Flat 3GPP Evolved Packet Core

Improvement for multiple network connections

Wolfgang Hahn
Corporate Research
Nokia Siemens Networks
wolfgang.hahn@nsn.com

***Abstract*—The paper discusses the 3GPP Evolved Packet Core (EPC) as an implementation of a flat data plane architecture. The number of data/user plane nodes might be not minimal in certain situations when multiple connections to different networks are established. Improvements are proposed to achieve an optimal flat network architecture.**

Keywords - 3GPP; Gateway; Evolved Packet Core, flat architecture, LTE

I. INTRODUCTION

A. Evolution of the 3GPP Network Architecture

A main target for the development of the Evolved Packet System (EPS) was to achieve a flat network with a low number of nodes in the data path. This improves the network cost structure and the connection delay. In the 3GPP Rel.8 EPS a main step was the removal of the RNC for the LTE access system. But even before this improvement the introduction of the Direct Tunnel and I-HSPA in 3GPP R7 resulted in a flat architecture: only one core node - the GGSN - is serving the data path. See figure 1 for the NW evolution.

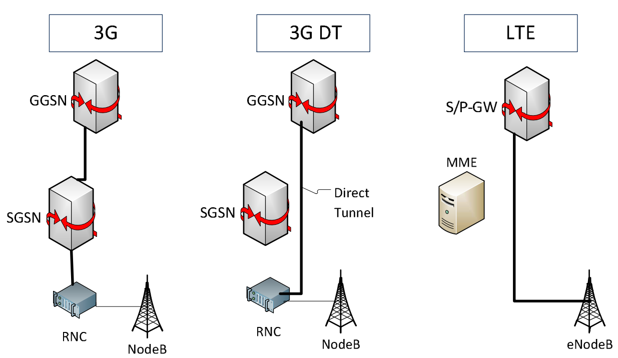

Figure 1. Evolution of the 3GPP packet core architecture

B. Consideration of an optimal flat Network / EPC

The different architectures raise the question on the optimal flat schema. A common definition is the minimal number of user plane (UP) nodes in the data path. It can be also observed that data path nodes for routing and switching in the transport network are usually implemented with more hardware support then the mobility specific devices. Consequently the last category will contribute more to the packet transport delay and the following considerations are focused to them. Obviously the base station node (BS, in EPS eNodeB) is necessary. A second node needed is a mobility anchor. This results from the fact that the 3GPP mobility concept is based on IP /Layer 3 mobility. IP tunnels (basically the GTP protocol) are connecting the anchor with the current serving BS. This has the following advantages: it allows the UE to keep its IP address while moving (service continuity) and in addition no special mobility related requirements are expected from the transport network, only IP routing. Mobility can span different networks domains independent of the underlying transport NW technology. From this it can be concluded that with the mentioned assumptions an optimal flat mobile network architecture has only two UP nodes in the traffic path.

C. Function split between BS and Mobility Anchor

3GPP has decided to co-locate certain network functions with the mobility anchor. This is mainly caused by security considerations as the mobility anchor are located deeper in the operator core network and is not exposed to unsecure remote BS sites. These network functions comprise

- Bearer management for real-time and non real-time QoS
- Data buffering and paging trigger for IDLE mode UEs
- Traffic volume reporting for charging and fair use policy
- Lawful interception
- Policy enforcement

Due to these functions and the fact that the mobility anchors provide the border between mobile and fixed network the mobility anchor nodes are termed gateways (GW). In the EPS evolved packet core (EPC) these functions are separated and distributed to a Serving GW (SGW) and a Packet Data Network GW (PGW). Only if SGW and PGW are collocated in one physical node (also termed SAE GW or S/P-GW) the architecture is considered as optimal flat.

II. FLAT ARCHITECTURE AND MULTIPLE PDN CONNECTIONS

A. Trend to multiple PDN connection

Already with the start of mobile packed data with GPRS it was possible to establish packet data connections (PDN connections) to different networks. This has been and is controlled by an Access Point Name (APN) that leads to a GW selection (GGSN or PGW in EPC) with connectivity to the target network.

In the EPC a UE may have different PDN connections in parallel that are terminated in different PGWs but only one SGW connects with the eNodeBs and serves as a mobility anchor for the eNodeBs.

A feature that makes use of the different PDN connections is the Selective IP Traffic Offload (SIPTO) what was introduced in 3GPP Rel.10. The solution is based on an enhanced GW selection that has the capability to select a mobile core network GW near to a RAN node (RNC, eNB). This is also called a local or distributed GW. The SIPTO function enables an operator to offload traffic that is destined to the Internet close to the UE's point of attachment to another network bypassing the operator service core network. In recent research the benefit of a distributed mobility architecture is a matter of increased interest, see for example [1].

