Multi-channel Communications Fall 2022

Lecture 15
Multiuser MIMO

Dr. R. M. Buehrer

Introduction

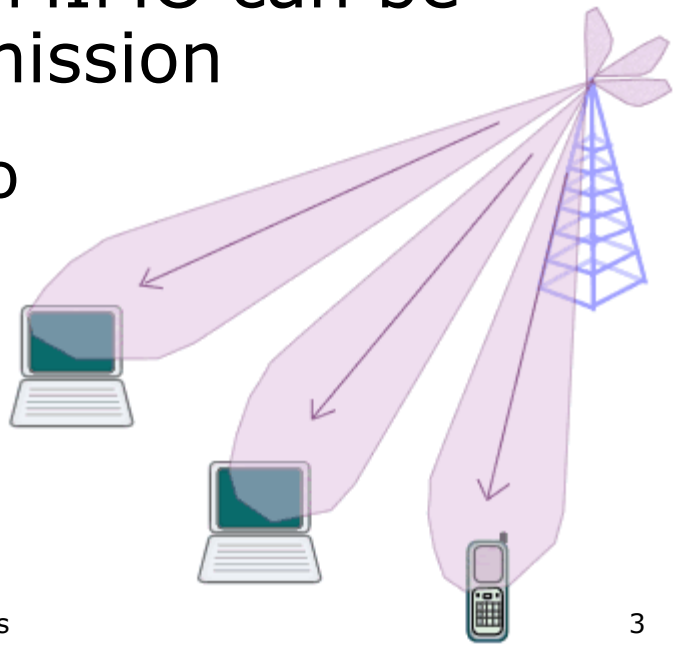
- The gains due to multiple antennas in **single user link capacity** can be extended to **multiple user** settings

- These gains are realized through
  - Signal processing,
  - Scheduling, and
  - Channel feedback (CSIT)

- Today we will explore these gains.

Motivation

- MU-MIMO is the combination of MIMO theory which tells us that multiple spatial channels exist in rich scattering environment with the concept of Space Division Multiple Access
- The degrees of freedom in MIMO can be applied to multiuser transmission
- Delay-tolerant traffic can also allow *multiuser scheduling*
- The combination of the two concepts provides a powerful tool for improving capacity and performance

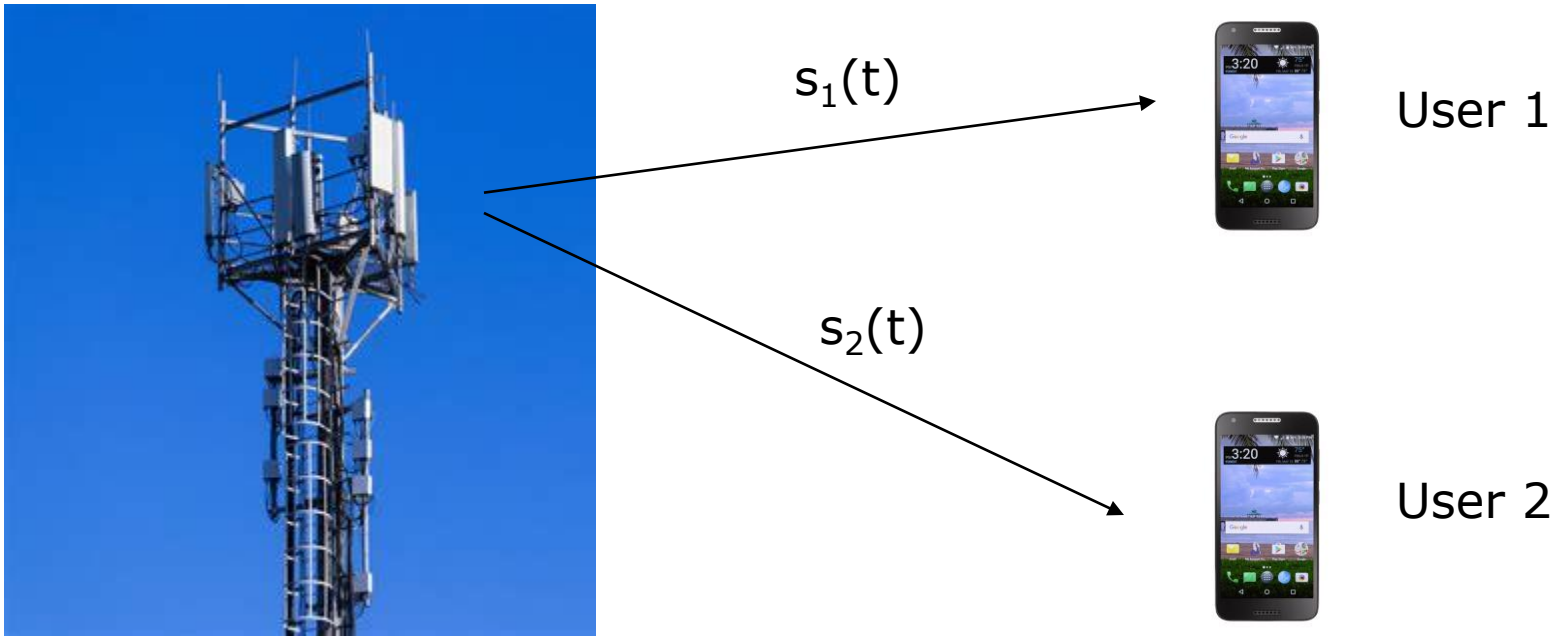


Multiuser Communications

- There are many forms of multiuser communication, but there are two specific cases of interest:
  - **Broadcast Channel (we focus on this)**
  - Multiple Access Channel
- In the case of multiple users, capacity is not a specific value, but a *region* which defines the set of possible rates.
- We will focus on the two-user case for simplicity, but concepts extend to $K > 2$

Broadcast Channel - SISO

- Broadcast Channel is from one transmitter to many receivers (e.g., base station to mobile users)

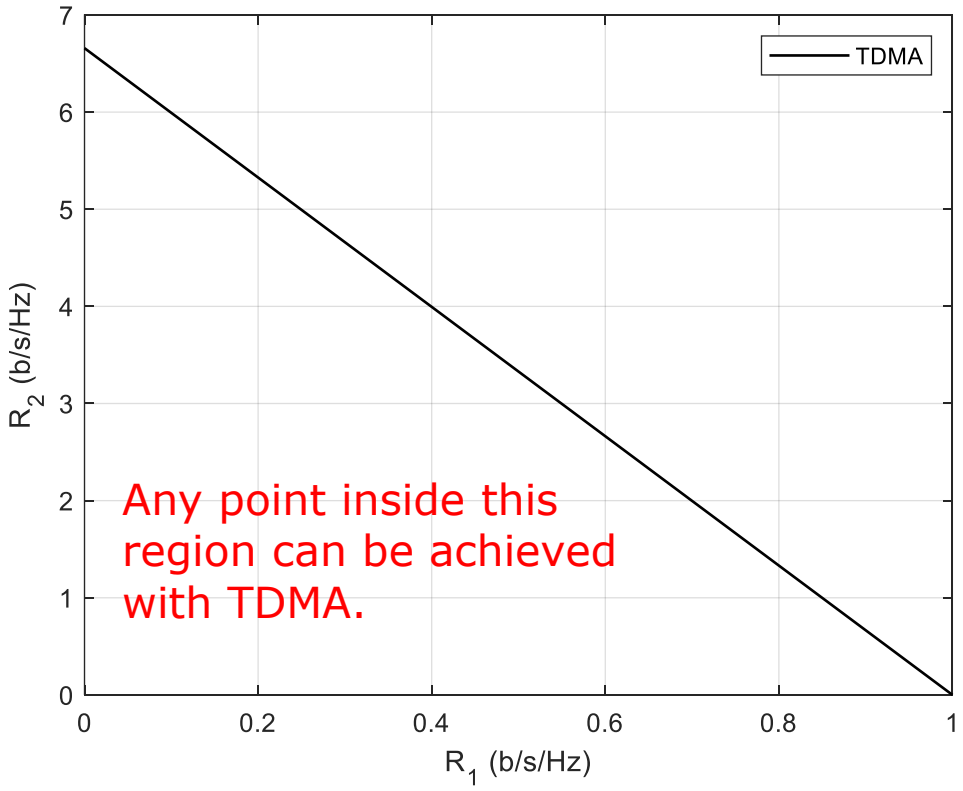


TDMA

- The simplest method of multiple access is to transmit over the entire bandwidth using all of the available power to one user at a time
- We transmit to user 1 a fraction τ of the time and transmit to user 2 a fraction (1-τ) of the time where $0 \leq \tau \leq 1$
- The capacity of this scheme is

$$C_{TD} = \bigcup_{\{\tau:0\leq\tau\leq1\}} \left\{ \frac{R_1}{B} = \tau \log_2\left(1+\mathrm{SNR}_1\right), \frac{R_1}{B} = (1-\tau)\log_2\left(1+\mathrm{SNR}_2\right) \right\}$$

