Scheduling for Multi-user MIMO Systems

Marios Kountouris and David Gesbert

Mobile Communications Department
Eurecom Institute, Sophia - Antipolis

February 2005

Outline

- Introduction
  - Challenges and design objectives in wireless networks
  - Cross-layer design
  - Capacity of MIMO Broadcast Channel

- Part I

  Memory-based Opportunistic Multiuser Beamforming

- Part II

  QoS-based Scheduling for Multiuser MIMO systems

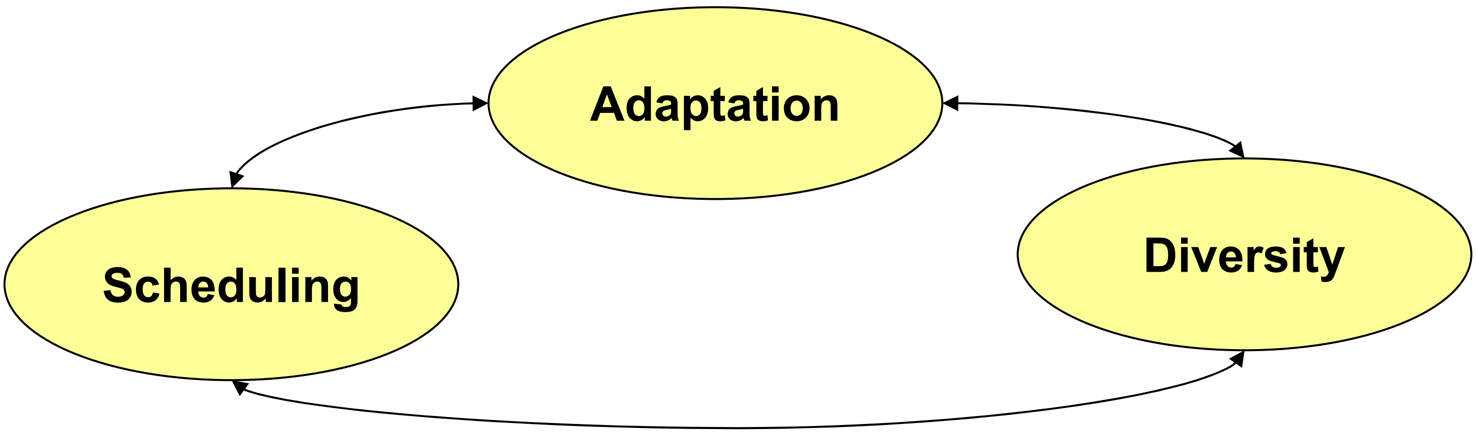
Challenges in Wireless Networks

- Wireless channel is a demanding, capacity-limited and shared communication medium
- Traffic patterns, user locations, and user channel conditions are constantly changing
- Demand for a wide variety of data services with a broad range of Quality of Service (QoS) requirements
- Some applications are heterogeneous with hard constraints that must be met by the network
- Energy and delay constraints change design principles across all layers of the protocol stack

Design Objectives

End-to-end QoS

- QoS satisfaction is an interplay among:



- Degrees of Freedom:
  - Space
  - Time
  - Frequency
  - Multi-user

MIMO technology

Employs multiple transmit and receive antennas, to exploit channel fading effects constructively, and achieve different kinds of gains:

– Spatial diversity gain (combat fading, stabilize link quality, increase coverage)

– Spatial multiplexing gain (transmit multiple independent data streams, increase link capacity)

– Array gain (capture more received energy, improve average SNR, increase coverage)

– Co-channel interference reduction gain (attenuate interference from adjacent cells, increase cellular capacity)

Dichotomy MAC - PHY

- Physical Layer

  QoS is synonymous to an acceptable SNR level, minimum rate or bit error rate (BER) at the receiver

- MAC Layer

  QoS is usually expressed in terms of maximum delay guarantees or delay jitter for a certain minimum rate

→ Need for a cross-layer approach

Cross-layer Design

- Applications
- Network
- Link
- MAC
- Physical

Delay Constraints
Rate Constraints
Energy Constraints

Adaptation across design layers
Efficient scheduling for capacity, fairness and QoS satisfaction
Provide robustness via diversity

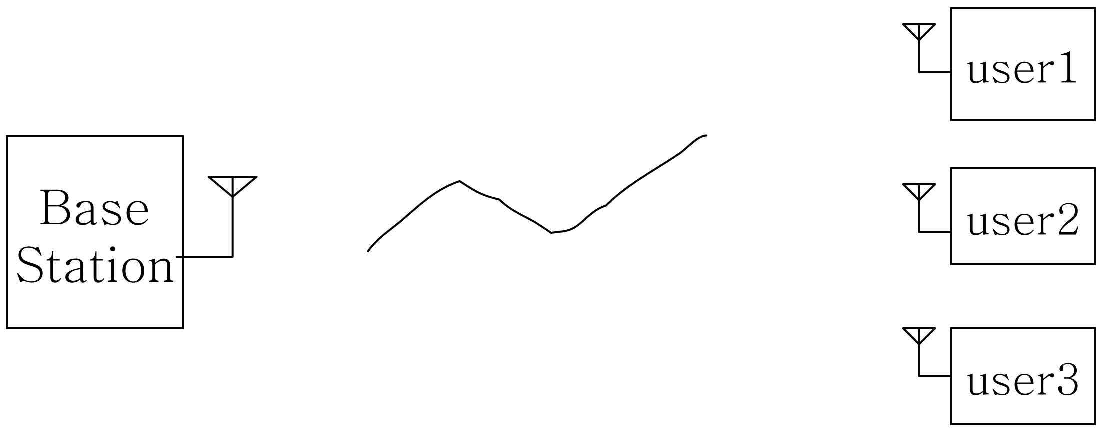
Cross-layer Techniques

- Adaptive techniques
  - Link, MAC, network, and application adaptation
  - Resource management and allocation (power control)
  - Synergy with diversity and scheduling
- Diversity techniques
  - Link diversity (antennas, channels, etc.)
  - Access diversity
  - Route diversity
  - Application diversity
  - Content location/server diversity
- Scheduling
  - Application scheduling/data prioritization
  - Resource reservation
  - Access scheduling

