MARKlab: Distributed Measurements in Public Mobile Networks at International Scale

Viktoria Vomhoff*, Stefan Geissler*, Steffen Gebert‡, Tobias Hossfeld*
*Chair of Communication Networks, University of Würzburg, Germany
Email: {firstname.lastname}@uni-wuerzburg.de
‡emnify GmbH, Germany
Email: {firstname.lastname}@emnify.com

***Abstract*—As demand for reliable mobile connectivity in IoT has surged, assessing expected real-world QoS has become crucial. However, due to their complexity, large-scale mobile network performance studies, especially including outbound roamers, remain scarce. To this end, we introduce MARKlab, a scalable framework for measuring mobile network KPIs such as latency, packet loss, throughput, and energy consumption in production environments. Conducting case study across four European countries, 12 operators and five radio access technologies – 2G, 3G, LTE Cat-1, LTE-M, and NB-IoT – we analyze each parameter's impact on performance. Our results highlight significant differences in latency, reliability, and energy efficiency, particularly for low-power IoT. The acquired measurement data is made publicly available to both researchers and industry.**

***Index Terms*—Internet of Things, mobile network, LTE-M, NB-IoT, 4G, 2G, 3G, international roaming, measurement platform**

I. Introduction

In today's hyperconnected world, mobile networks have evolved into vital infrastructure for societal, cultural, and economic development. They serve as the backbone of modern communication, enabling a wide range of verticals from emergency services and remote healthcare, to smart cities and global commerce. As society increasingly relies on these technologies, the global infrastructure supporting them grows more complex. Ensuring their robust performance and efficiency is a technical necessity. It is also crucial for understanding the shortcomings and limitations of the current mobile landscape.

In the past decade, mobile connectivity has transformed how individuals interact, businesses operate, and governments deliver services. At the same time, mobile network standards have evolved—from 2G and 3G to LTE, 5G, and specialized technologies like LTE for Machines (LTE-M) and Narrowband IoT (NB-IoT). This ongoing adaptation reflects the dynamic landscape where technology evolves to meet the growing demands of our hyperconnected society.

However, the complexity introduced of multiple mobile generations, and their respective Radio Access Technologies (RATs), is only part of the challenge in assessing the performance and efficiency of modern mobile networks. Instead, key factors, such as international roaming [1], protocol selection [2], and monitoring data availability [3], are crucial for a holistic view. This is especially relevant for applications in the Internet of Things (IoT). While smartphones generally benefit from connecting to the mobile network using the most modern technology available, IoT devices exhibit significantly different usage patterns. Instead of mainly optimizing for latency and throughput, energy efficiency, transmission distance, and general coverage is more relevant for IoT applications.

While previous and recent studies have focused on these aspects [4]–[7], many studies are conducted under non-realistic conditions, or are not broad enough to establish a holistic understanding of the wide range of influencing factors when it comes to the performance of mobile networks. In addition, many works suffer from the limited visibility of KPIs within production mobile networks, which has been already discussed in the past [3]. To address these limitations, researchers need appropriate tools and methodologies in order to conduct large scale, reproducible and reliable measurement studies in real-world production networks under real-world conditions.

To this end, we are developing and operating MARKlab, a distributed, multi-RAT mobile network measurement platform that enables researchers to conduct experiments at scale in the real world with a focus on flexibility, reliability and reproducibility. MARK stands for **M**ulti-RAT **A**utonomous Measu**R**ement Framewor**K**. MARKlab consists of autonomous nodes, called MARKs, that attach to a central management platform, collect scheduled measurement jobs, execute them and return the gathered data. At the time of writing, MARKlab is hosting eight nodes across four European countries, with more MARKs currently being deployed in additional countries in Europe as well as North and South America.

In this work, we introduce the technical details on the MARKlab platform as well as our measurement nodes, highlight the capabilities of the framework, and present a case study showcasing the application of MARKlab when conducting research on mobile networks at an international scale. Specifically, we make the following contributions in this work.

- We introduce the MARKlab framework to conduct mobile network measurements at scale in production environments under real-world conditions.
- We conduct a case study evaluating the QoS across multiple countries, mobile network operators and radio technologies for outbound roamers.
- We provide the raw measurement data obtained during the performed case study discussed in this work.[1]

978-3-903176-74-4

[1]https://doi.org/10.5281/zenodo.15420422

The remainder of this work is structured as follows. Section II provides related work on measurement platforms and studies in the mobile landscape. Section III presents MARKlab and introduces its components. Its functionality and measurement opportunities are explained in Section IV. We then evaluate and present the results of a case study performed in MARKlab in Section V. A discussion of the platform and its limitations, as well as directions for future research, is done in Section VI before Section VII provides concluding remarks.

II. Related Work

Regarding previous research, two directions need to be considered. First, positioning our contribution among existing mobile network measurement platforms and, second, classifying QoS and energy measurements in mobile environments.

A. Measurement Platforms

Several platforms have been developed to measure and analyze mobile network performance. MobileAtlas provides large-scale crowdsourced network performance data with a focus on security aspects [8]. The MONROE platform—developed under the European project of the same name and later maintained by the MONROE Alliance—offered a large-scale, multi-RAT testbed for mobile broadband experimentation across Europe [9], though it is no longer operational at the time of writing. Mobilyzer is a smartphone-based measurement library designed to help researchers run network performance tests efficiently and in an application-layer context [10]. RIPE Atlas and PlanetLab provide large-scale distributed infrastructures for network measurement with a focus on residential and WAN-based networking rather than cellular service [11], [12]. Wehe is primarily designed to detect traffic differentiation but does not perform broader QoS analysis [13]. The AmiGos platform offers QoS monitoring tailored for mobile services by exploiting travelers equipped with smartphones running custom software [14]. Finally, BISmark consists of home gateways running custom firmware, focusing on fixed-broadband measurements at the residential level [15].

Compared to previous approaches, our proposed system specifically targets distributed mobile measurements with identical hardware on every node to ensure comparability. It emphasizes multi-RAT QoS, provides sustainability metrics, and supports dynamic SIM selection at runtime, exceeding the flexibility of existing platforms. Unlike MobileAtlas and MONROE—which primarily focus on broadband or security-related metrics—our system is built for multi-technology radio access, fine-grained energy profiling, and application-layer protocol evaluations, making it well-suited for, but not limited to, IoT deployments. The differences of all the measurement platforms can be seen in detail in Table I.

Smartphone-based crowdsourcing excels at large-scale and application-layer assessments of realistic conditions with heterogeneous end user devices. However, it lacks control of the measurement conditions and the used devices in the tests and therefore, repeatability needed for technology-specific benchmarking—especially in IoT and roaming contexts [16].

Table I
FEATURE COMPARISON OF MEASUREMENT PLATFORMS.

