Measuring the Reliability of Mobile Broadband Networks

Džiugas Baltrūnas
Simula Research Laboratory
dziugas@simula.no

Ahmed Elmokashfi
Simula Research Laboratory
ahmed@simula.no

Amund Kvalbein
Simula Research Laboratory
amundk@simula.no

ABSTRACT

Mobile broadband networks play an increasingly important role in society, and there is a strong need for independent assessments of their robustness and performance. A promising source of such information is active end-to-end measurements. It is, however, a challenging task to go from individual measurements to an assessment of network reliability, which is a complex notion encompassing many stability and performance related metrics. This paper presents a framework for measuring the user-experienced reliability in mobile broadband networks. We argue that reliability must be assessed at several levels, from the availability of the network connection to the stability of application performance. Based on the proposed framework, we conduct a large-scale measurement study of reliability in 5 mobile broadband networks. The study builds on active measurements from hundreds of measurement nodes over a period of 10 months. The results show that the reliability of mobile broadband networks is lower than one could hope: more than 20% of connections from stationary nodes are unavailable more than 10 minutes per day. There is, however, a significant potential for improving robustness if a device can connect simultaneously to several networks. We find that in most cases, our devices can achieve 99.999% ("five nines") connection availability by combining two operators. We further show how both radio conditions and network configuration play important roles in determining reliability, and how external measurements can reveal weaknesses and incidents that are not always captured by the operators' existing monitoring tools.

Categories and Subject Descriptors

C.4 [**Performance of systems**]: Measurement techniques; C.4 [**Performance of systems**]: Reliability, availability, and serviceability

IMC'14, November 5–7, 2014, Vancouver, BC, Canada.
Copyright 2014 ACM 978-1-4503-3213-2/14/11 ...$15.00.
http://dx.doi.org/10.1145/2663716.2663725.

General Terms

Experimentation; Measurement

Keywords

Mobile broadband; reliability; robustness

1. INTRODUCTION

Cellular Mobile Broadband (MBB) networks are arguably becoming the most important component in the modern communications infrastructure. The immense popularity of mobile devices like smartphones and tablets, combined with the availability of high-capacity 3G and 4G mobile networks, have radically changed the way we access and use the Internet. Global mobile traffic in 2012 was nearly 12 times the total Internet traffic in 2000 [4]. MBB traffic is estimated to keep growing at a compound annual rate of 66% towards 2017. An increasing number of people rely on their MBB connection as their *only* network connection, replacing both a fixed broadband connection and the traditional telephone line.

The popularity of MBB networks has given them a role as *critical infrastructure*. The reliability of MBB networks is important for the daily routines of people and business, and network downtime or degradations can potentially impact millions of users and disrupt important services. More importantly, failures can also affect emergency services and people's ability to get help when they need it.

Given the importance of MBB networks, there is a strong need for a better understanding of their robustness and stability. Regulators need data in order to make informed policy decisions and determine where extra efforts are needed to improve robustness. Today, regulators are often left with a posteriori incident reports from the operators, and lack a true understanding of the many smaller events that affect the reliability of services. Providers of mobile services that run on top of MBB networks need reliable data on reliability in order to predict the performance of their own services. End users can use such information to compare different operators and choose the provider that best fills their needs.

The ambition of this work is to *measure the experienced reliability in MBB networks*, and to compare reliability between networks. We believe that reliability in MBB networks is too complex to be understood only through static analysis of the components involved, and that the most promising approach for assessing and predicting the reliability of the offered service is through long-term end-to-end measurements. We argue that reliability must be characterized at several levels, including the basic connection between the user equipment and the base station, the stability of the data plane, and the reliability of application level performance. In this work, these aspects of reliability are assessed through long-term active measurements from a large number of geographically distributed measurement nodes. By looking at measurements from individual connections, we are able to identify important differences between networks and to characterize the reliability of each network as a whole. In summary, this paper makes the following contributions:

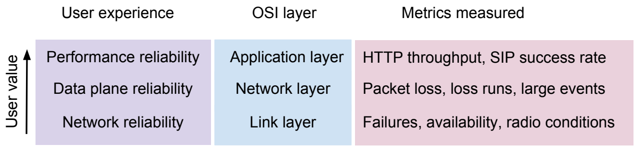
1. We propose a *framework for measuring robustness in MBB networks.* The framework captures aspects of reliability on several layers, from a basic registration in the network to a stable application performance over time. Within this framework, we define metrics and measurement experiments that describe reliability on the connection level, the data plane level, and the application level.

2. We present the *first large-scale measurement study of MBB reliability*, from a dedicated measurement infrastructure. The measurement experiments are performed on Nornet Edge (NNE) [15]. NNE is the largest infrastructure of its kind, with dedicated measurement nodes distributed in over 100 Norwegian municipalities. The data used in this work is captured from a total of 938 MBB connections from 341 distinct nodes and 5 different operators over a period of 10 months. Through long-term monitoring of a large number of connections, we find that a significant fraction of connections (15-38% depending on the operator) lose their network attachment more than 10 minute per day. We also observe clear differences in reliability characteristics between networks. While one network experiences frequent but short-lived connection failures, other networks have a longer time between failures but a higher overall downtime.

3. By capturing a rich set of metadata that describes the context of the measurements, this study *increases the value of end-user measurement data.* The metadata allows us to explain measurement results by looking at factors such as signal quality, radio state, network attachment, connection mode, etc. In many cases, we are also able to distinguish between problems in the radio access network and the mobile core network. We find a clear correlation between signal conditions, connection failures and loss, but we also discover that many failures can not be explained by signal quality. We further find that the inability to obtain dedicated radio resources is a common cause of application failures in some networks.

4. Thanks to the multi-connected nature of NNE measurement nodes, we can directly compare the performance and reliability of different networks at the same location, and thereby *quantify the potential gain in robustness from end-device multi-homing.* We find that there is mostly good diversity in radio conditions between operators, and that downtime can be reduced significantly if multiple networks can be used in parallel. In fact, most measurement nodes can achieve 99.999% ("five nines") connection availability when combining two operators.

The rest of this paper is organized as follows. Section 2 introduces our framework for measuring reliability in MBB networks. Section 3 presents the measurement infrastructure and data that forms the basis for our analysis. Sections 4 - 6 analyses reliability at the connection-, data- and application layers respectively. Section 7 looks at correlations between the different networks, and discusses the potential gain in robustness through multi-homing in light of this. Section 8 discusses related work, and finally, section 9 sums up and discusses the lessons learned from this study.



User experience	OSI layer	Metrics measured
Performance reliability	Application layer	HTTP throughput, SIP success rate
Data plane reliability	Network layer	Packet loss, loss runs, large events
Network reliability	Link layer	Failures, availability, radio conditions

Figure 1: Framework for measuring experienced reliability in MBB networks.

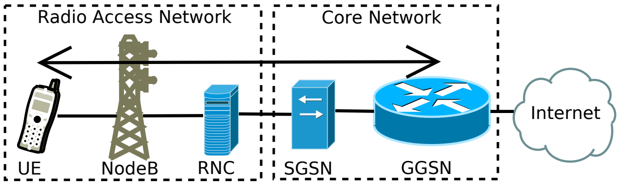


Figure 2: Simplified architecture of an UMTS MBB network.

2. A FRAMEWORK FOR MEASURING MOBILE BROADBAND RELIABILITY

Reliability is a complex notion, which relates to several stability and performance related metrics. Here, we propose a model where the reliability of a network is measured at different levels, reflecting increasing value for the user. A high level picture of the framework is shown in Fig. 1. The proposed model is a generic framework for describing the experienced reliability in MBB networks. In this work, we select a few relevant metrics at each level, and use these to characterize reliability of the measured networks. Other metrics can later be added to give an even more complete picture.

UMTS basics. Fig. 2 shows the main components of a UMTS network, divided into the Radio Access Network (RAN) and the Core Network (CN). Before any data can be transmitted, the User Equipment (UE), which can be a modem or a smartphone, must attach itself to the network and establish a Packet Data Protocol (PDP) context towards Gateway GPRS Service Node (GGSN). The PDP context is a data structure that contains the IP address and other information about the user session. This state is a prerequisite for any communication between the UE and the Internet. Once a PDP context is established, the Radio Network Controller (RNC) controls the Radio Resource Control (RRC) state of a user. Depending on the traffic pattern, RNC allocates a shared or dedicated radio channel for a user. If the user is not sending any data, RRC sets the state to IDLE or CELL_PCH. Otherwise, based on the bit rate, a user can be assigned a CELL_FACH state (shared channel, low bit rate, low power usage) or a CELL_DCH state (dedicated channel, high bit rate, high power usage). The principles are similar in networks based on the CDMA2000 architecture.

