Demystifying Resource Allocation Policies in Operational 5G mmWave Networks

Phuc Dinh[1*], Moinak Ghoshal[1*], Dimitrios Koutsonikolas[1], Joerg Widmer[2]
[1]Institute for the Wireless Internet of Things, Northeastern University, USA
[2]IMDEA Networks Institute, Spain
{dinh.p,ghoshal.m,d.koutsonikolas}@northeastern.edu, joerg.widmer@imdea.org

***Abstract*—5G mmWave is being rapidly deployed by all major mobile operators. With the technology still in its infancy, several early research works analyze the performance of operational 5G mmWave networks. Nonetheless, these measurement studies primarily focus on single-user performance, leaving the sharing and resource allocation policies largely unexplored. In this paper, we fill this gap by conducting the, to our best knowledge, first systematic study of resource allocation policies of current 5G mmWave mobile network deployments through an extensive measurement campaign across four major US cities and two major mobile operators. Our study reveals that resource allocation among multiple flows is strictly governed by the cellular operators and flows are not allowed to compete with each other in a shared queue. Operators employ simple threshold-based policies and often over-allocate resources to new flows with low traffic demands or reserve some capacity for future usage. Interestingly, these policies vary not only among operators but also for a single operator in different cities. We also discuss a number of anomalous behaviors we observed in our experiments across different cities and operators.**

I. Introduction

5G mmWave is being rapidly deployed by all major mobile operators and a large number of user equipment (UE) devices support the new technology. However, as commercial 5G mmWave deployments are still in their infancy, the actual performance experienced by cellular users remains largely unknown. A few recent works [1]–[4] conducted measurement studies of 5G mmWave networks in terms of performance, coverage, energy consumption, and the impact on application QoE. A common lesson out of these studies is that, although today's mmWave deployments indeed offer Gbps throughput and lower latency than 4G LTE, their performance is often suboptimal, coverage is sporadic, the handover process is not optimized, and applications cannot always take advantage of the full potential of 5G mmWave.

Interestingly, all these studies focus almost exclusively on single-user performance, leaving the sharing and resource allocation policies at 5G mmWave base stations (BS) largely unexplored. In the 5G mmWave landscape, where operators promise multi-Gbps data rates and bandwidth-hungry applications demand Gbps data rates, it is critical to understand how flows share the available resources. Since even a single flow can occupy a substantial fraction of the operator's resources at a BS, timely and efficient resource (re-)allocation is extremely important to avoid unfairness and starvation of flows. In this work, we conduct the, to our best knowledge, first systematic study of resource allocation policies in operational 5G mmWave networks. Through an extensive measurement campaign across 4 major US cities and 2 mobile operators, we shed light on the policies used by mobile operators to allocate wireless capacity to flows with diverse traffic demands.

*Phuc Dinh and Moinak Ghoshal are co-primary authors.

TABLE I: Dataset statistics.

Number of individual iperf3 tests	800+
Amount of cellular data used (GB)	3600+
Number of cities	4
Number of operators	2
Number of cloud servers	3
Cumulative time of measurements traces (minutes)	450+

Our study faces several challenges. First, unlike WiFi networks, cellular networks are "black boxes" from the UE's point of view; we have no direct insight into the operations performed on the BS side. Second, as previous studies have shown, 5G mmWave performance is affected by a variety of factors including the environment (transient and permanent blockages) and the transport layer protocol. Third, during our experiments, we have no control over the other users who might be sharing the same 5G mmWave cell. To address the first challenge, we design a systematic set of experiments that allows us to uncover the sharing policies in an incremental fashion, starting with scenarios involving backlogged flows and gradually moving towards heterogeneous scenarios with diverse traffic demands. To address the other two challenges, we repeat our experiments sufficiently often to carefully filter out those impacted by external factors, in order to isolate the performance impact due to the presence of only the flows controlled by us. Our main contributions and findings can be summarized as follows:

- We conduct the first systematic study of resource allocation and sharing policies in operational 5G mmWave networks across 2 major mobile operators and 4 major US cities. We perform a total of 800+ iperf3 tests and collect over 3600+ GB worth of data. Key statistics of this work are summarized in Table I.

• We find that resource sharing is strictly governed by the 5G operator and multiple flows do not directly compete against each other in a shared queue. Mobile operators neither employ the well-known proportional fairness policies (which are considered the de facto standard for opportunistic schedulers in cellular networks) nor do they aim at maximizing the total throughput. Instead, they leverage threshold-based resource allocation policies based on user traffic demands and often over-allocate resources to new flows or reserve some capacity for future use. Interestingly, these policies vary not only among operators but also for a single operator in different cities.
• We discuss a number of anomalous behaviors across cities and operators. We observe cases where the operator delays the update of the resource allocation of existing flows for several seconds or does not update it at all, when a new flow is added to the network. In some cities, we also observe that new flows may not start at all in the presence of existing flows, indicating that the operator never allocates any resources to them. Since the 5G mmWave deployments are still in their infancy, we believe that such anomalous behaviors will be eliminated as the technology becomes more mature.

II. Methodology

A. 5G UE, Carriers, Locations, and Cloud Servers

5G UE. We primarily use rooted Google Pixel 5 phones as UEs. The Pixel 5 radio supports the 5G mmWave bands n260/261 and 4-CC (4x100 MHz)/1-CC downlink/uplink carrier aggregation. For some tests, we also use a Samsung Galaxy S21 phone to measure the peak network performance via Ookla's Speedtest [5]. The S21 supports 8-CC carrier aggregation and thus achieves higher download speeds than the Pixel 5. However, since the S21 phone cannot be rooted, we cannot use it for our main measurements that require root tools such as iperf3 and tcpdump.
Carriers. We use Verizon's and AT&T's 5G mmWave services, two of the largest US operators that have a widespread mmWave deployment all over the country. Verizon's 5G mmWave service works in the 28 and 39 GHz frequency bands (n260/261), whereas AT&T works only in the 39 GHz (n260) band. Both operators adopt the 5G Non-Standalone (NSA) architecture, which shares the packet core with the 4G infrastructure. We also tried to do measurements with T-Mobile, the third major mobile operator in the US, but the Pixel 5 we used for our measurements did not support T-Mobile's 5G mmWave service. We ran a few speed tests with the Samsung S21 phone and measured an average throughput of 800-900 Mbps on T-Mobile, significantly lower than the throughput achieved over Verizon and AT&T (1.5-2 Gbps). Hence we do not report any further results with T-Mobile.
Locations. We conducted extensive measurements across 4 major cities in the US: Boston, Chicago, Indianapolis, and Atlanta. Details are shown in Table II. To identify mmWave coverage areas in each city, we consulted the coverage maps provided by the cellular operators. In Chicago and Indianapolis, both operators adopt a dense deployment, with BSs mounted on top of traffic lights, around 50-60 ft from the ground, mostly a block away from each other. Fig. 1a shows an example of a BS deployed in Chicago. In Atlanta, we conducted measurements in a large park, spanning an area of approximately 1.9 km^2. The structural design and density of the AT&T BSs is similar to that in Chicago and Indianapolis but they are mounted on top of light poles in the park. In contrast, the Verizon 5G mmWave deployment in Boston is very different. BSs are mounted on the walls of high buildings as shown in Fig. 1b in a sparse manner with two BSs being up to a few miles apart from each other.

