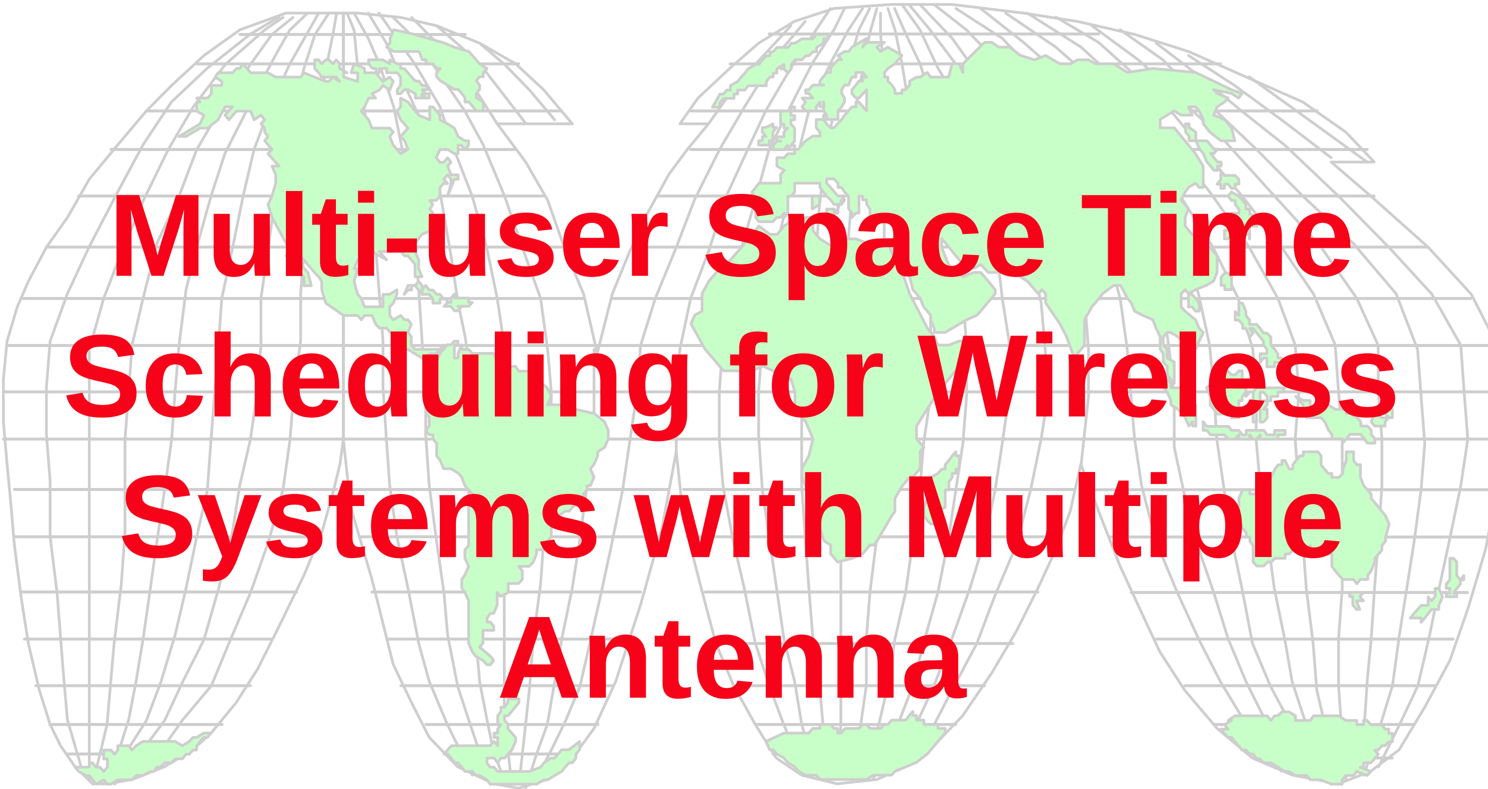


Multi-user Space Time Scheduling for Wireless Systems with Multiple Antenna

Vincent Lau

Associate Prof., University of Hong Kong

Senior Manager, ASTRI

Agenda

- **Background**
  - Link Level vs System Level Performance
  - Contributions of the Research Works.
- **Multiuser Downlink MIMO Space-Time Scheduling**
  - System Model
  - The multi-user space-time scheduling problem – information theoretical formulation & solution.
    - Single Cell Analytical Formulation & Optimal Scheduling Solution
    - Greedy-based Scheduling Algorithm
    - Genetic Scheduling Algorithm
  - Conclusion and Future Work

Background

“Link Level” versus “System Level”

- **Traditional layered approach in designing communication systems**
  - Isolated Optimization within layers without cross optimization.
  - Results in sub-optimal design, especially in wireless system where the physical channel is time varying.
- **Link Level Design for Wireless Channels:**
  - Focus on physical layer design to optimize the link capacity at given bandwidth and power budget.
  - Multiple transmit and receive antenna used to increase the capacity of the wireless link (at a given power and bandwidth budget) by forming “spatial channels”.
- **System Level Design for Wireless Channels:**
  - System level refers to the situation when we have multiple users.
  - Since data source is usually very bursty, packet scheduling is a very important component in the higher layer to achieve statistical multiplexing.
  - Achieving link level optimization does not always achieve system level optimization. → Joint design is important to exploit the time varying physical channel in wireless system.

