Asynchronously Coordinated Multi-timescale beamforming architecture for multi-cell networks

Antonios Michaloliakos, Weng C. Ao, Konstantinos Psounis and Yonglong Zhang

***Abstract*—Modern wireless devices like smartphones are pushing the demand for higher wireless data rates. The ensuing increase in wireless traffic demand can be met by a denser deployment of access points (APs), coupled with a coordinated deployment of advanced physical layer techniques to reduce inter-cell interference. Unfortunately, advanced physical layer techniques, e.g. multi-user (MU) MIMO found in 802.11ac and LTE-advanced, are not designed to operate efficiently in a coordinated fashion across multiple densely deployed transmitters.**

In this paper we introduce a new coordination architecture which can achieve high performance gains without the high overhead and deployment cost that usually comes with coordination, thus making the vision of high capacity wireless access via densely deployed transmitters practical. The basic idea is to loosely coordinate nearby transmitters using slow varying channel statistics, while keeping all the functionality which depends on fast varying channel state information and has tight time deadlines locally. We achieve this via a smart combination of analog and digital beamforming using inexpensive front ends, a provably efficient algorithm to select compatible users and analog beams across all transmitters, and backward compatible protocol extensions. Our performance results, which include analysis, simulations, and experiments with software defined radios (SDRs) and directional antennas, show that our approach can achieve the 10x gains of the theoretically optimal coordinated MU-MIMO approach, without the need to either tightly coordinate the clocks of the remote transmitters or meet tight delay constraints.

I. Introduction

The dramatic increase of the demand for wireless data, for wireless connectivity of devices in the internet of things, and for robust low-cost wireless networked systems, necessitates rethinking the future wireless broadband access. The industry's response to this challenge has been to come up with successive generations of wireless standards achieving continuous performance improvements. These standards incorporate advanced physical layer techniques which manage to push more bits through the same time/frequency slots. For example, the most recent such technique, multiuser MIMO (MU-MIMO) [1], [2], realizes large multiplexing gains by allowing multiple transmitting antennas to transmit signals to multiple receivers concurrently, and has been incorporated in the newest WLAN standard, 802.11ac, and in LTE-Advanced.

This fast technology transfer of the most promising ideas from academia into standards has resulted in sizable gains. However, these rate gains are not nearly enough to address the tremendous increase for wireless bandwidth demand. As a result, it is the general understanding of both academia and industry that a significant increase in wireless traffic demand can be met only by deploying more access points (APs) per square km. For example, reducing the APs range by a factor of 10, while maintaining full coverage, yields a cell density increase of 100x. Provided that the throughput per cell remains unchanged, this yields a 100-fold spectral efficiency increase by straightforward spectrum spatial reuse.

Unfortunately, things are not so simple. Setting aside deployment challenges, e.g. it would be hard to find available real estate for APs in densely populated areas where the problem is more acute, and it would be costly to interconnect all those stations with a wired back-haul,[1] there are fundamental system design challenges. Most notably, increasing the density of APs would tremendously increase the inter-cell interference and eventually cancel any performance gains. Indeed, it is well known that inter-cell interference is much more significant at short distances than at large distances, exemplified by the much smaller pathloss exponent in the former case [3]. Thus, packing cells more and more densely and reusing the same frequency bands in space results in too much inter-cell interference and the predicted gain vanishes. This is precisely the situation in today's dense deployments of enterprise WiFi networks, explaining the mediocre wireless access speeds experienced in city centers, airports, campuses, conference halls, stadiums, and other instances with high density of end users.

Both academia and industry have realized that the answer to inter-cell interference is coordination between nearby wireless stations. There have been a number of research works on coordination, most notably the recent testbeds implementing what is commonly referred to as distributed or coordinated MU-MIMO, where nearby APs form a virtual single transmitter and jointly MU-MIMO towards a large number of users thus eliminating inter-cell interference and achieving, in theory, the best possible performance [4]–[7]. The cellular industry has also been investigating various coordination approaches under the umbrella term of coordinated multipoint (CoMP). However, it has become apparent that coordination comes with high overhead and deployment cost, often requiring tight time and frequency synchronization among remote transmitters which is impractical. Thus, the real challenge is how to do "CoMP" in a way that achieves the high gains without the need for unrealistic synchronization requirements and unreasonably high overhead in todays commodity WiFi networks.

All authors are with the University of Southern California.

This work was supported by NSF grant ECCS-1444060, and by grants from Futurewei Technologies and CISCO Systems.

[1]There are some signs that the deployment challenge is manageable. For example, small cell technology, a major theme of 5G, is finding its way to the homes via carriers, to the streets at lampposts, and to city centers. And, WiFi coverage is becoming available even at super dense places like stadiums.

Motivated by this, in this paper we introduce a new wireless coordination architecture which can achieve high performance gains without the high overhead and deployment cost that often comes with coordination. The basic idea is to use a smart combination of analog and digital beamforming to separate coordination functionality and local instantaneous operations in two different timescales. Coordination lives in a slow varying world and only depends on slow varying channel statistics, collected and updated continuously over time. It allows concurrent transmissions from nearby APs while suppressing inter-cell interference by appropriately configuring analog front-ends. Local operations like digital precoding implementing MU-MIMO and other advanced PHY techniques live in a fast changing world as they depend on fast varying channel state information (CSI). The two worlds are connected by jointly selecting analog beamforming configuration parameters and the users to be served by each participating AP. Interestingly, despite the fact that there is no need for tight time or frequency synchronization, our architecture, which we refer to as ACME (for its Asynchronous, Coordinated Multi-timE scale nature), can achieve the same performance in practical dense deployments as that of a distributed, fully coordinated MU-MIMO system.

The outline of the paper is as follows. In the next section we briefly discuss related work. In Section III we present the system architecture in detail, paying special attention to the MAC operations required to support a cluster of coordinated transmitters. In Section IV we provide 3 algorithms to solve the joint scheduling problem of analog front-ends and end-users within a cluster of transmitters. In Section V we develop an analytical framework to study the performance of ACME against a fully coordinated MU-MIMO system as well against today's non-coordinated architectures. Next, in Section VI we conduct extensive large scale simulations and establish that ACME can achieve the performance level of a fully coordinated MU-MIMO system under dense scenarios of relevance. We also conduct WiFi simulations and show that our system can be implemented in a backward compatible manner. Last, in Section VII we use a small scale testbed consisting of SDRs and directional antennas, and establish the basic performance features of our system under real world conditions. Section VIII concludes the paper.

II. Related Work

Recently, there has been a wide variety of papers on hybrid digital/analog beamforming [8]–[10]. In these works, the authors' motivation stems from the need to take advantage of a large number of antennas without using an equally large number of RF chains, in the context of massive MIMO. With this in mind, these works focus on the combination of a fully reconfigurable analog front-end with MU-MIMO capabilities in a macrocell BS that exists in isolation. In contrast, our motivation is to suppress inter-cell interference in a dense multi-cell deployment of coordinated APs and we adopt a more restricting but significantly cheaper and thus practical switched-beam antenna front-end allowing for interference mitigation via smart inter-cell coordination. In [11], a multi-cell JSDM system is theoretically analyzed in the context of a cellular network where base stations (BSs) are equipped with fully reconfigurable analog front-ends and the goal is to reduce inter-cell interference at the edge of the cells. Instead of considering large reconfigurable antenna arrays whose cost is prohibitively expensive for WiFi APs, and, as a matter of fact, even for cellular mini BSs, we consider switched-beam antennas whose small number of useful configurations yield a fundamentally different system and a very different scheduling problem when jointly selecting analog configurations and users.

In addition to theoretical studies, there have been a few simulation/experimental works on hybrid digital/analog beamforming. The authors in [12] apply the theory of JSDM presented in [8] on switched-beam antenna front-ends and show via simulations multiplexing gains in a toy indoor deployment of two APs which do not coordinate with each other, further motivating our work. In [13], an implementation of a hybrid digital/analog multi-cell beamforming system is presented where the main contribution of the paper is a neat algorithm to infer the Angles of Arrival needed for analog beamforming. However, the system assumes a global clock for scheduling purposes which is incompatible with WiFi, allows only a single user per beam which is suboptimal, and its performance is not studied in large scale scenarios of relevance.

In the cellular industry CoMP (Coordinated Multi-Point) systems with looser coordination have been proposed, namely Coordinated Scheduling (CS) and Coordinated Beamforming (CB) (see [14] for an overview of CoMP transmission/reception techniques). CS refers to the scenario where transmissions for single users are scheduled without MU-MIMO capabilities [14]. In the case of CB, MU-MIMO can also be used to serve the users. There are multiple approaches suggested for CB in cellular networks. In [15] a coordination scheme is described with no tight synchronization required between BSs, however exchange of CSIT is required for the computation of the optimal beamformers. In [16] only local CSIT is required for computing the transmission pattern but this solution is tailored for fully-reconfigurable front-ends, which can't be used in the context of commodity WiFi networks as mentioned earlier.

The gains of using directional antennas in a cellular setting and outdoor environments are well established. The literature has also evaluated the applicability and exploit the use of directional antennas in indoor environments [17], [18]. Building on that, recent work has proposed MAC layers that can efficiently operate directional PHY layer transmissions [19]–[21]. The proposed MAC layers either allow for some cooperation between APs for efficient selection of users and beams to mitigate interference [19], [20] or extend the 802.11 CSMA RTS/CTS handshake to take into account directional transmissions [21]. All these works focus on analog beamforming and neglect the advances in the digital precoding domain, e.g. MU-MIMO, thus, consequently, also neglecting the sizable gains to be made by jointly optimizing directional transmissions with digital precoding.

Advances in the PHY layer have made it possible to operate a multi-cell network without inquiring any interference between BSs/APs. The technology, coined distributed / coor-

dinated MU-MIMO, assumes a tight synchronization between stations such that they operate as a single, massive station [4]. Such deployments have been recently demonstrated to be very effective in achieving the full multiplexing gains of a network [5]–[7] however they come with impractical challenges in terms of the level of synchronization they require [22]. Our work is further motivated from these impractical challenges.

Last, multi-time scale schemes are clearly not a new concept in system design and have been proposed before in wireless networking outside the context of hybrid digital/analog beamforming. For example, in [23] the authors use long-term channel statistics and knowledge of global topology for routing purposes while using short-term channel statistics for local beamforming.

III. System Architecture

The proposed system architecture consists of a network of loosely coordinated APs, to be referred to as APs or stations interchangeably. While our architecture is applicable to both cellular and WiFi networks, the motivation to avoid tight time and frequency synchronization and to use inexpensive and thus constrained analog front-ends is very strong in the case of WiFi. Thus, any protocol-specific discussion will be limited to WiFi.

Every station in our system is equipped with multiple RF chains attached to smart antenna front-ends, which provide the ability to specify to some extent the direction of power radiation. Considering the cost and capabilities trade-off, we envision such an architecture to employ the use of switched-beam smart antennas, which can select from a predetermined set of relatively narrow and overlapping beams that cover the whole surrounding area (more details on this selection can be found in Section III-A3). Thus, every station has the ability to perform directional and MIMO transmissions or a combination of the two, depending on the specific choice of the beams for each front-end and the users to be served.

