mmWave have shown highly parameter-sensitive transitions from bandwidth-limited to power-limited behavior, suggesting potentials for multi-hop communications.

## APPENDIX A
## PROOF OF THEOREM 4

We start with the analysis of ISH, which also forms the foundation of the IMH and IRH protocols. We describe the downlink proof in detail, whereas the uplink proof follows by minor changes. In MU-MIMO downlink BS $t$ assigns to each user node $r$ in the same subchannel a signature unitary vector $\mathbf{u}_{t,r}$ and transmits $\mathbf{x} = \sum_{r=1}^{\ell} \mathbf{u}_{t,r} x_{t,r}$, satisfying the power

allocation $E\left[|\mathbf{u}_{t,r}x_{t,r}|\right] = P_{BS}\frac{m}{n}$ by the protocol description of ISH (Sec. IV). Out-of-cell interfering transmissions in the same subchannel and transmissions by the BS not canceled by the linear precoder are treated as noise:

$$y_{t,r} = \underbrace{d_{t,r}^{-\frac{\alpha}{2}}(\mathbf{h}_{t,r}^H\mathbf{u}_{t,r})x_{t,r} + \sum_{r'}d_{t,r}^{-\frac{\alpha}{2}}(\mathbf{h}_{t,r}^H\mathbf{u}_{t,r'})x_{t,r'}}_{I_1\ (\text{Same BS})}$$
$$+ \underbrace{\sum_{(t',r')\in I_{t,r}^{\text{ISH}}} d_{t',r}^{-\frac{\alpha}{2}}(\mathbf{h}_{t',r}^H\mathbf{u}_{t',r'})x_{t',r'} + z_r}_{I_2\ (\text{Other BSs})} \quad (19)$$

We represent by $I_1$ the self-interference of the signals transmitted by the same BS towards other nodes, and by $I_2$ the out-of-cell interference from the set of interferers $I_{t,r}^{\text{ISH}}$ consisting in all transmitter-receiver pairs $(t', r')$ allocated to the same subchannel as $(t, r)$ by some other BS $t'$. The design of the transmit vectors $\mathbf{u}_{t,r}$ is studied in [42, 10.3]. We apply only linear transmit precoding vectors $\mathbf{u} = \mathbf{h}/\sqrt{\ell}$ that are optimal at low-SNR and suboptimal at high-SNR. However, since we are only interested in scaling, less power efficiency in the bandwidth-limited regime does not affect the scaling exponent. This makes the channel gain at the desired user

$$|\mathbf{h}_{t,r}^H\mathbf{u}_{t,r}|^2 = \left|\frac{\mathbf{h}^H\mathbf{h}}{\sqrt{\ell}}\right|^2 = \ell$$

Due to the fact that all transmissions are uniformly allocated across the available bandwidth and spatial dimensions, and there are many interferers, the second form of interference can be approximated as white noise by the central limit theorem. Since Gaussian is the worst distribution a noise of known covariance can have [50], we lower bound the rate by also modeling the self-interference within the MU-MIMO scheme as Gaussian with variance $E\left[|I_1|^2\right]$. The variance of the total MU-MIMO self-interference can be characterized as

$$E\left[|I_1|^2\right] = E\left[|\sum_{r'}d_{t,r}^{-\frac{\alpha}{2}}(\mathbf{h}_{t,r}^H\mathbf{u}_{t,r'})x_{t,r'}|^2\right] = \Theta(\ell\frac{m}{n}P_{BS})d_{t,r}^{-\alpha}$$

For the out-of-cell interference, we introduce a notation that applies to all protocols. For protocol x we denote the noise plus out-of-cell interference Power Spectral Density (PSD) for transmitter $t$ and receiver $r$ as

$$N_I^x \triangleq \frac{E\left[|I_2 + z_r|^2\right]}{W\frac{m}{n}\ell} = \frac{\sum_{t'\in I_{t,r}^x}d_{t',r}^{-\alpha}P_{t'}}{W} + N_0, \quad (20)$$

where $I^x(t, r)$ denotes a set of out-of-cell interference transmitters affecting $t$. Here $P_{t'}$ denotes the total power of interferer $t'$. In our particular case of ISH downlink $I(t, r)$ is the set of all BSs except $t$, and $P_{t'} = P_{BS}$.

