Proceedings of 7th Transport Research Arena TRA 2018, April 16-19, 2018, Vienna, Austria

End-to-End latency in HAD applications using cloud technology

Gottfried Allmer a*, Bernd Datler b, Manfred Harrer c, Peter Hrassnig d, Felix Pletzer e, Vijay Mudunuri f, Dominik Figl g, Oliver Hunger h, Georg Joo i

[a-d]*ASFINAG Maut Service GmbH, Am Europlatz 1, 1120 Vienna, Austria*
[e-f]*Intelliroad, Lakeside B01b, Klagenfurt, 9020 Austria*
[g]*Tieto Austria GmbH, Handelskai 94-96, 1200 Vienna, Austria*
[h-i]*Cellent GmbH, Lassallestraße 7B, 1020 Vienna, Austria*

Abstract

In the drive towards truly automated driving infrastructure data will play a substantial role, as it enhances the event horizon of the autonomous vehicle and enables the road operator to communicate strategic routing information. As infrastructure data is basically an aggregation of large data source systems, the guaranteed latency with which relevant information can be conveyed to the vehicle poses a challenge. This paper breaks up the downstream data chain from the infrastructure to the vehicle into its generic building blocks and focusses on the data throughput rate of the infrastructure database element. The achievable throughput rates are determined experimentally in a real life productive system during standard operation, the traffic information system of Austrian highway operator ASFINAG. The throughput rates through the main data gates have been made configurable and the timestamps for data passing through the individual software modules are recorded. Measurement results for the configuration with the highest throughput rate show a mean latency of 2 to 6 seconds for traffic messages from infrastructure into the vehicle, excluding the time for event detection. The concept will be expanded to eventually determine and monitor latency through all building blocks of the data chain.

Keywords: automated driving; infrastructure; latency

1. Introduction

The aim of the paper was to determine guaranteed latency values for traffic messages in real-life connected car systems. This is the start of a radically differential approach in determining data quality, where previous papers have taken an integral approach (Bogenberger and Hauschild 2009). Quality parameters, in this case latency, are investigated apart from any others.

Automated driving applications may require data from the infrastructure to the vehicle to achieve an acceptable driving experience and for controlling the bulk dynamics of vehicles following a vehicle initially affected by a traffic event. Latency is the prominent parameter in the transmission of infrastructure data as reaction time must be minimized for an automated vehicle to minimize the probability of erratic or dangerous automated driving behaviour.

A software model was defined to represent every connected car system in a generic way. A system was implemented incorporating all the elements of the software model and put into operation. We refer to this as the "operational reference system". Testing facilities were inherently included, consisting of the "Technical Exercise" to insert a traffic message manually in place of a real source, and a timestamp functionality in every generic software module to produce a time-log of every traffic message when passing through the modules. The tests are done in the operational system during real-life operation.

The software model assumes all data transmission to be frame-based, so that throughput rates are defined by frame rates. Two gates where data throughput is controlled are identified, an internal interface and an interface to an external cloud. The throughput rates of these gates were made configurable in the operational reference system.
We can now measure latency through each and every of the generic software modules, analyze the result and compare the results with the same generic module in other connected car systems in order to find optimal optimization implementations. Each of the elements of the generic connected car data chain will undergo a separate analysis. The element for this paper is the database.

Measurement results show that in the configuration with the highest frame rates, a traffic message can run through a connected car system in 2-6 seconds, excluding the source detection time.

This paper marks the beginning of a series of publications where every element of the data chain is measured including the transmission element where ITS-G5 will be compared to cellular G4/LTE/G5.

2. Background

2.1. ITS-G5 versus cellular 5G

As the move towards Highly Automated Driving (HAD) intensifies, so does the discussion about the preferred means of communication between vehicles, roadside infrastructure and traffic control centres. This paper aims to present specific cloud architectures which are being setup and tested by ASFINAG and industry partners as deployment projects in real traffic, and to characterize them by using end-to-end latency as a benchmark parameter.