TABLE II: Cities and the 5G operators used in this work.

City	Operators
Boston	Verizon
Chicago	Verizon, AT&T
Indianapolis	Verizon, AT&T
Atlanta	AT&T

(a) Chicago (b) Boston

Fig. 1: 5G mmWave BS deployments in different cities.

Cloud Servers. We tested three different cloud services – Google Cloud, Amazon Web Services (AWS), and Amazon Wavelength 5G Edge Services. The Google Cloud server we used in this work is located in Washington, DC and has 32 GB of memory, 8vCPUs and Ubuntu 18.04. The AWS server is located in Ashburn, Virginia and has similar specifications and performance as the Google Cloud server. However, we noticed some AWS performance outliers with drastically reduced network throughput. In contrast, the Amazon Wavelength servers are hosted in so-called Wavelength zones, allowing the UE application traffic to reach the application servers without exiting the cellular network. This provides ultra low latency (almost half of what we achieve with Google Cloud or AWS servers) but the performance in terms of throughput is similar. Since the Wavelength zones are not available in all of the four cities and work only with Verizon, all of the measurements reported in this work are done with Google Cloud.

B. Experiments

Since we are interested in performance in the presence of multiple flows with heterogeneous traffic demands, we use UDP for our experiments. This gives us more control over the flows' traffic rates and eliminates the impact of transport layer rate control (reaction to loss, congestion

control, slow start) on the measured performance, making it easier to isolate and understand the operator resource allocation policies. We use iperf3 to generate traffic logged every 100 ms and run tcpdump on the phone to capture packet traces. The Google Cloud server used for the measurements supports up to a total of 16 Gbps for all egress flows to external IP addresses. Hence, we can run multiple iperf3 instances on the same cloud server to send traffic to multiple clients.

We designed a set of systematic iperf3 tests with several UEs that allowed us to uncover the operator sharing policies in an incremental fashion, starting with scenarios involving backlogged flows and gradually moving towards heterogeneous scenarios with diverse traffic demands. *Experiment 1* involved two and three backlogged clients. *Experiment 2* involved two clients, one with backlogged traffic and another one with intermittent traffic of gradually increasing rate. *Experiment 3* involved three clients, one with backlogged traffic and the other two downloading at different fixed rates. Finally, *Experiment 4* also involved three clients, the first with backlogged traffic, the second with continuous fixed-rate traffic, and the third with intermittent traffic of gradually increasing rate. The details of these experiments are described in §V and §VI. The duration of the measurements varied from 20 s to 230 s based on the nature of the experiments we performed. For all experiments, we first run a downlink UDP session for 10 s on one of the phones, before the other clients start receiving data traffic. For each operator-city combination, we extract this 10 s worth of throughput and define it as the baseline throughput in Sec. III.

We use automated scripts to control the start of the iperf3 sessions and impose certain delays between them. Since all phones have individual system clocks, we need to synchronize them to observe the effect of resource sharing between the different flows over time. To this end, we use the ClockSync [6] app, which synchronizes the device system clock with atomic time from the Internet via NTP (Network Time Protocol). We then match the timestamps from the tcpdump traces collected during the experiments to align the throughput of the respective phones.

In each of the 4 cities, we carefully chose a measurement location with strong mmWave coverage for our experiments. With the exception of a small number of mobile experiments to understand the nature of resource sharing (contention-based vs. operator-controlled) described in §IV, all experiments are static with the user in line-of-sight (LOS) of the BS and the UE facing the BS.

To eliminate or at least mitigate the impact of external factors on the measured performance and isolate the impact of the operator resource allocation policies, we took the following steps. In Boston, Atlanta, and Indianapolis, we performed the experiments at times and places with minimal human and vehicle presence. This ensured that factors like transient blockage or background data usage did not affect our measurements. In Chicago, the BSs

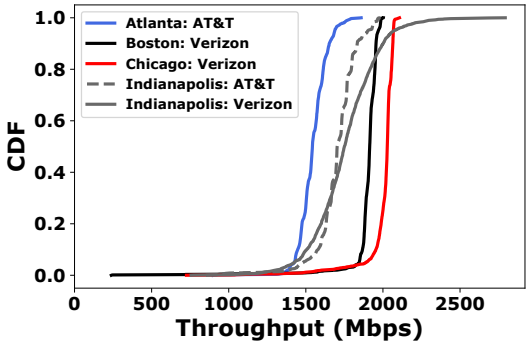


Fig. 2: UDP single flow throughput across cities and operators.

are deployed in a very crowded part of the city and are surrounded by tall buildings and trees, while cars and humans are always moving around. To mitigate the impact of such external factors on the measured performance, we stood very close to the BS. While measuring the network performance with a single client in all cities, we occasionally observed prolonged periods where throughput was low, often dropping below 75% of its expected capacity. We conjecture that the operator allocates fewer resources as it might have competing traffic in the back-haul from other bands, such as sub-6 GHz 5G networks. We carefully filtered out such cases and (to the extent possible) used only a clean set of measurements to study the 5G mmWave resource allocation policies. Also, we observed cases with momentary throughput drops due to channel disruptions. Since such situations are beyond our control and do reflect the behavior of the today's 5G mmWave networks, we kept these scenarios in our measurement dataset.

We faced two additional challenges with our AT&T experiments. In Chicago, although AT&T has a strong mmWave coverage, it employes a rate limiting policy after one or two sessions of backlogged downlink traffic, reducing the average throughput from 1000 Mbps to less than 100 Mbps for the next 10-15 minutes. Also, we were unable to perform any measurements involving multiple clients as always one or more flows failed to receive any traffic from the cloud server. This behavior was consistent, regardless of the time of day. In contrast to Chicago, in Indianapolis we were able to complete a partial set of multi-client experiments with AT&T. We were able to complete Experiments 1, 2, and 3.1, but we could not get any successful runs for Experiments 3.2 and 4. For Experiments 1, 2, and 3.1, the likelihood for an experiment to fail was much higher than for it to succeed, with roughly one successful run for every 5 failed attempts.