Scheduling for MIMO Broadcast Channel

- Capacity of MIMO Broadcast Channel
- Role of Channel Knowledge

If BS has one antenna

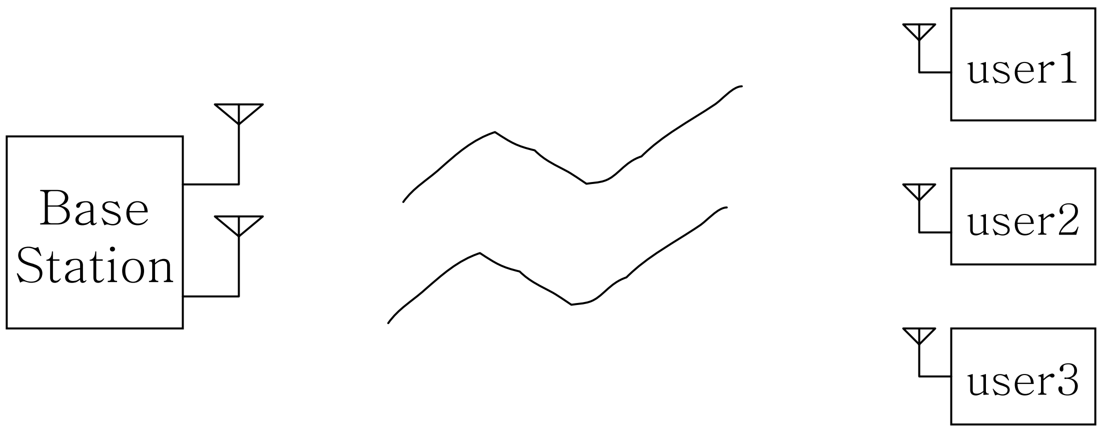


Both the base station and the receiver has one antenna

→ Degraded broadcast channel

Selecting the best user and send data only to that user is optimal in the sense of maximizing the sum rate

If BS has more than one antenna?



→ Non-degraded Broadcast channel

Selecting only one best user is not anymore optimal

Intuition: The BS can support two users simultaneously (STMA)

MIMO BC with full CSI

- Sum Capacity with full CSI

$$Csum = \max_{Q_k \geq 0, \sum_k Tr(Q_k) \leq P_T} \log_2 \det(I + \sum_k H_k^+ Q_k H_k)$$

$$H_k : \text{channel matrix user k}, Q_k : \text{covariance matrix}$$

$$C_{sum} \sim M_T loglogK \text{ (for } M_R\text{=1)}$$

- Achieved by Superposition Coding
  - Dirty Paper Coding – QR decomposition [Caire, Shamai ’00]
  - Lattice Strategies [Erez, Shamai, Zamir ’00], [Windpassinger et al. ‘04]
  - Trellis Precoding [Yu, Cioffi ’01]
  - Vector Pertubation [Peel, Hochwald, Swindlehurst ’03]
  - “Greedy” ZF beamforming [Tu, Blum ’03]
  - ZF and semi-orthogonal user selection [Yoo, Goldsmith ’05]
  - Costa Precoding [Airy, Forenza, Heath ’04]

MIMO BC with no CSIT

- The Gaussian MIMO BC is *degraded*

Sum Capacity is limited by the single-user capacity
(superposition coding is equivalent to time-sharing)

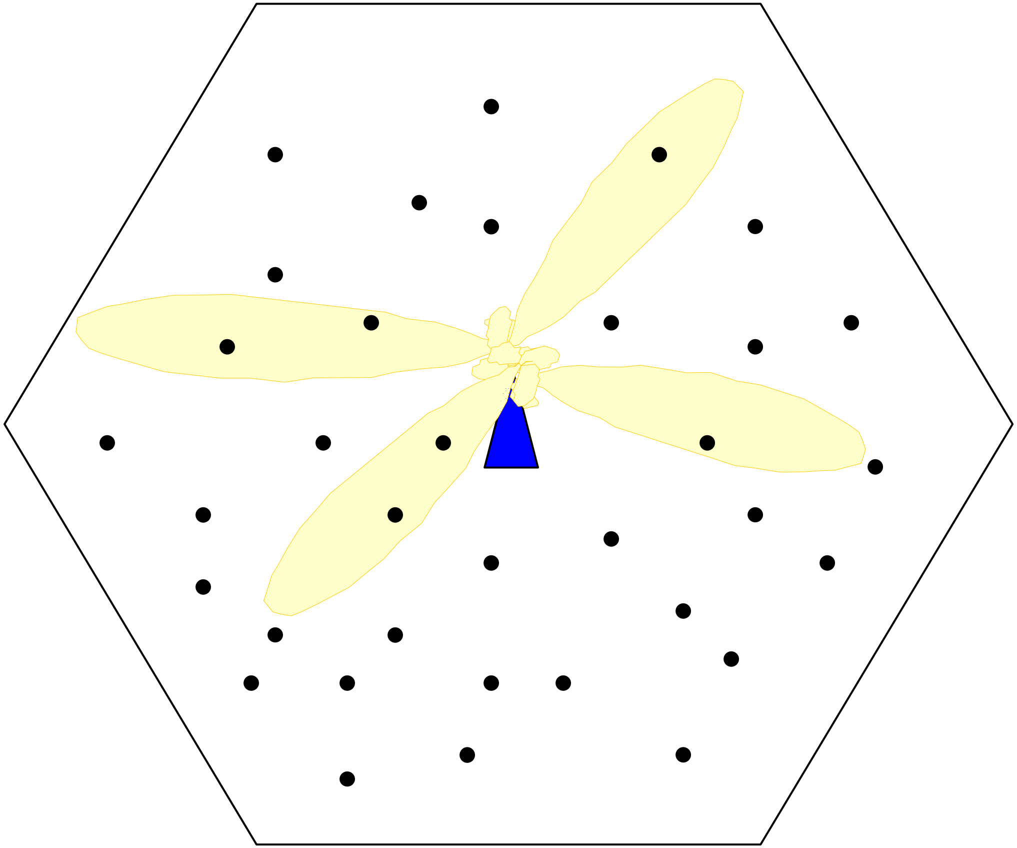
$C \sim log M_T + O(1)$ (M_R=1 and fixed P_T per antenna)

Sum Capacity is independent of the number of users
(no multiuser diversity)