Therefore the rate in the ISH downlink link $(t, r)$ can be lower bounded as

$$R_{t,r}^{\text{ISH}} \geq \frac{m}{n}\ell W \log\left(1 + \frac{\ell\frac{m}{n}P_{BS}d_{t,r}^{-\alpha}}{E\left[|I_1|^2\right] + W\frac{m}{n}\ell N_I^{\text{ISH}}}\right), \quad (21)$$

The achievable rate scaling must by definition be guaranteed to all the nodes in the network, so that

$R^{\text{ISH}}(n) = \min_{t,r}R_{t,r}^{\text{ISH}}$ where each $R_{t,r}^{\text{ISH}}$ is lower bounded by (21). Also, (21) decreases monotonously with $d_{t,r}$, and the worst-case distance satisfies $\max_{t,r}d_{t,r} = \Theta(r_{\text{cell}})$ with high probability. Therefore

$$R^{\text{ISH}}(n) \geq \begin{cases} \Theta(\frac{m}{n}\ell W) & \text{iff } r_{\text{cell}}^{-\alpha} \gtrsim WN_I^{\text{ISH}} \\ \Theta(\frac{m}{n}\ell r_{\text{cell}}^{-\alpha}) & \text{iff } r_{\text{cell}}^{-\alpha} \ll WN_I^{\text{ISH}}. \end{cases} \quad (22)$$

Finally, we show that the threshold of (22) is equivalent to $\psi > (\beta - \nu)\frac{\alpha}{2}$, producing (15). The ISH uplink analysis is identical except the transmitted power budget in uplink scales with $n/m$ per cell, yielding (16).

We examine necessary and sufficient conditions for $r_{\text{cell}}^{-\alpha} \ll WN_I^{\text{ISH}}$ separately. For a necessary condition for $WN_I \ll r_{\text{cell}}^{-\alpha}$, we upper bound the PSD by

$$N_I^{\text{ISH}} \leq \Theta\left(n^{((\beta-\nu)\frac{\alpha}{2}-\psi)^+}\right) \quad (23)$$

so, if $\psi \leq (\beta - \nu)\frac{\alpha}{2}$, then the scaling exponent of $WN_I^{\text{ISH}}$ is always lower than or equal to $r_{\text{cell}}^{-\alpha} \leq d_{t,r}^{-\alpha}$, and the rates of all users (21) are bandwidth-limited. Therefore, $\psi > (\beta - \nu)\frac{\alpha}{2}$ is necessary for the link rates to become power-limited.

To prove (23) we upper bound the distance sum in (20). In the ISH protocol we can upper bound the distance $d_{t',r}$ by the distance between the border of the cell of BS $t'$ and the cell where $r$ belongs. Considering the geometry of the hexagonal layout it can be shown that $6k$ cells form a ring with index $k$ separated a distance greater or equal than $\frac{3}{2}k - 1$ cell radii from the border of the cell of $r$. The network is finite and a maximum $k$ exists, but we can get rid of border effects and further bound the interference by extending the sum all the way to $k \to \infty$.

$$\sum_{t'\in I^{\text{ISH}}}d_{t',r}^{-\alpha} \leq r_{\text{cell}}^{-\alpha}\sum_{k=1}^{\infty}(6k)(2k - 1)^{-\alpha}$$
$$\leq r_{\text{cell}}^{-\alpha}\sum_{k=1}^{\infty}\frac{12}{(2k - 1)^{\alpha-1}}$$
$$= 12r_{\text{cell}}^{-\alpha}\left(\frac{3}{2}\right)^{\alpha-1}\zeta(\alpha - 1, \frac{1}{3}) \quad (24)$$

where the last equality holds for $\alpha > 2$ and the result is expressed using the generalized Riemann Zeta function $\zeta(s) = \sum_{x=1}^{\infty}\frac{1}{x^s}$, which is a constant with regard to $n$. Combining with $W$ this shows that interference PSD scales as $n^{(\beta-\nu)\frac{\alpha}{2}-\psi}$, while noise PSD $N_0$ is constant, so (20) for ISH downlink scales as (23).