Contributions of the Research Work

Q1) What is the optimal scheduling performance for multi-user MIMO?

Ans 1) Based on the proposed analytical framework, optimal space time scheduling performance is obtained as a performance reference.

Q2) How good is the widely used "greedy-based" space-time scheduling algorithms in 3G1x, EV-DO, EV-DV, HSDPA?

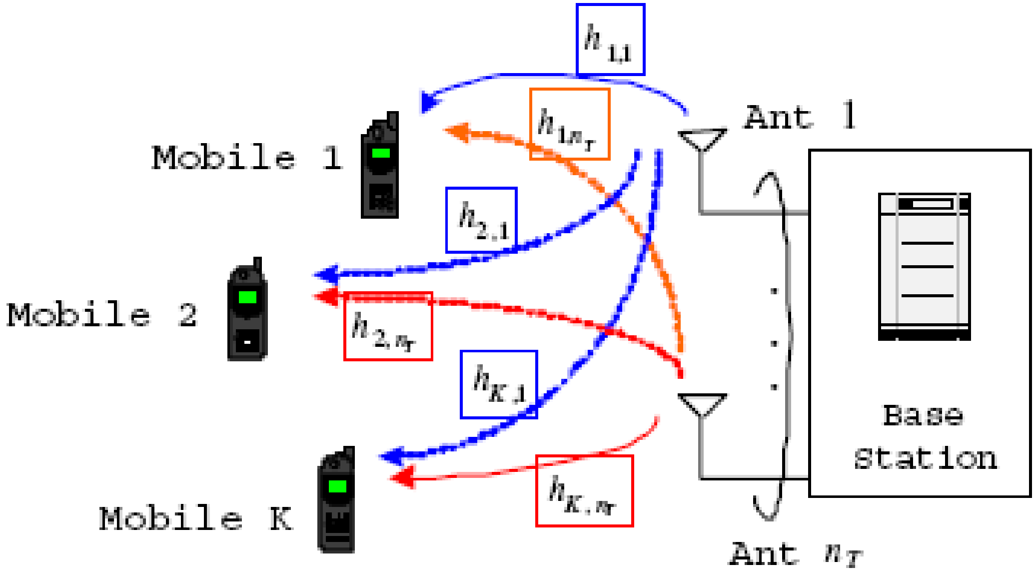
Ans 2) The "greedy-based" algorithms are widely used in existing systems and they achieve optimal performance for nT=1. Yet, there is a significant performance gap for nT>1.

Q3) Any better scheduling heuristics that could achieve better complexity – performance tradeoff?

Ans 3) Propose a low complexity genetic scheduling algorithm.

PART A:
Multi-User MIMO scheduling – Downlink, Single Cell:

System Model - Downlink

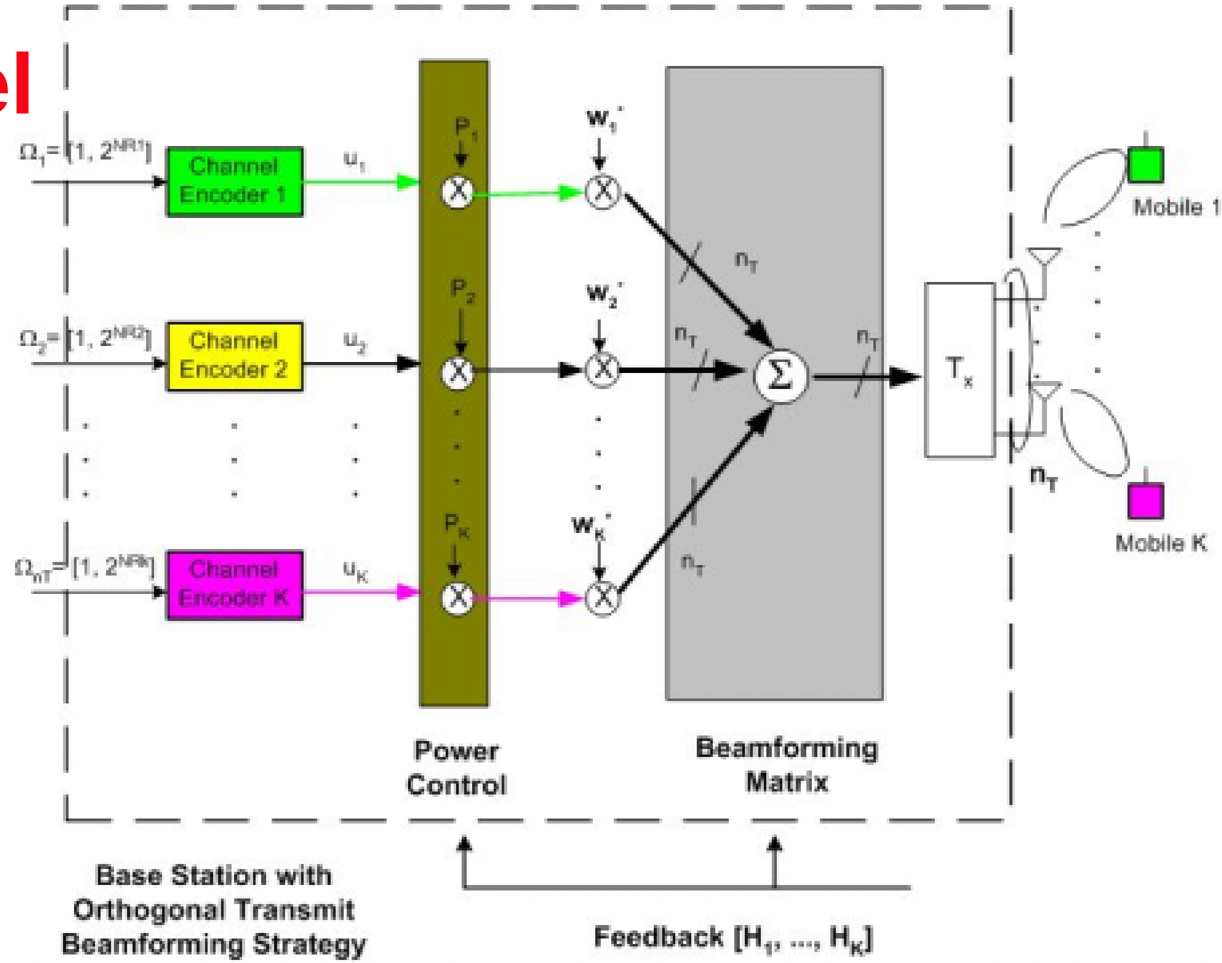


Design Constraints

- Linear Processing Constraint at Base Station
  - Orthogonal Transmit Beam-Forming
- Complexity Constraint at Mobiles
  - Single-Antenna mobile + Simple single-user processing capability
- Transmit Power Constraint
  - Total Transmitted power at base station at most P_{tx}

System Model

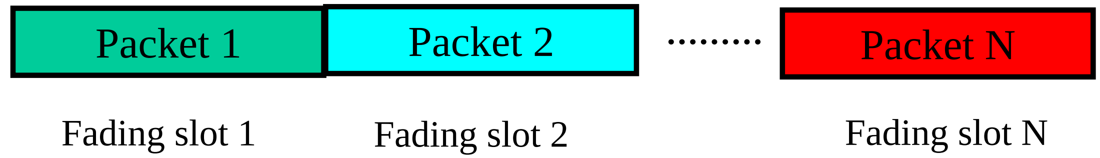
- **Orthogonal Transmit Beam-forming Structure (OTBF)**
  - Isolated Encoding Per User → U_k
  - Selectively switch on and off a branch by setting $p_k = 0$



- **Base Station Transmitted Signal:**

$$\underbrace{\begin{bmatrix} x(1) \\ \vdots \\ x(n_T) \end{bmatrix}}_{\text{Transmitted Signal } \mathbf{X}} = \underbrace{\begin{bmatrix} \mathbf{w}_1 & \cdots & \mathbf{w}_K \end{bmatrix}}_{\text{beam-forming matrix } \mathbf{W}} \underbrace{\begin{bmatrix} \sqrt{p_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{p_K} \end{bmatrix}}_{\text{Power Control Matrix } \mathbf{P}} \underbrace{\begin{bmatrix} U_1 \\ \vdots \\ U_K \end{bmatrix}}_{\text{Encoding Symbols } \mathbf{U}} = \sum_{k=1}^{K} \sqrt{p_k} U_k \mathbf{w}_k$$

System Model



- **Channel Model:**
  - Short burst duration + pedestrian mobility
  - Quasi-static fading → channel fading remains approximately constant within an encoding frame.
  - TDD → downlink channel matrices could be estimated at the uplink side without explicit feedback.
- **Source Model:**
  - To decouple the problems, we assume saturated analysis
  - Infinite buffer size at base station → Every mobile always has packets to transmit at every fading slot.
  - Performance of system is based on throughput and is therefore independent of source model.
- **Physical Layer Model:**
  - Based on information theoretical capacities to decouple the performance from specific implementations of channel coding and modulation.
  - Standard random codebook & Gaussian constellation → arbitrarily low error probability for data rate less than Shannon's capacity.
  - These assumption could be approximated for turbo-coded systems.