→ Huge gap to be filled between no CSIT and full CSIT

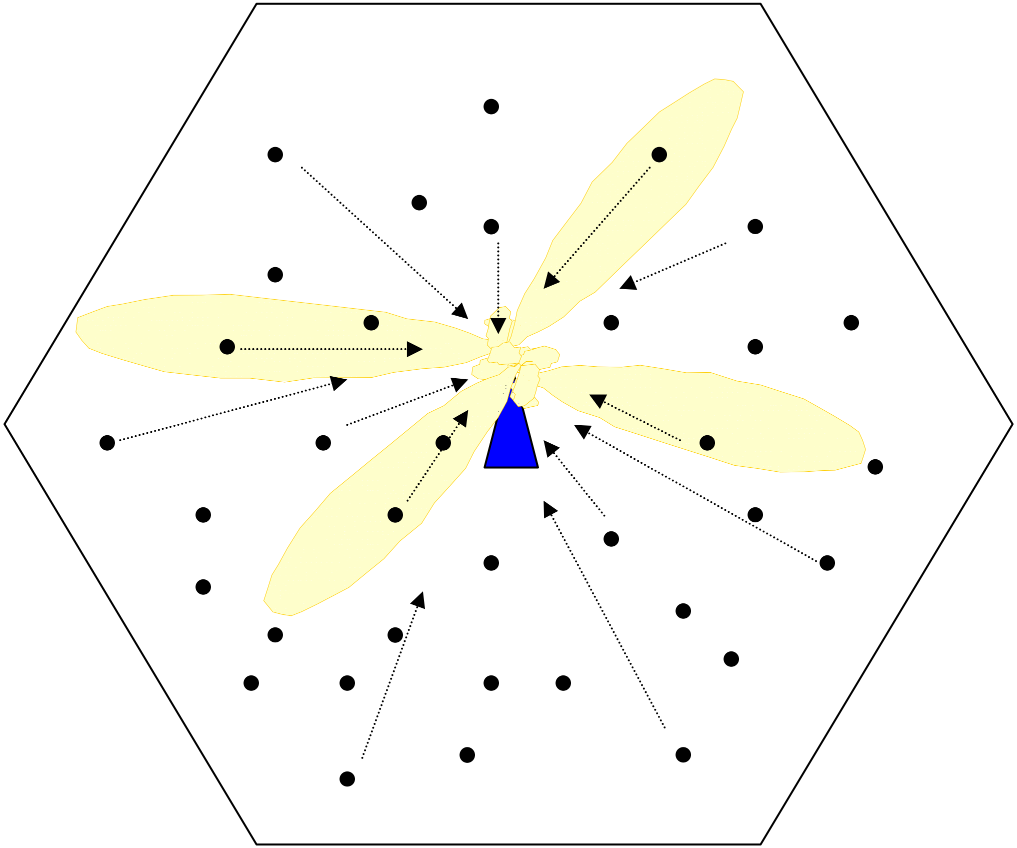
MIMO BC with partial CSIT

Send Random beams

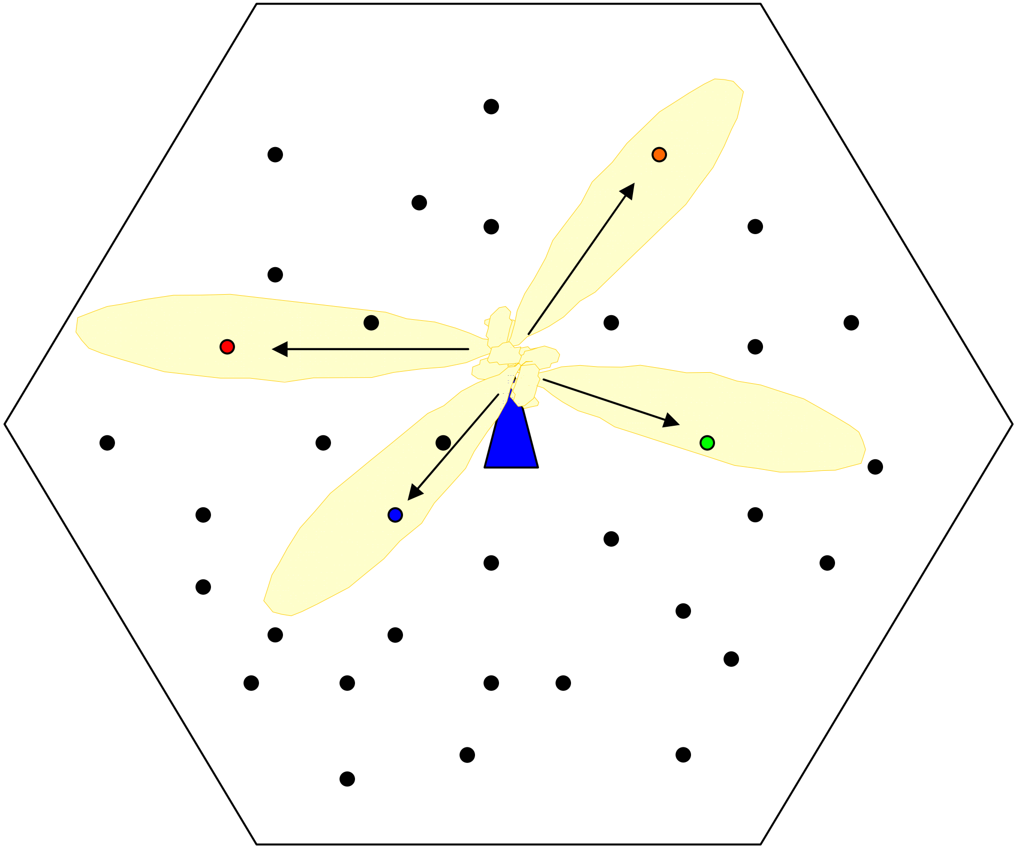


MIMO BC with partial CSIT

Users feeds back their maxSINR



MIMO BC with partial CSIT



BS assigns each beam to the user with the maxSINR for this beam

MIMO BC with partial CSIT

[Sharif & Hassibi '03] – Random Beamforming

- Send M random beams (orthonormal vectors $Q_m, m=1,\ldots,M$)
- Each i^{th} receiver can compute M SINRs: $$SINR_{i,m} = \frac{\left|H_i Q_m\right|^2}{1/SNR + \sum_{k \neq m} \left|H_i Q_k\right|^2}$$
- Each user feeds back its best SINR obtained on each particular beam & the transmitter assigns the beams to the user with the best SINR for that beam
- The Sum Rate is calculated as: $$SR = E\left\{\sum_{m=1}^{M} \log_2 (1 + \max_{1 \le i \le K} SINR_{i,m})\right\}$$
- Asymptotically equivalent to optimal BF (same scaling as having full CSI using DPC)

