MASSIVE MIMO: AN INTRODUCTION

Demand for wireless throughput, both mobile and fixed, will always increase. One can anticipate that, in five or ten years, millions of augmented reality users in a large city will want to transmit and receive holographic video more or less continuously, say 100 megabits per second per user in each direction. Massive MIMO—also called Large-Scale Antenna Systems—is a promising candidate technology for meeting this demand. Fifty-fold or greater spectral efficiency improvements over fourth generation (4G) technology are frequently mentioned. A multiplicity of physically small, individually controlled antennas performs aggressive multiplexing/demultiplexing for all active users, utilizing directly measured channel characteristics. By leveraging time-division duplexing (TDD), Massive MIMO is scalable to any desired degree with respect to the number of service antennas. Adding more antennas is always beneficial for increased throughput, reduced radiated power, uniformly great service everywhere in the cell, and greater simplicity in signal processing. Massive MIMO is a brand new technology that has yet to be reduced to practice. Notwithstanding, its principles of operation are well understood, and surprisingly simple to elucidate.

Thomas L. Marzetta

Introduction

Two timeless truths are evident: first, demand for wireless throughput will always grow; second, the quantity of available electromagnetic spectrum will never increase. Wireless communications is fundamentally different from optical fiber communications in that more fiber can always be manufactured and laid down, so irrespective of the cleverness of optical researchers there is no doubt that any future optical demand will always be met. In contrast, there is no easy solution for wireless throughput.

The fundamental and perennial wireless problem is a physical layer problem: how to provide ever-increasing total wireless throughput reliably and uniformly throughout a designated area [1, 2]. All proposed solutions seem to fall into one of three categories: 1) exploitation of spectrum that is currently unused or underutilized; 2) deployment of ever more access points, each covering a commensurately smaller area; and 3) use of access points and/or terminals with multiple antennas. The first two activities as epitomized by millimeter wave technology and small cells are more than adequately treated elsewhere and will not be discussed in this paper. The third activity is known as MIMO (multiple input multiple output), and it further divides into its original form, Point-to-Point MIMO [3–6], and the later (in terms of its theoretical development) Multi-User MIMO [7–10] of which Massive MIMO is emerging as its ultimate and most useful form [11–14].

Figure 1 illustrates some of the features of Massive MIMO. This could constitute one cell of a network of cells, or it could be an isolated (*single-cell*) site. An array of physically small, non-directive antennas serves a multiplicity of autonomous users (*terminals*). The terminals typically would have only a single antenna each. Downlink operation, shown in the figure, entails transmitting a different data stream to each user. The central task is to ensure that each user receives only the data stream intended for him, with minimal interference from the other data streams. Contemporary systems typically accomplish this multiplexing by some combination of sending the various data streams at different times (*time-division multiplexing*) and over different frequencies (*frequency-division multiplexing*). In contrast, Massive MIMO uses

1538-7305/14©2015ALCATEL-LUCENT

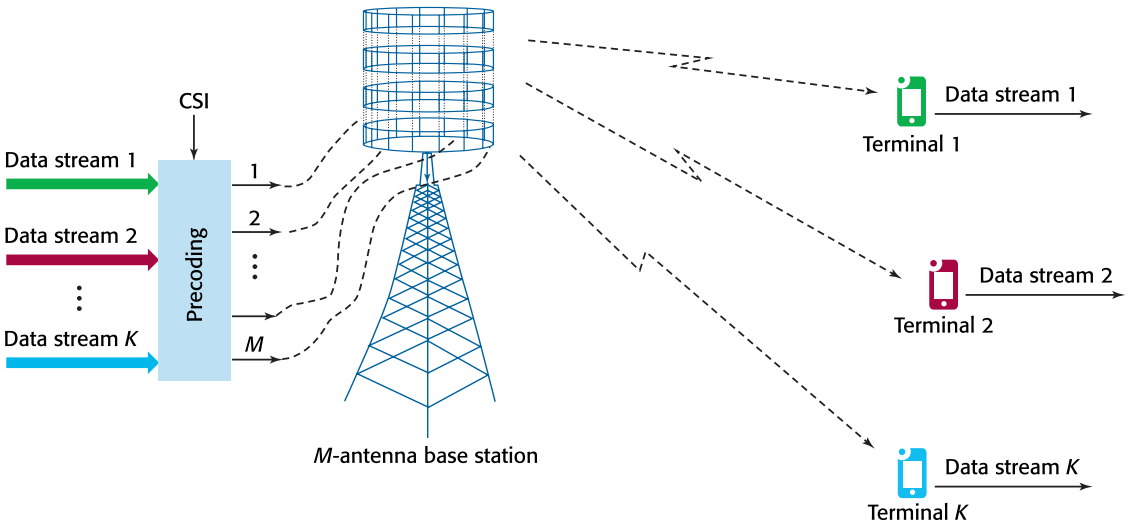

FIGURE 1. **Downlink operation of a Massive MIMO link. The antenna array selectively transmits a multiplicity of data streams, all occupying the same time/frequency resources, so that each user receives only the data stream that it intended for him.**

spatial-division multiplexing such that the different data streams occupy the *same* frequencies and times. A key element to performing wireless spatial multiplexing is an array of independently-controlled antennas. Under line-of-sight propagation conditions, the data streams are carried on focused beams of data. In a cluttered propagation environment, the data streams can arrive from many directions simultaneously. The streams tend to reinforce each other constructively where desired, and interfere destructively where they are unwanted. To carry out the multiplexing, the array needs to know the frequency response of the propagation channels between each of its elements and each of the users. This channel state information (CSI) is utilized in the precoding block (shown in the figure) and it is here that the data streams are mapped into the signals that drive each of the antennas. By increasing the number of antennas, the beams can be focused more selectively to the users.

Demand for wireless throughput will always grow, but the quantity of available electromagnetic spectrum will never increase.

The uplink operation of the Massive MIMO system, shown in Figure 2, is substantially the reverse of downlink operation. The users transmit data streams at the same time and over the same frequencies. The antenna array receives the sum of the data streams as modified by their respective propagation channels, and the decoding operation, again utilizing CSI, untangles the received signals to produce the individual data streams.

Portions of the preceding description of Massive MIMO also apply to a generic Multi-User MIMO system. What distinguishes Massive MIMO is that certain activities make it a scalable technology. The multiplexing and demultiplexing operations are accomplished using directly measured—rather than assumed—channel characteristics. (In contrast, sectorization as well as open-loop angle-selective beam-forming assume line-of-sight propagation conditions, and as explained later, are ultimately not scalable.) If measured channel responses are utilized, increasing the number of antennas always improves the performance of the system irrespective of the noisiness of the measurements. The benefits of growing the number of service antennas relative to the number of active users include greater selectivity in transmitting and receiving the data streams, which leads to greater throughput, a reduction in the required radiated power, effective power control that provides uniformly good service throughout the cell, and greater simplicity in signal processing.

