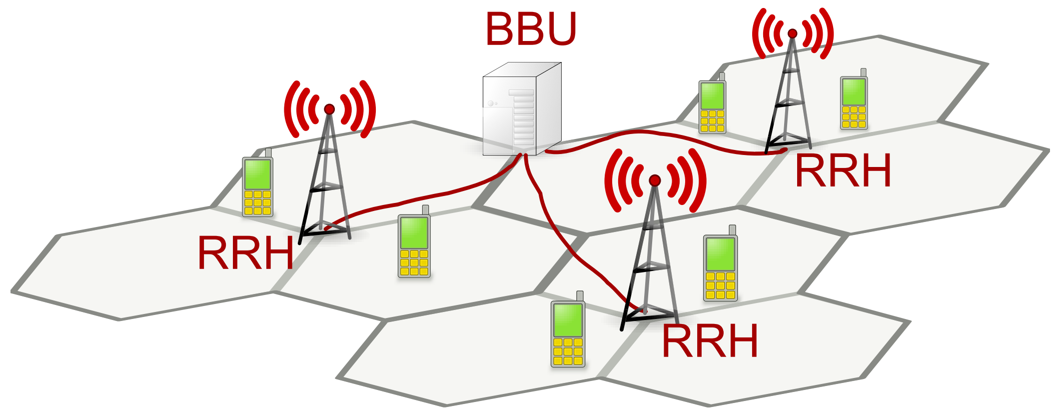
Adaptive Allocation of Compute Resources for LTE Phy Layer Processing

47. Treffen der VDE/ITG-Fachgruppe 5.2.4

Thomas Werthmann
thomas.werthmann@ikr.uni-stuttgart.de
25.06.2015

Universität Stuttgart
Institute of Communication Networks
and Computer Engineering (IKR)
Prof. Dr.-Ing. Andreas Kirstädter

Cloud Radio Access Network (C-RAN)



Principle

Split of the functional units
of a cellular base station

- Remote Radio Head (RRH)

  Antenna, power amplifier
  and AD/DA converters

- Base Band Unit (BBU)

  Processors for singal processing,
  coding, MAC layer and other tasks

Benefits

→ Simpler maintenance, reduced OPEX

→ Easier implementation of inter-site coordination

→ Multiplexing gains for the processor load

Scalable Compute Effort

LTE Requirements

- Strict real time constraints
- Significant signal processing effort

Cloud RAN Vision

- Implement C-RAN in software running on off-the-shelf hardware
- Use as much standard-IT as possible
  - Operating systems
  - Virtualization techniques
  - Communication networks
- Avoid overprovisioning