Platform	Mobile Focus	Crowd-sourced	Uniform HW	Multi-RAT	Dyn. SIM Selection	App-Eval.	Sustain-ability
MobileAtlas	✓	✓	✓	✓	✓	✗	✓
MONROE	✓	✗	✓	✓	✗	✗	✓
Mobilyzer	✓	✓	✗	✗	✗	✓	✗
RIPE Atlas	✗	✓	✓	✗	✗	✗	✗
PlanetLab	✗	✗	✗	✗	✗	✓	✗
Wehe	✓	✓	✗	✗	✗	✓	✗
AmiGos	✓	✓	✗	✗	✗	✗	✗
BISmark	✗	✗	✓	✗	✗	✗	✗
MARKlab	✓	✗	✓	✓	✓	✓	✓

MARKlab has a different focus and overcomes these limitations with dedicated modems, remote SIM decoupling, and full control over hardware and software. It enables systematic testing across RATs and countries via predefined procedures and integrates custom measurements without requiring new smartphone apps. With built-in current and voltage sensors, it also provides accurate energy profiling—something smartphones cannot offer. Thus, MARKlab complements crowdsourced approaches by enabling controlled, repeatable, and in-depth mobile network analysis for IoT, see the discussion on use cases for crowdsourced network measurements [16], [17].

B. Measurement Studies in Mobile Networks

Public mobile network measurement studies often lack the scale for generalizable results. Below, we review key works on global mobile infrastructure, emphasizing QoS and energy efficiency, to highlight existing research limitations.

Turk et al. [18] explore the impact of converged networks on QoS but limit their study to a single MNO. Baltrunas et al. [19] assess mobile broadband reliability using the NorNet Edge platform [20], though they overlook energy efficiency and broader QoS aspects. El-Saleh et al. [21] analyze QoS in a Malaysian city, focusing on specific operators. Loh et al. [22] evaluate 5G QoS via OpenSignal data [23]. Nikravesh et al. [24] conduct a 17-month global latency study using Android devices, exposing MNO discrepancies. Alay et al. [9] extend NorNet Edge with large-scale MONROE platform measurements. Lutu et al. [25] analyze roaming performance across six European countries and 16 MNOs but focus on mainstream 3G/4G technologies, excluding IoT solutions like NB-IoT or LTE-M. Geissler et al. [4] examine signaling traffic in European networks. Lutu et al. and Vomhoff et al. [3], [26] identify key IPX network characteristics for global roaming. Finally, several studies evaluate mobile energy consumption under various conditions [2], [27]–[29], though these are limited to small-scale lab environments, overlooking real-world factors affecting energy use. To address these gaps, MARKlab aims to deploy measurement nodes across diverse locations, supporting multiple access technologies and energy sensing alongside network and application QoS measurements.

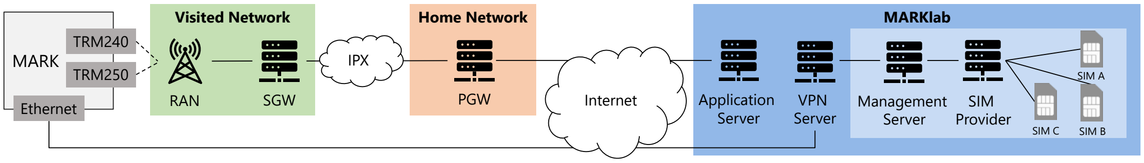

Figure 1. Schematic overview of MARKlab architecture; Measurement Nodes (left), Home and Visited Network (middle), and Platform Infrastructure (right).

III. The MARKlab Platform

In line with previous mobile measurement platforms like MONROE, MARKlab is built for scalability and maintainability. Its modular design enables independent scaling, updates, and replacements of components without affecting the rest of the platform. Figure 1 presents an end-to-end overview of the key measurement components, each of which is briefly introduced below. A more detailed technical description is available on the platform website.[2]

A. MARKlab Core

The core components of the platform, shown in Figure 1 in blue, represent the centrally hosted backend services. These include the Management, VPN, and Application Server, as well as the SIM Provider.

The *Management Server* represents the heart of the entire platform and is responsible for the orchestration of measurements and result collection. In addition, the management server provides the interface to the central database and data storage to persist both metadata and actual measurement results. Finally, it provides the automation API as well as the web UI. In order to enable seamless connectivity between the management server and the MARK nodes, all components are connected to the same *VPN Server*. The management API and the UI are the only core services reachable outside the VPN.

The final management component, the *SIM Provider* is responsible for hosting and tunneling available SIM cards to the measurement nodes. The SIM provider is adapted from the open source solution developed in the MobileAtlas project [8]. The SIM tunneling feature enables us to host all SIM cards locally in the lab, without any SIMs actually being deployed in the field next to the measurement nodes. Instead, SIM identities are tunneled to the corresponding node at runtime, effectively allowing a remote SIM swap at any time. This allows us to both reduce the number of total SIMs in the system, and enables a single SIM card to "travel" between nodes, thereby reducing cost and increasing flexibility.

Moving to the data plane components, the *Application Server* is simply an Internet-facing server hosting endpoints for various IoT-focused protocols, such as MQTT, CoAP, AMQP, XMPP or HTTP. We operate this service to have a consistent target and are able to capture performance metrics on the server side, should an experiment require it. Note, however, that experiments can be conducted to any arbitrary target reachable via the Internet.

[2]https://lsinfo3.github.io/marklab-website/

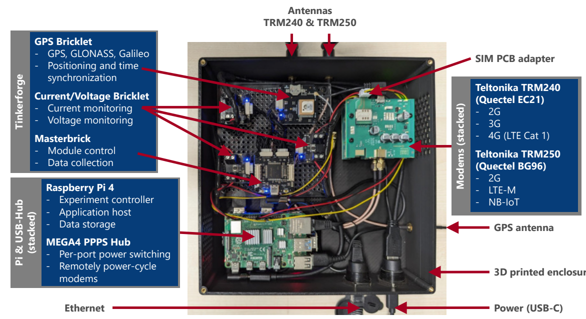

Figure 2. Internals of a MARK Node: Raspberry Pi 4, Teltonika Modems TRM240 and 250, Tinkerforge GPS tracker and energy sensors.

B. MARK Nodes

The *MARK Nodes* are the actual measurement nodes deployed in the field. At the time of writing, we have deployed a total of eight nodes in four European countries, with four more planned or already shipped and scheduled to come online.

Figure 2 shows the interior of a MARK node and annotates its key technical components. Built around readily available hardware, the nodes can be manufactured relatively cheaply and, aside from the enclosure, don't include any proprietary hardware. With a Raspberry Pi 4 forming the processing core of each node, additional key components are its mobile modems as well as the included energy measurement sensors. Based on the Quectel EC21 and BG96 chips, the two included Teltonika modems (TRM240, TRM250) offer a wide range of compatible RATs. At the same time, these two chips are widely used in IoT verticals. In combination, the two modems allow us to conduct measurements using EDGE (2G), DC-HSPA+ (3G), LTE Cat-1, LTE-M, and NB-IoT. Note that the selection of modems is not limited by the platform. The selection has been made based on the focus of the platform on IoT use cases. An extension by, e.g. 5G connectivity or faster LTE capabilities can be easily done, even on a per-node basis.