III. Baseline Measurements

To analyze the operators' resource allocation policies, we first have to establish as baseline performance the maximum throughput achieved by a single client. Fig. 2 shows the CDF of 100 ms downlink throughput samples of a single client for all city-operator combinations. We

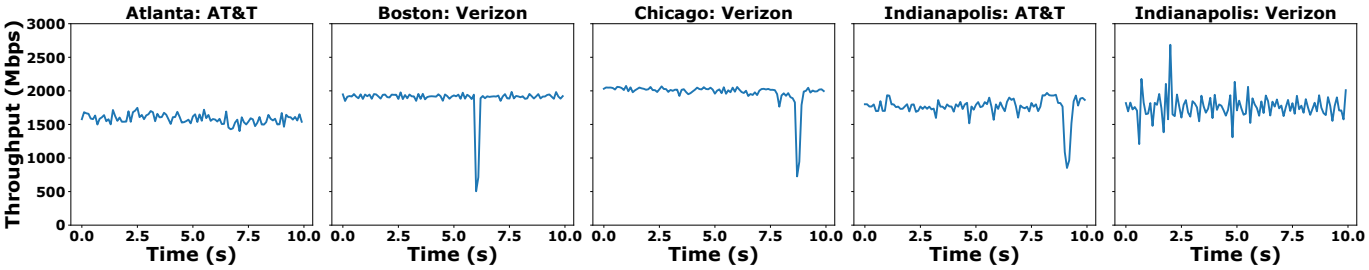


Fig. 3: Single client throughput patterns across cities and operators.

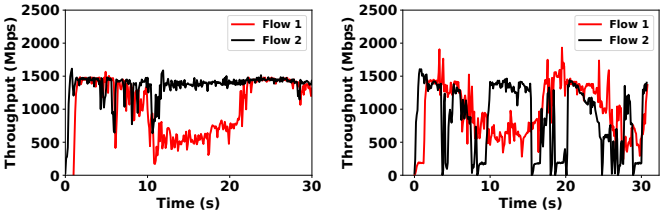
observe that both operators provide multi-Gbps downlink throughput in all four cities but the performance varies between the two operators and even for the same operator across different cities. The median throughput for Verizon in Chicago, Boston, and Indianapolis is 2 Gbps, 1.9 Gbps, and 1.75 Gbps, respectively. The AT&T throughput is generally lower, with median values of 1.7 Gbps in Indianapolis and 1.5 Gbps in Atlanta. For some operator-city combinations, we observe a long tail of very low throughput values (700 Mbps or lower), which we attribute to short-term channel fluctuations. Fig. 3 shows three example timelines exhibiting such instantaneous throughput drops for Verizon in Boston and Chicago and AT&T in Indianapolis. Interestingly, we observed no tail for AT&T in Atlanta – the city-operator combination with the lowest average performance in Fig. 2 (see again Fig. 3 for an example timeline). We also notice a distinct saw tooth pattern for Verizon in Indianapolis, where the throughput alternately rises and drops, as shown in Fig. 3. This pattern is consistent in all our traces. Since the experiments in Indianapolis were conducted in LOS in the absence of any pedestrians or cars, it is unlikely that fluctuating channel conditions are the reason for these large throughput variations. Instead, we conjecture that these fluctuations are due to buffering at the BS, causing packets to queue up and be released in bursts.

IV. Resource Sharing in 5G Networks

Next, we analyze whether the resource sharing in 5G mmWave networks is contention-based or controlled by the operator. We conducted two sets of experiments with Verizon in Boston. In the first experiment, one user held the UE, facing the BS and another user walked towards and away from the BS in a pseudo random manner. Both users had a backlogged UDP iperf3 session downloading traffic from the cloud server. Fig. 4a shows that each flow is allocated a fixed bandwidth, which never exceeds 1.5 Gbps, even though their maximum achievable throughput can be much higher, as we saw in Fig. 2. In particular, Flow 1 never gets more than 1.5 Gbps even when the throughput of Flow 2 drops significantly due to mobility. This suggests that (i) *the two flows do not compete against each other in a shared queue* and (ii) *the operator neither employs proportional fairness (considered the de facto standard in cellular networks) for allocating resources nor tries to maximize the total throughput*. Instead, the resource allocation is solely controlled by the operator's policy, which allocates a fixed capacity to each flow.

In the second experiment, we let both users download backlogged UDP traffic walk in a pseudo-random pattern. Fig. 4b shows that most of the time, the throughput of each flow is again limited to 1.5 Gbps, confirming our previous hypothesis that the operator allocates a certain amount of resources to each flow, independent of the channel conditions. Nonetheless, we observe a few instances where the throughput of Flow 1 exceeds 1.5 when the throughput of Flow 2 drops to very low values, e.g., at 4 s or from 18-20 s. We conjecture that the operator sometimes employs a threshold based resource allocation policy; if the throughput for a particular flow drops below a certain threshold for a given amount of time, it reacts by allocating more resources to the other flow.

We conclude that in 5G mmWave networks flows are treated independently and resource sharing is mainly driven by the operator's threshold-based policies. The experiments above were conducted with Verizon in Boston, and we verified that the same behavior holds for Verizon as well as AT&T in the other cities through the controlled experiments discussed in §V and §VI. Having established the general type of resource sharing policies employed by operators, we proceed to shed light on the details of these policies across operators and cities.



(a) Flow 1: static, Flow 2: mobile.

(b) Flow 1: mobile, Flow 2: mobile.

Fig. 4: Independent resource sharing between two clients under fluctuating channel conditions.

V. Verizon Resource Allocation Policies

In this section, we describe the four experiments mentioned in §II-B and use their results to uncover the details of Verizon's resource allocation policies in three different cities – Boston, Chicago, and Indianapolis. The bar graphs

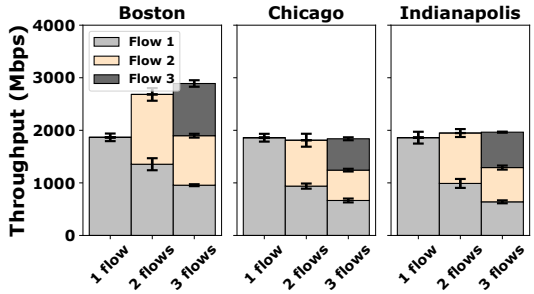


Fig. 5: Verizon, Experiment 1: Backlogged clients.

in the figures in this and the next section show average values and the error bars denote the standard deviations.

A. Experiment 1: Backlogged clients

We conduct Experiment 1 to estimate the network capacity. We introduce 3 backlogged UDP flows (Flow 1, Flow 2, Flow 3) from the server to three Pixel 5 clients, starting at 0 s, 10 s, and 20 s, respectively. Flow 1 lasts for 60 s, Flow 2 lasts for 40 s, and Flow 3 lasts for 20 s.

