Demystifying Resource Allocation Policies in Operational 5G mmWave Networks

Phuc Dinh*, Moinak Ghoshal,* *Student Member, IEEE,* Yunmeng Han, Yufei Feng, Dimitrios Koutsonikolas, *Senior Member, IEEE,* and Joerg Widmer, *Fellow, IEEE*

***Abstract*—Five years after the initial 5G rollout, several research works have analyzed the performance of operational 5G mmWave networks. However, 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 first systematic study, to our best knowledge, 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 observe in our experiments across different cities and operators.**

***Index Terms*—5G, access bandwidth measurements, resource allocation policies.**

I. Introduction

5G mmWave is being rapidly deployed by major mobile operators, especially in urban environments. During the past four years, several research works [1]–[11] have 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 may 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 fill this gap by conducting the first systematic study, to our best knowledge, of resource allocation policies in operational 5G mmWave networks. Through an extensive measurement campaign, 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	1050+
Amount of cellular data used (GB)	5000+
Number of cities	4
Number of operators	2
Cumulative time of measurements traces (minutes)	660+

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 base station (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 and 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 perform a systematic study of resource allocation and sharing policies in operational 5G mmWave networks across 2 major mobile operators and 4 major US cities. Our measurement suite includes tests performed in downlink and uplink traffic directions with both UDP traffic, acting as the proxy for peak network capacity, and TCP traffic, which is used by the majority of real-world applications. We performed a total of 1050+ iperf tests and used 5000+ GB worth of cellular 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 poli-

cies (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 take a detailed look into 5G mmWave resource sharing from the perspective of Carrier (or Cell) Aggregation (CA). Our results indicate that, regardless of the traffic sending rate, multiple UEs time-share the resources of multiple carriers of a BS, rather than having a single carrier allocated exclusively to one UE. Network operators try to max out the primary carrier's capacity first, but they often allocate more than one carrier, even if the primary carrier's capacity is sufficient to satisfy the traffic demand. These policies are based on a combination of per-flow sending rate and the number of flows.

• 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. 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 used rooted Google Pixel 5 phones as UEs. The Pixel 5 radio supports the 5G mmWave bands n260/261 and four component-carrier (4-CC) downlink (i.e., 4x100 MHz) and 1-CC uplink carrier aggregation. Since information regarding CA is not exposed via the Android API, we use an Accuver XCAL-Solo 5.0 device [12] for our CA study in §IV-E1, §IV-E2, and §V-A. XCAL Solo is a standalone commercial tool that is attached to a smartphone via the USB-C port and taps into the diagnostic interface of the smartphone to log all the PHY-layer KPIs. Since XCAL-Solo only works with Samsung phones, we use Samsung Galaxy S21 phones instead of Google Pixel 5 in §IV-E1, §IV-E2, and §V-A, which support 8-CC downlink and 2-CC uplink carrier aggregation.

Operators. 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. The third major US mobile operator, T-Mobile, deploys its high-speed 5G services almost exclusively in the midband and provides minimal mmWave coverage across the country [8], hence, we do not consider it in this work.

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

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

Locations. We conducted extensive measurements across 4 major cities in the US: Boston, Chicago, Indianapolis, and Atlanta. Details are shown in Table II. In each of the 4 cities, we carefully choose 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 §III, all experiments are static with the user in line-of-sight (LOS) of the BS and the UE facing the BS.

Cloud Server. All our experiments use a Google Cloud server located in Washington, DC, with 32 GB of memory and 8vCPUs, running Ubuntu 18.04. It supports up to a total of 16 Gbps for all egress flows to external IP addresses. This ensures that the bottleneck link is always the wireless link between the UE and the BS (see Fig. 1).

B. Experiments

Since we are interested in the performance in the presence of multiple flows with heterogeneous traffic demands, we use UDP for most of 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.

We design a set of systematic iperf3 tests with several UEs that allow 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. In total, we conduct 5 experiments, each repeated multiple times.

Experiment 1 involves two and three backlogged clients. *Experiment 2* involves two clients, one with backlogged traffic and another one with intermittent traffic of gradually increasing rate. *Experiment 3* involves three clients, one with backlogged traffic and the other two downloading at different fixed rates. *Experiment 4* also involves three clients, the first with backlogged traffic, the second with continuous fixed-rate traffic, and the third with intermittent traffic of gradually increasing rate. *Experiment 5*, explores resource allocation in the context of carrier aggregation as opposed to throughput. All these experiments are performed with UDP downlink traffic to provide a straightforward assessment of allocated bandwidth for each UE. We also replicate these experiments with uplink UDP traffic in §V and downlink TCP traffic in §VI. The latter is in fact the transport protocol that carries most of the application traffic in current networks.

The measurement duration varies from 20 s to 230 s for different experiments. For all experiments, we first run

a traffic 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 §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 [13] 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.

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 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 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 employs 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 Experiments 1, 2, and 3, but we could not get any successful runs for Experiment 4. For Experiments 1, 2, and 3, 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.

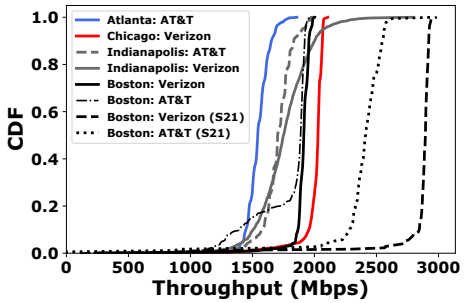


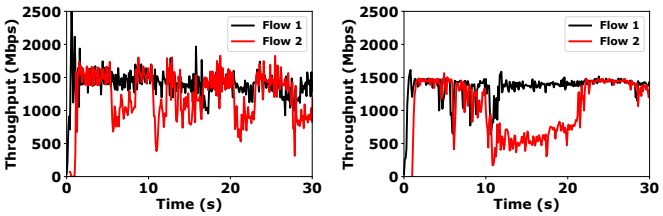
Fig. 1: UDP single flow throughput across cities and operators.

III. Resource Sharing in 5G mmWave Networks

To analyze the operators' resource allocation policies, we first have to measure as baseline performance the maximum throughput achieved by a single client. Fig. 1 shows the CDF of 100 ms downlink throughput samples of a single client for all city-operator combinations. We 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 with the Pixel 5 phone 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. The median throughput with the S21 phone is higher – 2.8 Gbps with Verizon and 2.4 Gbps with AT&T – owing to the higher carrier aggregation (8-CC vs. 4-CC). 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.

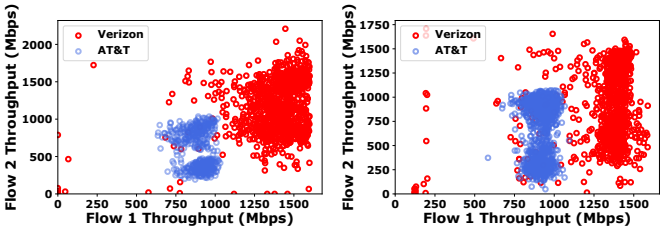
Next, we analyze whether the resource sharing in 5G mmWave networks is contention-based or controlled by the operator. We conduct two sets of experiments with Verizon and AT&T in Boston. Two users have a backlogged UDP iperf3 session each, downloading traffic from the cloud server. In the first experiment, one user faces the BS and another user faces away from the BS inflicting self-blockage. In the second experiment, one user again remains standing while facing the BS while the second user walks towards and away from the BS in a pseudo random manner.

Figs. 2a, 2b show representative timelines from the two experiments with Verizon. We find that, in both experiments, each flow is allocated a fixed bandwidth, which does not exceed 1.6 Gbps (with the exception of a few instantaneous spikes), even though their maximum achievable throughput can be much higher, as shown in Fig. 1. In particular, Flow 1 does not get more than 1.6 Gbps even when the throughput of Flow 2 drops significantly due to self-blockage or mobility. Additionally, in Fig. 3 we show scatterplots of the 100 ms throughput samples of the impaired link (Flow 2) vs. the throughput samples of the link facing the BS (Flow 1) for all the runs and both operators. For Verizon, we observe that the throughput of Flow 1 never exceeds 1.6 Gbps – about half



(a) Flow 1: LOS, Flow 2: blockage. (b) Flow 1: static, Flow 2: mobile.

Fig. 2: Representative timelines of resource sharing between two clients under fluctuating channel conditions.



(a) Flow 1 = LOS, Flow 2 = blockage. (b) Flow 1 = LOS, Flow 2 = mobile.

