3G/4G Mobile Communications Systems

Dr. Stefan Brück
Qualcomm Corporate R&D Center Germany
Chapter VIII: MAC Scheduling
MAC Scheduling

- Principle of a Shared Channel
- Classical Scheduling Approaches
  - Max C/I, Round Robin, Proportional Fair
- Scheduling Utility Functions
- QoS aware Scheduling
  - VoIP, Streaming
- Commonalities and Differences in HSPA and LTE
Motivation of a Shared Channel

- Dedicated channels and fast power control simplify the maintenance of quality of service (QoS) like
  - Voice quality for circuit-switched voice applications
  - Delay requirements for video telephony
  - Data rate requirements for streaming applications

- Conversely, dedicated channels waste resources
  - With fast power control, the Tx power becomes the larger the deeper the channel fade is
  - Fast power control contradicts the information-theoretic water-filling principle
  - In CDMA, the DL channelization codes for high data rates are limited (e.g. SF8 for 384 kbps)

*Capacity can be increased (at least) for best effort traffic without QoS constraints if channels/resources are shared between users instead of dedicated to a user*
Principle of a Shared Channel

- Data transmission takes place to those users with good channel conditions
  - The figure shows an DL example
- Also the data rate can be adapted to the channel conditions
- Such a fast adaptation requires a set of new algorithms/channels
  - Scheduling for user selection
  - Modulation and coding scheme adaptation
  - UL feedback of instantaneous DL channel information

In order to exploit fast fading, the round trip delay (feedback + scheduling) is the key
Dynamic Scheduling

- “Statistical multiplexing” of data packets from different data flows over one shared medium
- Optimised usage of radio resources
- Exploitation of the short-term variations on the radio channels
- Provides certain degree of QoS
UMTS Release 4: Downlink Shared Channel

- DSCH was introduced in 2001 to overcome channelization code limitations
  - It supported data rate adaptation with a rate of up to 10ms
  - Tx power was coupled with Tx power of associated dedicated channel
- MAC layer functionality remained in RNC
  - MAC-c/sh supports scheduling
- Main drawbacks
  - Scheduling delay of 10ms – 100ms
  - Power control in soft handover of associated dedicated channel
- At the end, the DSCH was never deployed in the field
  - It was removed again from Release 5 specification onwards
Fast Scheduling

- Channels are uncorrelated → Multi-user diversity
- Fading is good in multi user environment
- Assign the resources to the best user(s) in time to maximise throughput
  - Short round trip delay is required
- Also the data rate can be adapted to the radio channel conditions
  - Link adaptation
- With HSDPA Scheduling function is moved from RNC to Node B
Classical Scheduling Disciplines

- **Round Robin**: Allocate the users consecutively
  - Advantage:
    - Offers fair time allocation
    - One of the simplest solutions
  - Disadvantage:
    - Low cell and user throughput

- **Max C/I scheduler**: Prefer the users with good channel conditions
  - Advantage:
    - Highest system throughput and easy to implement
  - Disadvantage:
    - Starvation to users with low C/I

- **Proportional Fairness**: Equalise the *channel rate / throughput* ratio
  - Advantage:
    - Higher throughput than Round Robin
  - Disadvantage:
    - Does not use QoS information
Usually scheduling aims to maximize an utility function $U(R_1, \ldots, R_n)$ of the form

$$U(R_1, \ldots, R_n) = \begin{cases} 
\sum_{i=1}^{n} \left( \frac{R_i}{1 - \alpha} \right)^{1-\alpha} & \text{for } \alpha \neq 1 \\
\sum_{i=1}^{n} \log(R_i) & \text{for } \alpha = 1 
\end{cases}$$

over the throughput $R_i$

- $R_i$ is the long term average allocated throughput
- $R_i$ is a function of the instantaneous throughput $r_i(t)$
- $\alpha$ is a constant form factor
This function is tried to be maximized by allocating rates into the direction of the steepest ascent:

$$\sum_{i=1}^{n} \frac{\partial U}{\partial R_i} \cdot r_i(t) = \sum_{i=1}^{n} \frac{r_i(t)}{R_i^\alpha} \rightarrow \max_{r_1(t), \ldots, r_n(t)}$$