Fig. 5 shows the average throughput values for each flow when there are one, two, and three co-existing flows in the network. The total height of the highest bar in each city represents an estimate of the network capacity. We observe that the network capacity is not the same in all three cities. In Boston, the capacity is 2.9 Gbps (sum throughput of 3 backlogged flows), while two backlogged flows achieve a slightly lower sum throughput (2.8 Gbps). Interestingly, a single flow only achieves a throughput of 1.9 Gbps, as we also saw in Fig. 2, due to the low carrier aggregation (4CC) supported by the Pixel 5. The achieved throughput with a Samsung S21 phone with 8CC carrier aggregation is 2.9 Gbps, equal to the network capacity. In contrast to Boston, in the other two cities, the Verizon network capacity does not exceed 2 Gbps. Regardless of the capacity value, all backlogged flows are allocated an equal share of the capacity in all three cities.

B. Experiment 2: Two clients downloading backlogged and varying non-backlogged traffic

In Experiment 2 we study how the network reacts to the presence of both backlogged and non-backlogged traffic. Here, we introduce two flows. Flow 1 continuously injects backlogged traffic for 230 s, while Flow 2 injects traffic of gradually increasing rate in 10 s intervals, with gaps of 10 s between subsequent traffic intervals. Fig. 6a shows the measured per-flow throughput as a function of the injected rate of Flow 2. We make three observations.

First, in Boston, the sum throughput decreases initially, when the traffic for Flow 2 is low (10-200 Mbps), but starts increasing as Flow 2's rate increases (> 500 Mbps), until it reaches the capacity (around 2.9 Gbps). This indicates that *the operator imposes a threshold-based policy to the allocated capacity*. It initially allocates only a fraction of the available capacity (about 2 Gbps out of 2.9 Gbps) and only removes this limitation when Flow 2's rate exceeds 500 Mbps. In contrast, in Chicago and Indianapolis the operator always allocates the full capacity of only 1.9 Gbps to the clients regardless of the traffic demand.

The second observation concerns the throughput of the backlogged flow (Flow 1). The arrival of a non-backlogged flow (Flow 2) reduces the throughput of Flow 1, and the reduction is higher than the rate of the non-backlogged flow, suggesting that *the operator allocates to the non-backlogged flow more capacity than it demands, presumably as a safety margin*. Table III shows the actual capacity allocated to Flow 2 for different traffic demands, calculated as the difference between the network capacity and the capacity allocated to Flow 1 (i.e., the measured throughput of Flow 1).[1] We observe that different safety margins are used in different cities; the safety margins in Chicago and Indianapolis are much higher than the ones in Boston for a given Flow 2 injection rate. However, as the traffic demand of Flow 2 increases beyond 1000 Mbps in Chicago and beyond 1200 Mbps Indianapolis, the safety margin cannot be maintained since the network capacity is limited to only 1.9 Gbps. Overall, we observe that the operator employs an *over-provisioning resource sharing policy* as long as the network capacity allows.

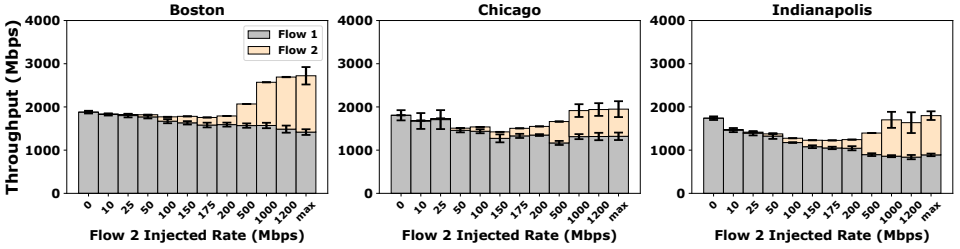
The final observation concerns our measurements in Chicago. The last bar for Chicago in Fig. 6a, corresponding to the case when Flow 2 also injects backlogged traffic, shows that the two flows do not share the capacity equally, contradicting Fig. 5. In fact, in our experiments in Chicago we observed both cases (equal and unequal sharing), which is reflected by the large standard deviation in Fig. 6a, but unequal sharing, which we consider abnormal, occurred more often. Second, Fig. 6a also shows that the throughput of Flow 1 does not decrease monotonically as the rate of Flow 2 increases, even though the operator still employs the over-provisioning allocation policy. Both these anomalous behaviors are due to a pathological scenario, which we call *failed allocation update*, where the network fails to adjust its allocation in accordance with changing traffic demands over time. We provide more details about this scenario in §VII.

C. Experiment 3: Three clients downloading backlogged and fixed-rate non-backlogged traffic

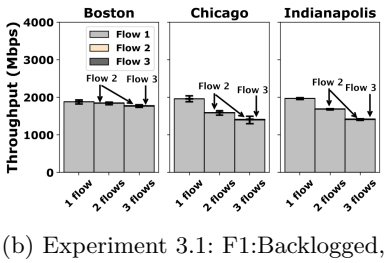
In Experiment 2, we established two principles for resource allocation in the presence of two clients: over-provisioning resource allocation for non-backlogged flows and threshold-based capacity limitation (only for Boston). The purpose of Experiment 3 is to verify these two principles in the presence of 3 clients. We consider two cases:

1) Experiment 3.1: Flow 1 starts at 0 s and generates continuous backlogged traffic for 70 s, Flow 2 starts at 10 s and generates fixed-rate traffic of 10 Mbps for 50

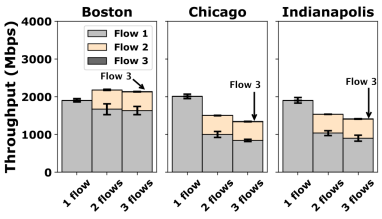
[1]Note that the values shown in Table III do not always agree with Fig. 6a. Fig. 6a shows the average values over all experiments, some of which having transient channel fluctuations as mentioned in Sec. I, which we remove from Table III.



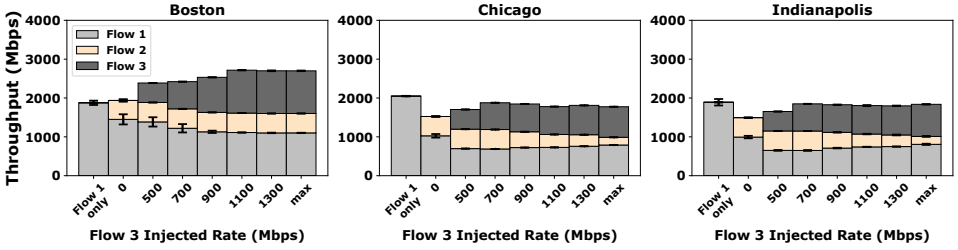
(a) Experiment 2: F1:Backlogged, F2:Non-backlogged.



(b) Experiment 3.1: F1:Backlogged, F2:10 Mbps, F3:10 Mbps.



(c) Experiment 3.2: F1:Backlogged, F2:500 Mbps, F3:10 Mbps.



