Engineering Wireless Multimedia Networks for QoS Differentiation

Dr. Mooi Choo Chuah (chuah@lucent.com)
Dr. On-Ching Yue (oyue@lucent.com)
Bell Laboratories, Lucent Technologies
Outline

• Introduction to QoS Engineering
• Overview of UMTS Architecture
• QoS Engineering of Core Network
• QoS Engineering of Radio Access Network
• Future Design Challenges
Customer and Supplier Perspectives

User’s Needs and Expectations
Deliver my contents with the following end-to-end QoS requirements:
- Speed (throughput, bit rate)
- Accuracy (error rate, loss rate)
- Latency (average, 95%, jitter)
- Availability (blocking, setup time)
- Reliability (call dropping, outage)

With the $ I pay you, I expect
- **QoS Differentiation**
- Fairness

Operator’s Views and Objectives
Meet the different QoS requirements of all admitted users with my network resources:
- RF spectrum and base stations
- Links and protocol processing
- Switches, routers and multiplexers
- Processors and controller software

Maximize the revenue using the infrastructure equipment
- Resource control and management
Circuit Switching Example
(The good old days)

• User needs a voice connection between CA and NJ.
• Operator has a collection of TDM switches between CA and NJ, connected by SDH/SONET (>153Mb/s) links.
  – Speed - Assign a DS0 (64kb/s) slot in all the connecting links.
  – Accuracy - All links have been designed to meet required BER.
  – Latency - Propagation delay (msec) + processing delay (usec)
  – Availability - Call processor speed, trunk size and connectivity
  – Reliability - HW/SW failure rate and network restoration strategy
• QoS differentiation is by resource reservation.
Techniques for Circuit Switching QoS Engineering

- Traffic characterization – voice arrival patterns, holding time, speech activity
- Admission control - block only if no trunks are available in alternate routes
- Call processing - proper sizing of processor, link and database for signalling
- Performance and reliability - budgeting of end-to-end requirements to individual network elements and protocol layers
- Traffic engineering - plan traffic routing for load balancing, provision link capacity for specific services: voice, private lines, etc
- Network design - to minimize link/element cost while meeting performance and reliability requirements
Challenges for QoS Engineering of Wireless Multimedia Networks

- User needs to set up a netmeeting (voice, streaming video, and whiteboard) with mobile colleagues.
- Operator has a 3G (UMTS) network with base stations, radio network controllers, specialized routers (GSNs), hosting environment, and a IP/ATM network.
  - **Speed** - Variable link bandwidth due to statistical multiplexing
  - **Accuracy** - Packet losses due to RF environment and buffer overflow
  - **Latency** - Variable queuing delay due to scheduling and RF condition
  - **Availability** - Control plane signaling and processor capacity
  - **Reliability** - Mobility management and network restoration strategy
QoS Engineering for Packet-based Wireless Network

- Traffic characterization – voice, web-browsing, WAP, streaming video
- Admission control and network resource allocation
- Server/link scheduling and traffic policing/shaping
- QoS mapping between networks and between layers
- Service level agreements (SLA) and policy management
- Radio channel selection and bandwidth allocation
- RF power control and rate control among users
- RF resource set-up and tear-down strategies
- Traffic engineering and network design
Outline

- Introduction to QoS Engineering
- **Overview of UMTS Architecture**
- QoS Engineering of Core Network
- QoS Engineering of Radio Access Network
- Future Design Challenges
Protocol Stacks for User Plane for UMTS R99

- **Application**
  - E.g., IP, PPP, OSP

- **L1**
  - L1
  - ATM

- **MAC**
  - MAC
  - AAL5

- **RLC**
  - RLC
  - UDP/IP

- **PDCP**
  - PDCP
  - GTP-U

- **Relay**
  - UDP/IP
  - L2
  - L1

MS - UTRAN - 3G-SGSN - 3G-GGSN
Layered Bearer Services Architecture

End-to-end Service

Classification of User traffic

Local Bearer Service

UMTS Bearer Service

Radio Access Bearer Service

Radio Bearer Service

Radio resource management

UTRA FDD/TDD Service

Encapsulated into GTP packets

Fragment into ATM cells

ATM Traffic Management

Backbone Bearer IP Service

Note: Can be over Ethernet, ATM, SONET etc.