For scalability, stations are grouped into coordination clusters, such that, by means of frequency reuse, inter-cluster interference is low. Stations that belong to the same coordination cluster are scheduled to transmit at the same time during a coordinated downlink slot. A cluster head is responsible to orchestrate these coordinated downlink transmissions, to be referred to as cluster transmissions henceforth. The cluster head is also the representative of the cluster when it comes to accessing the channel via random access, competing against uplink traffic and, in case of non-negligible inter-cluster interferences, other downlink cluster transmissions. For each cluster transmission, the cluster head will jointly select the users that each station will serve and configure the RF front-ends of all participating stations, based on long term statistics collected slowly over time. At the same time, each individual station will locally perform digital beamforming operations within the space available to it from the front-end configuration, using instantaneous channel state information that will be collected per the existing standards, e.g. 802.11ac.

In Figure 1 we give a schematic example of the proposed architecture.

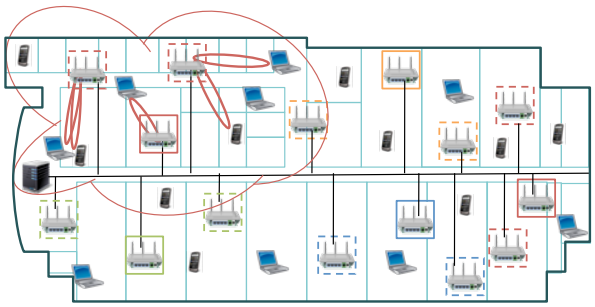


Fig. 1: ACME Architecture Example Deployment. Five clusters with 3 stations each are shown, cluster heads are outlined with a solid line and all stations are connected to a server.

A. Main architectural decisions

We will proceed with the presentation of the key architectural decisions that were made for the proposed system.

1) Multi-time scale feature: The ACME architecture operates in two different time-scales in parallel. Coordination and interference suppression between stations within a cluster rely on long-term statistics of channel characteristics such as the average amplitude of the received power and the average Direction of Arrival (DOA) of incoming transmissions. These channel properties change slowly with time in the nomadic moving WiFi environments that we examine [3]. Interference suppression based on these metrics is implemented in the analog domain by appropriately forming beams at the antenna front-ends. At the same time, scheduled users are served from their assigned stations utilizing advanced PHY layer schemes such as conjugate beamforming and MU-MIMO. These digital precoding schemes require instantaneous CSI at the transmitter which is collected through the regular procedures defined in the WiFi standards (that is, either through explicit or implicit feedback when channel reciprocity allows). It is important to note that the two decisions (beamforming in the analog and the digital domain) are not orthogonal to each other rather they are made jointly taking into account both the slow varying channel statics and the multiplexing gains from digital precoding schemes as will be seen in Section IV.

2) Asynchronous coordination feature: As stated already, ACME's inter-station coordination depends only on slow varying channel characteristics and no tight synchronization of stations is required. It is important to note that contrary to what happens in a fully coordinated MU-MIMO system where a user receives useful signal from multiple stations, in ACME each station serves solely its own users. Therefore, ACME can operate in a very loosely synchronized fashion while still unlocking a big portion of the interference suppression and multiplexing gains of tightly synchronized fully-coordinated MU-MIMO technologies.

One might argue that since there is no joint serving of the users, no time coordination is required. However, clearly the largest multiplexing gains are achieved when the stations of a cluster transmit "concurrently". Otherwise, the duration of a coordinated downlink transmission will last longer than necessary reducing the system gain. Such a loose synchronization requirement can be easily achieved with simple commodity

radios programmed to wait for a "send now" signal from a cluster head, see, for example, [24] for a recent implementation of ours in WiFi-enabled WARP SDRs. As long as clustered stations start their transmissions within say a few microseconds a coordinated downlink transmission will last for less than 1% longer than a regular transmission, a negligible overhead. In contrast, while a fully coordinated MU-MIMO system still enjoys the luxury of somewhat asynchronous transmission times thanks to the cyclic prefix, it does require picoseconds accuracy to keep carrier frequency offset at small levels [5], [22].

3) Intra-cluster coordination using smart antennas: In order to enable intra-cluster interference suppression in the analog domain we employ the use of smart antenna front-ends. Using analog beamforming and exploiting the slow varying DOA channel statistics, we manage to minimize the interference caused from stations to users currently being served by other stations in the same cluster. Each station is assigned users and beams appropriately such that no transmission will cause significant interference to users served by neighboring stations (see Section III-B for a more detailed presentation and the required MAC for such a coordination).

In general, smart antennas are antennas with multiple elements, where signals can be combined by an adaptive algorithm in an "intelligent" way to exploit the directional properties of the channel [3]. Smart antennas can be of two types: i) adaptive arrays and ii) switched-beam antennas. The first category provides full flexibility in the linear combinations of the signals of the different antenna elements and thus the biggest capacity gains, however it comes with a significant cost since it requires baseband processing and thus down-conversion chains. Thus, for a commodity network the practical and viable choice is that of switched-beam antennas that provide a more restricted choice of selectable transmission/reception directions, but can be realized with inexpensive phase-shifters in the Radio Frequency (RF) domain.[2]

4) Local MIMO-based precoding: The ACME system does not rely solely on the inter-station interference suppression described in the previous subsection for capacity increase. It combines it with local digital precoding schemes to extract the multiplexing and/or power gains of the MU-MIMO channel. Thus, after the joint selection of station beams and users to be served based on the available statistics, every station serves potentially multiple users, in a Point-to-Point/MU MIMO fashion. The decision of the local digital precoding scheme is taken in parallel with the beam scheduling through the algorithm described in Section IV.

For completeness, let us briefly state that Point-to-Point MIMO [25] came as an advancement from smart antennas and extracts higher capacity from a given bandwidth by adding say M antennas to the transmitter and N_r to the receiver. One of the advantages of having multiple transmit/receive antennas is the ability to increase the effective SNR, known as a *power gain*, through techniques such a conjugate beamforming and maximal ratio combining [3]. In addition, a variety of techniques known as *spatial precoding* [3] enable multiplexing in the spatial domain, such that the capacity at high SNR is increased by a multiplicative factor up to $\min(M, N_r)$, known as a *multiplexing gain*. For the downlink where N_r may be small, a set of techniques known as MU-MIMO [1] can be used to attain higher multiplexing gains up to $\min\{M, KN_r\}$, for a system where a station with M antennas serves K users with N_r antennas each.

5) Collection and integration of slow varying statistics: The ACME architecture collects, stores and monitors the average received power for every user/beam pair. In order to track this statistic without having to sample every single beam in each station we may rely on the DOA of the signal components to infer the corresponding beam that is active. Since DOA and average receive power are channel characteristics that vary slowly over time and frequency, channel reciprocity can be exploited to sample the uplink (UL) transmissions and populate a table similar to the one in Figure 2c for the downlink (DL) channel.

In order to infer the DOA from UL transmissions, we envision that stations will use a standard high resolution algorithm for this purpose [3] or state of the art approaches for specific types of antennas [13]. Since the tracked channel characteristics vary in a very slow pace, the duty cycle of this process is very long compared to the millisecond timescale transmission cycle, and thus adds a negligible overall overhead in the system since it can happen at times when a station is silent and overhears neighboring user transmissions. To identify the users from whom the UL transmissions originate, various approaches can be used such as decodable messages that are sniffed during transmissions from nearby users, or ideas like the LTE midamble [26].

B. MAC Layer

Central in our architecture is an efficient MAC layer that will enable the loose cooperation we envision and allow co-existence with legacy devices. There are specific MAC requirements that would need to be fulfilled for such system:

First, as already discussed, the stations which are to be coordinated need to collect long term channel statistics, in our case signal strength values and signal DOA values between a beam of a station and a user, and communicate them to a central server. This server can be a dedicated server serving multiple clusters, or the cluster head. Such a process is not particularly time sensitive since indoor channels change in the order of 100s of ms.

Second, based on these collected channel statistics, the dedicated server decides on a schedule of compatible user / beam tuples, such that the coordinated stations can transmit concurrently while causing minimal interference to each others' served users. This schedule is communicated to the stations asynchronously and not necessarily before every downlink slot, since channel coherence time allows such relative flexibility.

Third, a backward compatible signaling scheme is required such that a coordination cluster can grab the channel and restrict other devices from interfering during the ensuing coordinated downlink slot. Cluster head selection, thus, has to account for the requirement of being able to reach every device in the region of the coordination cluster with high probability.

[2]In particular, antenna manufacturers offer switched beam antennas for a few dollars, whereas adaptive arrays cost thousands of dollars.

Last, a signaling scheme is required that will allow a loosely synchronized cluster transmission, along the lines of the "send now" message mentioned earlier. We proceed by specifying the above concepts for the most prominent WiFi standard, 802.11.

In the WiFi regime both UL and DL traffic use the same channel frequencies. This allows for exploitation of UL/DL reciprocity for channel state information and long term channel statistic collection, but also increases the contention for the wireless medium between AP clusters and user terminals. Given the directional nature of DL transmissions, it is more likely that a user may falsely presume the channel to be free and initiate an omnidirectional UL transmission. In order to protect coordinated DL transmissions from such UL interference, we use an unsolicited Clear To Send (CTS) packet as a precursor to the DL slot. Note that such unsolicited CTS packets are part of the existing WiFi standard (CTS-to-self packages). When a CTS packet is overheard from a WiFi device and it is not destined for that device, the device understands that a transmission will take place and holds back from transmitting for an interval predetermined in the CTS packet. We want this CTS message to be overheard by all devices within the range of the coordination cluster preparing for transmission, thus we appropriately size clusters and choose the head of the cluster station that transmits this signal.

As an example, Figure 1 shows five AP clusters (denoted by different colors) with 3 APs each. Frequency re-use is done using 4 channels and thus 2 of the clusters share the same channel. The coverage area of one of the cluster heads is depicted, showing that all users and APs in its cluster are within its range (cluster heads are denoted by solid lines whereas other APs with dashed lines). Potential interference between clusters sharing the same frequency is minimized by appropriately assigning cluster channels the same way as enterprise WiFi networks assign channels to APs today. Even when inter-cluster interference does exist, users will be impacted only when they fall into the direction of a beam from a neighboring cluster using the same channel, and the interference power will be minimal thanks to the larger inter-cluster distances (as compared to inter-AP distances).

We end this section with a timeline of the operations that take place to complete a coordinated DL transmission. We first describe the required steps in the absence of any pre-communicated information about user-beam configurations and then describe an optimized operation.

First, a cluster head grabs the channel per the WiFi standard. Second, using the available slow varying channel statistics, a joint selection of beams and users takes place (as discussed in Section IV) or a pre-computed selection is used. Third, this configuration is broadcasted to all participating stations, or a simple index indicating a pre-loaded selection is communicated, optionally with a field that allows it to be reused for a number of consecutive transmissions. Fourth, if needed, stations independently acquire instantaneous CSI and compute precoding matrices as would happen in existing systems. Fifth, an unsolicited CTS message, that also acts as the "send now" signal, is broadcasted by the cluster head and triggers the coordinated DL transmission.