Fig. 3: Resource sharing between two flows under fluctuating channel conditions.

of the capacity (3 Gbps, Fig. 4) when the other link is impaired. Similarly for AT&T the throughput of Flow 1 is capped at 1 Gbps – again about half of the capacity (2 Gbps, Fig. 5).[1]

The above results suggest 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. We verified that the same behavior holds for both operators in the other cities through the controlled experiments discussed in §IV. 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.[2]

IV. Resource Allocation Policies

In this section, we describe the five experiments mentioned in §II-B and use their results to uncover the details of the resource allocation policies of the two operators in four different cities – Boston, Chicago, Atlanta, and Indianapolis. For each experiment, we describe the observed behavior with Verizon in detail, followed by a comparison

[1]In the case of blockage experiments with AT&T, the impaired link kept switching to 5G sub-6 or LTE. We removed the LTE or 5G sub-6 throughput and as a result, the number of AT&T samples is smaller than the number of Verizon samples in Fig. 3a.

[2]Since both 5G mmWave bands use TDD, downlink and uplink traffic are fully isolated and do not compete with each other for resources. We verified this by conducting experiments with simultaneous downlink and uplink flows, where we observed that the throughput of each flow is the same as in the case of one-direction traffic only. Due to space limitations, we omit the details of these experiments.

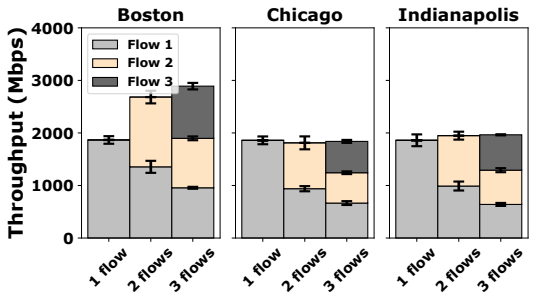


Fig. 4: Verizon, Experiment 1: Backlogged clients.

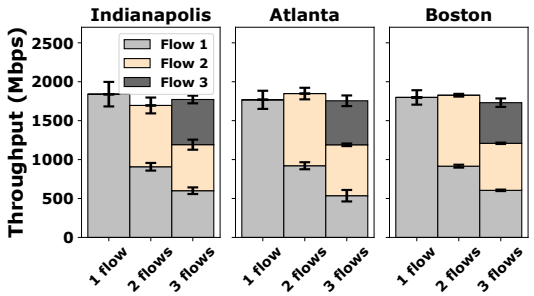


Fig. 5: AT&T, Experiment 1: Backlogged clients.

with AT&T. The bar graphs in the figures in this and the following sections show average values and the error bars denote the standard deviations.

A. Experiment 1: Backlogged clients

1) Verizon: We conduct Experiment 1 to estimate the network capacity. We introduce 3 backlogged UDP flows 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. 4 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. 1, due to the low carrier aggregation (4CC) supported by the Pixel 5. The achieved throughput with a Samsung S21 phone with 8-CC is 2.9 Gbps, equal to the network capacity (see §IV-E1). 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.*

2) AT&T: Fig. 5 shows the results for Experiment 1. Here, we observe larger standard deviations compared to Verizon (Fig. 4), due to transient channel fluctuation as discussed in Sec. III. The estimated capacity is around 1.9 Gbps for all three cities, but the average throughput (for a single flow) or sum-throughput (for multiple flows) is lower compared to the maximum capacity, as was established in Fig. 1. Similar to Verizon, *backlogged flows obtain an equal share of the capacity in both cities.*

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

1) Verizon: 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).[3] 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. 4. 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.

2) AT&T: Fig. 7a and Table IV show that the allocated capacity for Flow 2 increases monotonically as its traffic demand 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.*

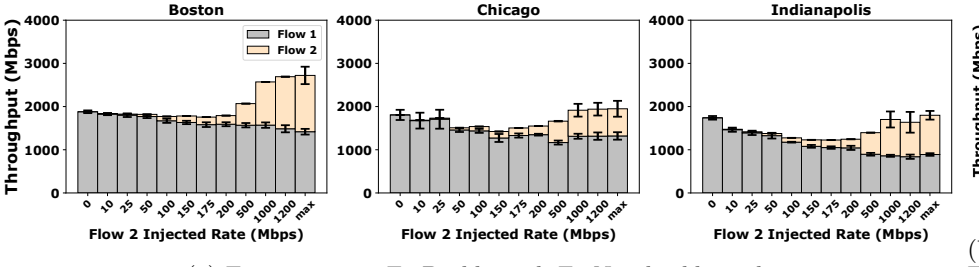
C. Experiment 3: Three clients downloading backlogged and fixed-rate non-backlogged traffic

1) Verizon: 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:

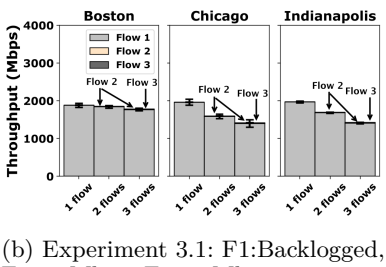
- *Experiment 3.1*: Flow 1 starts at 0 s and generates continuous backlogged traffic for 70 s, Flow 2 starts at 10 s and generates fixed-rate traffic of 10 Mbps for 50 s, and Flow 3 starts at 20 s and generates intermittent 10-s traffic bursts of 10 Mbps, with 10-s silence intervals between every two 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.
- *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.

2) AT&T: While the overall trend in Figs. 7b and 7c 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. 7b and 7c for Atlanta) and consistently for Experiment 3.2 in

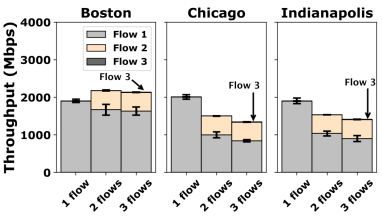
[3]Note that the values shown in Table III do not always agree with Fig. 6a. Fig. 6a shows the average values over all experiments, some of which having transient channel fluctuations as mentioned in Sec. I, which we remove from Table III.



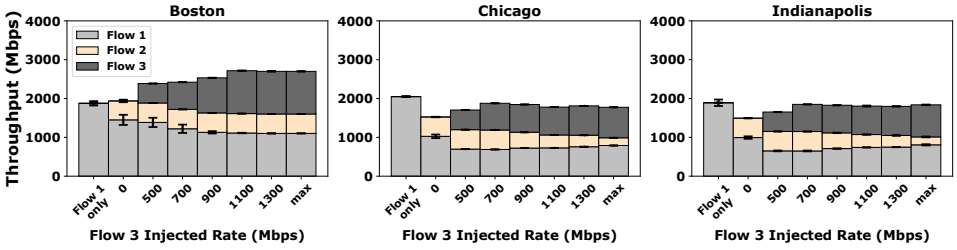
(a) Experiment 2: F1:Backlogged, F2:Non-backlogged.



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



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

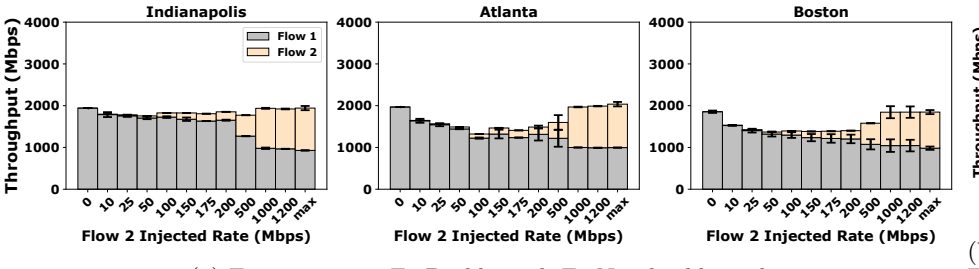


(d) Experiment 4: F1:Backlogged, F2:500 Mbps, F3:Variable.

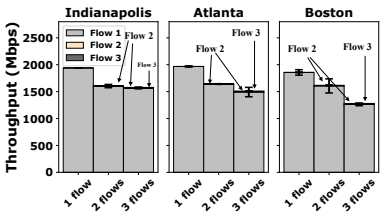
Fig. 6: Verizon, Experiments 2, 3, 4.