The paper is organized as follows. We begin with an overview of Point-to-Point MIMO, and explain in particular why it is not a scalable technology. Next, we consider the theory of Multi-User MIMO, which has fundamental advantages over Point-to-Point MIMO, but which in its original conception also is not scalable. We then provide a detailed description of Massive MIMO: why it is scalable, how CSI is acquired, the simplest effective types of multiplexing and demultiplexing, power control, and the ultimate limitations of Massive MIMO. We follow with a case study of a dense urban macrocellular Massive MIMO system in which each cell has a 64-antenna base station that serves 18 users. Despite the modest size of this system, it has amazingly good performance, particularly in that it can provide uniformly good service throughout the cell, even at the cell edges. We conclude with a brief

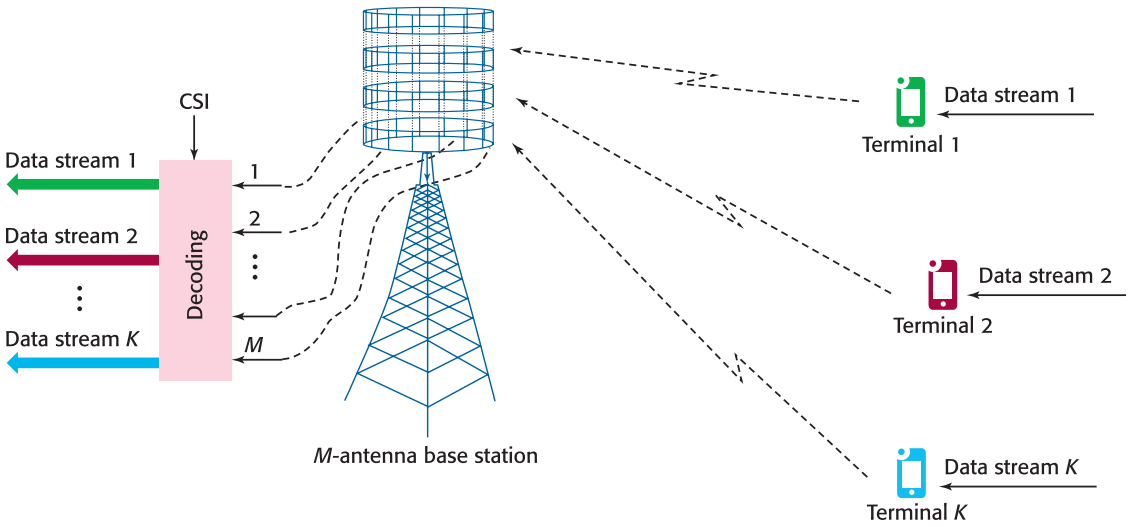

FIGURE 2. **Uplink operation of a Massive MIMO link. The users transmit data streams that occupy the same time/frequency resources, and the signals received by the elements of the antenna array are processed to recover the individual data streams.**

summary of current research issues and describe potential non-cellular applications for Massive MIMO.

Point-to-Point MIMO

Figure 3 depicts a Point-to-Point MIMO link in which a base station with a concentrated array of M antennas transmits data to a user who has a concentrated array of K antennas.

Different users are accommodated in disjoint time/frequency blocks via a combination of time division and frequency division multiplexing. Every use of the channel comprises transmitting a signal vector and receiving a signal vector, where every received signal is a linear combination of transmitted signals, and the combining coefficients are determined by the propagation between the two ends of the link. Subject to a number of assumptions, at sufficiently high signal-to-noise ratios, the spectral efficiency of the link expressed in bits/second/Hz, is approximately as follows:

$$C \propto \min(M, K)\log_2(\rho_{\mathrm{d}}), \quad \rho_{\mathrm{d}} \gg 1, \tag{1}$$

where ρ_{d} is the expected signal-to-noise ratio (SNR) at any receiver if full power were fed into one of the transmit antennas. Thus, without increasing either spectral bandwidth or radiated power, throughput can be increased by adding antennas to each end of the link.

Shannon theory yields a celebrated formula for system capacity, expressed in bits/s/Hz:

$$\begin{aligned} C &= \log_2 \det\left(\mathbf{I}_K + \frac{\rho_{\mathrm{d}}}{M}\mathbf{G}_{\mathrm{d}}^{\mathrm{H}}\mathbf{G}_{\mathrm{d}}\right) \\ &= \log_2 \det\left(\mathbf{I}_M + \frac{\rho_{\mathrm{d}}}{M}\mathbf{G}_{\mathrm{d}}\mathbf{G}_{\mathrm{d}}^{\mathrm{H}}\right), \end{aligned} \tag{2}$$

where $\mathbf{G}_{\mathrm{d}}$ is the $M \times K$ frequency response of the matrix-valued channel that connects the base station antennas and the user antennas, $\mathbf{I}_K$ denotes the $K \times K$ identity matrix, and the superscript "H" denotes "conjugate transpose." The validity of equation 2 depends on the receiver additive noise being complex Gaussian, and more importantly, on the receiver knowing the downlink channel matrix. The transmitter does not have to know the channel, although operation would be both simplified and improved

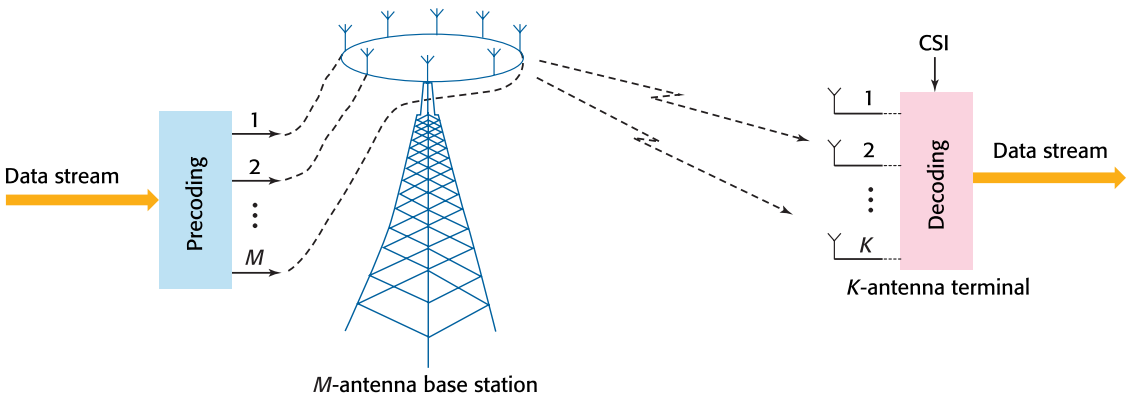

FIGURE 3. **Point to Point MIMO link: A base station equipped with an antenna array serves a user equipped with an antenna array. Different users are served over different time/frequency blocks.**

somewhat if both ends of the link knew the channel. If the elements of $\mathbf{G}_{\mathrm{d}}$ are independent, identically distributed, zero-mean complex Gaussian, unit-variance random variables (i.i.d. *Rayleigh fading*), then asymptotic random matrix theory yields equation 1, a formula which is valid for sufficiently high SNRs.

Equation 2 applies to downlink operation, with M transmit antennas and K receive antennas. For reverse operation on the same link we have K transmit antennas and M receive antennas, and the same capacity formula applies with the term ρ_{d}/M replaced by ρ_{u}/K, and with $\mathbf{G}_{\mathrm{d}}$ replaced by $\mathbf{G}_{\mathrm{u}}$. For FDD systems the uplink channels are different than the downlink channels, while for TDD systems the uplink channels are theoretically equal to the downlink channels.

The intuitive idea behind Point-to-Point MIMO is that the base station transmits a vector-valued signal whose components contain distinct pieces of information. This vector is multiplied by the channel matrix to produce the vector of received signals. If the channel matrix is well-conditioned then the receiver can reliably recover either the distinct transmitted signals, or at least some nontrivial linear combinations of transmitted signals.

In order for the receiver to learn the matrix-valued channel, the transmitter has to send known training signals (*pilots*) through the channel. Subject to peak power restrictions, the optimum training signals are mutually orthogonal [15]. Consequently the sample duration of the pilot sequences should be at least as great as the number of transmit antennas, $\tau_{\mathrm{d}} \geq M$. Uplink data transmission requires the base station to learn the uplink channel matrix which requires uplink pilots such that $\tau_{\mathrm{u}} \geq K$. Hence the total training burden (e.g., the amount of time required for training) for a complete system, either TDD or FDD, grows as

$$\tau_{\mathrm{d}} + \tau_{\mathrm{u}} \geq M + K. \tag{3}$$

Point-to-Point MIMO is an option under current wireless standards. The 802.11ac standard, for example, permits up to $(M, K) = (8, 8)$. For a variety of reasons, Point-to-Point MIMO is not readily scalable beyond 8×8. First, the propagation environment may not support eight data streams. Line-of-sight conditions present a particular challenge because, for compact arrays, the channel matrix has the minimum rank of one which permits only one data stream. Second, the activity of scaling up the number of antennas requires that proportional amounts of time have to be spent on training, as indicated by equation 3. Third, near the edge of the cell, SINRs are typically low, and multiplexing gains fall short of the promised $\min(M, K)$. Fourth, the user equipment is complicated, and independent electronics chains are required for each antenna. Fifth, achieving performance close to the Shannon limit requires rather involved signal processing on the part of both the base station and the user. To illustrate the third point, consider a user having $K = 4$ antennas, operating at an SNR of -3.0 dB, a typical value for cell-edge users. Table I shows the theoretical Shannon capacity (equation 2) for $M = 1, 2, 4, 8$ base station antennas (note that here we take the mean of equation 2 as consistent with ergodic coding over many independent realizations of the random channel). It is clear that barely two data streams are supported, and that most of the efficacy is due to the four receive antennas which collect four-fold signal power.