Connection level reliability. At the very basic level, the UE should have a reliable connection to the MBB network. By ”connection” in this context, we mean that there is an established PDP context in the CN. The stability of the PDP context depends on both the RAN and the CN; the PDP context can be broken by loss of coverage, failures in base stations or transmission, or by failures or capacity problems in the central components such as SGSN or GGSN. From the UE side, having a PDP context maps to having an assigned IP address from the mobile network. In Sec. 4, we measure reliability at the connection level by looking at the stability of the IP address assignment as a proxy for the PDP context. The metrics we look at are how often the connection is lost, and how long it takes before the node can successfully re-establish the PDP context. We also analyze how these metrics are related to underlying characteristics of the connections, such as signal strength and connection mode. The selected metric describes the stability of connections over time.

Data plane reliability. Having an established PDP context does not necessarily mean that the UE has well-functioning end-to-end connectivity to the Internet. Interference, drop in signal quality or congestion in either the wireless access or elsewhere in the mobile network may disrupt packet forwarding. This can cause periods of excessive packet loss, or ”gaps” where no data comes through.

In Sec. 5, we measure data plane reliability by looking at loss patterns in long-lasting continuous probing streams. We describe loss patterns in each network, and discuss how loss must be seen in relation with the radio condition of the MBB connection. We also use packet loss to identify abnormal events where packet loss is higher than normal for a significant number of connections.

Application layer reliability. Reliability also involves a notion of stability and predictability in the performance an application achieves over the MBB network. This stability depends of course on both the connection level reliability and the data plane reliability. Application layer performance varies depending on the specific application requirements. Some applications will perform well under a wide range of network conditions, while others have stronger requirements on available bandwidth or delay. In MBB networks, the experienced network performance depends on the *state* of the connection, since radio resources are assigned depending on the traffic load. It is therefore difficult to predict the performance of an application based on generic measurement probes. Instead, application performance should be assessed through experiments with actual application traffic.

In Sec. 6, we report on measurements with two typical applications: HTTP download using curl and Voice over IP (VoIP) using SIP/RTP. These applications have been selected because they are popular in MBB networks, and because they represent two quite different application classes in terms of traffic load. We measure the success rate, i.e., how often the download or VoIP call can be successfully completed. We also report on the stability of the achieved download rate.

This paper takes an important first step towards measuring the reliability of MBB networks through end-to-end measurements. An important aspect that is missing from this study, is mobility. All measurement nodes used in this work are stationary, and we can therefore not use these to describe how the stability of the offered service varies as you move.

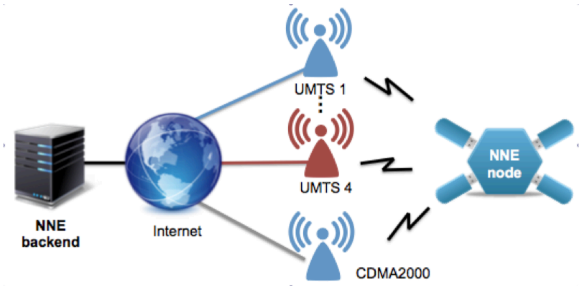


Figure 3: NNE overview.

There are, however, also advantages in doing measurements from fixed locations, since it removes a significant source of variation and uncertainty in the measurements. In future work, we plan to revisit MBB reliability in a mobile setting.

3. SYSTEM OVERVIEW AND DATA

This section presents the infrastructure that was used to run the measurement experiments described in this work, the MBB networks that were measured, the collected data, and how it is stored and post-processed.

3.1 The Nornet Edge measurement platform

NNE (Fig. 3) is a dedicated infrastructure for measurements and experimentation in MBB networks [15]. It consists of several hundred measurement nodes geographically distributed in more than 100 municipalities all over Norway, and a server-side infrastructure for management, processing and data storage. Figure 4 shows the placement of NNE nodes in Norway classified according to the number of MBB networks the node was connected to. NNE nodes are distributed to reflect the population density in Norway, with some bias towards urban areas. Nodes are placed indoors in small or large population centers, with a higher density of nodes in larger cities. More than half (177) NNE nodes are deployed in three largest cities, where 26.7%[1] of the coutry population lives.

An NNE node is a custom-made single-board computer, with a Samsung S5PV210 Cortex A8 microprocessor, one Fast Ethernet port, and 7 on-board USB ports. The node runs a standard Debian Linux distribution, giving large flexibility in the types of tools and experiments that can be supported. NNE also offers a set of tools for connection and configuration management, and a framework for deploying and managing measurement experiments. Each node is connected to 1-4 UMTS networks and 1 CDMA2000 1x Ev-Do network, using standard subscriptions. For the UMTS networks, connections are through Huawei E353 or E3131 3G USB modems. These modems support UMTS standards up to DC-HSPA and HSPA+ (”3.75G”) respectively, but not LTE (”4G”). They are configured so that they always connect to the 3G network where available, and fall back to 2G elsewhere. The same modem model is always used for all networks on the same node, to avoid differences caused by different hardware. For the CDMA2000 network, we connect to the Internet via a CDMA home gateway device over the Ethernet port.

[1] http://www.ssb.no/en/beftett/

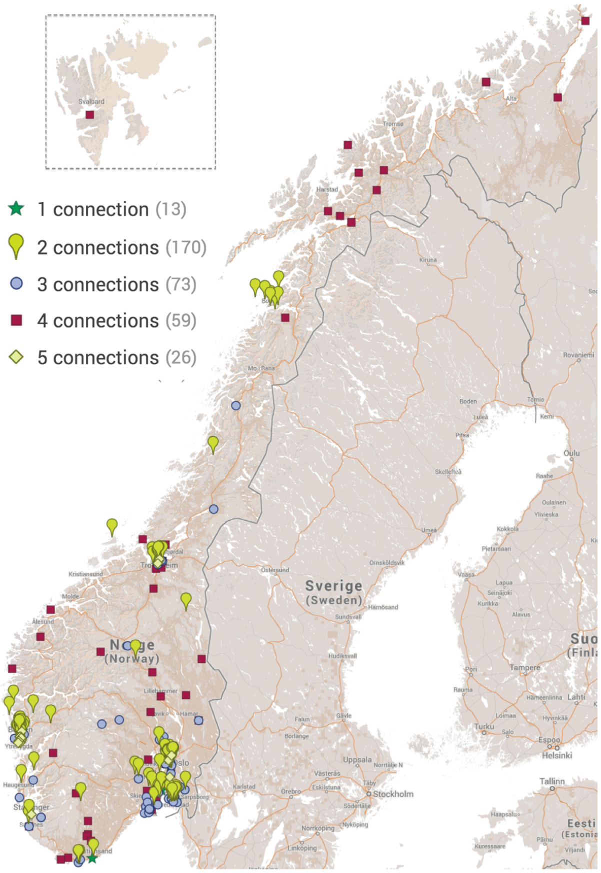


Figure 4: Placement of NNE nodes in Norway.

The NNE backend contains the server-side of the measurements, and is connected directly to the Norwegian research network UNINETT. The backend also contains servers for monitoring and managing the nodes, and for data processing.

3.2 Measured MBB networks

Nodes in the NNE platform are connected to up to five MBB networks. Four of these (Telenor, Netcom, Tele2 and Network Norway) are UMTS networks, while the last (Ice) is a CDMA2000 network operating in the 450 MHz frequency band. As shown in Fig. 5, Telenor and Netcom maintain their own nation-wide RAN. Tele2 and Network Norway are collaboratively building a third RAN, called Mobile Norway, which does not yet have nation-wide coverage. When outside of their home network, Tele2 customers camp on Netcom's RAN, while Network Norway customers camp on Telenor's RAN. This complex relation between the operators and RANs is an advantage for our measurement study. By looking at correlations between connections on the same RAN but in different operators (or vice versa), we can often determine whether an observed behavior is caused by the RAN or the CN.

3.3 Measurement experiments and data

The measurement experiments performed as part of this work are installed on the nodes using NNE's configuration management system. Measurements are then performed against measurement servers that are part of the NNE backend. The measurement servers are well provisioned in terms of bandwidth and processing power, to make sure they are not a performance limiting factor. Data from the measurements are uploaded to the backend database periodically. The data is post-processed to calculate aggregates and also to filter out time periods from problematic connections, NNE maintenance windows or when NNE experienced problems at the server-side due to hardware problems or problems with their network provider.

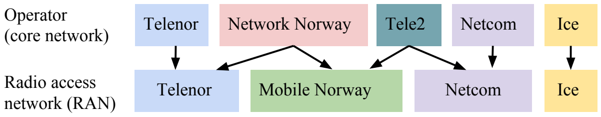


Figure 5: The operators and radio access networks measured in this study.

3.4 Metadata collection

The mode and RRC state of an MBB connection directly impacts its performance. To better explain the observed behavior, it is therefore important to collect state changes along with measurement results. The CDMA2000 gateway device provides only very limited diagnostic data, therefore we collect state information only for UMTS networks. The state attributes that are the most relevant to our measurements are connection mode (GSM/GPRS, WCDMA, LTE), connection submode (e.g. EDGE, WCDMA, HSPA+), signal strength (RSSI) and signal to noise ratio (E_c/I_o), RRC state and camping network operator. In addition, we also record when a connection comes up or disappears, i.e., when the PDP context is established or lost. As will be shown in the sequel, results can be very different depending on the network state.