(d) Experiment 4: F1:Backlogged, F2:500 Mbps, F3:Variable.

Fig. 6: Verizon, Experiments 2, 3, 4.

TABLE III: Verizon: Estimated capacity allocated to Flow 2 (in Mbps) as a function of Flow 2's rate.

	Flow 2 Injected Rate (Mbps)										
	10	**25**	**50**	**100**	**150**	**175**	**200**	**500**	**1000**	**1200**	**max**
Boston	74	107	138	236	275	326	316	1154	1151	1235	1304
Chicago	526	559	569	593	684	735	766	841	845	844	875
Indianapolis	396	579	644	795	891	920	926	989	1110	1131	1080

s, and Flow 3 starts at 20 s and generates intermittent 10-s traffic bursts of 10 Mbps, with 10-s silence intervals between every two traffic bursts. In Fig. 6b, we observe a reduction in the throughput of Flow 1 in the presence of Flow 2, consistent with the results in Experiment 2, and an additional reduction in the presence of Flow 3. Again, the allocated capacity to Flow 3 is higher than its traffic demand. The throughput of Flow 2 is not reduced in the presence of Flow 3.

2) Experiment 3.2: Experiment 3.2 is similar to Experiment 3.1, but Flow 2 generates traffic at a rate of 500 Mbps instead of 10 Mbps. As shown in Fig. 6c, the throughput of Flow 1 again reduces further in the presence of Flow 3, while the throughput of Flow 2 remains unchanged. The reduction in the throughput of Flow 1 is again higher than the traffic demand of Flow 3. Together, Experiment 3.1 and Experiment 3.2 confirm that the over-provisioning resource sharing policy also applies in 3-flow scenarios.

D. Experiment 4: Three clients downloading three different types of traffic

In Experiment 3.1 and Experiment 3.2, the third flow had a very low traffic demand (10 Mbps). We conduct Experiment 4 to observe the resource allocation policies in the case of three flows, when all flows have high traffic demands. Flow 1 starts at 0 s and generates backlogged traffic for 150 s. Flow 2 starts at 10 s and generates continuous traffic at a rate of 500 Mbps for 130 s. Finally, Flow 3 starts at 20 s and generates intermittent 10-s traffic bursts of gradually increasing rate, with 10-s silence intervals between every two traffic bursts. Interestingly, Fig. 6d shows that the operator employs two different policies, one in Boston and another one in Chicago and Indianapolis. In Boston (the city with the highest capacity), we observe that as the rate of Flow 3 gradually increases, the throughput of the backlogged flow (Flow 1) decreases, but the throughput of Flow 2 remains unchanged. In contrast, in the other two cities, the operator chooses to reduce the throughput of the non-backlogged flow (Flow 2), while allocating more resources to the backlogged flow, as the demand of Flow 3 increases.

VI. AT&T Resource Allocation Policies

In this section, we study the resource allocation and sharing policies of AT&T in two cities, Indianapolis and Atlanta. We conduct the same four experiments as in §V and contrast our findings against those for Verizon.

A. Experiment 1

Fig. 7 shows the results for Experiment 1. Here, we observe larger standard deviations compared to Verizon (Fig. 5), due to transient channel fluctuation as discussed in Sec. III. The estimated capacity is around 1.9 Gbps for both cities, but the average throughput (for a single flow) or sum-throughput (for multiple flows) is lower, as was established in Fig. 2. Similar to Verizon, backlogged flows obtain an equal share of the capacity in both cities.

B. Experiment 2

Fig. 8a and Table IV show that the allocated capacity for Flow 2 increases monotonically as its traffic demand

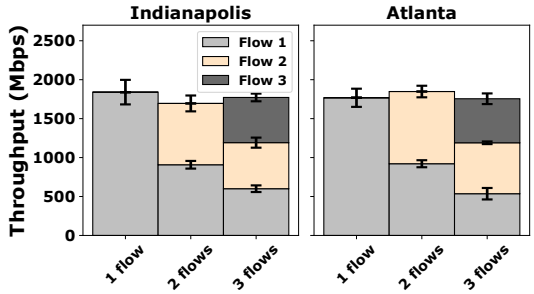


Fig. 7: AT&T, Experiment 1: Backlogged clients.

increases and the operator again maintains a safety margin to accommodate new flows, up to the point where the traffic demand of the new flow grows too large (around 1000 Mbps for both cities). The observed safety margin for AT&T in Indianapolis is much smaller than in Atlanta and is also smaller than the margins used by Verizon in Chicago and Indianapolis, but larger than the safety margin used by Verizon in Boston. Overall, we conclude that both operators use the same policy (over-provisioning resource allocation for small flows) but with different parameters (network capacity, safety margin) across operators and across different cities for the same operator.

C. Experiment 3

While the overall trend in Figs. 8b and 8c is similar to that observed for Verizon in Figs. 6b and 6c, we also see cases where adding a third flow does not reduce the throughput of the backlogged flow (Flow 1). This anomaly was observed occasionally for experiments 3.1 and 3.2 in Atlanta (see the large standard deviations in Figs. 8b and 8c for Atlanta) and consistently for Experiment 3.2 in Indianapolis. We call the anomaly of the network not reallocating capacity when a new flow is introduced *failed allocation update*, and discuss it further in §VII.

D. Experiment 4

Fig. 8d shows that, as we increase the sending rate of Flow 3, the operator starts reducing the throughput of the the non-backlogged flow (Flow 2), similar to the policy applied by Verizon in Chicago (Fig. 6d).

VII. Anomalous behavior

Since 5G mmWave deployments are still in their infancy, we encountered unexpected behavior that does not match with the identified sharing policies. In this section, we highlight three types of such anomalous behavior.

A. Delayed allocation update

In some experiments, we observed that when a new flow enters the network, the update of the resource allocation is delayed by a few seconds. Fig. 9 shows representative timelines of Experiment 3.2, described in §V-C and §VI-C, for selected operator-city combinations. In Figs. 9a and Fig. 9b, we observe that when Flow 3 is introduced, it takes 1-2 s for Flow 1's throughput to drop. In contrast, Fig. 9c and Fig. 9d show examples for Verizon in Indianapolis and AT&T in Atlanta with no visible delay.

B. Failed allocation update

In addition to delayed allocation updates, we sometimes observed that the operator did not update the resource allocation at all for an existing flow when a new flow was introduced. Fig. 10a shows an example timeline of Experiment 3.1, described in §V-C, for Verizon in Chicago. We observe that, when Flow 3 injects traffic into the network between 20-30 s, the throughput of Flow 1 remains unaffected. This behavior is very different from the general trend observed in Fig. 6b, where the operator always reduces Flow 1's throughput when a new flow arrives. However, in the same run we observe that on introducing Flow 3 again between 40-50 s, the operator reacts immediately and drops Flow 1's throughput. Fig. 10b shows a timeline of Experiment 3.1, where the operator allocates resources for existing flows properly every time a new flow enters the network. We also observed this anomalous behavior for AT&T in Atlanta (see Fig. 10c vs. Fig. 10d).