Note that the slow varying nature of the used channel statistics and the user traffic demands allow the system to pre-compute user-beam configurations for a number of future coordinated DL transmissions and pre-load them to the participating APs using over-the-wire communication. This allows to optimize the operations as follows: After the cluster head grabs the channel per the WiFi standard, a control signal is broadcasted over the air with a triple purpose: to indicate the pre-loaded configuration to be used, to act as an unsolicited CTS message and to also act as the "send now" signal. At this point, if needed, participating stations independently acquire instantaneous CSI, compute precoding matrices and transmit as would happen in existing systems. Note that while the CSI acquisition and the precoding matrix computation may take slightly different times among different participating stations, this is not a problem as transmissions do not have to be coherent.

In [27] we have implemented and tested the transmission of a "send now" signal triggering coordinated DL transmissions from the cluster head to other APs using WARP boards, see [27] for more details. Also, in Section VI-B we use the WiFi support of the NS-2 network simulator [28] to implement the WiFi specific aspects of the above operations, such as the broadcasting of unsolicited CTS messages, and verify the correct operation of the protocol, see Section VI-B for more details.

IV. The Joint User-Beam Scheduling Problem

Consider a cluster of N stations. Without loss of generality, consider K single-antenna users associated with these stations forming the set $\mathcal{S}$. Each station is equipped with M RF chains that allow it to transmit up to M independent streams simultaneously using MIMO techniques. Finally, every RF front-end is attached to a switched-beam antenna with B beams. We focus on the downlink scheduling problem which consists of jointly selecting the users to be served by each AP of the cluster under study, and, the beams to be used by each RF chain of each AP of the cluster, such that a function of the users' rates is maximized.

First, we formulate the scheduling decision as a network utility maximization (NUM) problem targeting proportionally fair rate allocation and propose a novel greedy algorithm to solve the computationally hard problem that arises. We then turn to graph theory and map our problem to a weighted maximum independent set (WMIS) problem with a cardinality constraint. The last formulation lends itself to an integer linear program (ILP) form that can be solved with off-the-shelve LP solvers, and, a graph-theory based approximation algorithm can be used to provide formal approximation bounds. Finally, we extend the results of [29] for the case of directional beams and map our problem to a maximization of the number of SINR-feasible links. We obtain formal approximation bounds for this formulation as well. The performance of the three schemes under real-world scenarios is investigated in Section VI.

A. NUM-based greedy user-beam selection

Following a similar approach like the one used for the traditional user selection problem in the context of MU-MIMO

Symbol		Definition
N	$\triangleq$	number of APs in cluster
M	$\triangleq$	number of antennas per AP
B	$\triangleq$	number of beams per switched-beam antenna
K	$\triangleq$	number of users in cluster
R_k	$\triangleq$	rate of user k
$\mathcal{S}$	$\triangleq$	set of compatible user-beam tuples
$\mathcal{I}$	$\triangleq$	family of independent sets
$\mathsf{SINR}_k(I_i)$	$\triangleq$	Signal to Interference plus Noise Ratio at user k from independent set $I_i \subseteq \mathcal{I}$
$C_k(I_i)$	$\triangleq$	instantaneous rate of user k for a given independent set $I_i \subseteq \mathcal{I}$
$Q_k(t)$	$\triangleq$	virtual queue weight for user k at time t
P_i	$\triangleq$	transmit power of AP i
R_{ik}	$\triangleq$	rate of user k from AP i
g_{ik}	$\triangleq$	pathloss from AP i to user k
S_i	$\triangleq$	number of downlink streams from AP/cluster i
K_i	$\triangleq$	number of users associated with AP i
$\mathcal{N}_i$	$\triangleq$	number of APs in cluster i

TABLE I: Notation Summary.

[30], we start with the objective to maximize some component-wise concave utility function $g()$ of the users' average rate vector $\overline{\mathbf{R}} = [\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_K]$, that is:

$$\begin{aligned} \max \quad & g(\overline{\mathbf{R}}) \\ \text{s.t.} \quad & \overline{\mathbf{R}} \in \mathcal{R}, \end{aligned} \tag{1}$$

where $\mathcal{R}$ is the achievable rate region.

We define the partition matroid $\mathcal{M} = (\mathcal{S}, \mathcal{I})$ [31] with ground set the set of user-beam tuples $\mathcal{S}$ that are compatible (that is, the user is within the area covered by the beam and is assigned to the station the beam belongs to) and $\mathcal{I}$ the family of subsets of $\mathcal{S}$, called independent sets, such that $\mathcal{I} = \{I : I \subseteq \mathcal{S},\ |I \cap \mathcal{S}_i| \leq M\ \forall\ i = 1, 2, \ldots, N\}$, where $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N$ is a partition of the user-beam tuple set $\mathcal{S}$ in N non-overlapping subsets associated with stations $1, 2, \ldots, N$ respectively, by some user-station association algorithm. The instantaneous rate for user k for a given active independent set of user-beam tuples $I_i \subseteq \mathcal{I}$, is given by

$$C_k(I_i) = \begin{cases} 0 & \text{for} \quad k \notin u(I_i) \\ \log\left(1 + \mathsf{SINR}_k(I_i)\right) & \text{for} \quad k \in u(I_i), \end{cases} \tag{2}$$

where $u(I_i)$ extracts all users from the tuples in I_i and $\mathsf{SINR}_k(I_i)$ is the Signal to Noise and Interference Ratio under some local beamforming scheme, e.g. MU-MIMO.

Following the stochastic network optimization theory [32] and assuming a proportional fairness objective we can formulate Problem (1) as a time evolving weighted sum-rate maximization problem, where the weights $Q_k(t), \quad k = 1, \ldots, K$ are derived from an appropriately updated virtual queue for every user k at time t. Letting $\mathcal{K}$ denote the set of all users, $K = |\mathcal{K}|$, Problem (1) becomes:

$$\begin{aligned} \max \quad & \sum_{k=1}^{K} Q_k(t) R_k(t) \\ \text{s.t.} \quad & R_k(t) \in \mathcal{R} \quad \forall k \in \mathcal{K}, \end{aligned} \tag{3}$$

where the virtual queues evolve according to: $Q_k(t+1) = \max\{0, Q_k(t) - R_k(t)\} + a_k(t)\}$ where $a_k(t)$ maximizes $Vg(\mathbf{a}(t)) - \sum_{k=1}^{K} a_k(t) Q_k(t)$, with $\mathbf{a}(t) : 0 \leq a_k(t) \leq A_{max}$, $\mathbf{a}(t) = (a_1(t), \ldots, a_K(t))$, and A_{max}, V are appropriately chosen constants.

Taking into account the partition matroid and the instantaneous rates from Equation (2) and focusing on a particular timeslot t we can rewrite the problem as:

$$\max_{I_i \in \mathcal{I}} \quad \sum_{k \in I_i} Q_k(t) C_k(I_i) \quad \forall k \in \mathcal{K}. \tag{4}$$

Problem (4) is a maximization of a set function that is neither sub-modular[3] nor super-modular under a partition matroid constraint. The independent sets we would have to compute in order to optimally solve the above problem are exponential in the number of stations in the network and thus such an approach is impractical. Instead, we apply a greedy approach (Algorithm 1) where we schedule users based on the weighted sum-rate they produce, adding the user that gives the higher marginal gains every time till we reach the maximum number of users per station or the weighted sum-rate decreases by adding additional users.

Algorithm 1 Greedy Algorithm for Problem (4)

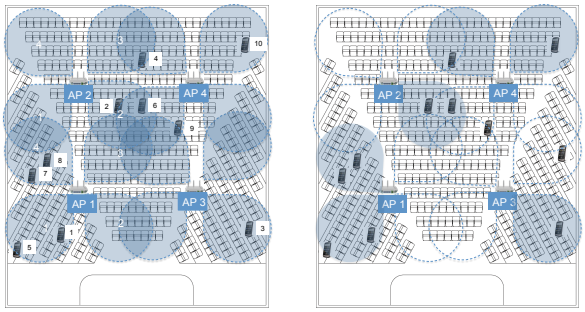
Initialization:
$\mathcal{J} = \emptyset$; $\mathbf{C}(\mathcal{J}) = 0$
while $\mathcal{J} \in \mathcal{I}$ **do**
  $k^* = \arg\max_{k \notin \mathcal{J}, (\mathcal{J} \cup \{k\}) \in \mathcal{I}} \sum_{k=1}^{K} Q_k(t) C_k(\mathcal{J} \cup \{k\})$
  if $\sum_{k=1}^{K} Q_k(t) C_k(\mathcal{J} \cup \{k^*\}) \leq \sum_{k=1}^{K} Q_k(t) C_k(\mathcal{J})$ or $k^* = \emptyset$ **then** break;
  else
    $\mathcal{J} \leftarrow \mathcal{J} \cup \{k^*\}$
  end if
end while

Figure 2 shows a simplified example of how the greedy scheme would operate. The example topology is a conference hall with 4 APs with 4 beams each and 10 users. The system collects the average received power statistics for every user-beam pair and greedily selects a viable set of user-beam tuples starting from the user that has the highest receive power from AP 1 (say user 1 via beam 1), excluding all beams that cause interference to the selected user (e.g. beam 1 of AP 2) and all users that see interference from the scheduled tuple (e.g. user 5, unless MU-MIMO is used to transmit to both users 1 and 5 from beam 1 of AP 1) and continuing in this fashion till no more users can be added. Notice that, although beam 2 of AP 2 is causing some interference to user 3, due to the large distance this interference is probably low and the scheduler may select user 3 as well to be served by beam 3 of AP 3.

B. Greedy max-weight independent set approach

Note that the above greedy algorithm does not have any theoretical performance guarantee. To design a greedy algorithm that can provide a performance guarantee we first adopt a simplified interference model using a conflict graph G. Specifically, the set of vertices of the conflict graph G is the set of user-beam tuples $\mathcal{S}$. An edge exists between a pair of

[3]A set function $f : 2^{\mathcal{S}} \rightarrow \mathbb{R}$ is sub-modular when for every $X, Y \subseteq \mathcal{S}$ with $X \subseteq Y$ and every $x \in \mathcal{S} \setminus Y$ we have $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$. When $-f$ is sub-modular f is called super-modular.



(a) Available transmission beams.

(b) Greedy beam selection.

User \ Beam	AP 1				AP 2				AP 3				AP 4			
	1	2	✖	4	✖	2	✖	4	✖	2	✖	✖	✖	2	3	4
1	19	-∞	-∞	-∞	12	-∞	-∞	-∞	13	-∞	-∞	-∞	10	-∞	-∞	-∞
2	-∞	-∞	13	-∞	-∞	19	-∞	-∞	-∞	-∞	-∞	12	18	-∞	-∞	-∞
3	-∞	10	-∞	-∞	-∞	9	-∞	-∞	-∞	18	-∞	-∞	-∞	10	-∞	-∞
4	-∞	-∞	12	-∞	-∞	-∞	18	-∞	-∞	-∞	-∞	13	-∞	-∞	-∞	18
5	17	-∞	-∞	-∞	12	-∞	-∞	-∞	12	-∞	-∞	-∞	10	-∞	-∞	-∞
✖	-∞	-∞	14	-∞	-∞	17	-∞	-∞	-∞	-∞	-∞	14	18	-∞	-∞	-∞
7	-∞	-∞	-∞	18	15	-∞	-∞	-∞	-∞	-∞	-∞	11	11	-∞	-∞	-∞
8	-∞	-∞	-∞	17	16	-∞	-∞	-∞	-∞	-∞	-∞	12	11	-∞	-∞	-∞
✖	-∞	-∞	13	-∞	-∞	14	-∞	-∞	-∞	-∞	-∞	16	18	-∞	-∞	-∞
10	-∞	-∞	10	-∞	-∞	-∞	12	-∞	-∞	-∞	13	-∞	-∞	-∞	17	-∞

(c) Avg. Received Power [dBm] for every beam/user pair.