All in all, our dataset consists of 10.1 billion entries in the database, gathered from 938 distinct connections at 341 distinct nodes. 327 of these are Telenor connections, 142 are Netcom, 75 are Tele2, 66 are Network Norway, and 328 are Ice[2]. The number of simultaneously active measurement nodes has varied in the range between 108 and 253 through the measurement period.

4. CONNECTION RELIABILITY

Data can only be sent over an MBB connection when there is an established PDP context in the CN. To establish a PDP context, the UE signals its presence to the respective signaling gateway (SGSN in UMTS networks), which then establishes the PDP context and returns a data session with an allocated IP address. This data session is essentially a tunnel connecting the UE to the Internet through intermediate gateways (GGSN in UMTS networks). The PDP context can be broken either by problems in the RAN (e.g., poor signal quality), or in the CN (e.g., failures or capacity problems in the SGSN). Failures can also be caused by the complex interaction between the OS running on the measurement node, the node's USB subsystem, and the MBB

[2] The varying number of connections per operator is caused by practical and economical constraints.

USB modem itself. We conjuncture, however, that if the majority of failures are caused by such artifacts, the differences between operators would be minor and hard to spot. In this section, we measure the frequency of PDP context losses, the time it takes before the PDP context is successfully restored, and the resulting downtime when no PDP context is available. We further correlate with signal quality and connection mode to gain an insight into what may have triggered the failure. The discussion in this section is limited to the UMTS networks, since we do not have the necessary logs from the CDMA2000 network.

4.1 Measuring connection failures

An NNE node continuously monitors the status of the PDP context for all UMTS connections, and tries to reestablish it as soon as it is broken. If it fails in doing that, the node keeps retrying until it eventually succeeds; we log all these attempts. There is no hold time between these consecutive reconnection attempts, so a new attempt is immediately initiated after the failure of the preceding attempt. A failure will therefore trigger a varying number of reconnection attempts depending on its duration (each attempt takes tens of milliseconds).

In some cases, the node manages to re-establish the PDP context for a short period, before it is again broken. To build a time series of failure events, we group consecutive reconnection attempts spaced by less than M minutes into the same event. In other words, a connection must keep its PDP context for at least M minutes before the reconnection was deemed successful and the failure event ends. Setting M to a high value underestimates the connection stability, while a low value will report a flapping connection at partially available. We experiment with different values for M in the range from 1 to 5 minutes. We detect a total of 154772 failures when setting M to 1 minute. Varying M from 1 minute to 3 minutes has a negligible impact on the number of detected failures. This number only drops by 0.019% when we set M to 3 minutes. It, however, decreases by 3% and 4.9% when we set M to 4 minutes and 5 minutes respectively. Based on this, we set M to 3 minutes when identifying PDP context failures. We believe that this grouping captures well what the user perceives as a usable connection, since a connection is not worth much if it flaps at a high frequency. The result of this grouping is a sequence of connection failure events of varying duration for each connection.

We impose two conditions to avoid overestimating the number and duration of connection failures by including measurement artifacts. First, we discard all failure events that were rectified either by rebooting the node or actively resetting the USB modem, since these may be caused by effects in the measurement platform. Second, to compensate for absent log files and failures that are not rectified by the end of our study period[3], we consider only failures that have well defined starting and ending points.

4.2 Analyzing connection failures

The stability of the tunnel that connects the UE to the CN depends largely on the RAN. Hence, to capture the effect of the radio access, we group our connections based on their respective RANs. Recall that, the measured four UMTS operators use three RANs as illustrated in Fig. 5. This gives us five logical networks in total, which are Telenor, Netcom, Mobile Norway (which includes Network Norway and Tele2 connections that use Mobile Norway's RAN), Network Norway@Telenor (which includes Network Norway connections that camp on Telenor's RAN), and finally Tele2@Netcom (which includes Tele2 connections that camp on Netcom's RAN). We use the camping information we collect from the modems to identify the connections that belong to the last three logical networks. For example, we classify a Network Norway connection as Network Norway@Telenor if it spends more than half of the time camping on Telenor's RAN, otherwise we classify it as Mobile Norway.

The three plots in Fig. 6 show the cumulative distribution function of the mean time between failures (MTBF), the mean time to restore (MTTR), and downtime percentage (due to PDP failures) for each connection in our data set, grouped by the five logical networks. We record distinct differences between operators, and observe a strong dependency between connection stability and the RAN. The statistics of Telenor connections and Network Norway@Telenor connections resemble each other. The same is true for Netcom connections and Tele2@Netcom connections. Although Mobile Norway is Network Norway's home RAN, the statistics of Network Norway@Telenor clearly differs from Mobile Norway's. The same is true for Tele2 and Mobile Norway, albeit to a lesser extent. This confirms the dominating role of the RAN in determining connection stability.

Differences between operators. Telenor and Network Norway@Telenor connections are less stable compared to the other three operators. About half of Telenor and Network Norway@Telenor connections fail at least once every day. For the other three operators this is the case for between one fourth (Mobile Norway) to one third of connections (Tele2@Netcom and Netcom). Telenor and Network Norway@Telenor, however, have much shorter MTTR compared to the other networks. Only 19% and 20% of Telenor and Network Norway@Telenor connections respectively have MTTR more than five minutes. The same numbers jump to 54% for Mobile Norway, 57% for Netcom, and 64% for Tele2@ Netcom. These differences suggest that the MTTR values for Netcom, Tele2@Netcom and Mobile Norway connections are influenced by a set of long lasting failures. To investigate whether these failures are the main factor behind the observed differences, we compute the median time to repair for all connections. While the median values are naturally smaller than the mean, the differences between operators remain consistent. For example, less than 9% of Telenor connections have a median time to repair longer than one minute compared to 33% for Netcom. Note that there are also slight differences, especially in the MTTR, between Network Norway@Telenor and Telenor. These differences can be attributed to the fact that many Network Norway@Telenor connections, though mainly camping on Telenor's RAN, spend some time camping on their home network as well. There are similar differences between Tele2@Netcom and Netcom but less pronounced. To check whether the observed difference between operators stem from varying coverage levels we measure the average RSSI for all connections. Figure 7 shows the CDF of mean RSSI for each connection in all operators. All curves collapse onto each other indicating no systematic differences between operators. The same is true also for E_c/I_o (not shown here).

[3] In some cases, measurement nodes were lost for varying periods of time which resulted in gaps in the logged data.

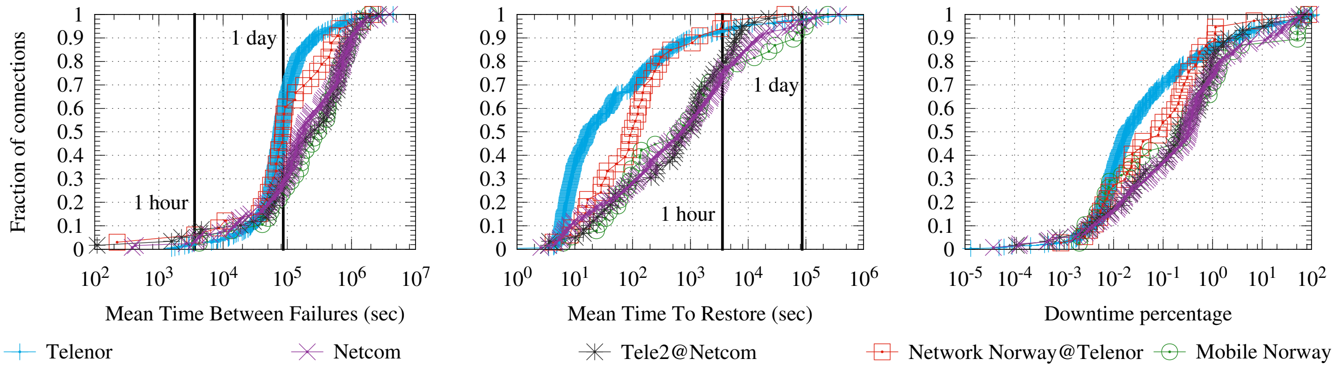


Figure 6: The statistics of connection failures.

Failure properties. Telenor and Network Norway@Telenor are dominated by frequent but short-lived failures compared to the other three networks. About half of Telenor and Network Norway@Telenor connections have MTTR less than 17 seconds and 90 seconds respectively. Looking closer at these short failures, we find that they are related to the RRC state of the connection, and they happen when the connection fails to be promoted from a shared channel (CELL_FACH) to a dedicated channel (CELL_DCH). This triggers the modem to reset the connection. As we demonstrate in Sec. 6, these short failures can have a drastic impact on applications performance. Netcom and Tele2@Netcom, on the other hand, have more long-lived failures that last for tens of minutes or even up to several hours. To gain a better insight into these long lasting failures, we investigate 157 failures in 27 distinct Tele2@Netcom connections which lasted for more than 1 hour. These connections are from NNE nodes that also have both a Netcom connection and a Telenor connection. Almost half of these failures (48.4%) affected the corresponding Netcom connections at the same time. The Telenor connections, however, remained stable. Hence, we feel safe that these long-lasting failures are not artifacts of our measurements. They seem rather related to the radio access availability, coverage, and possibly the interaction between the modems and the network. We plan to investigate the root cause of these failures in our future work.