When the traffic demand of the new flow is very low (as in Experiment 3.1, where Flow 3 only requests 10 Mbps), its throughput is not affected by a failed allocation update. However, such failed allocation updates can have a severe impact on the throughput of flows with high traffic demands. Figs. 10e and 10f show an example for Experiment 2, described in §V-B. In Figs. 10e, the two flows share resources as expected, whereas in Fig. 10f, the operator does not drop the throughput of Flow 1 and allocates a capacity of only 500 Mbps to Flow 2 when the traffic demand of Flow 2 is higher than 1000 Mbps.

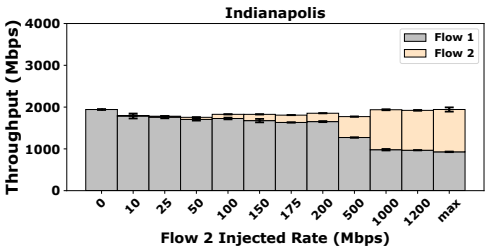
C. Flow startup failure

We also observed that sometimes a flow failed to start at all when there were other flows in the network. For instance, in Fig. 11a, Flow 3 does not start at 20 s, when Flow 1 and Flow 2 are already active. In Fig. 11b, Flow 2 does not start at 10 s, when Flow 1 is active, but Flow 3 starts properly at 20 s.

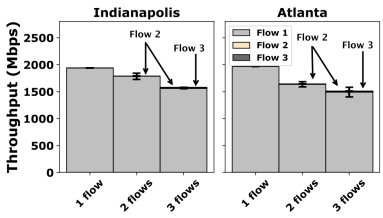
VIII. Discussion

In this section, we summarize common trends and major differences in the resource allocation policies of the two operators across different cities.

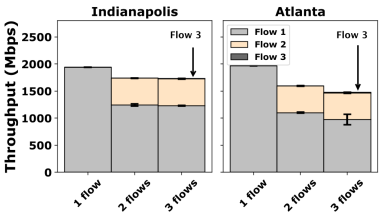
1) ***Network capacity:*** With the exception of Verizon in Boston, our measurements for both operators in other cities showed the network capacity to be below 2 Gbps. However, Verizon in Boston imposed a threshold-based limitation on the total capacity of 2.9 Gbps. As discussed in §V, this limitation is removed when there is at least one backlogged flow and another flow with a sending rate of 500 Mbps or higher; otherwise, flows are only allocated a total of 2 Gbps. We also notice that the network capacity may vary depending on the time of the day, dropping by up to 25%, as we mentioned in §II-B. We conjecture that



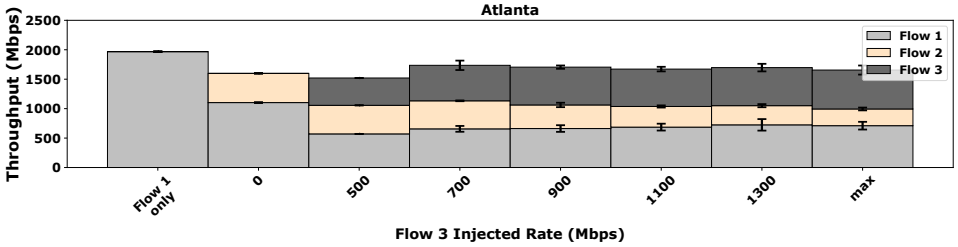
(a) Experiment 2: F1:Backlogged, F2:Non-backlogged



(b) Experiment 3.1: F1:Backlogged, F2:10 Mbps, F3: 10 Mbps.



(c) Experiment 3.2: F1:Backlogged, F2:500Mbps, F3:10 Mbps.

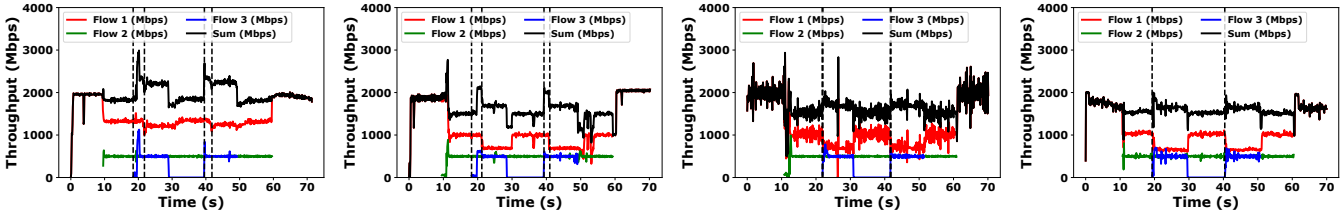


(d) Experiment 4: F1:Backlogged, F2:500 Mbps, F3:Variable.

Fig. 8: AT&T, Experiments 2, 3, 4.

TABLE IV: AT&T: Estimated capacity allocated to Flow 2 (in Mbps) as a function of Flow 2's rate.

	Flow 2 Injected Rate (Mbps)										
	10	**25**	**50**	**100**	**150**	**175**	**200**	**500**	**1000**	**1200**	**max**
Indianapolis	187	199	235	226	260	311	302	684	963	975	1014
Atlanta	331	425	526	745	723	733	760	961	970	976	972



(a) Verizon, Boston: Allocation update with delay.

(b) Verizon, Chicago: Allocation update with delay.

(c) Verizon, Indianapolis: Allocation update with no delay.

(d) AT&T, Atlanta: Allocation update with no delay.

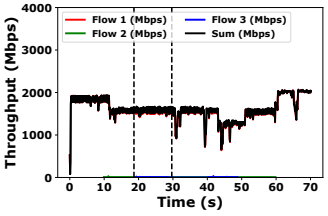
Fig. 9: Measurement timelines for experiments with and without delayed allocation updates.

this capacity variation is due to network dimensioning. We note that 5G mmWave clients still have to compete for resources at the backhaul level with sub-6 GHz clients.

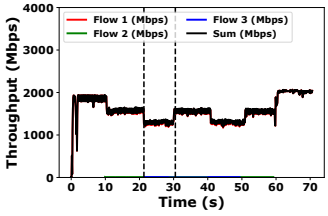
*2) **Equal sharing among backlogged flows**:* A common trend we observe throughout our measurements is that multiple backlogged flows are allocated an equal share of the capacity, regardless of their startup order. The only exception was in our two-flow sharing experiments for Verizon in Chicago (Fig. 6a), where Flow 2 was often allocated a smaller share of the capacity capacity than Flow 1 due to failed allocation updates.