Fig. 2: Greedy user/beam selection in an example 40x40m conference hall.

vertices (user-beam tuples) $v \in \mathcal{S}$ and $w \in \mathcal{S}$ if either the user of v is hit by the beam of w or the user of w is hit by the beam of v, unless the beam in both pairs is the same in which case no edge is introduced to allow for MU-MIMO transmissions within a beam. Note that in practice, whether a user is "hit" by a beam or not is decided by a received power threshold such that beams originating from very distant transmitters do not introduce edges on the graph (see, for example, Figure 2, where perhaps user 3 is far enough from AP 2 such that it is not "hit" by beam 2 of this AP).

In addition, we associate each vertex (user-beam tuple) $v \in \mathcal{S}$ with a weight $W(v)$ that depends on the transmission rate in isolation and the virtual queue size (introduced in the previous subsection) of the user v. In this way we pursue again a notion of proportional fairness for our problem. Then, the user-beam scheduling problem reduces to selecting the maximum weight independent set in the conflict graph G, respecting a cardinality constraint due to the fact that each AP can serve at most M users concurrently. Intuitively, what we are trying to achieve here is to pack as many beams as possible in space, while simultaneously regulating how much interference, i.e. overlap of the beams, is allowed.

The maximum weight independent set problem with cardinality constraints can be formulated as an integer programming (IP) problem:

$$\begin{aligned} \max_{x_u} \quad & \sum_{u \in V(G)} W(u) x_u \\ & x_u + x_v \leq 1, \ \forall (u,v) \in E(G) \\ & \sum_{u \in \mathcal{S}_i} x_u \leq M, \ \forall i = 1, \ldots, N \\ & x_u \in \{0,1\}, \ u \in V(G), \end{aligned} \tag{5}$$

where the binary decision variable $x_u = 1$ if the vertex (user-beam tuple) u is scheduled, and $x_u = 0$ otherwise. The first set of constraints corresponds to the conflict graph constraints, ensuring that no conflicting user-beam tuples are simultaneously scheduled. The second set of constraints corresponds to the cardinality constraints, where S_i denotes, as before, the set of user-beam tuples whose beams are generated by station i, ensuring that the number of user-beam tuples simultaneously scheduled under station i is less than or equal to the number of RF chains (the degrees of freedom) M that station i has.

We note that problem (5) has a linear objective function. In addition, both sets of constraints are linear. When the binary decision variables are relaxed, i.e. $x_u \in [0,1]$, problem (5) becomes an LP problem. Thanks to this, we can use CVX with Gurobi, which is a commercial solver for large scale mixed integer linear programming, to solve problem (5). The solver uses a mix of the branch and bound and interior point methods, and, when finding a solution, it also outputs the distance to the optimal solution of the relaxed problem. Section VI presents the output of the solver for specific scenarios of interest.

Another way to solve the maximum weight independent set problem is via a graph-based greedy algorithm. Let $V(G)$ denote the vertex set of the graph G, $d_G(v)$ denote the degree of a vertex v in graph G, $N_G(v)$ denote the neighborhood of v, $N_G^+(v) = \{v\} \cup N_G(v)$, $G[V']$ denote the subgraph of G induced by a vertex set V', and Δ_G denote the maximum degree of G. We propose the following graph-based greedy algorithm to solve the problem.

Algorithm 2 Graph-based Greedy Algorithm

Initialization:
$\mathcal{J} \leftarrow \emptyset$; $i \leftarrow 0$; $G_i \leftarrow G$;
while $V(G_i) \neq \emptyset$ **do**
Choose a vertex v_i in G_i such that $v_i = \arg\max_{u \in V(G_i), (\mathcal{J} \cup \{u\}) \in \mathcal{I}} \frac{W(u)}{d_{G_i}(u)+1}$;
$\mathcal{J} \leftarrow \mathcal{J} \cup \{v_i\}$; $G_{i+1} \leftarrow G_i[V(G_i) \setminus N_{G_i}^+(v_i)]$;
$i \leftarrow i + 1$;
end while
Return $\mathcal{J}$.

To prove a formal performance bound, ignore the cardinality requirement $(J \cup \{u\}) \in \mathcal{I}$ in step 2 and call this new algorithm Algorithm 2'. Then, the following bound holds:

Theorem 1. *Algorithm 2' is a* $\frac{1}{\Delta_G}$ *approximation algorithm.*

Proof. The crux of the proof is the creation of the aforementioned graph which allows us to use established results from graph theory. Specifically, we start by creating the conflict graph G as described above to encode the conflicts of user-beam tuples. As a result, the user-beam scheduling problem is transformed into finding the maximum weight

independent set of graph G. Then, using $\frac{W(u)}{d_G(u)+1}$ as the metric based on which greedy decisions take place, it can be shown that Algorithm 2' is a $\frac{1}{\Delta_G}$ approximation, see [33] for details. □

C. Greedy maximum cardinality SINR-feasible approach

Another approach to address the user-beam selection problem is to schedule a maximum cardinality set of user-beam pairs that are feasible. The transmission of a user-beam pair is feasible if the received SINR at the user is larger than some threshold β, denoting the minimum SINR required for a message to be successfully received. Let us consider a set of user-beam pairs $\mathcal{S} = \{l_1, \ldots, l_n\}$, where each user-beam pair l_v represents a beam from station (sender) s_v to user (receiver) r_v. The distance from the user-beam pair l_v to another user-beam pair l_w is defined as the Euclidean distance between l_v's sender s_v and l_w's receiver r_w, denoted by $d_{vw} = d(s_v, r_w)$. We assume that all beams transmit at the same power level P. In addition, we denote the main lobe gain by G_{ML} and the gain of side lobes by G_{SL}. We further assume a path-loss based radio propagation channel model with path-loss exponent α. As a result, the received power from transmitter s_v at receiver r_w , denoted by P_{vw}, equals PG_{vw}/d_{vw}^{α}, where G_{vw} equals G_{ML} when r_w is hit by the main lobe of s_v and it equals G_{SL} otherwise.

To facilitate the analysis, we define the following two concepts: relative interference and affectance, see [29] where the authors introduce these concepts while solving the maximum cardinality SINR-feasible problem under omnidirectional antennas. The relative interference of the user-beam pair l_w on another user-beam pair l_v is defined as $RI_w(v) \triangleq \frac{P_{wv}}{P_{vv}}$. Let $RI_v(v) \triangleq 0$ and $c_v \triangleq \frac{1}{1-\frac{\beta N}{P_{vv}}}$. The affectance of the user-beam pair l_v, caused by a set S of user-beam pairs, is defined as the sum of the relative interferences of the user-beam pairs in S on l_v, scaled by c_v, i.e.,

$$a_S(l_v) \triangleq c_v \sum_{l_w \in S} RI_w(v). \tag{6}$$

The transmission of the user-beam pair l_v is feasible if

$$\frac{P_{vv}}{\sum_{l_w \in S \setminus \{l_v\}} P_{wv} + N} \geq \beta, \tag{7}$$

where N is the background noise, β is the SINR threshold, and S is the set of concurrently scheduled user-beam pairs in the same slot as l_v. We say a set of user-beam pairs S is SINR-feasible if (7) is satisfied for each user-beam pair in S. We seek the maximum cardinality set of user-beam pairs that is SINR-feasible, i.e.,

$$\max_{S \subset \mathcal{S}} |S| \quad \text{s.t.} \ \frac{P_{vv}}{\sum_{l_w \in S \setminus \{l_v\}} P_{wv} + N} \geq \beta, \ \forall \ l_v \in S, \tag{8}$$

where $|S|$ is the cardinality of S.

Similar to [29], we construct a greedy algorithm (Algorithm 3) to solve problem (8) which uses the affectance value to make selection decisions.

Algorithm 3 SINR-based greedy algorithm

Require: A set of user-beam pairs $\mathcal{S} = \{l_1, \cdots, l_n\}$. A threshold T which depends on α and β. ($T = \frac{1}{\tau^{\alpha}}$ where $\tau = 2 + \max(2, (73 \cdot \beta \frac{\alpha-1}{\alpha-2})^{1/\alpha})$, see [29] for details.)
Ensure: A feasible set $S \subset \mathcal{S}$ of user-beam pairs
1: Initialize: sort the user-beam pairs $l_1, \cdots, l_n$ in decreasing order of received signal power P_{vv}; $S \leftarrow \emptyset$;
2: **for** $v = 1$ to n **do**
3: **if** $a_S(l_v) \leq T \cdot \frac{G_{SL}}{G_{ML}}$ **then**
4: add l_v to S;
5: **end if**
6: **end for**
7: Return S.

Theorem 2. *Algorithm 3 produces a schedule with cardinality that is smaller from the maximum cardinality by a factor of at most $\frac{G_{ML}}{G_{SL}}$. Thus, Algorithm 3 is a constant-factor approximation algorithm to Problem (8).*

Proof. When the transmit power is omni-directional and its value varies between $P_{\min}$ and $P_{\max}$ among the various transmitters, [29] proves that Algorithm 3 with affectance threshold $T \cdot \frac{P_{\min}}{P_{\max}}$ produces a feasible schedule with cardinality that is smaller from the maximum cardinality by a multiplicative factor of at most $\frac{P_{\min}}{P_{\max}}$. In the directional case with main lobe gain G_{ML}, side lobe gain G_{SL} and same transmit power from all transmitters, it suffices to use an affectance threshold $T \cdot \frac{G_{SL}}{G_{ML}}$ as in Algorithm 3 to get a constant-factor approximation which is at most $\frac{G_{SL}}{G_{ML}}$ from the optimal. □

V. Equation-based Performance Analysis

We wish to study the performance of our system under realistic large scale scenarios, e.g. typical Enterprise WiFi scenarios with hundreds or even thousands of APs densely deployed in crowded areas. Clearly, our experimental testbed cannot reach anywhere near those scales. What is more, we wish to avoid solely relying on large scale Monte Carlo simulations for the performance evaluation under such large scale scenarios, since they would take too long to run for some of the larger scale scenarios of interest. Motivated by this, we build upon a well-established analytical approach (see [34]–[37] and references therein) which allows us to compute user rates accurately and efficiently, as well as study the system performance as we vary key system parameters.

Assume a system with clusters of stations with set $\mathcal{N}_i$ denoting the stations of cluster i. In an architecture that employs proportional fairness to serve users in the DL, the average rate that user k gets from the station or cluster i it is assigned to is given by

$$R_{ik} = \frac{S_i}{K_i} \log\left(1 + \mathsf{SINR}_{ik}\right), \tag{9}$$

where S_i is the number of downlink streams, K_i the number of associated users, and SINR_{ik} is the SINR at user k, all in relation to station/cluster i.

We wish to compare the performance of ACME versus both state-of-the-art and forthcoming technologies. Specifically we compare it against SU-MISO (most common form of transmission in LTE and 802.11n/ac), local MU-MIMO (available in the second generation of 802.11ac chipsets), and a fully

coordinated MU-MIMO system where stations form a single "virtual" transmitter and MU-MIMO towards a large number of users. Next we briefly discuss how we arrive to useful approximations for these approaches.

