1

Basics of Performance Measurement in UMTS Terrestrial Radio Access Network (UTRAN)

Performance measurement represents a new stage of monitoring data. In the past monitoring networks meant decoding messages and filtering which messages belong to the same call. Single calls were analysed and failures were often only found by chance. Performance measurement is an effective means of scanning the whole network at any time and systematically searching for errors, bottlenecks and suspicious behaviour.

Performance measurement procedures and appropriate equipment have already been introduced in GSM and 2.5G GSM/GPRS radio access networks as well as in core networks, however, compared to the performance measurement requirements of UTRAN those legacy requirements were quite simple and it was relatively easy to collect the necessary protocol data as well as to compute and aggregate appropriate measurement results.

Nowadays even Technical Standard 3GPP 32.403 (*Telecommunication Management. Performance Management (PM). Performance Measurements – UMTS and Combined UMTS/GSM*) contains only a minimum set of requirements that is not much more than the tip of an iceberg. The definitions and recommendations of 3GPP explained in this chapter do not cover a wide enough range of possible performance measurement procedures, some descriptions are not even good enough to base a software implementation, and in some cases they lead to completely wrong measurement results. To put it in a nutshell it looks like the specification of performance measurement requirements for UTRAN is still in an early phase. This first part of the book will explain what is already defined by 3GPP, which additional requirements are of interest and which prerequisites and conditions always have to be kept in mind, because they have an impact on many measurement results even if they are not especially highlighted.

By the way, in the author's humble opinion, *the biggest error in performance measurement is the copy and paste error.* This results from copying requirements instead of developing concepts and ideas of one's own. As a result this book will also not contain ready-to-use performance measurement definitions, but rather discuss different ideas and

UMTS Performance Measurement: A Practical Guide to KPIs for the UTRAN Environment Ralf Kreher

offer possible solutions for a number of problems without claiming to cover all possibilities and having the only solutions.

1.1 GENERAL IDEAS OF PERFORMANCE MEASUREMENT

Performance measurement is fairly unique. There are many parameters and events that can be measured and many measurements that can be correlated to each other. The number of permutations is infinite. Hence, the question is: what is the right choice?

There is no general answer except perhaps the following: A network operator will define business targets based on economical key performance indicators (KPIs). These business targets provide the guidance to define network optimisation targets. And from network optimisation targets technical KPI targets can be derived, which describe an aspired behaviour of the network. Based on this, step by step, services are offered by operators. On a very common level these are e.g. speech calls and packet calls. These services will be optimised and detected errors will be eliminated. All in all it is correct to say that the purpose of performance measurement is to troubleshoot and optimise the network (see Figure 1.1).

However, whatever network operators do, it is up to the subscriber to finally evaluate if a network has been optimised in a way that meets customers' expectations. A rising churn rate (i.e. number of subscribers cancelling a contract and setting up a new one with a competitor operator) is an indicator that there might also be something wrong in the technical field.

Fortunately there is very good news for all analysts and market experts who care about churn rates: it is very difficult to calculate a real churn rate. This is because most subscribers in mobile networks today are prepaid subscribers, and since many prepaid subscribers are

