CHAPTER 2

STATE OF THE ART ANALYSIS

The 3GPP and 3GPP2 have standardized their own IMS specifications. Internet Engineering Task Force (IETF) also collaborates with them in developing protocols that fulfill their requirements. This chapter discusses the IMS architecture, Session Initiation Protocol, Policy Based Network and also the need to support the QoS in 3G networks. This chapter introduces the 4G networks, and its research opportunities and the thesis objective.

2.1 OVERVIEW OF IMS ARCHITECTURE

IMS was introduced in the 3GPP architecture release version 5 to support IP multimedia services over GPRS and UMTS access networks. IMS latest specification is release version 7 (TS 23.228, 2006). IMS uses SIP for the control of sessions (session signaling) and IPv6 for the network transport. It offers an excellent solution to integrate session setup at the application and network levels.

Figure 2.1 shows the IP Multimedia Core Network Subsystem reference architecture. The subsystem is simplified for comprehension purposes as shown in Figure 2.2. The common nodes included in the IMS are as follows:
Figure 2.1  IP Multimedia core network subsystem reference architecture

Figure 2.2  Simplified view of IMS
2.1.1 Call Session Control Functions (CSCF)

The CSCF acts as a call server and handles call signaling. It supports and controls the multimedia sessions, and provides the flexibility to add, modify or delete bearers used by the user’s service. The protocol that is used for the majority of the signaling is SIP.

CSCF (Frene et al 2001) handles the following functions:

- **Call control function**: This function executes call setup/termination and state/event management. This is an evolution of the MSC call control function.

- **Address translation function**: This function performs address analysis, translation, modification, and mapping.

- **Serving profiling database**: This function interacts with HSS to receive and cache user profile.

- **Incoming call gateway**: This function acts as an entry point and routes incoming calls.

The CSCF can be functionally decomposed to Proxy-CSCF (P-CSCF), Interrogating-CSCF (I-CSCF), and Serving-CSCF (S-CSCF).

2.1.1.1 Proxy - CSCF

The P-CSCF is the first contact point in the visited IMS network.

The main functions of P-CSCF (TS 23.228, 2002) include:
• Providing authorization of bearer resources and QoS management.

• Forwarding the SIP register request received from the UE to an I-CSCF determined using the home domain name, as provided by the UE.

• Forwarding SIP messages received from the UE to the SIP server (e.g. S-CSCF).

• Forwarding the SIP request or response to the UE.

• Terminating and independently generating SIP transactions in abnormal conditions.

2.1.1.2 Interrogating - CSCF

I-CSCF is the contact point within an operator’s network for all connections destined to a subscriber of that network operator, or a roaming user currently located within that network operator’s service area. There may be multiple I-CSCFs within an operator’s network. The main functions of I-CSCF (TS 23.228, 2002) include:

• Assigning an S-CSCF to a user performing SIP registration. It performs load balancing between the S-CSCFs with the support of the Home Subscriber Subsystem (HSS).

• Interrogating the HSS during mobile terminated session’s setup to obtain the address of the S-CSCF catering for the mobile, and then forwarding the SIP request or responding to it.
2.1.1.3 Serving - CSCF

The S-CSCF is the node that performs the session management for the IMS network. There can be several S-CSCFs in a network. They can be added as per the need of nodes capabilities or the network’s capacity requirements. The S-CSCF may be chosen differently based on the services requested or the capabilities of the mobile (TS 23.228, 2002). The main functions of S-CSCF include:

- Accepting registration requests from UE and making its information available through the HSS.
- Providing session control for the registered UE's sessions i.e. the S-CSCF in the home network is responsible for all session control. This means that the mobile is not restricted to the capabilities of the visited network as seen in the current wireless network.
- Accepting requests and servicing them internally or forwarding them.
- Terminating and independently generating SIP transactions.
- Interacting with services platforms for the support of services.
- Providing endpoints with service events related information (e.g. notification of tones/announcement together with location of additional media resources, billing notification).

2.1.2 Home Subscriber Subsystem

HSS is the centralized subscriber database evolved from the Home Location Register (HLR). The HSS interfaces with the I-CSCF and the
S-CSCF to provide information about the location of the subscriber and its subscription information. The HSS is responsible for holding the following user related information (TS 23.228, 2002):

- User identification, numbering and addressing information.
- User security information, which includes network access control information for authentication and authorization.
- User location information at inter-system level i.e. supporting user registration, and storing inter-system location information, etc.
- User profile information.

2.1.3 Media Gateway and Media Gateway Control Function

In an environment where all of the sessions are between IP capable end user devices, there is no need for anything other than the CSCFs and the HSS. In reality, there is a very long transition period to completely eliminate the legacy PSTN and mobile networks (Narayan Parameshwar and Chris Reece).

IMS supports several nodes for inter-working with legacy networks. They are Media Gateway (MGW), Media Gateway Control Function (MGCF), and Transport Signaling Gateway (TSGW.)

2.1.3.1 Media gateway control function

MGCF controls the call state for media channels in a media gateway. It communicates with the CSCF and performs protocol conversion
between legacy call control protocols and IMS call control protocols. For example, the MGCF receives a SIP message from the CSCF and converts it into appropriate ISUP messages and then sends it via IP to the Transport Signaling Gateway.

### 2.1.3.2 Media Gateway

The Media Gateway refers specifically to the IMS entity and can be termed as IMS-MGW. Its primary function is to convert media from one format to another. In IMS, this predominantly converts media between Pulse Code Modulation (PCM) in the PSTN and an IP based vocoder format networks (Narayan Parameshwar and Chris Reece).

### 2.1.3.3 Transport Signaling Gateway

T-SGW is the signaling end point of interworking with PSTN/legacy networks. It maps call-related signaling protocols from/to PSTN on an IP bearer and sends it to/from the MGCF. The T-SGW converts the lower layers of SS7 into IP. The application layer protocols (for example, ISUP) are not being affected. It is important to note that it is always an option to have the MGCF support SS7 and then, T-SGW will not be required.