It is currently most common to pit IEEE 802.11p technology against 4G/LTE/5G, or „Wifi vs. Cloud". But the „Cloud" version comes in several architectures, combining the exisiting clouds of the car manufacturers with clouds holding traffic management data and the traffic control center database. ASFINAG, the Austrian highway operator, owns and operates the entire highway infrastructure as well as all the highway traffic control centers in Austria. The company is in the favourable position of hosting HAD development projects with cars on a commissioned highway section in real traffic, and at the same time operating its own internal data transfer system to close the complete end-to-end data loop as foreseen in future HAD operations.

2.2. The role of infrastructure for automated driving

A paradigm of automatic driving is that the vehicle must be able to deal with all occurrences fully autonomously at least for any given immediate time frame.

Objects detected by the vehicle are moving towards it at great speed. Latency between the detection of the event and the entry of it into the processing unit of the vehicle is of utmost importance. If the object appears 250 m ahead of a vehicle, which is the event horizon of an automatic vehicle due to limitations of video detection, and the vehicle is travelling at 130 km/h, there are approximately 7 sec to react. To improve on that, any other system would have detect and transport the same information faster than 7 sec.

It may well not be feasible to detect and transport event information to a traffic control center in less than 7 s back and forth. In that case, true real time support can only come from communication technologies that not have to relay the information via a central station, e.g. ITS-G5, ITS-G5 V2V, LTE-V and LTE-V V2V.

Even if immediate warning is covered by other technologies, issues related to driver experience and the build up of dangerous bulk scenarios still remain. The user experience calls for the smooth handling of all traffic situations avoiding repeated abrupt braking action. This could be achieved by the vehicle having data from the infrastructure at hand, that allow the vehicle to see further than the on-board sensors allow. Bulk scenarios are inadvertantly involved during any traffic event. While the first individual vehicles on the scene may handle the situation perfectly well, the bulk of following vehicles may put unwanted strain on the individual vehicle's automatic control units because of multiple reaction threads. This could be significantly mitigated by relaying the correct data to the vehicle bulk.

sInfrastructure data will play a significant role for its ability to provide the greater picture to individual vehicles, which cannot be obtained by the on-board sensors. If latency and availability can be provided to meet industry needs, infrastructure can convey

- the complete regulatory information of the road network, thereby eliminating errors from on board cameras due to bad visibility, reflections or optical obfuscation by other vehicles
- unplanned events (accidents) as overall status over the complete network, including clearance forcasts, thereby enabling the vehicle to calculate an overall route strategy
- planned events (roadworks) as overall status over the complete network, including time plans and directions to negotiate large and complex roadworks, thereby enabling the vehicle to calculate an overall route strategy
- strategic re-routing information from a traffic control center taking into account the clean-up time forecast as well as inside information about the alternative routes.

3. The operational reference system of ASFINAG

ASFINAG owns and operates a real time traffic information system referred to as "the operational reference system" in this paper. Initially used to operate a smarthpone app, then enhanced with an Incident Management system to manage all incidents on the Austrian highway system, it now holds all traffic related data produced by ASFINAG's infrastructure.

From 2014-2016 a standardized DATEX II interface was added, so that industry partners can connect indiscriminately. The system, named „ASFINAG CONTENT", allows external agents to poll any number of specific data channels of interest independently via a standard REST (https) interface – e.g. current incidents, current and planned roadworks, current traffic sign settings, etc. Data is provided in real time. ASFINAG CONTENT is fully integrated into the live central system, operating day-in day-out to provide traffic information to the traffic control center operators, the ASFINAG mobile APP and industry partners who ultimately serve in-car applications.

An additional feature of ASFINAG CONTENT named „Technical Exercise" allows testers to insert simulated traffic messages via a dedicated user interface. The simulated messages are marked among the real ones in dedicated „Technical Exercise" infrastructure output files. The Technical Exercise data elements pick up time stamps as they travel through the ASFINAG system. In this way the latency contributions can be precisely attributed to the involved software modules.

4. The generic connected car data chain

Any connected car data chain consists of infrastructure management systems for the source of the data, an infrastructure database to collect and aggregate the data, a commercial traffic cloud the cars are connected to, and the mobile transmission of the data to the appropriate vehicles. The data scope of several terms in this paper is depicted in Figure 1.