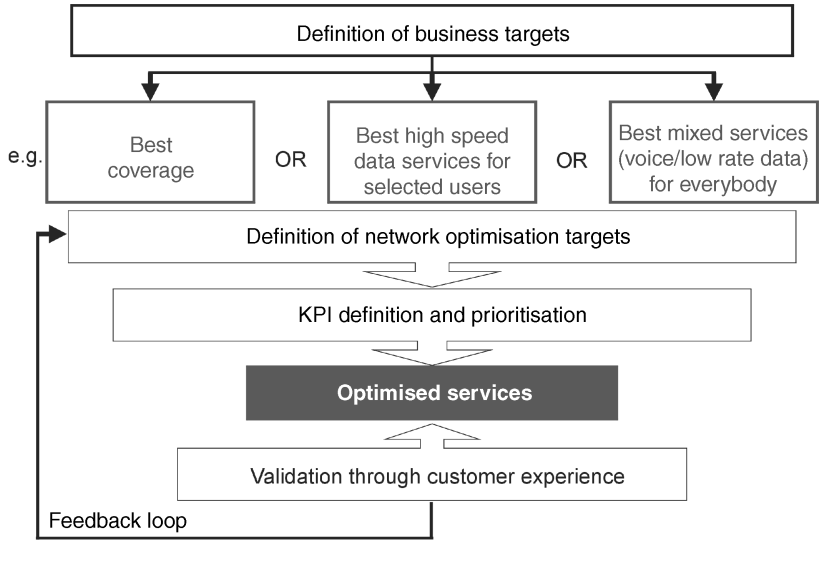

Figure 1.1 Network operator's optimisation strategy

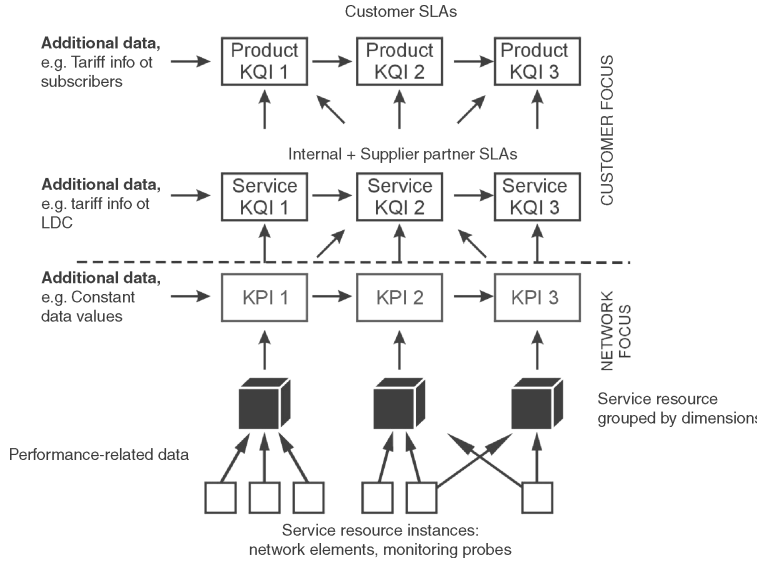

Figure 1.2 How to compute KPIs and KQIs

people who temporarily stay abroad, and based on the fact that prepaid tariffs are often significantly cheaper than roaming tariffs, such subscribers become temporary customers, so to speak. Once they go back to their home countries their prepaid accounts remain active until their contracts expire. Therefore not every expired contract is a churn. The actual number of churns is expected to be much less, but how much less? Additional information is necessary to find out about this.

The fact that additional information is necessary to compute non-technical key performance indicators based on measurement results (in this case based on a counter that counts the number of cancelled and expired contracts) also applies to the computation of technical KPIs and key quality indicators (KQIs). See Figure 1.2.

The general concept of these indicators is that network elements and probes, which are used as service resource instances, are placed at certain nodes of the network infrastructure to pick up performance-related data, e.g. cumulative counters of protocol events. In constant time intervals or in near real time this performance-related data is transferred to higher level service assurance and performance management systems. A typical example for such a solution is Vallent Corporation's WatchMark® software that is fed with performance data sent by radio network controllers (RNCs), mobile switching centres (MSCs) and GPRS support nodes (GSNs). For this purpose, e.g. an RNC writes the values of its predefined performance counters into a predefined XML report form every 15 minutes. This XML report file is sent via a so-called northbound interface that complies with the Tele Management Forum (TMF) CORBA specification to WatchMark® or any other higher level network management system. Additional data such as traffic and tariff models are provided by other sources and finally a complete solution for business and service management is presented.

As pointed out in www.watchmark.com the overall solution:

> . . . provides benefits across a service provider's entire customer base including pre-paid, post-paid and enterprise customers:
>
> - Service quality management provides an end-to-end visibility of service quality on the network to ensure that each service (e.g. MMS, WiFi, iMode, SMS and GPRS etc.) is functioning correctly for each user on the network.
> - Internal and 3rd Party service level agreements (SLAs) allow Service Providers to test, evaluate and monitor service levels within the organization to ensure that optimum service quality is delivered to customers.
> - Corporate SLAs enable Service Providers to establish specific agreements with their corporate customers where they undertake to deliver customized end-to-end levels of service quality.

However, there is one major problem with this concept: network elements that feed higher level network management systems with data are basically designed to switch connections. It is not the primary job of an RNC to measure and report performance-related data. The most critical part of mobile networks is the radio interface, and the UTRAN controlled by RNCs is an excellent place to collect data giving an overview of radio interface quality considering that drive tests that can do the same job are expensive (at least it is necessary to pay two people per day and a car for a single drive test campaign). Secondly, performance data measured during drive tests cannot be reported frequently and directly to higher layer network management systems. Therefore a great deal of important performance measurement data that could be of high value for service quality management is simply not available. This triggers the need for a new generation of measurement equipment that is able to capture terabytes of data from UTRAN interfaces, performs highly sophisticated filtering and correlation processes, stores key performance data results in databases and is able to display, export and import these measurement results using standard components and procedures.