To provide some background of the available technologies, 2G and 3G correspond to their regular implementation as standardized by the 3GPP [30], [31]. LTE Cat-1 works, from a technology point of view, identical to regular LTE (e.g. LTE Cat3-12), while the bandwidth is limited to $10\,\mathrm{Mbps}$ in the downlink and $5\,\mathrm{Mbps}$ in the uplink with the goal to reduce the energy costs for IoT devices. LTE-M (LTE Cat-M1) pushes this resource limitation even further, by reducing the available bandwidth from $20\,\mathrm{MHz}$ in LTE Cat-1 to $1.4\,\mathrm{MHz}$ in LTE-M, resulting in a throughput of $1\,\mathrm{Mbps}$in both uplink and downlink direction. Finally, on the extreme end of the spectrum, NB-IoT (LTE Cat-NB1) reduces the available resources even further, with throughputs of 66 kbps in the uplink and 26 kbps in the downlink. In addition, NB-IoT does not provide the full mobile feature set, as there is no mobility management available. Hence, devices are expected to re-register after moving, and are indented to remain stationary. For a full breakdown of the technologies, we refer to the corresponding standard documents [32]. In addition, we provide a brief comparison on the MARKlab website.

When it comes to both managing and measuring power within each MARK node, we include a MEGA4 PPPS USB Hub that allows switching individual ports on and off. This enables fine granular control of which components in the node are active at any given time, and allows us to remotely power cycle the attached modems to reset them into a known state after failed or interrupted measurements.

Spliced in-between the USB power delivery and the modems itself as well as the power delivery of the Raspberry Pi, three Tinkerforge Current/Voltage bricklets capture the energy consumption of both of the modems, and the system as a whole. Regarding the details on the measurements itself, we refer to Section IV, detailing the measurement procedure executed by each node during an experiment run.

Finally, each node is equipped with a GPS sensor to detect its current position to report back to the management server. While this is currently of limited relevance, mobile deployments are on the roadmap and are currently being worked on.

We open-source the code for the *Management Server* and MARK nodes, as well as a construction manual and the design files for the enclosure.[3]

C. Mobile Roaming

Finally, while not technically part of the MARKlab platform, the transmission path between MARK nodes, platform infrastructure, and the Internet is shown in Figure 1 (orange, green). With the SIM cards used in this study, MARK nodes register as outbound roamers in each visited network, meaning all measurements in this work occur under roaming conditions. Nodes connect to local mobile operators, and under Home-Routed (HR) roaming, traffic flows from the visited network through the IP-Exchange Network (IPX) to the home operator before reaching the Internet. Note that the use of roaming SIMs is not a platform limitation, as native SIMs can also be used. However, roaming was chosen to reflect real-world IoT deployments at international scale.

IV. Measurement Concept

In order to allow for consistent, reproducible and comparable measurements across all MARK nodes deployed in the platform, we define a strict procedure that all conducted measurements – called jobs – adhere to. To this end, the platform is designed around predefined measurements, defined as individual Docker images, each focusing on a specific experiment, such as QoS analysis, energy consumption, or application specific measurements. Additionally, the platform supports the deployment of custom Docker images, as long as the requirements for the platform structure are met. In general, we differentiate between three types of measurements: (i) *Measurement Tasks (M)* are active measurements that are run after a mobile connection has been established. Measurement tasks are always executed sequentially to eliminate interference between tasks. (ii) *Background Tasks (B)* are measurements that runs in the background, parallel to a measurement task. It starts before a connection to a mobile operator is established and ends after the connection is closed, making it particularly useful for energy measurements or other continuous experiments. Multiple background tasks can run simultaneously. Finally, (iii) *Enhanced Background Tasks (EB)* are hybrid measurements that start after a mobile connection is established. Multiple enhanced background tasks can run in parallel, making it suitable for tasks such as network packet capture using `tcpdump` or continuous monitoring of signal strength. Table II summarizes the predefined images supported by MARKlab, which, as of now, have been used to conduct more than 25 000 measurements.

Table II
Measurement tasks.

Name	**Type**	**Description**
AT Commander	M	Sends AT commands to the modem
fPing Delay	M	Measures network latency using fPing
Ping Delay	M	Measures ICMP ping delay
Curl Download	M	Downloads a file via Curl to test network speed
Network Report	M	Generates a report of network conditions
Traceroute Route	M	Traces network route to a target destination
Signal Quality	M	Measures and reports network signal quality
Curl Upload	M	Uploads a file using Curl to test upload speed
Get IP Address	M	Returns the device's public IP address
Current Voltage	B	Measures energy/voltage levels of the device
GPS Tracker	B	Tracks the GPS position of the device
System Information	B	Collects system-related information and metrics
Temperature CPU	B	Monitors the CPU temperature of the device
Packet Capture	EB	Captures network packets for analysis
Signal Strength	EB	Measures RSSI and the channel bit error rate

During platform operation, each node is continuously running a management loop, as illustrated in Figure 3. In randomized intervals with a mean sleep duration of 60 seconds, a node performs a cycle of the management loop, starting with a check if the node is currently performing a measurement. In case the node is busy, it immediately goes back to sleep to not interfere with running experiments. If no experiment is running, a cleanup job is performed to terminate any lingering processes and containers that might be left by a crashed or otherwise flawed previous job. After checking if the node is in a healthy state (e.g., low disk space or memory, poor connectivity to VPN), the node ensures that previous measurement results have been successfully transmitted, and retransmits data in case of issues or inconsistencies with a previous job. This preparation phase is shown in Figure 3 in green.

After preparation has succeeded, new jobs are pulled from the management server. Thereby, jobs of multiple users are

[3]https://github.com/lsinfo3/marklab

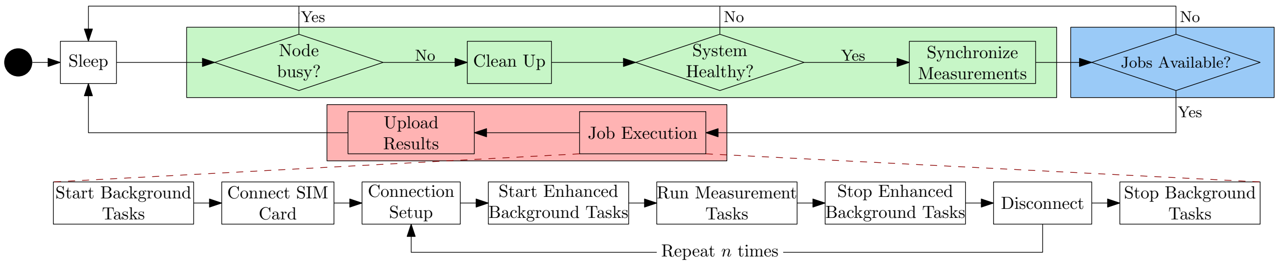

Figure 3. Sequence diagram of management routine executed by each MARK node.

scheduled based on fairness, prioritizing users who haven't recently run a task. If no tasks are available, the system returns to sleep. The scheduling step is shown in blue. If a job is scheduled for execution, the node queries task instructions and executes the individual tasks defined in the job, shown in red.