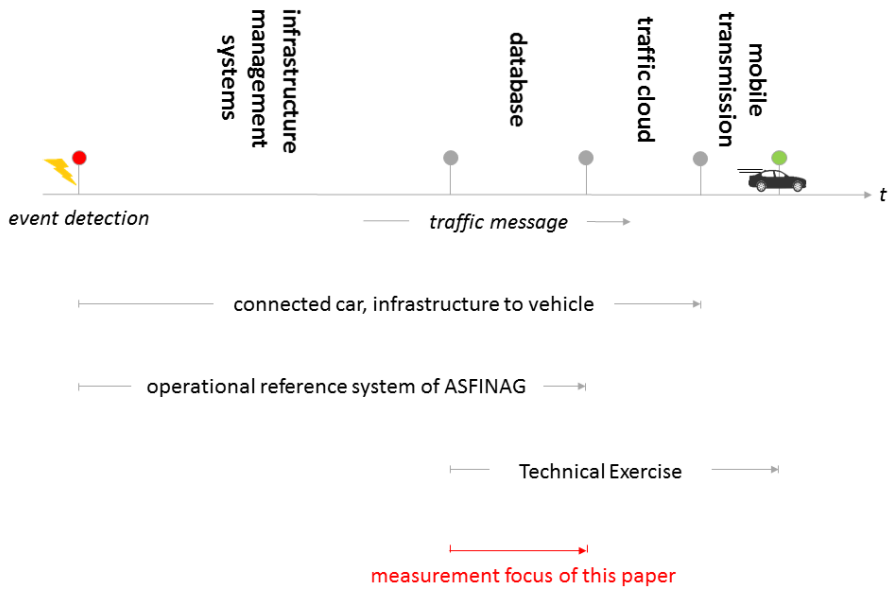

Fig. 1 generic connected car data chain

4.1. Infrastructure management systems

Infrastructure management systems are typically sizeable software rollouts which constitute stand alone systems in their own right. They are created to establish a standardized workflow for specialized tasks either because of the complexity of those tasks or the number of people involved.

Interfaces
Interfaces to other data systems are often only added at a later stage, so there is practically no standardization over the whole array of these systems.

Latency
These systems normally depend upon manual data entry which means latency class 10 min here. This data chain link has the potential to ruin an otherwise low latency. As a mitigation, try to automatize or circumvent the system.

Exemplified by the operational reference system

The main infrastructure management systems in operation are:

- Traffic control management system VMIS. Operators enter all regulatory measures into the system. Once the Variable Message Sign has altered its setting the information is transported through all the systems up the infrastructure interface without manual intervention. This time has been measured in a non-standardized setup to be typically 6 sec. So the latency contribution of VMIS is thought to be 3-4 sec.

- Incident management system EDB. All operators in the 10 Austrian traffic control centers are required to enter reported incidents as they happen. A total staff of about 100 people ensure 24/7 support. It is estimated that the mean time between an event happening and the EDB entry is around 5 min. Video detection systems

and the e-Call concept may help circumvent this manual step for a certain percentage of events. Latency contribution of EDB 10 m, maybe down to 5 sec in the future.

- Roadworks management system BMS. All project managers in construction projects are required to enter their projects and any changes to road and lane structure beforehand. Temporary maintenance is not entered however. In the near future roadwork trailers will be deployed that transmit their positions which could then be entered into the BMS automatically as if by a project manager. The latency loss manual step does not seem a problem, at least latency class 1 min should be achievable.

4.2. Database

Interfaces
Interface should and can be standardized according to EU directives.

Latency
In sophisticated modern database systems latency of class 1s should be achievable.

Exemplified by the operational reference system
The database of the ASFINAG traffic information system is called DDS (Data Distribution Server). It is a database cluster of 2 database servers programmed as standard database applications where data is aggregated for interfaces in “database views”. It has the potential to be split into smaller units for faster execution.
A multitude of data importers feed data of the infrastructure management systems into the database with various interfacing mechanisms. The output interface is a standardized DATEX II 2.3 interface, documented and always kept up-to-date on www.datex2.eu, DEPLOYMENTS, DII PROFILE DIRECTORY.

4.3. Traffic cloud

Interfaces
Interface should be standardized according to EU directives.

Latency
This is still under investigation. It is measurable with the operational reference system and will be focus of a later paper.