Before starting to discuss the architecture of such systems it is beneficial to have a look at some definitions.

1.1.1 WHAT IS A KPI?

Key performance indicators can be found everywhere, not just in telecommunications. A KPI does not need to deal with only technical things. There are dozens of economical KPIs that can be seen every day, for example the Dow Jones Index and exchanges rates. The turnover of a company should not be called a KPI, because it is just a counter value, however, the gross margin is a KPI. Hence, what makes the difference between performance-related data and a KPI is the fact that a KPI is computed using a formula.

There are different kinds of input for a KPI formula: cumulative counter values, constant values, timer values seem to be the most important ones. Also KPI values that have been already computed are often seen in new KPI formulas.

Most KPI formulas are simple. The difficulties are usually not in the formula itself, but e.g. in the way that data is first filtered and then collected. This shall be demonstrated by using a simple example. Imagine a KPI called *NBAP Success Rate*. It indicates how many

NBAP (Node B application part) procedures have been completed successfully and how many have failed.

NBAP is a protocol used for communication between Node B (the UMTS base station) and its CRNC (controlling radio network controller). To compute a *NBAP Success Rate* a formula needs to be defined. In 3GPP 25.433 standard for Node B Application Part (NBAP) protocol or in technical books dealing with the explanation of UMTS signalling procedures (e.g. Kreher and Ruedebusch, 2005) it is described that in NBAP there are only three kinds of messages: Initiating Message, Successful Outcome and Unsuccessful Outcome (see Figure 1.3).

Following this a *NBAP Success Rate* could be defined as shown in Equation (1.1):

$$\textbf{NBAP Success Rate} = \frac{\sum NBAP\ Successful\ Outcome}{\sum NBAP\ Initiating\ Message} \times 100\% \tag{1.1}$$

This looks good, but will lead to incorrect measurement results, because an important fact is not considered. There are two different classes of NBAP messages. In class 1 NBAP procedures the Initiating Message is answered with a Successful Outcome or Unsuccessful Outcome message, which is known in common protocol theory as acknowledged or connection-oriented data transfer. Class 2 NBAP procedures are unacknowledged or connectionless. This means only an Initiating Message is sent, but no answer is expected from the peer entity.

Since most NBAP messages monitored on the Iub interface belong to unacknowledged class 2 procedures (this is especially true for all NBAP common/dedicated measurement reports) the *NBAP Success Rate* computed using the above defined formula could show a value of less than 10%, which is caused by a major KPI definition/implementation error.

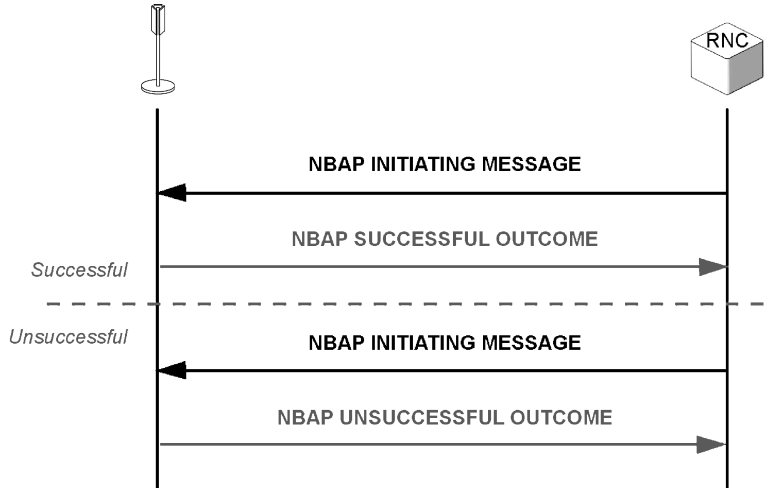

Figure 1.3 Successful/Unsuccessful NBAP call flow procedure

Knowing the difference between NBAP class 1 and class 2 procedures a filter criteria needs to be defined that could be expressed as follows:

$$\textbf{NBAP Class 1 Success Rate} = \frac{\sum NBAP\ Successful\ Outcome}{\sum NBAP\ Class\ 1\ Initiating\ Message} \times 100\% \qquad (1.2)$$

An exact definition is usually not expressed in formulas, but more often by fully explaining in writing the KPI definition. A couple of examples can be found in Chapter 2 of this book. The lesson learnt from the *NBAP Success Rate* example is that one cannot compare KPIs based on their names alone. KPIs even cannot be compared based on their formulas. When KPIs are compared it is necessary to know the exact definition, especially the filter criteria used to select input and – as explained in next chapter – the aggregation levels and parameter correlations.