Proper PHB

Encapsulated into GTP Packets

CN DS bearer Service

Iu ATM Bearer Service

ATM Service

Iu DS bearer Service

Uu ATM Bearer Service

Backbone Bearer Service

External Bearer Service

CGI

Iu DS bearer Service

Iu ATM Bearer Service

External Bearer Service

CGI

Iu DS bearer Service

Iu ATM Bearer Service

Backbone Bearer Service

External Bearer Service

CGI

Note: Can be over Ethernet, ATM, SONET etc.
Example of packet service session

1. RRC Connection Established (RRC)
2. Activate PDP Context Request (SM)
3. Establish Radio Bearers (RRC)
4. RAB Assignment Request (RANAP)
5. RAB Assignment Response (RANAP)
6. Create PDP Context Request (GTP-C)
7. Create PDP Context Response (GTP-C)
8. Activate PDP Context Accept (SM)

Source: 3GPP TS 23.060 - GPRS Service Description

GGSN = Gateway GPRS Support Node
SGSN = Serving GPRS Support Node
GTP-C = GPRS Tunneling Protocol - Control plane
SRNC = Serving Radio Network Controller
RAB = Radio Access Bearers
SM = Session Management
RANAP = Radio Access Network Application Part
UE = User Equipment
RRC = Radio Resource Control
PDP = Packet Data Protocol
# PDP Context QoS Profile Information Element

<table>
<thead>
<tr>
<th>Octet</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Quality of service IEI</td>
</tr>
<tr>
<td>2</td>
<td>Length of quality of service IE</td>
</tr>
<tr>
<td>3</td>
<td>Delay class</td>
</tr>
<tr>
<td>4</td>
<td>Reliability class</td>
</tr>
<tr>
<td>5</td>
<td>Peak throughput</td>
</tr>
<tr>
<td>6</td>
<td>Precedence class</td>
</tr>
<tr>
<td>7</td>
<td>Peak throughput</td>
</tr>
<tr>
<td>8</td>
<td>Mean throughput</td>
</tr>
<tr>
<td>9</td>
<td>Traffic Class</td>
</tr>
<tr>
<td>10</td>
<td>Delivery order</td>
</tr>
<tr>
<td>11</td>
<td>Delivery of erroneous SDU</td>
</tr>
<tr>
<td>12</td>
<td>Maximum SDU size</td>
</tr>
<tr>
<td>13</td>
<td>Maximum bit rate for uplink</td>
</tr>
<tr>
<td>14</td>
<td>Maximum bit rate for downlink</td>
</tr>
<tr>
<td>15</td>
<td>Residual BER</td>
</tr>
<tr>
<td>16</td>
<td>SDU error ratio</td>
</tr>
<tr>
<td>17</td>
<td>Transfer delay</td>
</tr>
<tr>
<td>18</td>
<td>Traffic Handling priority</td>
</tr>
<tr>
<td>19</td>
<td>Guaranteed bit rate for uplink</td>
</tr>
<tr>
<td>20</td>
<td>Guaranteed bit rate for downlink</td>
</tr>
</tbody>
</table>
Control Plane QoS Functions

MT
- Service Translator
- Admin./Cap. Control
- UMTS BS Manager
- Local BS Manager
- Radio BS Manager
- UTRA Phy. BS Manager

UTRAN
- GMM/SM
- RAB Manager
- UMTS GTP DS BS Manager
- Iu ATM BS Manager
- UTRA Phy. BS Manager
- Radio BS Manager
- ATM NS Manager

SGSN
- UMTS BS Manager
- UMTS GTP DS BS Manager
- Iu ATM BS Manager
- UMTS GTP DS BS Manager
- ATM NS Manager
- BB IP NS Manager

GGSN
- UMTS BS Manager
- UMTS GTP DS BS Manager
- Ext. BS Manager
- BB IP NS Manager

UTRAN Functions
Core Network Functions
Control primitives
Signaling primitives
Functions not required by Rel1v2
End User Applications mapping to UMTS QoS Classes

- **Conversational**: two-way transport between human users
  - real-time voice and real-time video
  - interactive games
- **Streaming**: one-way transport to human destination
  - real-time audio and real-time video.
- **Interactive**: two-way transport between user and server
  - web browsing
  - email
  - database retrieval
- **Background**: one-way transport to machine destination
  - fax and backdrop delivery of email
  - file transfer
  - database download
Application’s Qualitative QoS Requirements

