ERICSSON

HANDLING OF SIGNALING STORMS IN MOBILE NETWORKS

The Role of the User Data Management system

CONTENTS

1. EXECUTIVE SUMMARY

This paper describes the increasing challenges that mobile network operators face, due to the exponential rise of signaling introduced by the growth of IT based smartphone platforms, applications and devices.

Existing network protection has demonstrated not to be sufficient to face these challenges, and new strategies must be adopted to protect the networks and secure its carrier grade performance under any conditions.

A network congestion situation starts with a triggering event, which typically produces an escalation of the signaling traffic, mainly due to reconnection attempts. An e2e strategy in different layers to contain the overload surges is key to maximize end user service and ensure a quick recovery upon signaling storms. The e2e protection mechanism should include nodes at the edge of the radio network, signaling distributors and user data management systems. The overall strategy should be to contain the overload as close to the origin as possible.

This paper focuses on the role and mechanisms in the Ericsson user data management system, beyond traditional standardized mechanisms, which are instrumental to help the network handle this scenario efficiently, securing a quick and safe recovery.

2. INTRODUCTION

Current developments in telecommunication technologies and markets, stress the importance of a flexible and robust congestion control mechanisms in order to maximize performance and service availability.

The traditional approach of protecting individual nodes has been around since the introduction of GSM. The most relevant standardization bodies promoted the “optimistic” protection mechanisms, widely adopted by the industry. These mechanisms served its purpose successfully, until they were confronted with an increased complexity of mobile networks, and new usage scenarios, not considered in earlier specifications.

Nowadays, a constellation of different network access technologies, such as 2G, 3G, LTE, WiFi and fixed, coexist to provide seamless access to voice and data services. The huge penetration of smart phones dramatically increased the data consumption and bandwidth requirements and it is estimated that Smartphone subscriptions will more than double until 2020 (Ericsson Mobility Report, ref [1],). The Internet of Things (IoT), face the networks up to new usage scenarios, which in some cases increase drastically the signaling demands.

Telecom operators face extremely challenging network usage scenarios, leading to an increase of their service availability requirements, beyond the generally accepted 5-nines.

The continuous modernization of operator’s networks with increasing centralization of resources in higher capacity systems, such as Data Layered architectures in the user data management space, is leading to networks being more exposed to signaling storms, as incidents on centralized resources are likely to impact a larger amount of users.

The journey initiated by the industry towards cloud computing, with the transformation of the current network nodes into Virtualized Network Functions, seems at first sight to come as a handy solution to cope with the signaling storms in the network. Right the opposite; it may generate an illusion of infinite resources with the risk of underestimating the importance of the overload protection mechanisms. Scale-out mechanisms in the cloud will provide and increased flexibility to handle steady growths, although would not be fast enough to cope with sudden signaling peaks that escalate quicker than the network functions will do. This is particularly true in the user data management layer, where horizontal scalability requires a costly replication and / or redistribution process of the actual subscriber profiles.

Under these premises, it is easy to observe that the optimistic protection strategies, standardized and widely adopted in mobile networks are not valid anymore. High traffic peaks and network failure scenarios can lead to a massive signaling storm that can lead to long time outages of the offered network services.

According to Heavy Reading, ref [2], mobile operators are spending $15 billion a year to overcome network outages and service degradations. In average operators spend 1.5 percent of their annual revenues with some estimating it as high as 5 percent to deal with impacts from network outages and service degradations.

One of the biggest effects of network outages and service degradation is the increase in the rate of subscriber churn.

Advanced network models, later confirmed by the characterization of real implementations, demonstrate that end to end overload protection at network level maximize the system throughput and service availability under severe signaling storm scenarios, while minimizing the recovery time, compared with networks implementing the traditional protection mechanisms only.

Smartphone subscriptions will more than double until 2020

High traffic peaks and network failure scenarios can lead to a massive signaling storm that can lead to long time outages of the offered network services

[1] Ericsson, November 2014, Ericsson Mobility Report. Available at: http://www.ericsson.com/ericsson-mobility-report
[2] Heavy Reading, “Mobile network outages & service degradations”, Vol.11, No.11, October 2013

3. SIGNALING STORMS IN MOBILE NETWORKS

Mobile networks architectures, were developed under the assumption that connected devices, services and applications would be designed following some basic guidelines, optimizing the use of the scarce resources in their control plane.

This general assumption, common to all existing mobile network technologies, has been proven wrong with the massive growth of devices and applications designed under the traditional IT paradigms.

Also, assuming a “human behavior” behind a device is not valid anymore; terminals enters idle mode to save energy, and wake up periodically to sync data generating new authentications. Some applications send keep-alive messages without any user intervention. Millions of M2M devices may be configured to wake up exactly at the same time. These, among other usage cases, soon dismantled this belief.

At the same time, networks gradually increase their complexity. Multiple access technologies coexist, with the ambition of providing seamless access to data and voice services to consumers.

3.1 THE DEVICES

With the introduction of smartphones, telecom operators observed a great increase of signaling in their networks, which in some cases lead to signaling storms impacting a significant number of subscribers.

Major manufacturers of smartphone, come from the IT industry, and do not understand the constraints of the control plane in telecom networks. The result is mobile platforms which are constantly pooling the network automatically, rather than on a need basis

The pressure from the end users aggravated the situation, when complaints about usability, forced smartphone manufacturers to introduce additional features, which in some cases resulted in yet another increase of the signaling rate towards the network. That was the case for example of the strategies adopted to maximize the battery life for the terminals, like the non-standard “fast dormancy” feature introduced in 3G, which can multiply the signaling traffic in the network by a factor of 10, since it forces the device to initiate new connections after idle states. Exacerbated