Never trust the apparently endless lists of names of supported KPIs that can be found in marketing documents of network and measurement equipment manufacturers. Often these lists consist of simple event counters. Therefore, it must be kept in mind that additional data is always necessary as well as simple counter values to compute meaningful KPIs and KQIs.

1.1.2 KPI AGGREGATION LEVELS AND CORRELATIONS

KPIs can be correlated to each other or related to elements in the network topology. The correlation to a certain part of the network topology is often called the aggregation level.

Imagine a throughput measurement. The data for this measurement can be collected for instance on the Iub interface, but can then be aggregated on the cell level, which means that the measurement values are related to a certain cell. This is meaningful because several cells share the same Iub interface and in the case of softer handover they also share the same data stream transported in the same Iub physical transport bearer that is described by AAL2 SVC address (VPI/VCI/CID). So it may happen that a single data stream on the Iub interface is transmitted using two radio links in two or three different cells. If the previously mentioned throughput measurement is used to get an impression of the load in the cell it is absolutely correct to correlate the single measurement result with all cells involved in this softer handover situation.

To demonstrate the correlation between mobile network KPIs an example of car KPIs shall be used (see Figure 1.4). The instruments of a car cockpit show the most important KPIs for the driver while driving. Other performance-relevant data can be read in the manual, e.g. volume of the fuel tank.

The first KPI is the speed, computed by the distance driven and the period of time taken. Another one is the maximum driving distance, which depends on the maximum volume of fuel in the tank. Maybe the car has an integrated computer that delivers more sophisticated KPIs, such as fuel usage depending on current speed, and the more fuel needed to drive a certain distance influences the maximum driving distance. In other words, there is a correlation between fuel usage and the maximum driving distance.

Regarding mobile telecommunication networks like UTRAN similar questions are raised. A standard question is: How many calls can one UTRA cell serve?

Network equipment manufacturers' fact sheets give an average number used for traffic planning processes, e.g. 120 voice calls (AMR 12.2 kbps). There are more or less calls if

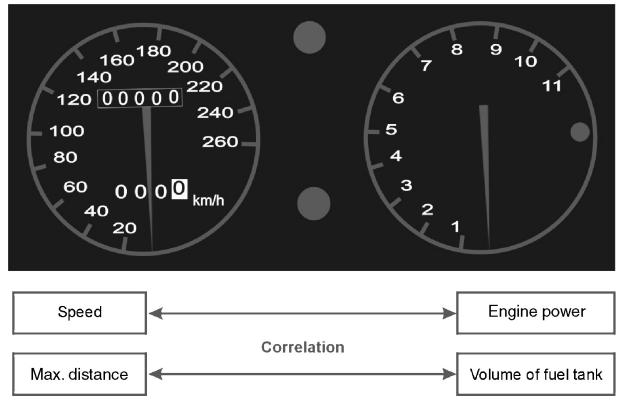

Figure 1.4 Correlation between car KPIs

different services such as 384 kbps data calls or different AMR codecs with lower data transmission rates are used. The capacity of a cell depends on the type of active services and the conditions on the radio interface, especially on the level of interference. Hence, it makes sense to correlate interference measurements with the number of active calls shown per service. This combination of RF measurements requires sophisticated KPI definitions and measurement applications. The first step could start with the following approach: Count the number of active connections per cell and the number of services running on those active connections in the cell.

Before continuing with this example it is necessary to explain the frame conditions of this measurement, looking at where these counters can be pegged under which conditions and how data can be filtered to display counter subsets per cell and per service.

