

# Nokia Siemens Networks

## Cell load and application-aware traffic management

End-to-end Quality of Service differentiation

White paper

Nokia Siemens  
Networks



# Executive summary: Differentiation is the key to managing growth

## Table of contents

Executive summary: Differentiation is the key to managing growth	2
Fundamental requirements for managing traffic	3
Key drivers for operators	4
Planning traffic differentiation: Key considerations	6
Solutions to suit different mobile networks	7
5.1 Bearer separation	8
5.2 Dynamic traffic priority adaptation in application-aware RAN	10
5.3 Core-based traffic management	12
Conclusions	14
Glossary	15

Today, an avalanche of innovation is transforming the communications industry as mobile broadband (MBB), smartphone apps and the mobile Internet become a worldwide phenomenon. Global mobile broadband is expected to reach more than 2.5 billion subscriptions by 2015, generating 60 Exabytes per month of traffic.

With the introduction of high-speed radio standards such as HSPA (+) and LTE and all-IP transport, many operators are ready to meet the rising demand for network capacity. Mobile broadband success is mainly based on a wide variety of services available on the Internet. Operators no longer need to invent all the “killer data applications” for all customer segments, with over-the-top (OTT) service providers performing this role instead. Profitable mobile broadband requires that operators have wide network coverage and capacity for all subscribers as well as tiered data plans for different customer segments. Furthermore, operators have the opportunity to provide added value and to differentiate based on the user experience.

Many operators have responded by focusing increasingly on offering users a personalized, differentiated quality of experience (QoE). These operators need solutions that allow them to manage the user experience end-to-end, turning network performance into a value generator in the era of smart mobile broadband applications.

One approach is to distinguish between the traffic from different applications, offering a range of quality of service (QoS) standards as appropriate. These QoS differentiation solutions must be able to adapt dynamically to match variable traffic loads end-to-end throughout the core and radio access network (RAN). Ultimately, LTE will introduce powerful mechanisms for dynamic QoS differentiation management. However, legacy devices, the installed 2/3G network base and different operator deployment strategies all need to be addressed when devising solutions for QoS differentiation management.

This white paper presents a consistent set of end-to-end QoS differentiation management solutions to fit the relevant deployment scenarios.

# Fundamental requirements for managing traffic

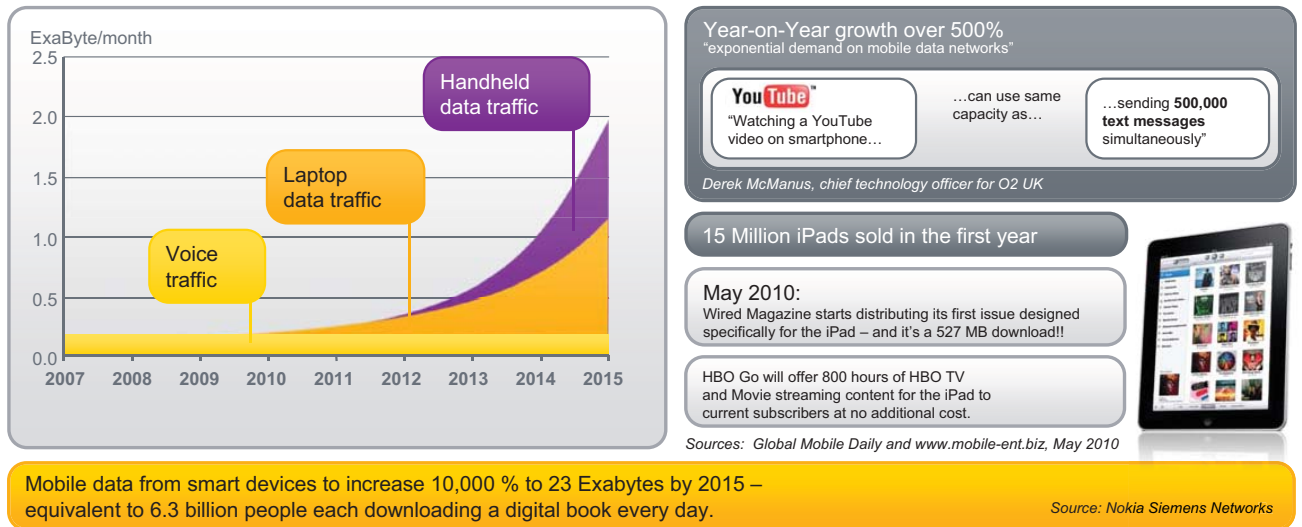


Figure 1. Mobile data growth until 2015

The age of anytime, anywhere access to data services has already arrived for many of us. This is driving a deluge of new traffic, as shown in Figure 1.

Evolving radio access technology such as HSPA and LTE can be combined with all-IP transport to meet the demand for capacity from MBB. Data rate and latency evolution in HSPA and LTE networks guarantees that even the most demanding of today's and future applications can be used over mobile access. While capacity expansion is a must, operators can optimize their network efficiency with QoS tools, which both guarantee high QoE and maximize utilization of network resources.

Traffic growth and respective network capacity expansion has a positive

impact on the network cost per delivered GB due to economies of scale. However, in order to guarantee profitability, the operator must also make sure that traffic growth is based on new subscriptions as well as being linked to the actual data consumption. An attractive service offering provides affordable plans for all customer segments with clear differentiation. Therefore, charging and policy control together with QoS differentiation are crucial tools for an operator to develop a profitable MBB business.

Thus, operators need a more segmented approach in order to boost MBB profitability.

Introducing efficient and intelligent traffic management can improve the user experience, increase revenues

and reduce churn. Only operators have the ability to provide the end-to-end QoS needed to support many MBB services. Differentiating QoS according to different types of traffic can improve the QoE for end users and manage the demand for scarce network resources. One approach is to match the QoS to the particular requirements of different applications.

QoE assurance mechanisms must span all the network elements from the core to the user device, as well as providing traffic detection using deep packet inspection (DPI), radio load-aware QoS, traffic separation and support for legacy devices.

No single solution suits all operators. An analysis of the existing network and business operation is needed to determine the right choice.