### 2.1.4 IMS Protocols

IMS works with a number of protocols. In designing IMS protocols, 3GPP leveraged the work of other Standards Development Organizations (SDOs) such as the IETF and ITU-T by reusing existing protocols. The following are some of the protocols:
• **SIP (RFC3261):** IETF defined this protocol and is chosen as the session control protocol for the IMS.

• **DIAMETER (RFC3588):** This provides an authentication, authorization, and accounting (AAA) framework for billable communications and accounting services.

• **COPS (RFC 2748):** This is a query and response protocol for exchanging policy information. This IETF protocol is used for communication of QoS within the IMS architecture.

• **H.248/MEGACO:** H.248 is an International Telecommunications Union Telecom Standardization Sector (ITU-T) standardized Gateway Control Protocol (GCP). IETF endorses this protocol and refers to it as Media Gateway Controller (MEGACO). IMS Media Gateway (MGW) uses H.248 for media conversion and provides end-to-end communication.

• **RTP/RTCP (RFC3550):** This is used in IMS as the media transmission protocol for end-to-end delivery of services. RTCP is used for feedback on the transmission and reception quality of data carried by RTP. IMS relies on these protocols for transfer of real time media such video and audio.

• **SCTP:** This is designed to transport PSTN signaling messages over IP networks. As a reliable transport protocol, SCTP has other applications such as delivery mechanism for multimedia and for wireless.
2.2 SESSION INITIATION PROTOCOL

In telecommunication networks, there are two categories of traffic. First one is the traffic related to control signaling (hereafter referred as signaling), which is used to establish, manage and terminate communication sessions. Second one is the actual data traffic (e.g.; voice). SIP signaling follows the concept of common channel signaling, in which the path used for the signaling traffic is independent of the path used for the actual data traffic. Separation of signaling traffic from media makes the session management more efficient, and is also more adaptive to functional changes.

SIP is an application-layer protocol for initiating, modifying, or terminating communication and collaborative sessions over IP networks. A session can be an IP telephony call, a multi-user conference that incorporates voice, video and data, instant messaging chat, or multi player online game. SIP can be used to invite participants to a scheduled or already existing session. Participants can be a person, an automated service or a physical device such as a handset. It can also be used to add or remove media to/from a session.

SIP signaling supports the following facets of multimedia session management:

- **User location:** This feature enables users to access telephony or other application features from remote locations.

- **User availability:** This feature determines the willingness of the called parties to engage in communication sessions.

- **User capabilities:** This feature determines the media and its parameters to be used for communication sessions.
- **Session setup:** This feature establishes the session parameters for point-to-point and multi-party sessions.

- **Session management:** This feature enables the transfer and termination of sessions, modification of session parameters, and the invocation of session services.

### 2.2.1 SIP Component Architecture

This section describes the architectural components of SIP - the user agent and the network servers. SIP messaging and the extensions, which are introduced in section 2.2.2, show how these components participate in call flow to support SIP functionality. The SIP user agents are the peer components that initiate and answer calls. SIP architecture defines the following functional elements:

- **User Agent:** A SIP User Agent (UA) is an end device, which can originate and receive SIP calls. It can be a phone terminal (mobile, PDA or Laptop) or an endpoint such as an answering machine. SIP supports both peer-to-peer and client server architectures. User agents act as peers; they find each other and negotiate session characteristics.

- **User Agent Server:** In the client server model, when sending requests or receiving responses, a SIP UA acts as the client. In this case it is referred to as the User Agent Client (UAC). The receiving SIP UA acts as the server (receives requests and sends responses) and is referred as the User Agent Server (UAS). UAC and UAS are logical entities that are contained in every SIP User Agent.
• **Back-to-Back UA (B2BUA):** When a SIP entity acts both as a User Agent Client, as well as the User Agent Server it is referred as the Back-to-Back User Agent (B2BUA). It generates requests to determine how the incoming request is answered.

• **Proxy server:** The SIP proxy server is a key component of SIP infrastructure. Its role as an edge routing server is similar to that of a Web proxy server. It provides routing capabilities and supports functions such as authentication, accounting, registration, and security. The SIP proxy server is the first entity that receives all outgoing requests from a SIP UA. It routes the request traversing intermediate servers until it locates the server closest to the destination SIP UA, which forwards the request to the called SIP UA. In the most common scenarios there are usually two SIP proxy servers - one at the caller end and one at the callee end. Proxy Servers can be configured to transaction stateful or stateless. Stateless proxy servers receive incoming requests and simply pass on responses without retaining any information about the transaction. On the other hand, stateful proxy servers retain information on all incoming requests, the server's responses and outgoing messages from the server. A SIP infrastructure can contain a combination of stateful and stateless proxy servers. The stateful servers can be configured closer to the SIP user agents to collect billing and other user relevant data, whereas the stateless proxy servers form the backbone of the network. SIP proxy servers can perform 'forking' process, where it sends out SIP INVITE to multiple devices at once. In a case, where a user is registered at multiple locations, a
Forking Proxy Server would send an incoming SIP INVITE message (for a session) to each registered location. When the user responds from one of the locations (upon receiving a SIP OK message from that location) the proxy server sends a SIP CANCEL message to the other locations. In order to perform forking process, the proxy server needs to be configured to transaction stateful.

- **Registrar:** The Registrar is a repository of user agent’s location information. The registrar accepts registration requests from user agents and places the information (the SIP address and associated IP address) in location database. A SIP REGISTER message tells the Registrar (and the network) at which address (or multiple addresses) the user is available henceforth (say office phone during the day). Once the location or device changes the user agent has to send another SIP REGISTER message to the Registrar. SIP proxy servers (and redirect servers) make use of the location information stored in the repository to obtain the callee user agent location(s).