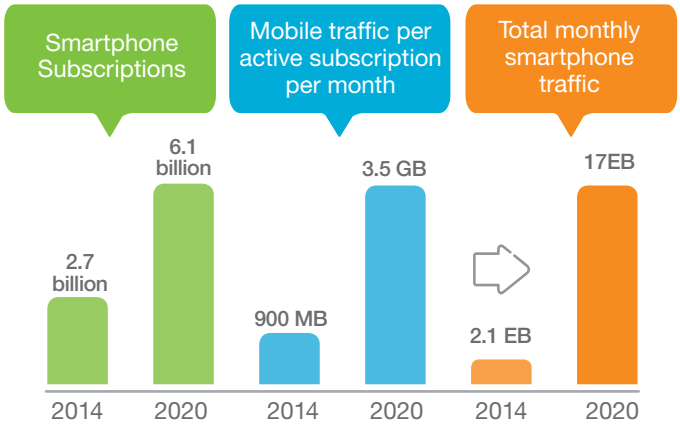


Figure 1: Smartphone traffic outlook

The huge variety of network applications and access technologies increase the risk of signaling storms

As it can be seen in Figure 1, extracted from Ericsson Mobility Report, ref. [1], the rising traffic generated by smartphones will continue growing exponentially. At the same time smartphone subscriptions are expected to more than double by 2020, on par with an increase of smartphone traffic over mobile networks around 8 times

In addition to the end user's terminals, the Internet of Things comes to bring another source of burden to telecom operators. Again, concerns do not come from the bandwidth consumption of those devices but from its signaling behavior.

Most of M2M devices nowadays are still using 2G (Figure 2) and almost 80% of these devices are GSM only, and its number will continue increasing in absolute terms, although the share of 3G / 4G devices will increase over time, to represent 50% by 2018. As a consequence, signaling issues associated with M2M devices may still come from both, MAP and Diameter based networks in the foreseeable future.

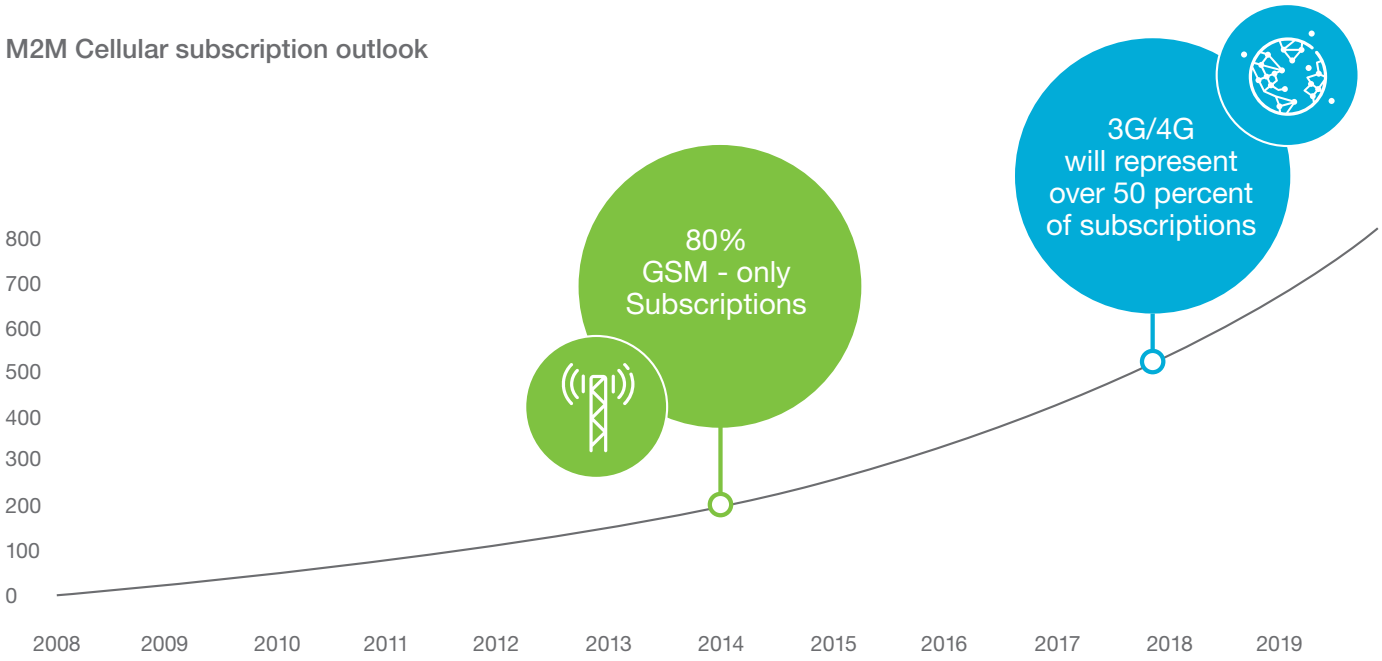


Figure 2: M2M Cellular subscription outlook

M2M devices are much more predictable in its usage patterns than devices controlled by humans. There have been cases reported, of a synchronized connection of millions of these devices to the network, in order to send small amounts of data. The authentication avalanche created peaks up to 40 times the expected signaling for the busy hour. Failed authentications induce the devices to follow exactly the same re-attempt patterns, which brings the network into an increasing degradation, which is difficult to recover.

Subscriber data bases are vital contributor of the network signaling traffic

3.2 THE NETWORK

The problems of network congestion, is not much about the data traffic that smart phones and other devices create, but in the underlying signaling traffic they generate.

The poor design of devices, platforms and applications is enough to explain most of the outages suffered by mobile operators around the world, due to signaling storms. On top of that, we could add the threat that malware or malicious attacks represent to the integrity of mobile networks, by exploiting the peculiarities of their control plane.

Telecom operators are experiencing signaling storms in two fronts; the radio access network (RAN) and the MAP and Diameter signaling traffic in the core network (CN). While both are different in nature, the end result is the same, and eventually RAN congestion may add on to the signaling avalanches in the core network.

In the core network, the control plane for both protocol families, MAP and Diameter create a strong interdependence between different nodes participating in the signaling flows, which may create a domino effect, propagating the overload to different parts of the network.