Downtime. Telenor and Network Norway@Telenor connections have less overall downtime compared to the other networks. The percentage of connections experiencing more than 10 minutes of downtime per day ranges from 38% for Tele2@Netcom to 15% for Network Norway@Telenor. Failures that last more than 10 minutes are between 5.1% and 13.5% of all failures depending on the operator. They are, however, responsible for between 96.4% and 98.7% of the overall downtime. Besides characterizing the overall connection downtime, we also investigate how connection stability has varied during our study period. To this end, we calculate the median daily downtime percentage per network measured as the median downtime across all connections available on each day. Figure 8 shows the time series of this metric for all networks throughout the study period. For all networks, the median daily downtime remains stable hinting at no significant changes in connection stability during our measurement period. Further, the time series are in line with the observations we made earlier in this section. Networks that share the same RAN exhibit similar median daily downtime. Also, Telenor and Network Norway@Telenor are a characterized by a frequently observed median daily downtime of $5e-5$%, which corresponds to a single outage of 4.32 seconds. This higher downtime percentage for both networks is consistent with our observation that they suffer more frequent short-lived failures compared to the other networks.

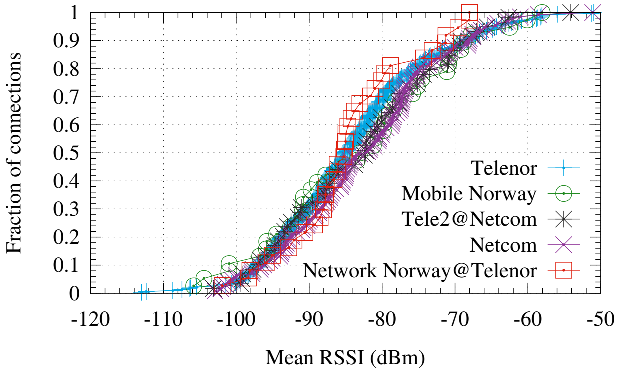


Figure 7: The CDF of the average RSSI per operator.

4.3 Correlating with metadata

To understand what may trigger the connection to be broken, we correlate the downtime due to PDP failures in a certain hour with the connection mode (e.g., 2G or 3G), the average RSSI, and the average E_c/I_o in that hour. To correlate with the connection mode, we say that a connection is in 3G (2G) mode in a given hour if it stays in 3G (2G) mode for at least 70% of its available time. Further, to construct a meaningful correlation with the signal quality, we group the average RSSI values into five categories that correspond to the standard mobile phones signal bars: 1 bar (-103 dBm or lower), 2 bars (-98 dBm to -102 dBm), 3 bars (-87 dBm to -97 dBm), 4 bars (-78 dBm to -86 dBm), and 5 bars (-77 dBm or higher). We also group the average E_c/I_o values

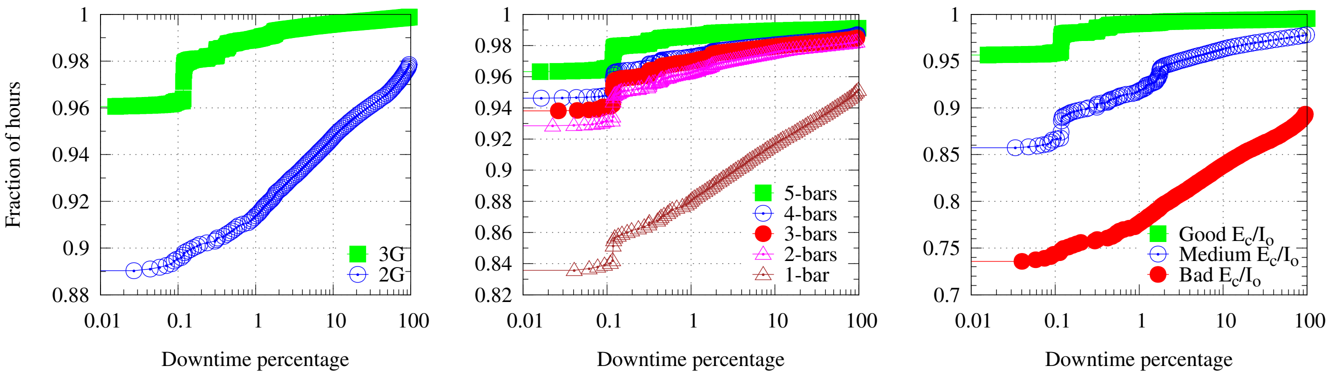


Figure 9: Downtime correlation with connection mode, RSSI and E_c/I_o.

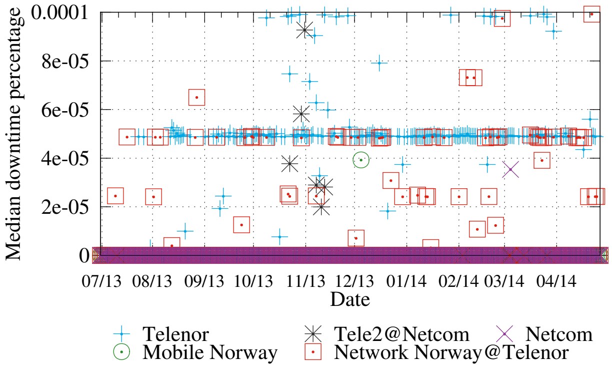


Figure 8: The daily mean downtime percentage for each MBB operator.

into three commonly used categories: Good (0 dB> E_c/I_o >-8 dB), Medium (-8 dB> E_c/I_o >-15 dB), Bad (-15dB> E_c/I_o <-33 dB) [10].

The left panel in Fig. 9 shows the CDF of the downtime percentage per hour, split according to the connection mode. This plot includes all connections from all operators. The center and right panels show the CDF of the downtime experienced at different RSSI levels and at different E_c/I_o levels respectively. The downtime is markedly higher for 2G connections than for 3G connections. We record 1% downtime or more in 1% of the hours with 3G connectivity compared to 9% of the hours with 2G connectivity. Further, as expected, downtime is influenced by the signal quality. Connections with an average RSSI equivalent to one signal bar have significantly higher downtimes. Beyond that, the differences between signal bar levels are less significant. Downtime also correlates strongly with E_c/I_o categories. We further observe that a sizable fraction of hours with downtime exists even when having good E_c/I_o or RSSI. For example, we experience downtime in 5% of the hours characterized by a good E_c/I_o. This indicates that radio quality can not alone explain all connection failures. To explain the stronger correlation between downtime and E_c/I_o as opposed to RSSI, we note that the RSSI measures the received signal strength which include all received components (i.e. signal and noise). Hence, a high RSSI does not necessarily translate into a good radio condition. E_c/I_o on the other hand measures the signal quality (i.e. the signal to noise ratio) capturing both interference with ambient and surrounding noise as well as interference from cross traffic at adjacent frequencies. We further correlate these three parameters to investigate whether a single measure is sufficient to describe the radio condition and consequently connection stability. Across operators, we do not observe clear correlation between RSSI and connection mode. Poor E_c/I_o, however, strongly correlates with RSSI of one bar as well as with 2G connectivity. This suggests that E_c/I_o can be picked as a predicator of connection stability.

The above explained correlations are visible in all operators, but the relation between downtime and metadata is not always linear. For instance, the correlation between different E_c/I_o categories and connection stability is more evident in Telenor and Network Norway@Telenor than in Netcom and Tele2@Netcom. This suggests that disconnects in Telenor and Network Norway@Telenor are more often caused by the radio conditions, matching well with the short MTTRs discussed above. While such failures also exist for Netcom and Tele2@Netcom, they are masked by the dominating long-lasting failures.

Summary of findings. The results presented in this section show that many connections have a downtime that can be problematic for critical applications, such as alarms or payment systems. 15-38% of connections are unavailable more than 10 minutes per day on average. There are also clear differences in connection stability between operators. While Telenor experiences relatively frequent but short-lived failures caused by the failure to acquire a dedicated radio channel, other operators have less frequent but longer lived failures giving a higher overall downtime. We further find that the connection level reliability is highly dependent on the RAN. In particular there is a high correlation between downtime and the signal-to-noise ratio of the connection. Still, there is also a significant number of connection failures that can not be explained by radio conditions. These may be caused by congestion or central failures in the network.

5. DATAPLANE RELIABILITY AND PERFORMANCE

This section looks at the networks' ability to deliver uninterrupted packet forwarding with an acceptable loss rate. Based on continuous end-to-end probing, we report on packet loss and on the duration of unavailable periods where no

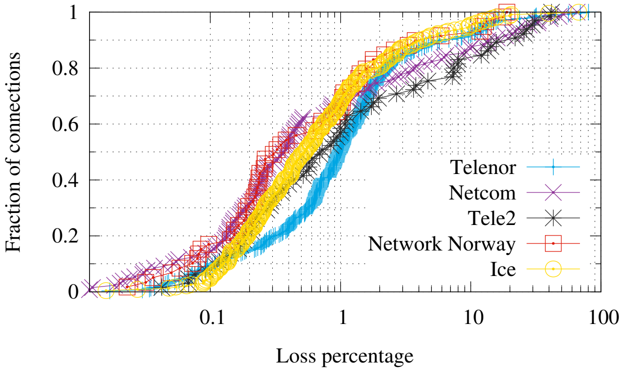


Figure 10: Loss rate for each MBB operator.