1.1.3 BASIC APPROACH TO CAPTURE AND FILTER PERFORMANCE-RELATED DATA IN UTRAN

The scope of this book is UTRAN performance measurement. Within UTRAN four interfaces exist where performance-related data can be captured: the Iub interface between Node Bs and RNC; the Iur interface between different RNCs; the IuCS interface between RNCs and the CS core network domain; and the IuPS interface between RNCs and the PS core network domain. For each interface a specific protocol stack is necessary to decode all layers of captured data as explained in detail in Section 1.2, which deals with the functions and architecture of performance measurement equipment. Usually this equipment is able to automatically detect to which specific interfaces a probe is connected and which protocol stacks are necessary to decode captured data. If necessary it can also detect on which particular channel data is transmitted. This especially refers to dedicated and common transport channels on the Iub interface. In addition, it can be assumed that the same equipment also provides a function that is commonly known as *call trace*, which allows for the automatic detection and filtering of all messages and data packets belonging to a particular connection between a single UE and the network. For a detailed overview of all interfaces, channels and call procedures it is recommended to read the appropriate chapters

in Kreher and Ruedebusch (2005). From a performance measurement expert's perspective it is expected that these functions are provided and work as required to decode and aggregate performance-related data. Nevertheless, in this chapter a few basic network procedures need to be explained, that apply to all scenarios, because they may be relevant for any call or at any time during an active connection.

Our approach is as already defined in the previous section. Count all connections in a cell and provide a set of sub-counters that is able to distinguish which services are used during these connections. From a subscriber's perspective this scenario is simple. They switch on their mobile phones, set up calls, walk around or drive by car (which could result in a couple of mobility management procedures) and finish their calls whenever they want.

Now from a network operator's perspective it is necessary to find out in which cell the calls are active and identify the type of service related to each particular call. It sounds easy, but due to the specific nature of the UTRAN procedure it is indeed a fairly complicated analysis.

When the term 'service' is used in the context of performance measurement this usually applies to end-user services such as voice calls, data calls and – if available in network and if the UE is capable – video-telephony calls. All kinds of supplementary services such as conference calls or multi-party connections are seen as special cases of the above categories and are not analysed in detail. However, when looking at data calls the type of service can also be determined from the TCP/IP application layer, e.g. file transfer (FTP) or web browsing (HTTP). These specific services are beyond the scope of this basic approach for two reasons. Firstly they require a more complex correlation of measurement data, secondly it makes no sense to define a TCP/IP analysis at cell level, because even the smallest email or website is segmented by the RLC into a number of different transport blocks and theoretically each transport block set can be transmitted using a different cell.

There is another well-known service from GSM, which is also available in UMTS. This service is called short message service (SMS). A short message is not sent using a dedicated traffic channel, it is sent piggybacked on signalling messages. Plain signalling is also necessary to register a mobile phone to the network after being switched on. There is no payload transmitted between subscriber and network, but nevertheless signalling is essential and for this reason another service type called 'signalling' will be defined in addition to 'voice', 'data' and 'video-telephony' in this basic approach.

Now the question is how to distinguish the four different services by monitoring protocol messages.

A CS call set up always starts with a Call Control Setup message as specified in 3GPP 24.008. The 'decision maker' that distinguishes between voice calls and video-telephony calls is the value of the bearer capability information element within this Setup message. If the bearer capability information element shows the value 'unrestricted digital info' the call is a video-telephony call. Another indicator is the signalling access protocol I.440/450 and rate adaptation following H.223 & H.245 mentioned in the same message. See Figure 1.5.

It is difficult to explain what a bearer is. Maybe the following definition is the best one: A bearer is a temporary channel used to transport a data stream (user or network data) with a defined quality of service. (All definitions in this book are given by the author using his own words. Standard definitions may be more exact, but are often not very understandable.)

This is true for both GSM and UMTS, but in UMTS the bearer concept covers all possible data streams in each part and layer of the network while in ISDN/GSM it is only used to

TS 24.008 Call Control V3.11.0 (CC-DMTAP) SETUP (= Setup)	
Setup	
Protocol Discriminator	call control, call related SS message
Transaction Id value (TIO)	TI value 0
Transaction Id flag	message sent from orig TI
Message Type	5
Send Sequence Number	Send Sequence Number = 2
Bearer Capability	
IE Name	Bearer Capability
IE Length	10
Info transfer capability	Unrestricted digital info
Transfer mode	Circuit mode
Coding standard	GSM standardized coding
Radio channel requirement	Full rate channel
Extension bit	No Extension
Establishment (Octet 4)	Demand
Neg of Intermed Rate Req	No meaning is associated with this value
Configuration	Point-to-point
Duplex Mode	Full duplex
Structure	Unstructured
Compression	data compr.not possible
Extension bit	No Extension
Signalling access protocol (Oct. 5)	I.440/450
Rate Adaption	Other rate adaption
Access ID	Octet identifier
Extension bit	Extension
Spare	0
Other rate adaption	H.223 & H.245
Other ITC	restricted digital information
Extension bit	No Extension
Synchronous/asynchronous (Oct. 6)	Synchronous
User Info L1 Protocol	Default layer 1 Protocol
Layer 1 ID	Octet identifier