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

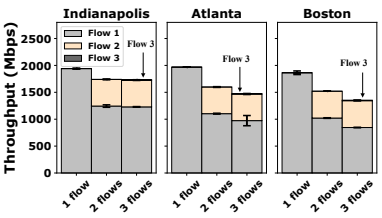
	Flow 2 Injected Rate (Mbps)										
	10	**25**	**50**	**100**	**150**	**175**	**200**	**500**	**1000**	**1200**	**max**
Boston	74	107	138	236	275	326	316	1154	1151	1235	1304
Chicago	526	559	569	593	684	735	766	841	845	844	875
Indianapolis	396	579	644	795	891	920	926	989	1110	1131	1080



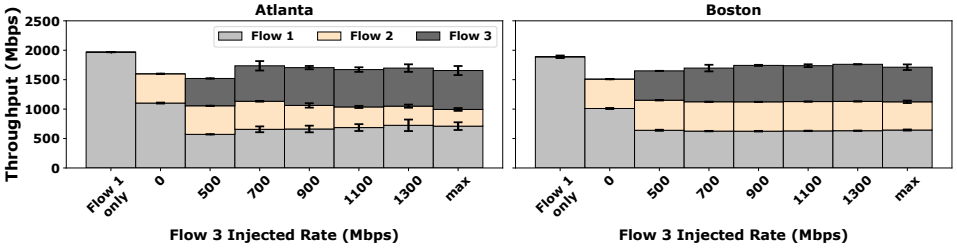
(a) Experiment 2: F1:Backlogged, F2:Non-backlogged



(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. 7: AT&T, Experiments 2, 3, 4.

Indianapolis. We call the anomaly of the network not re-allocating capacity when a new flow is introduced *failed allocation update*, and discuss it further in §VII.

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

1) Verizon: 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-

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
Boston	386	508	589	614	675	695	706	835	867	866	928

backlogged flow (Flow 2), while allocating more resources to the backlogged flow, as the demand of Flow 3 increases.

2) AT&T: Fig. 7d shows that, as we increase the sending rate of Flow 3, *the operator again employs two different policies in determining which flow(s) to penalize* in order to accommodate the increased demand of Flow 3. In Boston, similar to Verizon (Fig. 6d), it decreases the throughput of the backlogged flow only. In contrast, in Atlanta, it also decreases the throughput of the the non-backlogged flow (Flow 2), similar to the policy applied by Verizon in Chicago and Indianapolis.

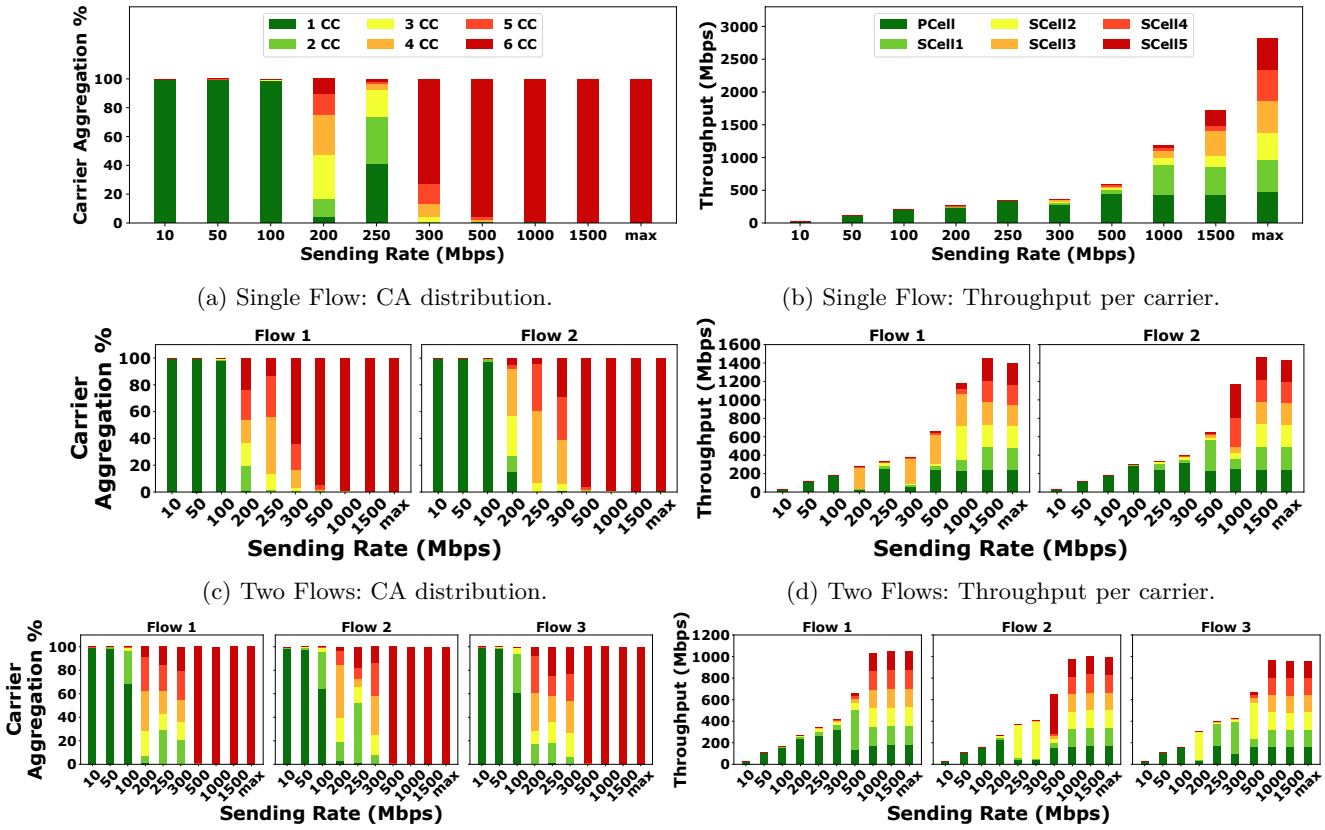
E. Experiment 5: Carrier Aggregation

While our previous experiments focused on resource sharing through the lens of application throughput, the next set of experiments delves into the low-layer operations of carrier aggregation (CA) in 5G mmWave networks. In this work, CA refers to the aggregation of different 5G n261/n260 channels, each channel having 100 MHz bandwidth. In this context, one of the 5G CCs, over which the 5G RRC signalling messages are transmitted along with user data, is marked as a Primary Cell (PCell), while other CCs (Secondary Cells or SCells) are added or removed dynamically for data transmission only. We do not consider CA with respect to 5G NSA Dual Connectivity, where the 5G connection acts as a secondary cell (SCell) by anchoring over a 4G LTE primary Cell (PCell). As noted in §I, this set of experiments is conducted with S21 phones connected to XCAL Solo devices. The experiments are conducted in Boston with 1, 2 and 3 UEs. In the experiments, all UEs receive data at a given rate for 20 seconds. The rates are gradually increased from 10 Mbps to backlogged traffic (max).

1) Verizon: Figs. 8a, 8c, and 8e show the distribution of the level of carrier aggregation (i.e., the percentage of time a certain number of carriers are used) over 5 runs for different sending rates in 1-flow, 2-flow, and 3-flow scenarios. We observe that Verizon in Boston aggregates a maximum of 6 carriers per UE – one primary carrier (PCell) and up to 5 secondary carriers (SCell1, ..., SCell5) – even though the S21 phones support 8-CC in the downlink. In Figs. 8b, 8d, and 8f, we pick an example run for each scenario and show the amount of traffic flowing through each carrier.[4] Note that, in contrast to all previous figures, in this section we plot the MAC layer throughput, as XCAL cannot provide a breakdown of the application layer throughput across different carriers. Also note that the MAC throughput is higher than the actual sending rate specified by iperf3, as it also includes retransmissions and header overhead. In Figs. 8a, 8c, and 8e and 8b, 8d, and 8f, we observe three distinct patterns based on the per-flow sending rate F.