Multiuser Capacity - TDMA



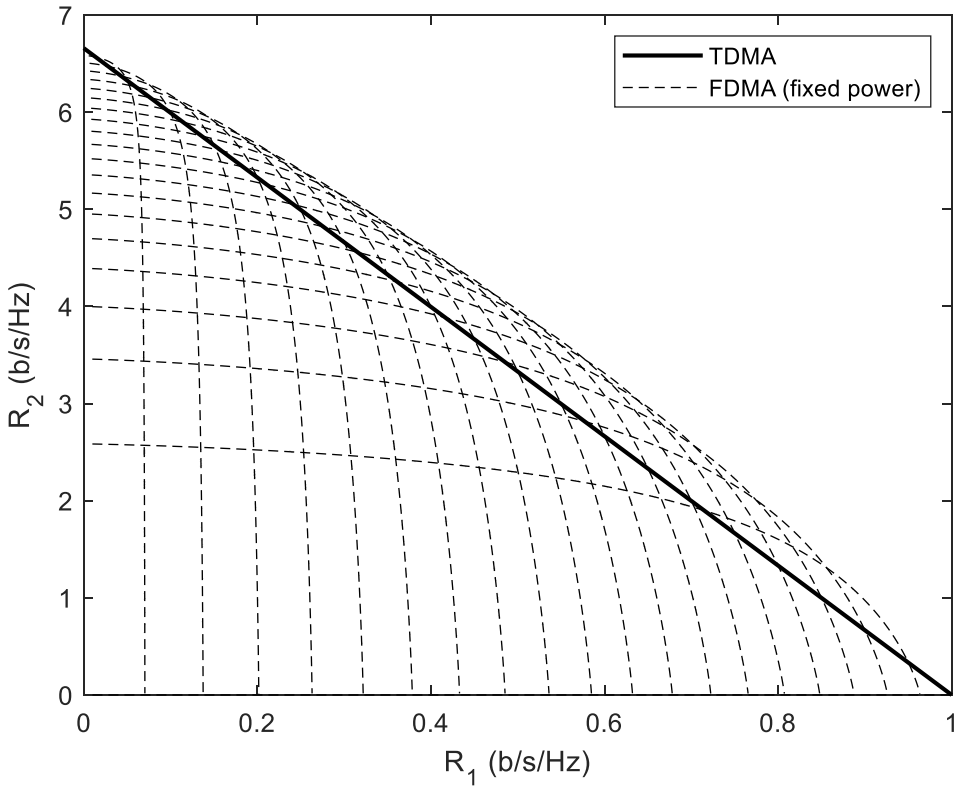
- $SNR_1 = 1$
- $SNR_2 = 100$

FDMA

- Another orthogonal method of multiple access is to transmit over a fraction of the total bandwidth using a fraction of the available power to each user
- We transmit to user 1 with βB Hz and εP Watts to user 2 with (1- β)B Hz and (1-ε)P Watts

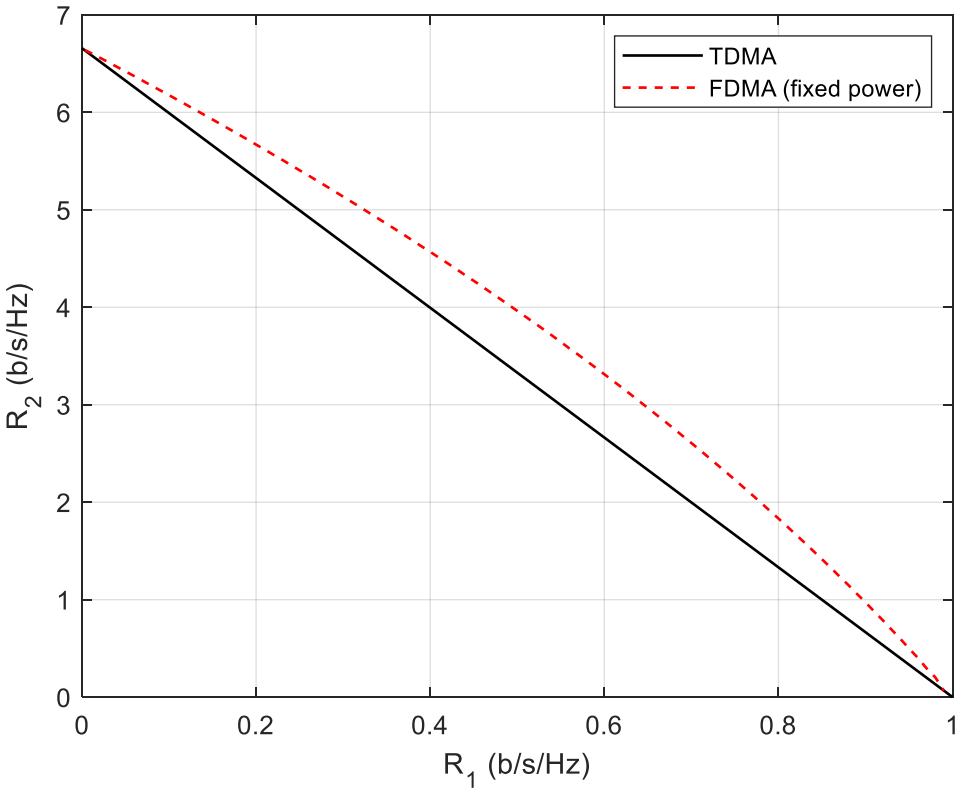
$$C_{FD} = \bigcup_{\{\beta,\eta:0\leq\beta,\epsilon\leq 1\}} \left\{ \frac{R_1}{B} = \beta \log_2\left(1+\frac{\epsilon}{\beta}\mathrm{SNR}_1\right), \frac{R_1}{B} = (1-\beta)\log_2\left(1+\frac{1-\epsilon}{1-\beta}\mathrm{SNR}_2\right) \right\}$$

Multiuser Capacity - FDMA



- $SNR_1 = 1$
- $SNR_2 = 100$
- ε, β variable
- Each dashed curve represents a different power allocation as bandwidth allocation is varied

Multiuser Capacity - FDMA



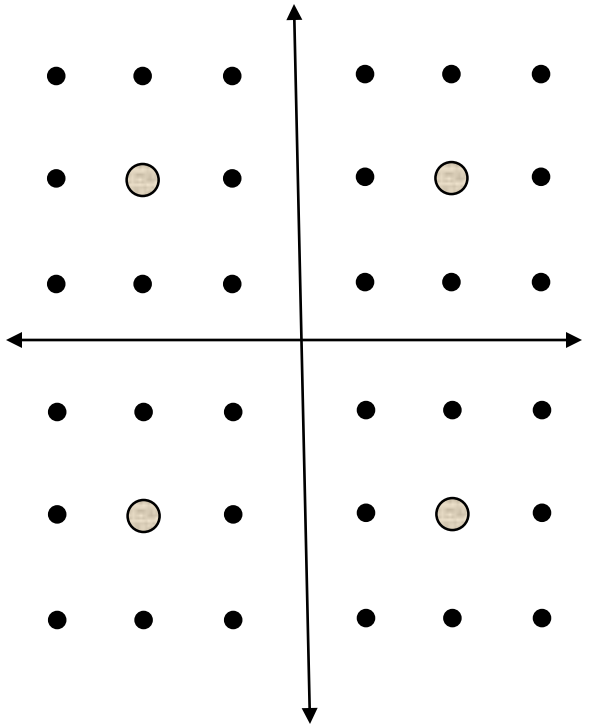
- $SNR_1 = 1$
- $SNR_2 = 100$
- For each desired value of R_1 we choose a bandwidth & power allocation to maximize R_2
- FDMA can obtain larger capacity region than TDMA since we can adjust the bandwidth and power based on the SNR
- If the two SNRs are equal, FDMA = TDMA