$$C^{RBF} \sim M_T \log \log K \quad (\text{provided } M_T \sim O(\log K))$$

PART I

Memory-based Opportunistic Multi-user Beamforming

M. Kountouris, D. Gesbert, "Memory-based Opportunistic Multi-user Beamforming " ***submitted to IEEE ISIT 2005***

Sharif & Hassibi's scheme evaluation

- Pros
  - Scaling as $M_T \log \log K$ (full CSI using DPC)
  - Low complexity
  - Low amount of Feedback
    (1 scalar value [maxSINR] & an index per user)

- Cons
  - It needs a high number of users to perform well
  - If the number of users is small, unitary matrices are not the best choice
  - No exploitation of Doppler spread

Proposed scheme

Motivation

Assuming the channel has memory $M = T_{coh} / T_{slot}$

Can we exploit past feedback knowledge to enhance the performance in slow time-varying channels?

Our Approach

Keep the “best” past random beamforming matrix in a set of ‘preferred’ beamforming matrices, compare it with a new randomly generated beam pattern and apply only the one that maximizes the Sum Rate

Algorithm Description

<u>1st Phase</u> (*'best' beamforming matrix selection*)

- Initialize Set with random beamforming matrices Q_i, each one with sum rate $SR(Q_i)$

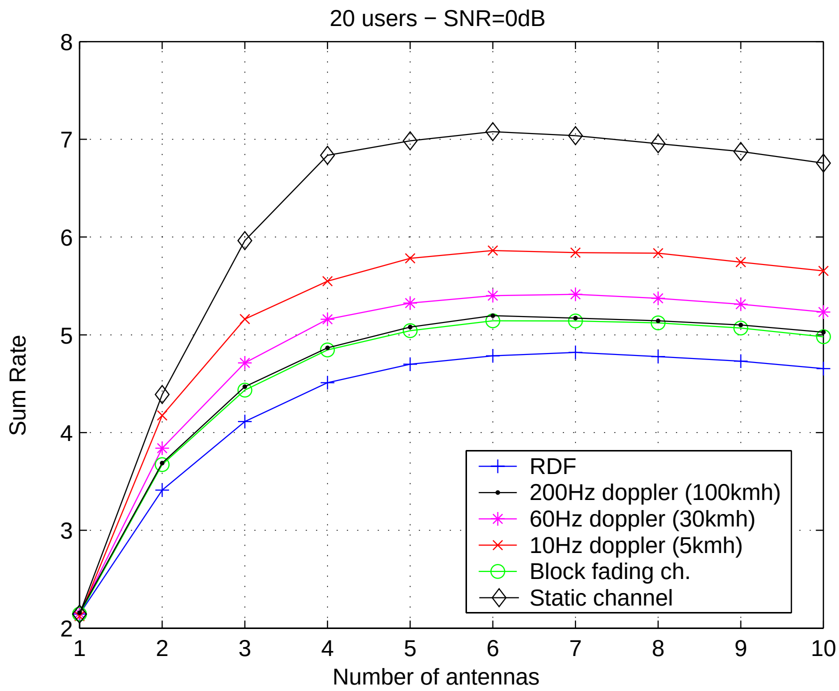
At each slot:

- Generate a new random beam pattern Q_{rand} with sum rate given by $SR(Q_{rand})$
- Choose $i^*=$*argmax*$\{SR(Q_i)\}$, apply and calculate $SR(Q_{i^*})$ given new channel
- If ($SR(Q_{i^*}) > SR(Q_{rand})$) use Q_{i^*} else use Q_{rand}

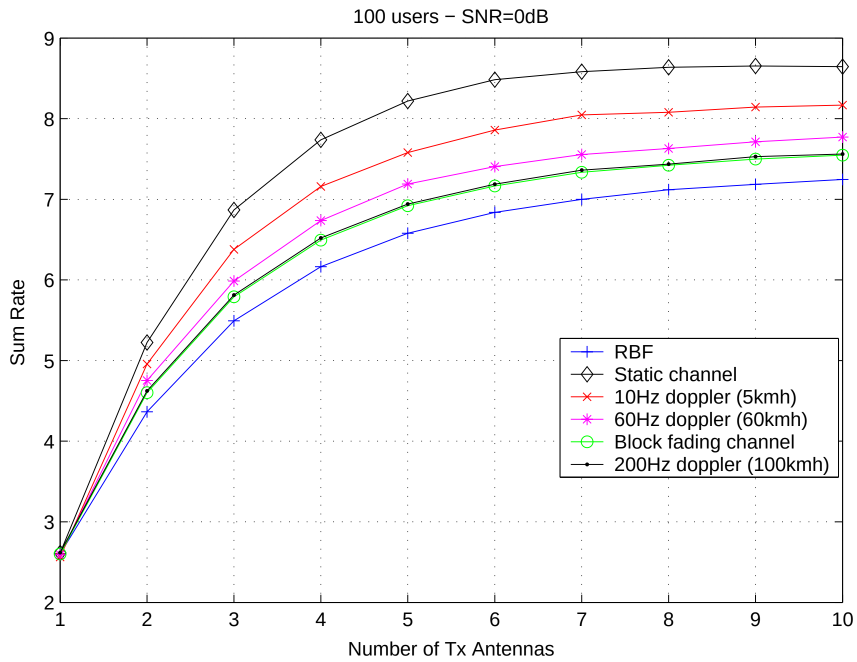
<u>2nd Phase</u> (*update of the set*)

- Update $SR(Q_{i^*})$ value of the set
- If ($SR_{rand} > \min(SR(Q_i))$) replace Q_{imin} (giving minimum sum rate) by Q_{rand}