TABLE I. Shannon Capacity (bits/s/Hz) vs. Number of Base Station Antennas for a Four-Antenna User Having an SNR of −3.0 dB

M	1	2	4	8
C	1.51	1.83	2.06	2.19

As we show in the next section, Multi-User MIMO mitigates some of the drawbacks of Point-to-Point MIMO, but as originally conceived, it still is not scalable.

Multi-User MIMO

A comparison of Figures 1 and 2 with Figure 3 implies that Multi-User MIMO is equivalent to starting with a Point-to-Point MIMO link, and breaking the single K antenna user into K autonomous single-antenna users. Since the users are not capable of communicating with each other, it is evident that the throughput achievable for individual users on poor quality channels can be severely impacted by the break-up. The remarkable and non-intuitive fact is that, subject to certain assumptions about CSI, the *sum throughput* (the sum of the individual capacities to and from the users) is not reduced by the break-up! It is interesting to take note of a paper from 1919 in which directional beamforming from a multiple-antenna array is mentioned as a possible solution to the perceived problem of inadequate spectrum [16]. (At that time it was believed that transoceanic wireless communication was only possible at wavelengths between 10 and 20 km and that the spectrum could only support about 200 worldwide stations.)

The Shannon sum-capacity for uplink Multi-User MIMO is identical to that of uplink Point-to-Point MIMO,

$$C_{\mathrm{sum\,up}} = \log_2 \det\left(\mathbf{I}_K + \frac{\rho_{\mathrm{u}}}{K}\mathbf{G}_{\mathrm{u}}^{\mathrm{H}}\mathbf{G}_{\mathrm{u}}\right), \tag{4}$$

and in both cases, only the base station has to know the uplink channel matrix.

The formula for downlink Shannon sum-capacity requires the solution of a convex optimization problem,

$$C_{\mathrm{sum\,down}} = \sup_{\mathbf{a}}\left\{\log_2 \det\left(\mathbf{I}_M + \rho_{\mathrm{d}}\mathbf{G}_{\mathrm{d}}\mathbf{D}_{\mathbf{a}}\mathbf{G}_{\mathrm{d}}^{\mathrm{H}}\right)\right\}, \quad \mathbf{a} \geq \mathbf{0},\ \mathbf{1}^{\mathrm{T}}\mathbf{a} = 1, \tag{5}$$

where $\mathbf{D}_{\mathbf{a}}$ is a diagonal matrix whose diagonal elements comprise the $M \times 1$ vector, $\mathbf{a}$, and $\mathbf{1}$ denotes the $M \times 1$ vector of ones. Crucially, this capacity is predicated on both ends of the link knowing the downlink channels.

(More precisely, the transmitter has to know the full downlink channel, while each user has to know only its own downlink channel.) A comparison of equation 5 with equation 2 reveals that the downlink multi-user sum capacity actually exceeds that of the point-to-point link which can be attributed to the additional channel state information assumed in the case of the multi-user system. To achieve near-capacity performance requires so-called dirty paper coding/decoding whose computational burden grows exponentially with the size of the system. Moreover, the success of dirty paper coding depends on having very accurate channel estimates.

The base station can acquire uplink CSI through uplink pilots of duration $\tau_u \geq K$. If TDD is employed on the uplink, pilots will also provide the base station with the downlink CSI, because reciprocity implies that the uplink and downlink channels are mathematically identical. Downlink CSI for the users (required for dirty paper decoding) is acquired from orthogonal downlink pilots of duration $\tau_d \geq M$. Hence complete CSI acquisition occupies time (assuming the more favorable TDD) proportional to $\tau_d + \tau_u \geq M + K$.

On the positive side, Multi-User MIMO has two advantages over Point-to-Point MIMO. First, it is less vulnerable to the propagation environment. It can function well even under line-of-sight conditions provided the typical angular separation between users is greater than the angular resolution of the base station array. Second, only single-antenna terminals are required.

What prevents the Shannon-theoretic version of Multi-User MIMO from being scalable is first, the exponentially growing complexity of dirty paper coding/decoding, and second, the time spent acquiring CSI which grows with both the number of service antennas and the number of users.

Massive MIMO: A Scalable Technology

Massive MIMO breaks the scalability barrier by *not* attempting to achieve the full Shannon limit, and, paradoxically, by increasing the size of the system. It departs from Shannon-theoretic practice in three ways. First, only the base station learns the downlink channel. In a TDD system the time required to acquire CSI is independent of the number of base station antennas. Second, the number of base station antennas is typically increased to several times the number of users. Third, a simple linear precoding multiplexing is employed on the downlink, coupled with linear decoding demultiplexing on the uplink. As the number of base station antennas increases, linear precoding and decoding performance can approach the Shannon limit.

Massive MIMO breaks the scalability barrier by *not* attempting to achieve the full Shannon limit, and, paradoxically, by increasing the size of the system.

1. Linear Precoding and Decoding

Figure 4 and Figure 5 illustrate matched filter decoding for uplink data transmission, and conjugate beamforming for downlink data transmission, the simplest type of linear decoding and precoding.

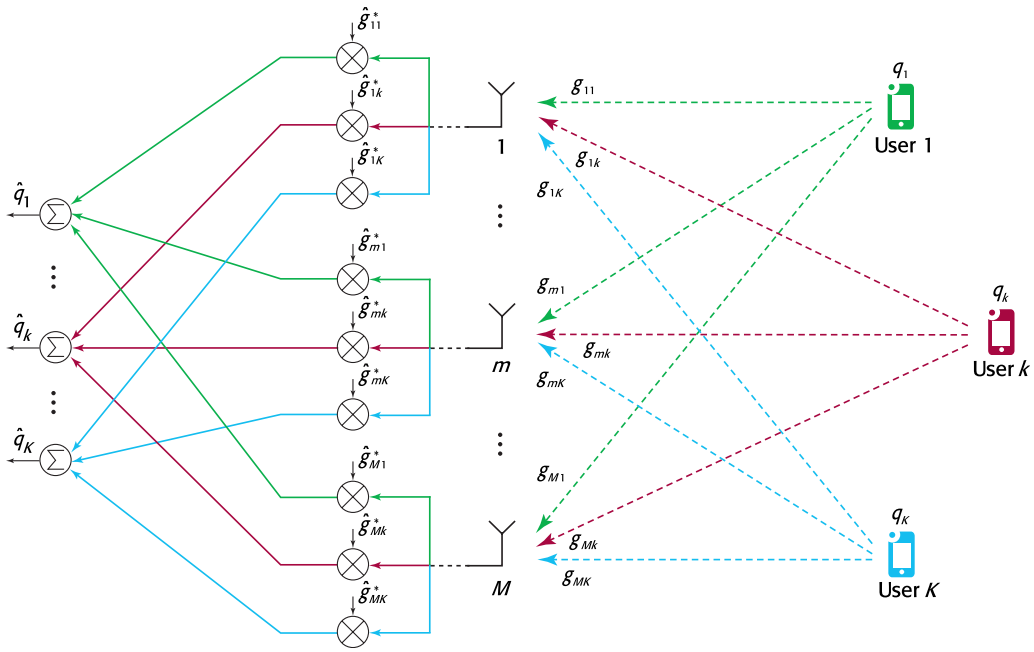

FIGURE 4. Matched filter decoding for uplink Massive MIMO.

