Modeling regional reliability of 2G, 3G and 4G mobile data networks and its effect on the German automatic tolling system

Bernd Pfitzinger1,2, Tommy Baumann3, Dragan Macos4, Thomas Jestädt1
1Toll Collect GmbH, Linkstr. 4, 10785 Berlin, Germany. Email: bernd.pfitzinger,thomas.jestaedt@toll-collect.de
2FOM Hochschule für Oekonomie & Management, Zeltnerstraße 19, 90443 Nürnberg, Germany.
3Andato GmbH & Co. KG, Ehrenbergstraße 11, 98693 Ilmenau, Germany. Email: tommy.baumann@andato.com
4Beuth Hochschule für Technik, Luxemburger Str. 10, 13353 Berlin, Germany. dmacos@beuth-hochschule.de

Abstract

Using an existing large-scale realistic discrete event simulation model of a mobile distributed system we model the infrastructure and dependencies of 2G, 3G and 4G mobile data networks. We investigate the effects of the different network generations and of transitioning connections between networks on the typical behaviour of the distributed system. In particular we enable the model to simulate regional outages of networks to predict the effects of shared back-end network components and the effects of mobile endpoints entering and leaving regional coverage.

1. Introduction

The ubiquitous availability of mobile data networks (both temporal and spatial) is often assumed in the design of distributed systems. However, over the years most major mobile networks were subject to outages ranging from local and short-term to substantial network-wide problems. While substantial network problems are rare [1] the consequences could be far-ranging. Looking beyond the personal and individual use of mobile devices we are particularly interested in the effect of the network availability on the performance of machine-to-machine systems that rely on mobile networks for their ongoing operations. In particular we investigate the dependency of the German automatic tolling systems on the mobile data network and possible benefits of using several different network generations (2G, 3G and 4G).

In our case, local network outages are of no particular concern – since the typical use case of an automatic tolling system is that the mobile endpoint (on-board-unit, OBU) is traveling along the road network and will therefore quickly leave behind any local outages. Yet, poor network connectivity will lead to an increase in retries required to complete the communication – which in turn leads to a higher server load in the central system, more data transmitted and a substantial increase in latency.

The performance of the tolling system could only be affected by regional or national network outages. In this article we use an existing simulation model of the German tolling system [2] to analyze the dependency on the mobile data network and the effects of regional and national network outages. Section 2 introduces the interaction between the tolling system and the mobile data network which is followed in section 3 by a discussion of the 2G, 3G and 4G core network architectures. In particular we look at the steps taken to achieve a high level of availability and the ability to increase availability through national roaming. Section 4 summarizes the changes introduced to the existing simulation model. Taking this model section 5 discusses the effect of regional network outages on the performance of the tolling system.
tolling system and possible ways to mitigate network outages.

2. System overview: The role of mobile data networks in GNSS-based tolling

Road user charging is well established in many countries using a wide variety of charging models and technologies [3]. Regardless of the technical architecture tolling systems are liability-critical systems and must therefore be dependable (in the sense of [4]): Tolling has to be accurate across the whole road network and at all times. From an ecological point-of-view any tolling system should be ‘free-flow’ i.e. work without impacting the flow of traffic. One type of (modern) electronic tolling systems is based on global navigation satellite systems (GNSS) to determine the tolls based on the factors distance, time or place.

The emergence of nationwide tolling schemes on large road networks (e.g. in Europe) led to an increasing interest in GNSS-based tolling systems [5]: These systems operate without road-side equipment and scale economically to large and complicated road networks. The fundamental process of a GNSS-based tolling system is of course the combination of the vehicle location determination (by an OBU using GNSS), the calculation of the tolls and the communication between the OBU and the other parts of the tolling system [6]. At this level of abstraction there exist already two distinct implementations (see figure 1): Is the toll calculation performed by the OBU (“thick client”) or by a backend-system (when the OBU becomes a “thin client”)? The first architecture can be observed in the German automatic tolling system for heavy-goods vehicles (HGVs) [2] and the second one in the newly constructed French system (eco-tax), [7]. Of course, both architectures rely on a mobile data network – typically GSM using 2G or 3G networks – to transfer data between the OBUs and the backend systems. They differ in the amount of data necessary to be transferred: Either the results of the toll calculation (and infrequent updates of the geo and tariff data residing on the OBU) or a sequence of location data with sufficient geographical and temporal resolution to allow the subsequent toll calculation in the backend system.