Background tasks are started first, each running in separate, isolated environments. These background tasks continue to run throughout the entire measurement process. After initiating the background tasks, a measurement iteration begins. This starts with the connection of a SIM card to the node via the SIM provider. Once a SIM is assigned, the respective modem is activated and a data connection is established. Following successful configuration, the device starts any configured enhanced background tasks before running the specified measurement tasks. After the tasks are completed, the enhanced background tasks are stopped, and the modem is deactivated to disconnect the connection to the selected provider, ensuring a clean state for the next iteration. Depending on job configuration, each experiment can be repeated multiple times. Thereby, only the job execution is repeated, and the same SIM card is used for all repetitions. After all iterations are completed, the background tasks are terminated and the collected data is packaged and sent to a central server for storage and analysis.

The mobile connection setup follows a structured process. The modem is enabled, the antenna activated, and automatic registration initiated. Configuration varies by modem (e.g., TRM250 settings for IoT). Network selection can be automatic, random, or manual. Once connected, the APN and IP settings are applied, and an ICMP ping test verifies connectivity. If packet loss exceeds 50 %, the connection attempt fails. Only after the node can verify an established connection, any measurement tasks are started.

Errors may arise at any stage, e.g. due to poor network conditions. The node continuously checks for known issues and retries each step up to three times. If failures persist, the modem is power cycled, and the process restarts. Should errors still persist after power cycling, the entire node is rebooted to ensure a clean slate. In this case, the job is marked as failed and needs to be manually rescheduled.

V. Case Study

To showcase the exemplary experiments made possible in MARKlab, we conducted a case study of key QoS metrics across multiple RATs (2G, 3G, LTE Cat-1, LTE-M, and NB-IoT) in four countries (Croatia, Germany, Italy, and Norway) and 12 mobile operators. The evaluation focuses on latency, packet loss, throughput, and energy consumption, highlighting differences between RAT, operators, and countries.

A. Latency

To assess device latency, we conduct a *Ping Delay* experiment (cf. Table II), measuring Round Trip Time (RTT) to Google's DNS (8.8.8.8) from all available nodes. The measurements were collected over the course of a year, from May 2024 to March 2025, at various times of day and on different days of the week, resulting in nearly 2 million RTT values. Figure 4 combines violin and box plots to depict RTT distributions in ms for various RATs (2G, 3G, LTE Cat-1, LTE-M, and NB-IoT) across operators in Croatia, Germany, Italy, and Norway. The violin shape reflects RTT density, with wider sections indicating frequent values and longer tails suggesting variability. Outliers are removed for clarity. Missing RAT-operator-country combinations imply unsupported RATs or SIM connection issues due to roaming constraints.

As expected, 2G has the highest latency among traditional mobile technologies, ranging from 100 ms to 706 ms with substantial fluctuations. Latency varies significantly between operators, especially in Italy and Norway, where Operator A in Norway and Operator C in Italy experience notably higher delays. In contrast, 3G networks offer lower and more stable RTTs (147 ms to 515 ms), but measurements are limited to Croatia and Italy, as Germany and Norway have phased out 3G. In Croatia, Operator B stands out with significantly lower 3G latency than its counterparts. As all measurements within a country have been performed using the same MARK node, all experiments also use the same transmission path in home-routed roaming. Hence, a significant improvement (e.g. Croatia Operator B vs A) is either a result of lower radio latency, or better IPX connectivity.

Moving on, LTE Cat-1 networks exhibit substantially lower RTTs than both 2G and 3G, with values between 35 ms and 136 ms. However, in Norway, Operator A shows significantly higher RTT than other operators with values up to 309 ms. Interestingly, the same operator also exhibits a higher delay in 2G as well as LTE-M connectivity. With RTT values between 48 ms and 221 ms, general RTTs are again significantly lower compared to Norway's Operator A, which has values ranging from 229 ms and 353 ms. Finally, NB-IoT

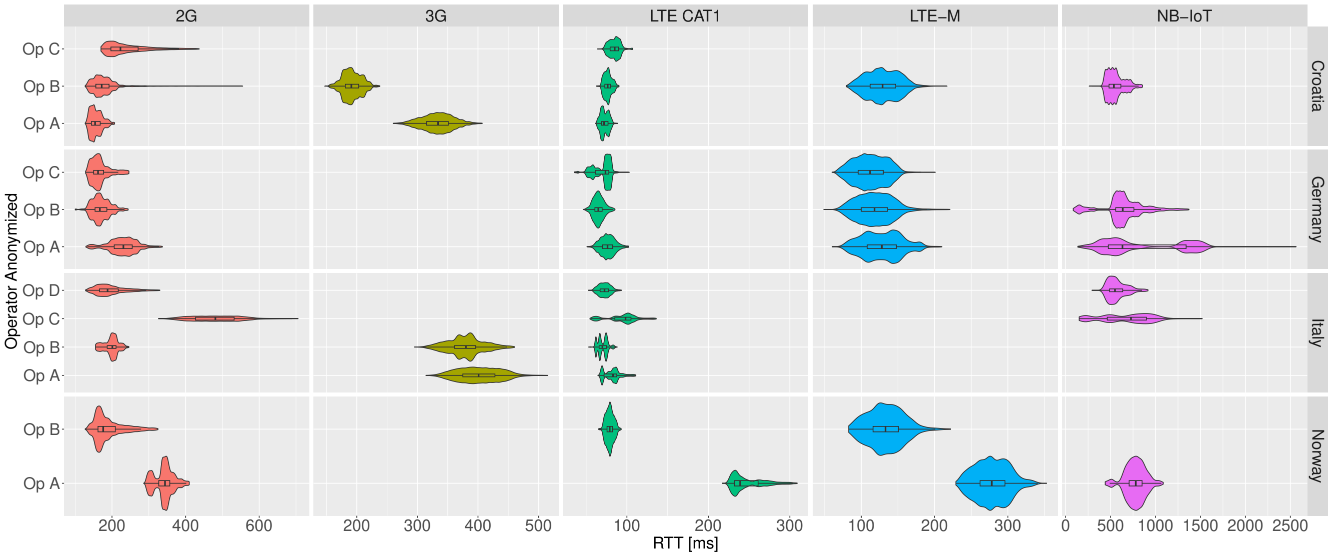

Figure 4. Latency distributions across 12 anonymized operators in Croatia, Germany, Italy and Norway for all investigated RATs.

exhibits the highest and most variable RTT among all technologies, with many values exceeding 1000 ms and reaching over 2500 ms. The violin plots for NB-IoT are notably wide, reflecting inconsistent and unpredictable latency. This behavior is largely due to NB-IoT's design, which prioritizes coverage and energy efficiency over latency, resulting in longer transmission delays and greater signaling overhead. In addition, NB-IoT often operates on shared or re-farmed spectrum with limited scheduling resources, contributing to variability. The lack of support for advanced features such as RRC resume, to reconnect a device faster, or optimized DRX cycles can further increase delay, especially in roaming scenarios. Compared to the other countries, NB-IoT in Croatia performs notably better, which may indicate more efficient local network configurations or reduced congestion.