<table>
<thead>
<tr>
<th>Error tolerant</th>
<th>Conversational voice and video</th>
<th>Voice messaging</th>
<th>Streaming audio and video</th>
<th>Fax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error intolerant</td>
<td>Telnet, interactive games</td>
<td>E-commerce, WWW browsing,</td>
<td>FTP, still image, paging</td>
<td>E-mail arrival notification</td>
</tr>
<tr>
<td>Conversational (delay &lt;&lt;1 sec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactive (delay approx 1 sec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streaming (delay &lt;10 sec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background (delay &gt;10 sec)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# UMTS Delay Objectives

<table>
<thead>
<tr>
<th>Bearer</th>
<th>Conversational</th>
<th>Streaming</th>
<th>Interactive</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMTS Bearer (RAB + CN Bearer)</td>
<td>150ms (from 23.107)</td>
<td>250ms (from 23.107)</td>
<td>400 ms (back of envelope calculation based on 10x1500 bytes packets per webpage)</td>
<td>1000ms (guess)</td>
</tr>
<tr>
<td>Radio Access Bearer (RAB) (UTRAN + Iu)</td>
<td>120ms (from 23.107)</td>
<td>200ms (80% of UMTS bearer)</td>
<td>320ms (80% of UMTS bearer)</td>
<td>800ms (80% of UMTS bearer)</td>
</tr>
<tr>
<td>CN Bearer (from MSC or SGSN to Gateway)</td>
<td>30ms</td>
<td>50ms</td>
<td>80 ms</td>
<td>200ms</td>
</tr>
<tr>
<td>Iu Bearer</td>
<td>24 ms (20% of RAB)</td>
<td>40ms (20% of RAB)</td>
<td>64 ms (20% of RAB)</td>
<td>160ms (20% of RAB)</td>
</tr>
<tr>
<td>Radio Bearer</td>
<td>96 ms (80% of RAB)</td>
<td>160ms (80% of RAB)</td>
<td>256 ms (80% of RAB)</td>
<td>640ms (80% of RAB)</td>
</tr>
</tbody>
</table>
Web-Browsing Traffic Modeling

Packet-call = individual web pages
Packets = objects contained in a web page (e.g. HTML code, .gif, .wav, etc.)
Outline

- Introduction to QoS Engineering
- Overview of UMTS Architecture
- QoS Engineering of Core Network
- QoS Engineering of Radio Access Network
- Future Design Challenges
SGSN is the Brain in the Core Network

MT
- Service Translator
- Adm./Cap. Control
- UMTS BS Manager
- Local BS Manager
- Radio BS Manager
- UTRA Phy. BS Manager

UTRAN
- Adm./Cap. Control
- RAB Manager
- UMTS GTP DS BS Manager
- Iu ATM BS Manager
- UTRA Phy. BS Manager
- Radio BS Manager
- ATM NS Manager

SGSN
- Adm./Cap. Control
- Subscriber Control
- UMTS BS Manager
- UMTS GTP DS BS Manager
- Iu ATM BS Manager
- BB IP NS Manager
- ATM NS Manager
- UTRA Phy. BS Manager

GGSN
- Adm./Cap. Control
- Service Translator
- UMTS BS Manager
- Ext. BS Manager
- BB IP NS Manager

UTRAN Functions
Core Network Functions
Control primitives
Signaling primitives
Functions not required by Rel1v2

Sept. 11, 2001 - CUHK 20
**SGSN is the Brain in the Core Network**

- It receives the connection setup request with the QoS requirement.
- It instructs the RNC to set up the Radio Bearer.
- It sets up a GTP connection between GGSN and the RNC.
- QoS engineering of the GTP tunnels
  - Throughput
  - Latency
  - Loss
  - Availability
GPRS GTP Tunnels

- GPRS defines GTP in 3GPP TS 29.060
- GPRS transports TCP/IP or UDP/IP packets within UDP/IP packets through the GPRS network
- GPRS allows layer 2 PPP frames to be transported transparently by encapsulating PPP frames within UDP/IP packet.
- A GTP tunnel in the user plane is defined for each PDP context in the GSNs and/or each RAB in the RNC.
- A GTP tunnel in the control plane is defined for all PDP contexts with the same PDP address and APN.
GPRS Tunnel Protocol (GTP)

- UDP/IP based tunnel
- GTP Header has Tunnel Endpoint ID

![Diagram of GTP structure]

IP Destination Address for the Tunnel
UDP header
GTP Header
The tunneled IP packet