The congestion situation could easily propagate to adjacent networks. Telecom operators usually offer 2G, 3G, 4G and Wi-Fi service under the same subscription. Devices experiencing problems on a particular access will try to fail over to another one, with the risk of creating a signaling avalanche in the new target network.

On top of that, the centralization of resources in newer generations of network architectures has been very successful on simplifying the network topology, facilitating the introduction of new advanced user serives and reducing OPEX, although made the networks more exposed to signaling build-up issues.

Any strategies to address signaling storms should consider an optimization of the end to end signaling handling in the network, and mechanisms to fight the signaling peak as close to the origin as possible.

Since most of the network signaling requires accessing some kind of subscription data, failing to contain a signaling peak will cause an escalation of the overload, and its propagation to the subscriber data management system.

In the user data management space, 3GPP standardized a data layered architecture named as User Data Convergence, UDC, where traditional network databases in Core Network, such as HLR, HSS, AAA, policy controller, etc. are split into Application Front Ends, handling the business logic, and a User Data Repository (UDR) storing the user data.

This architecture enables data consolidation, simplification of the network design, the provisioning flows, and maintenance, resulting on significant CAPEX and OPEX savings. The User Data Repository plays a central role in all key functions, allowing operators to monetize on their network usage.

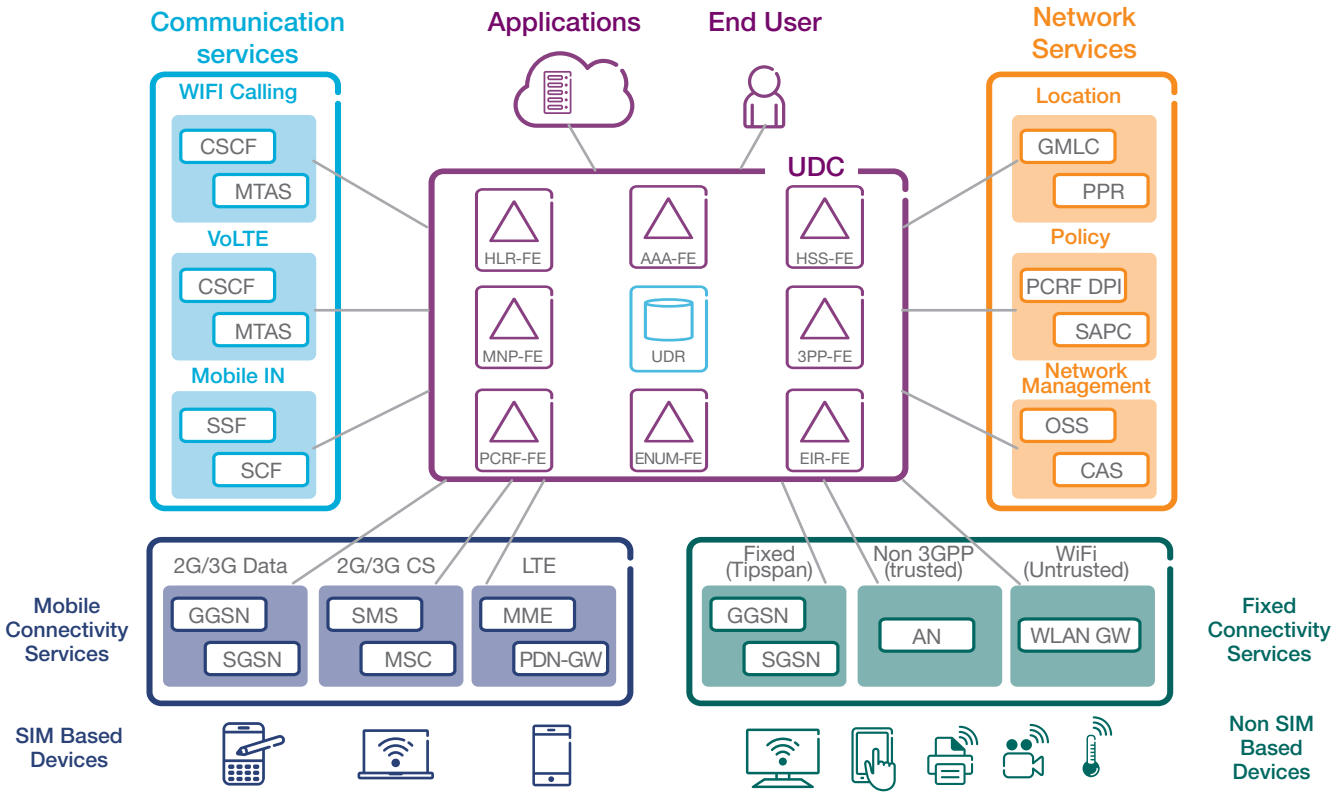


Figure 3: UDC in the network

Securing a smooth operation of the data store, by implementing advanced protection mechanisms against network signaling storms, guarantees that the network will produce the maximum useful traffic in any conditions, reducing the time required to recover from a signaling overload. This paper describes the Ericsson effective strategy to contain signaling storms in its user data consolidation solution.

3.3 THE APPLICATIONS

Similar to the smartphone platforms, popular applications such as instant messaging, social networking, and even games, are designed with no awareness about mobile network signaling efficiency principles.

It is well known, that even an apparently innocent game such as Angry Birds, can generate up to 2500 signaling events in one hour, and unfortunately it is not the exception.

Messaging, or social networking applications, launch automatically when the terminal is started, and silently send hundreds, or even thousands of keep-alive messages while the phone is turned on.

Cloud based applications also increased its popularity, and come to worsen the problem, since they must keep its local data in constantly in sync with the server, resulting on frequent additional signaling in the network.

4. ADVANCED HANDLING OF SIGNALING STORMS

4.1 OVERLOAD PROTECTION IN USER DATA MANAGEMENT SYSTEM

Signaling and data access protocols used in today's mobile networks, offer various mechanisms for a particular node to handle an overload situation.

A server receiving a signaling level exceeding its current capacity may react by explicitly rejecting the surplus requests with an error code, or silently discarding the message.