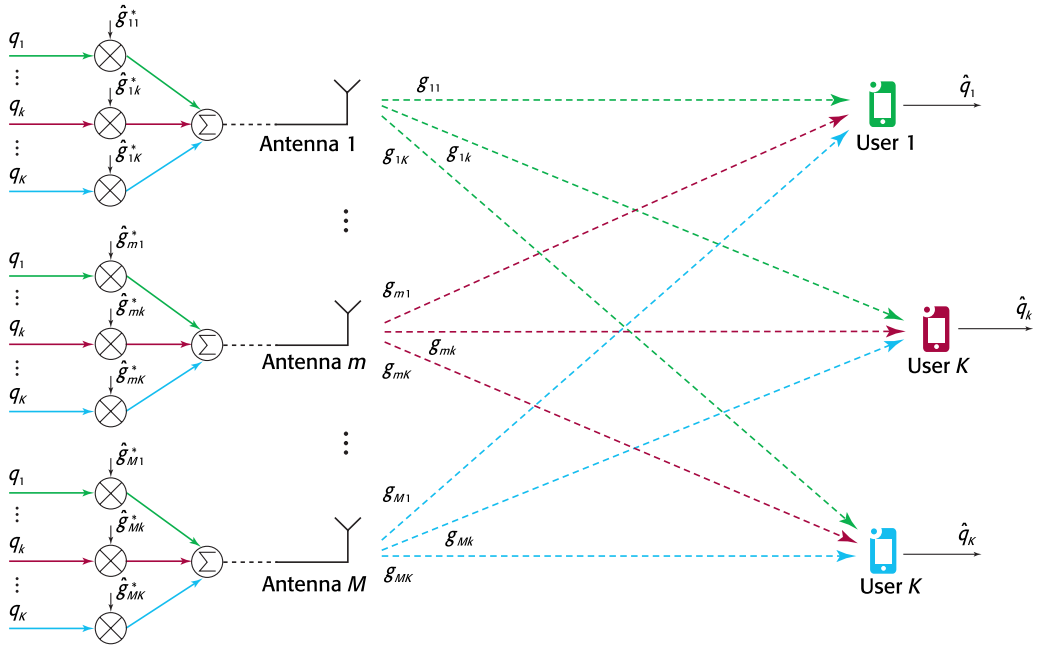

FIGURE 5. **Conjugate beamforming for downlink Massive MIMO.**

On the uplink, the users transmit their data bearing (QAM) symbols simultaneously, with no MIMO-specific signal processing, except for optional power control weighting (which is not shown in the figure). The k-th user transmits a QAM symbol denoted by q_k. The frequency response between the k-th user and the m-th base station antenna is denoted by g_{mk} where the dependence on frequency is not explicitly shown. To recover the k-th QAM symbol, the base station weights its m-th received signal by the complex conjugate of its estimate for the mk-th channel coefficient, $\hat{g}_{mk}$, and sums this weighted signal over the M antennas to produce $\hat{q}_k$. With only one user, this processing would be optimal for enhancing SNR, but multiple users will generate interference (crosstalk). Under *favorable propagation conditions*, which experimentally have been shown to prevail [12, 17], the vector-valued channels from one user to another are asymptotically orthogonal as M increases, and the expected power of the desired signal grows M times faster than the power of the interference.

On the downlink, the k-th QAM symbol, q_k (with optional power weighting), is multiplied by the complex conjugate of the estimate for the mk-th channel coefficient, $\hat{g}_{mk}$, and the weighted signals are summed over the K users to produce the signal that is fed into the m-th antenna. The weighting ensures that the components of the M transmitted signals that are associated with the k-th QAM symbol arrive in phase at the k-th user. For a single user, this weighting maximizes the received power for a total given transmitted power. Again, the asymptotic orthogonality of channels under favorable propagation implies that power for desirable signals grows M times faster than the power of the crosstalk interference.

A natural question follows: How do imperfect channel estimates affect the performance of linear precoding and decoding? In particular, as we add more antennas, does the quality of the CSI have to improve in order to maintain coherent gains proportional to M? The definitive answer is favorable: irrespective of the noisiness of the channel estimates, the coherent gain grows in proportion to M [11], albeit with a proportionality factor that decreases with increasing estimation error.

It is also possible to use other types of linear precoding and decoding. In general, linear decoding consists of multiplying the received vector by a $K \times M$ matrix, and linear precoding multiplies the vector of QAM symbols by an $M \times K$ matrix, where the matrices are functions of the channel estimates. For matched filtering and conjugate beamforming, the matrices are complex conjugates of the estimated channel matrix. Zero-forcing, instead, uses the pseudo-inverse of the estimated channel matrix. Under high SNR conditions, zero-forcing may perform significantly better than matched filtering and conjugate beamforming [18]. An advantage of matched filtering and conjugate beamforming is that the Massive MIMO signal processing can be performed locally at each antenna, as is apparent from Figures 4 and 5. This in turn permits a decentralized architecture for the antenna array, which lends great resilience to the system. For example, if half the antennas are lost from a lightning strike, the remaining antennas do exactly what they did before. Likewise, during periods of slack demand, some antennas can be put into sleep mode, for improved energy efficiency, without affecting the operations of the others.

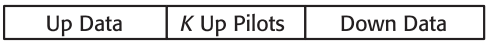

FIGURE 6. Slot structure for TDD Massive MIMO.

2. Channel Estimation

Massive MIMO relies on measuring the frequency responses of the actual propagation channels. To that end, either the users or the base station transmit known training signals and the receiver opposite then estimates the frequency response. Once the channels have been estimated, the CSI has to be utilized in a timely manner before the motion of the users significantly changes the channels. Hence there is only a limited amount of time available for training. In a Massive MIMO system both training and data transmission take place in a slot whose duration is chosen so that nobody moves more than a fraction of a wavelength within the slot.

In a TDD system, the users transmit orthogonal pilot sequences, of sample duration $\tau_u \geq K$. It is convenient to think of the frequency response as piece-wise constant over intervals defined by the Nyquist sampling interval—equal in Hertz to the reciprocal of the channel delay-spread. Hence, over each Nyquist interval the terminals transmit orthogonal pilot sequences. For typical OFDM parameters (symbol interval $T_s = 1/14$ ms, usable interval $T_u = 1/15$ ms, guard interval $T_g = 1/(14 \cdot 15)$ ms), if the delay-spread is assumed to be equal to the guard interval, the Nyquist interval is equivalent to $\frac{T_u}{T_g} = 14$ tones. In a slot of duration (in seconds) T_{slot}, there are $\frac{T_{slot}}{T_s}$ OFDM symbols. We define the sample duration of the slot as the number of OFDM symbols times the tone duration of the Nyquist interval, $T = \frac{T_{slot} T_u}{T_s T_g}$; this is the number of channel uses in each Nyquist interval of each slot. Shorter slots permit higher mobility: at 1.9 GHz a 280 km/h user would move 1/4 wavelength during a 500 μs slot. Thus a 500 μs slot has an equivalent sample duration of $T = 7 \times 14 = 98$, so as many as 98 terminals could be trained. However this upper limit would leave no time for transmitting data, and it can be shown that under general conditions no more than half of the slot should be expended for training. Figure 6 illustrates the TDD slot structure.

An FDD system requires considerably more time for training than a TDD system [19]. First, the M base station antennas transmit orthogonal downlink pilots with $\tau_d \geq M$. Next, the K users have to transmit their M received pilot signals on the uplink, which expends an additional τ_d samples of resources. Finally, the K users have to transmit ordinary uplink pilot sequences with $\tau_u \geq K$. Figure 7 shows the structures of the FDD downlink and uplink slots. The entire process requires a minimum of $2\tau_d + \tau_u \geq 2M + K$ channel uses, in contrast to $\tau_u \geq K$ channel uses for TDD.

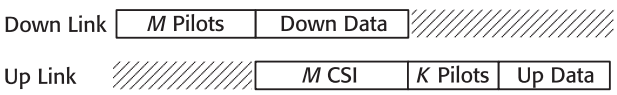

FIGURE 7. Slot structure for FDD Massive MIMO.