- The allocated rates $r_i(t)$ can be written as

$$r_i(t) = \sum_{j=1}^{\# PRB} \delta_{ij} \cdot \rho_{ij}$$

- $\delta_{ij} \in \{0,1\}$ indicates whether PRB $j$ is allocated to user $i$

- A PRB is a physical resource block in LTE of 180 kHz
  - In HSDPA a resource block can be regarded as the entire 5 MHz bandwidth

- $\rho_{ij}$ is the spectral efficiency of PRB $j$ for user $i$ (in an appropriate unit)
  - This assumes that a CQI $\equiv$ spectral efficiency per PRB is available
Then the maximization task becomes

\[
\max_{\delta_{ij}} \sum_{i=1}^{n} \frac{1}{R_i^\alpha} \sum_{j=1}^{\#PRB} \delta_{ij} \cdot \rho_{ij}
\]

This maximization can be re-written as

\[
\max_{\delta_{ij}} \sum_{i=1}^{n} \frac{f(\rho_{i_{PRB1}}, \ldots, \rho_{i_{PRBn}})}{R_i^\alpha} \sum_{j=1}^{\#PRB} \frac{\delta_{ij} \cdot \rho_{ij}}{f(\rho_{i_{PRB1}}, \ldots, \rho_{i_{PRBn}})}
\]

Now the priority \( P_i \) of user \( i \) can be defined as \( f(\rho_{i_{PRB1}}, \ldots, \rho_{i_{PRBn}}) / R_i^\alpha \)

The term \( \rho_{ij} / f(\rho_{i_{PRB1}}, \ldots, \rho_{i_{PRBn}}) \) can be named the resource weight \( C_{ij} \) for user \( i \) and PRB \( j \)

The allocation of resources can now be done in the following steps:

1. Rank the users according to the user priority \( P_i \)
2. Determine the \( n' \leq n \) users with highest priority \( P_i \)
   - \( n' < n \) can be required to limit processing time
3. Allocate PRB \( j \) to ranked user \( i \leq n' \) such that \( P_i \cdot C_{ij} \) is maximal
Comments to the previous Approach I/III

- The previous description holds both for LTE and HSDPA.
  - HSDPA scheduling can be seen as a special case of LTE with $\#\text{PRB} = 1$

- From the previous formula it is seen that for each PRB $j$ it is required to find the best user $i$ with highest $P_i \cdot C_{ij}$

- Therefore the number of required rankings is $\#\text{PRB} \cdot \#\text{eligible users}$
  - An user is called eligible if it can be scheduled, i.e. if it has data in its buffer and HARQ process not waiting for an ACK/NACK

- Scheduling approaches that exploit the CQI for the PRBs individually are called frequency-selective
  - LTE offers the possibility of frequency selective scheduling
  - HSDPA in 5 MHz does not allow frequency-selective scheduling since the 5 MHz bandwidth cannot be split in smaller chunks
    - See DC-HSDPA in chapter “Current and Future Trends in 3GPP – HSPA+ and LTE-A”

- The scheduling complexity in LTE does not only increase with the number of users but also with the available frequency bandwidth
The function $f(\rho_{i\text{PRB}1}, \ldots, \rho_{i\text{PRB}n})$ is arbitrary since it does not impact the maximization of the utility function.

If it is required to build a user list of length $n' < n$, user ranking is required and the utility function $U(R_1, \ldots, R_n)$ is not necessarily maximized anymore.

In this case the choice of the function $f(\rho_{i\text{PRB}1}, \ldots, \rho_{i\text{PRB}n})$ can have a big impact on the performance metrics (cell and user throughput, delay, etc).

For best effort traffic, the function $f(\rho_{i\text{PRB}1}, \ldots, \rho_{i\text{PRB}n})$ can be defined as

$$f \left( \rho_{i\text{PRB}1}, \ldots, \rho_{i\text{PRB}n} \right) = \frac{1}{n} \sum_{j=1}^{n} \rho_{ij}$$

i.e. the linear average over the spectral efficiencies.
The form factor $\alpha$ allows tuning the scheduler

For $\alpha = 0$ the scheduling rule is the max C/I scheduler. This rule maximizes the cell throughput at the expense of user fairness

For $\alpha = 1$ the scheduling rule is the proportional fair scheduler

If $\alpha$ is increased beyond one the fairness in terms of equal user throughput is increased further
Scheduler Inputs

QoS & Subscriber Profile
User 1: Best effort, silver class
User 2: High priority, platinum class

History
How long had the user been waiting?

