Downlink Scheduling and Resource Allocation for OFDM Systems

J. Huang\textsuperscript{1} \quad V. Subramanian\textsuperscript{2} \quad R. Agrawal \textsuperscript{2} \quad R. Berry\textsuperscript{3}

\textsuperscript{1}Princeton University

\textsuperscript{2}Motorola Inc.

\textsuperscript{3}Northwestern University

CISS, March 2006
“Channel Aware” Scheduling and Resource Allocation

- Dynamically schedule users based on channel conditions/QoS.
  - Use frequent channel quality feedback & adaptive modulation/coding.
  - Exploit multi-user diversity.
- Key component of recent wireless data systems
  - e.g. CDMA 1xEVDO, HSPDA, IEEE 802.16.
Scheduler needs to balance users’ QoS and global efficiency. Many approaches accomplish this via gradient-based scheduling. Assign each user a utility, \( U_i(\cdot) \), depending on delay, throughput, etc. Scheduler maximizes first order change in total utility. i.e. choose rate \( r = (r_1, \ldots, r_N)^T \) to solve:

\[
\max_{r \in \mathcal{R}(e)} \nabla U(\mathbf{X}(t)) \cdot r = \max_{r \in \mathcal{R}(e)} \sum_i \dot{U}_i(X_i(t)) r_i,
\]

- Myopic policy, requires no knowledge of channel or arrival statistics.
\section*{Gradient-based Scheduling Examples}

- **\(\alpha\)-fairness** - utility function of average throughput \(W_i\):

\[
U_i(W_i) = \begin{cases} \frac{1}{\alpha}(W_i)^\alpha, & \alpha \leq 1, \alpha \neq 0 \\ \log(W_i), & \alpha = 0 \end{cases}
\]

- \(\alpha = 0 \Rightarrow\) Prop. fair.
- \(\alpha = 1 \Rightarrow\) Max. throughput.

- Utility may also be function of delay/queue size.
  - e.g. Stabilizing policies.
State-dependent Feasible Rate Regions

- Optimization is over feasible rate region $\mathcal{R}(e_t)$.
  - $e_t = \text{available channel state information}$.
  - Region depends on physical layer technology and multiplexing used.

- Examples:
  - TDM systems
  - CDMA (HSDPA), OFDMA (802.16)
  - Requires allocating physical layer resources among scheduled users.
State-dependent Feasible Rate Regions

Optimization is over feasible rate region $\mathcal{R}(e_t)$.

- $e_t =$ available channel state information.
- Region depends on physical layer technology and multiplexing used.

**Ex: TDM systems**

- $\mathcal{R}(e_t) =$ simplex with max rate $r_i$ for each user $i$.
- Gradient-policy $\Rightarrow$ schedule users with max $\dot{U}_i(X_i)r_i$. 
State-dependent Feasible Rate Regions

- Optimization is over feasible rate region $\mathcal{R}(e_t)$.
  - $e_t = \text{available channel state information}$.
  - Region depends on physical layer technology and multiplexing used.

**Ex: TDM systems**
- $\mathcal{R}(e_t) = \text{simplex with max rate } r_i \text{ for each user } i$.
- Gradient-policy $\Rightarrow$ schedule users with $\max \dot{U}_i(X_i)r_i$.

- In many systems, additional multiplexing within a time-slot.
  - e.g. CDMA (HSDPA), OFDMA (802.16).
  - Requires allocating physical layer resources among scheduled users.
OFDMA systems

- Frequency band divided into $N$ subcarriers/tones.
- Resource allocation:
  - assigning tones to users
  - allocate power across tones.
OFDMA rate region

- Initially, allow users to time-share each tone.
  - In practice, one user/tone.
- Assume rate/tone $= \log(1 + SNR)$.
OFDMA rate region