The dependency of a GNSS-based tolling system on the availability of the mobile data network can be mitigated by buffering toll data on the OBU for later transmission thereby introducing considerable latency to the tolling process: Toll charging and billing run asynchronously. The “thick-client” solution depends on timely updates to its geo and tariff data (in the German toll system typically once every three months) where the “thin-client” generates considerably more data to be buffered locally. We note in passing that the thin-client would be unable to display the toll charge in real-time when the mobile data network is not available. To summarize, when toll data can be buffered the OBU can wait for a considerable time period (days to weeks) to (re-)establish network connectivity. Considering that HGVs are commercial, professional users it is obvious that they will typically move far enough during that time period to leave any local network outages behind. In effect the toll system depends on the availability of the core network of the mobile data network. The typical configuration of a core network with respect to the network availability is discussed in the next section.

3. The architecture of 2G, 3G and 4G networks

This section discusses the core network architecture of the three different GSM generations. In particular we look at the way in which mobile network operators ensure the availability of the network at a national or regional level. From the perspective of the tolling system the availability of mobile data connections is most easily improved through roaming i.e. the ability to connect to networks of different operators. These networks are operated independently in most respects (due to regulation) and share only some infrastructure [8]. However, at the moment national roaming is typically not offered in the EU [8].

![Figure 2: Simplified architecture of a 2G/3G packet data core network](image-url)
Looking at a single operator we investigate whether the use of different network generations increases network availability. In fact, the architecture of the core network differs considerably between the different generations. In the packet data domain of a 2G or 3G network the base transceiver stations are still controlled by base station controllers (BSC) as it is the case in the voice domain of the 2G network (figure 2, based on [9]). The core of the packet data network consists of two types of servers: The Serving GPRS Support Node (SGSN) and the Gateway GPRS Support Node (GGSN). The GGSN controls the network resources (e.g. PDP addresses and routing information) and is the interface to other networks (e.g. the internet). In contrast the SGSN manages the mobile station (e.g. the location within the cellular network and security or authentication) for a given set of BSCs i.e. within a certain region. In addition the Home Location Register (HLR) holds the subscriber information that is required for a mobile station to access the network (another register is necessary for visiting mobile stations, not shown in figure 2).

A typical network uses many BSCs and Base Transceiver Stations (BTS) and it is common for equipment to fail locally (one study found the MTTF of the BTS hardware to be several years and for the BTS and BSS software to be less than a year [1]). The use of redundant components can lead to availabilities better than 99.999% [10], [11].

Considering regional and national network availability we look at the SGSN, GGSN and HLR. The setup used in real-world networks is to our knowledge not public. However, looking at the data sheets of typical equipment (e.g. the Cisco ASR 5000 [12] or the Alcatel-Lucent 7500 SGSN [13]) we can estimate some properties: A typical SGSN can manage millions of subscribers or PDP contexts – the major German networks would need on the order of 10 units each to manage some 30 million subscribers. In reality the number should be higher due to spare capacities and the use of a high-availability architecture. With a system availability (of one SGSN) of better than 1000 years (mean time to failure of the hardware platform [12]) outages are rare and in the case of the SGSN limited to a region. In fact, on all levels of the core network units can be configured to take over in case of a failure.