Figure 1.5 Call Control Setup message for video-telephony call

define the characteristics of traffic channels between subscribers. A service from the point of view of UTRAN is always bound to a certain type of (radio) bearer and hence, analysing characteristics of UMTS bearer services is another possible definition of 'call type' and is completely different from the approach given in this chapter which is based on NAS signalling analysis.

Looking back to the specific signalling used between the UE and the CS core network domain it emerges that in contrast to video-telephony calls voice calls have the bearer capability value 'speech' in the Call Control Setup message. A PS connection (data call) always starts with a Service Request message. This Service Request indicates that there is data (IP payload) to be transmitted, but it should be noted that this definition might not always fit to the user's perspective of an active PS call.

Imagine a subscriber starting a mobile web-browsing application. For this purpose a PDP context is established between the UE and the SGSN and a traffic channel, which is called the radio access bearer (RAB) is provided. Now a website is downloaded and the user starts to read its contents. This may take a while. Besides the user may switch to another application while keeping the web-browser open. This is not a problem in fixed data networks. IP data is only transmitted when necessary, if there is no data transfer no network resources of the fixed line are occupied. That does not apply to UTRAN. Here dedicated resources (these are the codes used to identify channels on the radio interface) need to be provided for each RAB. And those resources are limited. That is the reason why the network needs to identify which resources are really used. All other resources are released to prevent

shortage and guarantee subscriber satisfaction. This leads to a situation that a PDP Context that is bound to the open web-browsing application remains active in the UE and SGSN while a RAB is released if the network detects that no data is transmitted for a certain time. Based on this a PS connection in UTRAN is defined, whereas an active PS RAB and RAB assignment is always triggered by a Session Management Service Request message.

RABs are also set up for CS connections, but for conversational calls they are active as long as a call is active. Indeed, there are several ways to count the number of active connections, which means that there are different protocol messages from different protocol layers. The advantages of the method described in this chapter are:

1. Non-access stratum (NAS) signalling messages can be found on both Iub and Iu interfaces.
2. NAS messages contain information elements that allow direct identification of the call type. In the case of e.g. an RANAP RAB Assignment Request it can only be guessed from the UL/DL maximum bit rate and traffic class mentioned in this message which call type is related to the RAB. This requires additional mapping tables running in the background of the performance measurement application (note that this is an alternative option).
3. Setup and Service Request messages may contain user identifiers that allow further filter options (e.g. count all active connections per cell, per call type, per UE) and are helpful for troubleshooting.

To complete the call type definition, 'signalling' constitutes all call flows between the UE and core network domains that do not contain Setup or Service Request messages. It is also necessary to define another category that is usually called 'Multi-RAB' and describes a UE that has at least two active connections (RABs) simultaneously. Multi-RAB calls can be a combination of CS and PS services for one UE, but multiple PS RABs are also possible, for instance if PS streaming video requires the set up of a secondary PDP context that triggers the establishment of a second PS RAB for the same UE. This second RAB provides a different traffic class (= different delay sensitivity) and different maximum bit rates. An example for such a kind of Multi-RAB PS+PS would be a GPRS session management message Activate Secondary PDP Context Request. Figure 1.6 shows the different filter options.

Protocol events used to determine the call type cannot immediately be used to count the number of active connections, because they only describe connection attempts. Therefore, it is necessary to check if the attempted connection has been set up successfully. This can be done on the RRC, RANAP or NAS layer. On the Iub interface the RRC Radio Bearer Setup Complete message indicates that a traffic channel has been established successfully. Following this the RANAP RAB Assignment Response is sent on the Iu interface while the NAS layer indicates that the connection between A-party and B-party has been established. For PS calls the session management Service Accept and PDP Context Activation Accept messages could be used as additional indicators for a successful connection. It should be noted that in the case of video telephony calls via the CS domain in-band signalling is also necessary to really get the service running. This in-band signalling is transmitted using the radio (access) bearer and the example proves that there are different perspectives of user and network and it clarifies the need to have different KPIs for those different perspectives.