- **Redirect Server:** Redirect Servers respond to SIP request with an address where the SIP message should be redirected. It maps a destination address (in the SIP message) to one or more addresses and returns the new address list to the originator of the SIP request. The location of the intended recipient is retrieved from the location database maintained by the SIP Registrar. Redirection is used for Call Forwarding and it also helps to reduce the processing load on proxy servers by pushing the processing back to the requesting clients.
2.2.2 SIP Messages

SIP signaling - the setting up, modification, and termination of communication and collaboration sessions - is realized through the exchange of messages. There are two types of messages: requests and responses. Requests are sent to initiate some action and responses are sent as replies to requests. Replies acknowledge receipt of requests and indicate the processing status. Requests and responses share a common message format, which consist of a start-line, one or more header fields, an empty line indicating the end of the header fields, and an optional message-body. The SIP message structure is illustrated in Figure 2.3.

![SIP Message Structure Diagram]

Figure 2.3 Structure of SIP message

The start-line in SIP messages can be either a request or a status line. Request messages use the request line to set the type of request. Response messages indicate whether the processing of a request is successful or not in the status line. SIP message headers consist of fields with name-
value pairs, where some header fields are optional, such as content type and length, while others are mandatory for every SIP message. Table 2.1 lists the mandatory header fields for SIP messages.

Table 2.1  SIP message mandatory header fields

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>To</td>
<td>The request destination SIP address</td>
</tr>
<tr>
<td>From</td>
<td>Indicates the originator of the request</td>
</tr>
<tr>
<td>Cseq</td>
<td>The command sequences that ensure messages are dealt with in the order they were generated.</td>
</tr>
<tr>
<td>Call-ID</td>
<td>A randomly generated string that uniquely identify SIP sessions.</td>
</tr>
<tr>
<td>Via</td>
<td>Used to route responses in the reverse direction</td>
</tr>
<tr>
<td>Contact</td>
<td>Contains the actual location of the callee, which might be different from the address of the originator in the From header.</td>
</tr>
</tbody>
</table>

### 2.2.2.1 SIP Requests

SIP requests have a Request-line for their start-line. The format of a Request-line is illustrated in Figure 2.4. It consists of three fields that are separated by a single space (SP) character.

![Figure 2.4 Format of a request message start line](image)
• **Method:** This field indicates the method to be performed. RFC 3261 identifies six methods: REGISTER, INVITE, ACK, CANCEL, BYE and OPTIONS. SIP extensions documented in other RFCs have defined additional methods.

• **Request-URI:** The request URI field holds a SIP or SIPS URI. It is used to indicate the user or service to which the request is addressed.

• **SIP-version:** The SIP version field identifies the version of SIP protocol that is in use.

Table 2.2 presents a list of basic request methods and a brief description of what each method does.

### 2.2.2.2 SIP responses

SIP response message is sent as a receipt of a request or when a proxy server triggers a response. SIP response messages are distinguished by the fact that they have Status-line in their start-line. The Status-line consists of three fields SIP-version, Status-code, and Reason-phrase. These fields are separated by a single space (SP) character. The format of the Status-line is illustrated in Figure 2.5.

![Figure 2.5 Format of a response message start line](image)

**Figure 2.5 Format of a response message start line**
Table 2.2 List of SIP basic methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>RFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>REGISTER</td>
<td>This method is used to provide information to the Registrar to specify the UA’s location and its availability for incoming SIP requests. When the user agent’s location changes, another REGISTER message is sent to update the Registrar’s database.</td>
<td>RFC 3261</td>
</tr>
<tr>
<td>INVITE</td>
<td>This method is used to initiate a communication session between two UA peers. Message is sent by a user to initiate a session with another peer user. INVITE can also be used to initiate a multi-party call.</td>
<td>RFC 3261</td>
</tr>
<tr>
<td>ACK</td>
<td>This method is used for acknowledgement. It indicates that the final response has been received.</td>
<td>RFC 3261</td>
</tr>
<tr>
<td>BYE</td>
<td>This method is used to indicate the termination of a session.</td>
<td>RFC 3261</td>
</tr>
<tr>
<td>CANCEL</td>
<td>This method is used to terminate pending requests. A calling party can cancel an INVITE message before it receives the final response.</td>
<td>RFC 3261</td>
</tr>
<tr>
<td>OPTIONS</td>
<td>This method is used to query a server on its capabilities. For example; it can be used to query if a to-be-called party can support a particular type of media.</td>
<td>RFC 3261</td>
</tr>
</tbody>
</table>
Table 2.2 Continued

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>RFC</th>
</tr>
</thead>
<tbody>
<tr>
<td>INFO</td>
<td>This method communicates additional information about an active session. For example, it can be used to inform a user agent when available pre-paid balance is approaching to zero.</td>
<td>RFC 2976</td>
</tr>
<tr>
<td>UPDATE</td>
<td>This method is used to change session information (such as a change in CODEC) before a final response to a SIP INVITE message has been received. Typically, UPDATE messages contain an SDP body that lists the modified session parameters.</td>
<td>RFC 3311</td>
</tr>
<tr>
<td>SUBSCRIBE/NOTIFY</td>
<td>A user can subscribe to events such as the Presence information of another user. Subscribe Method is sent to the SIP Presence Server. When the second user becomes available, the Presence Server replies with a NOTIFY Method.</td>
<td>RFC 3265</td>
</tr>
<tr>
<td>REFER</td>
<td>This method provides a third party’s contact information to the recipient. This method can be used for call transfers.</td>
<td>RFC 3515</td>
</tr>
</tbody>
</table>
- **SIP-version**: The SIP version field identifies the version of SIP protocol that is in use.

- **Status-code**: The status-code is a three-digit code, which represents the outcome of request processing. The range of values is between 100 and 699. The first digit indicates the class of the response (SIP/2.0 allows for 6 possible classes).

- **Reason-phrase**: This is a short textual description of the Status-code. Status-code is intended for machine processing, whereas the reason-phrase is a human-readable message that is rendered to the user by the user agent.

Table 2.3 provides a brief overview of the response classes.