Traffic Model
Morning  Afternoon
Evening  Off peak

Feedback from UL (CQI, ACK/NACK)

UE capability

Radio resources
Power, OVSF codes, PRBS

Buffer Status

Scheduler

Scheduled Users & Packet Formation Strategy
Scheduler Outputs

- **Goals**
  - Chosen utility function is maximized
  - QoS/GoS constraints are satisfied
  - Maintain fairness across UEs and traffic streams
QoS Scheduling

- The classical scheduling approaches max C/I, round robin, and proportional fair do not take Quality of Service (QoS) requirements into account.
- QoS requirements are typically given as guaranteed bit rate (GBR) or delay requirements.
- A common approach is to modify the utility function with a priority weight $p_{wi}$

$$
\sum_{i=1}^{n} p_{wi} \frac{r_i(t)}{R_i^\alpha} \rightarrow \max_{r_1(t), \ldots, r_n(t)}
$$

- Examples:
  - GBR constraints
    - $p_{wi}$ is increased when $R < R_{\text{min}}$
    - $p_{wi}$ is decreased when $R > R_{\text{max}}$
  - Delay constraints
    - $p_{wi}$ is increased if packet waiting time approaches delay requirement

- (Modified) First In First Out (FIFO) approaches can also be applied for very delay sensitive services.
Comparison of Schedulers for HSDPA

- Simple Round Robin doesn’t care about actual channel rate
- Proportional Fair offers higher cell throughput
- QoS aware algorithm offers significantly higher user perceived throughput than PF with similar cell throughput
  - QoS scheduler: $R_{\text{min}} = 40$ kbps, $R_{\text{max}} = 500$ kbps
Scheduling for Mixed Services – I/II


- VoIP service is scheduled by modified FIFO type scheduler
  - Channel quality is not taken into account
  - Introduction of a wait time $T_{\text{wait}}$:
    - Before expiration VoIP is ranked behind the background service
    - After timer $T_{\text{wait}}$ has expired VoIP is ranked on top
  - Services with the same state w.r.t. $T_{\text{wait}}$ are scheduled with FIFO metric

- VoIP service can also be scheduled by a channel aware metric with additional rule to fulfill a minimum rate of service

- Background traffic is scheduled by proportional fair metric

- Delay and rate requirements
  - For a voice codec AMR 12.2 typically 320 bits need to be transmitted with IPv4
  - Total delay budget including HSPA delay, transport delay, core delay: 280ms
    - HSPA MAC delay: 90ms
FIFO type schedulers provide the required QoS of VoIP traffic.

By adjusting the wait time $T_{\text{wait}}$, a trade-off between VoIP and background traffic can be achieved:
- Higher $T_{\text{wait}}$ provides better throughput for the background service.
- Lower $T_{\text{wait}}$ improves frame loss performance for the VoIP service.
- With $T_{\text{wait}} > 0$ there is an impact of increasing background traffic onto VoIP quality.

Min rate scheduler is unable to balance QoS requirements between VoIP and background traffic.

Results taken from paper by J. Mueckenheim et al.
Scheduling and Resource Allocation

- Basic unit of allocation is called a Resource Block (RB)
  - 12 subcarriers in frequency (= 180 kHz)
  - 1 sub-frame in time (= 1 ms, = 14 OFDM symbols)
  - Multiple resource blocks can be allocated to a user in a given subframe

- The total number of RBs available depends on the operating bandwidth

<table>
<thead>
<tr>
<th>Bandwidth (MHz)</th>
<th>1.4</th>
<th>3.0</th>
<th>5.0</th>
<th>10.0</th>
<th>15.0</th>
<th>20.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of available resource blocks</td>
<td>6</td>
<td>15</td>
<td>25</td>
<td>50</td>
<td>75</td>
<td>100</td>
</tr>
</tbody>
</table>
LTE Downlink Scheduling & Resource Allocation

- Channel dependent scheduling is supported in both time and frequency domain → enables two dimensional flexibility
  - CQI feedback can provide both wideband and frequency selective feedback
  - PMI and RI feedback allow for MIMO mode selection
  - Scheduler chooses bandwidth allocation, modulation and coding set (MCS), MIMO mode, and power allocation