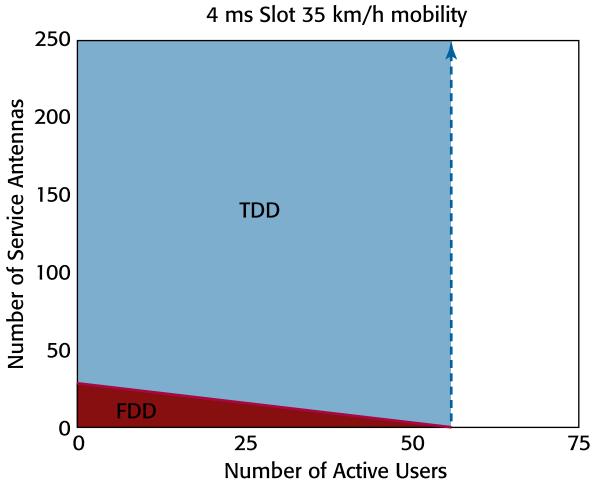

FIGURE 8. TDD enables Massive MIMO scalability for mobile users, while FDD is strictly limited to small systems.

Figure 8 shows the overwhelming advantage of TDD over FDD for mobile users. The vertical axis is the number of base station antennas, and the horizontal axis is the number of users. The light-to-medium blue region shows the system dimensions obtainable with TDD versus the much smaller red region for FDD. The slot duration is 4 ms which permits 35 km/h mobility at 1.9 GHz, and to mitigate pilot contamination (explained later) the orthogonal pilot sequences are made seven times longer than necessary.

3. The Importance of Using Measured Rather Than Assumed Channel Characteristics

Massive MIMO utilizes channel information that is directly measured rather than assumed. This, in turn, makes Massive MIMO a scalable technology: any number of base station antennas can be usefully employed with no tightening of array tolerances. Extra antennas always help.

In contrast, if an assumed channel response is used, the technology is ultimately not scalable. One technique for downlink beamforming is to form open-loop beams, transmit a downlink pilot through each beam, and, on the uplink, have each user report back which beam is the strongest. The base station then transmits downlink data to each user through its preferred beam. The open loop beams are directional beams subject to the assumption of line of sight propagation. While this scheme may work up to a point, eventually the addition of more antennas yields little further improvement. This happens for two reasons. First, any real propagation environment has some non-zero angle spread: one can transmit ever-finer beams, but the propagation medium itself will broaden the beams. To put it another way, there is a mismatch between any of the fine open loop beams and the actual propagation to the user. The second reason for lack of scalability is that even if the propagation were perfect line-of-sight, the activity of forming ever-finer open loop beams would require increasingly tight array tolerances. Figure 9 illustrates the

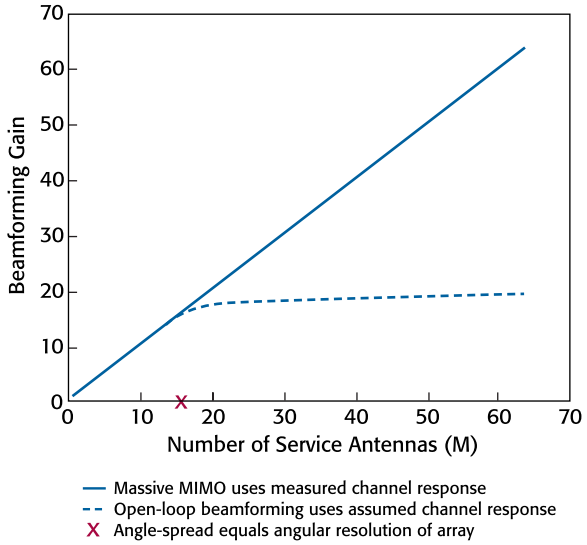

FIGURE 9. **The array gain of Massive MIMO always grows linearly with the number of antennas. Open-loop beamforming eventually yields only logarithmic improvements in array gain.**

divergence of array gain achievable by Massive MIMO based on the measured channels and the array gain achieved by open-loop beamforming. Massive MIMO gain increases linearly with the number of antennas, but under open-loop beamforming, when the beamwidth becomes less than the angle spread of the medium, the array gain increases only logarithmically with the number of antennas.

4. Power Control

As indicated in an earlier section, power control, in both uplink and downlink, entails multiplying the original set of K QAM symbols by power control coefficients,

$$\text{downlink}: q_k \rightarrow \sqrt{\eta_k} q_k,\ k = 1, \cdots, K, \quad \sum_{k=1}^{K} \eta_k \leq 1$$

$$\text{uplink}: q_k \rightarrow \sqrt{\eta_k} q_k,\ \eta_k \leq 1,\ k = 1, \cdots, K. \qquad (6)$$

One of the nice properties of Massive MIMO is that the large number of antennas makes the beamforming gains virtually constant over frequency, and moreover, dependent only on large-scale (*slow*) fading coefficients, which themselves are independent of frequency and of antenna index. Hence the power control coefficients can be made independent of frequency and their effect on the data rate attained by an individual user may be computed without regard to the short term channel estimates obtained from the pilots. It happens that, in the mathematical expressions for the effective SINRs of the precoded/decoded signals, the power control coefficients enter the expressions in a way such that inequality constraints on SINRs are equivalent to linear inequality constraints on power control coefficients. So max-min power control, which gives equal throughput to everyone in the cell, is obtained by solving a set of linear equations for the power control coefficients.

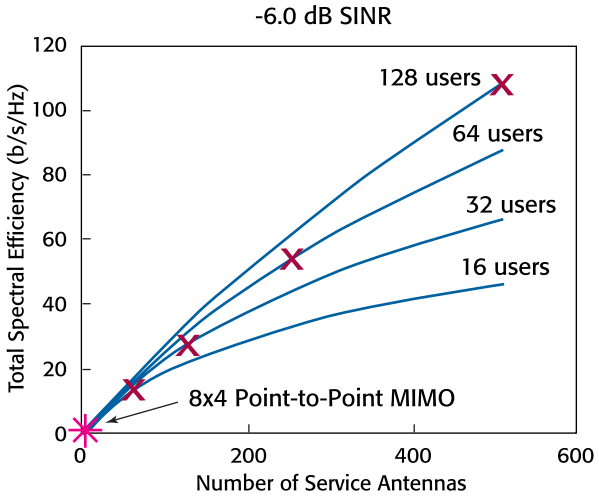

FIGURE 10. **Total spectral efficiency versus number of base station antennas for K equal to 16, 32, 64, 128 users operating at a minus 6 dB SINR. Asterisk indicates performance of 8 × 4 Point-to-Point MIMO. Crosses indicate $M = 4K$ operating points.**

Other strategies are possible. For example one could specify desired rates for a subset of users, and subject to these constraints use linear programming to find the power control coefficients that yield max-min throughput for the remaining users.

5. Performance of Massive MIMO

Massive MIMO can operate in a regime unavailable to Point-to-Point MIMO, as illustrated in Figure 10, comprising plots of sum spectral efficiency for $K = [16\ \ 32\ \ 64\ \ 128]$ users as functions of M for an SINR of -6.0 dB. The red X's correspond to the dimensions, $M = 4K$. The point $(M, K) = (64, 16)$ yields a total spectral efficiency of 13.6 bits/s/Hz which is doubled for every simultaneous doubling of (M, K). In contrast, 8 × 4 Point-to-Point MIMO has a spectral efficiency