Across all RATs, some operators show long tails, indicating frequent high RTT values. Multiple peaks in the distributions suggest external influences like congestion, location, or routing inefficiencies. This highlights NB-IoT as the least reliable technology regarding latency, despite its IoT-focus. However, high RTT values do not necessarily mean an operator has inherently high latency, as only roaming connections are considered. In some cases, increased delay results from suboptimal IPX routing as RTTs reflect the full end-to-end transmission path.

B. Packet loss

We evaluate packet loss using data from the *Ping Delay* task, which includes packet loss ratios. Figure 5 shows mean packet loss rates by country and RATs, while the error bars indicate 95% confidence intervals. We omit individual operator data due to marginal differences between operators.

Overall, Croatia exhibits the lowest packet loss at 0.034%, reflecting highly reliable networks. Germany, Italy, and Norway exhibit higher packet loss (0.23–0.40%), though confidence intervals suggest no significant differences. Packet loss varies across RATs as 2G exhibits the highest loss and variability, underscoring its reliability issues, likely due to its radio technology, limited spectral efficiency, and reduced error correction capabilities. In contrast, 3G, LTE Cat-1, and LTE-M experience negligible loss. NB-IoT shows moderate packet loss – higher than LTE-based technologies, but still significantly lower than 2G. Overall, a non-significant trend suggests higher packet loss in Germany, Italy, and Norway compared to Croatia, though all measured loss rates remain below 1%.

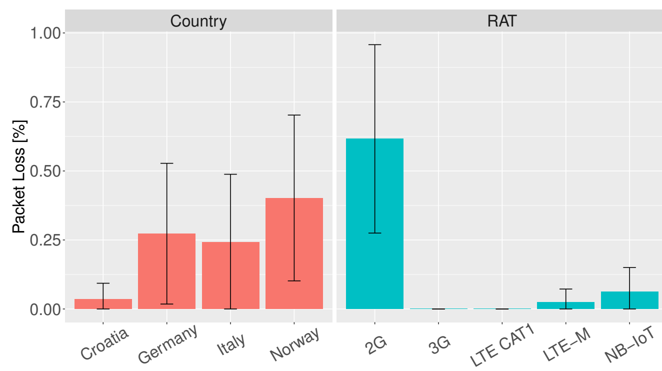

Figure 5. Packet loss across all investigated countries and RATs.

C. Throughput

Moving on to throughput, we conduct experiments using the *Curl Download* and *Curl Upload* measurement tasks. The measurements were conducted in March 2025, with a total of 1500 runs. Due to the expected throughput differences of the different RATs, we transmit differently sized files for each technology, as shown in Table III. These filesizes have been chosen to be large enough to overcome TCP slow start, but are

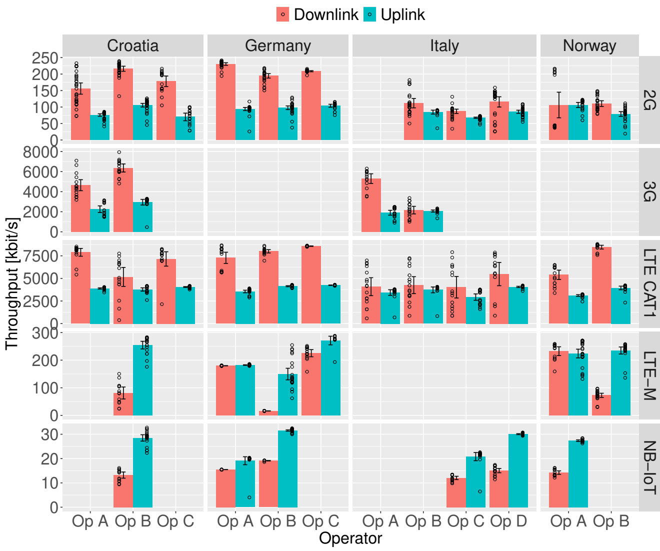

Figure 6. Throughput for downlink (red) and uplink (blue) across operators, countries and RATs.

kept as small as possible to be as cost-effective as possible. Figure 6 illustrates the downlink and uplink throughput (in kbps) for various RATs, separated by operators and countries. Downlink throughput is shown in red, uplink in blue. The bars represent the mean throughput, with individual data points and error bars indicating the 95% confidence intervals.

As expected, 2G exhibits the lowest throughput of the traditional technologies, with both downlink and uplink values generally below $250\,\mathrm{kbps}$. In Italy and Norway, the downlink speeds are slightly higher than uplink, but the differences remain minor. In Croatia and Germany, the mean uplink throughput is only half of the downlink throughput. In contrast, 3G offers significantly improved performance, with a maximum observed downlink throughput of $6.3\,\mathrm{Mbps}$ (Croatia, Operator B). Uplink speeds are lower but still significantly better than 2G. Note the missing 3G values for Germany and Norway, as 3G mobile networks are no longer available there.

LTE Cat-1 demonstrates even higher throughput, with mean values frequently above $7\,\mathrm{Mbps}$. The uplink speeds remain lower than the downlink, but are still considerably better than previous generations. Performance remains relatively stable across different operators for Germany and Italy, although some variability can be observed for Croatia and Norway. Moving to LTE-M, the throughput is significantly lower than LTE Cat-1 and 3G, with both downlink and uplink values reduced. However, with a focus on IoT use cases, LTE-M exhibits significantly larger uplink than downlink, as it is useful for sensing, tracking or other upload-focused verticals.

Table III
USED FILESIZES IN THROUGHPUT MEASUREMENTS.

Direction	2G	3G	LTE Cat 1	LTE-M	NB-IoT
Uplink	100 kB	2 MB	2 MB	200 kB	50 kB
Downlink	500 kB	5 MB	5 MB	500 kB	100 kB

NB-IoT exhibits the lowest throughput among all RATs, with both downlink and uplink speeds staying below 35 kbit/s. Similar to LTE-M, the uplink throughput is higher than the downlink, which aligns with NB-IoT's design for low-power, low-data-rate applications.

Typically, for traditional RATs downlink throughput exceeds uplink throughput due to asymmetric resource allocation in most RATs. However, we observe that while this pattern generally holds, Italian operators exhibit notably smaller differences between uplink and downlink. This could be due to balanced scheduling policies tailored for IoT traffic, backhaul limitations, or bottlenecks resulting in more symmetric throughput.

D. Energy Efficiency

Finally, investigating the energy consumption during operation for all combinations of country, operator, and RAT, we conduct a *Curl Upload* measurement task combined with a *Current Voltage* background task (cf. Table II). Figure 7 presents a violin plot with an embedded box plot illustrating the distribution of observed current values during upload in mA. Outliers have been removed to enhance readability. The current values are gathered every $5\,\mathrm{ms}$ during the upload process. For each parameter combination, we transmit a file according to Table III, in order to have a measurement last at least a couple of seconds to gather the different current values. Note that it is expected, to observe multi-modal distributions for all RATs, due to the periodic nature of radio signal modulation. During active transmission, there will be phases of low power draw with intermittent spikes in periodic intervals. We omit the technical details here and refer to [2] for an exemplary timeseries of such current values during operation.