System Model

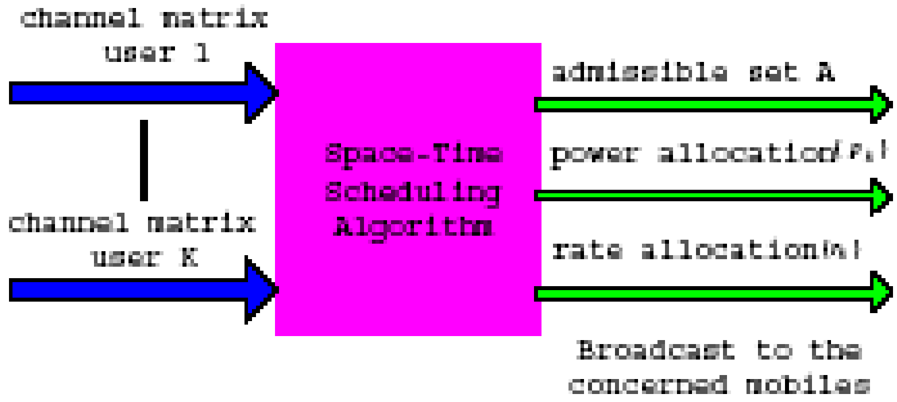
- **Received signal at the k-th mobile (in a fading slot):**

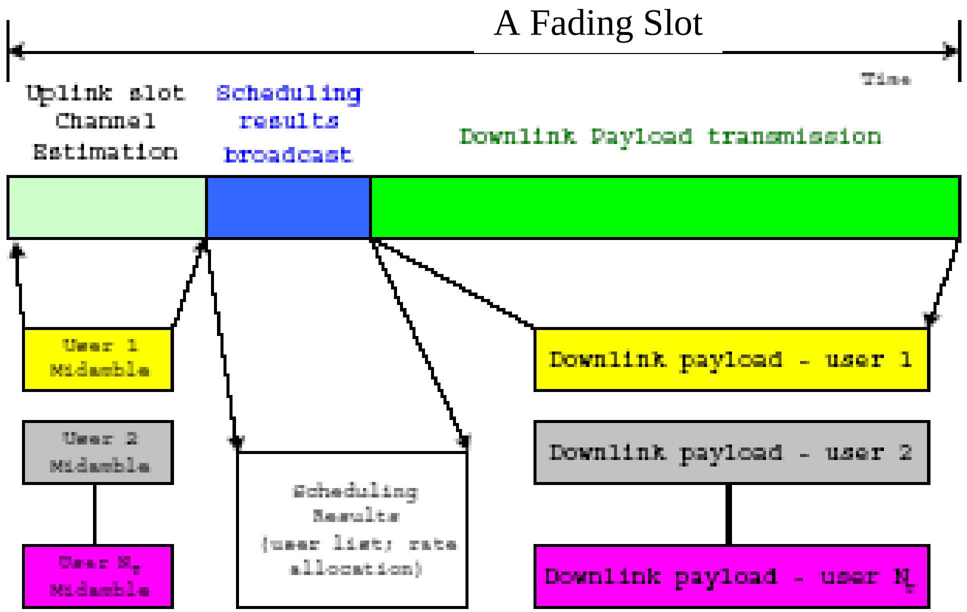
$$Y_k = \mathbf{h}_k \mathbf{X} + Z_k = \underbrace{\sqrt{p_k}\,\mathbf{h}_k \mathbf{w}_k U_k}_{\text{Information}} + \underbrace{\sum_{m \neq k} \sqrt{p_m}\,\mathbf{h}_k \mathbf{w}_m U_m}_{\text{Multi-beam Interference}} + \underbrace{Z_k}_{\text{Channel Noise}}$$

- **Admissible User Set:** $\mathbf{A} = \left\{ k \in [1, K] : p_k > 0 \right\}$
  - Set of users selected for transmission in the current fading slot
- **Beam-Forming Weight Selection:**
  - Eliminate multi-beam interference: $\begin{cases} \mathbf{w}_k^* \mathbf{w}_k = 1 & \forall k = 1, .., K \\ \mathbf{h}_k \mathbf{w}_m = 0 & \forall m \in \mathbf{A}, m \neq k \end{cases}$
- **Cardinality of Admissible User Set**
  - Due to limited degree of freedom with n_T transmitted antennas, the maximum cardinality of Admissible set is: $|\mathbf{A}| \leq n_T$.
  - In other words, at most n_T simultaneous transmission is allowed at any fading slot.

System Model

- **MAC Layer Model**:
  - Base station estimates the channel matrices of all users (per fading slot)
  - Set of channel matrices are passed to the scheduling algorithm
  - Output of scheduler = admissible set, power allocation, rate allocation.
  - Scheduling results are broadcast to all users (per fading slot).
  - Payload transmission takes place in the payload field of the downlink frame.





System Performance → System Utility

- System Performance – General Convex Utility Function

$$U\left(R_1,\ldots,R_K\right)=E\left[G\left(r_1,\ldots.,r_K\right)\right] \qquad R_k=E\left(r_k\right)$$

$$r_k = \text{instantaneous throughput of user-k}$$

  – Expectation is taken over various fading slots.
  – Scheduling Algorithm → optimize a given system utility function.