- Initially, allow users to time-share each tone.
  - In practice, one user/tone.
- Assume rate/tone = \( \log(1 + SNR) \).
- **Rate region** (similar to Li/Goldsmith, Wang, et al, etc.):

\[
\mathcal{R}(e) = \left\{ r : r_i = \sum_j x_{ij} \log \left( 1 + \frac{p_{ij} e_{ij}}{x_{ij}} \right), \sum_{ij} p_{ij} \leq P, \sum_i x_{ij} \leq 1, \forall j, (x, p) \in \mathcal{X} \right\},
\]

where
- \( \mathcal{X} := \{(x, p) \geq 0 : x_{ij} \leq 1, \forall i, j\} \).
- \( x_{ij} \) = fraction of subchannel \( j \) allocated to user \( i \).
- \( p_{ij} \) = power allocated to user \( i \) on subchannel \( j \).
- \( e_{ij} \) = received SNR/unit power.
Model Variations

1. **Maximum SINR constraint:** $s_{ij}$ (limit on modulation order)

   - Let
   
   $\mathcal{X} := \left\{ (x, p) \geq 0 : 0 \leq x_{ij} \leq 1, 0 \leq p_{ij} \leq \frac{x_{ij}s_{ij}}{e_{ij}} \forall i, j \right\}$.  \hspace{1cm} (1)
Model Variations

1. **Maximum SINR constraint:** $s_{ij}$ (limit on modulation order)
   - Let
   $$\mathcal{X} := \left\{ (x, p) \geq 0 : 0 \leq x_{ij} \leq 1, 0 \leq p_{ij} \leq \frac{x_{ij}s_{ij}}{e_{ij}} \forall i, j \right\}.$$  

2. **Sub-channelization** (bundle tones to reduce overhead)
   - Possible channelizations:
     - Interleaved (802.16 standard mode)
     - Adjacent (Band AMC mode)
     - Random (e.g. frequency hopped)
   - Can accommodate by letting $x_{ij} =$ allocation of subchannel $j$.
   - View $e_{ij}$ as “average” SNR/subchannel.
The optimal gradient-based scheduling algorithm must solve:

\[
\max_{x_{ij}, p_{ij} \in \mathcal{X}} V(x, p) := \sum_i w_i \sum_j x_{ij} \log \left( 1 + \frac{p_{ij} e_{ij}}{x_{ij}} \right)
\]

subject to: \( \sum_{i,j} p_{ij} \leq P \), and \( \sum_i x_{ij} \leq 1, \forall j \in \mathcal{N} \),

- Need to solve every scheduling interval.
- We consider optimal and suboptimal algorithms for this.
Optimal algorithm

- Scheduling problem (OPT) is convex and satisfies Slater’s condition. ⇒ No duality gap.
- Consider Lagrangian:

\[
L(x, p, \lambda, \mu) := \sum_i w_i \sum_j x_{ij} \log \left(1 + \frac{p_{ij} e_{ij}}{x_{ij}}\right) + \lambda \left(P - \sum_{i,j} p_{ij}\right) + \sum_j \mu_j \left(1 - \sum_i x_{ij}\right).
\]

- Associated dual function:

\[
L(\lambda, \mu) = \max_{(x, p) \in \mathcal{X}} L(x, p, \lambda, \mu)
\]

- By duality, optimal solution to (OPT) is:

\[
V^* = \min_{(\lambda, \mu) \geq 0} L(\lambda, \mu)
\]
Dual Function

- Can explicitly solve for the dual function.
- Fixing $x, \lambda, \mu$, optimizing over $p_{ij} \Rightarrow$ “water-filling” like solution.

\[ p_{ij}^* = \frac{x_{ij}}{e_{ij}} \left[ \left( \frac{w_i e_{ij}}{\lambda} - 1 \right)^+ \right] \wedge s_{ij} \]
Dual Function

- Can explicitly solve for the dual function.
- Fixing $x, \lambda, \mu$, optimizing over $p_{ij} \Rightarrow \text{“water-filling” like solution.}$

$$p_{ij}^* = \frac{x_{ij}}{e_{ij}} \left[ \left( \frac{w_i e_{ij}}{\lambda} - 1 \right)^+ \wedge s_{ij} \right].$$

