Chapter 4

Scheduling

http://mobnet.epfl.ch

Some slides derived from the book by Miao et al.
Background

• Wireless networks have evolved from analog, small-capacity, voice services to digital, large-capacity, data services.
• Users have different service requirements:
  – With unlimited wireless resources: all problems solved
    • Never going to happen!
  – A flexible service architecture to integrate different types of services on a single air-interface;
    • Optimized only for one type of service, other types will experience poor service qualities.
  – Effective QoS management schemes
    • QoS metrics differ between different applications
      – Video telephony has strict delay requirements
      – With different resolutions of videos, the delay requirements differ.
Resource Allocation

• Contention-based access protocols
  – Protocol itself decides the resource allocation, e.g. time slots when a packet should be transmitted
  – Example: Pure ALOHA determines when a packet transmitted and how much spectrum, the whole sys bandwidth, will be used.
  – The design determines the efficiency of resource utilization
    • Inefficient in practice (50% ~65% at most)
    • Risk of terrible performance with heavy traffic
  – Not efficient when many terminals access the network
    • Frequent in cellular networks
    • High efficiency needed in cellular networks

• Reservation-based access protocol with centralized scheduling
  – Most commonly used access protocol in wireless cellular networks
  – High efficiency as well as flexibility in managing wireless resources
A Reservation-based Protocol

- Physical Downlink Control Channel (PDCCH): conveys control information for each user.
- Physical Downlink Shared Channel (PDSCH): multiplex the data of all terminals:
  - Each user will transmit on a unique set of Orthogonal Frequency Division Multiplexing (OFDM) symbols and frequency blocks.
- Reservation phase: PDCCH. Data phase: PDSCH.
Role of Scheduling

• Reservation phase:
  – Estimate channel $\rightarrow$ Feedback $\rightarrow$ Scheduling
Issues in Wireless Scheduling

- Support of mixed classes of traffic desiring different QoS

<table>
<thead>
<tr>
<th>Traffic class</th>
<th>Characteristics</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conversational</td>
<td>Preserve time relation (variation) between information entities of the stream. Conversational pattern (stringent and low delay)</td>
<td>voice</td>
</tr>
<tr>
<td>Streaming</td>
<td>Preserve time relation (variation) between information entities of the stream.</td>
<td>streaming video</td>
</tr>
<tr>
<td>Interactive</td>
<td>Request response pattern. Preserve payload content</td>
<td>web browsing</td>
</tr>
<tr>
<td>Background</td>
<td>Destination is not expecting the data within a certain time. Preserve payload content</td>
<td>emails</td>
</tr>
</tbody>
</table>
Perceived QoS

Best Effort Data

Voice Service

Provided Link Quality

QoS_{min}
Issues in Wireless Scheduling

• Flexibility in allocating network resources
  – Achieved using scheduling algorithms
Issues in Wireless Scheduling

• Channel Variation
  – Shadowing, fading, noise, interference, and user mobility
  – Unstable, error-prone, and hard to predict
  – Capacity of each link varies significantly in different time periods and locations

• Even if the scheduler knows QoS requirements
  – Difficult to estimate the amount of resources needed
  – An adaptive procedure is needed to assure QoS, considering the requirement and channel variation
Uplink capacity with M users

- In general, when there are M uplink users, the sum of the data rates is bounded by (from Shannon–Hartley theorem):

\[
\sum_{i=1}^{M} r_i \leq W \log_2 \left( 1 + \frac{\sum_i P_i g_{i0}}{N_0 W} \right),
\]

where:
- \( W \) is the frequency bandwidth
- \( P_i \) is the transmission power of user \( i \)
- \( g_{i0} \) is the power gain of the channel of user \( i \)
- \( N_0 \) is the noise spectral density
Uplink capacity with two users

• Consider the uplink transmission with two users whose channels are static. The capacity region is the set of all rates \((r_1, r_2)\) that satisfy the following three constraints:

\[
\begin{align*}
  r_1 &\leq W \log_2 \left( 1 + \frac{P_{1g10}}{N_0 W} \right) \\
  r_2 &\leq W \log_2 \left( 1 + \frac{P_{2g20}}{N_0 W} \right) \\
  r_1 + r_2 &\leq W \log_2 \left( 1 + \frac{P_{1g10} + P_{2g20}}{N_0 W} \right)
\end{align*}
\]
Uplink capacity with two users

\[ W \log_2 \left( 1 + \frac{P_1 g_{10} + P_2 g_{20}}{N_0 W} \right) \]

\[ W \log_2 \left( 1 + \frac{P_2 g_{20}}{N_0 W} \right) \]