*3) **Over-provisioned capacity for non-backlogged flows**:* Another common trend is that operators allocate more capacity than required by the actual traffic demand of small (mice) flows, presumably as a safety margin. However, this policy also reduces the performance of the existing large (elephant) flows and wastes capacity.

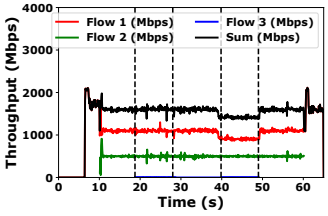
*4) **Resource allocation update policies**:* When a new flow is introduced in the network, the network has to update its resource allocation. Since the network capacity is limited, such allocation updates can result in reduced allocated capacity for some of the existing flows. When there is only one previous flow in the network, its allocated capacity will be reduced to accommodate the new flow, as seen in Fig. 6a. A more interesting case is when there are multiple existing flows with different traffic patterns. Here, we observed two trends. When the traffic demand of the new flow is low, the capacity allocated to the backlogged flow is always reduced. Figs. 6b and 6c are two typical examples. In contrast, when the traffic demand of the new flow is high, we saw that the operator typically chooses to also penalize the (lower rate) non-backlogged flow, as demonstrated in Fig. 6d for Verizon in Chicago and Indianapolis and Fig. 8d for AT&T in Atlanta, resulting in a less fair resource allocation. Verizon in Boston is the only exception to this second trend, where the backlogged flow is always penalized regardless of the traffic demand of



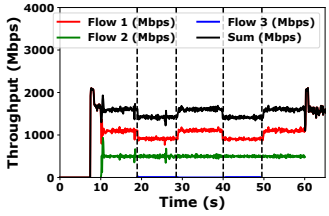
(a) Verizon, Chicago: Failed allocation update.



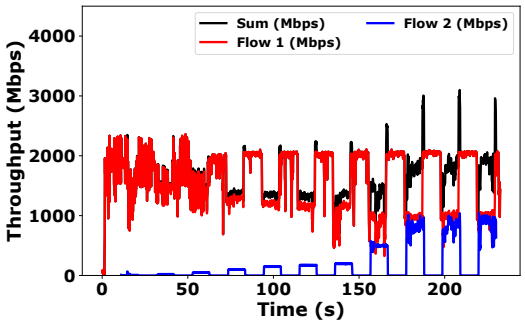
(b) Verizon, Chicago: Successful allocation update.



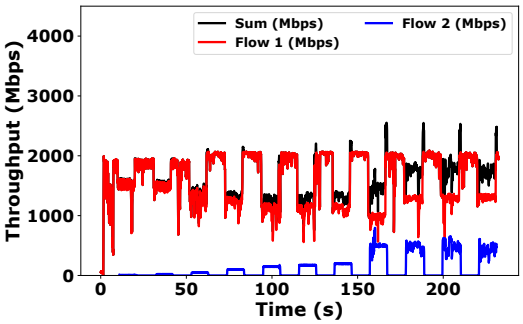
(c) AT&T, Atlanta: Failed allocation update.



(d) AT&T, Atlanta: Successful allocation update.

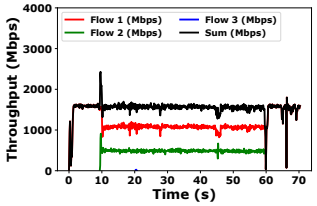


(e) Verizon, Chicago: Successful allocation update.

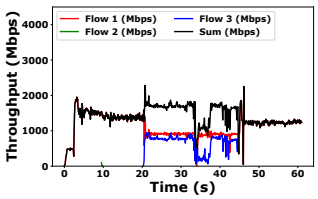


(f) Verizon, Chicago: Failed allocation update.

Fig. 10: Measurement timelines for experiments with successful and failed allocation updates.



(a) AT&T, Indianapolis: Flow startup failure.



(b) AT&T, Indianapolis: Flow startup failure.

Fig. 11: Measurement timelines for experiment with flow startup failure.

the new flow, and the (smaller) rate of the non-backlogged flow remains unchanged.

IX. Related Work

5G NR Resource Allocation. Resource allocation in 5G networks has been extensively studied [7]–[23]. However, this large body of research primarily focuses on mathematical modeling and optimization frameworks to analyze and improve 5G network resource allocation policies. The authors in these works generally examine different categories of resource allocation problems including computational resource allocation, backhaul resource allocation, power allocation, or bandwidth allocation. Besides, performance evaluation in these works is conducted via simulations under idealistic network settings. Due to the sheer volume of such literature, we refer interested readers to some related surveys [24]–[29] for a more in-depth discussion of the topic. As opposed to the aforementioned works, our work is the first to study the resource allocation policies of operational 5G mmWave networks in an empirical manner, through a systematic measurement study.

5G NR Measurements. There has been a limited number of measurement studies on the performance of 5G NR, especially at mmWave bands. In 2019, Qualcomm released a white paper [30] as one of the first reports on 5G performance profiling with the main focus on physical layer performance and coverage. The works in [1]–[4] conducted measurement studies of 5G mmWave networks exploring performance, coverage, energy consumption, and the impact on application QoE and the work in [31] conducted a similar study for sub-6 GHz 5G in China. Interestingly, all these studies focus almost exclusively on single-user performance, leaving the sharing and resource allocation policies employed by mobile operators largely unexplored. To the best of our knowledge, the work in [4] is the only one that conducted an experiment involving two phones, each downloading backlogged traffic, and concluded that the two phones achieve comparable performance, which we also confirmed in this study. In contrast to that work, our work conducts the first systematic study of resource allocation policies across different operators and cities in a variety of scenarios involving different numbers of clients and different traffic patterns.

X. Conclusion

In this work, we conducted the first systematic measurement study to demystify resource allocation and sharing policies in operational 5G mmWave networks. Our study is comprised of extensive measurements across four different cities with the two largest 5G mmWave mobile operators in the US. We first established the allocated capacity for a single client with our baseline measurements, observing different performance patterns both city-wise and operator-wise. Then, we investigated the resource allocation strategies for multi-client scenarios, from which

we drew a number of conclusions concerning the general trends and differences in the policies of the deployed 5G mmWave networks. Despite common policies such as over-provisioning resource sharing and equal sharing among backlogged flows, the different networks also exhibit very distinct characteristics in terms of overall capacity, timeliness of the resource reallocation, success rate of flow establishment, and fairness. Among all, a typical difference is the configured safety margin when over-allocating resources for small flows. Furthermore, we observed and categorized occasional unexpected suboptimal network behavior throughout the entire measurement campaign, which we refer to as anomalous behavior. Overall, the operator policies appear to be simple and, at times, unstable and unpredictable. The instability in the network operations may lead to adverse impact on real 5G mmWave users with more complex traffic patterns and needs to be addressed in future research.