Local SU-MISO: In the SU-MISO case only a single user, user i, can be served at a time ($S_i = 1$). In the regime where M is large enough such that the effect of small-scale fading disappears [38], one may show that the achieved SINR is given by:

$$\mathsf{SINR}_{ik} = \frac{g_{ik} M P_i}{1 + \sum_{j:j\neq i} g_{jk} P_j}, \tag{10}$$

where g_{ik} is the channel gain from station i to user k and P_i is the transmit power of station i.

Local MU-MIMO: In MU-MIMO, under the regime where both M and S_i become large while keeping the ratio $S_i/M \leq 1$ fixed, one may use the so called diversity gain result [39] as well as the random matrix theory trace lemma [40] to obtain a deterministic approximation for the SINR. Specifically, it can be shown that the SINR achieved by user k when served from station i converges to [34], [36]:

$$\mathsf{SINR}_{ik} = \frac{(M - S_i + 1) g_{ik} P_i / S_i}{1 + \sum_{j:j\neq i} g_{jk} P_j}. \tag{11}$$

Note that for simplicity this analysis assumes equal power allocation per user/stream, which is typical in real world implementations of the MU-MIMO mode in 802.11ac chipsets.

Coordinated MU-MIMO: Assuming the same regime as in MU-MIMO and special symmetric conditions in the pathloss coefficients such that an underlying system of fixed point equations can be uncoupled [35], the SINR achieved by user k when served from the coordinated cluster of stations in the set $\mathcal{N}_i$ is approximated by [35]:

$$\mathsf{SINR}_{ik} = \frac{(M|\mathcal{N}_i| - S_i + 1) \frac{\sum_{l\in\mathcal{N}_i} g_{lk}}{|\mathcal{N}_i|} \frac{\sum_{l\in\mathcal{N}_i} P_l}{S_i}}{1 + \sum_{j:j\notin\mathcal{N}_i} g_{jk} P_j}. \tag{12}$$

Note that despite the somewhat restrictive nature of the symmetric conditions required to get this expression, its accuracy is quite good, see [34], [36] for validation.

ACME: Finally, following the assumptions of large M and S_i and fixed $S_i/M \leq 1$ for the ACME system as well, a good asymptotic approximation of $\mathsf{SINR}_k(I_i)$ (extending the local MU-MIMO results to take into account beams serving users), where I_i is the independent set of user/beam tuples scheduled, is given by:

$$\mathsf{SINR}_k(I_i) = \frac{(M - S_{a(k)} + 1) g_{b(k)k} P_{a(k)} / S_{a(k)}}{1 + \sum_{j\in u(I_i)\setminus k} g_{b(j)k} P_{a(j)} / S_{a(j)}}, \tag{13}$$

with S_n and P_n representing the number of users served with MU-MIMO and the power for station n respectively, and functions $a(k)$ and $b(k)$ returning the station and the beam for user k. Finally, g_{bk} denotes the pathloss between beam b and user k, with b taking values $1, 2, \ldots, B \cdot N$ for a given ordering of the beams.

We proceed to validate the above formulas by comparing the analytic results with a Monte Carlo simulation. We derive random channels based on Rayleigh fading for the small-scale fading of the channel between the stations' antennas and the users' antennas [3], compute channel gains based on the widely accepted WINNER-II model [41] which captures the distance between a given station and a user and other "large-scale" effects, such as blocking objects, walls and trees, and implement MU-MIMO using ZFBF [3]. Figure 3 depicts a 30x30m conference hall with 200 users and 20 stations grouped in 4 clusters operating at non-overlapping 20 MHz channels, which we use for the comparison. In Figure 4 we see the results of the simulations and notice that the analytic CDF closely tracks the simulation one for every PHY technology with an error around 5-10%.

A. A motivating comparison for the low interference regime

To further motivate the use of the ACME system versus a tightly synchronized fully coordinated MU-MIMO system we examine the low interference regime and show that in this case, the ACME system can perform better. To simplify the comparison between the two systems we assume that every station is associated with an equal number of users ($|\mathcal{S}_i| = K/N$ for all $i = 1, \ldots, N$), every user is assigned to the station with the maximum gain in one of its beams ($b(k) = \text{argmax}_{i=1,\ldots,BN} g_{ik}$), and all stations use equal transmission power ($P_i = P$ for all $i = 1, \ldots, N$).

By combining Equations (9) and (13) for ACME, and Equations (9) and (12) for the fully coordinated system, assuming that the station density, channel reuse, and the shape of the beams allows users to be served with no inter- or intra-cluster interference, and forcing the stations to serve the maximum number of users possible (that is $S_i = \min\{M, K/N\} = M$, for a relative large number of users, for all $i = 1, \ldots, N$) we can see that the ACME system outperforms the fully coordinated system. Indeed:

$$R_k^{acme} = \frac{MN}{K} \log\left(1 + \mathsf{SINR}_k^{acme}\right), \tag{14}$$

with $\mathsf{SINR}_k^{acme} = \frac{g_{b(k)k} P}{M}$, whereas for the coordinated system:

$$R_k^{coor} = \frac{MN}{K} \log\left(1 + \mathsf{SINR}_k^{coor}\right), \tag{15}$$

with $\mathsf{SINR}_k^{coor} = \frac{\sum_{i=1}^N g_{ik} \frac{\sum_{i=1}^N P_i}{N}}{MN} = \frac{\sum_{i=1}^N g_{ik} P}{MN} \leq \mathsf{SINR}_k^{acme}$, where the last inequality comes from the fact that $\sum_{i=1}^N g_{ik}/N \leq g_{b(k)k}$ from the choice of $b(k)$ as the one maximizing the gain for user k. Thus, since $\mathsf{SINR}_k^{coor} \leq \mathsf{SINR}_k^{acme}$ and the multiplexing gains in (15) and (14) are equal, in this scenario ACME outperforms a coordinated MU-MIMO setup.

In the scenario above we force both systems to serve the maximum number of possible users which is not necessarily optimal throughput-wise [3]. Nevertheless, as we show in Appendix A, even when optimizing for the number of served users and under typical values of pathloss coefficients, still the ACME system may outperform the fully coordinated one. This statement might seem counterintuitive as the fully coordinated system is optimal in theory. However, it is justified by the assumptions in our model since, first, we compare the performance under a low interference regime where the superior interference suppression of a fully coordinated MU-MIMO system is less important, and second, we assume an

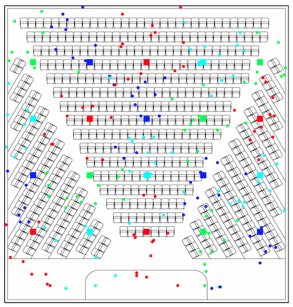


Fig. 3: Example topology.

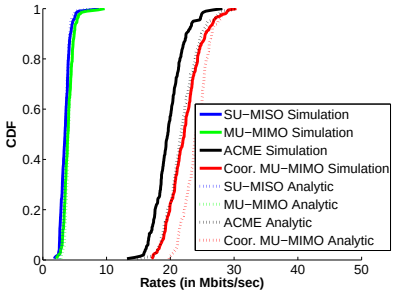


Fig. 4: Validation of analytic formulas.

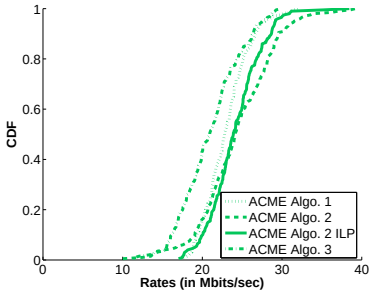


Fig. 5: Comparison of the different ACME scheduling algorithms.

equal transmission power per user. While this later assumption is usually true in practice, in theory it is suboptimal.

VI. Simulation Results

In this section we present the results of the evaluation of the ACME architecture compared to the technologies presented in Section V. We present both small scale and larger scale scenarios of interest. For larger scale scenarios where simulation results quickly become impractical we only have analytical results to show. For small scale scenarios we have both analytic and simulation results. For exposition we plot only the analytic results for consistency and since they are around 5% of the simulation results anyway (see, for example, Figure 4). Note that to obtain the analytical results we simply insert Equations (10)-(13) into (9) which gives us a rate that is used to generate the empirical CDFs that we present. Finally, we conclude this section with simulation results for the proposed MAC layer extensions using NS-2 network simulator [28].

We assume an enterprise WiFi setting where multiple APs with 3 RF chains each are placed in conference halls of various dimensions and user densities (see Figure 3 for an example topology). For this set of experiments we assume an 80MHz-wide band divided into 4 non-overlapping channels as is the case in the 802.11 2.4 GHz band and a transmit power of 90dB above the noise floor. The channel gain, g_{ik}, is a function of the distance, d_{ik}, between AP i and user k, the carrier frequency, f_c, and other "large-scale" effects, such as blocking objects, walls and trees. It is modeled using the B3 (indoor hotspot) scenario of the WINNER-II model [41]. According to this model the pathloss, g_{ik}, is given in dB from the formula below:

$$g_{ik}[dB] = A\log_{10}(d_{ik}[m]) + B + C\log_{10}(f_c[GHz]/5) + X,$$

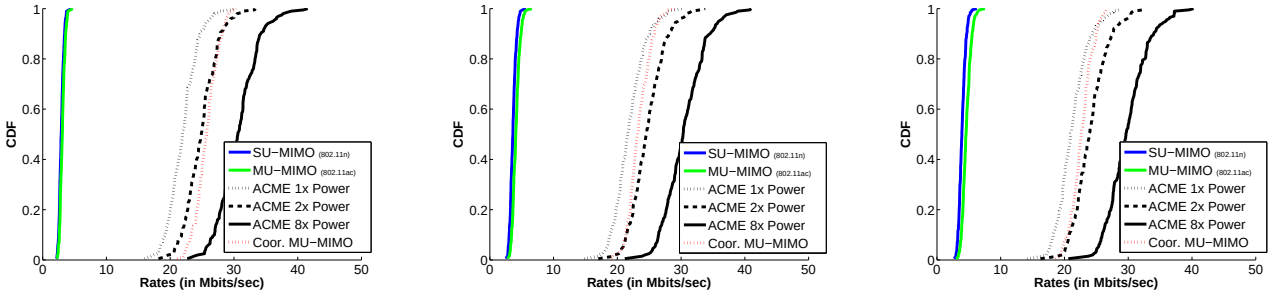
where A, B, C and X are scenario-dependent parameters which in our case take the values: $A = 13.9$, $B = 64.4$, $C = 20$, and $X = 0$.

For the coordinated MU-MIMO and ACME schemes, we group APs in four clusters, each operating at one of the four channels. For the ACME results, we assume that each RF chain is connected to a switched-beam antenna which is able to select from a fixed number of directional beams. Unless otherwise stated, we assume there are 16 45-degree beams available with a 22.5 degree step. We provide 3 alternatives for the directionality gain, 0 dB, 3 dB and 9 dB, which translate to 1x, 2x and 8x transmit power respectively. The 1x case corresponds to transmitting at the same power per degree as in the omnidirectional case. The 2x case follows FCC rules which allow to double the power per degree when using directional transmissions. The 8x case corresponds to transmitting the same total power as in the omnidirectional case.