# Key drivers for operators

Understanding what end users want is critical for operators faced with the fast development of innovative Internet offerings. How can operators stay relevant to their customers? Network performance still matters, but the way in which it's perceived to add value for end users has changed. Improved QoE, loyalty and customer experience are some of the key strategic goals for operators looking for a sustainable business, supporting all the business aims identified in Figure 2.

Consider the example of MBB, where heavy users can create bottlenecks that leave everyone else unable to access their applications. Operators

need to find an optimum way of ensuring fair access for everyone.

The ability to manage the user experience thus becomes a core differentiator for operators. Help users to access information and services for a fair price, give them fair access to content and be context-aware in a trusted, secure environment and you've got a winning formula.

Mobile operators are therefore looking to offer differentiated mobile service packages to different user segments, ranging from professionals to teens and other casual users. The policy control and charging (PCC)

architecture from 3GPP offers a powerful technical solution.

The bigger challenge is to differentiate QoS according to the needs of each application, whether they're narrowband, broadband, real-time or value-added offerings such as video on demand. Some operators have already invested in SAE/LTE technology, while others are looking for solutions to optimize traffic management in legacy networks. Solutions for both types of deployment may be needed, depending on the specific networks, business strategy and customer base of each operator.

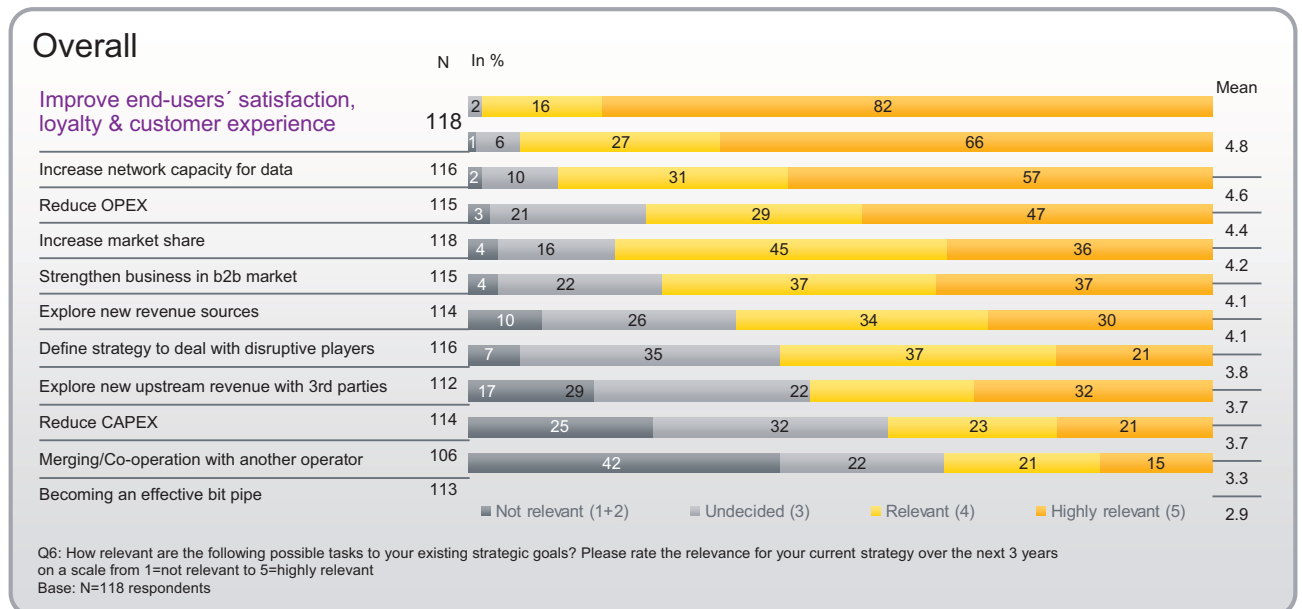


Figure 2. How QoS supports operators' strategic goals

Business Evolution Study 2010

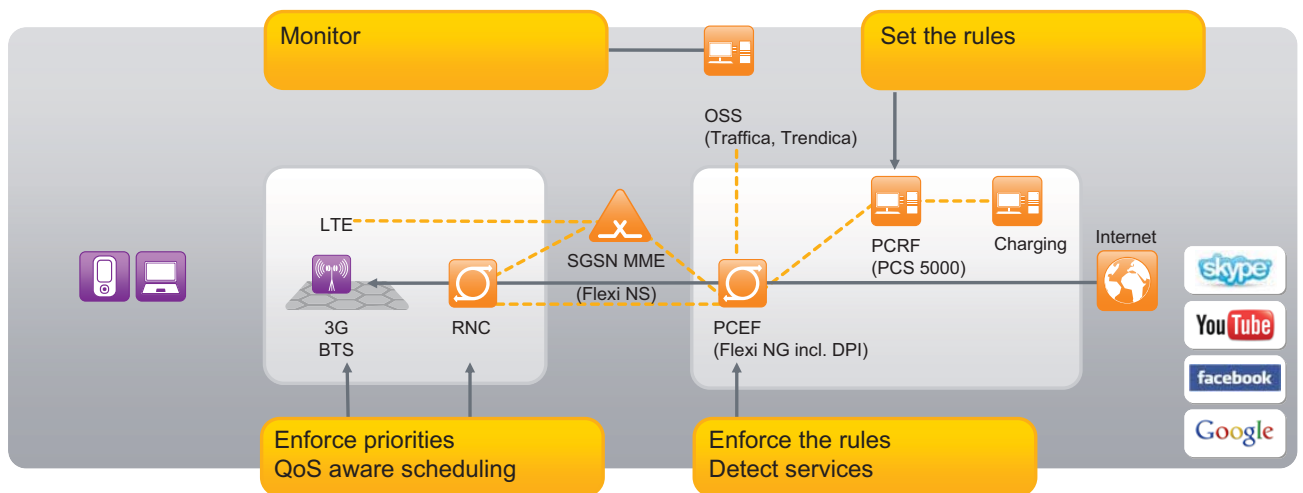


Figure 3. Nokia Siemens Networks end-to-end QoS differentiation solution

Four steps are needed in any solution that aims to protect QoE as traffic levels increase (see Figure 3).