Exemplified by the operational reference system
For the measurement runs the operational reference system was connected to the Nordic Way Traffic Cloud. The vehicles report their position into a Geo-Location-Management (GLM sever) which are located into vehicle company clouds. For the measruement runs a GLM server located in Nordic Way was supplied to ASFINAG.

4.4. Mobile transmission

Interfaces
Interface should be standardized according to EU directives.

Latency
This is still under investigation. It is measurable with the operational reference system and will be focus of a later paper.

Exemplified by the operational reference system
The GLM server transmits data to all vehicles which enter a 1 km radius around the location of the traffic message. It is sent as data transmission via cellular systems of Austria’s three mobile network operators, A1, T-Mobile and Hutchison 3.

5. Methods

5.1. The connected car data chain decomposed into software modules

A model for the generic modular structure for the connected car data chain is suggested below (laid out in Figure 2.). The intent is to formulate the most generic architecture that encompasses any architecture of a real life connected car data chain, while being sufficiently specific to perform latency measurements on the structure elements for comparison and optimization.

The core is a real-time database for all traffic information and a succession of two abstractions in the data path, the abstraction of the data from the database in the form of a database facade, and the abstraction of the infrastructure output interface from the facade in the form of a database connection manager.

The operation reference system was implemented according to this model. There is potential to optimize hardware and software in this implementation as it runs on standard off-the-shelf hardware and the software was implemented by non-specialized application programmers from scratch and not yet derived from existing time critical implementations. For this the latency potential was made measurable in letting each software module apply a timestamp to a traffic message passing through, thereby effectively creating log with module-level resolution for the complete cata chain.

5.1.1. Software modules

- Database (DDS)
  The database is fed by numerous infrastructure data importers and holds all the traffic information in real time. This is called Data Distribution Server (DDS) in the operation reference system.
- Database Facade (DF)
  The database supplies the data into the database facade with minimum latency loss. The data is stored in a manner that abstracts the complex database structure from any client wishing to access the data. The façade concept also prevents direct access of the database thereby eliminating any data clogging by excessive retrieval requests. The facade is in volatile memory thereby also adding only minimal latency to the overall data transmission.
- Database Connector Manager (DCM)
  The database connector manager serves as standardized client. It accesses the data from the data facade and aggregates it into files which are supplied to the web servers where the data can be accessed by traffic clouds.
- Technical Exercise (TE)
  A module for manual input of data via a web interface circumvents the real infrastructure sources to enable extensive test Frame-based data transmission sessions.

5.2. Frame-based data transmission

Latency is predominantly governed by data transmission frame-rates, disussed here in greater detail.

Data transmission through an interface can be implemented as either frame-based or event-based. Frame-based is to be understood as sending the data in pulses at fixed intervals, e.g. once every 60 seconds or once every 10 milliseconds. The detailed structure of a frame is not discussed here. Specific data states which occur in between frame transmissions are lost. In contrast, in event-based data transmission data is transmitted as soon as it is updated at the sender. All data states are transmitted.

For the generic model of connected car data chain, data transmission at all interfaces is assumed to be frame-based. This is a simplification for which a case is made below. The stated arguments imply that data transmission cannot merely be abstracized as being frame-based in any case, but that they should ideally be implemented as such for stability, measurability, and therefore ability to optimize latency throughout the total system.

5.2.1. Frame-based pros and cons

Frame-based transmission provides for maximum system stability. The interface is operated at full load all the time. Whether there is a change at all on the interface sender or there is a change for every frame makes no

difference. Drawbacks are the efficiency decrease in the usage of transmission channel resources with decreasing data change rate, e.g. in sending hundred unchanged frames for every changed frame. This drawback can be mitigated if the frame is checked before sending and is only sent if changed from previous frame. In this case the transmission channel is not intrinsically tested for full load. Test runs for full load have to be run with data in which data changes with every frame. The drawback can be further mitigated if the frame is only sent when changed from previous and only the changes are sent. In this case Test runs for full load have to be prepared with data where data changes with every frame and every data item in the frame changes.

5.2.2. Event-based pros and cons

Event-based transmission provides for minimum latency. Drawbacks are that there is no check at the receiver as to whether the interface is working. Also as event load increases this method is prone to overload issues as there is no load limiter like in the frame-based method (where only the last event state of events of the same type occuring within a frame is transmitted). So the method has hidden issues and can only be used if there is a guaranteed maximum of event rate and even then it can lead to problematic scenarios when events are sent in time window to small for the receiver to handle.