Capacity region
Uplink capacity with three users

• With three users, the capacity region is the set of all rates \((r_1, r_2, r_3)\) that satisfy the following constraints:

\[
\begin{align*}
  r_1 &\leq W \log_2 \left( 1 + \frac{P_{1g10}}{N_0 W} \right) \\
  r_2 &\leq W \log_2 \left( 1 + \frac{P_{2g20}}{N_0 W} \right) \\
  r_3 &\leq W \log_2 \left( 1 + \frac{P_{3g30}}{N_0 W} \right) \\
  r_1 + r_2 &\leq W \log_2 \left( 1 + \frac{P_{1g10} + P_{2g20}}{N_0 W} \right) \\
  r_2 + r_3 &\leq W \log_2 \left( 1 + \frac{P_{2g20} + P_{3g30}}{N_0 W} \right) \\
  r_1 + r_3 &\leq W \log_2 \left( 1 + \frac{P_{1g10} + P_{3g30}}{N_0 W} \right) \\
  r_1 + r_2 + r_3 &\leq W \log_2 \left( 1 + \frac{P_{1g10} + P_{2g20} + P_{3g30}}{N_0 W} \right).
\end{align*}
\]
Uplink capacity with three users
Downlink capacity with $M$ users

- In the downlink, the base station sends independent data streams to multiple users.
- Assuming $g_{01} \leq g_{02} \leq \ldots \leq g_{0M}$, the capacity region is given by:

$$r_m \leq W \log_2 \left( 1 + \frac{P_m g_{0m}}{\sum_{i=m+1}^{M} P_i g_{0i} + N_0 W} \right), \quad \forall m$$

for all possible power allocations $\sum_m P_m = P_0$, where $P_0$ is the total transmission power of the base station.
Downlink capacity with two users
Duality of the capacity regions

![Diagram showing the capacity regions of downlink and uplink multiple access channels.](image)
Wireless Packet Scheduling Algorithms

• Benefit of using scheduling in wireless networks
  – Flexible in adaptation to QoS and channel
  – A single air-interface to integrate various service types
  – Significant performance improvement w/o more spectrum

• Model: $M$ users served on a single channel; TDMA, one user in one slot; each user has a buffer (to store the packets to be sent)
Downlink Scheduling
Uplink Scheduling

\[ \eta_0[t] \rightarrow q_0[t] \rightarrow \lambda_0 \]

\[ \eta_1[t] \rightarrow q_1[t] \rightarrow \lambda_1 \]

\[ \eta_{M-1}[t] \rightarrow q_{M-1}[t] \rightarrow \lambda_{M-1} \]
Queue Modeling

- At timeslot $t$, the queue of user $i$ changes as follows:

$$q_i[t + 1] = q_i[t] + \delta_i[t] - \eta_i[t]$$

Number of bits in the queue

Number of new bits arriving

Number of bits scheduled to transmit, determined by scheduling.
Round-Robin (RR) Scheduling

- Users are scheduled in a round robin, i.e. cyclic order
- $i[t]$: user scheduled at time $t$.
- RR scheduler: $i[t+1] = i[t] + 1 \pmod{M - 1}$.
- Fair: all users scheduled the same amount of resources.
Round-Robin Scheduling

- Performance
  - All users are allocated the same amount of network resources
  - What is the throughput of all users in the following network?
Max Throughput Scheduling

- Objective: maximize total network throughput
- If user \( i \) is scheduled, the expected data rate is

\[
\hat{r}_i[t] = \frac{\hat{n}_i[t]}{T_s}
\]

Expected number of bits that can be successfully delivered
Slot length

- The total expected network throughput is

\[
\hat{r}[t] = \sum_{i=0}^{M-1} \hat{r}_i[t]I(i)
\]

\( I(i) \): Scheduling indicator: 1 scheduled, 0 otherwise.
Max Throughput Scheduling

- Schedule the user with the highest expected data rate
  - one way of estimating $\hat{r}_i[t]$ is:

  $$\hat{r}_i[t] = W \log_2 \left( 1 + \frac{\Gamma_i[t]}{\theta} \right)$$

where:
- $W$ is the frequency bandwidth
- $\Gamma_i[t]$ is the signal-to-interference-plus-noise ratio (SINR) at time $t$ given the allocated power.
- $\theta$ is the SINR gap that defines the gap between the channel capacity and a practical coding and modulation scheme