- $F \geq 500$ *Mbps*: All 6 carriers are used nearly 100% of the time, as shown in Figs. 8a, 8c, and 8e. In case of backlogged traffic (max), traffic is distributed equally among all 6 carriers in all three scenarios (1 flow, 2 flows, and 3 flows), as shown in Figs. 8b, 8d, and 8f. In case of non-backlogged traffic, the primary carrier is always maxed out with its capacity split equally among all flows, and then, depending on the sending rate and number of flows, one or more secondary carriers are prioritized for each flow. For example, Fig. 8d shows that in the case of 2 flows, each with a sending rate of 1000 Mbps, flow 1 uses SCell2 and SCell3 in addition to the PCell to distribute most of the traffic, while flow 2 uses SCell4 and SCell5 in addition to the PCell.
- $F < 200$ *Mbps*: In the case of 1 and 2 flows, the primary carrier is used almost exclusively (Figs. 8a, 8c and 8b, 8d). The same is true in the case of 3 flows for per-flow sending rates up to 50 Mbps. However, when 3 flows download 100 Mbps each, the operator allocates a second carrier for about 35% of the time and 2 or more secondary carriers for about 5% of the time (Fig. 8e), although the amount of traffic over the secondary carriers is negligible (Fig. 8f). Note that this additional allocation of secondary carriers is unnecessary; a single carrier can easily accommodate the total traffic of 300 Mbps, since the per-carrier capacity is about 500 Mbps, as can be inferred from the case of a single backlogged flow (Fig. 8a).
- 200 *Mbps* $\leq F <$ 500 *Mbps*: This is the most interesting case, where the number of carriers used varies over time, as shown in Figs. 8a, 8c, and 8e. In scenarios with 1 flow, we found that the PCell is always used and carries most of the traffic; other carriers are occasionally added but they carry a negligible amount of traffic (Fig. 8b). Fig. 9a shows an example timeline with a sending rate of 200 Mbps. Again here, the operator *over-allocates* carriers, although the traffic can be accommodated by the primary carrier alone. In 2-flow and 3-flow scenarios, the PCell is again used 100% of the time, and other carriers are occasionally added and removed, similar to the single-flow scenario. However, the amount of traffic sent over secondary carriers varies for different flows. Figs. 9b, 9c show two example timelines in the 2-flow scenario, with a per-flow sending rate of 200 Mbps. In the first example (Fig. 9b), Flow 1 uses the PCell and SCell1 during the entire run to deliver the traffic, whereas Flow 2 uses the PCell almost exclusively. On the other hand, in the second example (Fig. 9c), Flow 1 almost exclusively uses SCell3 all the time, and Flow 2 uses the Pcell. Similar behavior is observed in 3-flow scenarios; Figs. 9d, 9e show two example timelines, again with a per-flow sending rate of 200 Mbps. Fig. 9 also shows

[4]While the behavior across runs is consistent, which specific SCells are used for a given flow changes from experiment to experiment, making it difficult to show aggregate results.



(a) Single Flow: CA distribution.

(b) Single Flow: Throughput per carrier.

(c) Two Flows: CA distribution.

(d) Two Flows: Throughput per carrier.

(e) Three Flows: CA distribution.

(f) Three Flows: Throughput per carrier.

Fig. 8: Verizon: Carrier Aggregation distribution over 5 runs (8a, 8c, 8e) and throughput per carrier in a single run (8b, 8d, 8f) in different scenarios.

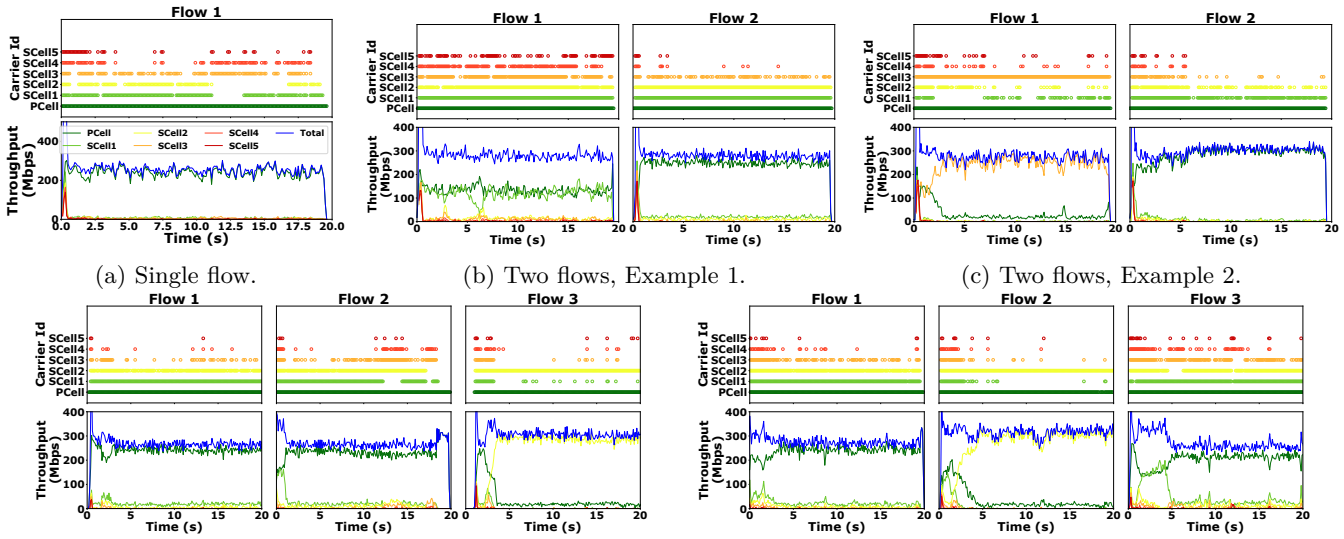
that the traffic distribution across carriers can change over the course of a run and there is typically a transitory behavior in the first 2-5 seconds. For example, in Fig. 9e, the traffic for Flow 3 is distributed equally across the PCell and SCell1 during the first 5 s, but is mostly sent over the PCell for the remaining 15 s.

From the above observations, we make two conclusions: (i) *Verizon often over-allocates carriers even though the traffic demand can be satisfied with a single carrier, but prefers to send most of the traffic over the PCell whenever possible.* (ii) *The carrier allocation appears to be based on a combination of per-flow sending rate and the sum rate of all flows.* In 1-flow and 2-flow scenarios, the allocation is based on the per-flow sending rate rather than the total traffic demand. For example, in 1- and 2-flow scenarios with 200 Mbps per flow, multiple carriers are allocated to each flow (Figs. 8a, 8c), whereas in 2-flow scenarios with 100 Mbps per flow (200 Mbps total demand) only one carrier is allocated (Fig. 8c), the same as in 1-flow scenarios with 100 Mbps (Fig. 8a). On the other hand, in 3-flow scenarios, the total traffic increases to a point where it starts to also be taken into account. For example, with a per-flow sending rate of 100 Mbps, more than one carrier is allocated to each flow, as shown in Fig. 8e.

2) AT&T: Fig. 10 shows that AT&T uses up to 8 carriers in Boston unlike Verizon, which only uses up to 6 carriers. Intriguingly, its total network capacity is not proportionally higher than Verizon's (3.4 Gbps compared to 3 Gbps) (see Figs. 10b, 10d, 10f vs. Figs. 8b, 8d, 8f in the case of backlogged traffic and also Fig. 1). Since both operators use carriers with the same bandwidth of 100 MHz, this observation suggests a lower spectrum efficiency for AT&T, characterized by a reduced throughput per carrier (425 compared to 500 Mbps). Additionally, for the case of backlogged traffic, Figs. 10b, 10d, 10f show a behavior similar to what we observed for Verizon in Boston with the Pixel 5 phone (Fig. 4); a single flow can get a MAC layer throughput of only 2.5 Gbps (lower than Verizon's single flow throughput of 3 Gbps), but in the case of 2 or 3 flows, the sum throughput rises to 3.4 Gbps.

However, both operators share similar trends with respect to CA policies. Fig. 10 shows that AT&T exhibits the same three patterns based on the per-flow sending rate F, as those seen for Verizon in Fig. 8: a) $F \geq 500$ $Mbps$; b) $F < 200$ $Mbps$; and c) 200 $Mbps \leq F < 500$ $Mbps$. We make the following observations:

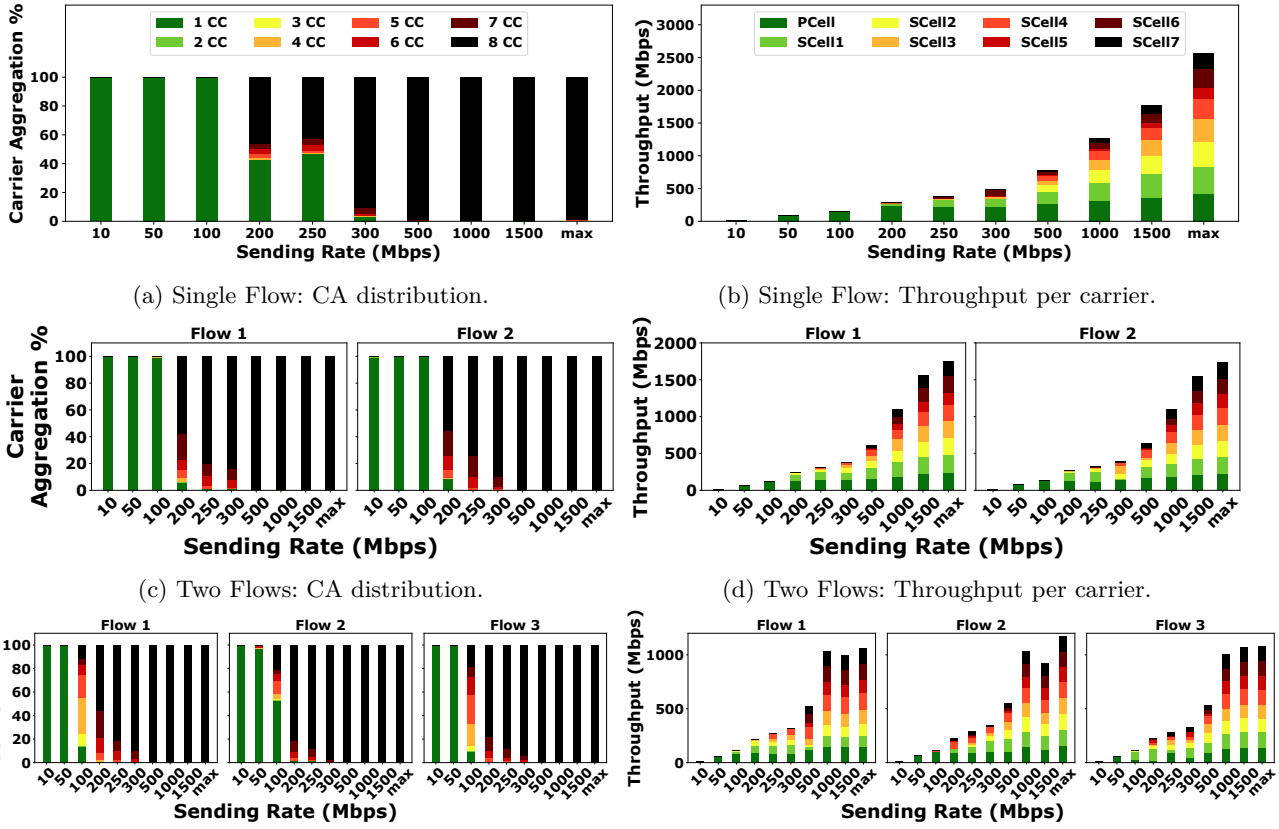
- $F \geq 500$ $Mbps$**:** Similar to Verizon, all 8 carriers are always activated. However, the traffic is distributed among the 8 carriers in a more balanced way compared to Verizon that distributes most of the traffic over the PCell and one or two SCells in the case of non-backlogged traffic (compare Figs. 8b, 8d, and 8f against Figs. 10b, 10d, 10f for a sending rate of 500 Mbps, 1000 Mbps, and 1500 Mbps).
- $F < 200$ $Mbps$**:** Similar to Verizon, AT&T allocates only the primary carrier in all cases, except for the 3-flow scenario with a per-flow sending rate of 100 Mbps (Fig. 10e



(a) Single flow. (b) Two flows, Example 1. (c) Two flows, Example 2.

(d) Three flows, Example 1. (e) Three flows, Example 2.

Fig. 9: Verizon: Timelines showing the carriers used and the corresponding per-carrier throughput in 1-flow, 2-flow, and 3-flow scenarios with a sending rate of 200 Mbps.



(a) Single Flow: CA distribution. (b) Single Flow: Throughput per carrier.

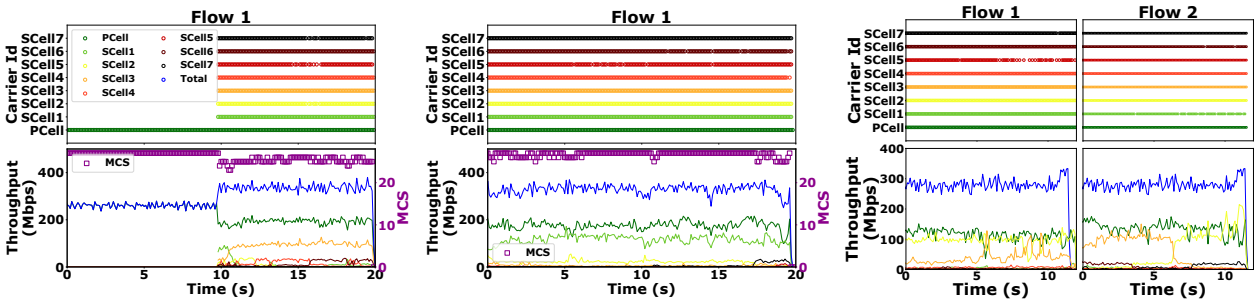
(c) Two Flows: CA distribution. (d) Two Flows: Throughput per carrier.

(e) Three Flows: CA distribution. (f) Three Flows: Throughput per carrier.

Fig. 10: AT&T: Carrier Aggregation distribution and throughput per carrier in different scenarios.

and 10f), where it performs over-allocation; up to 8 carriers are used but the amount of traffic over most secondary carriers is negligible.

• 200 $Mbps \leq F < 500$ $Mbps$**:** Similar to Verizon, the number of carriers used varies over time (Figs. 10a, 10c, and 10e). However, we observe a higher degree of over-allocation compared to Verizon, where all 8 carriers are used for a significant percentage of time. For example, with a per-flow sending rate of 200 Mbps, 8 carriers are used 50% of the time in 1-flow scenarios and this percentage increases when more flows are added, while in the same case with Verizon, all 6 carriers are only used for 10-20% of the time. Additionally, while the PCell is again always used similarly to Verizon, secondary carriers with AT&T carry a non-negligible amount of traffic even in 1-flow scenarios and traffic is distributed over multiple carriers.



(a) Single flow, Example 1.

(b) Single flow, Example 2.

(c) Two flows.

(d) Three flows.

Fig. 11: AT&T: Timelines showing the carriers used and the corresponding per-cell throughput in 1-flow, 2-flow, and 3-flow scenarios with a sending rate of 200 Mbps.

Fig. 11 shows example timelines for a per-flow sending rate of 200 Mbps. For the single flow scenario, in contrast to Verizon, we found that each run can be very different in terms of carrier allocation. For example, compare Fig. 11a and Fig. 11b. In the first example (Fig. 11a), the operator activates only a single carrier (PCell) for the first 10 seconds but, after the 10th second, all 8 carriers are activated and a significant amount of the traffic flows through SCell3. By looking at the XCAL traces, we found that the PCell MCS dropped from 28 (highest) and kept fluctuating till the end of the trace, forcing the network to activate other carriers. In the second example (Fig. 11b), the PCell MCS keeps fluctuating from the beginning of the trace and the operator activates all 8 carriers right away. Figs. 11c, 11d show examples in 2-flow and 3-flow scenarios where AT&T activates all 8 carriers most of the time and traffic is distributed over multiple carriers, in contrast to Verizon, which sends most of the traffic over 1 or 2 carriers (Figs. 9b-9e).

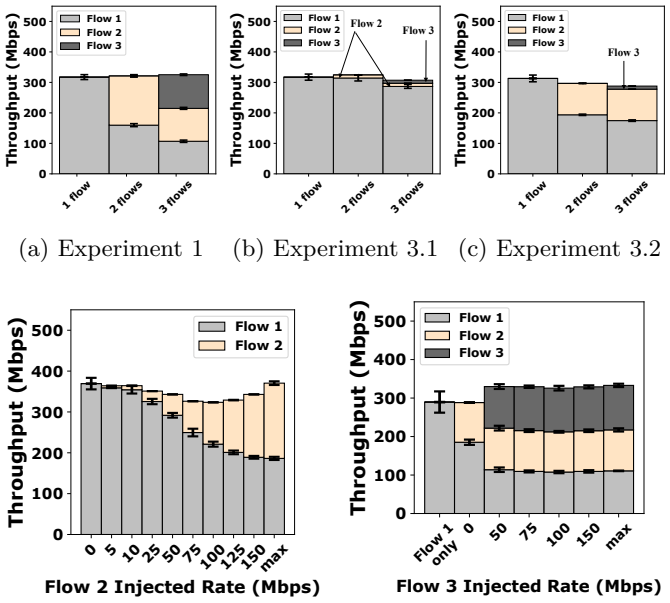
Overall, we conclude that *similar to Verizon, AT&T often over-allocates carriers even though the traffic demand can be satisfied with a single carrier* and *uses a similar policy based on a combination of per-flow sending rate and the sum rate of all flows*. However, unlike Verizon, which tries to distribute the traffic over a small number of carriers even if it activates all of them, *AT&T distributes traffic in a more balanced way over multiple carriers.*

V. Uplink experiments

In §IV, our experiments focused exclusively on the downlink direction. In this section, we replicate all experiments from §IV, but for uplink scenarios. Notably, for experiments involving non-backlogged flows, we reduce the injected rate compared to the downlink experiments, as the uplink bandwidth is significantly lower.