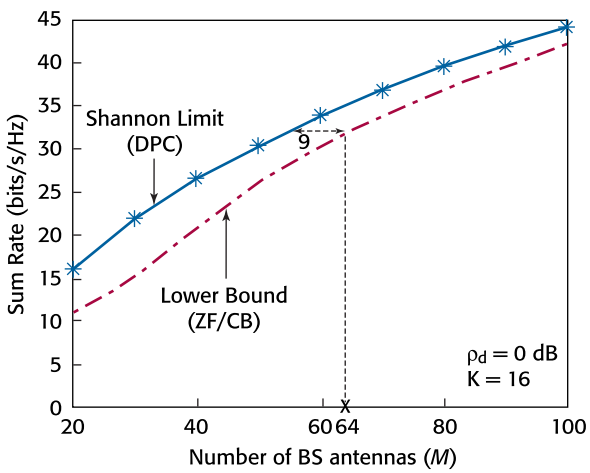

FIGURE 11. **Total spectral efficiency versus number of base station antennas for $K = 16$ users and 0.0 dB SINR. Blue curve: Shannon limit (dirty-paper coding); Red curve; Massive MIMO (linear pre-coding).**

of only 1.3 bits/s/Hz. For these results perfect CSI is assumed. The Point-to-Point MIMO performance is ergodic Shannon capacity according to equation 2. Massive MIMO performance is computed as a capacity lower-bound for conjugate beamforming according to a formula derived in [18]:

$$C_{\text{sum cb}} > K \log_2\left(1 + \frac{M\rho_{\text{d}}}{K(1+\rho_{\text{d}})}\right). \qquad (7)$$

Figure 11 shows that by employing additional base station antennas, the linear precoding used in Massive MIMO is highly competitive with the dirty-paper coding mandated by Shannon theory. For $K = 16$ users and an SINR of 0.0 dB, the blue curve represents the Shannon total spectral efficiency, and the red curve is the total spectral efficiency for Massive MIMO with linear precoding, as a function of the number of base station antennas. For $M = 64$ antennas, Massive MIMO gives the same performance as dirty-paper coding for $M = 55$. Here the Shannon limit is computed according to equation 5, while the linear precoding lower bound is computed as the greater of the conjugate beamforming lower bound (equation 7) or the zero-forcing lower bound [18]:

$$C_{\text{sum zf}} > K \log_2\left(1 + \frac{(M-K)\rho_{\text{d}}}{K}\right). \qquad (8)$$

6. Ultimate Limitation of Massive MIMO

When serving mobile users, there is a finite limit to the number of users that can be served simultaneously because of the overhead required for CSI acquisition. Of course one can add as many base station antennas as desired which will increase proportionally the SINR experienced by the users, but this ultimately yields only logarithmic improvements in throughput. Short of a breakthrough in channel estimation, extremely large arrays (say conformal arrays on the sides of skyscrapers) should be reserved primarily for fixed terminals, where theoretically, there is an unlimited amount of time for training.

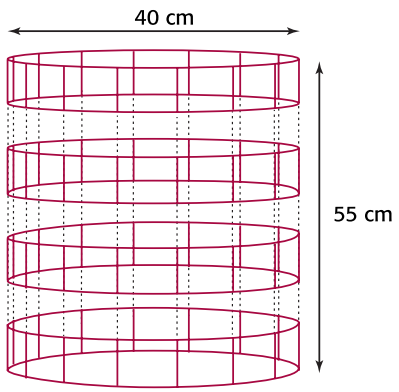

FIGURE 12. Antenna array for 1.9 GHz: Each of 64 antennas is 1/2 wavelength by 1/2 wavelength in size (8 cm × 8 cm), spaced 1/2 wavelength apart circumferentially, and spaced one-wavelength apart vertically.

Case Study: Dense Urban Macro-Cellular Massive MIMO

We next examine a multi-cellular scenario in which each cell, containing on average 18 randomly located users, is served by a base station with a 64-element array. The cell radii (center to vertex) are 500 meters, the carrier frequency is 1.9 GHz, and the spectral bandwidth is 20 MHz. The 64-element array, comprising 1/2 wavelength patch antennas arranged in a cylindrical configuration, is only 40 cm in diameter and 55 cm high, as shown in Figure 12. A persistent myth—easily refuted by Figure 12—is that Massive MIMO is not practical at anything but millimeter wavelengths because of the allegedly huge physical size of the array. The maximum total downlink radiated power of the base station is one Watt (16 mW per antenna), and the maximum uplink radiated power of each terminal is 200 mW.

1. Training, Pilot Contamination, and Slot Structure

The TDD slot duration is 2 ms, permitting 71 km/h mobility, and according to the OFDM parameters discussed in the section on Massive MIMO channel estimation, the sample duration of the slot is $T = 392$. Pilot sequences of duration 18 would be sufficient to ensure orthogonality within each cell, but re-use of the same 18 pilots from cell to cell gives rise to *pilot contamination* [14]. Pilot re-use implies that the pilot-derived estimate for the channel between the home base station and one of its users is contaminated by channels between the base station and users in other cells which share the same pilot sequence. On downlink, the base station inadvertently transmits coherent interference to the users in other cells. This impairment does not decrease with the addition of more antennas. A

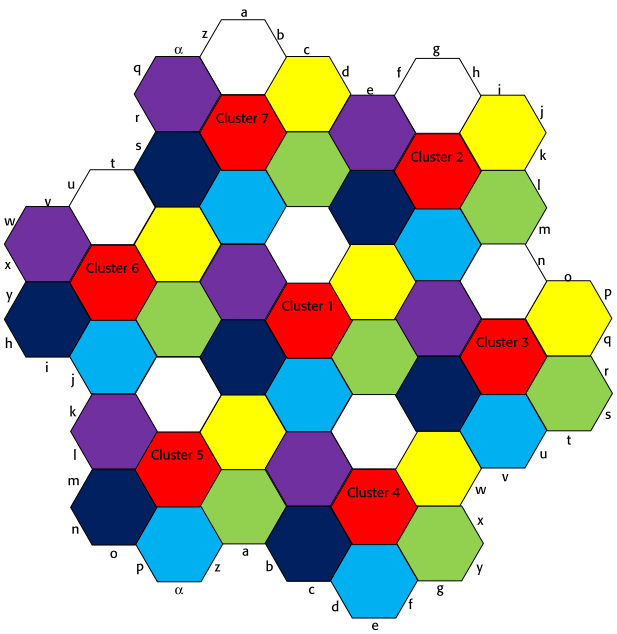

FIGURE 13. Pilot re-use seven. Cells having different colors have mutually orthogonal pilot sequences.

similar effect occurs on the uplink. A simple measure to mitigate pilot contamination is to use pilot sequences of duration $7 \times 18 = 126$ which enables a cluster of seven cells to assign mutually orthogonal pilot sequences to all users. As shown in Figure 13, the home cell is surrounded by two concentric rings of non-contaminating cells. Pilot contamination does arise from six cells in the third ring, but the coherent interference is reduced by 40 dB or so [20, 21]. The fraction of time spent on training is $\frac{126}{392} = .32$. The remaining fraction of the slot is divided evenly into uplink and downlink data transmission intervals.

2. Propagation Modeling

The propagation model comprises slow fading (long scale) and fast fading (short scale). Slow fading accounts for range-dependent attenuation according to the COST231 dense urban model in combination with log-normal shadow fading. The fast fading is independent Rayleigh.

3. Multiplexing Precoding/Decoding and Power Control

Precoding and decoding consist of conjugate beamforming and matched filtering. Max-min power control gives equal rate service to the users in each cell.

4. Semi-Analytical Simulations

Our performance calculations are based on extensions of the capacity bound (equation 7) that account for receiver noise, random channel instantiations, channel estimation error, the overhead associated with pilot transmissions, the imperfections of conjugate beamforming and matched filtering, power control, non-coherent inter-cell interference, and coherent inter-cell interference (due to pilot contamination) [21]. These capacity bounds only involve slow fading coefficients. One hundred simulations are performed.

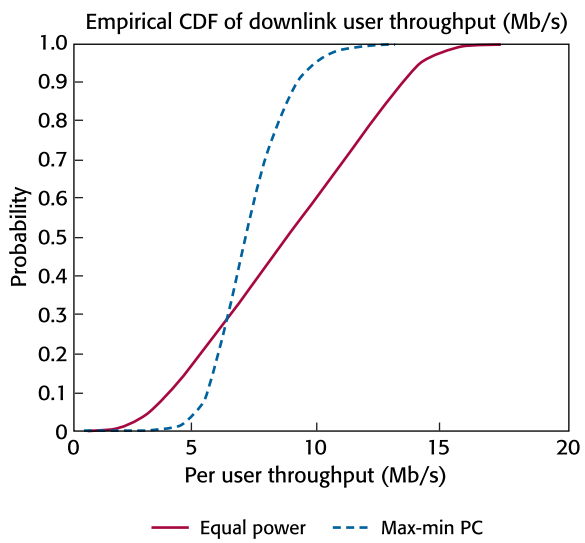

FIGURE 14. Cumulative distribution of downlink per-user net throughput (Mb/s). Red solid curve: equal power for all users; black dashed curve: max-min power control.