Multiuser Capacity - Optimal

- The optimal broadcast method is to allow users to overlap in time and frequency
- Users are jointly encoded such that the user with the higher SNR can decode and subtract out the signal for the other user.
- The signal for the higher SNR user appears as noise to the lower SNR user

Simple example

- 32-QAM with embedded 4-PSK
- Lower SNR user sends two bits which determine quadrant
- Higher SNR user sends three bits which chooses one of eight constellation points within the quadrant
- Higher SNR user easily detects quadrant and removes it before detecting the resulting 8-QAM signal
- 8-QAM signal appears as noise to the lower power user



Optimal Broadcast Capacity

- Using superposition encoding with SIC the capacity can be shown to be

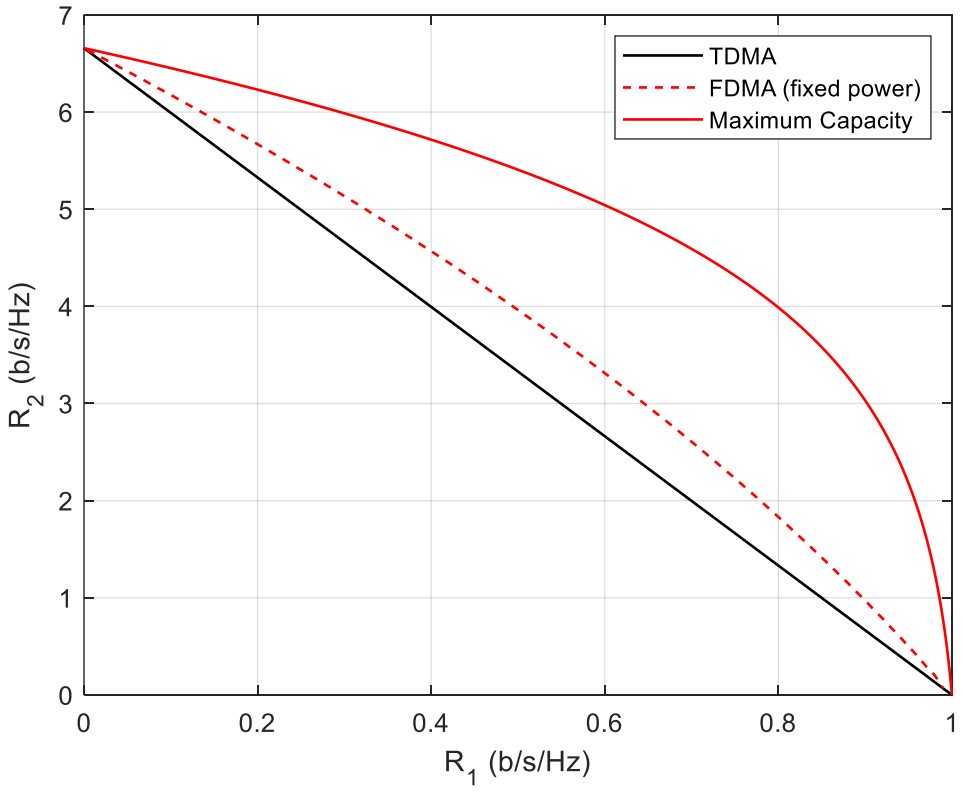
$$C_{BC} = \bigcup_{\{P_1, P_2 : P_1 + P_2 = P\}} \left\{ \frac{R_1}{B} = \log_2\left(1 + \frac{P_1 g_1}{N_o B}\right), \frac{R_2}{B} = \log_2\left(1 + \frac{P_2 g_2}{N_o B + P_1 g_2}\right) \right\}$$

where g_1 and g_2 are different gains with $g_1 > g_2$, N_o = noise PSD and $P_1 + P_2 = P$ (tx power).

- Defining $\mathrm{SNR}_i = \dfrac{g_i P}{N_o B}$ and $P_1 = \varepsilon P$ we have

$$C_{BC} = \bigcup_{\{\epsilon : 0 \le \epsilon \le 1\}} \left\{ \frac{R_1}{B} = \log_2\left(1 + \epsilon \mathrm{SNR}_1\right), \frac{R_2}{B} = \log_2\left(1 + \left(\frac{1}{(1-\epsilon)\mathrm{SNR}_2} + \frac{\epsilon}{1-\epsilon}\right)^{-1}\right) \right\}$$

Multiuser Capacity - Optimal



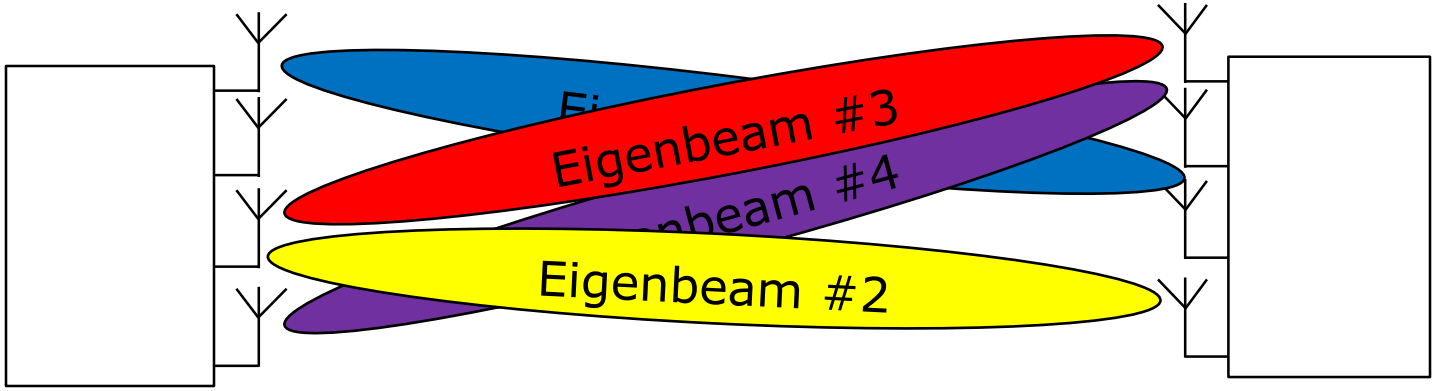
- $SNR_1 = 1$
- $SNR_2 = 100$
- For each desired value of R_1 we choose a power allocation to maximize R_2
- Superposition encoding with SIC can provide substantial gains
- Technique can be expanded to *K* users

Multiuser MIMO

- With MIMO, like in the SISO case, we can use TDMA or FDMA, but non-orthogonal sharing is optimal

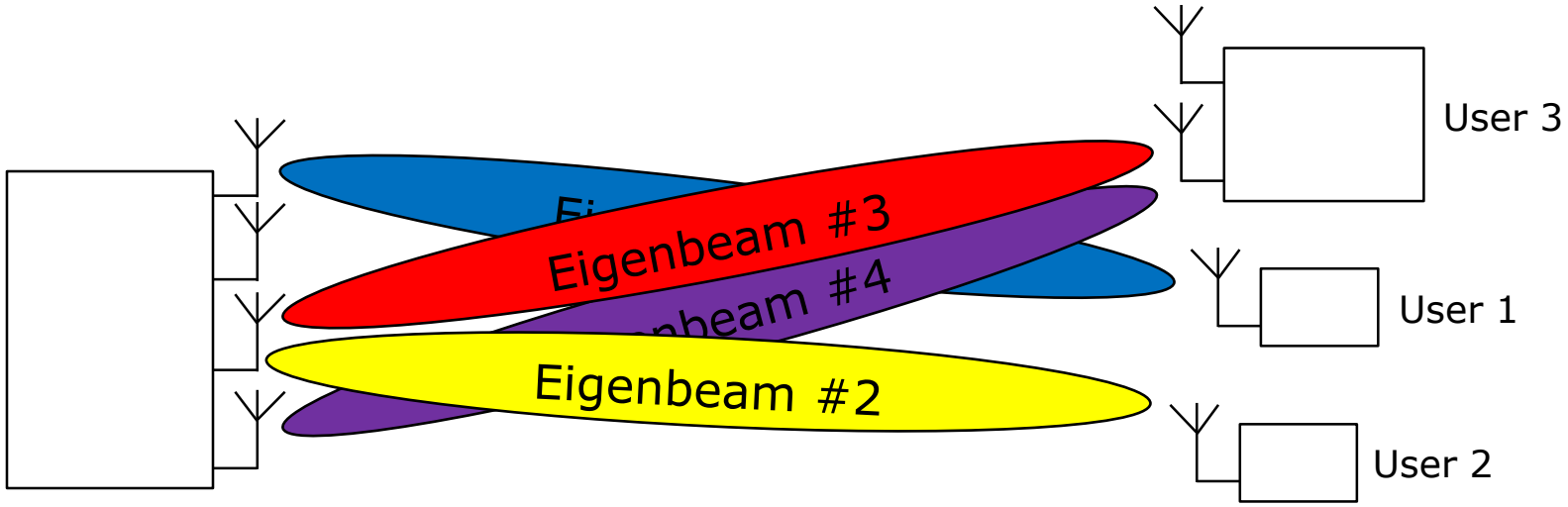
- Thus, with multi-user MIMO, we seek to use the spatial modes available in the channel to transmit to multiple users simultaneously