- (A) Maximal Throughput $U_{\text{maxthp}}\left(R_1,\ldots,R_K\right)=E\left[\sum_{k=1}^{K} r_k\right]$
- (B) Proportional Fair $U_{PF}\left(R_1,\ldots,R_K\right)=\sum_{k=1}^{K}\log\left(R_k\right)$
  – **Lemma 1:** A scheduler that maximizes $\tilde{G}_{PF}\left(R_1,\ldots,R_K\right)$ would also maximizes $U_{PF}\left(R_1,\ldots,R_K\right)$ where $\tilde{G}_{PF}\left(R_1,\ldots,R_K\right)=\left[\sum_{k=1}^{K}\frac{r_k}{R_k}\right]$
  – We further approximate R_k with moving window average

$$R_k\left(t+1\right)=\left(1-1/t_c\right)R_k\left(t\right)+1/t_c\, r_k\left(t+1\right)$$

Scheduling Problem

- Over a large number of fading slots, choose the admissible sets $\{\mathbf{A}\}$ & power allocation policy $\mathbf{P} = \{(p_1, p_2, ..., p_K)\}$ so that the system utility function is maximized.

$$\max_{\{\mathbf{A}\},\{(p_1,...,p_K)\}} \{U(R_1,...,R_K)\} = \max_{\{\mathbf{A}\},\{(p_1,...,p_K)\}} \{E_{\mathbf{H}}[G(r_1,...,r_K)]\}$$

$$= E_{\mathbf{H}}\left[\max_{\mathbf{A},(p_1,...,p_K)} \{G(r_1,...,r_K)\}\right]$$

Analytical Formulation – per fading slot

- Define a binary vector $\left(\alpha_1, \ldots, \alpha_K\right)$ where $\alpha_k = \begin{cases} 1 & k \in \mathbf{A} \\ 0 & k \notin \mathbf{A} \end{cases}$

- The scheduling problem is given by:

Given a channel matrix realization for all K users, $\left\{\mathbf{h}_1, \ldots., \mathbf{h}_K\right\}$, find the optimal binary vector $\left(\alpha_1, \ldots, \alpha_K\right)$ such that the system utility function $G\left(r_1, \ldots., r_K\right)$ is maximized with the constraint

$$\sum_{k=1}^{K} \alpha_k p_k \leq P_{tx} \quad \text{(Power Constraint)} \qquad \sum_{k=1}^{K} \alpha_k \leq n_T \quad \text{(Degree of freedom Constraint)}$$

and the achievable throughput of user k is given by:

$$r_k = \log_2\left(1 + \frac{\alpha_k p_k \left|\mathbf{h}_k \mathbf{w}_k\right|^2}{\sigma_z^2}\right)$$

- The optimizing variables = power allocation (continuous) $\left(p_1, \ldots, p_K\right)$ & admissible set (discrete) $\left(\alpha_1, \ldots., \alpha_K\right)$

Optimal Solution – Mixed Integer Programming

- **Step I (Convex Optimization on power allocation)**
  - Given a specific admissible set A, the optimal power allocation is given by:

$$p_k^*(\mathrm{maxthp}) = \left(\frac{1}{\lambda} - \frac{1}{\left|\mathbf{h}_k \mathbf{w}_k\right|^2} \right)^+ \qquad p_k^*(\mathrm{PF}) = \left(\frac{1}{\bar{R}_k \lambda} - \frac{1}{\left|\mathbf{h}_k \mathbf{w}_k\right|^2} \right)^+$$

$$\lambda = \text{Lagrandge Multiplier chosen to satisfy } \sum_k \alpha_k p_k(\lambda) \le P_{tx}$$

- **Step II (Discrete Optimization on admissible set)**
  - Combinatorial search over all possible admissible set satisfying $\sum_{k=1}^{K} \alpha_k \le n_T$.
  - Search Space is huge:

$$\sum_{m=1}^{n_T} \binom{K}{m}$$

Heuristic Scheduling Algorithms – (A) Greedy-Based Baseline

- Greedy-based Scheduling Algorithm – Baseline
  - **Step I:** For k = 1: K,
    - Initialize $\boldsymbol{\alpha}(k) = \Big(0, 0, \ldots, \underbrace{1}_{\text{k-th element}}, 0, \ldots 0\Big)$ $\mathbf{p}(k) = \Big(0, \ldots., 0, \underbrace{P_{tx}}_{\text{k-th element}}, 0, \ldots., 0\Big)$
    - Calculate $G_k^* = G\left(0, \ldots, 0, r_k, 0, \ldots 0\right)$ where r_k is based on $\boldsymbol{\alpha}(k), \mathbf{p}(k)$
  - **Step II**: Sort in descending order of $\{G_k^*\}$ calculated in step I.
  - **Step III**:
    - The ***admissible set*** is given by the first n_T user indices from the sorted list in Step II.
    - The ***power allocation*** is given by equations in previous page.
- Computational complexity ~ linear in K
- Achieve optimal performance for $n_T = 1$
- Widely used in existing systems such as 3G1x, EV-DO, UMTS-HSDPA