→ Max SINR or Max C/I (Carrier to Interference) scheduler
Max Throughput Scheduling

• Main drawbacks
  – Unfairness
  – Coverage limitation
  – Most users may never be served
Proportional Fair Scheduling

• PF scheduling: a compromised policy to balance the competing interests of network throughput and minimum service level

• Objective: maximize \[
\sum_{i=0}^{M-1} \ln S_i
\]

• \(S_i\): long-run throughput for user \(i\) can be predicted using

\[
\hat{S}_i[t] = (1 - \frac{1}{\tau})S_i[t - 1] + \frac{1}{\tau}\hat{r}_i[t]I(i)
\]

where \(\tau >> 1\) is a constant defined by the scheduler.

\[\rightarrow \text{ schedule the user with the highest } \frac{\hat{r}_i[t]}{S_i[t - 1]}\]
Proportional Fair Scheduling

- Meet the proportional fairness criterion:
  Assuming $x_i$ is the optimal value; For any other feasible value, the sum of proportional changes is non-positive, i.e.

$$\sum_{i=1}^{n} \frac{\hat{x}_i - x_i}{x_i} \leq 0$$

- Intuition: when the scheduling result is already proportional fair, changing the scheduling such that the throughput of any user is increased by a percentage, the cumulative decrease of the throughputs of the other users will be higher.

- In other words, any attempt of improvement somewhere will generate higher damage elsewhere.
Proportional Fair Scheduling

• Why proportional fair

Assume $\delta_i \ll x_i$.

$$\sum_{i=1}^{n} \ln(x_i + \delta_i) = \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \ln \left(1 + \frac{\delta_i}{x_i}\right)$$

$$\approx \sum_{i=1}^{n} \ln x_i + \sum_{i=1}^{n} \frac{\delta_i}{x_i}$$

$$\leq \sum_{i=1}^{n} \ln x_i.$$
Max-Min Scheduling

- Objective: maximize the minimum user throughput
  \[
  \max \min_i S_i
  \]

- A scheduling result is max-min fair if and only if a further increase of throughput of one user will result in the decrease of a user with a smaller throughput
  \[
  \hat{S}_i[t] = (1 - \frac{1}{\tau})S_i[t - 1] + \frac{1}{\tau}\hat{r}_i[t]I(i)
  \]
  \[
  \max \min_i (1 - \frac{1}{\tau})S_i[t - 1] + \frac{1}{\tau}\hat{r}_i[t]I(i).
  \]

- Schedule the user with the minimum \((1 - \frac{1}{\tau})S_i[t - 1]\), i.e. the one with the smallest throughput at time \(t-1\).
Max-Min Scheduling

- $M$ empty cylindrical buckets (users), all with the same radius but different heights.
- Allocate water
Max-Min Scheduling

- $M$ empty cylindrical buckets (users), all with the same radius but different heights.
- Allocate water
Can these schedulers deal with QoS?
Max Utility Scheduling

- Previous schedulers do not consider QoS
- Utility-based scheduling
  - Utility quantifies the satisfaction of each user given the allocated resources
  - Model the QoS perception of users
  - Objective: maximize the sum utility of all users, i.e. total network satisfaction
- Utility functions: model how user perceives services
Max Utility Scheduling

$$\max \sum_{i=0}^{M-1} U_i(S_i)$$

- Different utility functions can be designed.
- In addition to QoS modeling, different utility functions can be designed to reflect fairness and efficiency.
- Max-throughput (highest efficiency):
  $$U(S) = S$$
- Proportional fair:
  $$U(S) = \ln(S)$$
Max Utility Scheduling - Alpha Fair Utility

- More generic definition: $\alpha$ fair scheduling

$$U_\alpha(S) = \begin{cases} \frac{S^{1-\alpha}}{1-\alpha} & \alpha \geq 0 \text{ and } \neq 1 \\ \ln(S) & \alpha = 1. \end{cases}$$

- $\alpha$ measures how fair the scheduling result is
  - 0: Max throughput;
  - 1: Proportional fair;
  - 2: equivalent to minimizing $\sum_{i=0}^{M-1} \frac{1}{S_i}$
    - Minimize the total potential delay
  - Infinity: Most fair, max-min scheduler.