Scheduling performance: We compare the performance of ACME under the three greedy scheduling algorithms presented in Section IV against the scheduling decisions obtained by expressing the conflict graph problem in Section IV-B as an integer linear program (ILP) and using Gurobi, a commercial integer programming solver which yields near-optimal results. In Figure 5 the CDFs of the user rates under different ACME scheduling approaches are presented, as obtained in a 40x40m conference hall where 20 APs serve 200 users. We notice that all greedy algorithms achieve a very large percentage of the rates achieved by solving the ILP. While Algorithm 2 yields slightly larger rates than Algorithm 1 (both do better than Algorithm 3), the later is a bit more favorable towards lower rate users and, without loss of generality, choose to use Algorithm 1 for scheduling purposes for the rest of the paper.

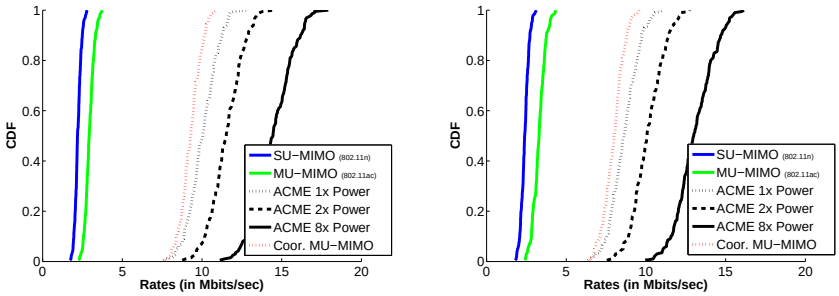
Upcoming dense deployments: In Table II we compare the avg. spectral efficiency of various schemes and their full CDFs can be seen in Figure 6. For these scenarios, a fixed number of APs and users (20 APs and 200 users) was used in different sized topologies, thus changing the density of APs and users. For example, in the 40x40m topology each AP covers an area of 80 sq.m whereas in the more dense 20x20m topology each AP covers an area of 20 sq.m. Also, assuming a typical seat setup in a conference hall, a 20x20m hall would fit around 500 people so the 200 users that we consider imply about 40% of the people are actively using the internet, whereas in the 40x40m hall only 10% of the people would be active. These values are inline with recent reports regarding densely populated venues [42]. As expected, the fully coordinated MU-MIMO transmissions benefit from a denser populated area since distances of APs and active users decrease without an interference increase. On the other hand, "local" 802.11n/ac take a performance hit from the increased interference.

Interestingly, the ACME system exhibits the same gains and average behavior as the interference free fully coordinated system. For example, in the 20x20m case it yields up to 10x improvement as compared to local 802.11ac (local MU-MIMO), with the coordinated MU-MIMO system trailing with



(a) 20x20 meters scenario. (b) 30x30 meters scenario. (c) 40x40 meters scenario.

Fig. 6: WiFi 200 users, 20 APs, user/AP ratio 10 scenarios.



(a) 20x20 meters scenario. (b) 40x40 meters scenario.

Fig. 7: WiFi 200 users, 8 APs, user/AP ratio 25 scenarios.

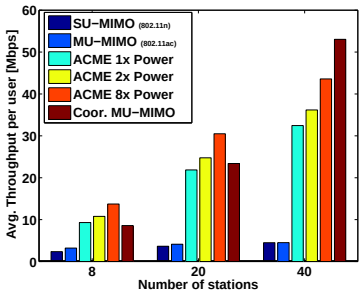


Fig. 8: Avg. Throughput for all 30x30 meters WiFi scenarios.

Topology	802.11n	802.11ac	ACME 1x	ACME 2x	ACME 8x	Coor. MU-MIMO
20x20	0.23	0.24 (1x)	1.69 (7x)	1.92 (8x)	2.36 (10x)	1.98 (8.5x)
30x30	0.28	0.32 (1x)	1.68 (5x)	1.90 (6x)	2.35 (7.5x)	1.85 (6x)
40x40	0.31	0.36 (1x)	1.63 (4.5x)	1.85 (5x)	2.30 (6.5x)	1.75 (5x)

TABLE II: Avg. spectral efficiency in bps/Hz for various topology dimensions. For all scenarios the number of APs is 20 and the number of users 200. Numbers in parenthesis show improvement over 802.11ac.

Topology	802.11n	802.11ac	ACME 1x	ACME 2x	ACME 8x	Coor. MU-MIMO
20x20	0.17	0.23 (1x)	0.77 (3.5x)	0.89 (4x)	1.11 (5x)	0.71 (3x)
30x30	0.18	0.25 (1x)	0.71 (3x)	0.83 (3.5x)	1.06 (4x)	0.66 (2.5x)
40x40	0.19	0.25 (1x)	0.66 (2.5x)	0.78 (3x)	1 (4x)	0.61 (2.5x)

TABLE III: Avg. spectral efficiency in bps/Hz for various topology dimensions. For all scenarios the number of APs is 8 and the number of users 200. Numbers in parenthesis show improvement over 802.11ac.

an 8.5x gain, and its spectral efficiency increases in denser AP deployments. Even without any power gains (1x case) ACME can still deliver a 7x improvement (see Figure 6a) which is very close to that of the fully coordinated system. This is explained by ACME's ability to serve many users concurrently almost interference free thanks to the smart combination of analog and digital beamforming.

Today's typical deployments: In Figure 7 and Table III, the number of APs was lowered increasing the user/AP ratio from 10 to 25 and decreasing the density of APs and thus the interference seen by local 802.11n/ac schemes. As a result, in this less dense scenario the gains of ACME and the coordinated MU-MIMO scheme over the local schemes are still sizable yet smaller than before, ranging from 2.5x to 5x. Also, in this lower interference regime, the ACME's approach to take advantage of MU-MIMO locally yields consistently higher gains than going for a fully coordinated MU-MIMO approach, see black (ACME) versus red (coordinated mu-MIMO) lines in Figure 7. As discussed in Section V-A already, the superiority of ACME over coordinated MU-MIMO even when no directional power gain is present (1x case) is in line with communication theory results because here we use equal power allocation per stream, a practical yet suboptimal choice. Also, section VI-A has a more general discussion about the difference between the practical coordinated MU-MIMO system considered here and what communication theory models assume when making optimality claims for a fully coordinated MU-MIMO system, further justifying this result.

Variable beam widths: Figure 9 shows the performance of ACME as we vary the width and number of beams that the front end supports, in a 40x40m hall with 20APs and 200 users. We proportionally decrease the beam width as the number of beams increases, thus creating narrower beams. As expected, a diminishing returns behavior is seen as the number of beams increases, slowly converging to the performance of a fully reconfigurable front end. We focused in most of the

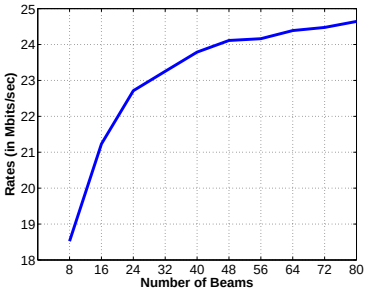


Fig. 9: ACME avg. user rate for variable number of beams.

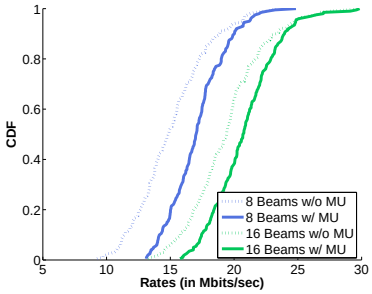


Fig. 10: MU-MIMO multiplexing gains in ACME.

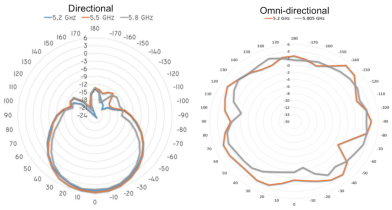


Fig. 11: Radiation patterns of the antennas used in the experiments.

simulations in the practical case of 16 45-degree beams with 22.5 degree steps since this is what switched-beam antennas can typically achieve today.

Multiple users per beam: ACME allows multiple users to be scheduled in the same beam by means of MU-MIMO when an AP uses multiple RF chains with the same/similar beam pattern. To investigate the additional gain from local MU-MIMO on top of the multiplexing thanks to the coordination across APs, in Figure 10 we compare the user rates of ACME with and without MU when the switched-beam antennas can select from either 45-degree beams as before or fatter 90-degree beams. With 90-degree beams local MU-MIMO gives an additional 15% gain. As expected, the gain is less pronounced with the 45-degree beams (10% gain) since not only less users are covered within a beam but also there is more opportunities to pack more non-interfering analog beams in space.

A. Real world scalability considerations

A word on density, clusterization and scalability is in order. In theory, one may envision a single mega-cluster of say 20 APs, collecting instantaneous channel state information for 60 antennas and computing a very large precoding matrix for this mega-cluster. This will clearly increase the performance of the coordinated MU-MIMO scheme by a factor of 4 (over the 4 clusters of 5 APs each scenario, see Fig 6) and of ACME by a somewhat smaller factor as the ability to pack beams in space does not scale equally well. But, in practice, it is virtually impossible to operate such a system. Along the same lines, in theory one may envision a situation where each AP consists of 4 "parallel systems" each operating in one of the four channels, forming 4 clusters in one channel each where now all APs participate in all clusters with one of their parallel systems. This system would again, in theory, quadruple the performance of the coordinated MU-MIMO scheme but clearly this is an even more impossible system to be found in the real world. Note that the ACME performance would increase by a smaller amount as the density of transmitting APs in a channel would increase by 4 and it won't be possible to pack that many beams in an interference free manner. The advantage of coordinated MU-MIMO in very dense scenarios like the above can be seen in Figure 8 when 40 APs are deployed in a 30x30 meter hall forming large clusters of 10 APs each. In accordance to well known results from wireless communication theory about the optimality of "global" MU-MIMO, the coordinated MU-MIMO system outperforms ACME. As we are interested in systems that have a chance of finding their way into real-world deployments, we do not consider mega-clusters or "parallel systems" any further.

B. MAC Simulations

In all the enterprise WiFi scenarios we presented in this section and for the typical power levels we have assumed, the unsolicited CTS message (CTS-to-self in the 802.11 standard) initiated by the cluster heads (see Section III-B) should be overheard and decoded by all users in the cluster, such that the interference from the UL is efficiently suppressed. The raw overhead inquired from the transmission of the CTS packet is in the order of ~ 67 μs (40 μs PHY header, ~ 17 μs MAC header and 10 μs a SIFS interval). This means that for typical downlink transmissions (a few ms that can grow larger with aggregation) this represents an overhead in the order of 1%.