### 2.2.3 SIP signaling procedures

A standard SIP call setup, as shown in Figure 2.6, consists of a three-way handshake. First the initiator sends an INVITE message to the recipient and the recipient can respond to the INVITE with a range of responses (provisional or final). The INVITE message includes the Session Description Protocol (SDP) data, which negotiates parameters for this particular call, such as, which CODEC and QoS parameters to use. Once the session is established, multimedia packets can be transferred between the two end points using RTP (RFC 3550). When the call is finished the BYE message is sent and OK 200 is finally sent back to complete the call termination.
### Table 2.3 List of SIP response code

<table>
<thead>
<tr>
<th>Code Range</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
</table>
| 1xx        | Provisional/Informative: Provisional response indicates that the associated request was received and being processed. Upon receipt of a provisional response, the request sender should stop retransmitting the request. | 100 Trying  
180 Ringing  
183 Progress  |
| 2xx        | Success: Success responses with Status-codes in the range from 200 to 299. This indicates that the request was received, understood, and successfully processed. | 200 OK  |
| 3xx        | Redirection: When further action such as a different location is needed to complete a request, redirection responses are used to provide the new location or an alternative service that would satisfy the request. | 301 Moved permanently  
302 Moved temporarily  |
| 4xx        | Client error: Client error response Status-codes are sent when requests cannot be processed. | 404 Not found  
407 Proxy unauthorized  |
| 5xx        | Server error: Server error response Status-codes are sent in cases where the request is valid but the server is unable to fulfill the request. | 503 Service unavailable  |
| 6xx        | Global failure: When any server cannot fulfill a request, the Global failure response Status-codes are returned. | 603 Decline  |
Figure 2.6 Basic SIP call between two UA’s

Figure 2.7 shows how the same call could look for a stateful proxy that acts as a server for client requests it receives and for a client to forward its requests.

Figure 2.7 Basic SIP call flow using SIP Proxy
2.2.4 Session Description Protocol

When a SIP is used to set the multimedia streams across the network there is a requirement for QoS to also be provided. In fact, SIP itself does not provide QoS mechanisms, but it does provide information via the SDP content about the service requirements. Each proxy server in the call path can determine the required bandwidth and request the QoS by looking at the SDP content in the INVITE and OK 200 messages. In Figure 2.7, the QoS procedure is negotiated in a basic SIP call but in this case the handshake will usually be three-way

1. UA1 offers UA2 a range of media streams (QoS parameters)
2. UA1 replies with answer confirming which media streams it would like to use as well as its QoS requirements
3. UA1 and UA2 perform QoS reservation.

2.3 POLICY BASED NETWORK

Several research and development efforts show the growing interest in policy-based networking. These efforts are in academia and industry, working groups leading standardization efforts, new technical conferences, and new commercial products supporting policy-based management. However, the idea of using policies in network management is not new. The original idea is known to have evolved in the early 1970s (Lewis 1996), to monitor and control the access rights of resources in large distributed systems. With the evolution of Internet, there is considerable increase in complexities and heterogeneity of underlying networking technologies and increase in the number of resources to be managed. One way to address the complexity of such a network is to use a policy base framework to automate the network management functions. The policy-based approach can be used to manage
different aspects of a network, commonly known as policy disciplines (Verma 2000). Some examples of policy disciplines are QoS, network security, and IP address allocation. Policy-based networking configures and controls the various operational characteristics of a network as a whole. This provides the network operator with a simplified, logically centralized, and automated control over the entire network.

In Westerinen et al (2001), a policy is defined as “a definite goal, course or method of action to guide and determine present and future decisions.” In general, policies can be seen as plans of an organization to achieve its objectives. This may involve a set of rules to govern the behavior of its network and its components (resources, users, applications, etc.), and the specification of a set of actions to be performed. Policies can be classified (Rajan et al 1999, Verma 2000) into different levels in a hierarchy allowing simplified abstraction of complex low-level policies to simple high-level policies that do not use networking jargon.

Business-level or high-level policies are those that express the overall goals of an organization. Network level policies are essentially business level policies mapped and expressed into networking terminology. These are defined and entered by a network operator with a high-level perspective of the network topology, objectives and network-wide utilization. Node-level policies are those that correspond to the objectives and requirements at the different network nodes. Device-level policies are device-specific instructions that facilitate implementation of algorithms. For example, classification, scheduling, buffer management etc. The node and device level policies typically constitute the low-level policies. In order to successfully deploy policies in a network, the policies need to satisfy certain requirements (Verma 2000). They should be precisely defined, easy to understand and enforced at a network element. The policies must be
compatible with the capabilities of the network element on which they are enforced. Furthermore, policies must be mutually consistent to avoid conflicts and ambiguous decision-making. Finally, the policies should be simple, intuitive, and easily understood at a higher-level by human operators. The network operator should also be able to specify them with ease.

IETF and the Distributed Management Task Force (DMTF) have been working together to define a policy framework. The IETF Policy Framework working group provides guidelines for defining a policy framework. It also defines an information model and schemata to define, store and retrieve policies (Westerinen et al 2001). The architectural elements typically found in a policy-based system are as shown in Figure 2.8.

![Figure 2.8 Key architecture elements of a policy-based management system](image)

A Policy Management Tool (PMT) provides the network administrator with an interface to interact with the network. A network administrator uses the policy management tool to define the various policies or policy groups. It is typically the function of the PMT to validate the syntactic and semantic correctness of the administrator input, to ensure consistency among the high-level policies and to check for compatibility of the various policies. Further, the PMT typically determines the association
between the policies and the various network elements where these policies are to be enforced. It also determines the low-level policies that can be used to support the specified high-level policies and ensures the specified policies are comprehensive enough to cover all the relevant scenarios. The policies specified at the PMT are then stored in a policy repository. A policy repository can be defined as a data store or a model abstraction that holds policy rules, their conditions and actions, and related policy data (Westerinen et al 2001). A Policy Information Base (PIB) can be considered as a type of policy repository. The concept of PIB is based on the Structure of Management Information (SMI) (McCloghrie et al 1999) to leverage the experience with the Simple Network Management Protocol (SNMP) (Stallings 1993) and related Management Information Bases (MIBs).