**Monitoring** the critical user experience and network performance indicators in (near) real-time is essential. The New Generation Operations Systems and Software (NGOSS) provides a suitable framework for developing, implementing and deploying the necessary Operations and Business Support Systems (OSS/BSS) collecting relevant KPI from the different network elements.

**Set the rules** based on insight into end-user behavior and network load. Define segmented experience parameters and create relevant marketing campaigns and QoS

policies, as well as directing targeted capacity extensions. This calls for consolidated business processes and a consistent information architecture / data model.

Operator campaigns can be validated by analyzing take-up rates and revenue contribution and the results can be fed back to campaign planning and tuning.

**Enforce the rules** using DPI to identify the service and end-to-end QoS control to define the right enforcement strategies in the core network. Team this with agile provisioning and charging processes to implement the marketed service offerings, policies, tariffs and capacities. This calls for consistent provisioning and rules

generation between the Policy Control Server (PCS), Online Charging System (OCS), OSS and Subscriber Data Management (SDM).

**Enforce priorities** by implementing QoS-aware scheduling in the radio access nodes.

The focus of this white paper is on the technical options offered by 3GPP. However, technical solutions alone cannot address the business or third-party application considerations that also play an important role in providing an enhanced user experience. This calls for a flexible and dynamic business process framework, as defined by the enhanced Telecom Operations Map (eTOM), for example. This enterprise architecture specified by the TeleManagement Forum (TM Forum) offers an industry standards-based transformation framework to promote a user-centric operator model.

# Planning traffic differentiation: Key considerations

Growth is generally good for any business, including mobile business. Today the main growth opportunity lies in mobile broadband. In order to boost the profitability, a more segmented approach is needed. Service differentiation and a personalized user experience are key ways for an operator to achieve a successful business.

The radio interface and the RAN backhaul are typical bottlenecks for MBB traffic, owing to the scarcity of spectrum resources and the widely varying user density per cell. Policies must be executed flexibly to cope with fluctuating capacity over-the-air due to mobility and interference, as well as unpredictable variation in traffic demands. In addition, high-volume services such as peer-to-peer file downloading negatively impact interactive services such as browsing, especially at peak times.

Furthermore, subscriber-level QoS - as supported in 3G with the primary PDP context - does not separate different applications within a single bearer, which makes it hard to implement fair usage at the application level. This is a particular problem for congested cells where the QoE of several applications may become unacceptable at once.

Bearer separation is a natural solution foreseen by 3GPP, where different applications are carried over different radio bearers. However, in the case of 3G, the solution requires the support of secondary PDP context in the user equipment, and that's rarely available in today's networks and mobile devices. Alternative solutions are therefore needed to support legacy devices.

Prioritization and charging for application differentiation requires the ability to perform specific packet enforcement actions, such as reducing the bandwidth for less-critical applications when there's congestion. Such policies require operators to analyze, detect and trigger appropriate actions in case of high traffic loads.

One challenge is how to implement traffic identification, load detection and policy enforcement in 3GPP network elements. Take video streaming, for example. Policy enforcement (PCEF) typically takes place in the gateway. However, if a streaming session suffers cell-load congestion, the gateway does not have complete information about other sessions in the same cell since, for example, handovers are happening transparently to the gateway. This is a common challenge in HSPA and LTE networks

Congestion handling is another issue for QoS-sensitive applications with bursty bit rates, such as video streaming. When setting up a connection between the user and the network, the 3GPP radio system can provide a Guaranteed Bit Rate (GBR) or a best-effort service. For a GBR bearer there are well-defined metrics indicating when the system is overloaded. But GBR bearers are ineffective for services with a bursty bit rate. There is no such congestion metric for best-effort bearers (non-GBR). The best-effort bearers get whatever is left after the GBR bearers have been served. This makes it essential to detect high traffic loads and have strategies to mitigate congestion for any non-GBR connections that carry QoS-sensitive, bursty services.

In conclusion, in order to translate the operator's strategy into a precise configuration, above considerations need to match technical, commercial and customer experience impact and planned targets. For this purpose, Nokia Siemens Network has developed its QoS Differentiation Introduction Solution, which offers support to operators who are planning to introduce intelligent traffic management and QoS Differentiation functionalities.

# Solutions to suit different mobile networks

The tool box to deliver a segmented MBB service offering is a new, dynamic end-to-end policy control solution. This enables MBB service differentiation by introducing a range of QoS traffic priorities to provide each user with the speed, access and quality they need and pay for, even during peak times. Traffic is prioritized so that higher-priority flows get access, throughput, latency and packet loss according to established SLAs, while lower-priority flows get a best-effort service.

DPI can be applied to identify the type of service being used. DPI ensures that even OTT services can be included in the policy decision process. It detects applications and notifies the policy

controller, which then manages the flow as agreed in the relevant SLA.

For example, if the subscriber's data package includes a premium sports channel, the DPI detects the connection and configures the delivery channel for HD video. It also stops data volume-based charging and triggers the per-view sports channel subscription fee. CSPs can differentiate their offers by prioritizing their own services and offering premium QoS for selected OTT content. With 3GPP Rel 11, the DPI function is part of the architecture and will be included in the TDF (Traffic Detection Function). The TDF can be included in the Packet Data Network (PDN) gateway or as a stand-alone function.

## Bearer separation

The most efficient way of differentiating between applications involves traffic separation using different bearers. This corresponds to the first solution described in Figure 4.

A bearer represents a virtual connection between the user terminal and a gateway node in the packet core. In HSPA the concept of the secondary PDP context exists for that purpose, while dedicated bearers are defined for LTE/SAE. Each bearer will be treated according to the QoS profile assigned to it by the scheduler.