A peer node, would typically handle discarded messages by queuing the request and applying a given reattempt policy or propagating the error back to the source which will perform a reattempt on certain operations, e.g. attach, location update, etc.

Figure 4 illustrates the expected overload performance of a standard user data management system. When the incoming traffic, offered load, exceeds the engineered capacity of the user data management system, the overload protection and load regulation mechanisms protects the system from crashing by rejecting the excess traffic.

This mechanism is very efficient to handle isolated overload events, and it is a must for every system in the network. The excess of traffic is quickly and effortless rejected, assuming that the overload is a temporary situation, and the message will be successfully handled is a few seconds upon reattempt. This mechanism is very efficient to handle isolated overload events, and it is a must for every system in the network. The excess of traffic is quickly and effortless rejected, assuming that the overload is a temporary situation, and the message will be successfully handled is a few seconds upon reattempt.

This strategy will however not be sufficient to handle the huge avalanches introduced by today's network usages patterns. A massive overload situation will force the node to devote an increasingly significant part of its processing capacity to handle the traffic rejection itself, reducing the amount of useful traffic it can process, as can be seen in Figure 4 with the declining processed throughput trend as offered load increases.

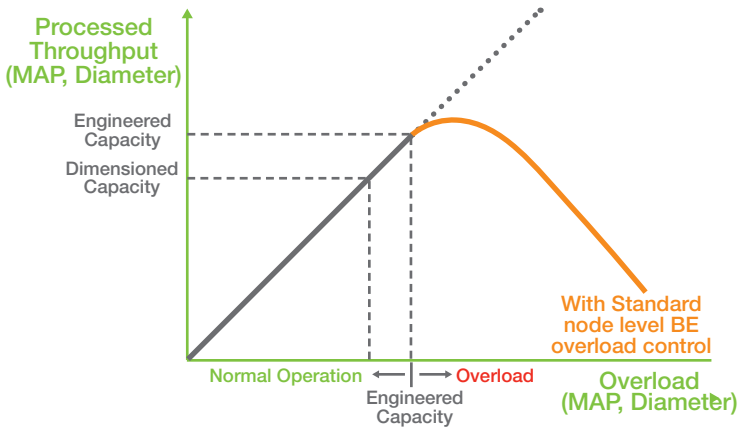


Figure 4: Typical overload performance behavior of a standard user data management system (FE + BE)

This situation leads to two main effects. On one hand, the obvious service degradation that the network and end users experience due to the rejected traffic. On the other hand, and more relevant for the overall overload situation, the consequences that traffic rejection has in certain operations that are either automatically and periodically reattempted by the network or terminal, such as network attach procedures, or as an initiative of the end user, like reattempting to establish a call.

The aforementioned automatic and periodic reattempts will exacerbate the overload situation, as it will create an exponential increase of load in the network on top of the already existing excess of traffic, inducing this way a snowball effect, that puts the whole network at risk.

This way, an overload situation in a node could quickly degenerate into generalized network congestion, a subsequent service outage and eventually the need for a manual intervention in order to recover.

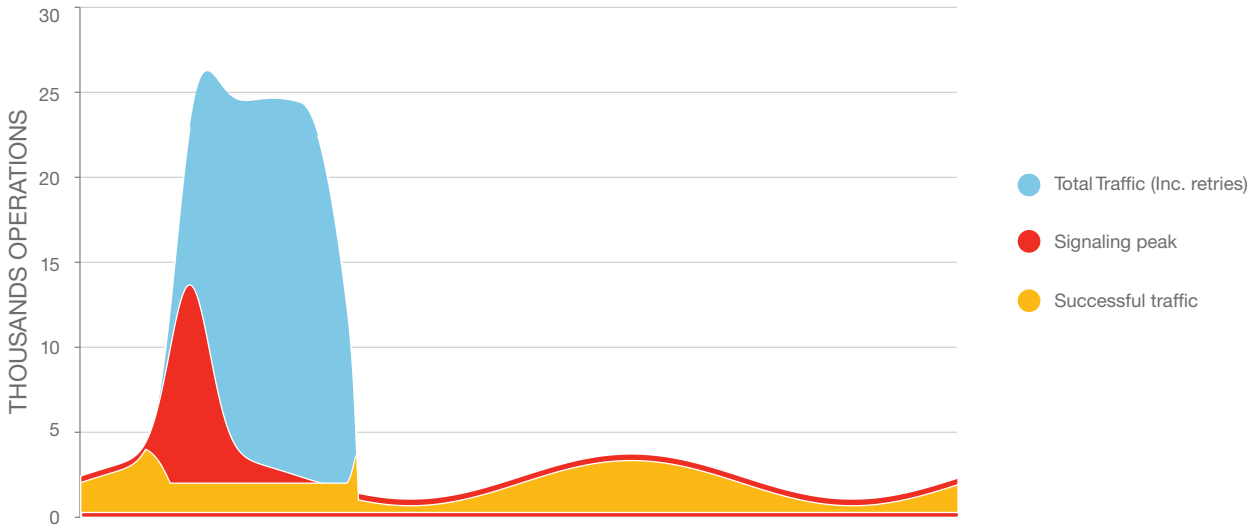


Figure 5: Network Congestion with overload protection at node level

The graph in Figure 5 shows the result of network congestion, where the servers use node level overload protection for the backend.