To prove the same condition is also sufficient we assume $\psi > (\beta - \nu)\frac{\alpha}{2}$. This allows to approximate the upper bound (23) by $\simeq N_0$, and since $N_I^{ISH} \geq N_0$ the bound is tight. Therefore, if $\psi > (\beta - \nu)\frac{\alpha}{2}$ then $R_{t,r}^{\text{ISH}} \simeq \frac{m}{n}\ell W \log\left(1 + \frac{P_{BS}d_{t,r}^{-\alpha}}{WN_0}\right)$, and since $\min_{t,r}d_{t,r}^{-\alpha} = \Theta(r_{\text{cell}}^{-\alpha}) \leq \Theta(W)$, this leads to $R^{\text{ISH}}(n) = \Theta(\frac{m}{n}\ell r_{\text{cell}}^{-\alpha})$.

## APPENDIX B
### PROOF OF THEOREM 2

The proof of Theorem 2 consists of using the arguments in Appendix A with minor variations. We only discuss the main differences here. Achievable rate scaling in IMH is given by the minimum of two constraints: we denote the BS-node link rates at the center of the cell by $R_{t,r}^{\text{IMH}(1)}$, and node-node link rates in the rest of the multi-hop system by $R_{t,r}^{\text{IMH}(2)}$. Thus

$$R_{DL}^{\text{IMH}}(n) = \min_{t,r} \min_{\{1,2\}} (R_{t,r}^{\text{IMH}(1)}, R_{t,r}^{\text{IMH}(2)}).$$

The scaling of $R_{t,r}^{\text{IMH}(1)}$, the BS-node link rate, is scaling using a MU-MIMO similar to ISH, with the exception that no frequency division is required. The BS of each cell transmits $\ell$ signals towards the nearest routing subcells with bandwidth $W$ using separate spatial signatures. There are a total of $n/m$ routes that need to be served by the BS, so the route towards each destination is time multiplexed by a factor $m\ell/n$ at the BS. Thus, rate between a BS and a node in its nearest routing subcell can be expressed as

$$R_{t,r}^{\text{IMH}(1)} = \frac{m\ell}{n} W \log\left(1 + \frac{P_t d_{t,r}^{-\alpha} \ell}{\mathrm{E}\left[|I_1|^2\right] + W N_I^{\text{IMH}}}\right) \quad (25)$$

where the power is $P_t = P_{\text{BS}}/\ell$ in DL and $P_t = P$ in UL.

The link rates in the rest of the node-node hops of each route are single antenna. Here, a factor of $\frac{m\ell}{n}$ is imposed on the rate because the routes inherit it from the BS.

$$R_{t,r}^{\text{IMH}(2)} = \frac{m\ell}{n} W \log\left(1 + \frac{P d_{t,r}^{-\alpha}}{W N_I^{\text{IMH}}}\right) \quad (26)$$

In order to determine the scaling of each term, the same arguments applied to the scaling of (22) can then be applied to (25) and (26). To study $N_I^{\text{IMH}}$, the noise plus out-of-subcell interference PSD for IMH using (20), we have to take into account that the set of interferers $I_{t,r}^{\text{IMH}}$ is the set of all transmitters in nearby routing subcells that transmit at the same time as the link $(t, r)$. In downlink the interferers may be either other nodes or BS, and the power of each interferer can be upper bounded by the worst case $P_{t'} \le \max(P_{\text{BS}}, P)$, whereas in UL only nodes transmit $P_{t'} = P$.

In order to derive an expression similar to (24) for IMH, $d_{t',r}$ is upper bounded by the distance between the border of the routing subcell that contains $t'$ and the border of the routing subcell that contains $r$. In order to guarantee a half-duplex collision-free routing, IMH uses a time-division where only 1 out of every 7 subcells in the hexagonal layout transmit at the same time. Following as in (24) we can show that

$$N_I^{\text{IMH}} = \Theta\left(n^{((1-\nu)\frac{\alpha}{2} - \psi)^+}\right)$$

Examining Table I both the BS to nodes and the node to node links can be shown to scale as

$$R_{t,r}^{\text{IMH}(1)} = \Theta(n^{\beta+\gamma-1+\min(\psi,(1-\nu)\frac{\alpha}{2})}) = R_{t,r}^{\text{IMH}(2)}. \quad (27)$$

where we distinguish the same two regimes as in ISH, but this time the threshold is $\psi < (1-\nu)\frac{\alpha}{2}$ for the bandwidth-limited rates, proving Theorem 2 for the downlink.