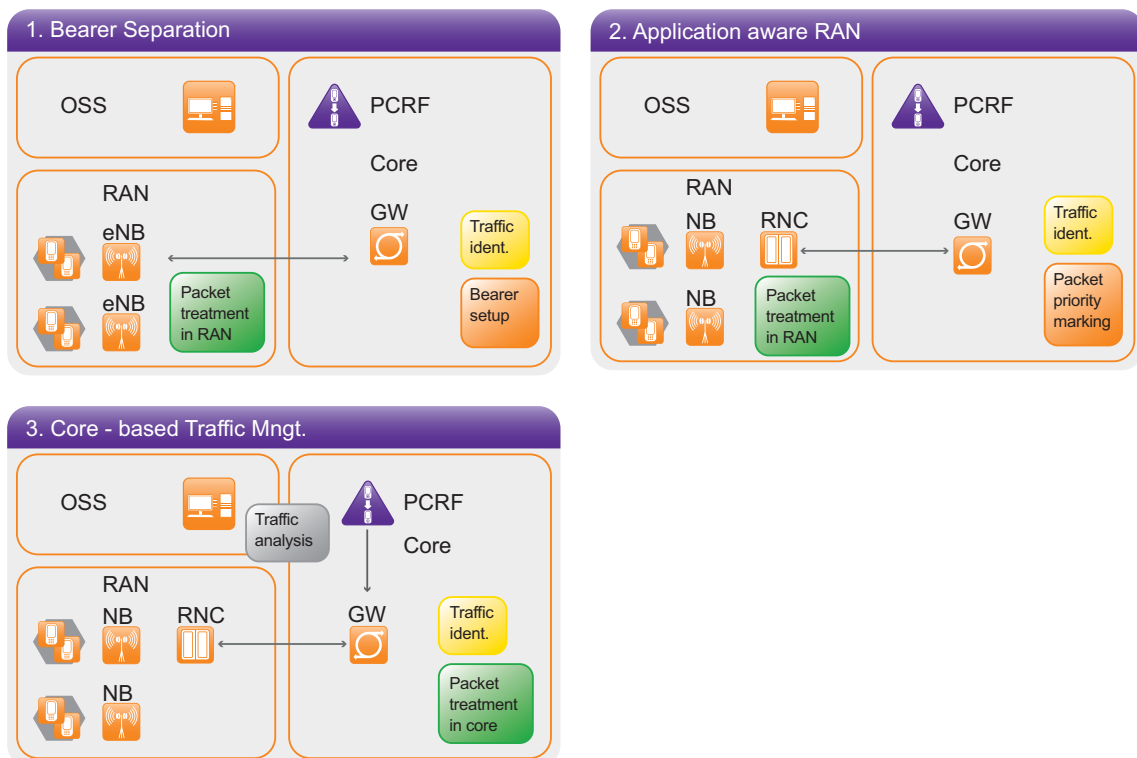


Figure 4. Solution space for cell-load and application aware traffic management

## Application-aware RAN

Since the secondary PDP context is not well supported in legacy 3G/HSPA networks and devices, application-aware RAN provides an alternative. User IP packets are inspected and marked in the mobile packet core according to the operator's application and subscription policies. The IP packet markings are monitored and evaluated in the RNC and fast radio QoS adaptation is applied according to predefined rules. The differentiated QoS treatment consists of in-bearer application flow differentiation combined with radio scheduling priority modification for the entire bearer. This provides reaction times on a millisecond timescale and leverages the radio resource management intelligence of the QoS-aware radio scheduler.

## Core-based traffic management

It's not always possible to apply a scheduler-based solution to avoid congestion. For example, a multi-vendor HSPA solution may not support packet marking for application-aware scheduling and packet treatment in the RAN. In this case, the recommendation is to perform traffic prioritization in the core based on traffic statistics. For example, high cell load conditions can be detected by measurements in the RAN and further insights can be gained by means of data gathered over several days or weeks. Corresponding policy control-enabled bandwidth management and enforcement can be triggered by the PCRF in the core, depending on the predicted load level.

The most suitable architecture for end-to-end QoS management depends on the existing network and the operator's business strategy and subscriber base. All the possibilities are considered in more detail in the following sections.

Whichever approach is taken, combining policy control and execution in a single, central control point is typically the most efficient implementation.

- 1 Policy management / QoS steering needs to be done centrally. The PCRF provides a common access point for all the management, charging and application functions that need to interface with the policy management function.
- 2 Conditioning the traffic flow for load-based distributed shaping can be achieved efficiently using traffic management at the service edge in the GGSN / packet gateway. This includes DPI for flow inspection and packet marking.
- 3 QoS-aware radio access has been introduced in base stations and the RNC. Fast-changing load conditions cause the RAN scheduler to make scheduling decisions in milliseconds for LTE and HSPA, taking into account the devices' radio-channel quality. This distributed policy execution enables the network to respond effectively to highly dynamic and fluctuating congestion levels in a radio cell and the rapidly changing quality of radio channels.

In summary, an efficient architecture for end-to-end QoS management is a functional split. Policy decisions are

made in the PCRF, traffic identification with DPI and bearer separation via dedicated bearers, secondary PDP context or packet marking takes place in the packet gateway and QoS-related packet treatments are carried out in the base station. The base station therefore decides how to make the best use of limited and oversubscribed resources. The efficiency of this approach translates into a higher sustainable load while maintaining the same QoE. Furthermore, it maximizes overall capacity, which is one of the key operating targets of a radio scheduler.

## 5.1 Bearer separation

This solution is based on the traffic separation and prioritization capabilities of the radio and core network using the EPS bearer functionality defined by 3GPP. Assigning radio resources to user traffic according to operator-defined traffic priorities and the actual load in the cell provides real-time reaction to varying cell load conditions. For example, the throughput of lower-priority applications (carried within dedicated bearers with appropriate QoS Class Identifier, QCI, values) will be reduced as the load increases. The solution combines the load-aware functionality in the RAN (eNB) with the application and policy awareness of the core. The solution complies with 3GPP LTE network architecture and requires the support of dedicated bearers.