The Policy Decision Point (PDP) or the policy server generally retrieves the policies from its repository. Then it performs complex policy interpretation and translation into a format that can then be used to configure one or more Policy Enforcement Points (PEPs) or policy clients. The PDP also needs to monitor any changes in the policies that might occur at the policy management tool (shown in Figure 2.8 as a dashed line) or repository. A policy management tool may not detect policy conflicts at a lower level, and such conflicts have to be handled by the PDP. The PEP is a network device (For example, end-host or router) where the policies are actually executed or enforced. The PEP is also responsible for monitoring any relevant information (such as installation/removal of policies, updates about its current status, etc.) and reporting it to the PDP to facilitate automated efficient network management.
2.4 QUALITY OF SERVICE

Quality of Service is a set of requirements to be met so that a service or application can be delivered to the end-user in a satisfactory manner. The QoS level can be quantified by packet loss probability, guaranteed bandwidth, end-to-end (E2E) delay and jitter, and reflects how the traffic flows through a network. In other words, QoS can be seen as the degree of satisfaction of an end-user for a delivered service.

The above concept leads to the basic idea of differentiating QoS according to different traffic types. For example, QoS may correspond to the different features and demands to the networks, or different services to be delivered to the customers based on tariffs. Consequently, different mechanisms must be implemented to provide/ensure the E2E features of applications matching their traffic type. Two main categories of mechanisms are:

- **QoS provision mechanisms**, which include parameters mapping, admission and resource reservations schemes.

- **QoS control mechanisms**, which include traffic shaping, scheduling, policing and resource control mechanisms.

Network operators offer QoS with Service Level Agreements (SLAs). A SLA is a contract to specify the transit of services through network domains.
2.4.1 Application Point of View

In packet-switched networks, the applications are viewed in two different ways with respect to QoS – real-time and non-real-time (elastic) as proposed by IETF. Real-time applications deliver time-sensitive information, where the data blocks must be displayed consecutively at predetermined time intervals, thus requiring specific delay, jitter and error parameters. On the other hand, the applications that don't include time-sensitive information are much more tolerable to delay and jitter but more sensitive to error parameters. Another way was proposed by 3GPP in UMTS network, where applications are classified into four types based on the generated traffic. They are conversational, streaming, interactive, and background as specified in Table 2.4.

Before stepping into more detailed discussion on IP QoS, UMTS QoS and IMS QoS viewpoints, it is necessary to first describe some well-known QoS metrics for presenting E2E service/application requirements, such as delay, jitter, loss rate, and throughput.

- **Delay**: It is the elapsed time for a packet to traverse the network from the source to the destination. At the network layer, the end-to-end packet latency is the sum of processing, transmission, queuing, and propagation delays.

- **Jitter**: It is defined as the variation in delay encountered by similar packets following the same route through the network. The jitter requirement only affects real-time streaming applications as this QoS requirement arises from the continuous traffic characteristics of this class of applications. Jitter is generally included as a performance parameter since it
is very important at the transport layer in packetized data systems, due to the inherent variability in arrival times of individual packets. Services intolerant of delay variation usually try to reduce the delay variation by means of buffering. However, late data arrivals make data useless, resulting in receiver buffer underflow, and early arrival can lead to receiver buffer overflow.

- **Loss Rate**: Loss rate refers to the percentage of data loss among all the delivered data in a given transmission time interval, which can be evaluated in frame level or packet level. Loss rate requirements apply to all classes of applications. In general, real-time applications tolerate a limited amount of data loss, depending on the error resiliency of the decoder, and the type of application. On the other hand, non-real-time applications typically have much more strict requirement on data loss.

- **Throughput**: It is defined as the rate at which packets are transmitted in a network. It can be expressed as a maximum rate or an average rate.

### 2.4.2 IP based QoS

The IP based QoS point of view can be best seen in this scenario: the transport of IP packet pass through a set of network and one of them is UMTS network. The known approaches for dealing with IP QoS are Integrated Services (IntServ) and Differentiated Services (DiffServ).

DiffServ is an attempt to design a simple architectural framework for QoS that can provide a variety of scalable end-to-end services across
multiple separately administered domains. These domains do not require complex behaviors in forwarding equipment. This approach is interesting in the sense that it overcomes many issues such as scalability, inter-operation and administration without using high signaling. Actually, unlike IntServ, DiffServ minimizes signaling by using Per-Hop-Behaviors (PHBs).

Integrated Service provides application requirements to the network, which have to maintain QoS mechanisms to ensure the promised QoS. The Resource Reservation Protocol (RSVP) transports the QoS requirements along the path from the sender to the receiver in order to make resource reservation. It sets state in router per flow and is supposed to gain a better E2E performance even in case the networks along the path are congested, since there are enough transmission resource reserved for all the existing data flows.

2.4.3 Policy-Based QoS

The IETF Resource Allocation Protocol (RAP) working group is active in the field of QoS policy. It has defined, among other standards, the policy-based admission control framework (Yavatkar 2000) and the Common Open Policy Service (COPS) protocol (Durham et al 2000) and its extension-1. It is noteworthy that a policy decision point (PDP) is typically a sub-component of a policy server (Strassner 1999). However, often the terms PDP and policy server are used synonymously (Westerinen et al 2001, Verma 2000, Kosiur 2001).

COPS for Provisioning (COPS-PR) (Chan et al 2001) is a simple query protocol that facilitates communication between the policy clients and remote policy server(s). Two policy control models have been defined:
outsourcing and provisioning, as illustrated in Figure 2.9. While COPS supports the outsourcing model, its extension COPS-PR integrates both the outsourcing and provisioning models. The outsourcing model is tailored to signaling protocols such as RSVP (Braden et al. 1997, Herzog et al. 2000), which requires traffic management on a per-flow basis. On the other hand, the provisioning or configuration model is used to control aggregate traffic-handling mechanisms such as the DiffServ architecture (Blake et al. 1998).