Some preliminary results

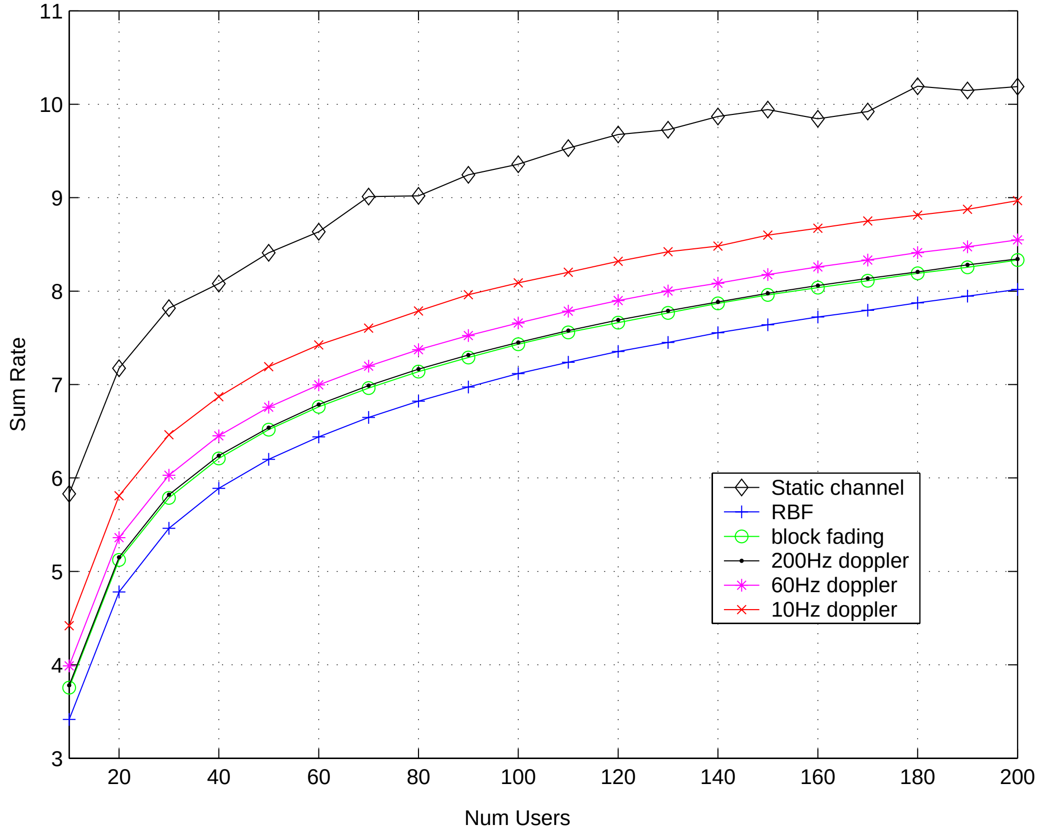


Sum Rate versus number of antennas for 20 users under different channel coherence time



Sum Rate versus number of antennas for 100 users under different channel coherence time

Some more results



Sum Rate as a function of number of users

Extension of PFS for SDMA scheduling

Let

S be the set of all active users

I with $|I| \leq N_t$ the set of simultaneously scheduled users

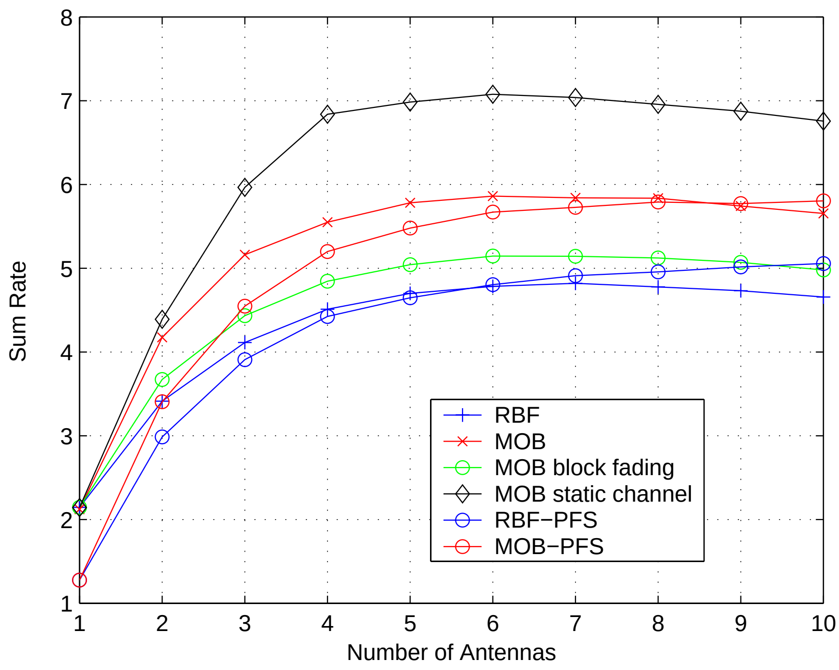
Under the proportional fair algorithm, in time slot t, the scheduling algorithm that guarantees the maximization of the sum of the logarithm of the long-term average throughput is defined by the following metric:

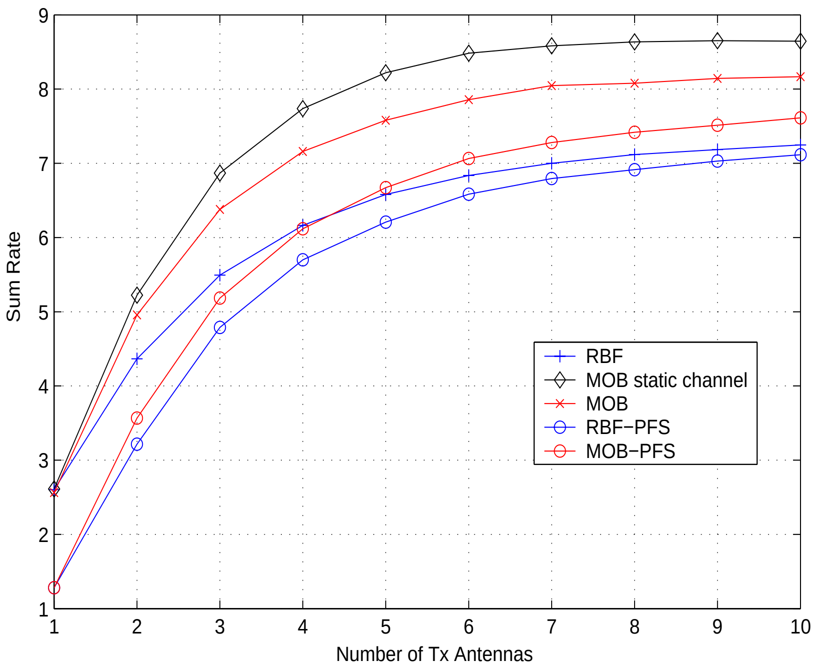
$$I = arg \max_{I \subset S} \sum_{k \subset I} \frac{R_{k|I}(t)}{T_{k|I}(t)}$$