XI. Acknowledgments

This work is partially supported by the Madrid Regional Government through the TAPIR-CM program (S2018/TCS-4496).

References

[1] A. Narayanan, E. Ramadan, J. Carpenter, Q. Liu, Y. Liu, F. Qian, and Z.-L. Zhang, "A First Look at Commercial 5G Performance on Smartphones," in *Proc. of ACM WWW*, 2020.

[2] A. Narayanan, E. Ramadan, R. Mehta, X. Hu, Q. Liu, R. A. K. Fezeu, U. K. Dayalan, S. Verma, P. Ji, T. Li, F. Qian, and Z.-L. Zhang, "Lumos5G: Mapping and Predicting Commercial MmWave 5G Throughput," in *Proc. of ACM IMC*, 2020.

[3] A. Narayanan, X. Z. R. Zhu, A. Hassan, S. Jin, X. Zhu, X. Zhang, D. Rybkin, Z. Yang, Z. M. Mao, F. Qian, and Z.-L. Zhang, "A Variegated Look at 5G in the Wild: Performance, Power, and QoE Implications," in *Proc. of ACM SIGCOMM*, 2021.

[4] M. I. Rochman, V. Sathya, N. Nunez, D. Fernandez, M. Ghosh, A. S. Ibrahim, and W. Payne, "A Comparison Study of Cellular Deployments in Chicago and Miami Using Apps on Smartphones," in *Proc. of ACM WiNTECH*, 2022.

[5] OOKLA, "Speedtest," https://www.speedtest.net/.

[6] S. Baranov, "ClockSync," https://clocksync.en.uptodown.com.

[7] M. Ismail, A. Abdrabou, and W. Zhuang, "Cooperative Decentralized Resource Allocation in Heterogeneous Wireless Access Medium," *IEEE Transactions on Wireless Communications*, 2013.

[8] L. Xu, H. Xing, A. Nallanathan, Y. Yang, and T. Chai, "Security-Aware Cross-Layer Resource Allocation for Heterogeneous Wireless Networks," *IEEE Transactions on Communications*, 2019.

[9] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wen, and M. Tao, "Resource Allocation in Spectrum-Sharing OFDMA Femtocells With Heterogeneous Services," *IEEE Transactions on Communications*, 2014.

[10] H. Dai, Y. Huang, and L. Yang, "Game Theoretic Max-logit Learning Approaches for Joint Base Station Selection and Resource Allocation in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communications*, 2015.

[11] S. Kim, B. G. Lee, and D. Park, "Energy-Per-Bit Minimized Radio Resource Allocation in Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, 2014.

[12] X. Sun and S. Wang, "Resource Allocation Scheme for Energy Saving in Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, 2015.

[13] Z. Chen, T. Li, P. Fan, T. Q. S. Quek, and K. B. Letaief, "Cooperation in 5G Heterogeneous Networking: Relay Scheme Combination and Resource Allocation," *IEEE Transactions on Communications*, 2016.

[14] I. Alqerm and B. Shihada, "Sophisticated Online Learning Scheme for Green Resource Allocation in 5G Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Mobile Computing*, 2018.

[15] M. Peng, Y. Wang, T. Dang, and Z. Yan, "Cost-Efficient Resource Allocation in Cloud Radio Access Networks With Heterogeneous Fronthaul Expenditures," *IEEE Transactions on Wireless Communications*, 2017.

[16] W. Hao, O. Muta, and H. Gacanin, "Price-Based Resource Allocation in Massive MIMO H-CRANs With Limited Fronthaul Capacity," *IEEE Transactions on Wireless Communications*, 2018.

[17] B. Xu, Y. Chen, J. R. Carrión, and T. Zhang, "Resource Allocation in Energy-Cooperation Enabled Two-Tier NOMA HetNets Toward Green 5G," *IEEE Journal on Selected Areas in Communications*, 2017.

[18] M. Moltafet, P. Azmi, N. Mokari, M. R. Javan, and A. Mokdad, "Optimal and Fair Energy Efficient Resource Allocation for Energy Harvesting-Enabled-PD-NOMA-Based HetNets," *IEEE Transactions on Wireless Communications*, 2018.

[19] M. Liu, T. Song, and G. Gui, "Deep Cognitive Perspective: Resource Allocation for NOMA-Based Heterogeneous IoT With Imperfect SIC," *IEEE Internet of Things Journal*, 2019.

[20] H. Dai, Y. Huang, J. Wang, and L. Yang, "Resource Optimization in Heterogeneous Cloud Radio Access Networks," *IEEE Communications Letters*, 2018.

[21] A. Mokdad, P. Azmi, N. Mokari, M. Moltafet, and M. Ghaffari-Miab, "Cross-Layer Energy Efficient Resource Allocation in PD-NOMA Based H-CRANs: Implementation via GPU," *IEEE Transactions on Mobile Computing*, 2019.

[22] M. Ali, Q. Rabbani, M. Naeem, S. Qaisar, and F. Qama, "Joint User Association, Power Allocation, and Throughput Maximization in 5G H-CRAN Networks," *IEEE Transactions on Vehicular Technology*, 2017.

[23] J. Li, M. Peng, Y. Yu, and Z. Ding, "Energy-Efficient Joint Congestion Control and Resource Optimization in Heterogeneous Cloud Radio Access Networks," *IEEE Transactions on Vehicular Technology*, 2016.

[24] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter Wave Communication: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 1616–1653, 2018.

[25] N. Xia, H.-H. Chen, and C.-S. Yang, "Radio Resource Management in Machine-to-Machine Communications—A Survey," *IEEE Communications Surveys Tutorials*, 2018.

[26] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A Survey on Resource Allocation for 5G Heterogeneous Networks: Current Research, Future Trends, and Challenges," *IEEE Communications Surveys Tutorials*, 2021.

[27] W. Ejaz, S. K. Sharma, S. Saadat, M. Naeem, A. Anpalagan, and N. Chughtai, "A comprehensive survey on resource allocation for CRAN in 5G and beyond networks," *Journal of Network and Computer Applications*, 2020.

[28] Y. Xu, G. Gui, H. Gacanin, and F. Adachi, "A Survey on Resource Allocation for 5G Heterogeneous Networks: Current Research, Future Trends, and Challenges," *IEEE Communications Surveys and Tutorials*, 2021.

[29] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M.-S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: advances and challenges," *IEEE Wireless Communications*, 2015.

[30] Signals Research Group, "A Global Perspective of 5GNetwork Performance," Qualcomm, Tech. Rep., 2019. [Online]. Available: https://www.qualcomm.com/media/documents/files/signals-research-group-s-5g-benchmark-study.pdf

[31] D. Xu, A. Zhou, X. Zhang, G. Wang, X. Liu, C. An, Y. Shi, L. Liu, and H. Ma, "Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption," in *Proc. of ACM SIGCOMM*, 2020.