Downlink Scheduling and Resource Allocation for OFDM Systems

J. Huang[1] V. Subramanian[2] R. Agrawal [2] R. Berry[3]

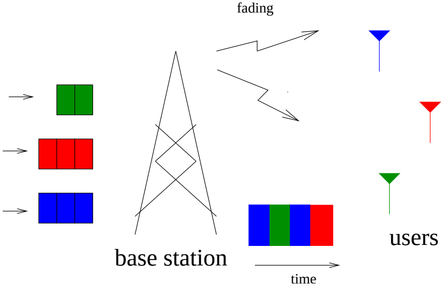
[1]Princeton University

[2]Motorola Inc.

[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.

Gradient-based Scheduling

- 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 $\mathbf{r} = (r_1, \ldots, r_N)^T$ to solve:

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

  - Myopic policy, requires no knowledge of channel or arrival statistics.

Gradient-based Scheduling Examples

- **α-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 $\mathcal{R}(\mathbf{e}_t)$.
  - $\mathbf{e}_t$ = available channel state information.
  - Region depends on physical layer technology and multiplexing used.

State-dependent Feasible Rate Regions

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

Ex: TDM systems

- $\mathcal{R}(\mathbf{e}_t)$ = 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

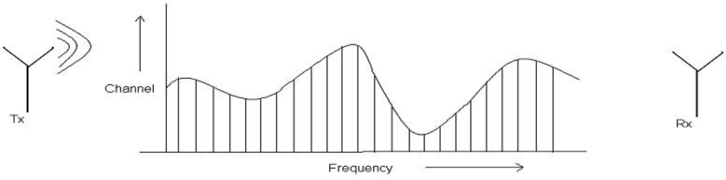
- Optimization is over feasible rate region $\mathcal{R}(\mathbf{e}_t)$.
  - $\mathbf{e}_t$ = available channel state information.
  - Region depends on physical layer technology and multiplexing used.

Ex: TDM systems

  - $\mathcal{R}(\mathbf{e}_t)$ = simplex with max rate r_i for each user i.
  - Gradient-policy $\Rightarrow$ schedule users with max $\dot{U}_i(X_i) r_i$.

- 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.):

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

where

- $\mathcal{X} := \{(\mathbf{x}, \mathbf{p}) \ge \mathbf{0} : x_{ij} \le 1, \ \forall i, j\}$.
- 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
$$\mathcal{X} := \left\{ (\mathbf{x}, \mathbf{p}) \geq \mathbf{0} : 0 \leq x_{ij} \leq 1, 0 \leq p_{ij} \leq \frac{x_{ij} s_{ij}}{e_{ij}} \ \forall i, j \right\}. \qquad (1)$$

Model Variations

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

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.

Optimal Scheduling algorithm

The optimal gradient-based scheduling algorithm must solve:

$$\max_{x_{ij}, p_{ij} \in \mathcal{X}} V(\mathbf{x}, \mathbf{p}) := \sum_i w_i \sum_j x_{ij} \log\left(1 + \frac{p_{ij} e_{ij}}{x_{ij}}\right)$$
$$\text{subject to:} \quad \sum_{i,j} p_{ij} \leq P, \text{ and } \sum_i x_{ij} \leq 1, \forall j \in \mathcal{N}, \tag{OPT}$$

- 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.
  $\Rightarrow$ No duality gap.
- Consider Lagrangian:

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

- Associated dual function:

$$L(\lambda, \boldsymbol{\mu}) = \max_{(\mathbf{x},\mathbf{p}) \in \mathcal{X}} L(\mathbf{x}, \mathbf{p}, \lambda, \boldsymbol{\mu})$$

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

$$V^* = \min_{(\lambda, \boldsymbol{\mu}) \geq \mathbf{0}} L(\lambda, \boldsymbol{\mu})$$

Dual Function

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

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

Dual Function

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

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

- Given optimum p_{ij}^*,

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

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

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

Minimizing the dual function

- Dual function:

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

- First minimize over $\boldsymbol{\mu}$:

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

    - Requires one sort of users per subchannel.

Minimizing the dual function

- Dual function:

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

- First minimize over $\boldsymbol{\mu}$:

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

  - Requires one sort of users per subchannel.

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

Optimal Primal Values.

- Given λ^*, $\boldsymbol{\mu}^*$, let

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

- If $(\mathbf{x}^*, \mathbf{p}^*)$ are primal feasible and satisfy complimentary slackness, they are optimal scheduling decision.
- Can find these as before, except multiple μ_{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 $(\mathbf{x}^*, \mathbf{p}^*)$ 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.

Single User per Subchannel Heuristic

- 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

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

- 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 λ.
- 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

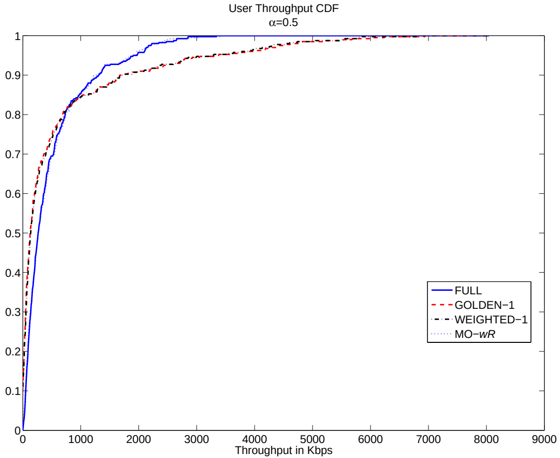
Simulation set-up:

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

Different choices of α

α	Algorithm	Utility	Log U	Rate(kbps)	Num.
0.5	FULL	1236	12.58	497.8	5.40
0.5	MO-$w\bar{R}$	1234	12.56	498.3	5.17
0	FULL	12.69	12.69	396.8	5.75
0	MO-$w\bar{R}$	12.68	12.68	393.0	5.47
1	FULL	716955	8.04	719.3	3.04
1	MO-$w\bar{R}$	716955	8.04	719.3	3.04

User throughput CDFs



$\alpha = 0.5$.

Different channelization schemes

Chan.	Algorithm	Utility	Log U	Rate (kbps)	Num.
Adj.	FULL	1236	12.58	497.8	5.40
Adj.	MO-$w\bar{R}$	1234	12.56	498.3	5.17
Ran.	FULL	1171	12.42	465.2	4.08
Ran.	MO-$w\bar{R}$	1167	12.40	465.5	3.64
Int.	FULL	1136	12.32	447.1	1
Int.	MO-$w\bar{R}$	1142	12.33	455.2	1

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.