The average throughput is updated as follows:

$$T_{ki}(t+1) = \begin{cases} (1-\frac{1}{t_c})T_{ki}(t)+\frac{1}{t_c}R_{ki|I}(t), ki \in I \\ (1-\frac{1}{t_c})T_{ki}(t), ki \notin I \end{cases}$$

Simulation Results





Comparison of performance of PFS in terms of Sum Rate versus number of antennas for 20 users for 10Hz

Comparison of performance of PFS in terms of Sum Rate versus number of antennas for 100 users for 10Hz

PART II

QoS-based User Scheduling for Multi-user MIMO Systems

M. Kountouris, A. Pandharipande, H. Kim, D. Gesbert, "QoS-based User Scheduling for Multiuser MIMO Systems " in *Proc. IEEE VTC Spring 2005, Stockholm, 30 May – 1 June,* 2005

QoS-based scheduling scheme

Motivation

- Users have different requirements in rate and delays
  - Users with different applications (e.g. voice calls, internet browsing, video streaming) – experiencing different channel conditions

  → Demand for different amount of data rate / different delay tolerance

- QoS defined in terms of throughput, error rate, delay

Goal

To identify scheduling schemes that provide a good *trade-off* between:

(i) higher system capacity by exploiting inherent multi-user diversity

(ii) fairness among users based on their instantaneous channel conditions relative to average channel conditions

(iii) tolerable latency requirements specified by the user applications

System Model

- MIMO Broadcast Channel: downlink scheduling wherein a BS with multiple antennas serves a multi-user system
- Different users can be served by the transmitter antennas at each slot → allocation of each BS antenna to users based on certain priority functions (joint, or independent scheduling across antennas)
- Priority functions capture the user QoS demands quantified in terms of throughput and delay
- Simultaneous transmission to multiple users (transmit beamforming)
  - *The allocation can be used also for Subcarrier Allocation in OFDMA*

Scheduling rules

- Delay definition (probabilistic packet delay bound):

$$\Pr\{W_k > D_k\} \leq \delta_k$$

W_k: HOL latency, D_k: maximum tolerable delay, δ_k: maximum dropping probability

- Five scheduling policies:
  - MAX-DELAY $\quad j = \arg\max_k (W_k(t))$
  - MAX-RATE $\quad j = \arg\max_k (R_k(t))$
  - PFS $\quad j = \arg\max_k \left(\frac{R_k(t)}{T_k(t)} \right)$
  - M-LWDF $\quad j = \arg\max_k \left(\gamma_k R_k(t) W_k(t) \right)$
  - Exponential rule $\quad j = \arg\max_k \left(\gamma_k R_k(t) \exp\left(\frac{a_k W_k(t) - \overline{aW}}{1 + \sqrt{\overline{aW}}} \right) \right)$

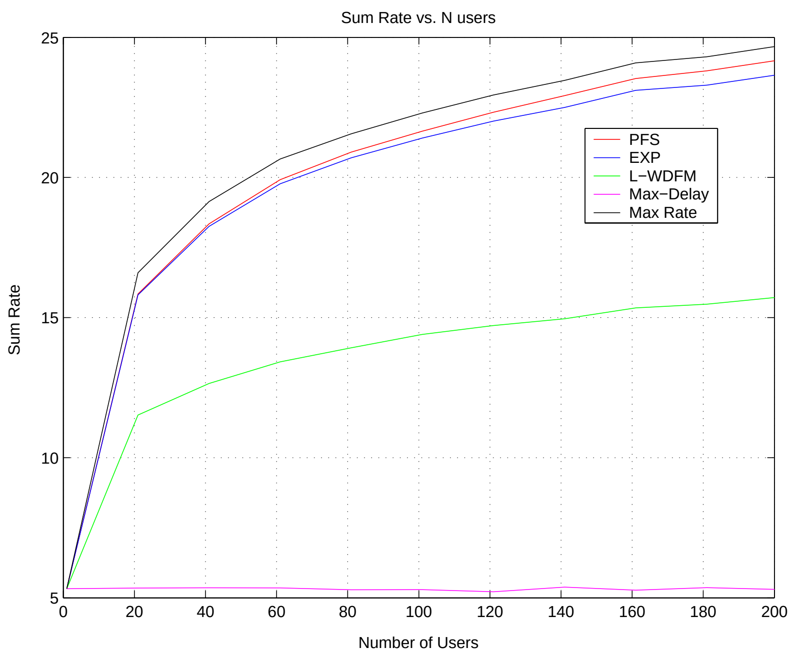
$R_k(t)$: actual rate supported by the channel of user k

T_k: mean data rate observed by user k over a long sliding window

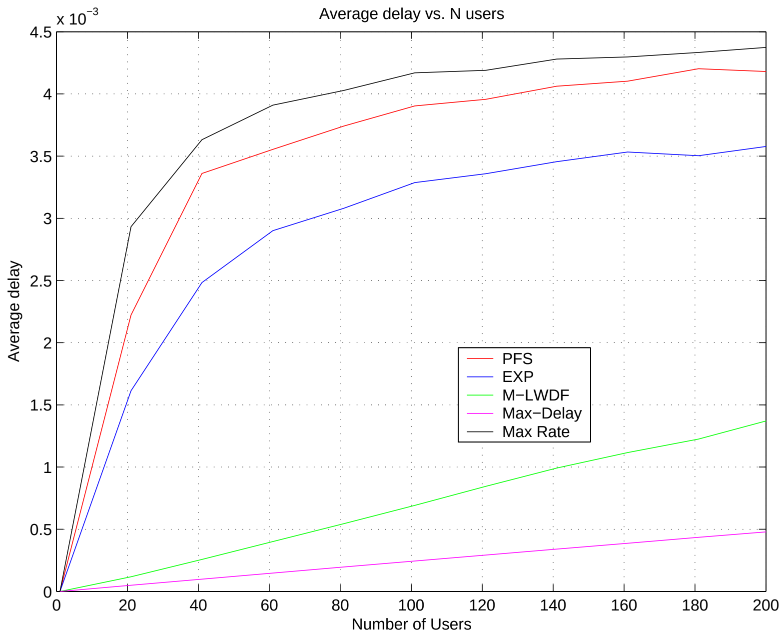
$\gamma_k = \alpha_k / T_k$, where $\alpha_k = -(\log \delta_k)/D_k$