The HLR as the central registry of subscriber information is the one component that may lead to a complete failure of the mobile data network. It is of course expected that the HLR is multiply redundant and the availability accordingly very high. However, HLR faults have been experienced in the past [14]. Unfortunately the HLR is serving both the voice and the data domain. In our case the German tolling system has a fallback option to use SMS (instead of GPRS network connections) with a proprietary application-level protocol between OBU and the backend systems. The use of SMS can mitigate any outages of the GPRS network but not of the HLR.

The 4G networks differ considerably (figure 3, based on [15]): The network handles the voice and data domains in the same way (as data) using only four different network elements: The Serving Gateway (SGW) and the Packet Data Network Gateway (PDN GW) on the user plane and the Mobility Management Entity (MME) and Home Subscriber Server (HSS) on the control plane.

In a sense the network logic moves closer to the User Equipment (UE). The tasks of the SGSN are typically taken over by the base stations (eNodeB) together with the MME. The SGW provides the IP-tunneling between PDN GW and the UE. Handovers are typically handled by the eNodeBs. Only when the two eNodeB involved in the handover are served by different MMEs the SGW will be involved.

In the same way the PDN GW is similar to the GGSN and the HSS to the HLR. We note that a 4G network implements its own HSS, i.e. in theory the central role of the HLR (spanning 2G and 3G

![Figure 3: Simplified architecture of a 4G network](image)

Considering regional and national network availability we look at the SGSN, GGSN and HLR. The setup used in real-world networks is to our knowledge not public. However, looking at the data sheets of typical equipment (e.g. the Cisco ASR 5000 [12] or the Alcatel-Lucent 7500 SGSN [13]) we can estimate some properties: A typical SGSN can manage millions of subscribers or PDP contexts – the major German networks would need on the order of 10 units each to manage some 30 million subscribers. In reality the number should be higher due to spare capacities and the use of a high-availability architecture. With a system availability (of one SGSN) of better than 1000 years (mean time to failure of the hardware platform [12]) outages are rare and in the case of the SGSN limited to a region. In fact, on all levels of the core network units can be configured to take over in case of a failure.

The HLR as the central registry of subscriber information is the one component that may lead to a complete failure of the mobile data network. It is of course expected that the HLR is multiply redundant and the availability accordingly very high. However, HLR faults have been experienced in the past [14]. Unfortunately the HLR is serving both the voice and the data domain. In our case the German tolling system has a fallback option to use SMS (instead of GPRS network connections) with a proprietary application-level protocol between OBU and the backend systems. The use of SMS can mitigate any outages of the GPRS network but not of the HLR.

The 4G networks differ considerably (figure 3, based on [15]): The network handles the voice and data domains in the same way (as data) using only four different network elements: The Serving Gateway (SGW) and the Packet Data Network Gateway (PDN GW) on the user plane and the Mobility Management Entity (MME) and Home Subscriber Server (HSS) on the control plane.

In a sense the network logic moves closer to the User Equipment (UE). The tasks of the SGSN are typically taken over by the base stations (eNodeB) together with the MME. The SGW provides the IP-tunneling between PDN GW and the UE. Handovers are typically handled by the eNodeBs. Only when the two eNodeB involved in the handover are served by different MMEs the SGW will be involved.

In the same way the PDN GW is similar to the GGSN and the HSS to the HLR. We note that a 4G network implements its own HSS, i.e. in theory the central role of the HLR (spanning 2G and 3G

![Figure 4: High-level model of the toll system](image)
networks of the same operator) does not extend to the 4G networks. In practice HSS and HLR are either the same or so closely coupled that we expect an outage to affect all network types. To summarize the dependency of a GNSS-based tolling system on mobile data networks: regional outages are apparently exceedingly rare and national outages of a single type of network are known (but also rare).

4. Simulation model of the German tolling system

From the perspective of the German tolling system it is not necessary to model the core network of 2G, 3G or 4G networks. It is sufficient to know that regional and national outages are possible (but rare) – and of course the mean time to repair is completely unknown. In this section we describe the modifications to the existing simulation model [2] that allow us to simulate regional and national network outages as well as the ability of evading network outages by switching to a different network type or operator (national roaming).