The proof for uplink follows the same principle but the BS receives multiple transmissions at the same time, increasing the received power, and the point to point links become a bottleneck where each route is served with rate

$$R_{UL}^{\text{IMH}}(n) = \Theta(n^{\beta+\gamma-1+\min(\psi,(1-\nu)\frac{\alpha}{2})}) \quad (28)$$

## APPENDIX C
### ANALYSIS OF IRH

We first assume that the IRH always makes use of the RNs, and from the resulting rates derive the scaling exponent thresholds where it is best to fall back to ISH. The IRH protocol implements RN multi-hop in two phases. The access phase is an ISH protocol with $m + k$ APs that consist of both BSs and RNs. We denote this protocol by ISH-R and its achievable rate scaling

$$R_{DL}^{\text{ISH-R}}(n) = \tau\Theta\left(n^{\rho-1+\min(\psi,(\rho-\nu)\frac{\alpha}{2})}\right). \quad (29)$$

The interconnection phase uses an IMH protocol where the BSs are the infrastructure and the RNs are the user nodes. We denote this by IMH-R with achievable backhauling rate per RN

$$R_{DL}^{\text{IMH-R}}(k) = (1-\tau)\Theta\left(k^{\beta'+\gamma'-1+\min(\psi',(1-\nu')\frac{\alpha}{2})}\right) \quad (30)$$

where

$$\beta' = \frac{\beta}{\rho}, \quad \gamma' = \frac{\gamma}{\rho}, \quad \psi' = \frac{\psi}{\rho}.$$

where we modify the exponents of area, bandwidth and number of antennas to account for the fact that IMH is evaluated with $k = n^\rho$ user nodes. Each RN serves $n/k$ nodes and must divide the backhauling rate it achieves equally among them, giving $n^{\rho-1} R_{DL}^{\text{IMH-R}}(n^\rho)$ per user node.

Finally, each node in the IRH network achieves the minimum between the rate of its link to its AP, and the fraction of the AP bachkauling rate that is assigned to the node.

$$R_{DL}^{\text{IRH}}(n) = \Theta\Big( \min\Big( (1-\tau)n^{\beta+\gamma-1+\min(\psi,(\rho-\nu)\frac{\alpha}{2})},$$
$$\tau n^{\rho-1+\min(\psi,(\rho-\nu)\frac{\alpha}{2})}\Big)\Big) \quad (31)$$

The optimal time-division in a decode-and-forward relay scheme with two links of known rates is well known to be the value that minimizes the total transmission time per bit, as verified for example in [51]. This is given by

$$\tau^* = \arg\min_\tau (1-\tau)\frac{1}{k/n R_{DL}^{\text{IMH-R}}(k)} + \tau\frac{1}{R_{DL}^{\text{IRH}}(n)}$$

$$= \frac{k/n R_{DL}^{\text{IMH-R}}(k)}{R_{DL}^{\text{IRH}}(n) + k/n R_{DL}^{\text{IMH-R}}(k)}. \quad (32)$$

If $R_{DL}^{ISH-R}(n) > \Theta(k/n R_{DL}^{IMH-R}(n))$, $\tau^*$ converges to zero. If $R_{DL}^{ISH-R}(n) < \Theta(k/n R_{DL}^{IMH-R}(n))$, $\tau^*$ converges to 1. If $R_{DL}^{ISH-R}(n) = \Theta(R_{DL}^{IMH-R}(n))$, $\tau^*$ does not affect the rate scaling. Putting everything together gives the expression in Theorem 5 for downlink. By careful comparison of Theorem 5 and Theorem 4, we deduce the threshold where it is better to use ISH is $\rho \ge \beta + \gamma + (\beta - \nu)\frac{\alpha}{2} - \psi$.