The solution uses the EPS bearer concept, which establishes dedicated bearers in addition to the default bearer. A different set of QoS parameters can be assigned for each dedicated bearer. This enables the radio scheduler to assign resources to each bearer

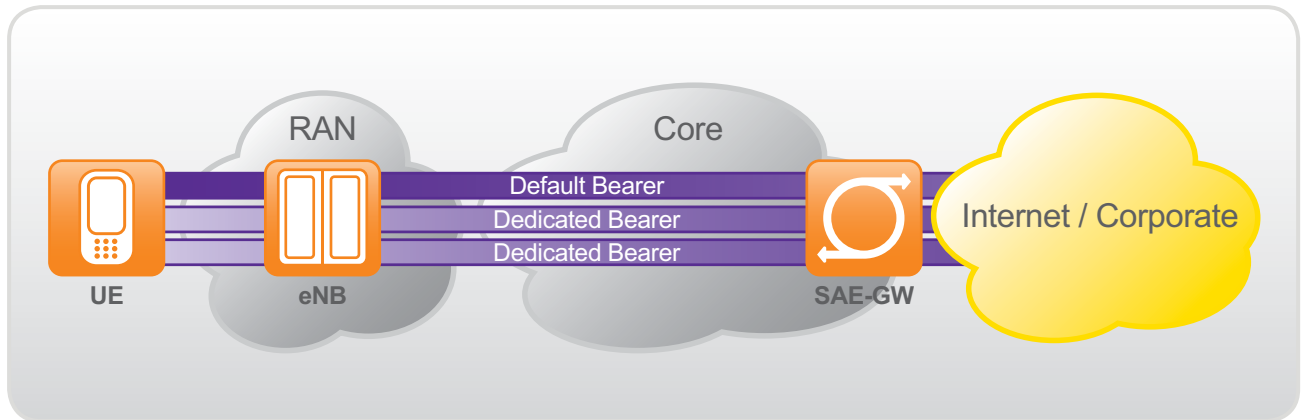


Figure 5. Bearer concept of EPS

according to the bearer's priority and the actual cell load, thus reducing the throughput of low-priority traffic to reduce the congestion probability. The bearer concept is shown in Figure 5.

The core provides the standard-compliant functionality for application differentiation using the TDF to determine the application and the PCRF plus PCEF functionality for the policy-controlled setup of dedicated bearers. A non-GBR dedicated bearer may carry several applications requiring the same QoS treatment.

### Detection and application awareness

Detection comprises two things: the detection of the application and the detection of the load in the cell.

DPI detects the application. The TDF may report the start and end of a detected application and the related service data flow descriptions to the PCRF, or it may act based on local policies. The PCRF may instruct the TDF on which applications to detect or the TDF may be pre-configured. It determines how to handle each packet flow for each subscriber in terms of the QoS parameters to be associated with the handling of that packet flow.

The radio scheduler in the eNB calculates the actual load (load detection functionality) and also provides enforcement functionality for this solution.

### Enforcement

Enforcement comprises the provision of a dedicated bearer in the core and priority-dependent scheduling in the RAN.

The PCRF sends policy rules to the PCEF (PDN-GW), which in turn are used as a trigger to establish a dedicated bearer. If a dedicated bearer with the requested QoS parameters exists already, the bearer will be modified by adding the Traffic Flow Template (TFT), which is the set of all packet filters associated with that bearer.

The radio scheduler in the eNB performs enforcement at the bearer level. It considers different parameters, including the bearer priority and the load to calculate which packets to send next. The QoS concept includes fast bearer prioritization to minimize the effects of instantaneous congestion. This excludes the lowest-priority bearers from the group of bearers that

can be scheduled. Thus the prioritized bearers can fulfill their required QoS, while low-priority users may suffer a temporary hold-up. The throughput of those users is boosted once the congestion is over.

### Key considerations

The standard LTE solution provides application differentiation at the bearer level, using scheduling to react in real-time to cell load conditions according to QoS priorities at the RAN level. It is a very flexible and efficient solution since no explicit congestion notification is needed. Instead, the relative priorities are assigned during bearer setup and the radio scheduler handles the different bearers according to the load.

Bearer separation provides very good support for QoE and efficiency management. The requirements of the different application types can be met in the radio and transport interface and the utilization of the most critical network resources can be maximized. The network planning and optimization can benefit from the understanding of the observed performance of the key applications in the radio and transport network as applications can be monitored separately.

The drawback of fast and efficient enforcement in the RAN is a potential waste of backbone resources, since packets sent via the backbone to the eNB might be dropped by the eNB under high load.

## 5.2 Dynamic traffic priority adaptation in application-aware RAN

This solution combines the real-time cell load awareness of the RAN with the application and subscription awareness of the mobile GW and DPI functionality in the mobile packet core. It optimizes the use of radio interface resources during high-load situations using IP packet marking in the mobile packet core and priority-based radio

scheduling and queuing in the RAN. This solution is supported end-to-end by Nokia Siemens Networks network elements (core network / gateway and RAN).

The system typically reacts to the current traffic mix within a bearer by dedicated IP packet flow priority queuing, and increasing or decreasing its radio scheduling priority on a short time scale. The effect of priority changes only becomes apparent during high-load situations. This approach can be used in 3G-HSPA networks without the support of a secondary PDP context.

The components of the solution are the application and policy-aware packet marking in the gateway and the priority-based scheduling in the RAN, see Figure 6.

Application and subscriber policies are taken into account, when downlink application-related packet flows are detected with DPI and the DSCP field of corresponding IP packets are marked by the mobile packet core.

Packet flow treatment in the RAN includes monitoring the DSCP fields of downlink user IP packets, in-bearer priority queuing of IP packet flows in the PDCP layer and dynamic adaptation of the radio scheduling priority according to DSCP monitoring evaluation rules.