\[\begin{array}{cccc}
\text{Max throughput} & \text{PF} & \text{Min potential delay} & \text{Max-min} \\
0 & 1 & 2 & \infty
\end{array}\]
Alpha Fair Utility
Utility Functions with QoS Consideration

- Determined based on traffic characteristics

  - (a) best effort;
  - (b) real time with tight delay requirement;
  - (c) real time with loose delay requirement.
How to Schedule Users

\[
\max \sum_{i=0}^{M-1} U_i(S_i[t]) = \max \sum_{i=0}^{M-1} U_i \left( (1 - \frac{1}{\tau})S_i[t-1] + \frac{1}{\tau} \hat{r}_i[t]I(i) \right)
\]

Since \((1 - \frac{1}{\tau})S_i[t-1] >> \frac{1}{\tau} \hat{r}_i[t]I(i)\)

We have

\[
U_i \left( (1 - \frac{1}{\tau})S_i[t-1] + \frac{1}{\tau} \hat{r}_i[t]I(i) \right) \\
\approx U_i \left( (1 - \frac{1}{\tau})S_i[t-1] \right) + U'_i \left( (1 - \frac{1}{\tau})S_i[t-1] \right) \frac{1}{\tau} \hat{r}_i[t]I(i)
\]

• Since \(U_i \left( (1 - \frac{1}{\tau})S_i[t-1] \right)\) is fixed at time t

→ Equivalent objective: \(\max \sum_{i=0}^{M-1} U'_i \left( (1 - \frac{1}{\tau})S_i[t-1] \right) \frac{1}{\tau} \hat{r}_i[t]I(i)\)

• Schedule the user with the largest: \(U'_i \left( (1 - \frac{1}{\tau})S_i[t-1] \right) \hat{r}_i[t]\)

or \(U'_i (S_i[t-1]) \hat{r}_i[t]\) (because \(T >> 1 \rightarrow 1/\tau \rightarrow 0\))
Performance Comparison

- In a cellular network, users requesting different types of services from the BS. Ten users dropped randomly in each cell.

<table>
<thead>
<tr>
<th>Traffic</th>
<th>Type</th>
<th>Basic rate requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>VoIP</td>
<td>Real Time</td>
<td>102 Kbps</td>
</tr>
<tr>
<td>Video Streaming</td>
<td>Real Time</td>
<td>580 Kbps</td>
</tr>
<tr>
<td>High Rate File Download</td>
<td>Best effort</td>
<td>1.74 Mbps</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Schedulers</th>
<th>Average throughput (Mbps)</th>
<th>Outage probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max-SINR</td>
<td>17.4300</td>
<td>5.8762</td>
</tr>
<tr>
<td>Round Robin</td>
<td>9.1745</td>
<td>2.9555</td>
</tr>
<tr>
<td>Proportional</td>
<td>12.7160</td>
<td>2.0584</td>
</tr>
<tr>
<td>Utility</td>
<td>9.5690</td>
<td>0</td>
</tr>
</tbody>
</table>
Scheduling in OFDMA Systems

• One more dimension of resources
  – Subcarrier allocation

• Different users experience, independent wireless channels and their subcarriers may experience substantially different channel gains because of the frequency selectivity in the channels.

• Theoretically it is possible to set the data rate for each subcarrier based on its channel quality and power allocated.
Scheduling in OFDMA Systems

- Subcarrier in deep fading for one user may not be used by this user as sending bits on these subcarriers costs too much power
  - May be in good channel conditions to other users and can be used by them instead
- Scheduling of subcarriers in an adaptive way based on the instantaneous channel qualities
- OFDMA systems typically use adaptive subcarrier assignment, power allocation, modulation and coding to exploit the diversity in multiple users and frequency to improve the network performance
System Model

Base Station

User m

Subcarrier assignment, power allocation, and adaptive MCS decision

CSI feedback from M users

Subcarrier and bit allocation

Adaptive modulation and coding 0

Adaptive modulation and coding 1

Adaptive modulation and coding K-1

IFFT

Frequency-selective channel

CSI feedback

Adaptive demodulation and decoding m_k

Adaptive demodulation and decoding m_k+1

Subcarrier allocation and bit loading information, channel estimation

Adaptive demodulation and decoding m_k
Signaling in OFDMA Scheduling

• Additional signal overhead is necessary as all users need to feed back CSI
• The base station has to inform all users the resource allocation result.
• The overhead is rather small, especially in slow-fading channels, and the resource allocation may be updated once every many OFDM symbols.
• Besides there are also many other techniques to further reduce the signaling overhead, like bundling of adjacent subcarriers.
Max-Throughput OFDMA Scheduling