The network has been modeled with the following parameters:

Voice

2G & SMS: 95%

VoIP / VoLTE: 5%

Data

3G: 50%

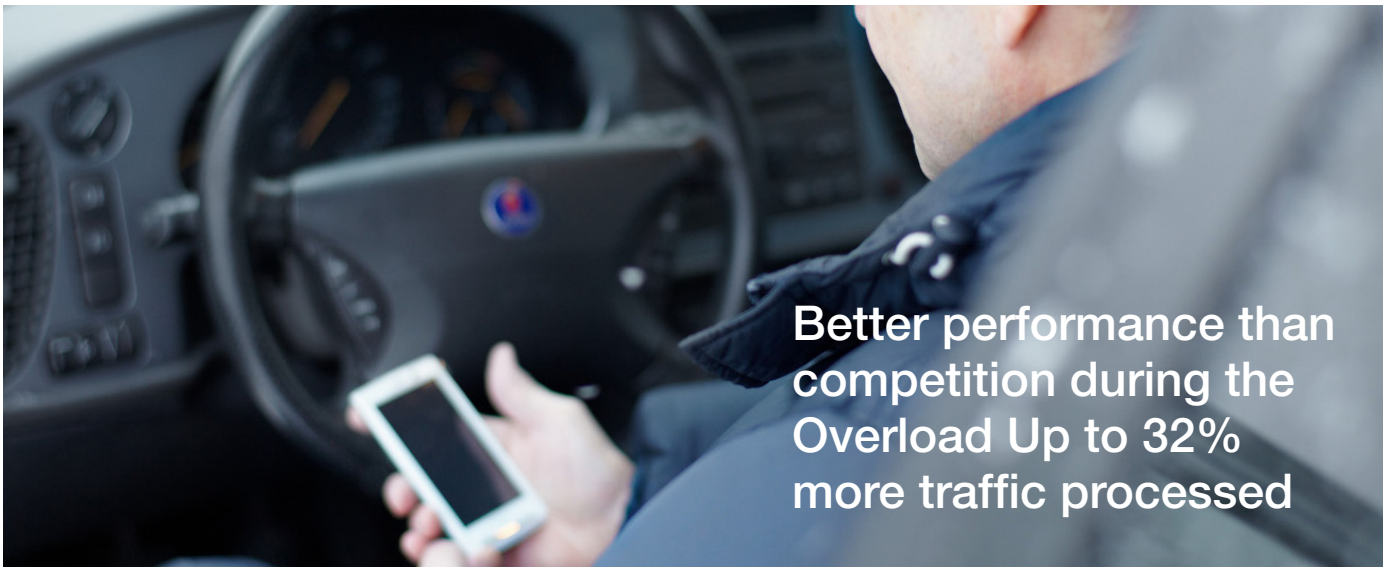
4G: 40%

Wi-Fi: 10%

A signaling avalanche, 3 times the network dimensioned capacity, is introduced in the busy hour. The impacted servers start rejecting traffic immediately, and soon the network accumulate retries (light blue in the picture) that escalate the signaling traffic up to 6 times the dimensioned capacity.

Once the introduced signaling peak declines, the network will still be congested for about 5.8 hours, trying to get rid of the snowball effect created by the retry policies and reconnection attempts. During the congestion period, the network loses its ability to process a normal amount of successful traffic and under these conditions, this mechanism, on its own, cannot give a network the “carrier-grade” consideration.

Cloud based applications also increased its popularity, and come to worsen the problem, since they must keep its local data in constantly in sync with the server, resulting on frequent additional signaling in the network.



4.2 THROUGHPUT ELASTICITY DURING OVERLOAD

In the core networks, signaling is processed by the contribution of many different network functions. The response time of each individual node in the process, contributes to the total response time of a network operation. For that reason, nodes in the control plane usually have very strict “real time” requirements, and the database is not an exception.

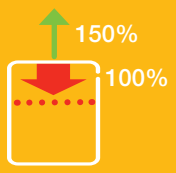
The response time is a function of the load level in the server. A strict policy, instructing the server to reject traffic when the response time reaches a certain threshold works great to keep the expected service level agreement (SLA) under normal traffic conditions. This is typically implemented through a rate limiter function in the system that secures fulfillment of the aforementioned SLA and keeps optimal e2e latency.

The down side comes under overload conditions where rejecting excess traffic to secure latency budgets will lead to an inferior network overload handling, as traffic rejection will eventually lead to a snowball effect due to reattempts as previously described.

Overload conditions calls for a different tradeoff between latency and throughput, where increasing throughput at the expense of increasing the latency will optimize the overall network handling of the overload situation, resulting in lower severity and lower duration of the overload event. This increase of latency can, by no means, be unbounded. Latency provided by the database must secure operations are answered within network timers, i.e. operations are useful for the network.

This throughput elasticity of the database upon overload conditions is key to promptly reduce the signaling traffic, and contributes to avoid the escalation of the traffic due to an excess of retries. At the same time, the response time will not be impacted under normal traffic conditions allowing optimal e2e latency as well

Extra processing capacity instead of degradation: Up to **150%** of capacity

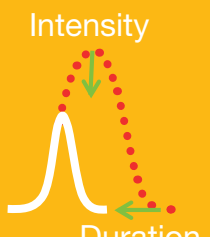


4.3 INTELLIGENT TRAFFIC THROTTLING

When fighting a signaling storm, the first priority is to secure that systems are kept safe, data integrity is not compromised and the total useful traffic processed by the network is maximized.

User data management systems and especially the user databases sit at the end of the signaling chain.

Less duration and intensity of the signaling storm Up to **83%** shorter & **20%** softer



Therefore, during a network overload scenario, user databases will naturally be under pressure, likely becoming the first overloaded component in the network. On the other hand user databases are the heart of telecommunications networks in the sense that are vital to deliver end user services, and a failure or degradation of the user databases might compromise the complete network.

As previously discussed in chapter 4.1 and shown in Figure 4, the standard overload protection and load regulation mechanisms used in user databases is proven not to be efficient enough, especially in severe overload conditions.

The amount of useful traffic processed by the network drops dramatically, multiplying the effects of the initial signaling peak.The user database plays a key role to keep the network going so the data stores cannot only rely on a node protection mechanism.

4.3.1 Ericsson's cooperative load regulation solution in the user data management systems

User data management systems require a more intelligent throttling of excess traffic to fully utilize the scarce resources in an overload context, with the ultimate goal to minimize both the severity and duration of the overload event. These objectives can be summarized in the following goals for a user data management system:

- Maximize e2e useful throughput, by minimizing the resources used in discarding excess traffic
- Securing that most important traffic operations are processed

The corner stone of this mechanism is the cooperation between the application Front Ends in the network, such as HLR, HSS, AAA, etc., and the user database.