In the outsourcing model, when the PEP receives an event (e.g. RSVP reservation request) which requires a new policy decision; it sends a request (REQ) message to the remote Policy Decision Point (PDP). The PDP then makes and sends a decision (DEC) message (e.g. accept or reject) back to the PEP. The outsourcing model is, thus, PEP-driven and involves a direct 1:1 relation between PEP events and PDP decisions.

Figure 2.9  Policy control models

On the other hand, the provisioning or configurations model (Chan et al. 2001) makes no assumptions of such direct 1:1 correlation between PEP events and PDP decisions. The PDP may proactively provision the PEP reacting to external events, PEP events, and any combination thereof (N:M correlation). Provisioning thus tends to be PDP-driven and may be
performed in bulk (For example, entire router QoS configuration) or in portions (For example, updating a DiffServ marking filter (Chan et al 2003).

### 2.4.4 QoS in UMTS

The 3GPP defines UMTS QoS classes as:

- Conversational class
- Streaming class
- Interactive class
- Background class

The main distinguishing factor between these QoS classes is how delay sensitive the traffic is. The Conversational class on the one end is meant for traffic that is very delay sensitive while the Background class on the other end is the most delay insensitive traffic class. Table 2.4 shows the application associated with the types of traffic types and their main characteristics.

<table>
<thead>
<tr>
<th>Traffic Class</th>
<th>Example Application</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversation (RT)</td>
<td>Voice/video telephony</td>
<td>Low E2E delay, jitter, two-way</td>
</tr>
<tr>
<td>Streaming (RT)</td>
<td>Streaming Video</td>
<td>Low jitter, one way</td>
</tr>
<tr>
<td>Interactive (BE)</td>
<td>Web browsing</td>
<td>Two way, Low loss rate/error</td>
</tr>
<tr>
<td>Background (BE)</td>
<td>Email, background download</td>
<td>One way, Low loss rate/error</td>
</tr>
</tbody>
</table>
To provide these QoS requirements, 3GPP proposed the concept named Bearer Service. To realize a certain network QoS, a Bearer Service with clearly defined characteristics and functionality is to be set up from the source to the destination of a service. A bearer service includes all aspects required to enable the provision of a contracted QoS. These aspects are the control signaling, user plane transport and QoS management functionality. UMTS bearer service layered architecture is depicted in Figure 2.10. Each bearer service on a specific layer offers its individual services using services provided by the layers below. The Bearer Service concept is important as it allows us to divide the E2E QoS problem into a set of sub-problems concerning the QoS provision inside a specific Bearer Service and the QoS profile mapping among them (between 2 Bearer Services in the same plane or between an upper layer and a lower layer).

Figure 2.10 UMTS QoS architecture
2.4.5 QoS in IMS

The IMS supports several end-to-end QoS models (3GPP TS23.207). Terminals may use link layer resource reservation protocols (PDP context activation), RSVP or DiffServ codes directly. Generally, the common model to support terminal systems is to use link-layer protocol. UMTS uses at the GGSN to map link-layer resource reservation flows. In the IMS, the P-CSCF must instruct the terminal to perform resource reservation. The QoS reservation decisions for the IMS are based on two criteria – resource and policy. The resource criterion simply states whether the network has enough capacity to accept the call and whether the reservation must be made locally at the point of control. The RNC controls radio resources in the RAN while the SGSN controls the resources of GPRS core network. For the policy criterion, the network allocates bandwidth for the subscriber. In IMS policy decision that is used in Common Open Policy Service (COPS) (RFC 2748) can be centralized and it is carried out by an entity called the Policy Decision Function (PDF) (3GPP TS29.207). In 3GPP Release 5, the PDF is a logical entity that can either be co-located with the P-CSCF or in a stand-alone unit. In 3GPP Release version 6, the P-CSCF and the PDF may be separated and the protocol is standardized as shown in Figure 2.11.
In COPS protocol, the GGSN acts as the policy enforcement point (PEP) and the PDF acts as the PDP. The role of PDF consists of the characters in the session by the GGSN from the user agent that is authorized to establish particular parameters. In the PDF, 3GPP provides a mechanism for the network to authorize the establishment of media streams. It is known as the Service-Based Local Policy (SBLP) (3GPP TS23.207) and is based on the network inserting a media authorization token. The media authorization token is generated by the PDF from the IMS terminals. When the terminals request the establishment of a media stream in the SIP message (3GPP TS29.208) the SBLP model follows the media authorization model (RFC 3521). This is shown in Figures 2.12 and 2.13. The import point of the PDF is generating an authentication token that cause the P-CSCF to add a P-media-Authorization header field (RFC 3313) in incoming INVITE message. The terminal adds and modifies this token (RFC 3520) in the messages when it sends to reserve network resources. On the other hand, the P-CSCF uses the single reservation flow (SRF) semantics (RFC 3524) of the
SDP grouping framework (RFC 3388). This framework allows grouping of media streams and describes the semantics of the group (3GPP TS24.229), as shown in Figures 2.12 and 2.13.

Figure 2.12 Authorization token transfers in an INVITE request
2.5 SESSION ESTABLISHMENT SCENARIO

In 3G IMS, the 3GPP defines the Go interface to facilitate the necessary communication between the PDF on behalf of the P-CSCF and the PEP on behalf of the GGSN. This realizes the policy based QoS. The 3GPP has chosen to use the Common Open Policy Service – Policy provisioning (COPS-PR) protocol (RFC 3084). It is an extended version of the base COPS protocol as the communication protocol on the Go interface. It provides a simple request-response framework to transport policy information reliably between the PEP and the PDF over a TCP connection. Figure 2.14 explains the COPS and COPS-PR in the 3G IMS to provide correlation of application layer signaling and network resources.
Figure 2.14 Session establishment scenario using 3GPP release-6 architecture
2.6 HETEROGENEOUS NETWORKS

In the last few years, many new wireless networks gained popularity. The main driver for these new networks is capacity. Also, as the wireless usage continues to expand, existing systems are reaching their limits. The next step requires the heterogeneity of the access networks. The wireless 4G must allow the coexistence of different mobile technologies and provide a differentiated set of services to the end user. However, the provisioning of differentiated services over heterogeneous wireless networks poses several challenges. This section defines the limitations of 3G and presents the benefits and issues of the internetworking of different access networks.