### Detection and application awareness

In most mobile packet core deployments, DPI makes it possible to detect application-related IP packet

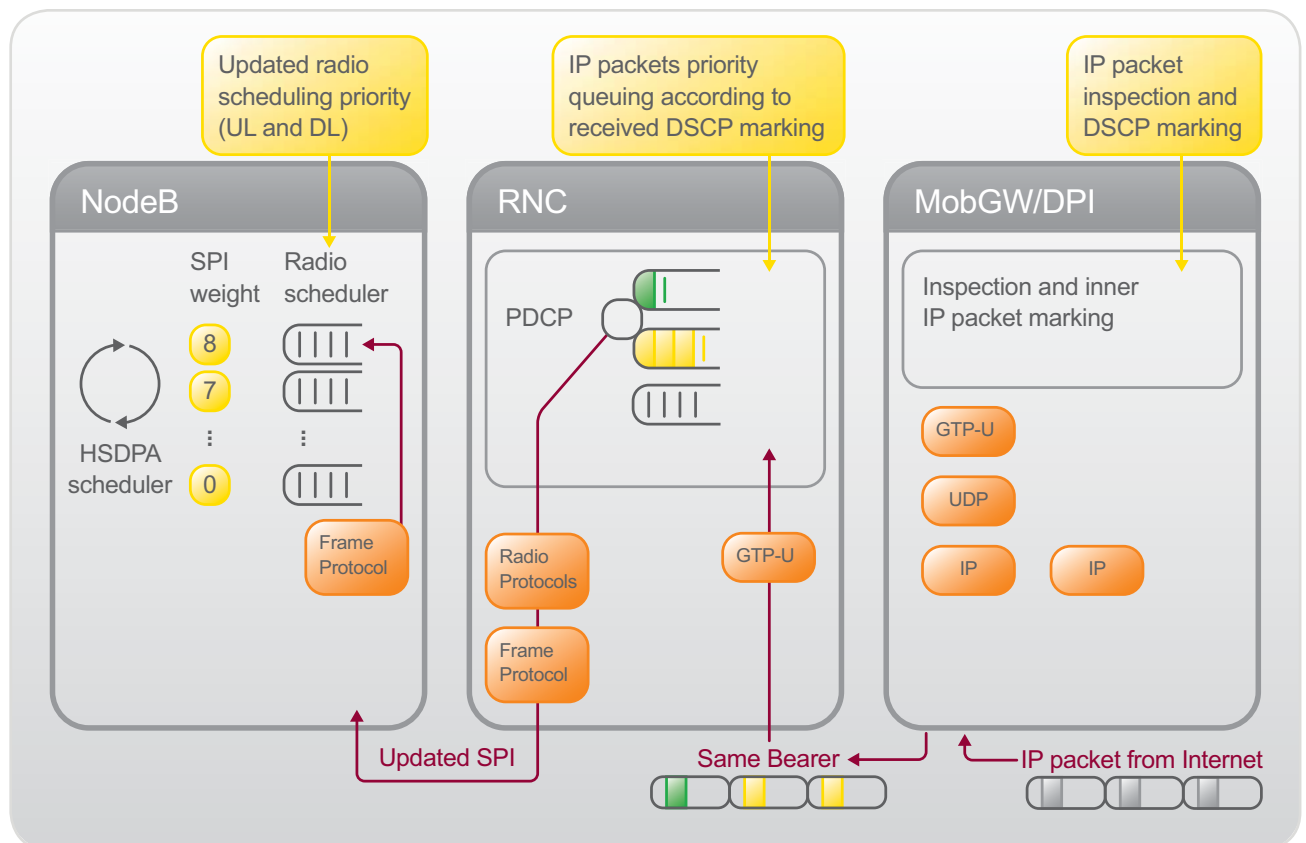


Figure 6. Application-aware RAN downlink traffic flow

flows. Furthermore, the mobile packet core provides full operator policy awareness via the Gx interface to the policy control server. This makes it possible to classify traffic according to policy and application. Thus, the combination of a mobile gateway and DPI can mark the user IP packet flows based on the detected applications and policies and send the information to the RAN. Packet marking is done by setting DSCP values for user IP packets according to predefined rules.

In the simplest scenario, packet flows could be classified as very important (“green”), less important (“yellow”) and normal (“white”). Since the entire range of DSCP values is available, more sophisticated schemes are possible in practice.

## Enforcement

In the RNC of a 3G RAN, the downlink user IP packets are monitored for DSCP marking. There are two mechanisms for treating dedicated traffic in the application-aware RAN.

First, application-aware in-bearer queuing provides PDCP packet buffering for each bearer. In the simple scenario outlined before, PDCP IP packet buffering takes place in three queues per bearer, serving green, yellow and white flows. The queues are populated based on IP packet marking and packets are forwarded to other protocol layers according to queue priorities. Green IP packet flows receive preferred treatment, reducing the latency. This especially helps in the case of multi-flow users, where priority queuing avoids “head of line” blocking by lower priority flows.

In addition, application-aware bearer prioritization uses the HSPA radio scheduler in the base station (NodeB) to allocate radio resources among competing radio bearers in real time, based on bearer-specific parameters such as QoS, radio condition and load situation. The scheduling priority indicator (SPI) is a key scheduling parameter, which is derived for each bearer by the RNC from the respective QoS parameters and conveyed to the base station, where it directly influences the share of radio resources. So adapting the SPI value of a bearer according to the traffic mix within the bearer will increase or decrease the respective bearer throughput over the radio and lub interface. In our example, the SPI value is up- or down-graded depending on the occurrence of green and yellow packets in the RNC. The new SPI value is applied in the up- and downlink to improve the TCP and/or application feedback loop latencies. During low cell load situations, the impact of an SPI change is negligible, whereas the impact can be considerable in high-load situations.

## Key considerations

This solution dynamically adapts the radio scheduling priority to the transient application mix and cell load conditions on a millisecond timescale. It increases the application throughput for a “promoted” flow and improves the QoE during high cell load situations. In-bearer priority queuing further improves the QoE for the promoted flows by providing them an increased share of the bearer bandwidth. Simulations have shown significant improvements for WEB

browsing sessions. They are sensitive to fast take-up of TCP flow bit-rates, what in-turn is the main condition for fast download of WEB pages. Application aware RAN’s short reaction time for bearer promotion combined with in-bearer preference for WEB session IP flows creates an optimal environment for TCP to quickly pick up speed. This results in significantly shorter WEB page download times also in high load conditions.

Application aware RAN does not set any specific QoS requirements for the UE. Furthermore, application differentiation within a single bearer reduces the capacity demand for simultaneously active bearers both in the RAN and core networks.

To consider operator application and subscription policies, IP packet marking rules and values must be consistently agreed between the mobile packet core and RAN. Additional standardization is not necessary since the DSCP concept is an existing IETF recommendation, widely supported by DPI equipment and mobile gateways. The specific DSCP values (for the “green” and “yellow” flows in our example) need operator and network-specific agreements.