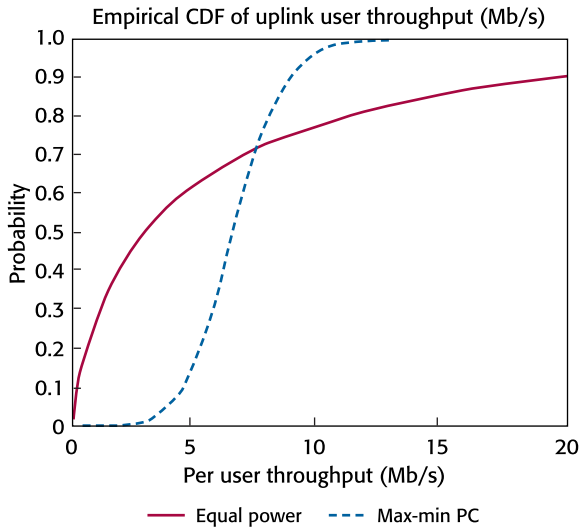

FIGURE 15. Cumulative distribution of uplink per-user net throughput (Mb/s). Red solid curve: equal power for all users; black dashed curve: max-min power control.

Each simulation involves 49 cells as shown in Figure 13 with periodic wrap-around and a total of 18×49 randomly located users with random shadow fading. Max-min power control parameters are derived from the slow fading coefficients. As is common in cellular practice, the 5% most troublesome users are dropped from service.

5. Numerical Results

Figures 14 and 15 show the cumulative distributions of per-user net throughput (megabits/s) for the downlink and the uplink respectively (excluding the 5% of the users who were dropped from service).

Consider first the downlink results captured in Figure 14. Note that max-min power control reduces the median net throughput per user, but advantageously flattens the cumulative distribution. The median net throughput is about 7 Mb/s per user, and the 95% likely throughput is 4.8 Mb/s. The extraordinary thing is that each of the users experiences the same high throughput irrespective of his location in the cell.

Power control is even more important on the uplink (modeled in Figure 15) than on the downlink. Users close to the base station are in a position to disrupt the transmissions of users far away. With max-min power control, median net throughput per user is about 6.5 Mb/s, and the 95% likely throughput is 3.3 Mb/s. Note that uplink performance is not significantly inferior to downlink performance. The total median net system spectral efficiency is 11.5 bits/s/Hz.

The system could be scaled-up in various ways. Extra antennas would increase the SINR of every user, resulting in higher throughput. Simultaneously doubling the number of service antennas, the number of users, and the slot duration (while reducing the mobility by a factor of two) would double the system throughput.

6. Downlink Energy Efficiency

Total energy efficiency for the Massive MIMO base station is estimated at 7.8 Mb/Joule, using GreenTouch parameters for internal power consumption [22], which comprises RF power generation, Massive MIMO computing, and the power consumed by internal electronics.

Research Issues and Non-Cellular Applications

Despite its great promise, Massive MIMO has not yet been reduced to practice. A number of issues still need to be tackled, and there is much ongoing research. Further, the potential for leveraging Massive MIMO extends beyond cellular deployments.

1. Issues

Massive MIMO replaces costly instrument-grade 40 Watt transceivers with a large number of low-power and possibly low-precision units. Ideally each antenna would be contained in an inexpensive module containing all electronics, signal processing, and a small power amplifier and as many of these modules as desired could be assembled like Lego bricks. This represents an entirely new design philosophy for which a low cost solution is needed.

Massive MIMO will likely require new standards.

Most Massive MIMO research has focused on the physical layer. Higher layers are a possible challenge. For example, when a new user joins the system, the base station cannot utilize the full selectivity of the array until after the preliminary handshaking tasks have been completed, and the user has been assigned a pilot sequence.

Massive MIMO automatically provides great gains in radiated energy efficiency. To achieve total energy efficiency there has to be a commensurate reduction in internal power consumption, for which there has been little incentive in the past.

Massive MIMO relies on *favorable propagation*—vector-valued channels to different users grow asymptotically orthogonal with the increasing numbers of antennas. Experiments so far support this hypothesis but many more experiments are required.

Although there is a great deal of theoretical and simulation evidence for the vast superiority of Massive MIMO over 4G technology, public demonstrations of Massive MIMO on a sufficiently large scale will be required to convince many people of its potential.

2. Massive MIMO Research Trends

Massive MIMO is dependent upon the ability to acquire channel information, and user mobility appears to impose a limit on the number of active users who can be served. Considerable research is underway to mitigate these limitations [23–30]. Signal processing algorithms and modulation formats are attracting much attention [31–36]. Additional research is aimed at CSI acquisition either with the object of making FDD work in mobile environments, or simply accommodating more users with TDD [37–40].

3. Non-Cellular Applications

The principles of Massive MIMO are not confined to concentrated antenna arrays, but could also be applied to distributed (*cell free*) deployments. Billboard-sized arrays in suburban and rural locations could provide high-speed fixed wireless access to hundreds or thousands of homes. Uplink Massive MIMO could transport data continuously from tens of thousands of sensors to an access point.

Acknowledgements

The author thanks Hien Ngo and Hong Yang for help with the numerical results and figures. The quality of the paper was greatly improved due to suggestions from anonymous reviewers, as well as Alexei Ashikhmin, Oliver Blume, Thierry Klein, Erik Larsson, Iraj Saniee, and Chris White. Special thanks to the Editor of BLTJ, Jane DeHaven.

References

[1] F. Boccardi, R. W. Heath, 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.

[2] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What Will 5G Be?" IEEE J. Sel. Areas. Commun., vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[3] G. J. Foschini and M. J. Gans, "On Limits of Wireless Communications in a Fading Environment When Using Multiple Antennas," Wireless Pers. Commun., vol. 6, no. 3, pp. 311–335, Mar. 1998.

[4] E. Telatar, "Capacity of Multi-Antenna Gaussian Channels," Eur. Trans. Telecommun., vol. 10, no. 6, pp. 585–595, Nov./Dec. 1999.

[5] G. G. Raleigh and J. M. Cioffi, "Spatio-Temporal Coding for Wireless Communication," IEEE Trans. Commun., vol. 46, no. 3, pp. 357–366, Mar. 1998.

[6] A. Paulraj and T. Kailath, "Increasing Capacity in Wireless Broadcast Systems Using Distributed Transmission/Directional Reception (DTDR)," U.S. Patent 5 345 599, Sep. 6, 1994.

[7] D. Gesbert, M. Kountouris, R. W. HeathJr., C. Chae, and T. Chae, "Shifting the MIMO Paradigm," IEEE Signal Process. Mag., vol. 24, no. 5, pp. 36–46, Sep. 2007.

[8] G. Caire and S. Shamai, "On the Achievable Throughput of a Multi-Antenna Gaussian Broadcast Channel," IEEE Trans. Inf. Theory, vol. 49, no. 7, pp. 1691–1706, Jul. 2003.

[9] P. Viswanath and D. N. C. Tse, "Sum Capacity of a Vector Gaussian Broadcast Channel and Uplink-Downlink Duality," IEEE Trans. Inf. Theory, vol. 49, no. 8, pp. 1912–1921, Aug. 2003.

[10] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, Achievable Rates, Sum-Rate Capacity of Gaussian MIMO Broadcast Channels," IEEE Trans. Inf. Theory, vol. 49, no. 10, pp. 658–2668, Oct. 2003.

[11] T. L. Marzetta, "How Much Training is Required for Multiuser MIMO," in Proc. 40th Asilomar Conf. Signals, Syst., Comput., Nov. 2006, pp. 359–363.