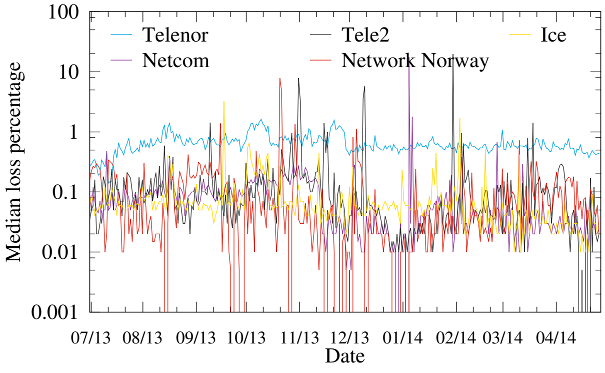


Figure 11: Median daily loss rate for each MBB operator.

packets come through. We look at how these metrics are related to mode, radio state and and signal quality. Finally, we identify particular events where a large fraction of connections simultaneously experienced abnormal packet loss rates.

5.1 Measurement description

Data plane reliability is measured by sending one 20 byte UDP packet to an echo server every second, and recording the reply packet from the server. A timestamp and an incremental sequence number is included in the packet payload for duplicate detection and round-trip time calculation. While the round-trip time normally is in the order of tens of milliseconds, we sometimes observe delays in the order of several seconds. Such high delays can sometimes be caused by excessive buffering [13]. We consider a packet to be lost if we do not receive a reply within 60 seconds.

This measurement test starts automatically on all network interfaces as they become active, and keeps running as long as the interface is up. When the connection is not available, the measurement script stops sending request packets and waits until the connection becomes available again. In total, more than 10 billion data points were collected from June 2013 to Apr. 2014. The measurement duration for each connection varies depending on how long the node was available and had working connections. In the following analysis, we require that we have at least 10 days of measurements to include a connection in the discussion. Due to space limitations we focus on packet loss in this section, and defer discussing delays and duplicates to future work.

5.2 Loss rate

The CDF in Fig. 10 shows the overall average loss rate for each connection in all operators. The loss rate is defined as (lost packets)/(sent packets) for the whole measurement period. Loss is relatively limited in all networks, and 50% of connections have less than 1% packet loss in all operators. We observe that relatively fewer Telenor connections have a very low packet loss rate compared to the other networks. 72% of Telenor connections have a loss rate higher than 0.5%, whereas this ratio is between 42 and 56% for the other networks. Telenor does not, however, have many connections with a high loss rate. Only 10% of connections have more than 5% loss, compared to 20% for Netcom and 23% for Tele2. Overall, Network Norway and Ice have the lowest packet loss rates.

There are diurnal patterns in packet loss in all networks, with higher loss rates in office hours when traffic is more intense. Ice, which has a larger fraction of home users, reaches their highest loss rates around 8PM. Packet loss in Telenor is consistently higher than in other networks throughout the day, also outside peak hours. We therefore believe that this higher packet loss is due to the RAN configuration rather than capacity limitations. To account for possible hardware, software or configuration changes over time in MBB networks, in Fig. 11 we plot the median daily loss rate for each MBB operator. We see that during the whole measurement period the median loss percentage remains stable and conforms with the results show in Fig. 10. Days with unusually high loss percentage are due to large events presented later in this section.

Networks differ in the thresholds they use to promote a connection from CELL_FACH to CELL_DCH. Telenor is more conservative than the other networks in such promotions [15], and hence the measured Telenor connections are more often in CELL_FACH. For instance, a Telenor connection spends on average 65% of its time in CELL_FACH compared to 34% for a Netcom connection. Unlike in the other networks, Telenor connections have a higher loss rate in CELL_FACH than in CELL_DCH, indicating that this channel is operating closer to its capacity limit[4]. Telenor does, however, have higher packet loss than other operators also when on CELL_DCH, so this can not alone explain the difference.

To further explain the observed differences, we have looked at loss rates combined with the metadata collected in parallel with the measurements. We first observe that loss rates are similar for 2G and 3G connections in all networks, with the exception of Netcom, where 2G loss is higher. A typical Netcom connection experiences more than 1% packet loss in 60% (25%) of all hours when it is on 2G (3G). Loss increases in periods when connections perform vertical handovers between 2G and 3G.

Not surprisingly, loss is higher in connections with poor signal quality. The Venn diagram in Fig. 12 shows that many

[4] Telenor has previously confirmed to us that they have had capacity issues in the FACH channel.

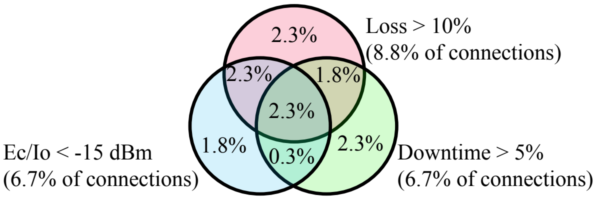


Figure 12: Loss, downtime and signal quality.

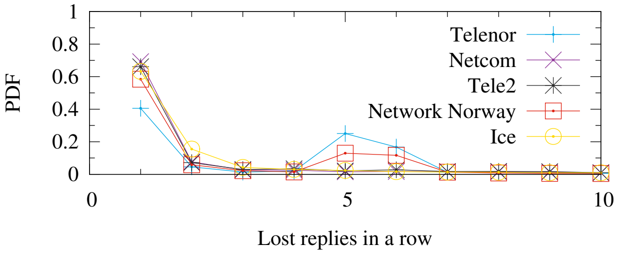


Figure 13: The distribution of loss run sizes across operators.

of the connections that experience a high loss rate also experience much downtime and have low E_c/I_o values. Out of the 341 connections where we have all the necessary metadata to make this comparison, 8.8% have an average loss rate higher than 10%. As seen in the figure, most of these (73%) have either a downtime ratio >5%, E_c/I_o <-15 dBm, or both. We can not identify any systematic differences between operators in the relation between loss, downtime and signal quality.

5.3 Loss runs

From our measurement data, we identify each sequence of packets that were lost in a row. We call such sequence a *l*oss run. Loss runs are important for how packet loss will affect the user experience. Since we send one packet every second, the size of a loss run is approximately equal to the number of seconds when no downlink traffic is received.

The distribution of loss run sizes is shown in Fig. 13. Not surprisingly, the dominating size of a loss run is just one packet, accounting for 60% of loss runs for three of the networks. Telenor and Network Norway, however, also have many loss runs of size 5 and 6. As explained in Sec. 3, Network Norway connections sometimes camp on Telenor's RAN, and we have confirmed that this is the case for connections with many loss runs of size 5 or 6. There is no clear periodicity in occurrence of loss runs of size 5 or 6, nor any dependence on signal quality, time of day or geographical area. Looking at the connection state during such loss runs, we find that they are normally caused by an RRC state demotion from CELL_FACH to IDLE. Re-establishing the CELL_FACH radio state and resuming packet forwarding takes 5-6 seconds, which is longer than an ordinary promotion [17]. While we do not know the exact cause of these demotions, we note that demotions can some times be caused by a network-initiated revocation of the radio access bearer. Such revocations can be triggered when capacity needs to be made available for other connections.

5.4 Large events

Next, we discuss *large events*, where many connections in an operator experiences abnormal packet loss at the same time. Such events will normally be caused by failures in the CN. To identify and analyze large events, we first divide our measurement time series into 5 minute intervals, and calculate the loss rate for each connection in each interval. Additionally, to account for downtime periods, all 5 minute intervals when a connection was unavailable are assigned a 100% loss rate. We consider that a connection has an abnormal loss rate in a particular 5 minute interval if more than 10% of the packets are lost. A large event is defined as a period of one or more intervals when more than 10% of all connections in an operator has abnormal packet loss.

Figure 14 shows a visual representation of all large events recorded during our measurement period. Each event is represented by a circle (multiple events in the same day are merged). The diameter of the circle reflects the severity of the event, and can be thought of as the total volume of lost traffic. This is calculated as the product of the fraction of affected connections, the average loss rate in the affected connections, and the duration of the lossy period. The fraction of affected connections is also represented on the y-axis.

As seen in the figure, networks experience large events with varying severity and frequency. Short-lived events with limited loss and affecting 10-20% of connections happen on a weekly basis in all networks. These events might be attributed to short-lived congestion, and may be considered part of normal operation. There are, however, also a number of larger events that can severely influence the user experience.

The collected measurements can normally give a good idea about the underlying cause for the more severe large events. By looking at the geographical distribution of the affected connections, the affected RAN(s), loss intensity and other parameters, we can often pin the root cause to either the transmission network, the CN, or to the interconnection between operators.