Single User MIMO



- With CSIT $\min(N_t, M_r)$ eigenbeams can be formed to transmit multiple spatial streams between the base station and a mobile

Multiuser MIMO



- $\min\left(N_t, \sum_k M_r^{(k)}\right)$ eigenbeams can be formed to transmit multiple spatial streams between the base station and multiple mobiles

MU-MIMO vs. SU-MIMO

- MU-MIMO
  - allows for **direct gain in multiple access capacity** (proportional to the number of base station antennas) due to multiuser multiplexing
  - **is more immune to propagation limitations** which impact SU-MIMO (i.e., low-rank channel loss or antenna correlation) due to decorrelated antennas across users and multiuser diversity gain
    - Even LOS propagation can be used with MU-MIMO!
  - **allows for spatial multiplexing gain** at the base station **without the need for multiple antennas at the user equipment**

Costs of MU-MIMO

- Two primary costs associated with MU-MIMO
  - Requirement of **channel state information at the transmitter** (CSIT)
    - Unlike SU-MIMO which can perform reasonably well without CSIT
    - Results in loss of reverse link capacity due to feedback
  - **Complexity** associated with scheduling algorithms
    - If K users connected to N_t-antenna base station we must examine all possible N_t-user selections and all possible orders those users

Uplink vs. Downlink

- MU-MIMO can be applied to either uplink or downlink of a wireless system
- Uplink (user equipment to base station or access point) is known as the ***Multiple Access Channel***
  - MU-MIMO techniques are generalizations of multiuser detection methods primarily studied in the context of CDMA
- Downlink (base station to user equipment) is known as the ***Broadcast Channel*** and is the more challenging link theoretically but also more interesting
  - Many applications have asymmetric traffic requirements with downlink being more demanding
  - More complexity allowed at base station
  - We will focus on this link

Key Pieces of MU-MIMO Downlink

- Precoding
  - Signal processing at the base station transmitter to mitigate interference between streams to each user (analogous to SVD-based eigenbeamforming in SU-MIMO)
  - Dirty Paper Coding is optimal
  - Linear and Non-linear sub-optimal approaches have been examined
- Feedback
  - CSIT needs to be intelligently fed back to the transmitter since perfect CSIT is impossible
- Scheduling
  - Gains depend on which users are selected for transmission at any given time and which order they are placed in the precoding process

Precoding

- Optimal – Dirty Paper Coding (DPC)
- Sub-optimal Linear
  - Linear MMSE
  - Zero-Forcing
- Sub-optimal Non-linear
  - Vector Perturbation
  - DPC-based Techniques
  - Tomlinson-Harishima Precoding

Feedback

- Many ways to feed back the channel state information
- Trades downlink performance for uplink capacity
- Common techniques
  - Vector Quantization
  - Adaptive feedback
  - Statistical feedback

Scheduling

- Optimal scheduling requires a search over all *K* users connected to the base station for the *best ordered set of N_u users*
  - N_u need not be equal to N_t → $1 \leq N_u \leq N_t^2$
  - Optimal search can be computationally impractical
- There are also many sub-optimal techniques to schedule users that provide acceptable trade between complexity and performance
  - Max-rate techniques
  - Greedy user selection
  - Random user selection
  - Proportionally Fair scheduling

Signal Models

- Uplink – Received base station vector (M_r x 1) from N_U UEs

$$\mathbf{y} = \sum_{k=1}^{N_U} \mathbf{H}_k \mathbf{x}_k + \mathbf{n}$$

$N_t^{(k)}$ x 1 transmit vector

M_r x $N_t^{(k)}$ channel matrix

- Downlink – Received UE vector ($M_r^{(k)}$ x 1) from base station

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x} + \mathbf{n}_k$$

N_t x 1 transmit vector $\quad \mathbf{x} = \sum_k \mathbf{x}_k$

$M_r^{(k)}$ x N_t channel matrix

Information Theory Perspective

- In multiuser systems capacity is not a single number – it is instead characterized by a *K*-dimensional rate region
  - Each point is a vector of achievable rates by all users simultaneously
- Capacity region (Gaussian BC) for unit noise power and channels $\mathbf{H}_i$:

$$C = \bigcup_{\sum_k P_k = P_T} \left\{ (R_1, R_2, \ldots R_{N_U}) \in \Re^{+N_U} R_i \leq \log_2 \frac{det\left[\mathbf{I} + \mathbf{H}_i(\Sigma_{j \geq i}\mathbf{Q}_j)\mathbf{H}_i^H\right]}{det\left[\mathbf{I} + \mathbf{H}_i(\Sigma_{j > i}\mathbf{Q_j})\mathbf{H}_i^H\right]} \right\}$$

$$\mathbf{Q}_k = E\left\{\mathbf{x}_k \mathbf{x}_k^H\right\}$$

Value of CSIT

- With CSIT and Dirty Paper Coding, the sum rate (all UEs have M_r antennas):

$$\lim_{N_U \to \infty} \frac{E\{R^{DPC}\}}{\log\log(N_U M_r)} = N_t$$

  i.e., sum rate scales with the number of transmit antennas

- Without CSI

$$E\{R^{NoCSIT}\} \approx min\{N_t, M_r\} \log(SNR)$$

Design Lessons from Info Thy

- Results advocate **serving multiple users simultaneously** in SDMA fashion (assuming suitable chosen precoding scheme at the transmitter)
- Multiplexing gain can be achieved while **receivers have single antenna** – low cost!
  - Multiple antennas at the receiver would provide capability of multiple streams per user or performance enhancements
- Gains are immune to ill-behaved single user channels
  - Full rank of global channel matrix nearly guaranteed by physical separation of users and large number of users
- Multiplexing gain of N_t comes with the condition of channel knowledge (unlike SU case)

Resource Allocation

- Resource allocation techniques help exploit gains of multiuser MIMO
  - PHY/MAC now tied together
- The primary resources considered at the transmitter (in addition to frequency channels) are spatial stream and power
  - These depend entirely on the channel conditions
- The number of simultaneous users served is upper bounded by N_t^2 but practically is limited to N_t with linear precoding.
- Scheduler depends on metrics used
  - Sum-rate, per-user rate targets, are often metrics
  - Multiuser diversity (SNR) and user separability impact these metrics

Signal Processing on the Downlink

- Uplink
  - Essentially requires multiuser detection. Not as interesting a problem and not studied as much
- Downlink
  - Optimal is Dirty Paper Coding
    - Pre-canceling interference at the transmitter using CSIT and knowledge of the signals
    - Difficult to implement in practice
- Linear precoding
  - Generalization of SDMA – users assigned different precoding matrices
- Non-linear precoding

Dirty Paper Coding

- The name dirty paper coding is attributed to Costa who showed that for a scalar discrete-time point-to-point memoryless channel,

$$y_i = x_i + s_i + n_i$$

- x_i and y_i are the transmitted and the received signals respectively, the interfering signal s_i is known to the transmitter but not to the receiver, and n_i is the unknown noise.
- If both s_i and n_i are i.i.d. Gaussian, and if the entire non-causal realization of s_i is known to the transmitter prior to transmission, the channel capacity is the same as that of the AWGN channel *i.e.,* as if the interference were not present.
  - This tells us that knowing the interference non-causally at the transmitter is as powerful as knowing it at both the transmitter and the receiver. Costa's result was generalized to a vector point-to-point memoryless channel.
  - Dirty Paper → channel with interference; Ink → power

Linear Precoding

- Zero-Forcing
  - Consider the received signal at the *k*th user:

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{W}_k \mathbf{s}_k + \mathbf{H}_k \sum_{l=1, l \neq k}^{N_u} \mathbf{W}_l \mathbf{s}_l + \mathbf{n}_k$$

  where $\mathbf{W}_k$ is the precoding matrix and $\mathbf{s}_k$ is the transmit symbol vector of the *k*th user.