In the 3GPP system benefits and drawbacks for both distributed and central GWs can be found:

Local GWs allow for Internet offload if a local Internet peering point is available or shortest path routing to local content sources or local correspondent hosts is enabled. But when the UE moves away from the point of attachment the routing may become non optimal. For this case the SIPTO feature contains the possibility for a GW reallocation: the MME deactivates the impacted PDN connections indicating "reactivation requested". The reestablishment of the PDN connection afterwards is associated with a new IP address allocation for the UE what causes consequently a service interruption.
In case of a central PGW the PDN connection benefits from the EPC packet bearer support that may provide a certain QoS compared to data traversing the Internet. This may be beneficial e.g. for VoIP services. During the movement of the UE the routing stays optimal as long the traffic is terminated in central operator service sites or Internet access points but the routing is not optimal if destinations local to the UE have to be reached.

Due to this situation it may be beneficial for the UE to establish different multiple PDN connections. Depending on the used services and applications the UE may keep a local or a central connection as "always on" connection and establish a second central or local PDN connection if needed. E.g. the UE may select the always on PDN connection with the central PGW and for certain Internet applications it could establish a local/SIPTO PDN connection on demand.
To avoid that UE implemented solutions would differ too much depending on the UE vendor and to allow for more operator control a dedicated effort in 3GPP Rel.11 was started with a work item to facilitate the traffic steering in the UE in case of multiple connections (Operator Policies for IP Interface selection – OPIIS [2]). It is also discussed that in case a suitable PDN connection is missing the policies could trigger the UE to establish a new one, e.g. a SIPTO connection for Internet traffic.

Note: this situation is very similar to multiple IP connections due to different access types like cellular and WLAN access.

The described multiplicity of data connections has impact on the "flatness" of the network usage what will be discussed in the next section.

B. Impact of multiple PDN connection to the flat architecture

During the UE attachment a PDN connection is established and the GW can be assigned in an optimal flat way, what means a co-located S/P-GW. If another PDN connection is established only a new PGW can be allocated e.g. according to SIPTO requirements (since there is no "mobility" trigger for a SGW relocation procedure). This may lead to not optimal routing situations as depicted on the left side of figure 2. To avoid this non optimal routing and to have the option to allocate a local/distributed PGW for any PDN connection it is needed to always select a local SGW - even in case a central PGW is chosen. This solution is depicted on the right side of figure 2. Although now non optimal routing for local traffic can be avoided a drawback of this solution is the forced separation of SGW and PGW. It introduces a limitation for the target to limit the number of involved mobile specific nodes in the data path. (The reason beyond is that in the EPC only one SGW is serving all the PDN connections of a UE, but for each PDN connection (so also for SIPTO connections) a different PGW can be selected.)

The question is if non optimal/ non flat configurations (when the SGW is located in a different site than the PGW) can be avoided? A solution could be that during establishing (or releasing) a PDN connection the SGW location would be changed.
The example on the left side of figure 1 may be considered again for the existing procedures: The UE may have a first established (always on) connection to an APN related to operator presence and voice services terminated at a central S/P -GW. Later the UE establishes a PDN connection to an "Internet APN" that could be subject to SIPTO. As the UE requested PDN connection procedure contains only a PDN-GW selection the SGW is unchanged and only a local PGW could be selected.

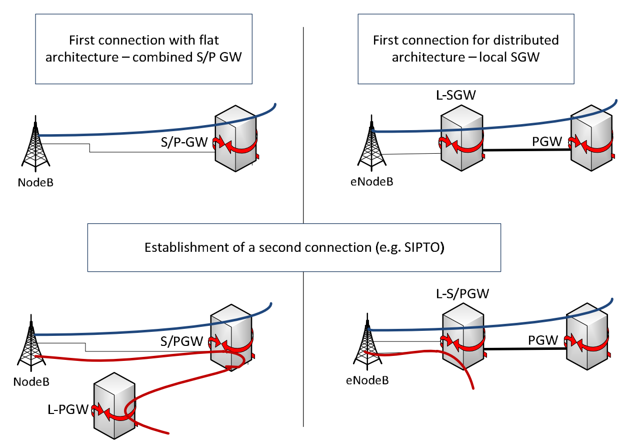

Figure 2. Non optimal a) routing and b) network flatness

It can be summarized that a) the existing procedure lead to either non optimal routing or b) a maximally flat architecture can't be achieved although it would be possible by the deployment of combined S/PGW nodes. It should be noted that this problem is not relevant for 3G networks to the same extend: With the direct tunnel feature only one core GW can be used and on the other hand the SGSN contains the control plane in addition to the GW functionality (compared to the GGSN a GW only node) which makes a co-location of both nodes not such attractive as in the EPC.