For example, on Sept. 10 2013, most Tele2 connections experienced 20-50% packet loss for around 40 minutes. Similar behavior repeated itself on Oct. 28 and on Nov. 16. These failures happened outside of maintenance windows, and affected connections from all parts of the country, and were therefore likely related to a component in the CN. Tele2 has later informed us that these events were probably caused by a failure in a load balancer in their CN.

One of the largest events recorded in our measurement period took place in Tele2 on Nov. 1-2, when 86% of connections were unavailable for more than 6 hours. 41% of the affected connections lost the PDP context during this period, while the others maintained a valid IP address but could not send or receive any data. The failure affected only Tele2 connections camping on Netcom's RAN, and Tele2 has confirmed that the root cause of the failure was a failing component in the peering between these two networks.

An interesting event took place on Oct. 21-22, and affected all Network Norway connections camping on Telenor's RAN. These connections experienced around 50% packet loss for more than 11 hours. We also observe that the packet loss rate was higher at times when traffic volumes are higher,

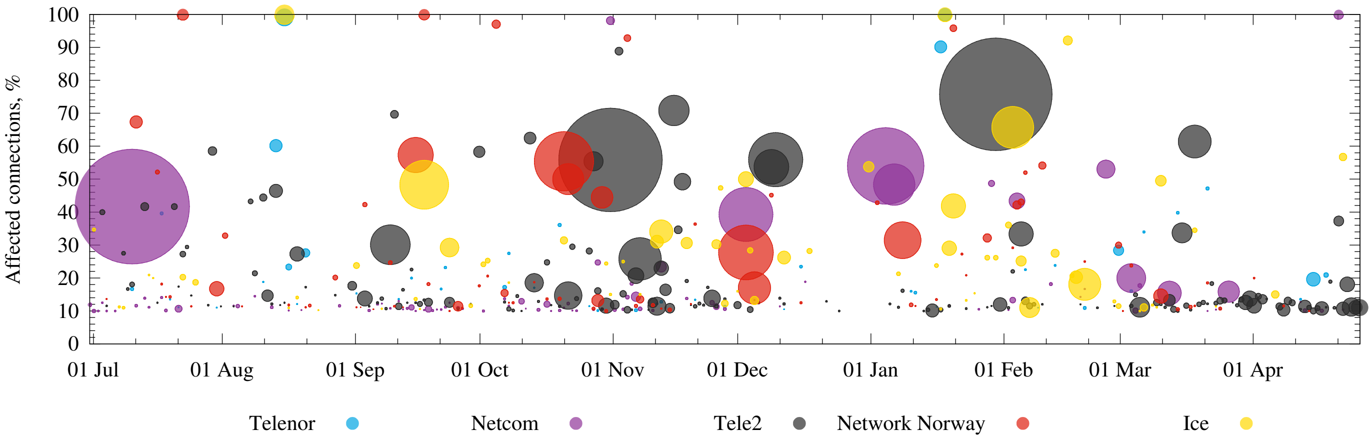


Figure 14: Large events since July 2013.

indicating a capacity problem in the connection between these two networks. Network Norway later confirmed that this event was caused by a routing problem that sent traffic through a roaming exchange point normally used for international roaming traffic only. This link is dimensioned for significantly lower traffic volumes, and could therefore not support the offered load.

In our discussions with the network operators, we have learned that several of the loss events identified in this study were not detected by their internal monitoring systems. These systems detect situations where many customers lose their Internet connections, but not necessarily situations where packet loss is less than 100%. This shows the potential of end-user measurements, and illustrates how they can help operators discover weaknesses and failures in their networks.

Summary of findings. In this section, we revealed clear differences in packet loss between operators. While Telenor has a non-negligible but low packet loss in most connections, other operators have higher variation in loss among connections. Telenor (and Network Norway@Telenor) also has a high number of loss runs of size 5 or 6, which tend to occur when connections lose their radio channel. Such short-lived gaps in packet forwarding can have a negative effect on applications such as interactive games, which often do not produce enough traffic to acquire a dedicated radio channel. Overall, the measured loss remains unexpectedly high since MBB networks rely heavily on link-layer retransmission to rectify errors. Our observations thus far, however, indicate that this unexpected loss is mainly related to state RRC transitions. We defer investigating the exact causes of loss to a followup study. Finally, we have used loss data to identify events that affect a significant fraction of connections in a single MBB network. An online alarm system based on end-to-end measurements that discovers such events can be useful for both operators and regulators.

6. APPLICATION LAYER RELIABILITY

An important aspect of reliability is a stable performance on the application layer. In this section, we look at the stability in performance of two representative applications: HTTP download and VoIP using the SIP protocol.

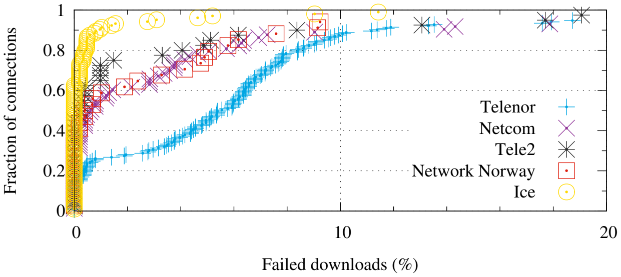


Figure 15: Failure rates in HTTP download tests.

6.1 HTTP download test

Much of the traffic in MBB networks goes over the HTTP protocol, which is used for both web browsing and streaming. Here, we report results from a simple experiment where a 1 MByte file is downloaded from a server using the HTTP GET method. For each test iteration, we record the time it took to start the data transfer, the total time to complete the download (if successful), throughput, and any error code. For each connection, the test was repeated once per hour for a duration of 3 weeks, giving a total of up to 504 runs per connection. Based on these measurements, we report on two different metrics: the probability of successfully completing the transfer, and the probability of achieving a download rate of at least 1 Mbps.

Note that our aim here is to look at the stability of the application layer performance, and not to measure the maximum achievable throughput in the connection, which would require a different approach.

Figure 15 shows the fraction of download attempts that could not be successfully completed. We observe that the fraction of failing attempts is significantly higher in Telenor than in the other networks. 55% of Telenor connections experience a failure rate higher than 5%, and 12% experience a failure rate higher than 10%. Ice, on the other hand, sees very few unsuccessful downloads, with only 9% of connections having a failure rate above 1%.

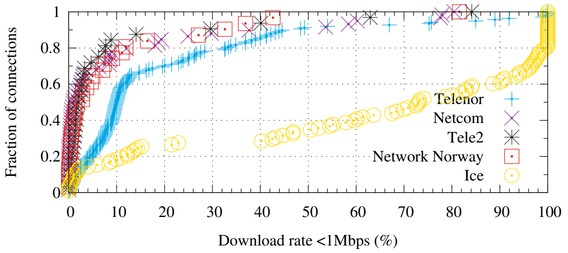


Figure 16: Probability of achieving less than 1Mbps HTTP throughput.

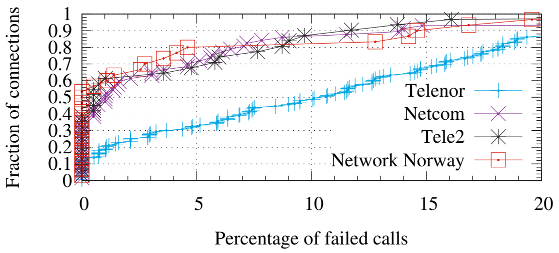


Figure 17: Failure rates in VoIP tests.

Looking at the error codes for the unsuccessful transfers, we find that the dominating reason for a failed download in all networks except Telenor is the failure to establish the connection. Correlating erroneous transfers with the data from our connectivity test we observed that these unsuccessful TCP handshakes happen during times with high packet loss. For Telenor, however, the connection broke *after* the initial TCP handshake in 74% of the failing attempts. Such errors are largely responsible for the difference between Telenor and the other networks. Looking closer at these events, we find that they happen when the connection can not be promoted from the CELL_FACH to the CELL_DCH state. The download traffic will normally trigger such a promotion, but when this does not happen, the modem responds by resetting the connection, and the download fails. There is a clear diurnal pattern in this type of failures, which makes it natural to believe that they are related to congestion in the FACH channel and/or insufficient resources for a promotion to CELL_DCH.

Figure 16 shows, for each connection, the fraction of runs where the achieved throughput was less than 1 Mbps. Only 3G connections are included in this plot. We observe that in all UMTS operators, most connections achieve this download rate most of the time. In Netcom, Tele2 and Network Norway, around 90% of connections achieve at least 1 Mbps 90% of the time or more. Telenor achieves this rate less often. We believe that this is caused by the higher loss rate observed in Telenor and how it interacts with TCP congestion control. Ice achieves far lower throughput than the UMTS operators. Around 20% of the connections never achieve the target download speed, while only 19% of connections meets the target in 90% of the cases. From Figs. 15 and 16, we conclude that Ice offers a slower but more stable download performance than the UMTS networks.