In order to validate that the unsolicited CTS can successfully serve its purpose we have performed network level simulations using NS-2 [28], since it has a full implementation of the 802.11 CSMA/CA protocol. A setup similar to Figure 3 was recreated assuming a 30x30 meters conference hall. We report results for a cluster of 5 APs serving 20 users. The basic rate was set to 1 Mbps (rate in which CTS and other control packets are transmitted) and the data rate at 11 Mbps.[4] The NS-2 simulator was modified such that the cluster head AP transmits omni-directionally a CTS-to-self message, which, when overheard from clients made them defer their uplink transmissions for a predefined amount of time that is carried within the CTS package (duration field), and, when overheard from the rest of the APs of the cluster made them transmit concurrently in defiance of CSMA. Traffic was configured as a constant bit rate (CBR) application over a UDP connection. For the APs the traffic was configured at 4.5 Mbps and for users at 500 Kbps. (In the best case scenario where all 5 APs transmit concurrently, a total of $4.5 \times 5/20 = 1.125$ Mbps per user DL rate and 0.5 Mbps per user UL rate are sustained, implying a bit more than a 2:1 DL/UL traffic ratio in accordance with real world measurements.)

We verified that when the cluster head uses the ACME scheme and sends unsolicited CTS messages, the users are

[4]NS-2 has an implementation of 802.11b and thus, without loss of generality, 802.11b rates were used for the simulation.

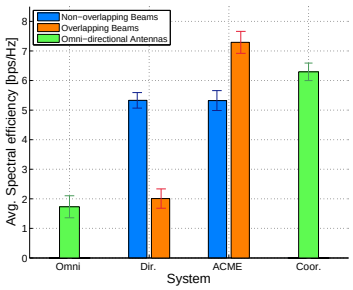


Fig. 12: Avg. spectral efficiency achieved for the different AP/user experimental setups.

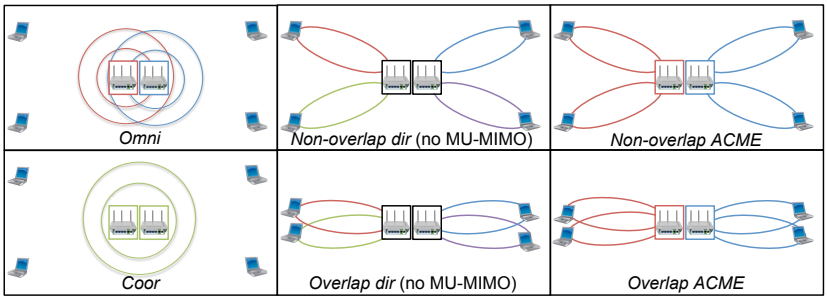


Fig. 13: Experimental setup topologies.

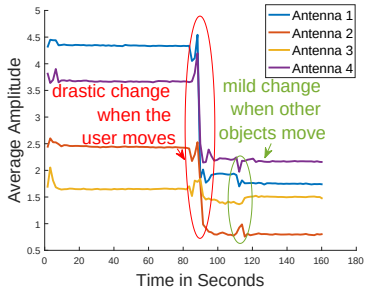


Fig. 14: Dynamics of long-term channel statistics.

silenced for the ensuing coordinated transmission. Also, the APs that belong to the ACME system overhear the CTS message carrying the cluster heads IP address and start transmitting to their respective users achieving this away a loose synchronization of their transmissions. Specifically, during a 40 sec simulation time, out of the 531 CTS-to-self messages the cluster head transmitted all of them were overheard and decoded from all users and APs in the cluster. Moreover, the CTS packets were decoded in a timely manner from all other APs in the cluster: the time offset between the CTS reception at the first AP to decode it and the CTS reception at the last AP to decode it was a mere ~ 70 ns, verifying that the duration of a coordinated downlink transmission is not unnecessarily elongated due to the asynchronous coordination feature, see discussion in Section III-A.

C. Overhead and Mobillity

A note on the MAC overhead experienced by the presented systems is in order. To enable MU-MIMO transmissions, CSI is required to be collected by both ACME and a fully coordinated system. By means of uplink-downlink channel reciprocity [3], this overhead can be kept comparable for both cases even as the number of users and APs is growing. Notice that using explicit feedback as per the 802.11ac standard, is almost prohibitive for the case of a fully coordinated system since each user would have to estimate the channel for each AP and explicitly send back this information to the corresponding AP. An overhead that is not shared from the ACME system and can significantly affect the goodput (effective throughput) of the fully coordinated system is that of timing and frequency synchronization (required for coherent downlink transmissions among remote APs) and distributed calibration (necessary for exploiting uplink-downlink channel reciprocity across multiple coordinated APs). An efficient protocol to achieve these goals has been introduced and analyzed in our previous work [22]. Accounting for this overhead would only increase ACME's relative advantage.

Notice that in the above results user mobility is consider nomadic since wireless devices connected to a WiFi network change their location in minute time-scales [43]–[45]. This leads to long-term channel statistics which remain stable for thousands of transmissions as long as such statistics do no change much in the presence of *other* moving objects. To verify this, we perform an experiment where we constantly measure the elements of the channel matrix for 3 minutes in a typical office room. In the first minute and a half the environment is relatively steady: about 6 people work on their desks. In the middle of the experiment we change the location of the measured user/device by about 1 meter. In the second half of the experiment, the 6 people begin to walk around in the space, opening and closing doors. As shown in Figure 14, the long-term channel gains are quite steady during the first half of the experiment, as expected. In the second half, the long-term gains vary very little due to moving objects, posing no issues to the selection of the right beams. Last, as expected, there is a sizable change when the user/antenna moves. Concluding, we confirm that in a typical (indoor) environment the long-term statistics of the channel matrix stay useful for a long time.

VII. Testbed Experiments

We, finally, verify the underlying concepts of ACME in a testbed setup. We use two WARP-3 SDN FPGAs [46] as our APs and connect two RF front-ends to each one of them. The directional and omni-directional antennas used were provided by Adant [47] and their radiation patterns can be seen in Figure 11. The following experiments (see Figure 13 for experimental setup schematics) were administered and the results can be found in Figure 12 where the average spectral efficiency in bps/Hz is depicted. The experiments were repeated 10 times for each setup and the average spectral efficiencies of the users were computed. Moreover, errors of 1 standard deviation are incorporated in the plot.

Omnidirectional with MU-MIMO (*Omni*): The two APs with two omni-directional antennas each serve two users each in an MU-MIMO fashion (without any coordination between the APs). This is representing today's uncoordinated 802.11ac setups. As expected, interference from the omni-directional antennas results in very low SINRs and a corresponding spectral efficiency smaller than 2 bps/Hz. Note that in this experimental scenario only one AP is interfering to a user. We have seen in the simulation results in Section VI that, as the APs density increases and the number of interfering APs grows, the SINR and spectral efficiencies reduce even more.

Non-overlapping directional without MU-MIMO (*Non-overlap dir*): The two APs, equipped with two directional antennas each, serve two users each. Users are carefully placed such that the interference to a user from the side and back lobes of antennas serving other users is low. For comparison purposes, the users are placed in the same locations as in the omnidirectional experiment (see Figure 13). The expectation that the directionality of the antennas mitigates a big part of the interference and thus leads to SINR gains compared to the *Omni* setup is met and SINRs $> 15dB$ were achieved for all 4 users leading to an average spectral efficiency of 5.3bps/Hz.

Overlapping directional without MU-MIMO (*Overlap dir*): We alter the previous setup by placing users assigned to the same AP closer to each other. As a result, those users cannot be targeted with two directional antennas without sizable interference to each other, unless MU-MIMO precoding is used. Indeed, the interference to a user from the lobe serving the other user assigned to the same AP yields a very low SINR and an average spectral efficiency lower than 2bps/Hz.

Non-overlapping directional with MU-MIMO (*Non-ovelap ACME*): Here we have the same setup and user placement as in the *Non-overlap dir* experiment but now MU-MIMO precoding abilities of the APs are used to mitigate any residual interference to one user assigned to an AP from the same AP's antenna serving the other user. Since the residual interference from side and back lobes was small, the gains from using MU-MIMO are, as expected, negligible. Specifically, the average spectral efficiency is close to 5.3bps/Hz like in the *Non-overlap dir* setup. It is interesting to note that the standard deviation here is larger, which is due to the dependence of MU-MIMO in instantaneous channel state information, which increases the variability among repetitions of the experiment.

Overlapping directional with MU-MIMO (*Overlap ACME*): We alter the previous setup by placing users assigned to the same AP closer to each other such that the beams serving them are overlapping. In the *Overlap dir* experiment, without the use of MU-MIMO, we found that users get $\sim 4dB$, however, using MU-MIMO to mitigate interference leads to an SINR of $\sim 22dB$ for all users leading to an average spectral efficiency of 7.3bps/Hz. It's interesting to note here that the average SINR achieved is greater than the case where the beams are not overlapping. This is a by-product of the fact that the directional antenna modules used have a very small back lobe but bigger side lobes apart from the main lobe (see radiation pattern in Figure 11). Thus, as can be seen in Figure 13 the interference from the neighboring AP is smaller in this case since the users are only in the back-lobe of the non-serving AP. Compare this to the non-overlap user placement scenario where users are affected also from the side lobes of the non-serving AP.

Coordinated MU-MIMO (*Coor*): The two APs are connected with a clock line and simultaneously serve the 4 users in an MU-MIMO fashion. This represents a coordinated MU-MIMO system with tight phase/frequency synchronization. As expected, due to the complete absence of interference this scenario will give high SINR for all users. Indeed, it achieves an SINR of $\sim 18dB$ and an average spectral efficiency of 6.3bps/Hz, which is a bit higher than in the *Non-ovelap ACME* scenario. If the users are placed closer to each other, then the performance of *Coor* lowers a bit because the corresponding channel matrix has a worse condition number (less channel diversity). Interestingly, notice that the *Coor* setup SINR even with favorable user placement is lower compared to the *Overlap ACME* setup. In addition to the directional power gain, the WARP v3 boards implement MU-MIMO using equal power allocation instead of waterfilling (which is also what all commercially available 802.11ac wave 2 chipsets do as well), providing an explanation for the superiority of ACME over the coordinated MU-MIMO system, in line with the simulation results presented in Section VI.

Finally, before concluding this section we draw some comparisons between simulation and experimental results. Table III results are appropriate for such a comparison since a system with 8 APs divided in 4 non-overlapping channels consists of 2 APs per cluster competing for the same channel. Figure 11 implies that the power gain between the directional and omnidirectional modes of the used antennas is about 3-4dB thus we will be focusing on the 2x directional power gain cases in the table. We see from the results comparison, that, although the absolute values of spectral efficiencies are different (due to differences in transmission powers, distances between APs and users, and number of antennas per AP) the relative results are comparable. For instance, in simulations, the ACME system with a 2x directionality gain has a 4x relative spectral efficiency gain over its 802.11ac counterpart, and, in experiments, the *Non-ovelap ACME* and *Overlap ACME* cases give about a 3-4x relatively gain over the *Omni* case. The coordinated MU-MIMO case also gives similar relative gains which are around 3-3.5x in both simulations and experiments.

VIII. Conclusion

We have presented and evaluated the performance of ACME, an asynchronous, loosely coordinated multi-time scale beamforming architecture. We have shown that it is highly efficient realizing the gains of a coordinated MU-MIMO deployment without the overhead that comes with it.