Just to get an idea…

Single-user scenario



Sum rate vs. Num. users for SU MIMO QoS scheduling



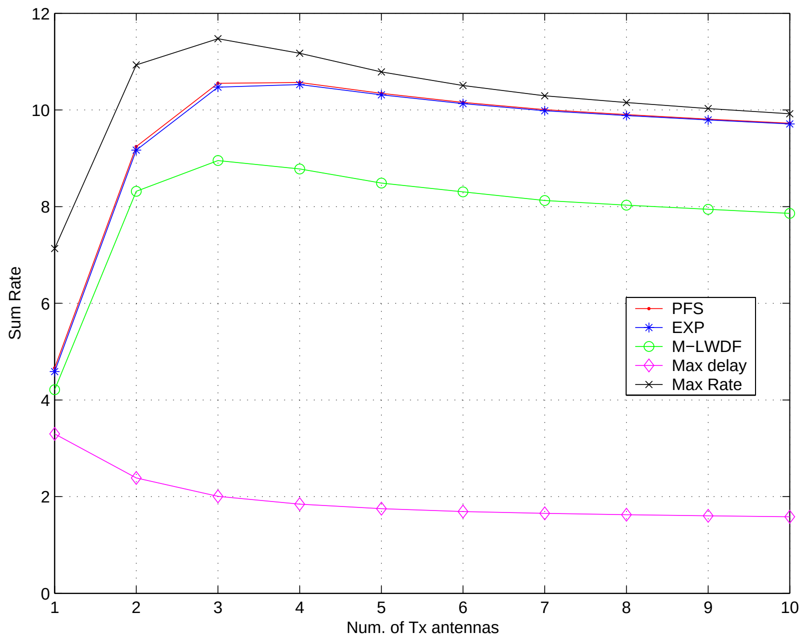
Avg.Delay vs. Num. users for SU MIMO QoS scheduling

Simulation Results for the Multi-user case

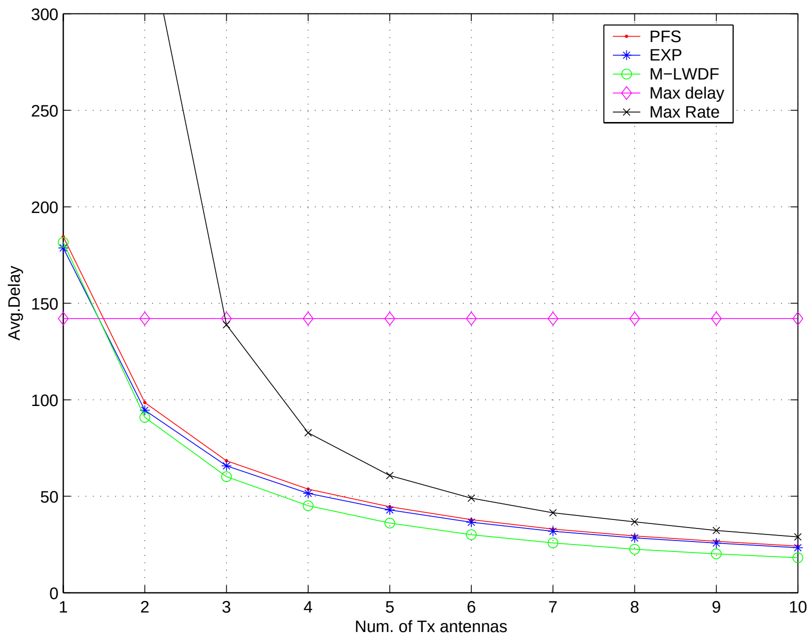
<u>Simulation Parameters</u>:

- Rayleigh fading channel – user speeds: 5kmh
- Average SNRs uniformly distributed from 3dB to 15dB
- Slot duration: *1 msec*
- PFS observation window: *tc = 500 slots*
- User QoS requirements:
  - 1st quarter: D_k = 4sec with probability $\delta_{k} =$ 0.2
  - 2nd quarter: D_k = 3sec with probability $\delta_{k} =$ 0.1
  - 3rd quarter: D_k = 2sec with probability $\delta_{k} =$ 0.1
  - 4th quarter: D_k = 1sec with probability $\delta_{k} =$ 0.05