Heuristic Scheduling Algorithm – Genetic Based

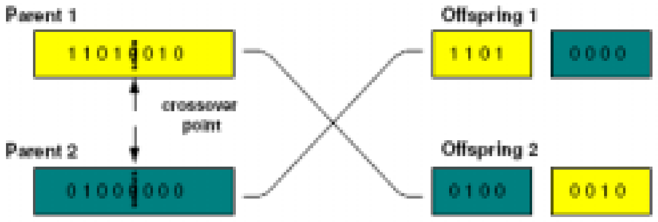
- Genetic-Based Scheduling Algorithm
  - Define a chromosome to be the binary vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \alpha_k \in \{0,1\}$
  - **Step I: Initialization**
    - Initialize a population of N_p chromosomes satisfying the constraint $\sum_{k=1}^{K} \alpha_k \le n_T$
  - **Step II: Selection**
    - Construct an intermediate population based on current population & a selection rule.
    - For each randomly selected (i-th) chromosome from the current population, evaluate it's fitness:

$$G^*_{\alpha,i} \big/ \bar{G} \quad : G^*_{\alpha,i} = \max_{(p_1,\ldots,p_K)} \left\{ G\left(r_1,\ldots,r_K \middle| \boldsymbol{\alpha}(i)\right) \right\}, \quad \bar{G} = \sum_i G^*_{\alpha,i}$$

    - The integral portion determines how many copies of the i-th chromosome are placed into the intermediate population.
    - The fractional portion determines the probability that an additional copy is placed.
    - The selection process carries on until all N_p slots have been filled up in the intermediate population.

Heuristic Scheduling Algorithm – (B) Genetic Based Scheduling.

- **Step III: Breeding**
  - Randomly select a pair of chromosomes in the intermediate population & combines the 2 parents into 2 off-springs according to a cross-over and a mutation rules.
  - There is a probability of p_c to perform cross-over.



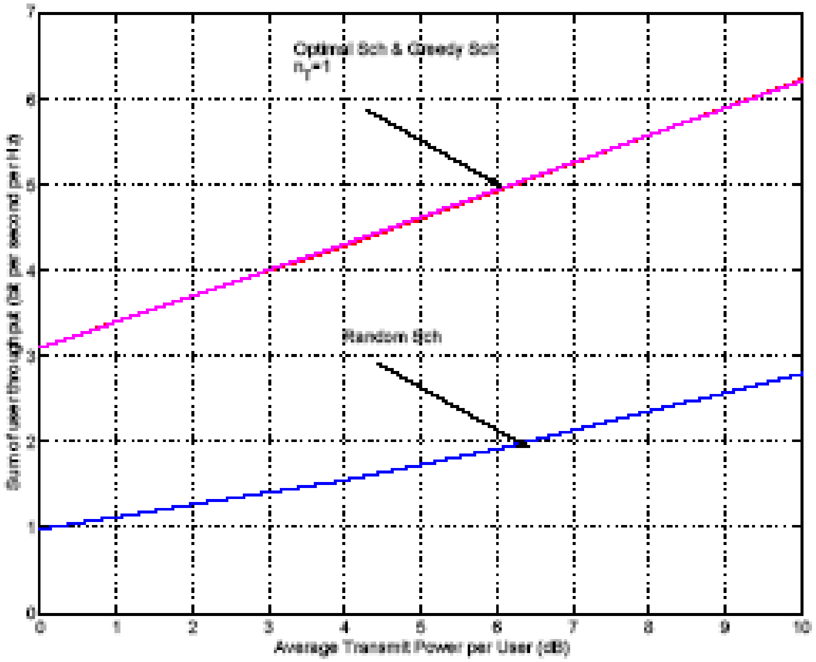
  - For every bit in the cross-over outputs, there is a p_m probability of performing mutation (bit toggling).
  - Dynamically adapts the mutation probability with the spread of the fitness.

$$p_m = \frac{1}{\beta_1 + \beta_2 \sigma_G / \overline{G}}$$

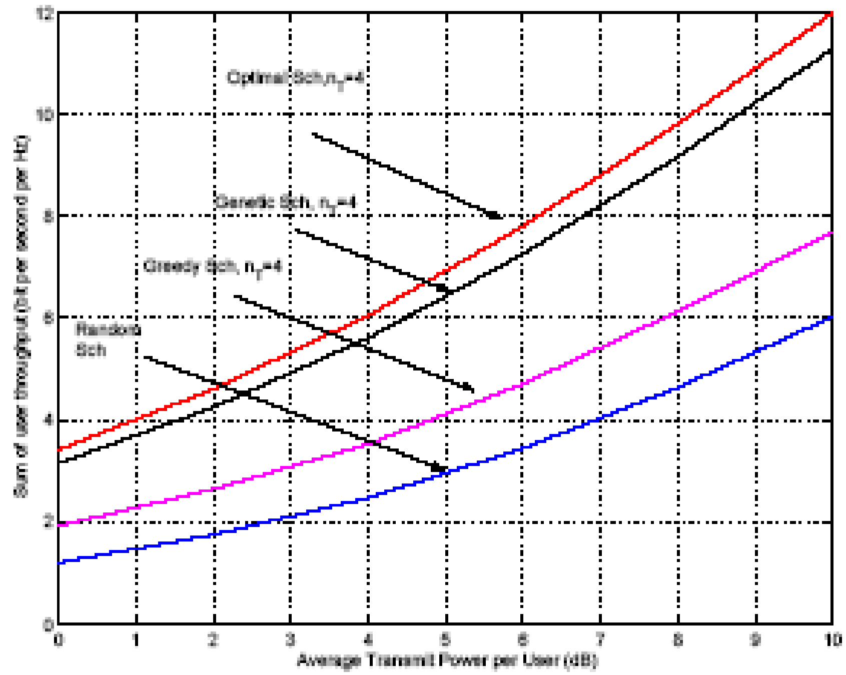
- **Step IV: Termination**
  - For processed chromosomes violating the constraint, '0' is randomly inserted into the chromosome until the constraint is satisfied. The intermediate population becomes the current population and step I-III are repeated for N_g times.