As in bearer separation, Application aware RAN provides very good tools for QoE and efficiency management. Since the essential majority of the 3GPP cellular data traffic is delivered today over HSPA, the benefits of Application aware RAN also apply widely.

### 5.3 Core-based traffic management

This solution focuses on HSPA networks in single- and multi-vendor environments and is based on functions in the core network. Bandwidth management in the core is triggered by means of load measurements and evaluations performed in the GGSN or using OSS-based measurements of typical key performance indicators (KPIs) from the different 3G radio network elements to retrieve even more information. For third-party network elements, probes can be deployed at standardized interfaces, such as the Iu-PS and Gn. Based on the measured values, the congestion in the RAN can be calculated and bandwidth management requirements are indicated by the OSS or GGSN to the PCRF function in the core network.

#### Detection

The Nokia Siemens Networks version of this solution comprises two OSS components - the NetAct Reporting

Suite and the Traffica system. NetAct and Traffica can provide detailed insight into recurring patterns of congestion (time, geographical location, terminal types, affected users, type of application, and so on). Data is collected from the RAN and core network and stored by the OSS for detailed trend analysis. The collection of KPIs from elements in the RAN and core network is managed by NetAct, which analyzes and correlates the KPI data from different counters to detect congestion. Meanwhile, Traffica monitors signaling between the user and RNC and identifies subscribers. NetAct uses counter thresholds to detect cell load congestion in real time. Congestion detection in a particular radio cell or service area can be triggered by a combination of counter events and reports collected by NetAct from radio cells. These allow NetAct to distinguish between congestion patterns and patterns caused by overloads in signaling systems or faults in the network (see Figure 7). Such events indicate the number of active users, the scheduler capacity limit in a base station, the common channel

average load, the uplink power, or the use of Iu-PS capacity. Load data relevant to congestion estimation can also be collected and analyzed by the GGSN.

NetAct only measures the load for radio bearers and per SPI. It can't identify the individual subscriber. With Traffica for RNC/ core deployed, the congestion trigger can also be sent to Traffica to identify the affected subscribers in the congested cell (active RRC IMSI detection). In addition, applications and data volumes can be measured by IP Flow Analyzer (IPFA) for multi-vendor core networks. Those values are analyzed to assess the congestion status and trigger subscriber and application-sensitive bandwidth control by the PCRF.

#### Enforcement

Application- and policy-aware traffic management maintains the user experience. Policy awareness is realized via the interface with the policy control server and application awareness via traffic identification.

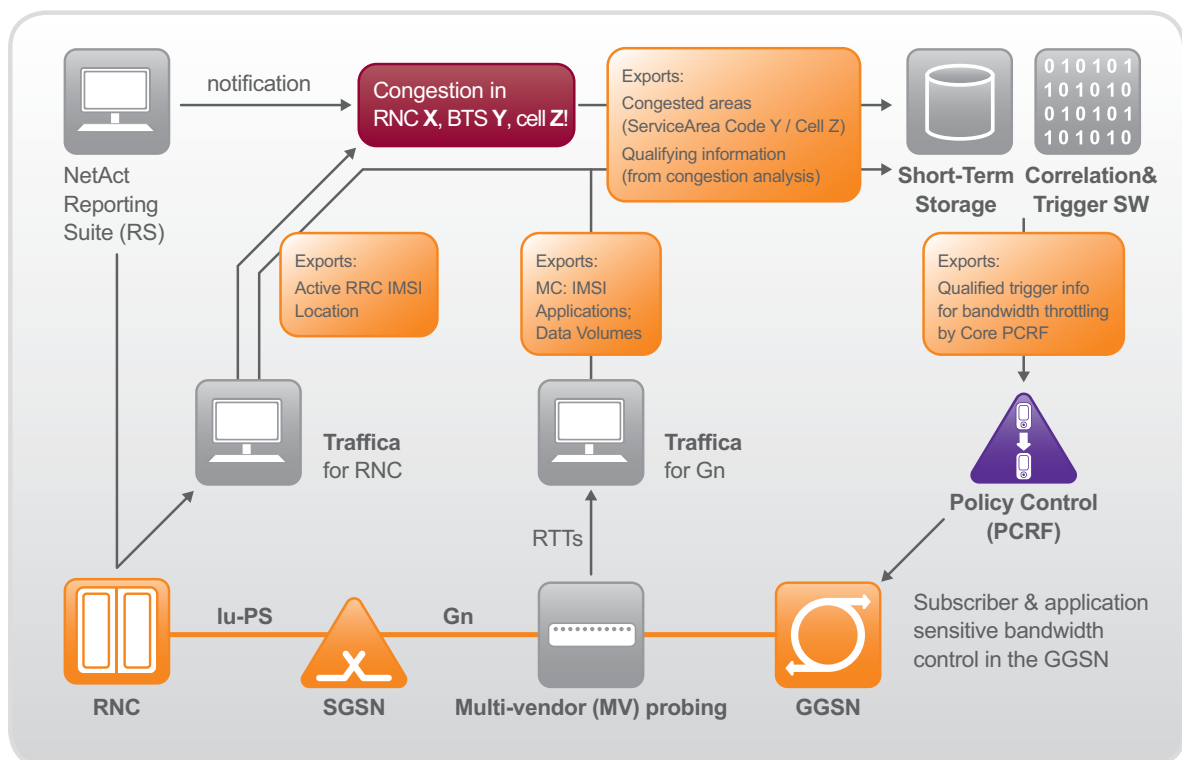


Figure 7. OSS-based congestion detection using NetAct and Traffica

Several user-based policies can be pre-configured for each PDN connection or retrieved via the Gx or SGi interfaces. An active PCC rule set can then be built within the GGSN for each session.