A. Verizon uplink resource allocation policies

Fig. 12 showcases the resource allocation policies for uplink traffic. A first observation is that, unlike in the downlink experiments, the total bandwidth in uplink experiments fluctuates notably across different tests, despite these experiments being conducted at times with no other UE devices in proximity. For instance, the total uplink bandwidth is approximately 310 Mbps in Experiments 1, 3, and 4, but it rises to around 360 Mbps in Experiment 2. Another distinction from the downlink tests in Boston is the absence of a threshold-based constraint on the total bandwidth; the network uplink bandwidth remains constant with growing total traffic demand. This trend is evident in Fig. 12a, where the bandwidth is consistently around 310 Mbps, regardless of whether 1, 2, or 3 UEs are uploading backlogged traffic. This is in contrast to Fig. 4 for the downlink, where bandwidth jumps from about 1.8 Gbps to 2.8 Gbps. However, certain parallels between uplink and downlink allocation policies are observable. For instance, *the network typically grants a share of the capacity that surpasses the actual transmission rate for non-backlogged flows, as illustrated in Experiments 2, 3.1, and 3.2 (indicative of over-allocation policy).* Another shared policy between uplink and downlink in Boston is *the prioritization of non-backlogged flows over backlogged ones.* As shown in Fig. 12e, when Flow 3 (non-backlogged) increases its transmission rate, the bandwidth designated



(a) Experiment 1 (b) Experiment 3.1 (c) Experiment 3.2

(d) Experiment 2 (e) Experiment 4

Fig. 12: Replications of experiments 1-4 for uplink

for Flow 1 (backlogged) reduces, while that for Flow 2 (non-backlogged) remains stable.

Next, we study resource allocation for uplink traffic in the context of carrier aggregation, which can be observed in Fig.13. First, we notice that as opposed to downlink with a total of 6 carriers, only 2 carriers are used for uplink. Since the total capacity for uplink is around 300 Mbps, the per-carrier capacity for uplink is around 150 Mbps. Similar to downlink, the primary carrier is used exclusively when the per-flow sending rate is small (<25 Mbps); for higher sending rates (25-75 Mbps per flow), the number of carriers used by each UE varies between 1 and 2 (over-allocation); and for sending rates above 100 Mbps per flow, both carriers are almost always used for both UEs. Another similarity to downlink is that, as the total traffic approaches the capacity, each UE shares each carrier's capacity roughly equally.

B. AT&T uplink resource allocation policies

We also attempted to replicate our previous experiments for AT&T in Boston. However, the uplink performance of AT&T's 5G mmWave proved to be very unstable and did not display a consistent trend. Consequently, it was not possible to derive conclusive observations from these experiments. Fig. 14 illustrates examples of inconsistent sharing behaviors, while trying to replicate Experiment 1 (3 backlogged flows). Specifically, in one of the runs (left), Flow 2 achieves a throughput significantly higher (around 180 Mbps) than its expected allocated bandwidth of approximately 120 Mbps (with a total capacity of around 360 Mbps), while Flow 3 achieves only around 40 Mbps. In another run (right), Flow 3's throughput is around 160 Mbps, while the other two flows have throughput close to 100 Mbps each. Moreover, in each run, the throughput for each flow varies wildly, highlighting the network's instability.

VI. TCP Experiments

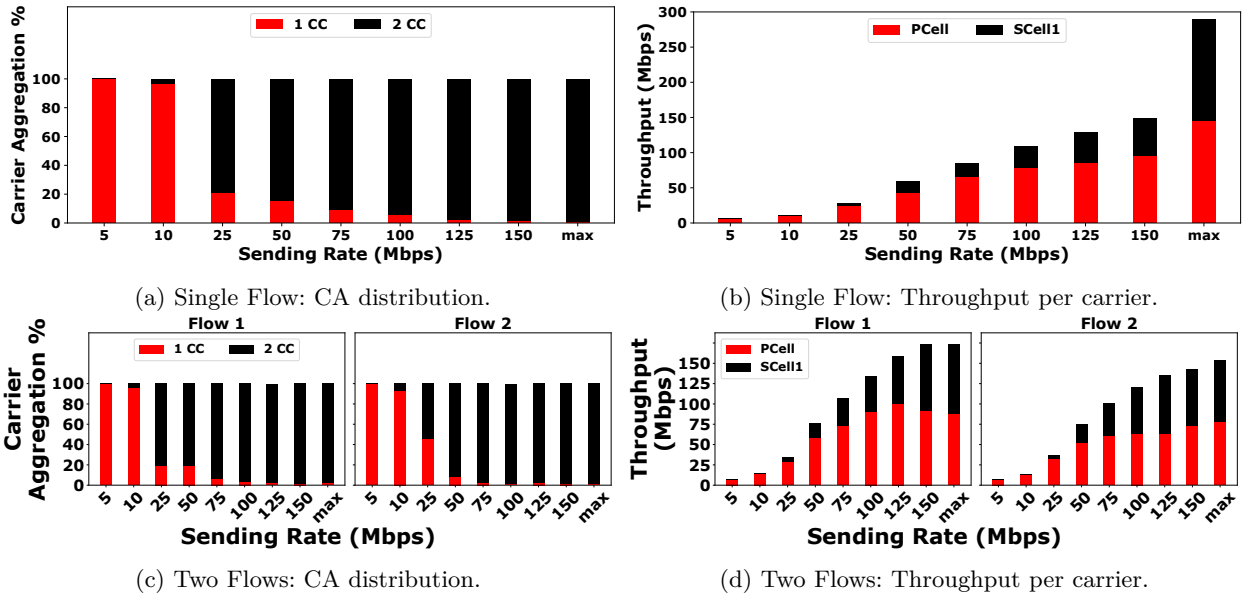
A. TCP's adaptive behavior and its potential impact on resource sharing

In all our previous experiments, we exclusively used UDP traffic. Our rationale is that UDP offers a more straightforward assessment of allocated bandwidth, since it transmits data at a constant rate determined by the sender without considering network feedback or receiver state. In contrast, TCP uses congestion control and flow control mechanisms that adapt the sending rate based on network conditions and receiver capabilities. At the same time, however, the vast majority of real-world applications rely on TCP as their transport protocol for data transmission. Given its prevalence, it is imperative to understand how TCP behaves under the same conditions that we tested with UDP. With this in mind, we replicate experiments 1-4 with TCP for the case of Boston using Verizon's infrastructure. We use TCP Cubic in all of our experiments as it is the default congestion control algorithm of most Linux distributions.

As we established in §IV, the bandwidth allocated to a UE is a function of the server's sending rates to the UE. This implies that any TCP mechanism affecting sending rates such as slow start and congestion avoidance will invariably influence resource sharing and the resulting TCP throughput. To illustrate this point, we analyze three real-time throughput examples from Experiment 2's replication with TCP. Despite using iperf3 to control the maximum sending rates for TCP flows, the actual rates vary in response to TCP's adaptive mechanisms. Fig. 15 shows substantial variability in TCP throughput from one replication to another. Specifically, in the rightmost example of Fig. 15, the backlogged Flow 1 did not reach its per-flow maximum allocated capacity of about 2 Gbps within the first 150 seconds, and its throughput significantly reduced even with minimal traffic on Flow 2. In the other instances, although the throughput of Flow 1 reduces proportionally to the sending rate of Flow 2, resembling the behavior of UDP, the two flows occasionally exhibit sharp drops in throughput to almost zero. This variability and the sometimes erratic TCP throughput patterns, with transient sharp drops, are indicative of the complex interplay between TCP's reaction to network conditions and the operator's bandwidth allocation decisions.

B. Throughput comparisons with UDP

Fig. 16 shows the average TCP throughput values for Experiments 1-4. We note that, in each experiment, TCP flows undergo varying slow start periods, often lasting several seconds, before reaching a steady state. Consequently, these periods have been omitted from our average and standard deviation calculations for a more accurate comparison with UDP throughput. Our first observation is that TCP throughput exhibits greater variability than UDP in most experiments, a result that is to be expected given TCP's reactive mechanisms in response to packet loss. Our second observation is that the behaviors we



(a) Single Flow: CA distribution.

(b) Single Flow: Throughput per carrier.

(c) Two Flows: CA distribution.