As expected, 2G exhibits high energy consumption with significant probability mass in the range of $500\,\mathrm{mA}$. 3G shows lower energy consumption than 2G, with values generally ranging between $100\,\mathrm{mA}$ and $400\,\mathrm{mA}$, although variations exist among different operators. These variations may be due to distance to the base station and the employed uplink power control mechanism [33]. LTE Cat-1 demonstrates an energy-profile with current values up to $800\,\mathrm{mA}$. However, the range of values varies substantially between operators.

LTE-M displays a lower and more consistent energy consumption pattern, with values predominantly between $50\,\mathrm{mA}$ and $300\,\mathrm{mA}$. The narrower violin shapes indicate less variability, suggesting a smaller difference between low phases and the aforementioned spikes. NB-IoT exhibits the lowest current consumption, with most values concentrated below $150\,\mathrm{mA}$, reinforcing its suitability for low-power applications.

Overall, differences can be observed between countries and operators. In some cases, operators within the same country exhibit considerable differences in power consumption, particularly for 3G and LTE Cat-1. This variability highlights the impact of network configuration, signal conditions, and operator-specific implementations. In contrast, NB-IoT and LTE-M present a more uniform power consumption pattern

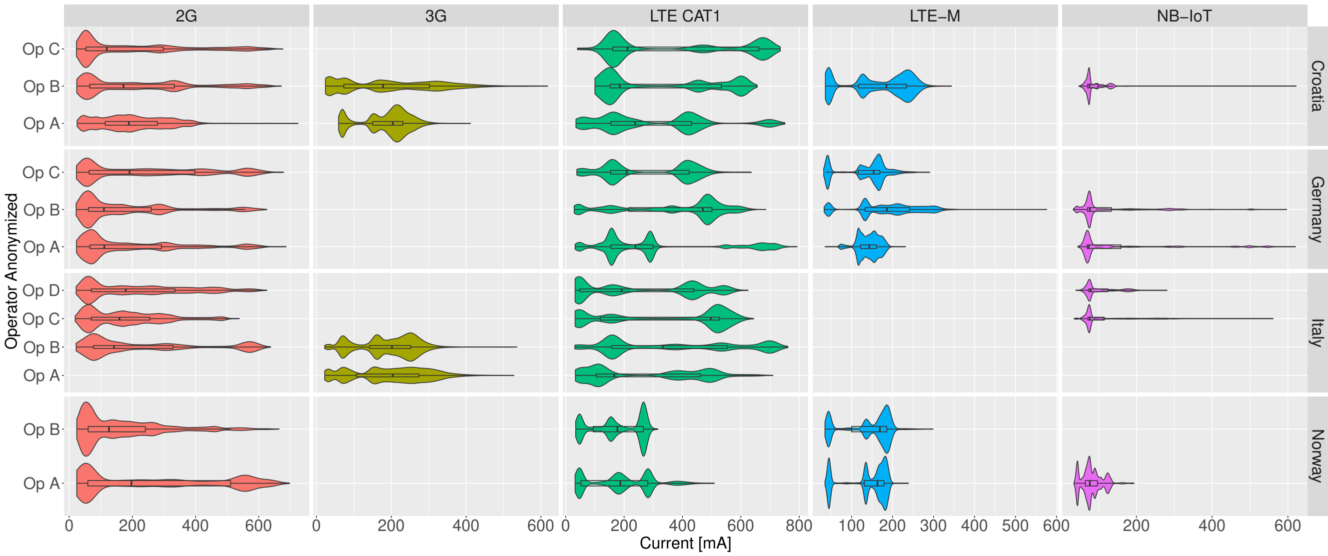

Figure 7. Violin plots of current values gathered every 5 ms during an upload.

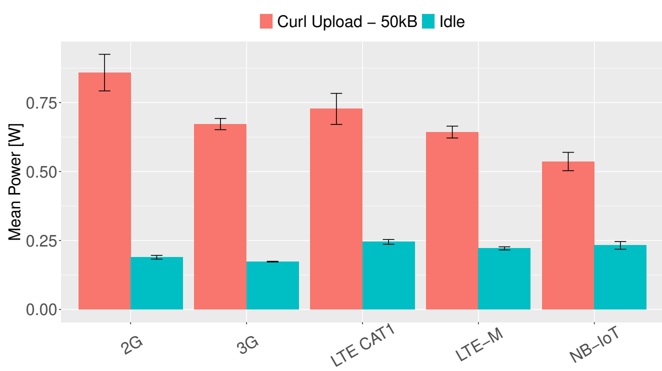

Figure 8. Mean power draw during *Curl Upload* (red) and idle phases (blue).

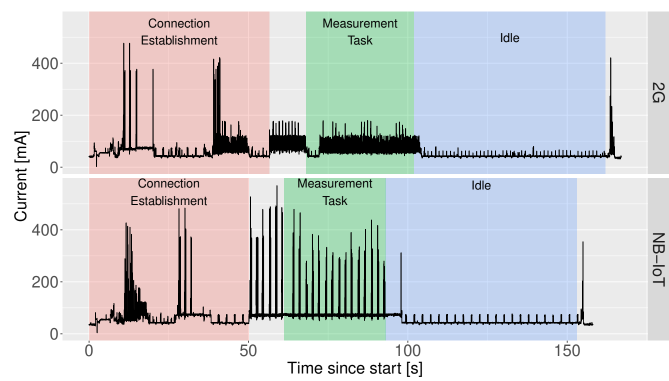

Figure 9. Time series of current [mA] during a typical measurement cycle for 2G (top) and NB-IoT (bottom).

across countries, aligning with their design goals for low-power IoT applications.

Finally, Figure 8 summarizes the power draws observed for each evaluated RAT. The figure shows the mean power draw observed during upload (red) and idle phases (blue). We define idle as being connected to the network and maintaining an open data tunnel, without transmitting payload data. For the upload in this case for all RATs we transmit a 50 kB file, to ensure comparability in the amount of performed work.

The data shows that, during upload, NB-IoT and LTE-M draw significantly less power than LTE Cat-1 and 2G, highlighting again their focus on low power applications. However, when combining these values with the observed throughput values in Figure 6, it becomes apparent that transmissions using LTE-M and NB-IoT will have a longer duration than transmitting the same amount of data with, e.g. LTE Cat-1 and LTE-M. Hence, the total energy consumption is highly dependent on the data volume to be transmitted. During idle phases, the energy consumption of all technologies are largely similar. The data shows LTE Cat-1 to draw significantly more power than other RATs, while 3G was observed to draw the least amount of power. Note that these values have been obtained without any energy optimization mechanisms such as power saving mode (PSM) or Extended Discontinuous Reception (eDRX) [34].