Numerical Results – Maximal Throughput Scheduler

System Throughput vs SNR (nT = 1)



System Throughput vs SNR (nT = 4)



- Greedy-based baseline algorithm achieved optimal performance at single antenna
- Performance gap between the greedy-based baseline scheduler and optimal scheduler is quite large for multiple antennas.
- Comparison w.r.t. random scheduler → multi-user diversity gain of scheduling.
- Genetic algorithm could fill in the performance gap.

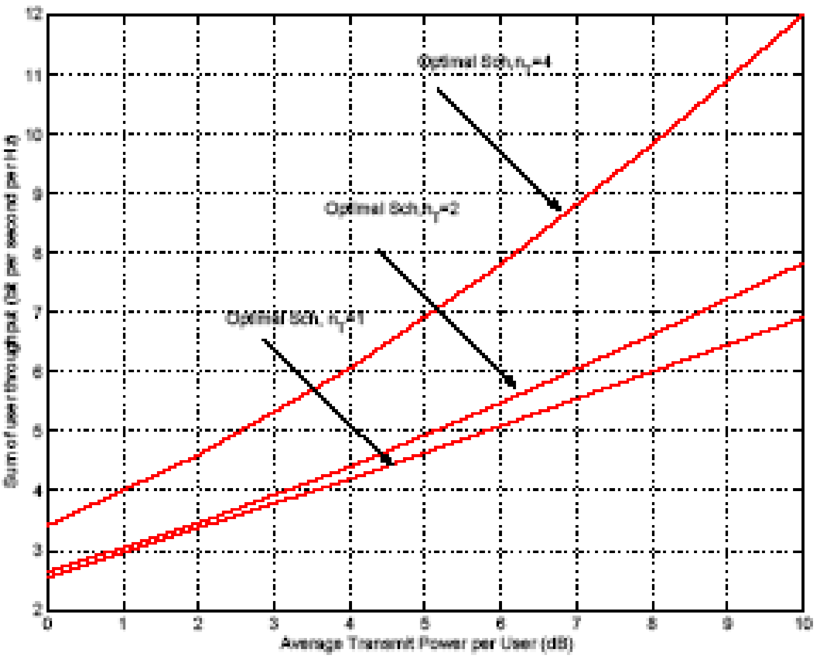
Numerical Results – Maximal Throughput Scheduler

- Complexity comparison
  - At 20 users and 4 transmit antennas, genetic algorithm is ~ ***36 times less complex*** than optimal algorithm. Yet, genetic algorithm is ~ 5 times more complex than the greedy-based baseline algorithm. → a reasonable performance – complexity tradeoff.

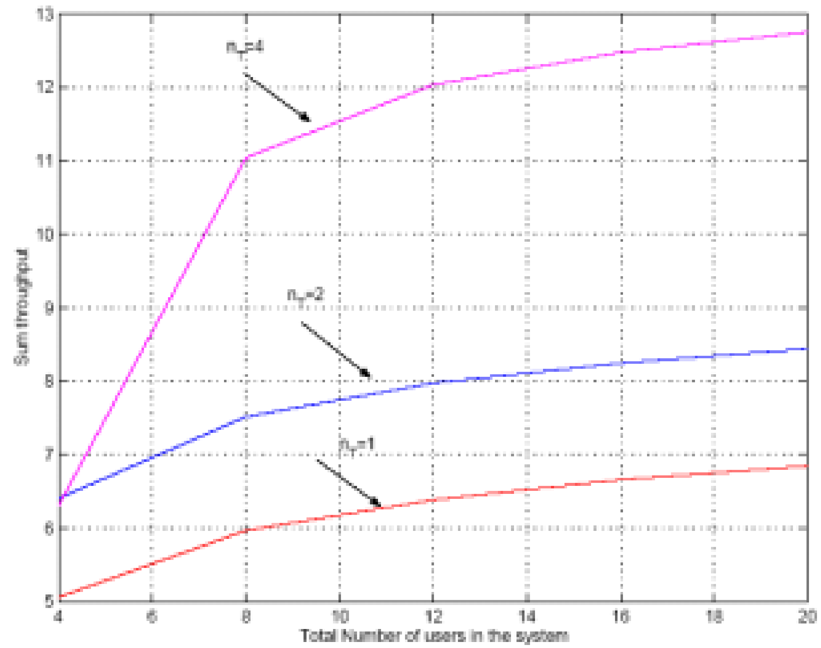
(K, n_T)	Greedy Algorithm	Genetic Algorithm	Optimal Algorithm
(10,2)	10 + sorting	10x2=20	55
(10,4)	10 + sorting	10x5=50	385
(20,2)	20 + sorting	10x5=50	210
(20,4)	20 + sorting	20x5=100	3645