- HARQ operation is asynchronous and adaptive
- Assigned PRBs need not be contiguous for a given user in the downlink
LTE Uplink Scheduling & Resource Allocation

- Channel dependent scheduling in both time and frequency enabled through the use of the sounding reference signal (SRS)
  - Scheduler selects bandwidth, modulation and coding set (MCS), use of MU-MIMO, and PC parameters

- HARQ operation is synchronous, and is non-adaptive

- PRBs assigned for a particular UE must be contiguous in the uplink (SC-FDMA)
  - To reduce UE complexity, restriction placed on # of PRBs that can be assigned
    - Number of allocated subcarriers must have largest prime factor less than or equal to 5 → can use radix-2,3,5 FFT for DFT-precoding (i.e., cannot assign 7, 11, 13, 17,… PRBs)
Semi-Persistent Resource Allocation for VoIP

- Semi-persistent allocation is introduced to support a large number of VoIP users without running into control channel bottleneck
  - RRC signaling configures time periodicity of persistent allocation (i.e. 20ms period) and a “persistent scheduling C-RNTI” (special identifier)
  - Scheduling grant on PDCCH used to activate a persistent allocation which applies to the first HARQ transmission
    - Scheduling grant assigns MCS and subframe location for persistent allocation
    - Resources implicitly released after inactivity
  - Retransmissions *maybe dynamically scheduled selectively to optimize packing of VoIP users*

![Diagram of persistent allocation](image)
Commonalities and Differences in HSPA and LTE

**Commonalities:**
- HSDPA, LTE PDSCH and LTE PUSCH are shared channels
- For DL channels, channel information is available by UE feedback
  - Exploiting the fast fading is possible
- For UL scheduling decisions, it is more difficult to take fast fading into account
  - LTE UL introduced sounding reference signal to allow frequency-selective uplink measurements in the eNB
- In UL scheduling, both power resources of the terminals as well as buffer status need to be taken into account
  - The terminal must provide this information to the Node B and eNB

**Differences:**
- HSUPA is not a shared channel
  - The transmit power of the E-DCH is coupled to the UL DPCCH
  - HSUPA can be in soft/softer handover
  - Exploiting the fast fading is not possible
- HSUPA is not synchronized
  - Intra-cell interference occurs between users
- HSUPA scheduling is similar to load control for R99 channels
  - In E-DCH the load is controlled by the Node B and not by the RNC
Node B Scheduling Principle

- E-DCH scheduler constraint
  - Keep UL load within the limit

- Scheduler controls:
  - E-DCH load portion of non-serving users from other cells
  - E-DCH resources of each serving user of own cell

- Non-E-DCH Load
  - Includes DCH, HS-DPCCH, non-scheduled E-DCH
  - Controlled by legacy load control in RNC

- Principles:
  - Rate vs. time scheduling
  - Dedicated control for serving users
  - Common control for non-serving users

- Note: Scheduler cannot exploit fast fading!
E-DCH Scheduling Options

- UEs are continuously active
- Data rate is incremental increased/decreased by relative scheduling grants
- No synch between UEs required
- Load variations can be kept low
- For low to medium data rates

- UEs are switched on/off by absolute scheduling grants
- UEs should be in synch
- Load variations might be large
- For (very) high data rates
E-DCH Scheduling

- UE maintains internal serving grant SG
- SG are quantized Maximum E-DPDCH/ DPCCH power ratio (TPR), which are defined by 3GPP
- Reception of absolute grant: SG = AG
  - No transmission: SG = “Zero_Grant”
- Reception of relative grants: increment/ decrement index of SG in the SG table
- AG and RG from serving RLS can be activated for specific HARQ processes for 2msec TTI
- UE selects E-TFC at each TTI
- Allocates the E-TFC according to the given restrictions
  - Serving grant SG
  - UE transmit power
- Provides priority between the different logical channels
Scheduling Grant Table