- Assume $M_r^{(k)} = 1 \rightarrow \mathbf{H}_k = \mathbf{h}_k$ is 1 x N_t row vector
  - Choose $\mathbf{W}_k$ as the *k*th column of the right pseudo-inverse of the composite channel $\left[\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots \mathbf{h}_{N_u}^T, \right]^T$

ZF Linear Precoding (cont.)

- ZF can result in poor SNR for some users
  - MMSE version can improve SNR
  - Scheduling algorithm can also help by choosing/grouping users that provide good SNR for all users
- If $M_r^{(k)} > 1$, previous algorithm is not suitable
- For $M_r^{(k)} > 1$, but $\sum_{k=1}^{N_u} M_r^{(k)} = N_t$ and $S_k = M_r^{(k)}$ we can use block diagonalization → choose

$$\mathbf{H}_l \mathbf{W}_k = 0 \quad \forall l \neq k$$

Streams of user k

Block Diagonalization ZF

- Define $\tilde{\mathbf{H}}_k = \left[\mathbf{H}_1^T, \mathbf{H}_2^T, \ldots \mathbf{H}_{k-1}^T, \mathbf{H}_{k+1}^T \ldots \mathbf{H}_{N_u}^T\right]^T$
- Chose $\mathbf{W}_k$ to lie in the null space of $\tilde{\mathbf{H}}_k$
  - Since $\mathbf{H}_k$ is a $M_r^{(k)}$ x N_t matrix, $\tilde{\mathbf{H}}_k$ is a N_t-$M_r^{(k)}$ x N_t matrix.
  - Using singular value decomposition we can form
$$\tilde{\mathbf{H}}_k = \tilde{\mathbf{U}}_k \tilde{\mathbf{D}}_k \left[\mathbf{V}_k^{(1)}, \mathbf{V}_k^{(0)}\right]^H$$
  - $\mathbf{V}_k^{(1)}$ and $\mathbf{V}_k^{(0)}$ are the right singular matrices corresponding to the non-zero and zero singular values respectively.
- $\mathbf{W}_k$ can be any linear combination of columns of $\mathbf{V}_k^{(0)}$

Signal-to-Leakage-and-Noise Ratio Precoding

- The SINR for *k*th user is

$$SINR_k = \frac{||\mathbf{H}_k \mathbf{W}_k||^2}{M_r^{(k)} \sigma_n^2 + \sum_{l \neq k} ||\mathbf{H}_k \mathbf{W}_l||^2}$$

- We would like to choose $\mathbf{W}_k$ for each user to maximize this for each user
- This is a very challenging optimization problem
- Instead it has been proposed to use signal-to-leakage plus noise ratio:

$$SINR_k = \frac{||\mathbf{H}_k \mathbf{W}_k||^2}{M_r^{(k)} \sigma_n^2 + ||\tilde{\mathbf{H}}_k \mathbf{W}_k||^2}$$

- This attempts to minimize the interference caused to other users rather than caused by other users while maximizing my rx power.
- $\tilde{\mathbf{H}}_k = \left[\mathbf{H}_1^T, \mathbf{H}_2^T, \ldots \mathbf{H}_{k-1}^T, \mathbf{H}_{k+1}^T \ldots \mathbf{H}_{N_u}^T\right]^T$ is termed the "leakage channel"

Performance Example*

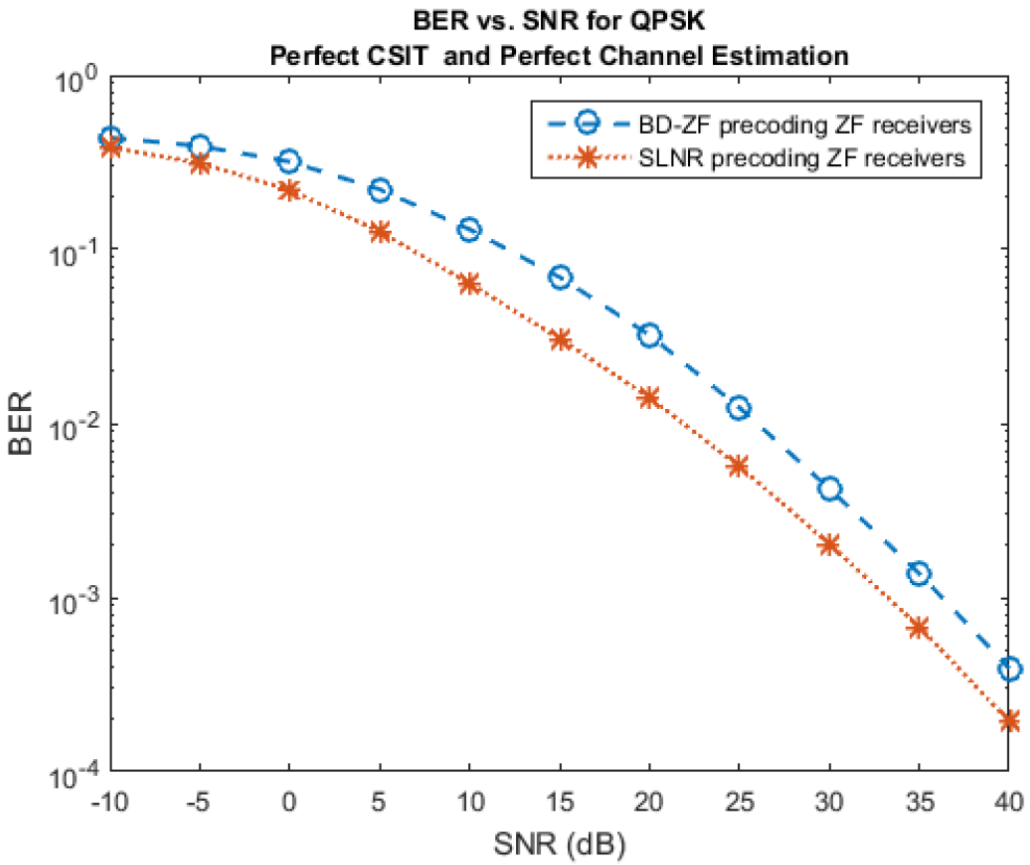


Figure 2.2: BER using BD-ZF and SLNR precoding schemes

- $N_t = 9$
- $N_u = 3$
- $M_r^{(k)} = 3$
- $S_k = 3$ (three streams per user)
- SLNR precoding provides better performance as it accounts for noise at the receiver

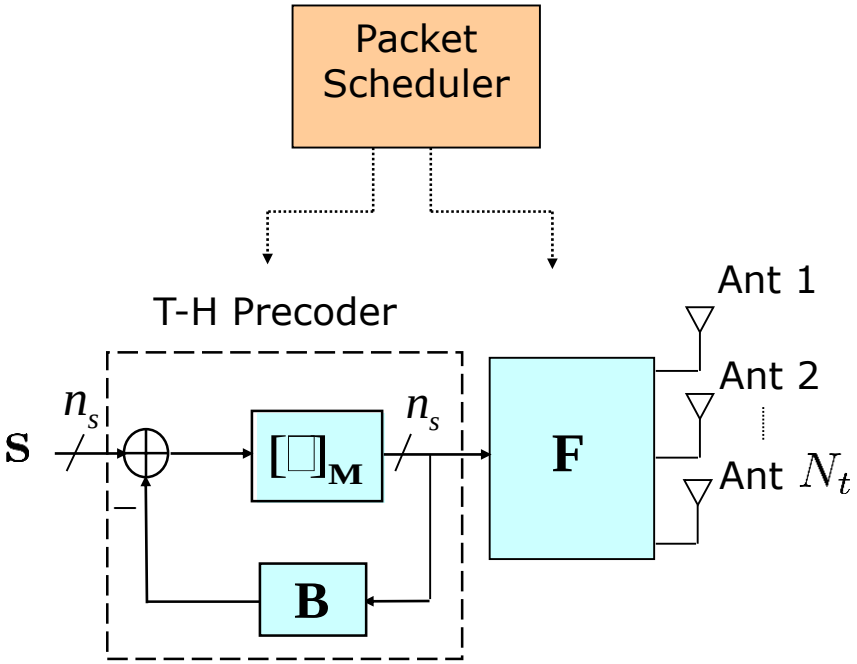
* Source: E. Sollenberger, "Iterative Leakage-Based Precoding for Multiuser-MIMO Systems," MS Thesis, Virginia Tech, 2016.