IP Source Address for the Tunnel
GTP QoS Implementation Options

- IETF DiffServ (DS) will be used for QoS differentiation in GTP tunnels.
  - DS specifies Per Hop Behaviors (PHB) by DS Code Points
  - Expedited Forwarding (EF): highest priority
  - Assured Forwarding (AF$_{ij}$): 4 lower priority classes ($i=1,2,3,4$) with different drop precedence ($j=1,2,3$)
  - Handles aggregate traffic rather than individual flows

- GTP/IP/*
  - L2 services which do not support QoS: Routers need to implement DS per class queueing (buffer management) and scheduling.
  - MPLS: map DSCP to LSP QoS classes
  - ATM: map DSCP to ATM QoS classes (CBR, rt-VBR, nrt-VBR, ABR, UBR)
  - Ethernet: map DSCP to 802.1D priority services
A Generic Resource Node for GTP/IP/L2

Arrivals

Buffer/Memory

Queue structure

Scheduling

Server/Link Complex

Departures

Blocked Arrivals
Examples of Buffer Management Schemes

• Random Early Discard (RED) - drops packets with a certain probability as a function of the average queue length.
• Multi-class RED - has single or multiple thresholds.
• RED with In/Out and Coupled Virtual Queues (RIO-C) - marks packets belonging to a stream as IN or OUT according to a specified policy and has thresholds for virtual queues.
Scheduler Design

• Examples
  – Weighted Fair Queueing: GPS, SCFQ, WF2Q
  – Earliest Due Date
  – Class-based Queueing

• Design Considerations
  – Worst case fairness
  – Fair Excess bandwidth sharing
  – Scalable – in terms of latency with increasing number of sessions
  – Implementation complexity
GTP/IP/MPLS Option

• Multi-Protocol Label Switching (MPLS) allows the operator to route packets along predetermined Label Switch Paths (LSP) instead of those selected by IGP, etc.
  – Allows Traffic Engineering: load balancing of links and network elements, and fast path restoration.
  – Uses RSVP as the signaling protocol to reserve resources and establish path states in the network nodes.

• QoS related attributes in LSP
  – FLOWSPEC object specifies service type (controlled-load or guaranteed service), traffic description (peak rate, token bucket for rate control) and service request (bandwidth, delay, loss).
  – SESSION_ATTRIBUTE object controls the LSP priority, preemption and fast-reroute features.
Outline

• Introduction to QoS Engineering
• Overview of UMTS Architecture
• QoS Engineering of Core Network
• QoS Engineering of Radio Access Network
• Future Design Challenges
Example of packet service session

1. RRC Connection Established (RRC)
2. Activate PDP Context Request (SM)
3. Establish Radio Bearers (RRC)
4. RAB Assignment Request (RANAP)
5. RAB Assignment Response (RANAP)
6.
7.
8. Activate PDP Context Accept (SM)

Source: 3GPP TS 23.060 - GPRS Service Description

GGSN = Gateway GPRS Support Node
GTP-C = GPRS Tunneling Protocol - Control plane
RAB = Radio Access Bearers
RANAP = Radio Access Network Application Part
RRC = Radio Resource Control
PDP = Packet Data Protocol
SGSN = Serving GPRS Support Node
SRNC = Serving Radio Network Controller
SM = Session Management
UE = User Equipment
Radio Interface Protocol Architecture

RRC: Radio Resource Control
RLC: Radio Link Control
MAC: Medium Access Control
PHY: Physical Layer
PDCP: Packet Data Conversion Protocol
BMC: Broadcast- & Multicast Control

C-plane signalling

U-plane information

Physical Channels

Logical Channels

Transport Channels

L1

L2/MAC

L2/RLC

L2/BMC

L2/PDCP

PHY: Physical Layer

RRC: Radio Resource Control
RLC: Radio Link Control
MAC: Medium Access Control
BMC: Broadcast- & Multicast Control
RLC Functions & Services

- The RLC Layer consists of the following protocol Modes:
  - There is one transmitting and one receiving entity for the Tr and the UM modes and one combined transmitting and receiving entity for the AM mode service.
  - **Transparent data transfer Service (Tr):**
    - Segmentation and re-assembly, Transfer of user data.
  - **Unacknowledged data transfer Service (UM):**
    - Segmentation and reassembly, Concatenation, Padding, Transfer of user data, Ciphering, Sequence number check.
  - **Acknowledged data transfer Service (AM):**
    - Segmentation and reassembly, Concatenation, Padding, Transfer of user data, Error correction, In-sequence delivery of higher layer PDUs, Duplicate detection, Flow Control, Protocol error detection and recovery, Ciphering.
  - QoS setting
  - Notification of unrecoverable errors.
Model of Acknowledged Mode (AM) RLC (TS 25.322)
Example of 64-Kbps Transmission in the Downlink