(d) Two Flows: Throughput per carrier.

Fig. 13: Uplink carrier aggregation distribution and throughput per carrier in different scenarios.

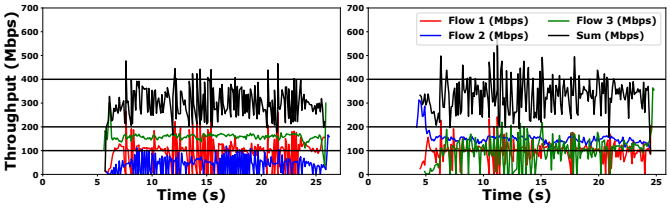


Fig. 14: AT&T failed uplink sharing timelines.

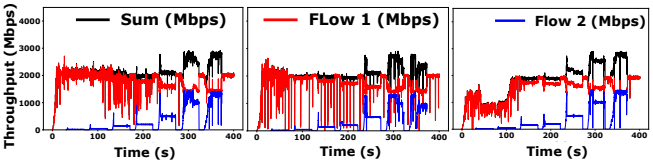
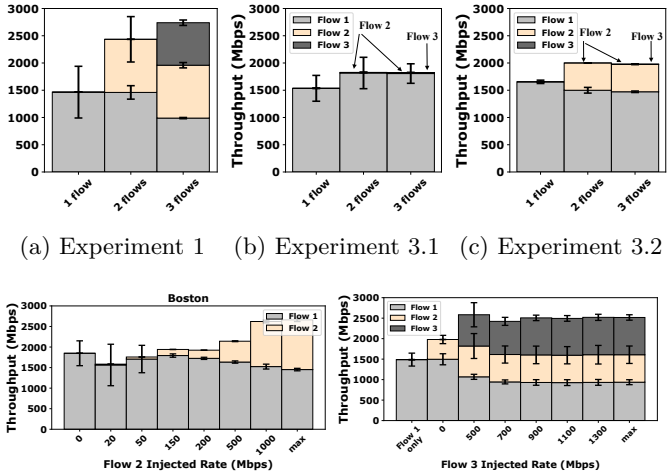


Fig. 15: Examples of different TCP throughput patterns in real time (Verizon: Experiment 2).



(a) Experiment 1 (b) Experiment 3.1 (c) Experiment 3.2

(d) Experiment 2 (e) Experiment 4

Fig. 16: Verizon: Replications of experiments 1-4 for TCP.

observed in the UDP experiments do not consistently apply to TCP. For example, in Experiment 1, two or three saturated TCP flows do not always obtain equal throughput. Additionally, in Experiment 2 and Experiment 3.1, we do not observe a consistent decrease in throughput for the existing flow (Flow 1); instead, we see irregular decreasing patterns (in Experiment 2) or even an increase (in Experiment 3.1). This discrepancy arises because TCP sending rates are inherently variable, leading to fluctuations in allocated bandwidth over time. Interestingly, in Experiments 3.2 and 4, TCP's performance closely mirrors that of UDP, likely due to the negligible loss events encountered during these tests.

VII. Anomalous behaviors

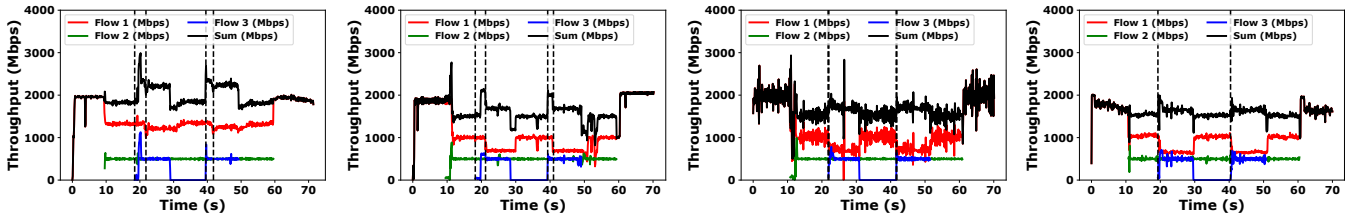
We highlight three types of anomalous behavior, which does not match the identified sharing policies.

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. 17 shows representative timelines of Experiment 3.2, described in §IV-C1 and §IV-C2, for selected operator-city combinations. In Figs. 17a and Fig. 17b, we observe that when Flow 3 is introduced, it takes 1-2 s for Flow 1's throughput to drop. In contrast, Fig. 17c and Fig. 17d 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. 18a shows an example timeline of Experiment 3.1, described in §IV-C1, 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



(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. 17: Measurement timelines for experiments with and without delayed allocation updates.

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. 18b 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. 18c vs. Fig. 18d).

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. 18e and 18f show an example for Experiment 2, described in §IV-B1. In Figs. 18e, the two flows share resources as expected, whereas in Fig. 18f, 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. 19a, Flow 3 does not start at 20 s, when Flow 1 and Flow 2 are already active. In Fig. 19b, 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. In addition, we also discuss how the general trends concluded from UDP downlink experiments translate to uplink and TCP cases.

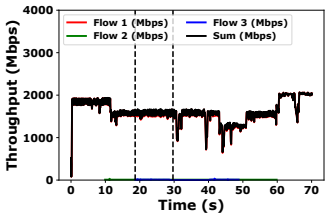
1) ***Network capacity****:* With the exception of Verizon in Boston, our measurements for both operators in other cities show the network capacity to be below 2 Gbps. However, Verizon in Boston imposes a threshold-based limitation on the total capacity of 2.9 Gbps. As discussed in §IV, this limitation 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 day, dropping by up to 25%, as we mentioned in §II-B. We conjecture that 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.

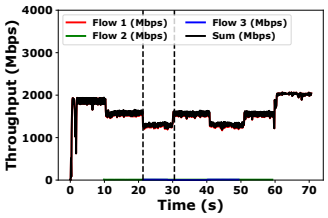
3) ***Over-provisioned capacity for non-backlogged flows****:* Another common trend is that operators allocate more capacity than required by the actual traffic demand of small (mice) flows, presumably as a safety margin. However, this policy also reduces the performance of the existing large (elephant) 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. 7d 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 the new flow, and the (smaller) rate of the non-backlogged flow remains unchanged.

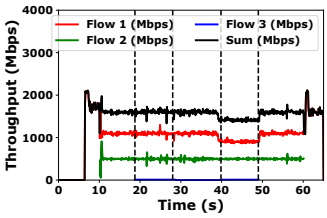
5) Carrier aggregation: A common trend across both operators is that multiple UEs time-share the resources of multiple carriers of a BS, rather than having a single carrier allocated exclusively to one UE. Another common trend is that both operators often over-allocate more than one carrier, even if the PCell's capacity is sufficient to



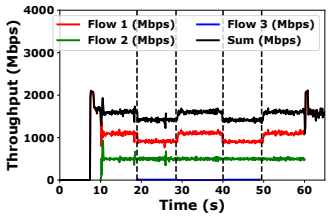
(a) Verizon, Chicago: Failed allocation update.



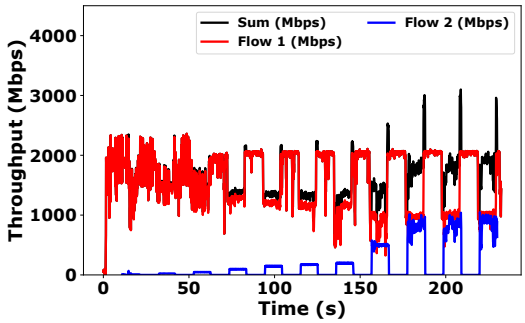
(b) Verizon, Chicago: Successful allocation update.



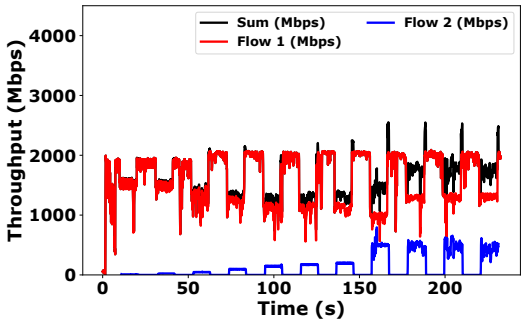
(c) AT&T, Atlanta: Failed allocation update.



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

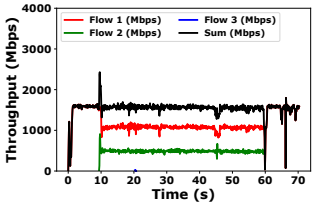


(e) Verizon, Chicago: Successful allocation update.