Performance vs. Num. Antennas

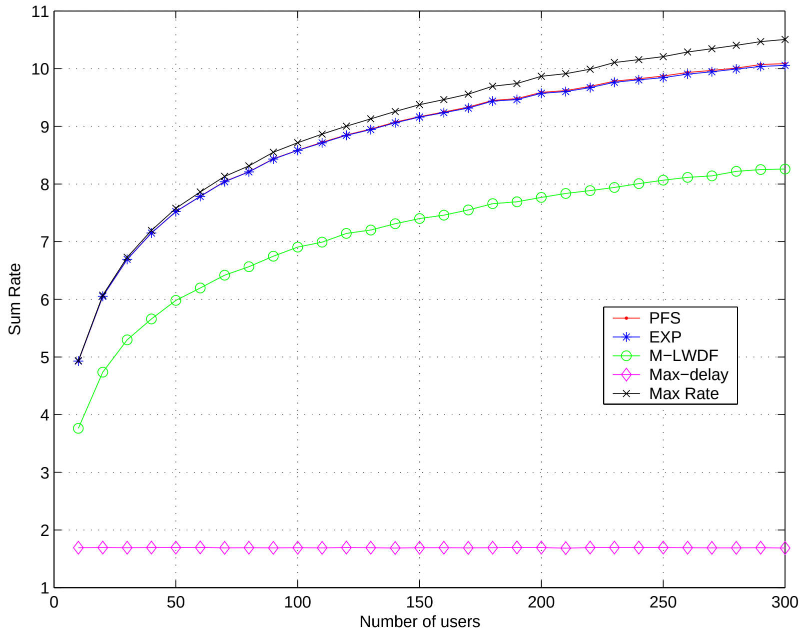


Sum rate vs. num. antennas for MU MIMO scheduling

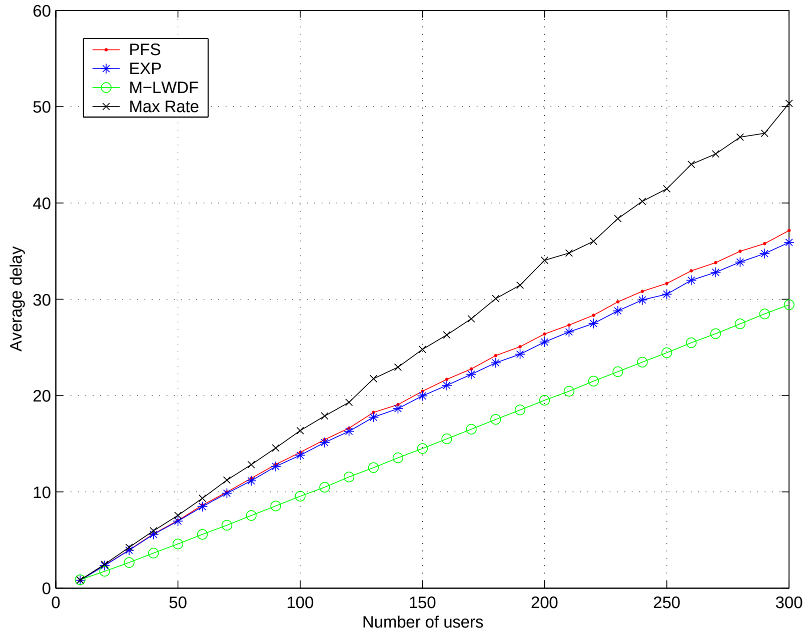


Avg. Delay vs. num. antennas for MU MIMO scheduling

Performance vs. Num. Users

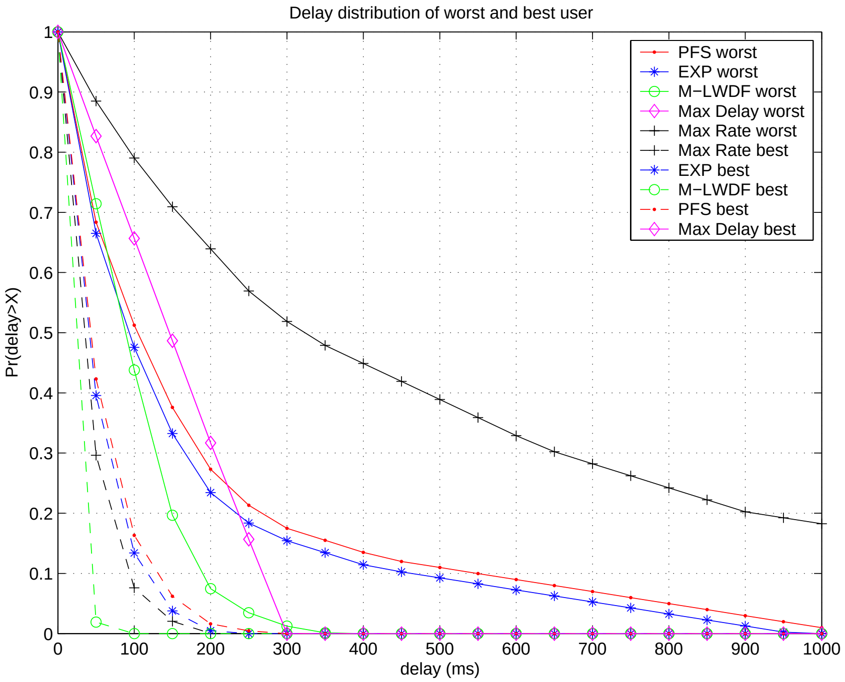


Sum rate vs. num. users for MU MIMO scheduling

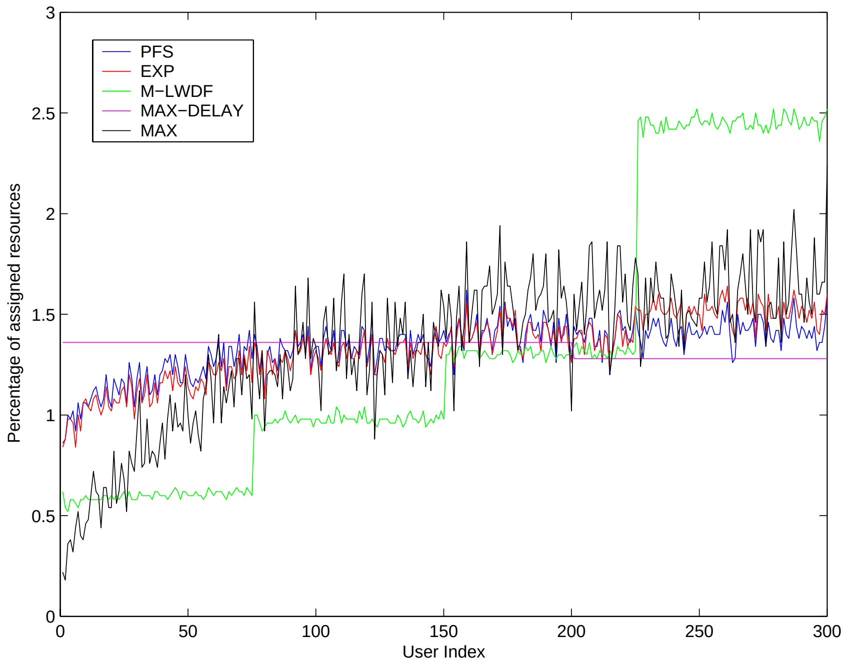


Avg. Delay vs. num. users for MU MIMO scheduling

User delay performance



Delay distribution of the strongest and the weakest users of the heterogeneous network



Frequency that each user with the corresponding SNR is chosen

Some practical issues for future work

In real-life systems, it should be assumed:

Scheduling delay

- Channel estimation - feedback
- Scheduling decision
- Notification of selected user

Outdated Channel Estimates

Questions??