<table>
<thead>
<tr>
<th>Index</th>
<th>Scheduled Grant</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>(168/15)^2*6</td>
</tr>
<tr>
<td>36</td>
<td>(150/15)^2*6</td>
</tr>
<tr>
<td>35</td>
<td>(168/15)^2*4</td>
</tr>
<tr>
<td>34</td>
<td>(150/15)^2*4</td>
</tr>
<tr>
<td>33</td>
<td>(134/15)^2*4</td>
</tr>
<tr>
<td>32</td>
<td>(119/15)^2*4</td>
</tr>
<tr>
<td>31</td>
<td>(150/15)^2</td>
</tr>
<tr>
<td>30</td>
<td>(95/15)^2*4</td>
</tr>
<tr>
<td>29</td>
<td>(168/15)^2</td>
</tr>
</tbody>
</table>

- Scheduling grants are max. E-DPDCH / DPCCH power ratio
  - Power Ratio is related to UE data rate

- Relative Grants
  - SG moves up/ down when RG = UP/ DOWN

- Absolute Grants
  - SG jumps to entry for AG
  - 2 reserved values for ZERO_GRANT/ INACTIVE
AG and RG associated with specific uplink E-DCH TTI, i.e. specific HARQ process

- Association based on the timing of the E-AGCH and E-RGCH. Timing is tight enough that this relationship is un-ambiguous.
- Example: 10msec TTI
Scheduling Information

- Happy bit signaling
- One bit status flag send on E-DPCCH at each TTI
- Criterion for happy bit
  - Set to “unhappy” if UE is able to send more data than given with existing serving grant
  - Otherwise set to “happy”

Scheduling Information Reporting

- Content of MAC-e report
  - Provides more detailed information (log. channel, buffer status, UE power headroom)
  - Will be sent less frequently (e.g. every 100 msec)
- Parameters adjusted by RRC (e.g. reporting intervals, channels to report)
HSUPA Resource Allocation

- Capabilities of the UEs
  - MAC-e PDU size limits
  - SF limits

- Node B resources
  - Decoding capability
  - Iub bandwidth capacity

- Cell resources
  - Admissible uplink noise rise/load
  - CAC via RNC
  - Number AGCH/RGCH

- QoS parameters
  - Throughput bounds

**E-DCH Radio Resource Management “E-RRM”**
- Keep uplink load within the limit
- Control E-DCH load portion from non-serving users of other cells
- Control E-DCH resources from each serving user of own cell
- Satisfy QoS/GoS requirements (Ranking PF/SW)
- Maximize HSUPA cell throughput

**Task:** assigns Serving Grants (relative or absolute grants) in terms of a power offset to the current DPCCH power to the UEs in order to control the maximum data rate

**Finally, the UE decides by itself on the used power ratio and the transport block size taking into account the restrictions sent by Node B**
Streaming over HSUPA


- There are two scheduling strategies to support streaming QoS on E-DCH
  - Scheduled transmission
    - A streaming user is scheduled in the Node B like a normal interactive and background user
    - Special priority is given to adjust the minimum guaranteed bit rate (GBR) for the streaming service
    - User can exploit any unused capacity to improve throughput
    - In case of soft/softer handover user might be downgraded by non-serving relative grant from other Node B
  - Non-scheduled transmission
    - Streaming user gets a “non-scheduled” grant assigned by the RNC
    - This provides a guaranteed data rate like R99 DCH
Non-scheduled Transmission (NST)

- Configured by the SRNC
  - UE is allowed to send E-DCH data at any time
  - Signaling overhead and scheduling delay are minimized
  - Support of QoS traffic on E-DCH, e.g. VoIP & SRB

- Characteristics
  - Resource given by SRNC:
    - Non-scheduled Grant = max. # of bits that can be included in a MAC-e PDU
    - UTRAN can reserve HARQ processes for non-scheduled transmission
  - Non-scheduled transmissions defined per MAC-d flow
    - Multiple non-scheduled MAC-d flows may be configured in parallel
    - One specific non-scheduled MAC-d flow can only transmit up to the non-scheduled grant configured for that MAC-d flow
  - Scheduled grants will be considered on top of non-scheduled transmissions
    - Scheduled logical channels cannot use non-scheduled grant
    - Non-scheduled logical channels cannot transmit data using Scheduling Grant
Three different priority schemes
- Both FTP and streaming have no GBR
- Streaming has GBR = 128 kbps, FTP has no GBR
- Non scheduled transmission for streaming

Priority schemes reduce throughput of FTP users

With NST the upload load target cannot be maintained → Call admission control required

Results taken from paper by J. Mueckenheim et al.