By similar arguments, the scaling of uplink IRH when the RNs are used can be shown to be

$$R_{\text{UL}}^{\text{IRH}}(n) = \Theta\Bigg( \min\Big( (1-\tau)n^{\beta+\gamma-1+\min\left(\psi,(\rho-\nu)\frac{\alpha}{2}\right)},$$
$$\tau n^{\rho-1+\min\left(\psi,(\rho-\nu)\frac{\alpha}{2}+1-\rho\right)}\Big)\Bigg) \quad (33)$$

and by comparison with Theorem 4 we can deduce that IRH only benefits from RNs if $\rho$ is high enough such that $\min\left(\psi-(\beta+\gamma-\rho)^+,(\rho-\nu)\frac{\alpha}{2}\right) \geq \min\left(\psi,(\beta-\nu)\frac{\alpha}{2}+1-\beta\right)$.

## REFERENCES

[1] F. Gómez-Cuba, S. Rangan, and E. Erkip, "Scaling laws for infrastructure single and multihop wireless networks in wideband regimes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, Jun./Jul. 2014, pp. 76–80.

[2] F. Gómez-Cuba, S. Rangan, E. Erkip, and F. J. González-Castaño, "Capacity scaling bounds in wideband cellular networks," in *Proc. 24th Int. Zurich Seminar Commun. (IZS)*, Zurich, Switzerland, Mar. 2016.

[3] F. Boccardi, R. W. Heath, Jr., A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 74–80, Feb. 2014.

[4] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, Jun. 2011.

[5] P. Pietraski, D. Britz, A. Roy, R. Pragada, and G. Charlton, "Millimeter wave and Terahertz communications: Feasibility and challenges," *ZTE Commun.*, vol. 10, no. 4, pp. 3–12, 2012.

[6] Y. Azar *et al.*, "28 GHz propagation measurements for outdoor cellular communications using steerable beam antennas in New York City," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2013, pp. 5143–5147.

[7] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[8] T. S. Rappaport, R. W. Heath, Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. London, U.K.: Pearson Education, 2014.

[9] J. Hoydis, S. ten Brink, and M. Debbah, "Massive MIMO in the UL/DL of cellular networks: How many antennas do we need?" *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 160–171, Feb. 2013.

[10] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[11] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[12] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, "Femtocells: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497–508, Apr. 2012.

[13] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[14] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the *K*-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[15] A. Ozgur, R. Johari, D. N. C. Tse, and O. Lévêque, "Information-theoretic operating regimes of large wireless networks," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 427–437, Jan. 2009.

[16] M. Médard and R. G. Gallager, "Bandwidth scaling for fading multipath channels," *IEEE Trans. Inf. Theory*, vol. 48, no. 4, pp. 840–852, Apr. 2002.

[17] V. Sethuraman, L. Wang, B. Hajek, and A. Lapidoth, "Low-SNR capacity of noncoherent fading channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 4, pp. 1555–1574, Apr. 2009.

[18] A. Lozano and D. Porrat, "Non-peaky signals in wideband fading channels: Achievable bit rates and optimal bandwidth," *IEEE Trans. Wireless Commun.*, vol. 11, no. 1, pp. 246–257, Jan. 2012.

[19] F. Gómez-Cuba, J. Du, M. Médard, and E. Erkip, "Unified capacity limit of non-coherent wideband fading channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 43–57, Jan. 2017.

[20] M. Chowdhury, A. Manolakos, F. Gómez-Cuba, E. Erkip, and A. J. Goldsmith, "Capacity scaling in noncoherent wideband massive SIMO systems," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr./May 2015, pp. 1–5.

[21] D. E. Knuth, "Big omicron and big omega and big theta," *ACM SIGACT News*, vol. 8, no. 2, pp. 18–24, Apr./Jun. 1976.

[22] A. Ozgur, O. Lévêque, and D. Tse, "Hierarchical cooperation achieves linear capacity scaling in ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3549–3572, Oct. 2007.

[23] M. Franceschetti, M. D. Migliore, and P. Minero, "The capacity of wireless networks: Information-theoretic and physical limits," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3413–3424, Aug. 2009.