- Given optimum $p_{ij}^*$,

$$L(x, p^*, \lambda, \mu) = \sum_{ij} x_{ij} (\mu_{ij}(\lambda) - \mu_j) + \sum_j \mu_j + \lambda P$$

  $\Rightarrow$ Optimizing over $x_{ij} \in [0, 1]$ is now easy.

$$\Rightarrow L(\lambda, \mu) = \sum_{ij} (\mu_{ij}(\lambda) - \mu_j)^+ + \sum_j \mu_j + \lambda P$$
Minimizing the dual function

- **Dual function:**
  \[
  L(\lambda, \mu) = \sum_{ij} (\mu_{ij}(\lambda) - \mu_j)^+ + \sum_j \mu_j + \lambda P.
  \]

- **First minimize over** \(\mu\):
  \[
  L(\lambda) := \min_{\mu \geq 0} L(\lambda, \mu) = \lambda P + \sum_j \max_i \mu_{ij}(\lambda).
  \]
  ▶ Requires one sort of users per subchannel.
Minimizing the dual function

- Dual function:

\[ L(\lambda, \mu) = \sum_{ij} (\mu_{ij}(\lambda) - \mu_j)^+ + \sum_j \mu_j + \lambda P. \]

- First minimize over \( \mu \):

\[ L(\lambda) := \min_{\mu \geq 0} L(\lambda, \mu) = \lambda P + \sum_j \max_i \mu_{ij}(\lambda). \]

  - Requires one sort of users per subchannel.

- \( L(\lambda) \) is convex function of \( \lambda \).
  - Can minimize using iterated 1-D search (e.g. golden section).
Optimal Primal Values.

- Given $\lambda^*, \mu^*$, let

$$ (x^*, p^*) = \arg \max_{(x, p) \in \mathcal{X}} L(x, p, \lambda^*, \mu^*). \quad (*) $$

- If $(x^*, p^*)$ are primal feasible and satisfy complimentary slackness, they are optimal scheduling decision.

- Can find these as before, except multiple $\mu_{ij}$’s may be tied at the maximum value.
  - $\Rightarrow$ Multiple $x_{ij}$’s can be $> 0$.
    - Not all choices result in feasible primal solutions.
Breaking ties - optimal time-sharing

- When ties occur, can show $L(\lambda)$ is not differentiable.
- Each $(x^*, p^*)$ that satisfy (*) and complimentary slackness give a subgradient of $L(\lambda)$.
- Simple sort can find max and min subgradients (one user/subchannel).
- Time-sharing between these gives a primal optimal solution.
  - At most 2 users/subchannel.
In practice typically restricted to one user/subchannel.

- Given optimal dual solution, if no “ties” this will be satisfied.
- When ties occurs, selecting one user involved in the tie corresponds to choosing one subgradient.
- In simulations, we choose the user that corresponds to the smallest negative subgradient.
  - Other mechanisms also possible.
  - Resulting power constraint will not be tight.
Re-optimizing the power allocation

- Given a feasible $\mathbf{x}$, consider

$$\max_{\mathbf{p}:(\mathbf{p}, \mathbf{x}) \in \mathcal{X}} V(\mathbf{x}, \mathbf{p}) \quad \text{s.t.} \quad \sum_{ij} p_{ij} \leq P$$

- Solution again “water-filling” like power allocation with a given Lagrange multiplier $\tilde{\lambda}$.
- Optimal $\tilde{\lambda}$ can be shown to satisfy fixed point equation

$$\lambda = f(\lambda),$$

$f(\lambda)$ is increasing, finite-valued (piece-wise constant).