The existing model (see figure 4) is split in two parts: The model of the technical system (consisting of a fleet of OBUs, the mobile data network and the central system) and a separate model of the external stimulus (called driving patterns, i.e. the user interaction with the HGV). The model was enhanced to allow regional network outages and national roaming:

- Central System: With roaming enabled a given OBU can have multiple network addresses.
- Mobile Data Network: Roaming is enabled and can be configured for each OBU; regions of a network can be switched on and off.
- OBU: For network outages the OBU tries to evade the loss of communications by switching to a different network operator.
- Driving Patterns: The transition between regions is included.

Previously the simulation model used to generate the driving patterns created only points in time (when the HGV was powered on or off, passed a section on the toll road signified by a toll event) and did not use any geographic information [16], [17]. For the current work we extended the driving patterns by adding regions (corresponding to the regions served by a SGSN or SGW of a mobile data network operator). Every time an OBU drives from one region to the next one the driving patterns model generates an event passes it to the OBU (see figure 5). The OBU processes this event by changing its current region to the newly provided one.

Lacking any detailed knowledge of the network configuration we assume that the network is split into regions (roughly corresponding to the 16 Länder of Germany). However, the HGV movement across the toll road network remains unchanged – i.e. a movement in time only. With the start of each power cycle the model randomly determines the point in time when the HGV passes from its current network region to a neighboring one. The probability depends on the size of the region and the probability of a transition between two regions is proportional to the length of the shared boundary connection the regions (see table 1). HGVs leaving the German mobile data network are modeled as powered-off, i.e. neither are any events generated nor any further transitions between regions.

Table 1: Probability [%] of an HGV to transition from the current region to another region within a fixed period of time.

| current region | next region | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ...
|---------------|-------------|---|---|---|---|---|---|---|---|---|---
| 1             | 30          | 30 | 30 | 10 |   |   |   |   |   |   |   |
| 2             | 30          | 30 | 30 | 10 |   |   |   |   |   |   |   |
| 3             | 20          | 30 | 15 | 15 | 15 | 5 |   |   |   |   |   |
| 4             | 5           | 10 | 10 | 25 | 15 | 25 | 10 |   |   |   |   |
| 5             | 20          | 50 | 30 |   |   |   |   |   |   |   |   |
| 6             | 15          | 30 | 25 |   |   |   |   |   |   |   |   |
| 7             | 20          | 30 | 30 | 10 |   |   |   |   |   |   |   |
| 8             | 5           | 20 | 10 | 15 | 30 |   |   |   |   |   |   |
| 9             | 20          | 40 | 20 |   |   |   |   |   |   |   |   |

Figure 5: The driving patterns contain the points in time for power cycles, toll events and when the HGV changes to another region of the mobile data network.

- HGV power
- toll events
- change of region
- time
Whenever an OBU needs to communicate with the central system (in the thick-client approach, the OBU typically initiates any communication) it first checks for network availability. As previously the network has a certain probability for local unavailability. We added that a network can fail completely in a region for a given network operator and the OBU can be provided with a list of roaming partners (whose networks the OBU can use as well). Once the OBU has a need for communication while the preferred network is not available the OBU tries to reach one of the alternate networks; with every new connection it tries to return to its preferred network.

5. Dependability: What are the effects of regional outages?

The simulation is initialized by distributing the OBUs to the regions proportionally to the size of the regions. Within every power cycle the OBU can change the region randomly (depending on the duration of the power cycle and the transition probability, see table 1). The simulation is run at a scale of 1:1 (i.e. with a fleet-size of 750 000 OBUs) over a simulated period of 30 weeks. During that time the typical processes of a toll system are executed: Tolling occurs continuously with a strong daily and weekly pattern, occasionally updates of the geo and tariff data are distributed across the fleet.