Numerical Results – Maximal Throughput Scheduler

Capacity vs nT



Capacity vs K

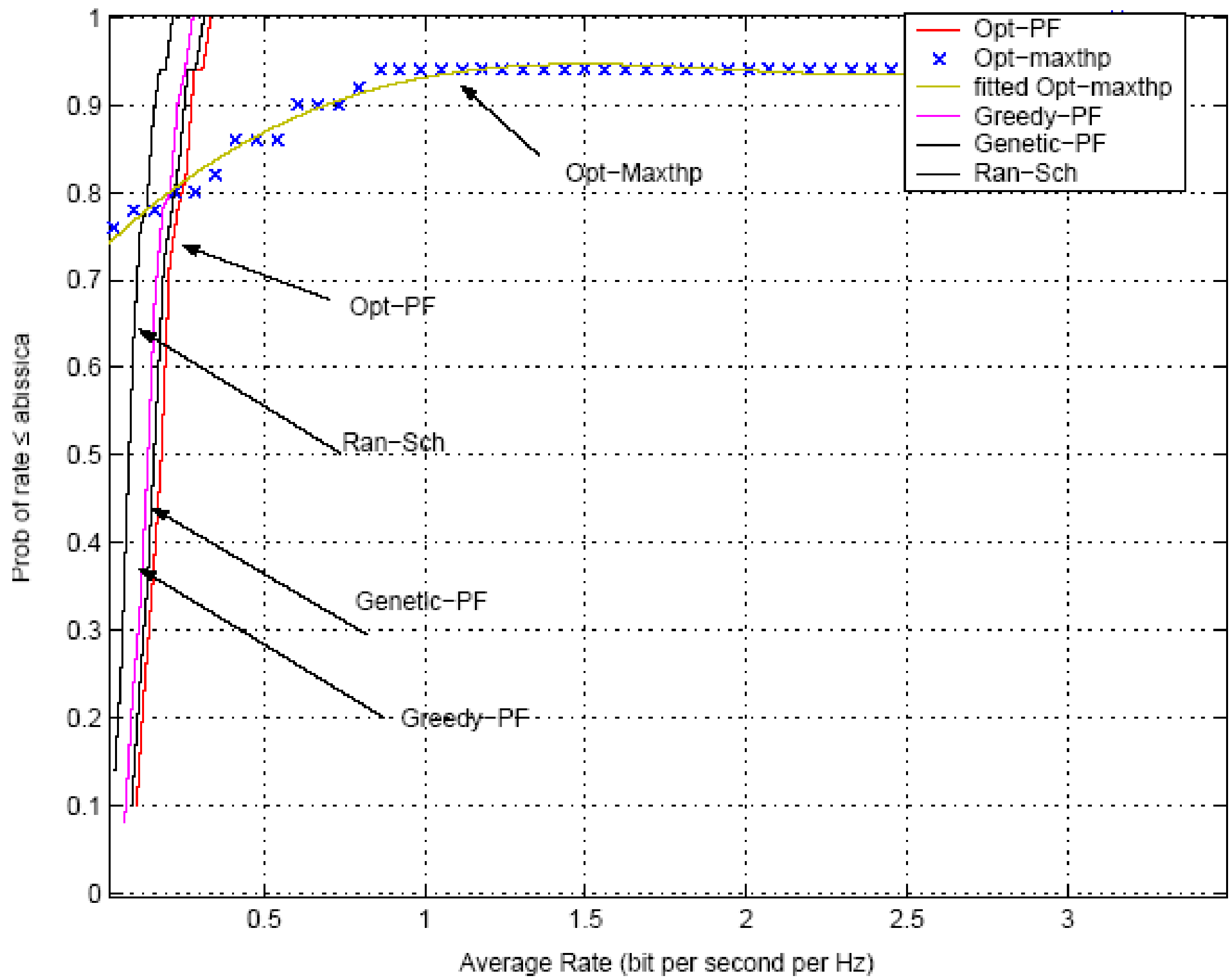


- Capacity gain vs nT
  - Increasing nT enhances system throughput at high SNR → due to multi-beam transmission (spatial multiplexing)
  - Capacity gain at small SNR is insignificant ~ limited by power splitting.
- .At moderate K~10, the multi-user diversity gain is already significant.

Numerical Results – PF Scheduler

User throughput c.d.f.

- K=50, nT=2.
- Genetic algorithm → Over 90% of users could achieve a throughput of 0.2
- Greedy-based baseline algorithm → Over 90% of users could achieve a throughput of 0.1.
- Random scheduler → Over 90% of users could achieve a throughput ~ 0.02.



Conclusion

- Analytical framework is proposed (based on information theory) to model the multi-user space-time scheduling problem (single cell) & obtain optimal scheduling performance as reference.

- Commonly employed greedy-based baseline algorithm → optimal only in single antenna, large performance gap at multiple antennas.

- Proposed a genetic based algorithm → reasonable complexity, performance tradeoff for multiple antenna scheduling.

- On-going works → robust scheduling w.r.t. channel estimation errors.