2.6.1 Limitations of 3G and WLAN

3G can support multimedia Internet type services at improved speed and quality as compared to 2G. As it has been pointed out in the previous chapter, for example, the W-CDMA based air interface is designed to provide improved high capacity for medium bit rate (384 Kbps) and limited coverage at up to 2 Mbps (in indoor environments). Statistical multiplexing on the air also improves the efficiency of packet mode transmission. However, there are limitations in 3G as follows:

- Extension to higher data rates is difficult with CDMA due to excessive interference between services
- It is difficult to provide a full range of multi rate services, all with different QoS and performance requirements due to the constraints imposed on the core network by the air interface standard.
The bandwidth available in the 2 GHz bands allocated for 3G will soon become saturated. It will be a major constraint for the combination of frequency and time division duplex modes imposed by regulators to serve different environments efficiently. Also, as users move farther from a base station, interference from other cells can weaken the signal and cause channel errors. In addition, a system that works with both voice and data will not get the maximum throughput. Users demand that wireless voice communications offer the same quality as wire line phone technology. Therefore, wireless systems must devote resources to voice communication’s quality of service under all circumstances, which in turn reduces maximum data performance.

2.6.2 Benefits of Internetworking Access Networks

The internetworking of access networks offers considerable benefits for the mobile stations as well as for the networks. The mobile station should be able to switch seamlessly from different access networks. For example, for a WWAN network, mobile station improves its quality of service and increases capacity by reducing interferences. Also, new services can be made available in the WLAN networks. The expected benefits for the cellular network are three-fold:

- Coverage area and interference reduction
- Improved transmission bit rate
- Support for hierarchical service area

2.6.2.1 Area coverage and interference reduction

In a 3G networks, terminals that are far away from the base station need to use proportionally large amount of transmission power. In the relay communication case i.e. when using an intermediate mobile or access point in
between, the transmission power for the cellular link can be reduced. This fact has a direct consequence on the reduction of interference. Based on the above observations, an increase in capacity of the system is expected.

2.6.2.2 Transmission bit rate improvement

The 3G systems achieve a maximum bit rate of 2 Mbps but the bit rate may decrease in a vehicular-speed environment. Wireless LANs and broadband wireless access systems use 5 GHz frequency bands. For example, 802.11 and HiperLAN offer greater than 30 Mbps transmission capability in an indoor environment. For the heterogeneous networks, 2 to 100 Mbps transmission is realized in an outdoor/vehicular environment.

2.6.2.3 Hierarchical service area

Although the expectation is that all devices will be connected to a network through wireless links, it may be difficult for small devices to be directly connected to the 4G system due to power consumption and antenna size. However, compact devices will be capable of exchanging wireless signals at short range. Therefore, compact devices will be able to access the 4G network through a miniature BS, which will act as a MT for the 4G system. By employing such a configuration, service areas will consist of multiple overlapping cells.

2.6.3 Internetworking Issues

Several issues associated with the concept of heterogeneous access technologies need to be addressed, including radio and IP resource allocation, handoff management, QoS provision, routing protocol, and security (e.g., user authentication in vertical handoff). This is because each of the access
networks has its own infrastructures, link characteristics (e.g., data rates, latency, delay, signal power), and mobility management process (e.g., registration, coverage discovery and handoff decision). These differences are highlighted below:

### 2.6.3.1 Signal quality

The signal powers received from the base stations of the involved networks are incomparable. For example, for a WWAN, due to the long distance between two units, the mobile stations and base stations have to transmit at higher powers in order to communicate. In WLAN or WPAN network, the limited range (100 m) implies low power.

### 2.6.3.2 Data rate

The networks support different data rate ranges. The data rates offered by a WLAN or a WPAN networks are significantly higher than those offered by UMTS networks. In fact, the data rate in UMTS depends on number of factors such as the type of mobility and the range of coverage. Theoretically, the best-case scenario allows UMTS to offer up to 2 Mbps. On the other hand, WLAN data rates offered by 802.11 range high from 11 to 54 Mbps.

### 2.6.3.3 Handoff decision

The handoff decision in heterogeneous networks appears to be more complex to manage. It should be based on application-specified policies and user quality of service requirements as well as quality periodic measurements of the underlying network connectivity.
2.6.3.4 Coverage discovery

WLAN and WPAN networks offer limited coverage and higher data access rates, while WWAN provides geographically wide area coverage but lower access bandwidth (Lungaro and Wallin 2003).

2.6.3.5 Security

Security mechanisms should guarantee that only corresponding parties have knowledge about the key and mobile station identity. Also, the user should maintain the same level of security when roaming across different access networks.

2.7 4G REQUIREMENTS AND CHARACTERISTICS

The wireless 4G networks will be a heterogeneous network consisting of different access networks, which may overlap with one another (i.e., wireless overlay networks). In this environment, a mobile will typically be equipped with multiple wireless interfaces (or a multi-mode interface). It will enhance and extend anytime, anywhere, mobility and accessibility, IP mobility, privacy and security, diversity of services while keeping low cost (Kim et al 2003).

2.7.1 QoS Requirements

In spite of different approaches, resulting from different visions of the future network currently under investigation, the main objectives of wireless 4G networks can be stated as being ubiquity, multi-service platform, and low cost per bit. Wireless 4G-service quality will be the collective effect of the performance of all system elements in combination with the user
expectations. The service provider’s perspective is characterized by the user service requirements, the user perception of QoS, the offered QoS, and the QoS actually delivered. Thus QoS modeling and QoS signaling will be crucial factors for wireless 4G networks. The main requirements for quality of service are the following:

2.7.1.1 **Seamless access**

Seamless access in wireless 4G networks will mean transparent connectivity to the end user across a wide range of access technologies and access networks.