5.2.3. Implementations

Push and pull
Whether the data is pulled (polled) from the receiver end or pushed from the transmitter makes no difference, in both instances data transport starts only at the frame start.

Notification implementations
Sender sends a notification when data has changed, receiver then fetches the data. This is frame-based when receiver polls at fixed intervals, so the fetching can only occur at the beginning of a frame. The notifications should also be sent frame-based, otherwise a notification burst could overload the interface.

5.2.4. System comparison

When comparing frame-based and event-based variants of an interface, the suggested procedure is to determine the minimum possible time resolution of the receiver and then treat the interface as frame-based with the determined frame-duration.

5.2.5. Summed up

In frame-based transmission, time resolution is sacrificed somewhat in order to keep the interface stable at all times. In event-based transmission is also a limit to the achievable time resolution in that there is only a finite minimum time interval between events that can actually be transmitted, as well as only a finite maximum event-rate that can be transmitted. As this is initially hidden when only few events occur, every interface should at least be tested with frame-based transmission.

The method of choice is to implement frame-based transmission with one of the two mentioned mitigations in place as required for the receiver. Then the frame-rate is increased to the point where the latency requirements are met and at the same time the receiving end can cope with the rate.

5.3. Data chain software modules and latency measurement points

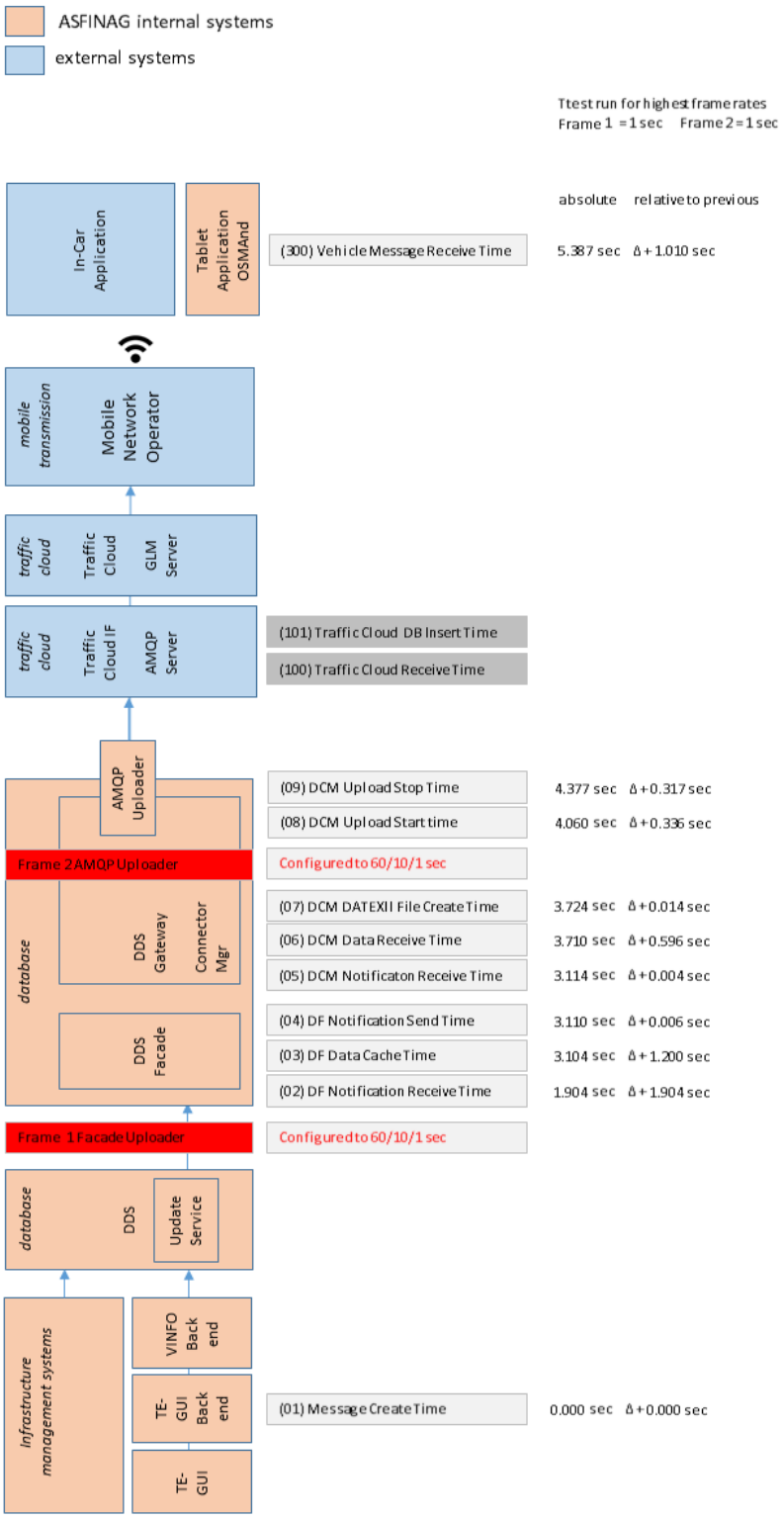

Fig. 2 Operational reference system software modules and latency measurement points

Figure 2 shows the position of the latency measurement points and includes the result of a typical test run for the highest current frame rate in the operational reference system.

Table 1. Latency measurement points

ID	Tag	Description
01	Message Create Time	Timestamp at which TE GUI backend created message
02	DF Notification Receive Time	Timestamp at which DF received a notification about the new messages
03	DF Data Cache Time	Timestamp at which DF read the new message and put it into the facade cache
04	DF Notification Send Time	Timestamp at which DF sent a notification to DCM about the update
05	DCM Notification Receive Time	Timestamp at which DCM received a notification from DF about the update
06	DCM Data Receive Time	Timestamp at which DCM fetched the data from DF