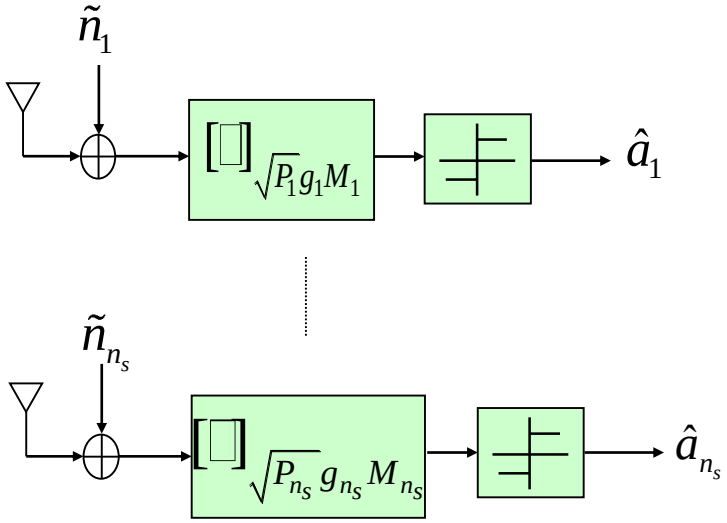
Non-linear Precoding

- Linear precoding provides reasonable performance but may remain far from DPC-line precoding strategies
- Vector Perturbation
- Tomlinson Harishima Precoding
  - Based on ISI cancellation techniques designed for landline modems
  - A transmit precoding analog to receiver successive interference cancellation
  - A symbol-by-symbol version of Costa's Dirty Paper Coding

Tomlinson-Harishima Precoding

- A spatial zero-forcing solution (ZF-THP)





$$\mathbf{A} = \mathbf{H}\mathbf{H}^H = \mathbf{L}\mathbf{L}^H$$

$$\mathbf{F} = \left(\mathbf{L}^{-1}\mathbf{H}\right)^H$$

$$\mathbf{L} = \mathbf{G}\cdot(\mathbf{B}+\mathbf{I}_{n_s}), \quad \mathbf{G} = \mathrm{diag}\{g_1,\ldots,g_{n_s}\}$$

Set: $n_s = N_t$

$$P_i = P_T / n_s, \; i = 1,\ldots,n_s$$

Precoding Performance

$(N_t, M_r^{(k)}, N_u)$

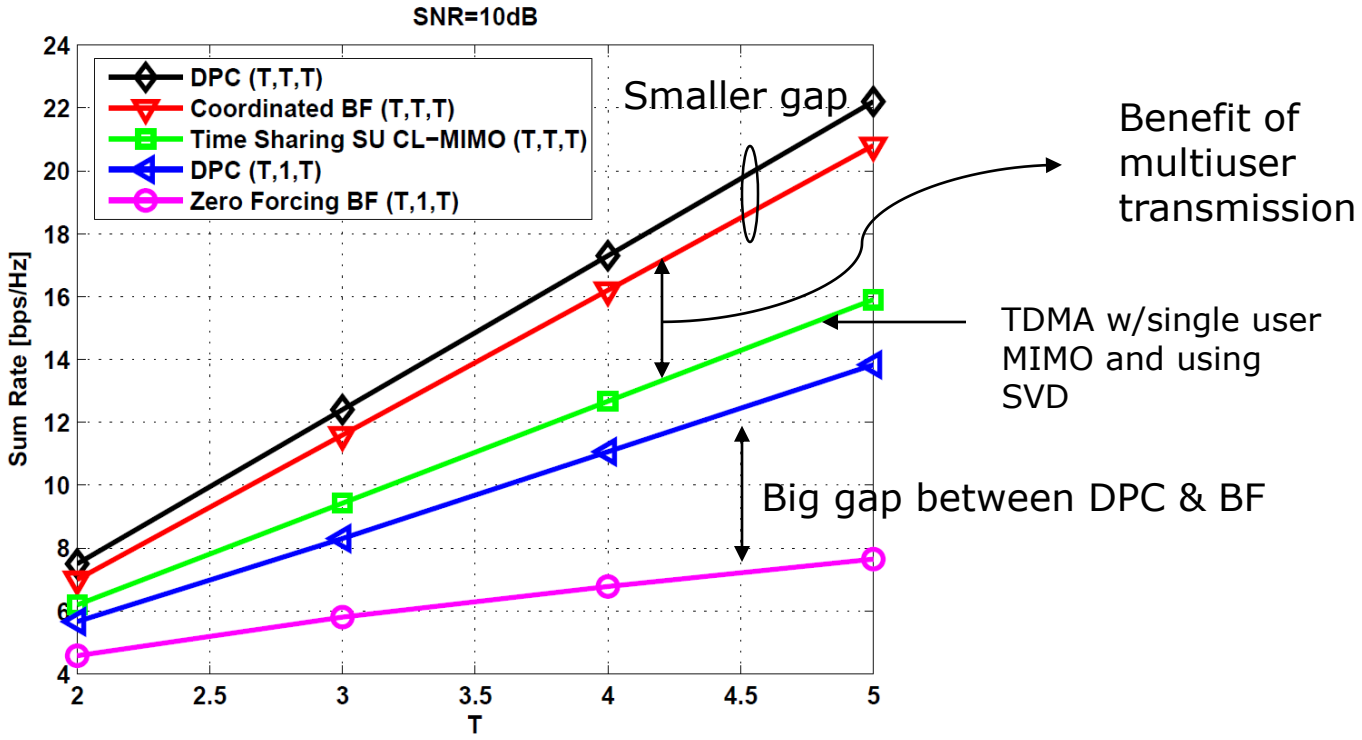


Fig. 2. Ergodic sum capacity and achievable sum rate as a function of the number of users, the number of transmit/receive antennas. (T_1, T_2, T_3) denotes the number of transmit antennas at the BS, the number of receive antennas at the user, and the number of active users in the network, respectively. Coordinated BF refers to the method presented in [19].

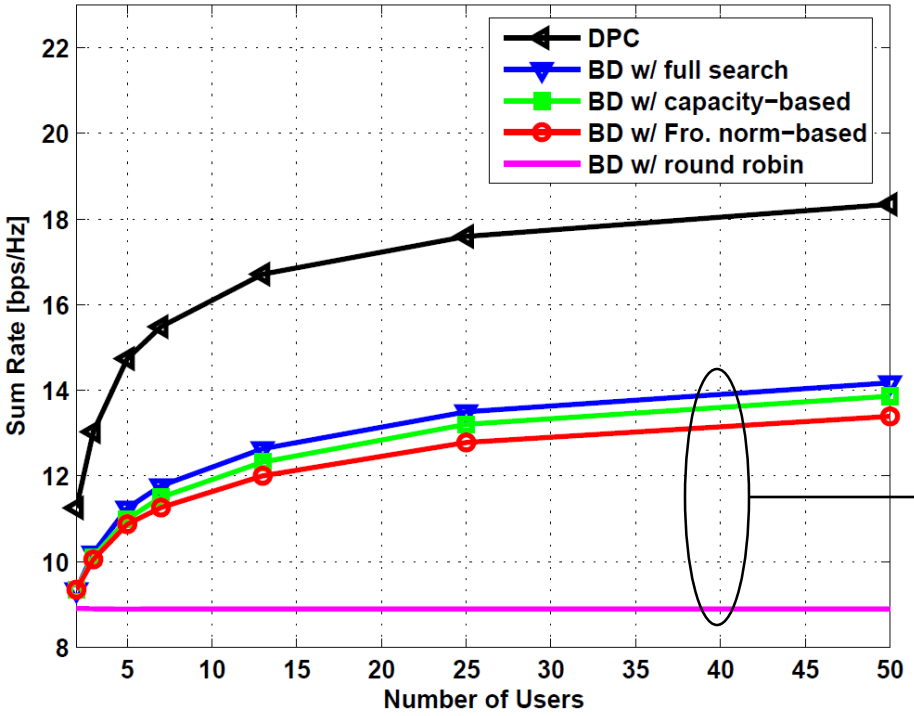
* D. Gesbert, M.Kountouris, R. W. Heath Jr., C.-B. Chae,and T. Sälzer, "Shifting the MIMO Paradigm," *IEEE Signal Processing Magazine,* pp. 36-46, Sept. 2007.