[24] A. Özgür, O. Lévêque, and D. Tse, "Linear capacity scaling in wireless networks: Beyond physical limits?" in *Proc. Inf. Theory Appl. Workshop (ITA)*, Jan./Feb. 2010, pp. 259–268.

[25] S.-H. Lee and S.-Y. Chung, "Capacity scaling of wireless ad hoc networks: Effect of finite wavelength," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2010, pp. 1713–1717.

[26] S. H. Lee and S. Y. Chung, "Capacity scaling of wireless ad hoc networks: Shannon meets Maxwell," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1702–1715, Mar. 2012.

[27] S. S.-N. Hong and G. Caire, "Demystifying the scaling laws of dense wireless networks: No linear scaling in practice," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 71–75.

[28] N. Lu and X. S. Shen, "Scaling laws for throughput capacity and delay in wireless networks—A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 642–657, 2nd Quart., 2013.

[29] U. C. Kozat and L. Tassiulas, "Throughput capacity of random ad hoc networks with infrastructure support," in *Proc. Int. Conf. Mobile Comput. Netw.*, 2003, pp. 55–65.

[30] W.-Y. Shin *et al.*, "Improved capacity scaling in wireless networks with infrastructure," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5088–5102, Aug. 2011.

[31] C. Jeong and W.-Y. Shin, "Ad hoc networking with cost-effective infrastructure: Generalized capacity scaling," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2014, pp. 1–27.

[32] P. Li, X. Huang, and Y. Fang, "Capacity scaling of multihop cellular networks," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 2831–2839.

[33] L. Zeger and M. Médard. (Feb. 2014). "On scalability of wireless networks: A practical primer for large scale cooperation." [Online]. Available: https://arxiv.org/abs/1402.1761

[34] X. Lin, J. G. Andrews, and A. Ghosh, "Spectrum sharing for device-to-device communication in cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 12, pp. 6727–6740, Dec. 2014.

[35] R. Negi and A. Rajeswaran, "Capacity of power constrained ad-hoc networks," in *Proc. 23rd Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, Mar. 2004, p. 453.

[36] X. Tang and Y. Hua, "Capacity of ultra-wideband power-constrained ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 54, no. 2, pp. 916–920, Feb. 2008.

[37] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2196–2211, Oct. 2015.

[38] M. Franceschetti, O. Dousse, D. N. C. Tse, and P. Thiran, "Closing the gap in the capacity of wireless networks via percolation theory," *IEEE Trans. Inf. Theory*, vol. 53, no. 3, pp. 1009–1018, Mar. 2007.

[39] M. Desgroseilliers, O. Lévêque, and E. Preissmann, "Spatial degrees of freedom of MIMO systems in line-of-sight environment," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 834–838.

[40] V. Raghavan and A. M. Sayeed, "Sublinear capacity scaling laws for sparse MIMO channels," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 345–364, Jan. 2011.

[41] M. K. Samimi and T. S. Rappaport, "Ultra-wideband statistical channel model for non line of sight millimeter-wave urban channels," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2014, pp. 3483–3489.

[42] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[43] M. Franceschetti, "A note on Lévêque and Telatar's upper bound on the capacity of wireless ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3207–3211, Sep. 2007.

[44] R. Baldemair *et al.*, "Ultra-dense networks in millimeter-wave frequencies," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 202–208, Jan. 2015.

[45] F. Gómez-Cuba and F. J. González-Castaño, "Improving third-party relaying for LTE-A: A realistic simulation approach," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 2344–2350.

[46] M. Rebato, M. Mezzavilla, S. Rangan, F. Boccardi, and M. Zorzi, "Understanding noise and interference regimes in 5G millimeter-wave cellular networks," in *Proc. 22th Eur. Wireless Conf.*, May 2016, pp. 1–5.

[47] H. S. Dhillon and G. Caire, "Information theoretic upper bound on the capacity of wireless backhaul networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 251–255.

[48] H. S. Dhillon and G. Caire, "Scalability of line-of-sight massive MIMO mesh networks for wireless backhaul," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 2709–2713.

[49] J. Yoon, W.-Y. Shin, and S.-W. Jeon, "Elastic routing in wireless networks with directional antennas," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun./Jul. 2014, pp. 1001–1005.