- Round-robin, max-throughput, PF, max-min, and max-utility scheduling can all be done for OFDMA.
- Take max-throughput scheduling as an example.
  - Schedule the users such that the total network throughput is maximized
- The expected throughput of user $i$ in the $t$-th OFDM symbol is

$$ R_i[t] = \sum_j W \log_2 \left( 1 + \frac{p_{ij}[t] \gamma_{ij}[t]}{\theta} \right) I(i, j) $$

where

- $W$ is the subcarrier bandwidth.
- $p_{ij}[t]$ is the power allocation on the $j$-th subcarrier of terminal $i$.
- $\gamma_{ij}[t]$ is the ratio between the channel gain and the interference, hence $SINR = p_{ij}[t] \times \gamma_{ij}[t]$
- $I(i, j) = 1$ if the $j$-th subcarrier is assigned to terminal $i$ and 0 otherwise.
Max-Throughput OFDMA Scheduling

• Each subcarrier is assigned to only one user and we have the following constraint

\[ \sum_i I(i, j) = 1, \forall j. \]

• Overall network throughput

\[ R[t] = \sum_{i=1}^{M} R_i[t] = \sum_i \sum_j W \log_2 \left( 1 + \frac{p_{ij}[t] \gamma_{ij}[t]}{\theta} \right) I(i, j) \]

• W/o and w/ power adaptation
Without power adaptation

• Assume no adaptive power allocation

\[ R[t] = \sum_{i=1}^{M} R_i[t] = \sum_{i} \sum_{j} W \log_2 \left( 1 + \frac{p_{ij}[t] \gamma_{ij}[t]}{\theta} \right) I(i, j) \]

• In order to maximize the network throughput, subcarrier \( i \) should be allocated to the user whose SINR/\( \theta \) on this subcarrier is the highest among all users.

\[ I^*(i, j) = \begin{cases} 1 & \frac{p_{ij[t] \gamma_{ij[t]}}}{\theta} \geq \frac{p_{mj[t] \gamma_{mj[t]}}}{\theta}, \forall m \\ 0 & \text{otherwise.} \end{cases} \]
With power adaptation

• Assume adaptive power allocation

\[
R[t] = \sum_{i=1}^{M} R_i[t] = \sum_i \sum_j W \log_2 \left( 1 + \frac{p_{ij}[t]\gamma_{ij}[t]}{\theta} \right) I(i, j)
\]

• Subcarrier allocation
  – Each subcarrier should be assigned to the user with the highest \( \gamma_{ij}[t] \) as the rate increase by using any amount of power on any subcarrier will be maximized if the subcarrier has the highest \( \gamma_{ij}[t] \).

\[
I^*(i, j) = \begin{cases} 
1 & \gamma_{ij}[t] \geq \gamma_{mj}[t], \forall m \\
0 & \text{otherwise}
\end{cases}
\]

• How much power should be used on each subcarrier?
With power adaptation

- The BS has a total power constraint

\[
P^*[t] = \arg_{P[t]} \sum_i \sum_j W \log_2 \left(1 + \frac{p_{ij}[t] \gamma_{ij}[t]}{\theta} \right) I^*(i, j)
\]

\[
s.t. \sum_{i,j} p_{ij}[t] \leq P_o,
\]

where \(P_o\) is the total transmit power limit of the BS power amplifier.

- Concave problem and the Lagrangian method can be used to find the optimal power allocation.

\[
p_{ij}[t] = \left[\lambda - \frac{\theta}{\gamma_{ij}[t]}\right]^+ I^*(i, j)
\]

with the constraint

\[
\sum_{i,j} p_{ij}[t] = \sum_{i,j} \left[\lambda - \frac{\theta}{\gamma_{ij}[t]}\right]^+ I^*(i, j) = P_o
\]
Other OFDMA Schedulers

• Max-throughput OFDMA scheduling also suffers unfairness and coverage limitations
  – Users closer to the BS have more subcarrier assignment
  – Users at cell edges may never be served
• Fair scheduling algorithms for OFDMA systems have been extensively studied
• QoS consideration further complicates the scheduling
  – Mixed types of traffic, utility based scheduling should be used
• Generally these schedulers have no closed-form expressions
  – Usually based on optimization techniques
  – Much more complicated than the schedulers introduced in this chapter
  – Refer to the references for more details