To better understand the high and variable delays of NB-IoT, Figure 9 shows the current consumption during a typical measurement. Unlike 2G, which quickly enters a low-power idle state, NB-IoT remains semi-active, with periodic spikes even in idle. This suggests deep sleep is not reached, likely due to failed Power Saving Mode (PSM) negotiation, firmware limits, or lack of network support—leading to prolonged connected states and increased latency.

These results reflect NB-IoT's trade-offs: it targets energy-constrained, delay-tolerant applications and offers deep coverage, but at the cost of higher and less predictable latency. In contrast, LTE-M strikes a better balance between energy use and responsiveness, better serving latency-sensitive IoT scenarios. Our analysis highlights how each technology aligns with its intended KPIs under real-world conditions.

VI. Discussion

The MARKlab platform introduced in this work and used to conduct the presented case study provides a powerful tool for analyzing global cellular connectivity, particularly in the context of IoT networks and machine-to-machine communication. However, as conducting reliable and valid measurements in real-world production mobile networks is inherently challenging, several limitations need to be discussed and kept in mind when conducting experiments using MARKlab. Moving forward, several of the mentioned limitations will be explored to enhance the platform's capabilities and evolve both our understanding of the global mobile ecosystem and the best practices and required methodologies to conduct large scale experiments in real-world mobile networks.

Hardware Limitations: MARK nodes are build on top of a uniform hardware platform with preselected components. Despite rigorous testing and following general best practices during development, experiments may still face hardware-imposed limitations. Identical node design ensures comparability but lacks hardware variation, which could affect results. This is especially relevant for energy-related experiments, where modem type, firmware, or experiment layout can influence outcomes. As maintaining a fleet of heterogeneous nodes is largely infeasible, this limitation can be worked around by performing lab experiments in addition to MARKlab experiments as well as compare results obtained in different platforms, such as MobileAtlas [8].

SIM Bias: SIM bias occurs in platforms like MARKlab when only a limited set of SIM identities is available, which may lead to skewed results. This bias can arise from factors such as regional restrictions, network provider limitations, or the general availability of a heterogeneous set of SIMs. As a result, the collected data may not accurately represent the broader mobile user base, potentially affecting conclusions about network performance, user behavior, or application functionality. Addressing SIM bias requires strategies such as diversifying SIM sources, randomizing test SIMs, and using statistical adjustments to correct for sampling distortions. As of the time of writing, we only employ a single SIM sponsor, meaning all experiments have been conducted using SIMs of the same international operator. We are already working out diversifying the available set of SIM cards.

Limited Geographical and Deployment Scope: An inherent limitation of our platform lies in the deployment of nodes, both in terms of total number of nodes and their geographic distribution. To obtain representative and generalizable data, a wide range of countries, operators, and locations needs to be covered by nodes. However, due to organizational and budget constraints, there is a natural limit to how many nodes can be deployed and where they can be deployed. Naturally, we are continuously working towards expanding the platform and have already several additional nodes in preparation to expand into additional countries (Spain, Switzerland, Brazil, USA) as well as expand our presence in existing countries, e.g. to cover rural or remote areas.

Inherent Measurement Variability: Conducting measurement in public, non-controlled environments results in an inherent variability in results due to dynamic factors such as user mobility, fluctuating network load, and environmental conditions. As, other than some technical metrics such as signal quality, no performance measures are known about the underlying infrastructure of the visited network, there is a large factor of uncertainty that can not be eliminated. To minimize this uncertainty, it's essential to optimize measurement methodologies. In the MARKlab platform, conducting experiments across many nodes, running multiple repetitions, and designing experiments with this uncertainty in mind, allows experimenters to partially work around this inherent limitation. In the long run, once mobile MARK nodes are deployed, techniques like drive testing, which involves systematic data collection under controlled conditions, can help reduce inconsistencies. Additionally, ensuring proper measurement gap configurations can enhance the accuracy of collected data. By implementing these strategies, experimenters can improve the reliability of their results.

VII. Conclusion

In this work, we introduced MARKlab, a scalable platform for large-scale mobile network measurements under real-world conditions. MARKlab supports diverse experiments, from QoS and energy assessments to monitoring application behavior. With its robust measurement framework and the ability to deploy experiments as Docker images, MARKlab serves as a valuable tool for mapping the global mobile landscape.

Using MARKlab, we performed a comprehensive evaluation of key QoS parameters across multiple countries, mobile network operators, and RATs, with a focus on outbound roamers. Our study revealed significant variations in network performance, influenced by both RAT and operator selection.

The results show that 2G exhibits the highest latency, packet loss, and power consumption. 3G improves latency and throughput while reducing packet loss but is being phased out in some regions. LTE Cat-1 provides a balance between energy efficiency and throughput, while LTE-M offers lower power consumption at the cost of substantially reduced throughput. Finally, NB-IoT is the most power-efficient, but suffers from high latency and moderate packet loss. Variations between operators and countries highlight the influence of deployment strategies and infrastructure on network performance.

In the future, we will expand the number of nodes to cover additional countries and explore the integration of new access technologies such as 5G, LoRa, and Wi-Fi. The platform's modularity also allows for potential inclusion of IoT sensors (e.g., weather data) to support more realistic application scenarios. We are developing mobile MARK nodes for deployment in non-stationary locations like public transport. Additionally, we aim to analyze time-dependent traffic patterns, incorporate non-roaming SIM cards for complementary measurements, and introduce multiple endpoints to enhance latency analysis. Ultimately, the platform will be opened to external researchers.

References

[1] A. Lutu, B. Jun, A. Finamore, F. E. Bustamante, and D. Perino, “Where things roam: Uncovering cellular iot/m2m connectivity,” in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 147–161.

[2] V. Vomhoff, S. Raffeck, S. Gebert, S. Geissler, and T. Hossfeld, “Nb-iot vs. lte-m: Measurement study of the energy consumption of lpwan technologies,” in *2023 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2023, pp. 403–408.

[3] V. Vomhoff, M. Sichermann, S. Geißler, A. Lutu, M. Giess, and T. Hoßfeld, “A shortcut through the ipx: Measuring latencies in global mobile roaming with regional breakouts,” in *2024 8th Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2024, pp. 1–10.

[4] S. Geißler, A. Lutu, F. Wamser, T. Favale, V. Vomhoff, M. Krolikowski, M. Mellia, D. Perino, and T. Hoßfeld, “Untangling iot global connectivity: the importance of mobile signaling traffic,” *IEEE Transactions on Network and Service Management*, 2024.

[5] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, “Classifying iot devices in smart environments using network traffic characteristics,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1745–1759, 2018.

[6] S. Hui, H. Wang, D. Xu, J. Wu, Y. Li, and D. Jin, “Distinguishing between smartphones and iot devices via network traffic,” *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1182–1196, 2021.

[7] F. Metzger, T. Hoßfeld, A. Bauer, S. Kounev, and P. E. Heegaard, “Modeling of aggregated iot traffic and its application to an iot cloud,” *Proceedings of the IEEE*, vol. 107, no. 4, pp. 679–694, 2019.