6.2 Voice-over-IP test

Initiating VoIP calls over MBB connections is becoming increasingly popular, thanks to the rise of applications such as Viber [24]. The ability to initiate and complete VoIP calls are therefore important aspects of the user-experienced reliability. To asses this ability we design a simple experiment that faithfully emulates a real VoIP call. Our test consists of a custom-made client that runs on the measurement node and an Asterisk PBX [6] hosted at the NNE backend. The client initiates a VoIP call to the PBX using SIP, then it uses RTP to play a one minute long audio file. Upon completion the PBX replays the same file back to the client and terminates the call. The audio file is encoded as G.711 [1] resulting in a sending rate of about 50 packets per second. For each connection, we run this experiment once per hour for one week, giving us 168 calls per connection.

Figure 17 illustrates the fraction of calls that could not be successfully completed. We observe that the failure rate is significantly higher for Telenor. 30% of Telenor connections have a failure rate higher than 15% compared to only 4% of Tele2 connections. 34% and 21% of the failures for Netcom and Tele2 respectively happened during the call initiation phase (i.e. the SIP INVITE fails). For Telenor and Network Norway, however, this percentage drops to 13% and 14.9% respectively. The remaining failures happened after the call started successfully. We believe that the explanation for these dropped calls is the same as for the unsuccessful HTTP downloads discussed above.

Summary of findings. This section has shown that short pauses in packet forwarding that are likely caused by the lack of available radio resources can lead to significant problems for applications, such as dropped calls and failed downloads.

7. ROBUSTNESS BY MULTI-HOMING

So far, our discussion has focused on the experienced reliability in each MBB network separately. A related question is how much reliability can be increased if an end device can be simultaneously connected to more than one MBB network. Such multi-homed devices are becoming increasingly popular, for example as mobile gateways that provide WiFi service on public transport.

The potential for increased reliability through end device multi-homing depends heavily on whether coverage and failure patterns are largely independent across operators. If connections to different operators tend to fail simultaneously due to the same underlying reasons, there is little to be gained from a redundant connection. Our discussion in this section focuses on three important aspects of cross-correlation between operators: coverage, connection failures and loss events.

7.1 Correlations in coverage

We first look at whether different operators are similar with respect to signal quality at a given location. NNE measurement nodes are distributed widely across Norway, and we believe they give a reasonable representation of the indoor signal quality that can be expected in each operator.

Figure 18 gives an overview of the E_c/I_o values for all UMTS connections in all nodes in our dataset. The values shown are averages across the whole measurement period (for most connections the variation is small). Since almost

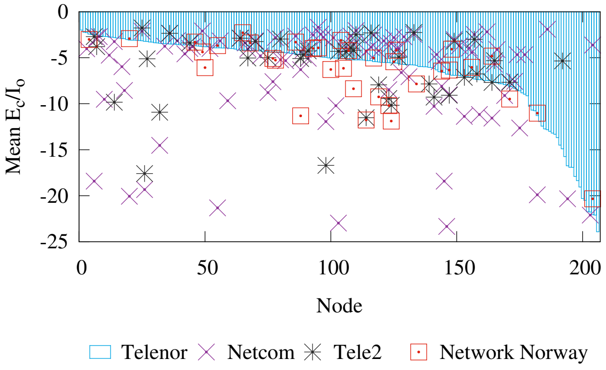


Figure 18: Average E_c/I_o values for connections in different operators.

all nodes have a Telenor connection, we sort the nodes based on the E_c/I_o of the Telenor connection, and plot E_c/I_o values for the other connections where available.

We observe that there does not seem to be a strong correlation between operators in general. As expected, we find that for some nodes, Netcom/Tele2 and Telenor/Network Norway have pairwise very similar E_c/I_o due to their national roaming arrangements. Calculating the Pearson correlation coefficient for each pair of operators confirms the visual impression from Fig. 18. The correlation coefficient for the pairs of operators that do not share the same RAN is between -0.10 (Netcom and Network Norway) and 0.25 (Telenor and Netcom). The correlation is evidently higher when the respective pair of operators share the same RAN, that is 0.47 for Tele2 and Netcom and 0.75 for Telenor and Network Norway. We also performed a similar analysis for RSSI values, which gave similar results. These findings are positive from a robustness perspective, since they indicate that there is a significant potential gain from provider multihoming.

7.2 Correlations in downtime

Next, we revisit the connection failures discussed in Sec. 4, to see how downtime could be reduced if the end user is able to connect to more than one operator. For each pair of connections from the same node, we identify the time periods when both connections were unavailable (i.e., they had no PDP context). The resulting downtime represents a lower limit on the downtime that can be achieved if the end system is able to exploit both connections in parallel.

Figure 19 shows the downtime for three operators (repeated from Fig. 6), and the combined downtime for all pairs of connections from the same operators. Other operators and pairs are excluded from the graph for readability. For comparison, we also plot the expected combined downtime for connection pairs under the (false) assumption that they fail independently. This value is calculated by multiplying the downtime rates of each connection, and serves as a lower limit on the downtime we can expect by combining operators.

We make two main observations. First, there is a large potential for reducing downtime through multi-homing. 60% of the nodes achieve 99.999% uptime when combining Telenor and Tele2 connections; and 55% of the nodes achieve the same when combining Netcom and Telenor connections. These reductions in downtime are arguably surprisingly high, given that connection from different operators often share many potential sources of failures, such as local weather conditions, cell towers, power, or even transmission. We see that the measured combined downtime is not very different from the theoretical downtime assuming independence. Second, the reduction depends heavily on selecting the right pair of connections. As expected, operators that often camp on the same RAN show a much larger correlation in their failure pattern, and the gain from combining such connections is limited. As shown in Fig. 19, there is often little or no gain in combining connections from Netcom and Tele2.

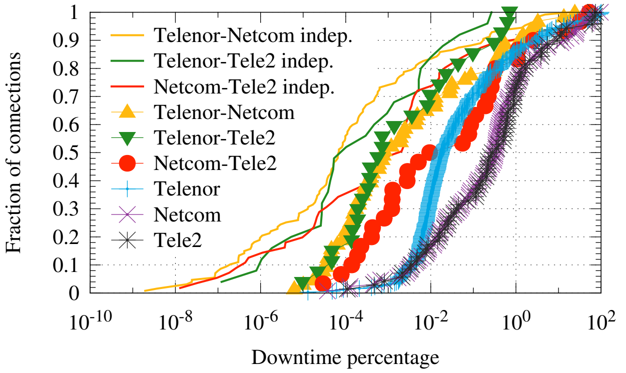


Figure 19: Common downtime across operators.

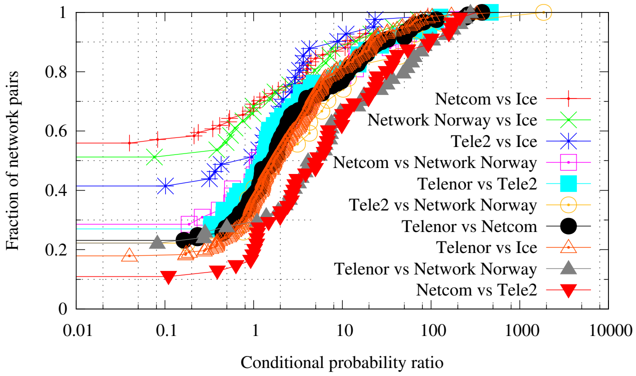


Figure 20: Conditional probability of loss events.

7.3 Correlations in loss

Finally, we look at correlations in loss between networks. We base our analysis on the same 5 minute intervals used in Sec. 5. Let $P(A)$ denote the (empirical) probability that a connection A has a loss rate higher than 10% in a given 5 minute interval, and $P(B)$ be the same for a connection B. We calculate the conditional probability $P(A|B)$ for each pair of connections from the same node, and compare it to the unconditional probability $P(A)$. If the conditional probability ratio $R = P(A|B)/P(A)$ is close to 1, it means that connection A and B fails largely independent, while a high R means that they tend to fail together. Note that by Baye's law, $P(A|B)/P(A) = P(B|A)/P(B)$.

Figure 20 shows R for all pairs of connections at the same node, grouped by operators. We require that A and B have at least one full week of overlapping measurements to be included in the analysis. Note that the number of events where both connections experience high loss may in some cases be very low, so the graph should be interpreted with some care. Some observations are still clear. First, connections from different networks are not completely independent. In between 25 and 70% of connection pairs, the likelihood of high packet loss in connection A more than doubles when connection B experiences high loss. In between 8 and 35% of the cases, the likelihood increases more than 10 times. There are, however, clear differences between operator pairs. Not surprisingly, the strongest pairwise dependence is for Netcom/Tele2 and Telenor/Network Norway. The weakest dependence is between Ice and the other operators. This might also be expected, since Ice operates with a different technology, a much lower frequency band, and has a different customer mix (and therefore traffic pattern) than the other operators.

Summary of findings. The results in this section indicate that there is a large potential for increased reliability through multi-homing to several MBB networks. There are generally clear differences between operators in signal quality at a given location, and exploiting two connections in parallel can potentially give 99.999% availability in many cases.