→ Computational jitter, no hard real-time guarantees

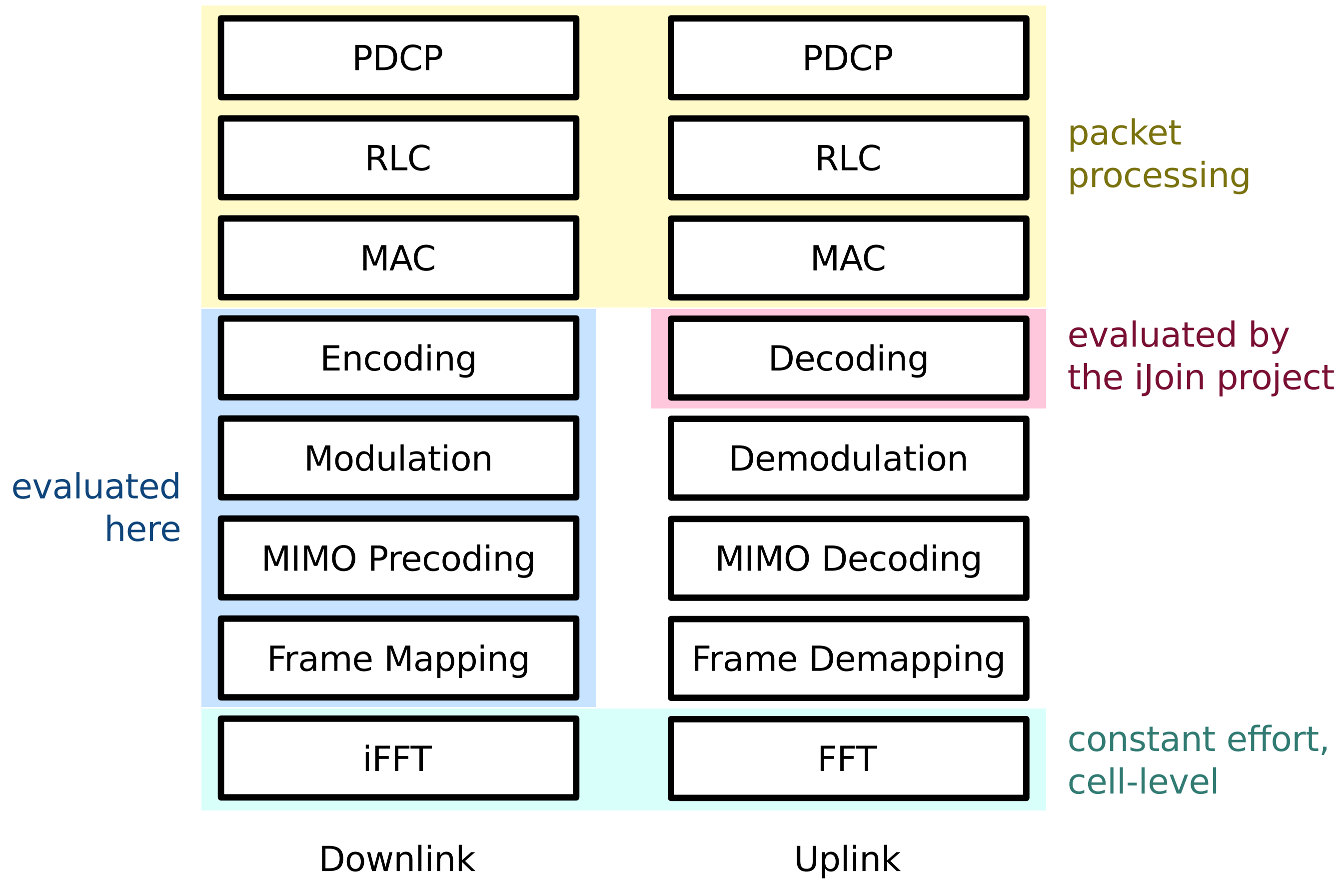
→ Cloud typically used for elastic applications

→ Better to reduce system performance than to break real-time requirements

→ Elastic scaling of C-RAN compute effort required

Where to Safe Compute Effort?

LTE Protocol Stack



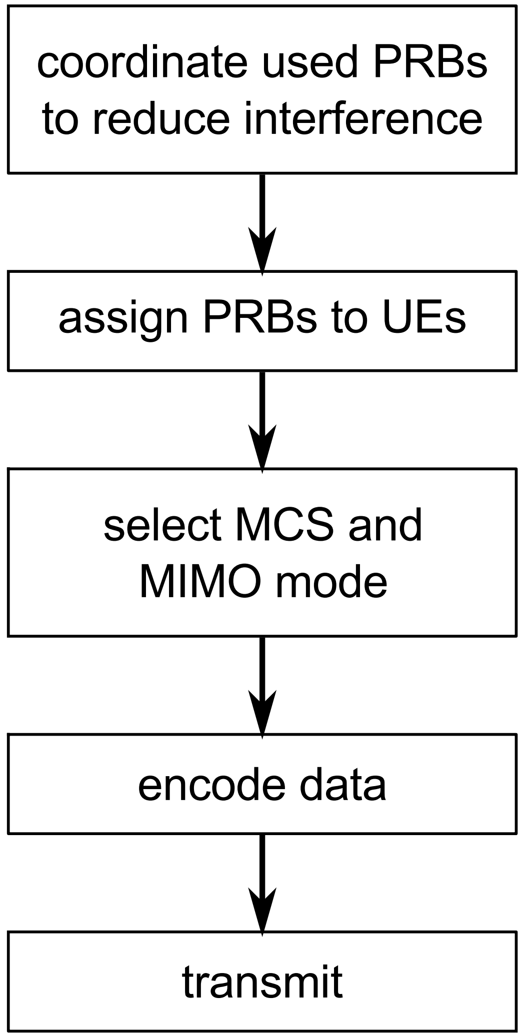
→ **Here concentrating on the downlink Phy layer**