[12] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and Challenges With Very Large Arrays," IEEE Signal Process. Mag., vol. 30, no. 1, pp. 40–60, Jan. 2013.

[13] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for Next Generation Wireless Systems," IEEE Commun. Mag., vol. 52, no. 2, pp. 186–195, Feb. 2014.

[14] T. L. Marzetta, "Noncooperative Cellular Wireless With Unlimited Numbers of Base Station Antennas," IEEE Trans. Wireless Commun., vol. 9, no. 1, pp. 3590–3600, Nov. 2010.
[15] T. L. Marzetta, "BLAST Training: Estimating Channel Characteristics for High-Capacity Space-Time Wireless," in Proc. 37th Annu. Allerton Conf. Commun., Control, Comput., Monticello, IL, USA, 1999, pp. 958–966.
[16] E. F. W. Alexanderson, "Transoceanic Radio Communication," Trans. Amer. Inst. Elect. Eng., vol. 38, no. 2, pp. 1269–1285, Jul.–Dec. 1919, reprinted Proc. IEEE, vol. 72, no. 5, May 1984.
[17] X. Gao, O. Edfors, F. Rusek, and F. Tufvesson, "Massive MIMO in Real Propagation Environments," IEEE Trans. Wireless Commun., Mar. 2014.
[18] H. Yang and T. L. Marzetta, "Performance of Conjugate and Zero-forcing Beamforming in Large-Scale Antenna Systems," IEEE J. Sel. Areas Commun., vol. 31, no. 2, pp. 172–179, Feb. 2013.
[19] T. L. Marzetta and B. M. Hochwald, "Fast Transfer of Channel State Information in Wireless Systems," IEEE Trans. Signal Process., vol. 54, no. 4, pp. 1268–1278, Apr. 2006.
[20] H. Yang and T. L. Marzetta, "A Macro Cellular Wireless Network With Uniformly High User Throughputs," in Proc. IEEE Veh. Technol. Conf., Sep. 2014, pp. 1–5.
[21] H. Yang and T. L. Marzetta, "Capacity Performance of Multicell Large Scale Antenna Systems," in Proc. 51st Allerton Conf. Commun., Control, Comput., Oct. 2013, pp. 668–675.
[22] H. Yang and T. L. Marzetta, "Total Energy Efficiency of Cellular Large Scale Antenna System Multiple Access Mobile Networks," in Proc. IEEE Online Conf. Green Commun., Oct. 2013, pp. 27–32.
[23] 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, Aug. 2011.
[24] J. Choi, D.J. Love, and P. Bidigare, "Downlink Training Techniques for FDD Massive MIMO Systems: Open-Loop and Closed-Loop Training With Memory," IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 802–814, Oct. 2014.
[25] K. T. Truong and R. W. Heath, Jr., "Effects of Channel Aging in Massive MIMO Systems," J. Commun. Netw.—Special Issue Massive MIMO, vol. 15, no. 4, pp. 338–351, Aug. 2013.
[26] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A Coordinated Approach to Channel Estimation in Large-Scale Multiple-Antenna Systems," IEEE J. Sel. Areas Commun., vol. 31, no. 2, pp. 264–273, Feb. 2013.
[27] J. Nam, J.-Y. Ahn, A. Adhikary, and G. Caire, "Joint Spatial Division and Multiplexing: Realizing Massive MIMO Gains With Limited Channel State Information," in Proc. CISS, Princeton, NJ, USA, Mar. 2012, pp. 1–6.
[28] S. Noh, M. D. Zoltowski, Y. Sung, and D. J. Love, "Pilot Beam Pattern Design for Channel Estimation in Massive MIMO Systems," IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 787–801, Oct. 2014.
[29] N. Shariati, E. Bjornson, M. Bengtsson, and M. Debbah, "Low Complexity Polynomial Channel Estimation in Large Scale MIMO With Arbitrary Statistics," IEEE J. Sel. Topics Signal Process., vol. 8, no. 5, pp. 815–830, Oct. 2014.
[30] K. T. Truong, A. Lozano, and R. W. Heath, Jr., "Optimal Training in Continuous Flat-Fading Massive MIMO Systems," in Proc. European Wireless Conf., Barcelona, Spain, May 2014, pp. 1–6.
[31] A. Pitarokoilis, S. K. Mohammed, and E. G. Larsson, "On the Optimality of Single-Carrier Transmission in Large-Scale Antenna Systems," IEEE Wireless Commun. Lett., vol. 1, no. 4, pp. 276–279, Aug. 2012.
[32] 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.
[33] H. Huh, G. Caire, H. C. Papadopoulos, and S. A. Ramprashad, "Achieving 'Massive MIMO' Spectral Efficiency With a Not-so-Large Number of Antennas," IEEE Trans. Wireless Commun., vol. 11, no. 9, pp. 3226–3239, Sept. 2012.
[34] S. K. Mohammed and E. G. Larsson, "Per-Antenna Constant Envelope Precoding for Large Multi-User MIMO Systems," IEEE Trans. Commun., vol. 61, no. 3, pp. 1059–1071, Mar. 2013.
[35] S. K. Mohammed and E. G. Larsson, "Constant-Envelope Multi-User Precoding for Frequency-Selective Massive MIMO Systems," IEEE Wireless Commun. Lett., vol. 2, no. 5, pp. 547–550, Oct. 2013.
[36] C. Studer and E. G. Larsson, "PAR-Aware Large-Scale Multi-User MIMO-OFDM Downlink," IEEE J. Sel. Areas Commun., vol. 31, no. 2, pp. 303–313, Feb. 2013.
[37] J. Hoydis, K. Hosseini, S. ten Brink, and M. Debbah, "Making Smart Use of Excess Antennas: Massive MIMO, Small Cells, TDD," Bell Labs Tech. J., vol. 18, no. 2, pp. 5–21, Sep. 2013.
[38] K. T. Truong, and R. W. Heath, Jr., "The Viability of Distributed Antennas for Massive MIMO Systems," in Proc. Asilomar Conf. Signals, Syst., Comput., Pacific Grove, CA, USA, Nov. 3–6, 2013, pp. 1318–1323.
[39] E. Bjornson, M. Kountouris, and M. Debbah, "Massive MIMO and Small Cells: Improving Energy Efficiency by Optimal Soft-Cell Coordination," in Proc. ICT, May 2013, pp. 1–5.
[40] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achieveable Rates of FDD Massive MIMO Systems with Spatial Channel Correlation," IEEE Trans. Wireless Commun., to be published.

(Manuscript approved January 2015)

Author

Thomas L. Marzetta heads the Large-Scale Antenna Systems Group in the Network Energy Program at Bell Labs in Murray Hill, New Jersey. He received a Ph.D. and an S.B. in electrical engineering from Massachusetts Institute of Technology in 1978 and 1972, and an M.S. in systems engineering from the University of Pennsylvania in 1973. He worked for Schlumberger-Doll Research in petroleum exploration and for Nichols Research Corporation in defense research before joining Bell Labs in 1995. He served as director of the Communications and Statistical Sciences Department within the former Mathematical Sciences Research Center. He is the originator of Massive MIMO, a promising technology for addressing the ever increasing demand for wireless throughput. Currently Dr. Marzetta serves as coordinator of the GreenTouch Consortium's Large-Scale Antenna Systems Project, and as member of the advisory board of MAMMOET (Massive MIMO for Efficient Transmission), an EU-sponsored FP7 project. Dr. Marzetta was the recipient of the 1981 ASSP Paper Award from the IEEE Signal Processing Society. He was elected a Fellow of the IEEE in 2003. He became a Bell Labs Fellow in 2014. For his contributions to Massive MIMO he received the 2013 IEEE Guglielmo Marconi Best Paper Award, the 2013 IEEE OnLineGreenComm Conference Best Paper Award, the 2014 GreenTouch Consortium 1000x Award, the 2014 Thomas Alva Edison patent award in telecommunications, and the 2015 IEEE W. R. G. Baker Award. ▼