Traffic management is based on the analysis already described and enforced in the GGSN according to the relevant policy. Different rules can be applied according to the local policy configuration in the gateway or from the PCRF through the Gx interface. The GGSN supports admission control, which restricts bearer activation using allocation and retention priority (ARP) levels and triggers the RAN to check internal resources before accepting PDP context modifications. Additional QoS mechanisms comprise packet classification, packet queuing and packet scheduling for incoming and outgoing packets. Traffic policing forces the traffic in the PDP context to adjust to the reserved resources, discarding downlink packets that exceed the negotiated bit rate. Traffic shaping decreases traffic bursts by delaying packets. Shaping is done for streaming, interactive and background traffic classes in the downlink. Operators may also configure the mapping between traffic classes and outgoing packet queues to enhance packet prioritization within the GGSN.

The ability also exists in the GGSN to modify the QoS of the bearer at the start of a session or at any time during the session, as instructed by the PCRF.

Policy enforcement applies different traffic treatments to each subscriber,

according to their subscriptions and session-related data, such as the volume used or their current location. At the same time, the operator can control and manage network resources efficiently. New business opportunities are opening up with user-specific charging.

### Key considerations

The core-based traffic management solution suits multi-vendor deployments since the necessary functionality for detection and enforcement reside in one location and have no impact on other network elements. Policies are also defined in the same place that traffic management decisions are taken, making it easier to apply charging rules. Another big advantage is that resources in the mobile backhaul are conserved.

NetAct Reporting Suite and Traffica OSS components together provide instantaneous congestion detection and bandwidth control. By learning from patterns across the various counters, OSS-based solutions enable congestion forecasting with complex analysis of the conditions that trigger the PCRF function. Network elements can also analyze traffic patterns and understand the best trigger points for traffic throttling.

This solution makes information about the congested cell and active users available in the core, enabling selective throttling of the traffic streams in cells with high load. It also avoids downgrading applications and wasting capacity in cells when sufficient resources are available.

Since the congestion detection and application information is available in the core, the individual traffic streams can be adapted quickly to avoid congestion in the mobile backhaul.

Furthermore, where many high-priority streams cannot be served within a cell, even under optimized scheduling conditions, traffic management for an optimized QoE performance can be initiated in the core. This is an advantage of combining the system view, the flow-specific view and centralized mechanisms for efficient traffic management within the core. In general in the core network the traffic management methods are used to reduce the user data load, using typically policing and shaping the bitrates of the individual applications, and in some cases also advanced content optimization such as e.g. video, webpage and image compression.

The OSS system has interfaces to network elements from different vendors and introduces probes on standard interfaces. The triggering of activities and enforcement takes place in a single location in the core using the functions and interfaces of the PCRF. This solution therefore supports multi-vendor scenarios and can be easily integrated into existing networks.

However, instantaneous high cell-load conditions are not spotted as quickly as with mechanisms in the RAN. Extra signaling traffic is generated when information from the OSS is used, so there is a tradeoff between the accuracy of detection of cell load congestion and the signaling load of probes.

# Conclusions

The success of MBB and the growth in data traffic requires operators to focus on profitability and the user experience. Solutions require end-to-end mechanisms spanning network elements from the gateway to the mobile device. As traffic volumes grow, application differentiation is one way to use the infrastructure efficiently while safeguarding QoS and the user experience.

All the solutions described aim to prioritize traffic so that higher-priority flows get the QoS they need, even during peak hours. Lower priority flows will receive a best-effort service. DPI is applied to identify applications and to ensure that the type of service influences any decision process.

In general the radio network QoS differentiation methods, bearer separation and Application Aware RAN, provide very good support for QoE and efficiency management. The requirements of the different application types can be met in the radio and transport interface and the utilization of the most critical network resources can be maximized. Also the network planning and optimization can benefit from the understanding of the observed performance of the key applications in the radio and transport network. However, it's not always possible to apply radio network QoS differentiation methods to avoid congestion, for instance in multivendor environments. In this case, the recommendation is to perform traffic management in the core, based on traffic statistics. The traffic management methods in the core are

used to reduce the user data load, using typically policing and shaping the bitrates of the individual applications, and in some cases also advanced content optimization such as e.g. video, webpage and image compression.

The best architecture for end-to-end QoS management depends on the operator's existing network and business strategy, as well as the subscriber base. Even a combination of some functions of the different solutions might be the preferred choice, especially in networks comprising a number of technologies.

Nokia Siemens Networks supports operators in finding and applying the right technologies and solutions for end-to-end QoS, including those with multi-vendor networks.

# Glossary

3GPP	Third Generation Partnership Project
BSS	Business Support System
CSP	Communications Service Provider
DPI	Deep Packet Inspection
DSCP	Differentiated Service Code Point
eNB	Enhanced NB
eTOM	enhanced Telecom Operations Map
GGSN	Gateway GPRS Support Node
HSPA	High Speed Packet Access
IP	Internet Protocol
MBB	Mobile Broadband
NB	Node B
NGOSS	New Generation Operations Systems and Software
OCS	Online Charging System
OSS	Operations Support System
OTT	Over-The-Top
PCC	Policy Control & Charging
PCRF	Policy and Charging Rules Function
PCS	Policy Control Server
PDN	Packet Data Network
QoE	Quality of Experience
QoS	Quality of Service
RAN	Radio Access Network
RNC	Radio Network Controller
SDM	Subscriber Data Management
SGSN	Serving GPRS Support Node
SPI	Scheduling Priority Indicator
TDF	Traffic Detection Function

Nokia Siemens Networks  
P.O. Box 1  
FI-02022 NOKIA SIEMENS NETWORKS  
Finland  
Visiting address:  
Karaportti 3, ESPOO, Finland

Switchboard +358 71 400 4000 (Finland)  
Switchboard +49 89 5159 01 (Germany)

Copyright © 2012 Nokia Siemens Networks.  
All rights reserved.

Product code: C401-00740-WP-201109-1-EN

Nokia is a registered trademark of Nokia Corporation,  
Siemens is a registered trademark of Siemens AG.  
The wave logo is a trademark of Nokia Siemens Networks Oy.  
Other company and product names mentioned in this document  
may be trademarks of their respective owners, and they are  
mentioned for identification purposes only.

This publication is issued to provide information only and is not  
to form part of any order or contract. The products and services  
described herein are subject to availability and change  
without notice.