Resource Saving Strategies

Sequence of Tasks in Existing Systems

- Interference coordination restricts usable PRBs (or transmit power)
- Scheduler assigns PRBs to balance fairness and throughput
- MCS and MIMO mode selection strives to achieve highest possible throughput
- Encoding and transmission follow unconditionally

→ **Adaptation required to cope with compute resource restrictions**



© 2015 Universität Stuttgart • IKR

Resource Saving Strategies

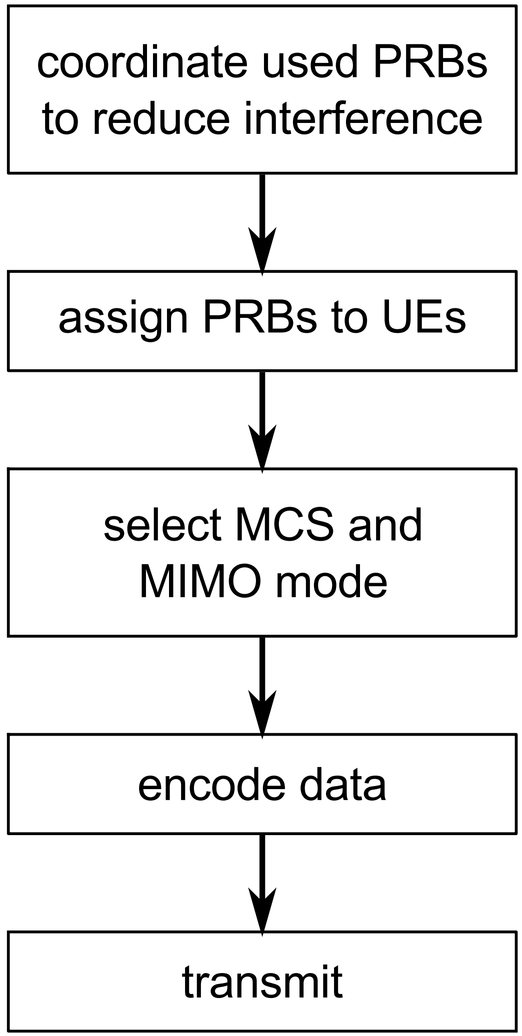
How to save Compute Resources?

- Adapt interference coordination
  - improved SINR can compensate for empty PRBs
- Adapt scheduling
  - assing PRBs to UEs with "simpler" channels
- Adapt MIMO mode selection
  - use less demanding MIMO modes, e.g. less Tx antennas
- Don’t adapt existing tasks,
  but reduce PRB allocations or skip UEs subsequently

Or just one large optimization problem?

→ **What is the most efficient strategy?**

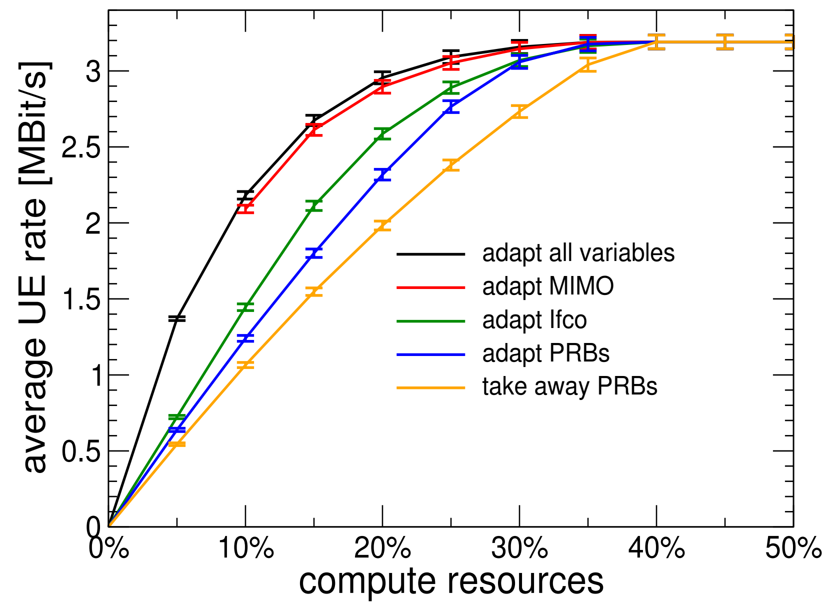
→ **How much performance is lost compared to a joint optimization?**