To see the effect of regional network outages we set up the simulation so that half of the OBUs prefer one network operator (A). During one week this network will have an outage of one region, 50% or 100% of all its regions. For the duration of the network outage we choose one day and repeat the simulation runs with an outage of one week both with and without national roaming enabled. This week coincides with the second week of an on-going update to the geo and tariff data.

By design, fleet-wide updates are spread over several weeks (see figure 6). As expected, short and regional network outages have no visible impact. The first simulation run includes the outage of one region in the network of operator A for one day (the Monday of the second week, marked with a red X in figure 6). No effect can be discerned (black line in figure 6): With an HGV driving on toll roads (at an average speed of approx. 80 km/h) a region can be traversed within a few hours and the network connection will be established in a different network region (even without resorting to roaming). Once the outage reaches a substantial part of the network (50% of all regions in the network A for the whole week, green line in figure 6), the impact is clearly visible on the progress of an ongoing update. A complete failure of network operator A for one week (blue line in figure 6) cuts the progress during that week in half.

![Figure 6: Progress of a fleet-wide update of the geo and tariff data as percentage of the fleet having the updated version when regional outages occur in one mobile data network during the second week (red X): Outages of one region for one day (black line), 50% of all regions for one week (green line) and 100% of one region for one week (blue line).](image)

![Figure 7: Relative server load during a fleet-wide update with negligible network outages (black line) and complete failure of network A during the first week (blue line).](image)
(since 50% of the fleet were subscribed to operator “A”).

We note that spreading the update across several weeks mitigates most of the network outage even when one network fails completely for a complete week (which is to date unprecedented): Most updates are made up during the week following the outage. This in turn leads to a significant, disproportionate increase in the peak server load (blue line in figure 7): The peak load on the following Monday almost doubles (although only 50% of the fleet were affected). In fact, this is an example of the emergent dynamic behavior: Only the combination of the technical system and the user interaction produces response of the system to normal or adverse operating conditions.

In section 3 we looked at different ways to mitigate network outages: The use of (supposedly) independent network generations (2G / 3G / 4G) or simply national roaming. For this discussion we choose to enable roaming so that OBUs subscribed to network “A” are able to connect to any available network as well. Taking the tolling process as an example we look at the rate at which the central system receives the toll data. The process of transmitting the tolls is very quick (on the order of 10 seconds) – in the “thick client” solution where the toll is calculated on the OBU and only the results need to be transferred.

The system is typically in equilibrium (see figure 8, the first three days) – i.e. toll events are generated (green line) at the same rate as they are delivered to the central system (blue line). Both rates diverge if the central system is unreachable (in our example network “A” is completely unavailable on Monday and no alternate mode of communications is possible). Barring any public holidays outages the typical weekly rhythm remains the almost unchanged (note in figure 7 that Monday and Friday show very similar HGV activity, a longer time series is in figure 9).

Can national roaming hide widespread network outages of a single network? Provided that roaming is both configured and still available (e.g. the HLR of the home network is still responding) we compare the tolls received in the central system for three consecutive weeks (see figure 9). During the second week the operator “A” (serving 50% of the fleet) is unavailable. Without roaming (blue line) the toll events gathered by the OBUs are not delivered (i.e. less tolls are received during the second week) but rather stored in the OBUs as long as sufficient memory (or credit) is available. Once network communication is reestablished in the third week, the tolls are sent to the central system (note the increase of received tolls on Monday).

Enabling national roaming we confirm that the availability of an alternate network completely hides the network outage of half the vehicle fleet (see figure 9, green line): The rate at which tolls are received is the same in all three weeks (e.g. looking at the dashed line in figure 9).

6. Summary and outlook

Designing or operating a system-of-systems with mobile endpoints should take into account large-scale network issues requiring considerable time-to-repair. In the case of the German tolling system we find through simulations that the inherent mobility of the
OBU mitigates regional outages fairly quickly. However, further study is required to calibrate the geographic spread of the driving patterns against real-world data [18]. Using such a model would allow identifying those HGVs that are used for local transportation only and therefore potentially affected by purely regional network outages.

7. References