$\Rightarrow$ finite time algorithm for finding $\tilde{\lambda}$. 
Single Sort Heuristic

- Optimal subchannel assignment is to user with max $\mu_{ij}(\lambda)$.
  - Requires iterating to find optimal $\lambda$.
- Instead consider single-sort using metric $w_{ij}\bar{R}_{ij}$,

$$\bar{R}_{ij} = \log[1 + (s_{ij} \wedge (e_{ij}P/N))].$$

Motivated by e.g. [Hoo, et al.].

- Then optimally allocate power as before.
- Other heuristics in paper.
Numerical Results

Simulation set-up:

- Single cell, $M = 40$ users.
- $e_{ij} = \text{(fixed location-based term)} \times \text{(frequency selective fast fading)}$
  - Fixed term = empirical distribution.
  - frequency selective term = block fading in time ($2\text{ msec coh. time}$);
    standard ref. mobile delay spread ($1 \mu\text{sec} \approx 250\text{MHz Doppler}$).
- 5 MHz BW, 512 tones.
- Initially adjacent channelization, 8 tones/subchannel.
- use $\alpha$-utility functions.
- Simulate full algorithm (with one user/subchannel) and single sort.
Different choices of $\alpha$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Algorithm</th>
<th>Utility</th>
<th>Log U</th>
<th>Rate (kbps)</th>
<th>Num.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>FULL</td>
<td>1236</td>
<td>12.58</td>
<td>497.8</td>
<td>5.40</td>
</tr>
<tr>
<td>0.5</td>
<td>MO-$\bar{w}\bar{R}$</td>
<td>1234</td>
<td>12.56</td>
<td>498.3</td>
<td>5.17</td>
</tr>
<tr>
<td>0</td>
<td>FULL</td>
<td>12.69</td>
<td>12.69</td>
<td>396.8</td>
<td>5.75</td>
</tr>
<tr>
<td>0</td>
<td>MO-$\bar{w}\bar{R}$</td>
<td>12.68</td>
<td>12.68</td>
<td>393.0</td>
<td>5.47</td>
</tr>
<tr>
<td>1</td>
<td>FULL</td>
<td>716955</td>
<td>8.04</td>
<td>719.3</td>
<td>3.04</td>
</tr>
<tr>
<td>1</td>
<td>MO-$\bar{w}\bar{R}$</td>
<td>716955</td>
<td>8.04</td>
<td>719.3</td>
<td>3.04</td>
</tr>
</tbody>
</table>
\( \alpha = 0.5 \).
Different channelization schemes

<table>
<thead>
<tr>
<th>Chan.</th>
<th>Algorithm</th>
<th>Utility</th>
<th>Log U</th>
<th>Rate (kbps)</th>
<th>Num.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj.</td>
<td>FULL</td>
<td>1236</td>
<td>12.58</td>
<td>497.8</td>
<td>5.40</td>
</tr>
<tr>
<td>Adj.</td>
<td>MO-wR</td>
<td>1234</td>
<td>12.56</td>
<td>498.3</td>
<td>5.17</td>
</tr>
<tr>
<td>Ran.</td>
<td>FULL</td>
<td>1171</td>
<td>12.42</td>
<td>465.2</td>
<td>4.08</td>
</tr>
<tr>
<td>Ran.</td>
<td>MO-wR</td>
<td>1167</td>
<td>12.40</td>
<td>465.5</td>
<td>3.64</td>
</tr>
<tr>
<td>Int.</td>
<td>FULL</td>
<td>1136</td>
<td>12.32</td>
<td>447.1</td>
<td>1</td>
</tr>
<tr>
<td>Int.</td>
<td>MO-wR</td>
<td>1142</td>
<td>12.33</td>
<td>455.2</td>
<td>1</td>
</tr>
</tbody>
</table>

Upperbound on rate/channel; looser for interleaved/random case.
Conclusions

- Presented optimal and sub-optimal algorithms for gradient-based scheduling in OFDM systems.
- Can accommodate different channelizations and max. SINR constraints.
- Found simple sort has near optimal performance.