8. RELATED WORK

Mobile broadband measurement. During the past few years, there have been a growing interest by regulators, policy makers and the networking community in measuring the performance of home and mobile broadband networks. Several regulators have translated this into ongoing nationwide efforts, examples of such efforts include US FCC's Measuring Broadband America initiative [7] and an annual activity by the Communications Regulatory Authority of Lithuania for measuring MBB performance [5]. A study by Sundaresan et. al [23], partly based on the FCC data, demonstrated that the long-term continuous monitoring from the edge is indispensable for understanding and assessing home broadband performance. Our work is the first to present a country-wide assessment of MBB reliability filling an important gap in this area.

Approaches for measuring MBB networks can be classified into three categories:

1. Crowd-sourced user initiated measurements. The most prominent example in this respect is Mobiperf [2], an Android application that measure several performance metrics including throughput and latency. Nikravesh et al. [16] used an extensive data set contributed by the users of two apps, Mobiperf and Speedometer, and performed a longitudinal analysis of MBB performance. They highlighted significant performance differences across operators, access technologies, and regions. They also showed that the performance of MBB is not improving overtime, which necessitates the continuos monitoring of MBB performance. Sommers and Barford used crowd-sourced data from Speedtest.net to compare the performance of MBB to WiFi [21]. Approaches based on user-initiated tests can complement measurements from dedicated infrastructures such as NNE. It is difficult, however, to rely on such tests for the continuous monitoring of MBB networks stability.

2. Measurements collected using dedicated infrastructure. Examples of such studies include [18] and [14]. The former used a few laptops mounted on public buses to compare the performance of three MBB operators, while the latter used low cost notebooks for comparing the performance of four operators in seven locations in India. Our work falls into this category. It is, however, based on a much larger deployment in terms of the number measured operators, the number of connections, geographical distribution and duration.

3. Measurements based on network-side data. Several measurements used network side logs to assess various aspect of MBB performance. Examples include, assessing RRC state machine impact on performance [17], characterizing the properties of cellular traffic [20], HTTP performance [9], performance during crowded events [19], and TCP performance over LTE [11]. Such logs are however only available to operators. This line of work and ours complement each other. Network-side data gives more insights into the network internals, while end-to-end measurements can help operators detecting anomalies that are not easily identifiable by only using network-side monitoring, as we demonstrated in Sec. 5.

Measurement framework and metrics. Several ongoing efforts are aiming to standardize the task of performing large scale measurements [3] and to define meaningful performance metrics for IP networks in general [22] and MBB networks in particular [12]. Sundaresan et. al [23] investigated the suitability of different metrics in characterizing home broadband performance, while Goga and Teixeira explored several approaches for estimating broadband access speed [8]. This paper presents a framework for assessing MBB reliability using a large scale deployment. We believe that our work is a timely input to the ongoing efforts to defining approaches and metrics for measuring MBB networks.

9. DISCUSSION AND CONCLUSIONS

This work has presented a framework for measuring reliability in MBB networks, based on the position that end-to-end measurements can give useful insights about performance and stability problems in the network as a whole. The main argument in the proposed approach is that reliability must be measured at several levels, from the stability of the network connection to the reliability of the data plane and application layer performance. We believe that this framework gives a good basis for describing the overall reliability of an MBB network. In the future, this should also be extended with measurements that capture the effect of mobility on the experienced reliability. Efforts are ongoing to extend the NNE infrastructure with mobile nodes, which would allow such measurements.

Using the proposed framework, we have presented a large-scale study of reliability in 5 Norwegian MBB networks. We have focused on a few selected metrics at each level in the framework. There are still many other metrics that can be relevant for understanding reliability. On the connection level, the ability to *establish* a PDP context when needed is an important aspect of reliability, which is different than the ability to maintain the connection uninterrupted for a long time. On the data plane level, further analysis can be made to describe the pattern of packet loss and delay variations under varying network conditions. An important topic is also to look at the reliability and performance of various transport layer protocols in MBB networks.

The measurements presented here have demonstrated that there are clear differences in reliability between operators, and that these can be identified and characterized by end-to-end measurements. Networks vary in the stability of connections, in packet loss patterns, and in their ability to support popular applications. We have also shown how end-to-end measurements can be used to identify failures and performance problems that are not necessarily captured by the operators' monitoring systems.

This study was performed on the NNE infrastructure, with dedicated measurement nodes. The framework is, however, also applicable for studies based on crowd-sourced data from mobile phones. Such approaches will, nevertheless, often be more limited in the availability of metadata, and in the ability to gather long uninterrupted time series under stable conditions.

This study has some limitations that we hope to address in future work. First, only 2 different 3G modem models are used in this study. While this makes it possible to compare connections across operators and regions with the same equipment, it may also introduce effects that we would not see with other types of user terminals. In the future, we hope to include also a few other modem types in our studies. Further, the UMTS modems used in the NNE platform support 3G protocols up to DC-HSPA ("3.75G"), but not LTE/4G. As LTE becomes more widespread, it will be important to measure also these networks.

10. ACKNOWLEDGEMENTS

We thank Amogh Dhamdhere and the anonymous reviewers for their insightful comments on this draft. Nornet Edge is financed by the Research Council of Norway through the INFRASTRUKTUR program. Nornet Edge is realized with additional generous support from Tele2, Ice.net.

11. REFERENCES

[1] ITU-T recommendation G.711. Pulse code modulation (PCM) of voice frequencies, 1988.

[2] Mobiperf. `http://www.mobiperf.com`, 2014.

[3] M. Bagnulo, P. Eardley, T. Burbridge, B. Trammell, and R. Winter. Standardizing Large-scale Measurement Platforms. *SIGCOMM Comput. Commun. Rev.*, 43, 2013.

[4] Cisco Systems, Inc. *Cisco visual networking index: Global mobile data traffic forecast update, 2012 - 2017*, February 2013.

[5] Communications Regulatory Authority of the Republic of Lithuania. *Annual report of the Communications Regulatory Authority (RTT) of the Republic of Lithuania*, 2012.

[6] Digium. Asterisk. `http://www.asterisk.org/`.

[7] FCC. 2013 Measuring Broadband America February Report. Technical report, FCC's Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, 2013.

[8] O. Goga and R. Teixeira. Speed Measurements of Residential Internet Access. In *Proc. of PAM*, 2012.

[9] E. Halepovic, J. Pang, and O. Spatscheck. Can you GET me now?: Estimating the time-to-first-byte of HTTP transactions with passive measurements. In *Proc. of IMC*, 2012.

[10] H. Holma and A. Toskala. *WCDMA for UMTS: HSPA Evolution and LTE*. John Wiley & Sons Ltd., 4th edition, 2007.

[11] J. Huang, F. Qian, Y. Guo, Y. Zhou, Q. Xu, Z. M. Mao, S. Sen, and O. Spatscheck. An In-depth Study of LTE: Effect of Network Protocol and Application Behavior on Performance. In *Proc. of SIGCOMM*, 2013.

[12] IEEE. P802.16.3 Project: Mobile Broadband Network Performance Measurements. `http://www.ieee802.org/16/mbnpm/index.html`.

[13] H. Jiang, Y. Wang, K. Lee, and I. Rhee. Tackling bufferbloat in 3g/4g networks. In *Proc. of IMC*, 2012.

[14] Z. Koradia, G. Mannava, A. Raman, G. Aggarwal, V. Ribeiro, A. Seth, S. Ardon, A. Mahanti, and S. Triukose. First Impressions on the State of Cellular Data Connectivity in India. In *Proceedings of the 4th Annual Symposium on Computing for Development*, ACM DEV-4 '13, 2013.

[15] A. Kvalbein, D. Baltrūnas, J. Xiang, K. R. Evensen, A. Elmokashfi, and S. Ferlin-Oliveira. The Nornet Edge platform for mobile broadband measurements. *Elsevier Computer Networks special issue on Future Internet Testbeds*, 2014.

[16] 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 *Procs. of PAM*, 2014.

[17] F. Qian, Z. Wang, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. Characterizing Radio Resource Allocation for 3G Networks. In *Proc. of IMC*, 2010.

[18] S. Sen, J. Yoon, J. Hare, J. Ormont, and S.Banerjee. Can they hear me now?: A case for a client-assisted approach to monitoring wide-area wireless networks. In *Proc. of IMC*, 2011.

[19] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang. A first look at cellular network performance during crowded events. In *Proc. of SIGMETRICS*, 2013.

[20] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang. Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices. In *Proc. of SIGMETRICS*, 2011.

[21] J. Sommers and P. Barford. Cell vs. WiFi: On the Performance of Metro Area Mobile Connections. In *Proc. of IMC*, 2012.

[22] M. Stiemerling. IP Performance Metrics charter-ietf-ippm-05. `http://datatracker.ietf.org/doc/charter-ietf-ippm/`, 2013.

[23] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescapè. Broadband Internet performance: A view from the gateway. *SIGCOMM Comput. Commun. Rev.*, 41:134–145, 2011.

[24] Viber-Media. Viber. `http://www.viber.com/`.