The user database is constantly monitoring the resource utilization levels, such as the response time, length of the buffers, etc. to comply with given service levels agreements. As soon as some of the resources reach its limit, the situation is reported back to the applications Front Ends as an overload indication.

As discussed in chapter 4.2, this mechanism should be elastic, allowing for certain flexibility for the thresholds triggering the traffic rejection.

The application Front-Ends keep on measuring the amount of overload signals received from the user database and proactively throttle the traffic accordingly. The throttling level is continuously adjusted based on a dynamic and real time feedback loop, to make sure the database is always performing up to its maximum capacity avoiding the throughput degradation caused the resources devoted to reject the excess traffic

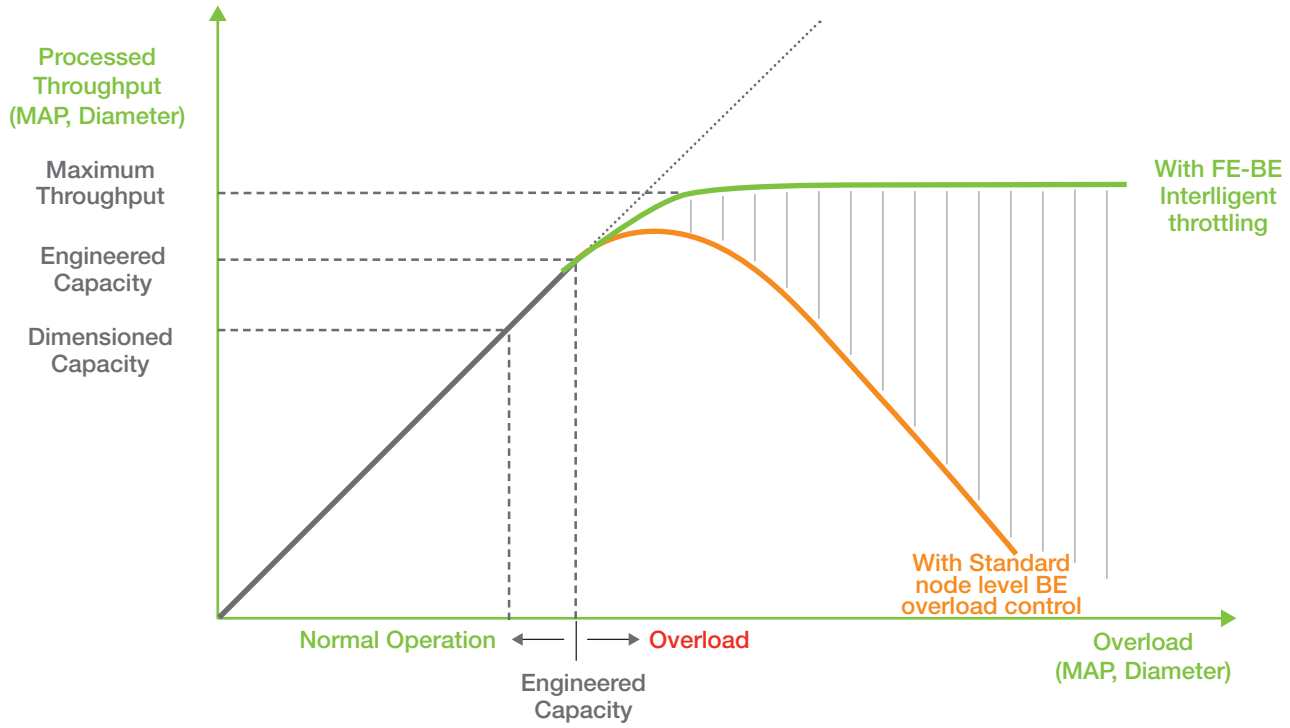


Figure 6: Overload performance behavior of a user data management system with cooperative load regulation (FE+BE)

Figure 6 shows the expected behavior of a user data management system with cooperative load regulation and a user database with throughput elasticity during overload as compared to a standard system.

When the system enters in overload, the first mechanism applied is the aforementioned throughput elasticity, which enables to provide additional throughput by modifying the tradeoff between throughput and latency applied in normal conditions.

This allows reaching the maximum throughput level for the user database. At that point, the user database will start to signal congestion indications to the Application Front Ends, which will start throttling traffic towards the user database, keeping the system working at its maximum possible capacity.

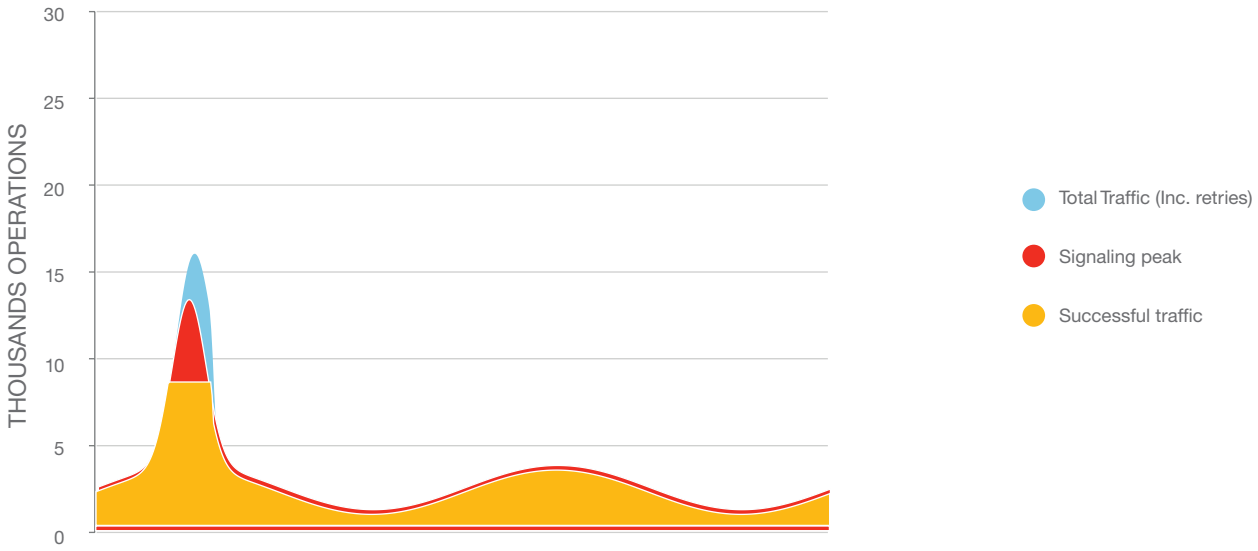


Figure 7: Network congestion with cooperative load regulation

Figure 7 represents the network behavior under de same conditions as introduced in chapter 4.1. This time, throughput elasticity and cooperative load regulation between the application front-ends and the user database was in place.