2.7.1.2 **Low handoff delay and loss rate**

Handoff introduces packet loss and delay, which can severely affect data communications. Handoff mechanisms must therefore be managed to minimize these aspects and maintain a good network performance (no disruption to user traffic):

- **Low handoff**: The time required to effect the handoff should be appropriate for the rate of mobility as well as the nature of data transferred.

- **Minimal additional signaling**: Reduction of control signals has been the fundamental design consideration. In fact, handoff signaling should have a minimal impact on network and quality.

- **Low packet loss rate**: The handoff algorithm should minimize the packet loss rate to zero or near zero.
Near-zero handoff blocking probability, and it should have:

- Near-zero call blocking probability.

### 2.7.1.3 Multi-service network

A multi-service network will be an essential property of the wireless 4G, not only because it is the main reason for user transition, it will also give telecommunication operators access to new levels of traffic. Voice will lose its weight in the overall user bill with the raise of more and more data services. The wireless 4G network will offer unlimited mobility and support high data rate, services with variable bandwidths, symmetrical and asymmetrical data transfer (e.g., voice, video, fax, Internet services). This broad array of services will be provided by supporting load balancing, connection priorities and guaranteed quality of service classes.

### 2.7.2 Characteristics

Wireless 4G networks utilize multiple radio access technologies including cellular networks (such as GSM, GPRS, and UMTS), satellite-based networks, and Wireless LANs (e.g., Bluetooth and Ad-hoc networks). They are seamlessly integrated to form a heterogeneous wireless network. The integration is typically based on mobile IP (Yokota et al 2002) and cellular IP. While mobile IP is used for providing macro-mobility management (mobility between different cells that fall under the administration of distinct organizations), cellular IP is used for micro-mobility management. With their complementary characteristics, especially in terms of data rate and radio coverage, it is foreseen that the heterogeneous 4G networks will offer an overlapping coverage to mobile users and enhance cellular network throughput. The defining features of wireless 4G networks are listed below:
2.7.2.1 **Diversified services and ease of use**

The 4G wireless networks will enhance the system performance and functionality to introduce a variety of services that include not only ordinary telephone services, but also services that transfer high-resolution multimedia traffic.

2.7.2.2 **High speed**

The 4G systems will offer a peak speed of more than 100 Mbps in stationary mode with an average of 20 Mbps when traveling.

2.7.2.3 **Fast and seamless handoff across multiple networks**

Wireless 4G networks will support global roaming across multiple wireless and mobile networks. A seamless handoff is defined as a handoff scheme that maintains the connectivity of all applications on the mobile device when the handoff occurs. Seamless handoffs aim to provide continuous end-to-end data service in the face of any link outages or handoff events. Network layer hierarchical mobility management based on mobile IPv6 and cellular IP can be used for fast and seamless handoff to terminals.

2.7.2.4 **Heterogeneous networks**

The mobile IPv6 presents a great contribution to the adaptability of heterogeneous networks.
2.7.2.5  **Multimedia support**

The underlying network for wireless 4G network must be able to support fast speed and large volume data transmission at a lower cost than today.

2.7.2.6  **Anytime, anywhere availability**

This means that mobile networks must be available to the user, any time, anywhere. To accomplish this objective, services and technologies must be standardized in a worldwide scale. Furthermore, the services to be implemented should be available not only to users as is the rule in previous systems, but also to everything that needs to communicate.

2.7.2.7  **Dynamic network resource allocation**

Network resource can be dynamically allocated to cope with varying traffic load, channel condition, and service environment. Traffic conditions will be dynamically monitored and controlled via techniques such as distributed and decentralized control of network functionalities.
2.7.2.8 Low cost per bit

It is an essential feature in a scenario where high volumes of data are being transmitted over the mobile network. In order to make feature wireless 4G services reasonably priced, there is a need to find more efficient methods to allocate and reserve resource for handoff, schedule packets, secure applications and networks.

2.8 RESEARCH OPPORTUNITIES AND THESIS OBJECTIVE

Considerable basic research has already been carried out in the fields related to this thesis, (for example Gross et al 2001, Joer Ott et al 2005). Some of the solutions are even standardized and have great commercial acceptance, SIP (RFC3261). The 3GPP-defined IMS (TS23228) framework integrates some of these solutions under a coherent “umbrella” and achieves a business model that places the network operator in the core of the business value chain.

However, the IMS is designed for 3G networks. Most of 3G solutions can be exported to 4G. Section 2.7 shows the required adaptations. The scientific community is addressing this, while dedicating strong efforts to the design of 4G networks (for example, Antonio Cuevas et al 2005). The session setup topic addressed in this thesis falls within this research area. This thesis presented the current research efforts and the solutions they provide and identified their shortcomings. Those shortcomings forms basis of the research opportunities to this thesis and justify the objectives presented in the introduction.

Section 2.1 explains that IMS assumes and is designed to support only one business model and one type of application: the ones using SIP and
IMS proxies. However, several business models and scenarios will co-exist in 4G networks and hence the goal of this thesis is to propose suitable mechanisms to support them in an integrated framework. SIP (RFC3263) and SDP (RFC2327) are good session setup techniques; RSVP (RFC2205) on its side is used to setup and request QoS resources at the network level. However, due to the shift in business paradigms, they have to be enriched to support new aspects. This also will be addressed in this thesis. Also, many 4G solutions are available (for example, Andrea Calvagna et al 2000, Shou-Chih Lo et al 2004) but, as per section 2.7, the challenge is to integrate them. This thesis proposes the PBQMS as a solution for session setup and QoS maintenance.

The main contribution of this thesis is to propose a new policy based QoS and session setup technique using a policy-based architecture called PBQMS for the 4G networks using IMS. The proposed PBQMS architecture extends the original policy based IMS into a multi-domain QoS policy architecture as per user’s preferences. The contributions of this thesis are:

- Propose a new policy based QoS architecture
- Propose a policy based session setup technique using a modified SIP and XCAP protocols.
- Modify the existing policy based QoS for roaming between different transmission technologies.
- Use an analytical model to compare the performance of policy based session establishment scenarios in a mobile environment.

In the following chapters, the solutions proposed to address these aspects are discussed.