User Scheduling

- Of the K total users that are connected to the base station, at any given scheduling instant, the scheduler must choose N_u users receive transmissions.
- The rate of the system can be maximized by choosing users intelligently
- Scheduling depends on
  - Rate maximization
  - Resulting SINR
  - Fairness
  - Priority

Scheduling Performance

$N_t = 4$
$M_r^{(k)} = 2$
$N_u = 2$

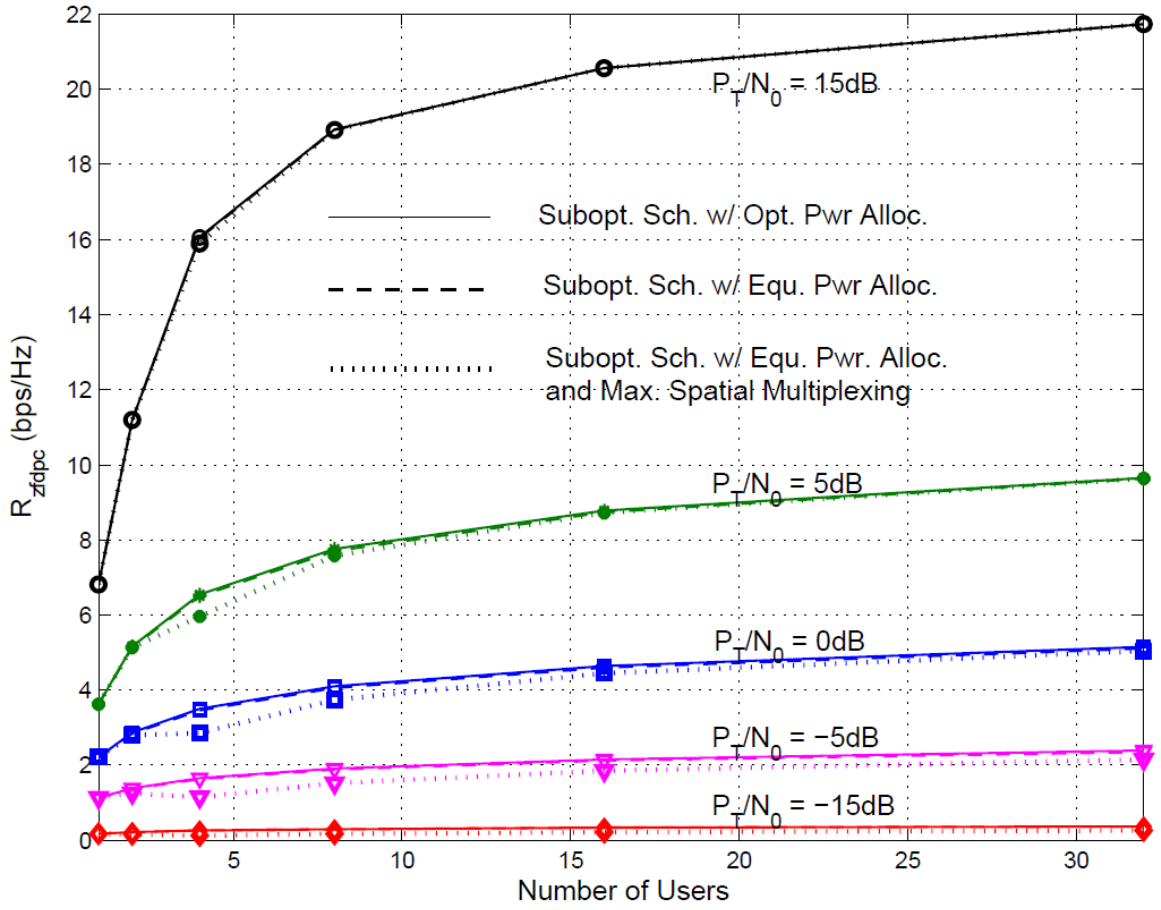


Linear precoding (Block Diagonalization)

Fig. 3. Sum rate as a function of the number of users, where the number of transmit antennas is 4, the number of receive antennas is 2, and the maximum number of selected users is 2.

* D. Gesbert, M.Kountouris, R. W. Heath Jr., C.-B. Chae,and T. Sälzer, "Shifting the MIMO Paradigm," *IEEE Signal Processing Magazine,* pp. 36-46, Sept. 2007.

Greedy Scheduling with ZF-DPC



- $N_t = 4$
- $M_r^{(k)} = 1$
- Benefit of MU-MIMO depends on # of users and SNR

- J. Jiang, R.M. Buehrer, and W.H. Tranter, "Greedy Scheduling Performance for a Zero-Forcing Dirty-Paper Coded System," *IEEE Trans. On Communications,* vol. 54, no. 5, May 2006.

Multiuser Packet Scheduling

- Greedy scheduler
  - Maximizes the instantaneous sum rate
  - Ignores fairness of time slot allocation
- Round-Robin (RR) scheduler
  - Allocates time slots to every subset of N_t users alternately
  - Complete fair, but not ignores channel state
- Proportional fair scheduler
  - A tradeoff between total throughput and fairness
  - Allocates slot t to user subset S^*

$$\mathrm{S}^* = \arg \max_{\mathrm{S}} \sum_{i=1}^{n_s} \frac{R_i(t)}{T_i(t)}$$

$R_i(t)$: achievalbe rate of user i at slot t;

$T_i(t)$: throughput of user i over a past window of length T_c, and updated slot-wise

$$T_i(t+1) = \begin{cases} (1-1/T_c)T_i(t) + R_i(t)/T_c, & i \in \mathrm{S}^* \\ (1-1/T_c)T_i(t), & \text{otherwise} \end{cases}$$

Maximum Sum Rate Scheduling

- ZF-THP with equal power allocation

$$\tilde{R}_{\text{sum}}^{\text{zfthp-max}} = \max_{S} \tilde{R}_{\text{sum}}^{\text{zfthp}}$$

over all ordered user subsets S with cardinality

$$|S| = N_t$$

- ZF-DPC with optimal power allocation

$$R_{sum}^{zfdpc} = \max_{S} \sum_{i=1}^{N_t} \left(\zeta \cdot \gamma_i^{zfdpc} \right) \qquad \text{with} \quad \sum_{i=1}^{N_t} \left[\zeta - \frac{1}{\gamma_i^{zfdpc}} \right] = P_t$$

over all ordered user subsets S with cardinality

$$|S| = N_t$$

A Suboptimal Packet Scheduler

- For $N_u \geq N_t$ a direct optimization of ZF-DPC and ZF-THP requires search over $N_u(N_u-1)\ldots(N_u-N_t+1)$ number of subsets
  - Prohibitive even for moderate N_u and N_t
- Observing that $g_i, i=1,\ldots,n_t,$ only depend on users j for $j \leq i$
  - At step $k,\ k=1,\ldots,n_t$, select the user k^*