As shown in the case before, a signaling avalanche that is 3 times the network dimensioned capacity is introduced in the busy hour. In this case, there is also an initial escalation of traffic up to 4 times the dimensioned capacity, due to reattempted messages, and reconnections.

The user data management systems keeps on processing useful traffic up to its maximum capacity, helping the network to avoid the snowball effect, and solve the congestion shortly after the signaling peak declines.

4.3.2 Traffic prioritization

Handling of excess traffic during overload in a user data management system should be performed with the ultimate goal to maximize e2e useful throughput. Consequently, dynamic throttling mechanisms must apply priorities to make sure that:

- The most important signaling messages are not discarded. The importance of messages needs to be assessed from the perspectives of the contribution to the overall overload situation (e.g. prioritizing messages that would otherwise be subject of periodic automatic reattempts) and also from end user service perspective (e.g. call related messages)
- An ongoing operation involving several messages will conclude once it was successfully initiated. This is to secure that scarce resources during overload are utilized to process traffic that effectively contributes to e2e operations.

4.3.3 Multi-application cooperative load regulation

Data layered architectures in user data management domain aim to consolidate user data from different applications into a centralized user database. This enables both an efficient management of user data for mission critical applications such as HLR, HSS, AAA, with high throughput and latency requirements in the user database, while also enabling other non-critical applications to access consolidated user profiles and even centralize its own profiles.

In data layered architectures, the surge of the overload might come from any of the domains served by applications that centralize their profiles into the user database.

In the context of a user data management system supporting communication services (circuit switched voice, VoLTE, Wi-Fi calling, messaging, etc…), data services (2G, 3G, LTE, Wi-Fi), etc… a comprehensive cooperative load regulation solution involving the Application FEs supporting mission critical services, such as HLR-FE, HSS-FE and AAA-FE, is key for a solid and efficient overload solution.

Better performance than competition during the Overload Up to
32% more traffic processed

5. CONCLUSION

Standard overload protection mechanisms used in mobile networks today, have demonstrated not to be sufficient to respond to the challenges introduced by the growth of IT based smartphones platforms, applications and M2M devices.

Networks are not sufficiently protected against, not only malicious attacks, but also signaling floods coming from misbehaving applications or devices.

An e2e strategy in different network layers to contain the overload surges is key to maximize end user service and ensure a quick recovery upon signaling storms.

In the Ericsson User Data Consolidation solution an effective protection against this type of events is achieved by introducing a mechanism where network elements cooperate:

- Keep the network servers safe from the effects of the overload
- Keep the data integrity at the user data management system
- Maximize the network throughput, avoiding a snowball effect

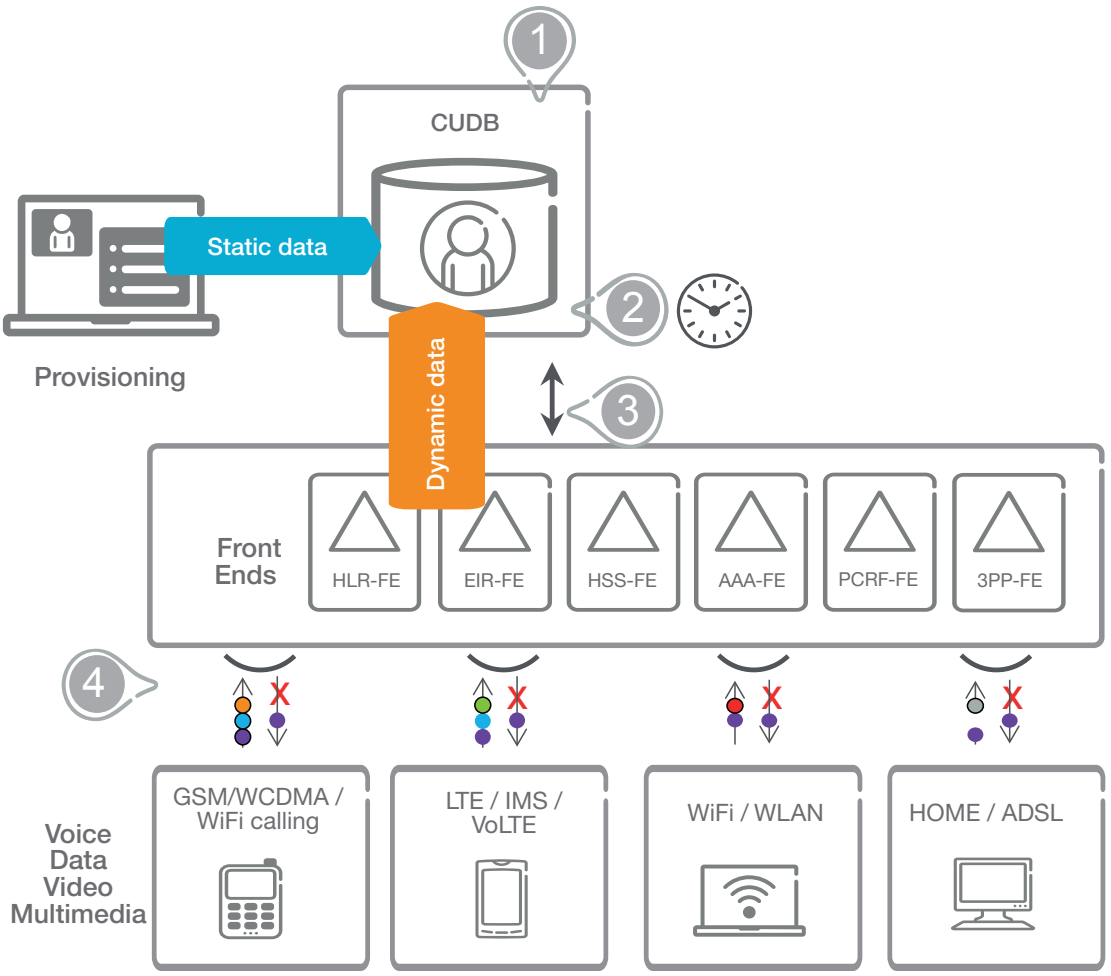
The corner stones for an efficient handling of the overload in Ericsson's solution, beyond standard mechanisms, can be summarized as:

- Throughput elasticity in the user database during overload with adequate latency tradeoff to deliver additional throughput when the network needs it most.
- Intelligent throttling of excess traffic, by means of a cooperative load regulation between Application FEs and user database with proper traffic prioritization
- Multi-application cooperative load regulation to secure optimal overload handling for mission critical applications

The result of implementing these mechanisms is a key contribution to minimize the severity and duration of an overload surge, increasing the resilience of the operator's network.

6. SIGNALING OVERLOAD HANDLING IN ERICSSON UDC

Balance between throughput and latency



1. System dimensioned according to traffic needs (static and dynamic data)
2. Congestion control level 1: throughput elasticity during overload, a tradeoff between latency and throughput providing extra capacity in the BE when it is needed most
3. Congestion control level 2: Cooperative Load regulation between CUDB and FEs – system level traffic regulation in high intense congestion cases
4. Intelligent traffic throttling between the FE and the Core Network

GLOSSARY

AAA - Authentication, Authorization and Accounting

CAPEX - Capital Expenditures

CN - Core Network

GSM - Global System for Mobile communication

HLR - Home Location Register

HSS - Home Subscriber Server

IoT - Internet of Things

LTE - Long Term Evolution

M2M - Machine to Machine

MAP - Mobile Application Part

OPEX - Operational Expenditures

RAN - Radio Access Network

SLA - Service Level Agreement

UDC - User Data Convergence

UDR - User Data Register

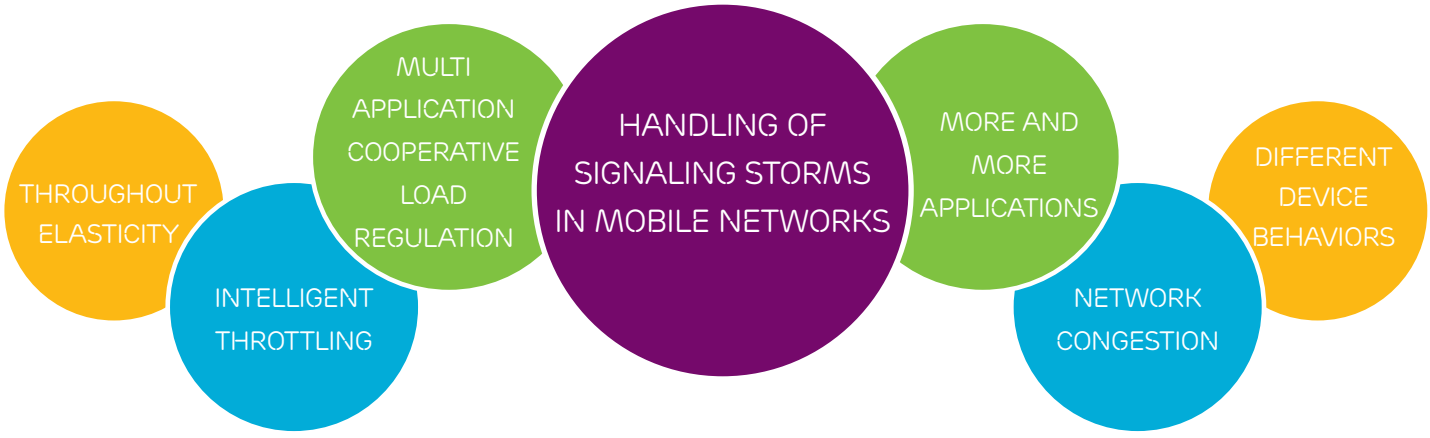
VoIP - Voice over IP

VoLTE - Voice over LTE

REFERENCES

[1] Ericsson, November 2014, Ericsson Mobility Report. Available at: http://www.ericsson.com/ericsson-mobility-report

[2] Heavy Reading, “Mobile network outages & service degradations” , Vol.11, No.11, October 2013



Ericsson is the driving force behind the Networked Society – a world leader in communications technology and services. Our long-term relationships with every major telecom operator in the world allow people, businesses and societies to fulfill their potential and create a more sustainable future.

Our services, software and infrastructure – especially in mobility, broadband and the cloud – are enabling the telecom industry and other sectors to do better business, increase efficiency, improve the user experience and capture new opportunities.

With more than 110,000 professionals and customers in 180 countries, we combine global scale with technology and services leadership. We support networks that connect more than 2.5 billion subscribers. Forty percent of the world's mobile traffic is carried over Ericsson networks. And our investments in research and development ensure that our solutions – and our customers – stay in front.

Founded in 1876, Ericsson has its headquarters in Stockholm, Sweden. Net sales in 2013 were SEK 227.4 billion (USD 34.9 billion). Ericsson is listed on NASDAQ OMX stock exchange in Stockholm and the NASDAQ in New York.

Ericsson

SE-126 25 Stockholm, Sweden

Telephone +46 10 719 00 00

www.ericsson.com

7/28701–FGB 101 807