Appendix A
Optimal number of users under low interference

The following proposition gives the optimal choice of scheduled users for the low interference regime:

Proposition 1. *Assuming $M \leq K/N$, the optimal number of users to be scheduled in the low interference regime is $A_{coor}(MN+1) - 1 \leq S_{opt} \leq A_{coor}(MN+1) + 1$ for the Coor. MU-MIMO and*
$A_{acme}(M+1) - 1 \leq S_{a(k),opt} \leq A^k_{acme}(M+1) + 1$

for the ACME system, where: $A_{coor} = \frac{GP}{e^{W((G-1)/e)+1}+GP-1}$ *and* $A^k_{acme} = \frac{g_k P}{e^{W((g_k-1)/e)+1}+g_k P-1}$, *with* $G = \sum_{i=1}^{N} g_{ik}$, $g_k = g_{b(k)k}$ *and* $W(x)$ *denote the Lambert W-function.* [5]

Proof. The proof follows by taking the derivatives for finding the number of users maximizing (9) for the different SINRs from (12) and (13) respectively and using the Lambert W-function to treat the resulting fixed point equation. □

Corollary 1. *The maximum achievable average rate for user* k *is given, under the assumptions of Proposition 1, by:* $R^{coor}_{k,opt} = \frac{A_{coor}(MN+1)}{K} \log\left(1 + \frac{(1-A_{coor})}{A_{coor}} \sum_{i=1}^{N} g_{ik} P\right)$ *and* $R^{acme}_{k,opt} = \frac{NA_{acme}(M+1)}{K} \log\left(1 + \frac{(1-A_{acme})}{A_{acme}} g_k P\right)$.

Proof. We can prove this by direct substitution of the results of Proposition 1 in Equations (14) and (15). □

It is worthwhile to notice that the spectral efficiency (in-log factor) does not depend on the number of APs and antennas per AP, but rather on the transmit power and the pathloss coefficients. Moreover, for typical values of pathloss coefficients in less dense deployments we can see that $R^{coor}_{k,opt} \leq R^{acme}_{k,opt}$.

[5]The Lambert W-function is defined by $W(x)e^{W(x)} = x$ for $x \geq -e^{-1}$.

REFERENCES

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691 – 1706, Jul. 2003.

[2] A. J. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 5, pp. 684–702, June 2003.

[3] A. Molisch, *Wireless Communications*. Wiley-IEEE, 2005.

[4] H. V. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire, "Achieving high data rates in a distributed MIMO system," in *IEEE MOBICOM*, 2012, pp. 41–52.

[5] H. Balan, R. Rogalin, A. Michaloliakos, K. Psounis, and G. Caire, "Airsync: Enabling distributed multiuser MIMO with full spatial multiplexing," *IEEE/ACM Trans. Netw.*, vol. 21, no. 6, pp. 1681–1695, Dec 2013.

[6] H. S. Rahul, S. Kumar, and D. Katabi, "JMB: scaling wireless capacity with user demands," in *ACM SIGCOMM*, 2012, pp. 235–246.

[7] S. Yun, L. Qiu, and A. Bhartia, "Multi-point to multi-point MIMO in wireless lans," in *IEEE INFOCOM*, April 2013, pp. 125–129.

[8] A. Adhikary, J. Nam, J.-Y. Ahn, and G. Caire, "Joint spatial division and multiplexing - the large-scale array regime," *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6441–6463, Oct 2013.

[9] A. Adhikary *et al.*, "Joint spatial division and multiplexing for mm-wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, June 2014.

[10] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, March 2014.

[11] A. Adhikary and G. Caire, "Jsdm and multi-cell networks: Handling inter-cell interference through long-term antenna statistics," in *Asilomar*, Nov 2014, pp. 649–655.

[12] M. Kurras, L. Thiele, and G. Caire, "Interference mitigation and multiuser multiplexing with beam-steering antennas," in *WSA*, March 2015, pp. 1–5.

[13] X. Xie, E. Chai, X. Zhang, K. Sundaresan, A. Khojastepour, and S. Rangarajan, "Hekaton: Efficient and practical large-scale MIMO," in *Mobicom*. ACM, 2015, pp. 304–316.

[14] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, "Coordinated multipoint transmission/reception techniques for LTE-advanced [coordinated and distributed MIMO]," *IEEE Wireless Commun. Mag.*, vol. 17, no. 3, pp. 26–34, June 2010.

[15] H. Dahrouj and W. Yu, "Coordinated beamforming for the multicell multi-antenna wireless system," *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1748–1759, May 2010.

[16] R. Zakhour, Z. K. M. Ho, and D. Gesbert, "Distributed beamforming coordination in multicell MIMO channels," in *VTC Spring 2009 - IEEE 69th Vehicular Technology Conference*, April 2009, pp. 1–5.

[17] M. Blanco, R. Kokku, K. Ramachandran, S. Rangarajan, and K. Sundaresan, "On the effectiveness of switched beam antennas in indoor environments," in *PAM*. Springer Berlin Heidelberg, 2008, vol. 4979, pp. 122–131.

[18] A. P. Subramanian, H. Lundgren, and T. Salonidis, "Experimental characterization of sectorized antennas in dense 802.11 wireless mesh networks," in *ACM MOBIHOC*, 2009, pp. 259–268.

[19] X. Liu, A. Sheth, M. Kaminsky, K. Papagiannaki, S. Seshan, and P. Steenkiste, "DIRC: Increasing indoor wireless capacity using directional antennas," in *ACM SIGCOMM*, 2009, pp. 171–182.

[20] ——, "Pushing the envelope of indoor wireless spatial reuse using directional access points and clients," in *ACM MOBICOM*, 2010, pp. 209–220.

[21] R. Choudhury, X. Yang, R. Ramanathan, and N. Vaidya, "On designing MAC protocols for wireless networks using directional antennas," *IEEE Trans. Mobile Comput.*, vol. 5, no. 5, pp. 477–491, May 2006.

[22] R. Rogalin *et al.*, "Scalable synchronization and reciprocity calibration for distributed multiuser MIMO," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 1815–1831, 2014.

[23] Q.-V. Pham, H.-L. To, and W.-J. Hwang, "A multi-timescale cross-layer approach for wireless ad hoc networks," *Computer Networks*, vol. 91, pp. 471 – 482, 2015.

[24] H. Qiu, K. Psounis, G. Caire, K. M. Chugg, and K. Wang, "High-rate WiFi broadcasting in crowded scenarios via lightweight coordination of multiple access points," in *ACM Mobihoc*, 2016.

[25] G. Foschini and M. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311–335, 1998.

[26] A. Kuchar, M. Tangemann, and E. Bonek, "A real-time doa-based smart antenna processor," *IEEE Trans. Veh. Commun.*, vol. 51, no. 6, pp. 1279–1293, Nov 2002.

[27] H. Qiu, K. Psounis, G. Caire, K. M. Chugg, and K. Wang, "High-rate WiFi broadcasting in crowded scenarios via lightweight coordination of multiple access points," in *ACM MobiHoc*. New York, NY, USA: ACM, 2016, pp. 301–310.

[28] T. Issariyakul and E. Hossain, *Introduction to Network Simulator NS2; 2nd ed.* Dordrecht: Springer, 2012.

[29] O. Goussevskaia, M. Halldorsson, and R. Wattenhofer, "Algorithms for wireless capacity," *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 745–755, Jun. 2014.

[30] G. Dimic and N. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 10, pp. 3857 – 3868, Oct. 2005.

[31] D. Welsh, *Matroid Theory*. Dover Publications, 2010.

[32] H. Shirani-Mehr, G. Caire, and M. Neely, "MIMO downlink scheduling with non-perfect channel state knowledge," *IEEE Trans. Commun.*, vol. 58, no. 7, pp. 2055–2066, July 2010.

[33] S. Sakai, M. Togasaki, and K. Yamazaki, "A note on greedy algorithms for the maximum weighted independent set problem," *Discrete Applied Mathematics*, vol. 126, pp. 313–322, Mar. 2003.

[34] A. Michaloliakos, R. Rogalin, Y. Zhang, K. Psounis, and G. Caire, "Performance modeling of next-generation WiFi networks," *Computer Networks*, pp. –, 2016.

[35] H. Huh, A. Tulino, and G. Caire, "Network MIMO with linear zero-forcing beamforming: Large system analysis, impact of channel estimation, and reduced-complexity scheduling," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 2911–2934, 2012.

[36] N. Anand, J. Lee, S. Lee, and E. Knightly, "Mode and user selection for multi-user MIMO wlans without csi," in *IEEE INFOCOM*, April 2015.

[37] Y. Lim, C. Chae, and G. Caire, "Performance analysis of massive MIMO for cell-boundary users," *IEEE Trans. Wireless Commun.*, vol. PP, no. 99, pp. 1–1, 2015.

[38] B. Hochwald, T. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sept 2004.

[39] K.-K. Wong and Z. Pan, "Array gain and diversity order of multiuser miso antenna systems," *Int. J. of Wirel. Inf. Netw.*, vol. 15, no. 2, pp. 82–89, 2008.

[40] R. Couillet and M. Debbah, *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

[41] P. Kyösti and et. al, "WINNER II channel models," Information Society Technologies, Tech. Rep., 2007.

[42] Extreme Networks, "Record-Breaking Wi-Fi Usage at Super Bowl LI," 2017. Available: https://content.extremenetworks.com/h/i/326569881

[43] M. Balazinska and P. Castro, "Characterizing mobility and network usage in a corporate wireless local-area network," in *ACM MobiSys*. New York, NY, USA: ACM, 2003, pp. 303–316.

[44] A. Balachandran, G. M. Voelker, P. Bahl, and P. V. Rangan, "Characterizing user behavior and network performance in a public wireless lan," in *ACM SIGMETRICS*. New York, NY, USA: ACM, 2002, pp. 195–205.

[45] D. Tang and M. Baker, "Analysis of a local-area wireless network," in *ACM MOBICOM*. New York, NY, USA: ACM, 2000, pp. 1–10.

[46] Mango Communications, "Warp v3 kit." Available: https://mangocomm.com/products/kits/warp-v3-kit

[47] Adant Technologies Inc. Available: http://www.adant.com/

Antonios Michaloliakos graduated in 2009 from the department of Electrical and Computer Engineering of the National Technical University of Athens. He received a PhD in Electrical Engineering from the University of Southern California in 2016. His interests span the areas of probabilistic modeling and design of networks and cross-layer optimization in new generation wireless networks.

Weng Chon Ao received the B.S. degree in computer science and the M.S. degree in electrical engineering from National Taiwan University, Taiwan. He is currently pursuing the Ph.D. degree in electrical engineering at University of Southern California, Los Angeles, USA. He received IEEE PIMRC 2013 Best Paper Award. His research interests include wireless communication systems and data center networking.

Konstantinos Psounis is an Associate Professor of Electrical Engineering and Computer Science at the University of Southern California. He received his first degree from the department of Electrical and Computer Engineering of National Technical University of Athens, Greece, in June 1997, and the M.S.and PhD degrees in Electrical Engineering from Stanford University, California, in January 1999 and December 2002 respectively. Konstantinos models and analyzes the performance of a variety of wired and wireless networks. He also designs methods, algorithms, and protocols to solve problems related to such systems. He is a senior member of both IEEE and ACM.

Yonglong Zhang graduated from the department of Electrical and Computer Engineering of the Ecole Supérieure des Sciences et Technologies de L'Ingénieur de Nancy in 2012. He is a PhD student in Electrical Engineering at the University of Southern California advised by Konstantinos Psounis. His research interests are performance analyzing of computer networks and WLAN resource management. He is currently working on user-AP association protocol for enterprise Wi-Fi networks.