C. Proposal for an enhanced GW allocation in the PDN connection procedures

To overcome the non optimal split into SGW and PGW in two different locations for many cases it is proposed to add the SGW reselection/relocation procedure to the PDN connectivity related procedures (establishment and release of a PDN connection). Currently only during mobility procedures (what include a tracking area update (TAU)) the SGW can be changed.

In the following it is described how to add a SGW change procedure to PDN connectivity procedures and a proposal for this new procedure is given. Basically the SGW relocation procedure in the TAU procedure can be used in the PDN connection procedure with minimal changes. Variants may differ in placing the slight modifications more to the MME or to the eNodeB.

As described above multiple PDN connections initiated by the "UE requested PDN connectivity procedure" (see in 3GPP specification 23.401 [3]) may be a preferred subject for traffic offload (SIPTO). The proposed solution would work also in the case of a TAU with a MME decision to force a re-establishment of the SIPTO connection initiated by a PDN deactivation with a special cause code for a reactivation request:
During the re-establishment of a new PDN connection the SGW could be newly allocated in case in the TAU procedure the optimal SGW has not been selected since the old PGW is still active. (No change is needed for the TAU procedure compared to the existing standard). After the MME has selected a new (SIPTO L-) PGW it considers in a second step a further path optimization by a SGW change…

Based on the PDN connection procedures in [3] the PDN establishment can be modified as follows:

After receiving the PDN connectivity message (1) the MME determines whether this connection may be subject to SIPTO and performs (L-)PGW and SGW selection procedure accordingly. This leads to the decision to relocate the Serving GW. In this case the MME allocates a new SGW and releases the old SGW (steps 2 – 6b):

2 The MME sends a Create Session Request (bearer context(s) with PDN GW addresses and TEIDs (for GTP-based S5/S8) or GRE keys (for PMIP-based S5/S8) at the PDN GW(s) for uplink traffic, eNodeB address(es) and TEIDs for downlink user plane for the accepted EPS bearers, the Protocol Type over S5/S8) message per PDN connection to the target Serving GW. The target Serving GW allocates the S-GW addresses and TEIDs for the uplink traffic on S1_U reference point (one TEID per bearer). The Protocol Type over S5/S8 is provided to Serving GW which protocol should be used over S5/S8 interface. If the PDN GW requested UE's location info, the MME also includes the User Location Information IE in this message.

3 The target Serving GW assigns addresses and TEIDs (one per bearer) for downlink traffic from the PDN GW. It sends a Modify Bearer Request (Serving GW addresses for user plane and TEID(s)) message per PDN connection to the PDN GW(s). The S-GW also includes User Location Information IE if it is present in step 2. The PDN GW updates its context field and returns a Modify Bearer Response (Charging Id, MSISDN, etc.) message to the Serving GW. The MSISDN is included if the PDN GW has it stored in its UE context. The PDN GW starts sending downlink packets to the target GW using the newly received address and TEIDs. These downlink packets will use the new downlink path via the target Serving GW to the target eNodeB.

4 The target Serving GW sends a Create Session Response (Serving GW addresses and uplink TEID(s) for user plane) message back to the target MME. The MME starts a timer, to be used in step 6.

5. The MME informs the eNB about the new SGW. This could be a new message e.g. Uplink Path Switch Request message or included in an existing message e.g. the bearer modification message. This message includes Serving GW addresses and uplink TEID(s) for user plane. The target eNodeB starts using the new Serving GW address(es) and TEID(s) for forwarding subsequent uplink packets.
The eNB may confirm the UL Path Switch request with a dedicated message or include the UL Path Switch request in the later conformation of the Bearer Setup response message for the new SIPTO PDN connectivity bearer.

6. When the timer has expired after step 4, the MME releases the bearer(s) in the source Serving GW by sending a Delete Session Request message (Cause), which is acknowledged by the Serving GW with Delete Session Response messages. The cause indicates to the Source Serving GW that the Source Serving GW shall not initiate a delete procedure towards the PDN GW.

The corresponding message flow can be found in figure 3. It is assumed in the figure that the new selected SGW and PGW are local ones (L-SGW, L-PGW).