07	DCM DATEXII File Create Time	Timestamp at which DCM finished converting to DATEX II and saved it
08	DCM Upload Start Time	Timestamp at which DCM started upload of DATEX II file to AMQP Server
09	DCM Upload Stop Time	Timestamp at which DCM completed upload to AMQP Server
100	Traffic Cloud Receive Time	Timestamp at which Traffic Cloud received the data
101	Traffic Cloud DB Insert Time	Timestamp at which Traffic Cloud completed data insertion
300	Vehicle Message Receive Time	Timestamp at which vehicle received the message

DCM	*DDS Connector Manager*
DDS	*Data Distribution Server*
DF	*DDS Facade*
TE	*Technical Exercise*

6. Measurement Results

Test runs were performed in 3 configurations where the frame rates of gates Frame 1, the internal interface between database and database façade, and Frame 2, the interface to the external traffic cloud, were adjusted.

- Standard frame rates: tests with initial system setting when the system was used to support smartphone apps
- Inreased database frame rate: tests to verify the mean latency changes when the frame rates are altered
- Highest frame rate: tests with currently highest frame rates an single traffic messages
- Highest frame rate with load: tests with highest frame rates and real-life system load

Table 2. Measured latency values.

Configuration	Frame 1 database [sec]	Frame 2 interface [sec]	Mean latency [sec]	Max latency [sec]
Standard frame rates	60	60	61.276	70.150
Increased database frame rate	10	60	42.702	53.060
Increased interface frame rate	10	10	14.418	17.063
Highest frame rate	1	1	2.011	2.533
Highest frame rate with load	1	1	5.939	9.117

The highest frame rate produced an achievable mean latency of 2 seconds. Applying real-life system load pushed that back to 6 seconds however. This will be investigated.

7. Conclusion

Any connected car data chain can be expected to achieve a mean latency of 2 to 6 seconds for traffic messages from infrastructure into the vehicle excluding the time for event detection. This was shown by measurements in an operational reference system. Due to its generic architecture it is suggested that any other connected car system would achieve at least the same performance.

Further work will investigate the influence of message load which pushed the mean latency to 6 seconds. The generic architecture model will be used to investigate latency contributions of specific elements in the operational reference system and will eventually try to determine the achievable optimum latency.

8. References

Bogenberger and Hauschild 2009. QFCD - A microscopic model for measuring the quality of traffic information. ITS World Congress 2009, Stockholm, Sweden.