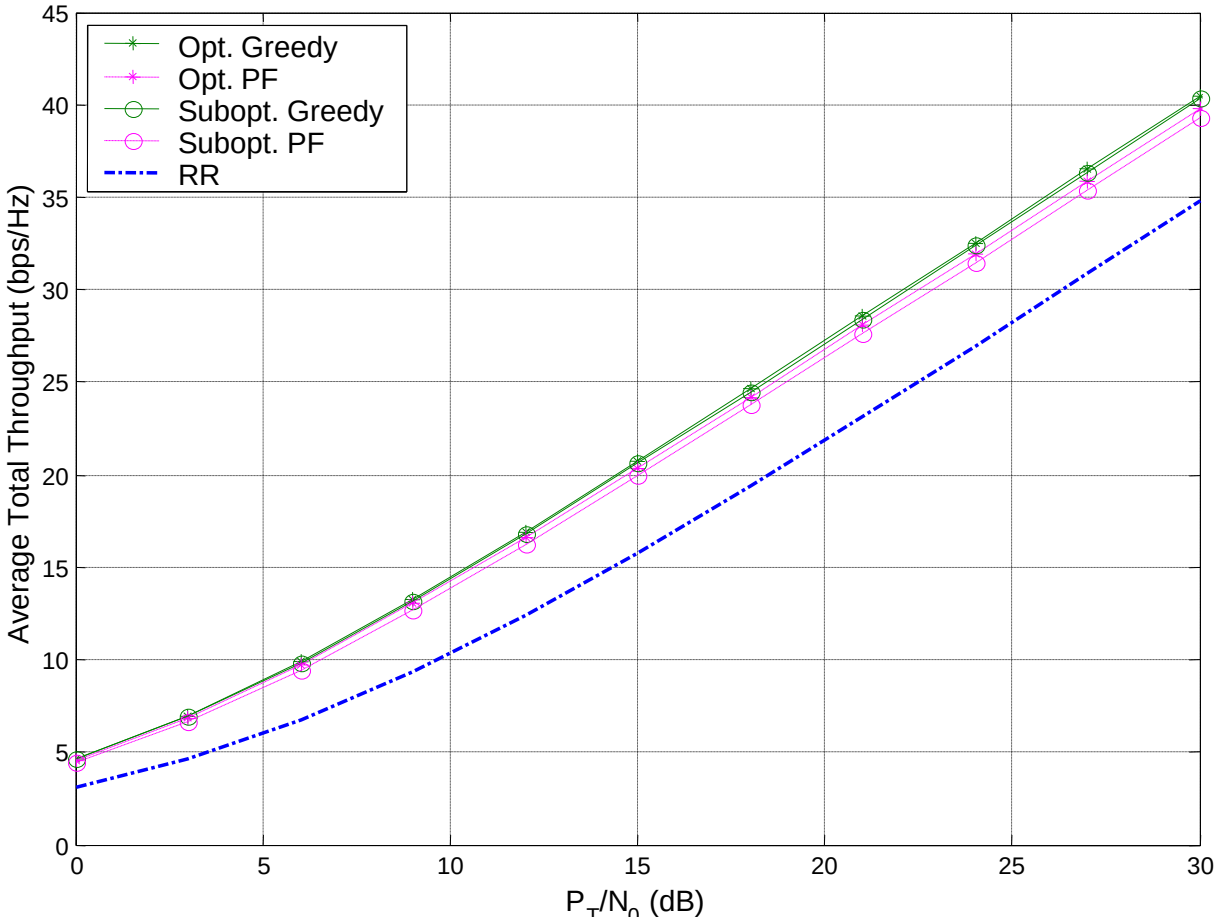
$$k^* = \arg \max_{k \in \{1^*,\ldots,(k-1)^*\}^c} \left[\sum_{i=1^*}^{(k-1)^*} \mu_i R_i(t) + \mu_k R_k(t) \right]$$

$\{1^*,\ldots,(k-1)^*\}^c$: complement of set $\{1^*,\ldots,(k-1)^*\}$

- The total number of user subsets under scheduling reduces to

$$N = \sum_{i=0}^{N_t-1} C_{N_u-i} = N_t(2N_u - N_t + 1)/2$$

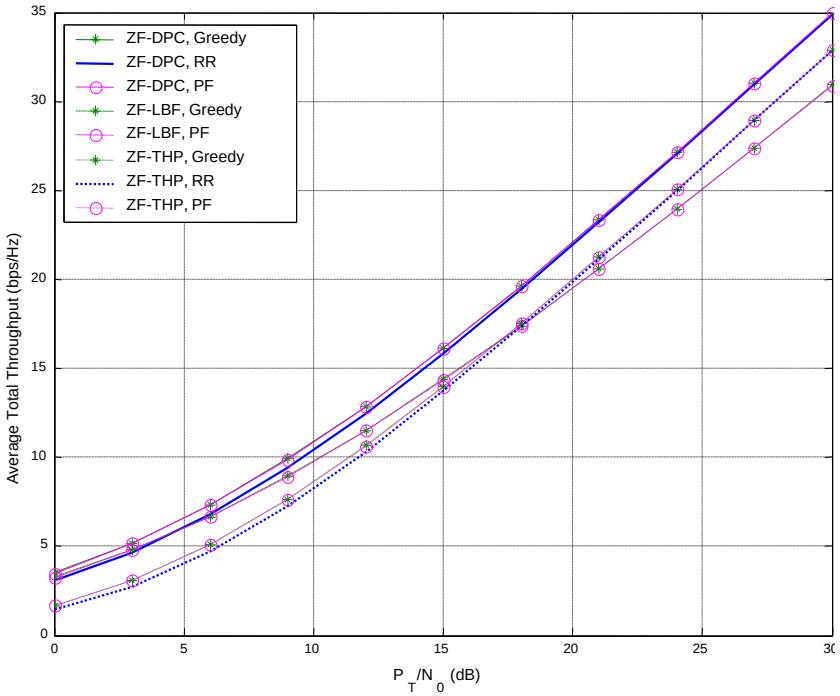
Optimal vs. Suboptimal Scheduling



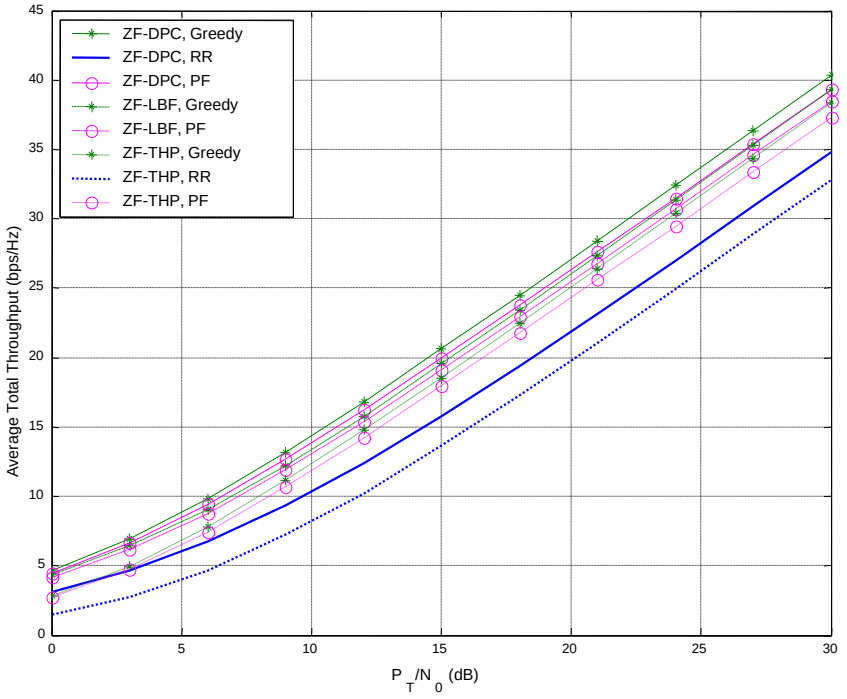
ZF-DPC for N_t=4 and N_u = *16* in i.i.d. flat Rayleigh fading with 5.6 Hz Doppler over 2000 slots.

Numerical Results

- Average maximum sum rate for $N_t = 4$ in i.i.d. flat Rayleigh fading



$N_u = 4$



$N_u = 16$

Scheduling and Precoding

- ZF-THP can provide high spatial multiplexing gain at medium to high total transmit power under channel-aware scheduling
- For ZF-DPC and ZF-THP, the suboptimal packet scheduler approximates the optimal scheduler closely and the complexity increasing linearly with the number of active users
- Spatial multiplexing over multiple transmit antennas enhances PF scheduling performance over slow fading with strong LoS components
- ZF-THP has a restricted output peak-to-rms due to the uniform distribution over the Voronoi region

Conclusions

- In this lecture we have examined how multiple antennas can improve *network* performance through signal processing and scheduling
- The primary benefit is on the downlink
- The cost is channel state information must be fed back to the transmitter for all users