DTCH

- Information data
- CRC detection
- Turbo code R=1/3
- Rate matching
- 1st interleaving
- Radio Frame Segmentation
- 2nd interleaving
- Slot segmentation
- 120kps DPCH (including TFCI bits)

DCCH

- Information data
- CRC detection
- Tail bit discard
- Viterbi decoding R=1/3
- Rate matching
- 1st interleaving
- Radio Frame Segmentation
- 2nd interleaving
- Slot segmentation
- 120kps DPCH (including TFCI bits)
Comparison of analytical and simulation results for mixed voice and data traffic

Comparison of analysis and simulation for mixed voice and data traffic
Performance of ARQ Schemes

Stop-and-Wait:
- Maximum Throughput = \( (1 - p)/(t + T) \)
  where \( p \) = packet loss probability
  \( t \) = packet transmit time
  \( T \) = feedback delay
- Delay = \( (t + T)/(1 - p) \)

Selective Retransmit:
- Maximum Throughput = \( (1 - p)/t \)
  where \( t \) depends on RF bandwidth
- Delay = \( t + pT/(1-p) \)
  where \( T \) depends on polling strategy
- Loss occurs due to timeouts.
A Generic Resource Node

Arrivals → Buffer/Memory

Queue structure

→ Server/Link Complex

Scheduling

→ Departures

Blocked Arrivals

New Challenge:
Link capacity is packet dependent.
Challenges in scheduler design

- Fairness: users belonging to same priority class should have similar service (delay and throughput)
- QoS: Requirements on maximum delay and minimum throughput
- Assume a slotted system: scheduler decides which user(s) to serve in each time slot.
- Time slot size: efficiency vs fill factor
- Signalling cost: fast tracking loop requires uplink bandwidth
Scheduling Options

- **Round-Robin**
  - select user packets in a round-robin fashion
  - number of slots allocated to a user is inversely proportion to the user’s rate.
  - equal number of bytes are transferred assuming equal size packets
  - provides equal delay (fairness) across all the users

- **Equal power**
  - users are assigned equal number of slots irrespective of the data rate
  - results in unequal delay
More Scheduling Options

- **Max C/I**
  - Attempts to take advantage of the variations in the wireless channel across users and attempts to serve the user with the best channel conditions (with the largest $R[n]$ in slot n).
  - Provides the highest multi-user diversity gains.
  - Can be unfair!

- **Proportional fair**
  - Maximizes the product of the throughputs delivered to all users.
  - Serve the user with the largest $R[n]/R_{av}$. Where $R[n]$ is the rate in slot n and $R_{av}$ is the average rate for the user.
Performance Metrics

$OTA = \frac{\text{Total good bits}}{(\text{Total slots with transmissions}) \times T_{\text{slot}}}$

Service throughput $= \frac{\text{Total good bits}}{\text{Total slots} \times T_{\text{slot}}}$

Packet Call Throughput $= \frac{1}{\text{Num. pkt calls}} \sum_{j=1}^{\text{Num. pkt calls}} \frac{\text{Bits in pkt call } j}{\text{Duration of pkt call } j}$

Utilization $= \frac{\text{Total slots with transmissions}}{\text{Total slots}}$
System Throughput

![System Throughput Graph]

- **Throughput [Kb/s]**
- **Number of UEs**
- **Throughput**
  - Packet Call
  - OTA
  - Service

<table>
<thead>
<tr>
<th>Number of UEs</th>
<th>Throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>37 UEs</td>
<td></td>
</tr>
<tr>
<td>50 UEs</td>
<td></td>
</tr>
<tr>
<td>75 UEs</td>
<td></td>
</tr>
<tr>
<td>150 UEs</td>
<td></td>
</tr>
</tbody>
</table>
Outline

• Introduction to QoS Engineering
• Overview of UMTS Architecture
• QoS Engineering of Core Network
• QoS Engineering of Radio Access Network
• Future Design Challenges
Release 5 - IP Multimedia Core Network Subsystem

IP Multimedia Subsystem

ANSI/MAP CS Domain

GPRS/MIP Domain

GERAN

UTRAN

CDMA2000

Sept. 11, 2001 - CUHK