[50] S. N. Diggavi and T. M. Cover, "The worst additive noise under a covariance constraint," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 3072–3081, Nov. 2001.

[51] K. Azarian, H. El Gamal, and P. Schniter, "On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4152–4172, Dec. 2005.

**Felipe Gómez-Cuba** received the Ingeniero de Telecomunicación degree (Telecommunication Engineering five year program) in 2010, the M.Sc. degree in signal processing applications for communications in 2012, and the Ph.D. degree from the University of Vigo, Spain, in 2015. He was a Researcher with the Information Technologies Group, Telematic Engineering Department, University of Vigo, from 2010 to 2011, and the Galician Research and Development Center in Advanced Telecommunications, Vigo, Spain, from 2011 to 2013. He was a Visiting Scholar with the NYUWireless Center, NYU Tandon School of Engineering, Brooklyn, NY, USA, from 2013 to 2014 and received the Ph.D. degree under the FPU Grant from the Spanish MINECO from 2013 to 2016 with the University of Vigo. He received the Marie Curie Individual Fellowship, Global Fellowship with the Dipartimento d'Engegneria dell'Informazione, University of Padova, Italy, and the Department of Electrical Engineering, Stanford University, USA, in 2016.

**Elza Erkip** (S'93–M'96–SM'05–F'11) received the B.S. degree in Electrical and Electronics Engineering from Middle East Technical University, Ankara, Turkey, and the M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, USA. She is currently a Professor of Electrical and Computer engineering with the New York University Tandon School of Engineering, Brooklyn, NY, USA. Her research interests are in information theory, communication theory, and wireless communications.

She is a member of the Science Academy Society of Turkey and is among the 2014 and 2015 Thomson Reuters Highly Cited Researchers. She has been a member of the Board of Governors of the IEEE Information Theory Society since 2012, where she is currently the First Vice President. She was a Distinguished Lecturer of the IEEE Information Theory Society from 2013 to 2014. She received the NSF CAREER Award in 2001 and the IEEE Communications Society WICE Outstanding Achievement Award in 2016. Her paper awards include the IEEE Communications Society Stephen O. Rice Paper Prize in 2004 and the IEEE Communications Society Award for Advances in communication in 2013.

Dr. Erkip has had many editorial and conference organization responsibilities. Some recent ones include the Asilomar Conference on Signals, Systems and Computers, MIMO Communications and Signal Processing Track Chair in 2017, the IEEE Wireless Communications and Networking Conference Technical Co-Chair in 2017, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Guest Editor in 2015, and the IEEE International Symposium of Information Theory General Co-Chair in 2013.

**Sundeep Rangan** (S'94–M'98–SM'13–F'16) received the B.A.Sc. degree from the University of Waterloo, Canada, and the M.Sc. and Ph.D. degrees from the University of California at Berkeley, Berkeley, CA, USA, all in electrical engineering. He has held post-doctoral appointments with the University of Michigan, Ann Arbor, MI, USA, and Bell Labs.

In 2000, he co-founded (with four others) Flarion Technologies, a spin-off of Bell Labs, that developed Flash OFDM, one of the first cellular OFDM data systems and pre-cursor to 4G systems, including LTE and WiMAX. In 2006, Flarion was acquired by Qualcomm Technologies, where he was the Director of Engineering involved in OFDM infrastructure products. He joined the ECE Department, NYU Tandon, Brooklyn, NY, USA, in 2010. He is the Director of NYU WIRELESS, an academic-industry research center researching next-generation wireless systems. His research interests are in wireless communications, signal processing, information theory and control theory.

**Francisco Javier González-Castaño** received the Ingeniero de Telecomunicación degree from the University of Santiago de Compostela, Spain, in 1990, and the Ph.D. degree in telecomunicación from the University of Vigo, Spain, in 1998. He is currently a Catedrático de Universidad (Full Professor) with the Telematics Engineering Department, University of Vigo, Spain, where he leads the Information Technologies Group. He has authored over 80 papers in international journals, in the fields of telecommunications and computer science, and has participated in several relevant national and international projects. He holds two U.S. patents.