After that the MME proceeds with the (Re-)establishing the PDN connection to the (local) PGW as defined in section 5.10 in [3]

III. CONCLUSION

A UE may use in the EPS several data connection simultaneously or in series. The paper elaborates that with the current procedures and for a UE changing between one and several PDN connections the data path will be configured in a significant number of cases with separated SGW and PGW. This leads to increased network resource usage and decreased service quality. A proposal is made to enhance the 3GPP standard. It is shown that the problem can be solved if the procedures that are creating and releasing data connections are enhanced with a SGW change capability. This way a maximally flat architecture with co-located S/PGW can be achieved.

Nokia Siemens Network related research has received funding from European CELTIC MEVICO project.

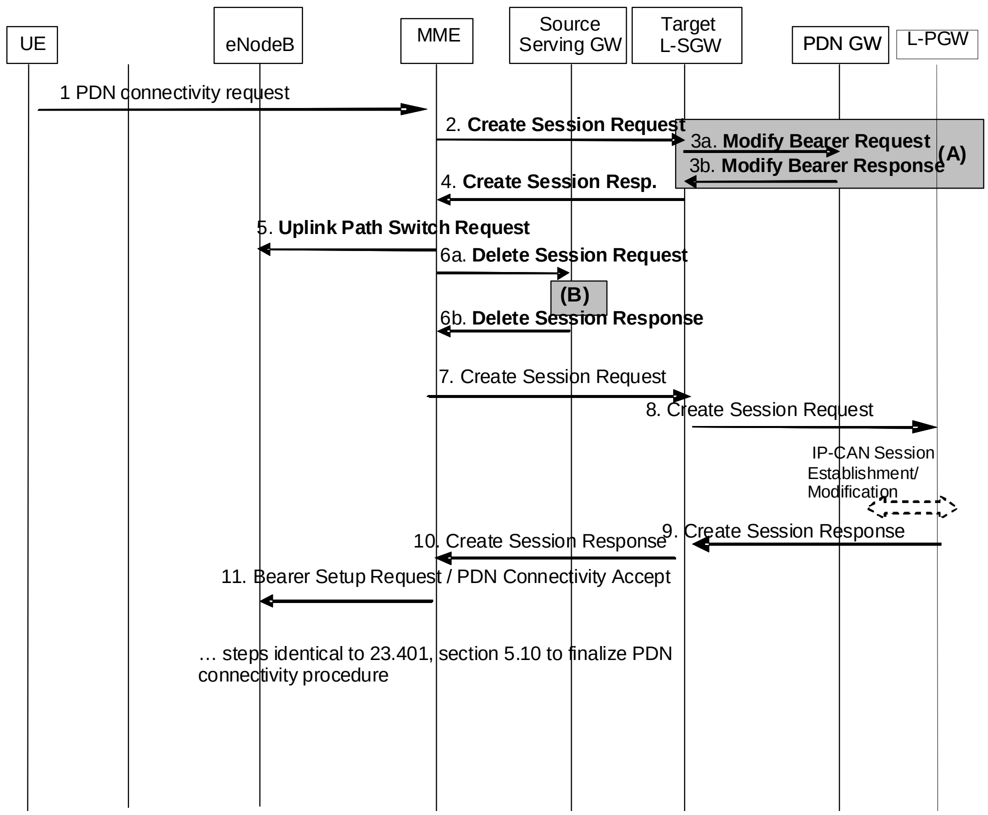

Figure 3. Enhanced PDN connection establishment procedure

ABBREVIATIONS

APN	Access Point Name
BS	Base Station
EPC	Evolved Packet Core (in the EPS)
EPS	Evolved Packet system, LTE RAN and EPC
GW	Gateway, GGSN or S/P-GW
SGW	Serving GW
PGW	PDN GW
L-GW	Local GW
LTE	Long Term Evolution
MME	Mobility Management Entity
PDN	Packet Data Network
PMIP	Proxy Mobile IP
SIPTO	Selective IP Traffic Offload
TAU	Tracking Area Update
TEID	Tunnel Endpoint Identifier
UE	User Equipment, mobile device
UL	Uplink
UP	User Plane

REFERENCES

[1] Dapeng Liu, Hidetoshi Yokota, Charles E. Perkins, Melia Telemaco, Pierrick Seite, H. Anthony Chan, Wassim Haddad, Hui Deng, Elena Demaria, Carlos Bernardos, Jun Song, “Distributed/Dynamic Mobility Management(DMM) - Problem Statement” IETF 79 Proceedings, Beijing, China November 07-12, 2010

[2] 3GPP™ Work Item Description “Operator Policies for IP Interface selection (OPIIS)” SA WG2 Meeting #83, TD S2-110668, 21 - 25 February 2011, Salt Lake City, Utah, USA

[3] 3GPP TS 23.401: "GPRS Enhancements for E-UTRAN Access"