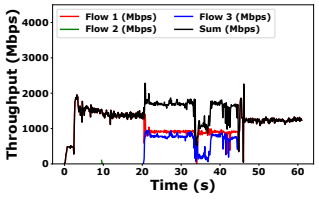


(f) Verizon, Chicago: Failed allocation update.

Fig. 18: 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. 19: Measurement timelines for experiment with flow startup failure.

satisfy the traffic demand. Finally, both operators allocate carriers to flows based on a combination of per-flow sending rate and the sum rate of all flows. On the other hand, while both operators try to max out the primary cell's capacity first, they take a different approach to allocating traffic over the secondary carriers. Verizon often limits the load to one or two carriers in addition to the primary carrier, while sending a negligible amount of traffic over the remaining secondary carriers. In contrast, AT&T often tries to balance the carriers' load by distributing the traffic over multiple carriers.

6) ***Resource allocation in uplink****:* Uplink resource allocation is also policy-based. Despite the much lower capacity, similar resource allocation policies are observed as in the downlink case, such as over-allocation policies (in terms of assigned capacity to flows with low traffic demands and in terms of assigned carriers), prioritization of non-backlogged flows over backlogged ones (in Boston), carrier allocation policies based on the per-flow sending rate.

7) ***Impact of TCP dynamics****:* TCP throughput is generally more variable than UDP due to the protocol's adaptive mechanisms in response to network conditions and established conclusions based on UDP behavior do not consistently apply to TCP scenarios. This variability is particularly evident when comparing throughput across different experiments, where TCP does not always follow the same patterns of change as those observed with UDP. As a result, some of our observations with UDP traffic do not always hold for TCP; for example, multiple saturated flows do not always share the network capacity equally and the throughput of a saturated TCP flow does not always decrease consistently as new flows are added.

8) ***Implications of the findings for operators and applications****:* Our findings have important implications for cellular operators and applications over 5G mmWave networks. For operators, a key consideration is the trade-off between the overprovisioning safety margins and responsiveness to real-time traffic demand. Although larger safety margins help ensure robust performance and responsiveness for applications that require non-backlogged traffic, they can lead to inefficient channel utilization, as redundant resources are allocated for less bandwidth-demanding applications. To avoid this, operators can set tighter safety margins to more closely match actual traffic demands, reducing the redundant allocation. However, this comes with the challenge of maintaining sufficient responsiveness for sudden traffic bursts or variations in user demand. Tightening safety margins benefits mostly applications with predictable or stable traffic patterns, while those with more dynamic requirements experience reduced QoE if resources cannot be quickly reallocated.

9) ***Potential improvements and optimization****:* With the current policies, one straightforward optimization is to balance the aforementioned trade-off between responsiveness and resource utilization by setting an optimal safety margin based on the traffic patterns of specific applications. Additionally, operators might also explore the adoption of proportional fairness in resource allocation,

which has already been widely used for LTE. Proportional fairness can dynamically adjust resources based on real-time changes in demand and channel conditions, which can lead to better bandwidth utilization and overall fairness. However, in mmWave networks, where channel conditions fluctuate significantly, such a policy may over-react to short-term variations, leading to unnecessary reallocations and instability. Moreover, proportional fairness introduces additional complexity in the implementation, especially compared to static resource splitting, which offers a simpler and more stable approach. Operators should carefully weigh these trade-offs, considering both the complexity of implementation and the need for efficient and fair resource distribution in their networks.

10) ***Resource allocation in LTE and 5G low/mid bands:*** As we mention in §IX, most research on resource allocation in cellular networks exclusively focuses on mathematical modeling and optimization frameworks with simulation-based evaluation, while the actual allocation policies/algorithms used by operators are often proprietary. Currently, we are not aware of any similar measurement studies of resource allocation for LTE and 5G low/mid bands in operational cellular networks to draw fair and direct comparisons, although it is known among the research and cellular practitioner community that the most popular scheduling mechanisms for LTE include round-robin, maximum-rate, and proportional fairness scheduling [14]. Due to space limitations, we focus on 5G mmWave in this work and leave a thorough investigation of resource allocation in other bands as future work.

IX. Related Work

5G NR Resource Allocation. Resource allocation in 5G networks has been extensively studied [15]–[31]. 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 [32]–[37] 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 [38] as one of the first reports on 5G performance profiling with the main focus on physical layer performance and coverage. The works in [1]–[8] conducted measurement studies of 5G mmWave networks exploring performance, coverage, beamforming, energy consumption, and the impact on application QoE and the works in [9]–[11], [39]–[42] conducted similar studies for sub-6 GHz 5G in the US, Europe, and China. Finally, the work in [43] presented a study of 5G in the UK (sub-6 GHz) from an operator's perspective. 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 of resource allocation and sharing policies in operational 5G mmWave networks. Our study comprises 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 resource allocation 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.

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] A. Narayanan, M. I. Rochman, A. Hassan, B. S. Firmansyah, V. Sathya, M. Ghosh, F. Qian, and Z.-L. Zhang, "A Comparative Measurement Study of Commercial 5G mmWave Deployments," in *In Proc. of IEEE INFOCOM*.

[6] M. Ghoshal, Z. J. Kong, Q. Xu, Z. Lu, S. Aggarwal, I. Khan, Y. Li, Y. C. Hu, and D. Koutsonikolas, "An In-Depth Study of Uplink Performance of 5G MmWave Networks," in *Proc. of the ACM SIGCOMM 5G-MeMU Workshop*, 2022.

[7] Y. Feng, J. Wei, P. Dinh, M. Ghoshal, and D. Koutsonikolas, "Beam Management in Operational 5G mmWave Networks," in *Proc. of ACm mmNets*, 2023.

[8] M. Ghoshal, I. Khan, Z. J. Kong, P. Dinh, J. Meng, Y. C. Hu, and D. Koutsonikolas, "Performance of Cellular Networks on the Wheels ," in *Proc. of ACM IMC*, 2023.

[9] R. Fezeu, C. Fiandrino, E. Ramadan, J. Carpenter, L. Freitas, F. Bilal, W. Ye, J. Widmer, F. Qian, and Z.-L. Zhang, "Unveiling the 5G Mid-Band Landscape: From Network Deployment to Performance and Application QoE," in *Proc. of ACM SIGCOMM*, 2024.

[10] W. Ye, X. Hu, S. Sleder, A. Zhang, U. Dayalan, A. Hassan, R. Fezeu, A. Jajoo, yungjin Lee, E. Ramadan, F. Qian, and Z.-L. Zhang, "Dissecting Carrier Aggregation in 5G Networks: Measurement, QoE Implications and Prediction," in *Proc. of ACM SIGCOMM*, 2024.

[11] I. Khan, M. Ghoshal, J. Angjo, S. Dimce, M. Hussain, P. Parastar, Y. Yu, C. Fiandrino, C. Orfanidis, S. Aggarwal, A. C. Aguiar, O. Alay, C. F. Chiasserini, F. Dressler, Y. C. Hu, S. Y. Ko, D. Koutsonikolas, and J. Widmer, "How Mature is 5G Deployment? A Cross-Sectional, Year-Long Study of 5G Uplink Performances," in *Proceedings of the IFIP/IEEE Networking Conference*, 2024.

[12] "XCAL Solo," https://accuver.com/sub/products/view.php?idx=1&s=4.

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

[14] S. Sesia, I. Toufik, and M. Baker, "Multi-User Scheduling and Interference Coordination," in *LTE – The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. Chichester, West Sussex, UK: Wiley, 2011, ch. 12.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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

[39] 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.

[40] K. Kousias, M. Rajiullah, G. Caso, O. Alay, A. Brunstorm, L. D. Nardis, M. N. U. Ali, and M.-G. D. Benedetto, "Implications of Handover Events in commercial 5G Non-Standalone Deployments in Rome," in *Proc. of ACM 5G-MeMU*, 2022.

[41] ——, "Coverage and Performance Analysis of 5G Non-Standalone Deployments," in *Proc. of ACM WiNTECH 2022*, 2022.

[42] R. Fezeu, C. Fiandrino, E. Ramadan, J. Carpenter, D. Chen, Y. Tan, F. Qian, J. Widmer, and Z.-L. Zhang, "Roaming across the European Union in the 5G Era: Performance, Challenges, and Opportunities," in *Proc. of IEEE INFOCOM*, 2024.

[43] P. Parastar, A. L. O. O. Alay, G. Caso, and D. Perino, "Spotlight on 5G: Performance, Device Evolution and Challenges from a Mobile Operator Perspective," in *Proc. of IEEE INFOCOM*, 2023.