[8] G. K. Gegenhuber, W. Mayer, E. Weippl, and A. Dabrowski, “{MobileAtlas}: Geographically decoupled measurements in cellular networks for security and privacy research,” in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 3493–3510.

[9] Ö. Alay, A. Lutu, R. García, M. Peón-Quirós, V. Mancuso, T. Hirsch, T. Dely, J. Werme, K. Evensen, A. Hansen *et al.*, “Monroe: Measuring mobile broadband networks in europe,” in *Building the Future Internet through FIRE*. River Publishers, 2022, pp. 155–187.

[10] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao, “Mobilyzer: An open platform for controllable mobile network measurements,” in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 389–404.

[11] R. N. Staff, “Ripe atlas: A global internet measurement network,” *Internet Protocol Journal*, vol. 18, no. 3, pp. 2–26, 2015.

[12] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, “Planetlab: an overlay testbed for broad-coverage services,” *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 3, pp. 3–12, 2003.

[13] F. Li, A. A. Niaki, D. Choffnes, P. Gill, and A. Mislove, “A large-scale analysis of deployed traffic differentiation practices,” in *Proceedings of the ACM Special Interest Group on Data Communication*, 2019, pp. 130–144.

[14] M. Varvello and Y. Zaki, “A worldwide look into mobile access networks through the eyes of amigos,” in *2023 7th Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2023, pp. 1–10.

[15] S. Sundaresan, S. Burnett, N. Feamster, and W. De Donato, “{BISmark}: A testbed for deploying measurements and applications in broadband access networks,” in *2014 USENIX Annual Technical Conference (USENIX ATC 14)*, 2014, pp. 383–394.

[16] M. Hirth, T. Hossfeld, M. Mellia, C. Schwartz, and F. Lehrieder, “Crowdsourced network measurements: Benefits and best practices,” *Computer Networks*, vol. 90, pp. 85–98, 2015.

[17] T. Hoßfeld, S. Wunderer, A. Beyer, A. Hall, A. Schwind, C. Gassner, F. Guillemin, F. Wamser, K. Wascinski, M. Hirth *et al.*, “White paper on crowdsourced network and qoe measurements–definitions, use cases and challenges,” *arXiv preprint arXiv:2006.16896*, 2020.

[18] Y. Turk and E. Zeydan, “An experimental measurement analysis of congestion over converged fixed and mobile networks,” *Wireless Networks*, vol. 26, no. 2, pp. 1017–1032, 2020.

[19] D. Baltrunas, A. Elmokashfi, and A. Kvalbein, “Measuring the reliability of mobile broadband networks,” in *Proceedings of the 2014 conference on internet measurement conference*, 2014, pp. 45–58.

[20] A. Kvalbein, D. Baltrūnas, K. Evensen, J. Xiang, A. Elmokashfi, and S. Ferlin-Oliveira, “The nornet edge platform for mobile broadband measurements,” *Computer Networks*, vol. 61, pp. 88–101, 2014.

[21] A. A. El-Saleh, A. Alhammadi, I. Shayea, W. H. Hassan, M. S. Honnurvali, and Y. I. Daradkeh, “Measurement analysis and performance evaluation of mobile broadband cellular networks in a populated city,” *Alexandria Engineering Journal*, vol. 66, pp. 927–946, 2023.

[22] F. Loh, F. Raithel, S. Raffeck, A. Seufert, R. Frohler, C. Heller, and T. Hossfeld, “Qos and qoe study of the european 5g mobile networks for next generation of applications,” *IEEE Communications Magazine*, 2025.

[23] OpenSignal, “Methodology Overview,” https://www.opensignal.com/methodology-overview, [Accessed 21-03-2025].

[24] A. Nikravesh, D. R. Choffnes, E. Katz-Bassett, Z. M. Mao, and M. Welsh, “Mobile network performance from user devices: A longitudinal, multidimensional analysis,” in *Passive and Active Measurement: 15th International Conference, PAM 2014, Los Angeles, CA, USA, March 10-11, 2014, Proceedings 15*. Springer, 2014, pp. 12–22.

[25] A. M. Mandalari, A. Lutu, A. Custura, A. S. Khatouni, Ö. Alay, M. Bagnulo, V. Bajpai, A. Brunstrom, J. Ott, M. Trevisan *et al.*, “Measuring roaming in europe: Infrastructure and implications on users’ qoe,” *IEEE Transactions on Mobile Computing*, vol. 21, no. 10, pp. 3687–3699, 2021.

[26] A. Lutu, B. Jun, F. E. Bustamante, D. Perino, M. Bagnulo, and C. G. Bontje, “A first look at the ip exchange ecosystem,” *ACM SIGCOMM Computer Communication Review*, vol. 50, no. 4, pp. 25–34, 2020.

[27] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, “Energy consumption in mobile phones: a measurement study and implications for network applications,” in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, 2009, pp. 280–293.

[28] V. Vomhoff, F. Poignee, F. Loh, S. Geissler, and T. Hossfeld, “Energy consumption in converged networks: A study of lte-m and non-terrestrial nb-iot,” in *2024 KuVS Fachgespräch - Würzburg Workshop on 6G Networks (WueWoWAS)*, 2024.

[29] M. Tawalbeh, A. Eardley, and L. Tawalbeh, “Studying the energy consumption in mobile devices,” *Procedia Computer Science*, vol. 94, pp. 183–189, 2016.

[30] 3rd Generation Partnership Project (3GPP), “3GPP TS 25.331: Radio Resource Control (RRC) Protocol Specification,” 3GPP, Tech. Rep., 2024, release 17. [Online]. Available: https://www.3gpp.org/DynaReport/25331.htm

[31] ——, “3GPP TS 43.022: Functions related to Mobile Station (MS) in idle mode and group receive mode,” 3GPP, Tech. Rep., 2024, release 17. [Online]. Available: https://www.3gpp.org/DynaReport/43022.htm

[32] ——, “3GPP TS 36.300: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Overall Description,” 3GPP, Tech. Rep., 2024, release 17. [Online]. Available: https://www.3gpp.org/DynaReport/36300.htm

[33] A. Simonsson and A. Furuskar, “Uplink power control in lte-overview and performance, subtitle: principles and benefits of utilizing rather than compensating for sinr variations,” in *2008 IEEE 68th Vehicular Technology Conference*. IEEE, 2008, pp. 1–5.

[34] A. K. Sultania, C. Blondia, and J. Famaey, “Optimizing the energy-latency tradeoff in nb-iot with psm and edrx,” *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 12 436–12 454, 2021.

Appendix

This research has been conducted with the highest regard for ethical standards, prioritizing the privacy and confidentiality of the user data gathered in the context of this work. The collected data points have been conscientiously cleaned and anonymized before exporting the data for the research conducted in this work. No personal identifiers can be linked to individuals, and no contact information was available or used at any stage of the research. Hence, this work does not raise any ethical issues.

This work was partly funded through the University sponsorship award of the IHK Würzburg-Schweinfurt (Dr. Stefan Geißler). The authors would like to thank Safa Matti for his efforts in implementing the platform.