Insights from Optimization Studies

Evaluation

- Studies with optimizer in simple model
- Optimization problem (ILP)
  – Variables: IfCo, scheduling and MIMO modes
  – Objective: optimal performance (prop. fair)
  – Restrictions: limited compute resources
- Here adapting only part of the variables to model realization strategies



Results

- Resources can be reduced to 40% without impacting system performance
- Gradually degradation of system performance, tolerates overload
- By adapting MIMO modes, close-to-optimal performance can be achieved

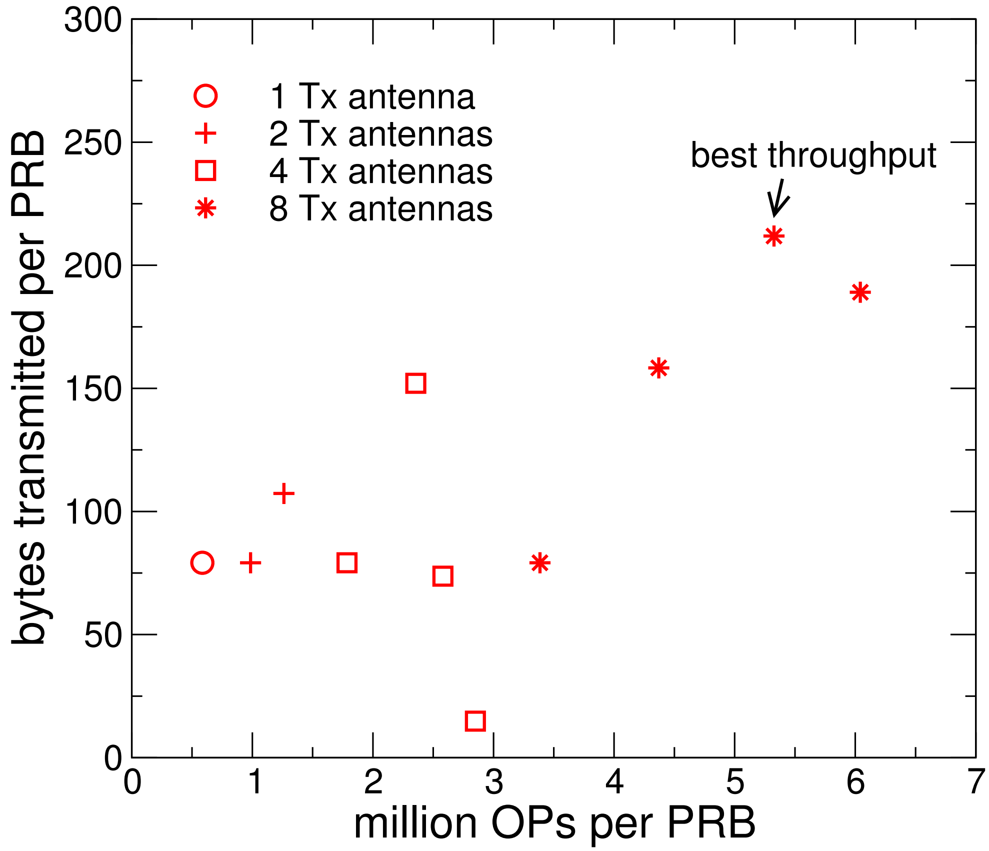
→ **Design heuristic which adapts MIMO modes to scale effort**

Thomas Werthmann, 2015: Approaches to Adaptively Reduce Processing Effort for LTE Cloud-RAN Systems (Workshop on Cloud-Processing in Heterogeneous Mobile Communication Networks)

Proposed System

MIMO modes of a Single User

- Throughput (and effort) depend on channel characteristics
- With unrestricted compute resources, best throughput mode would be used

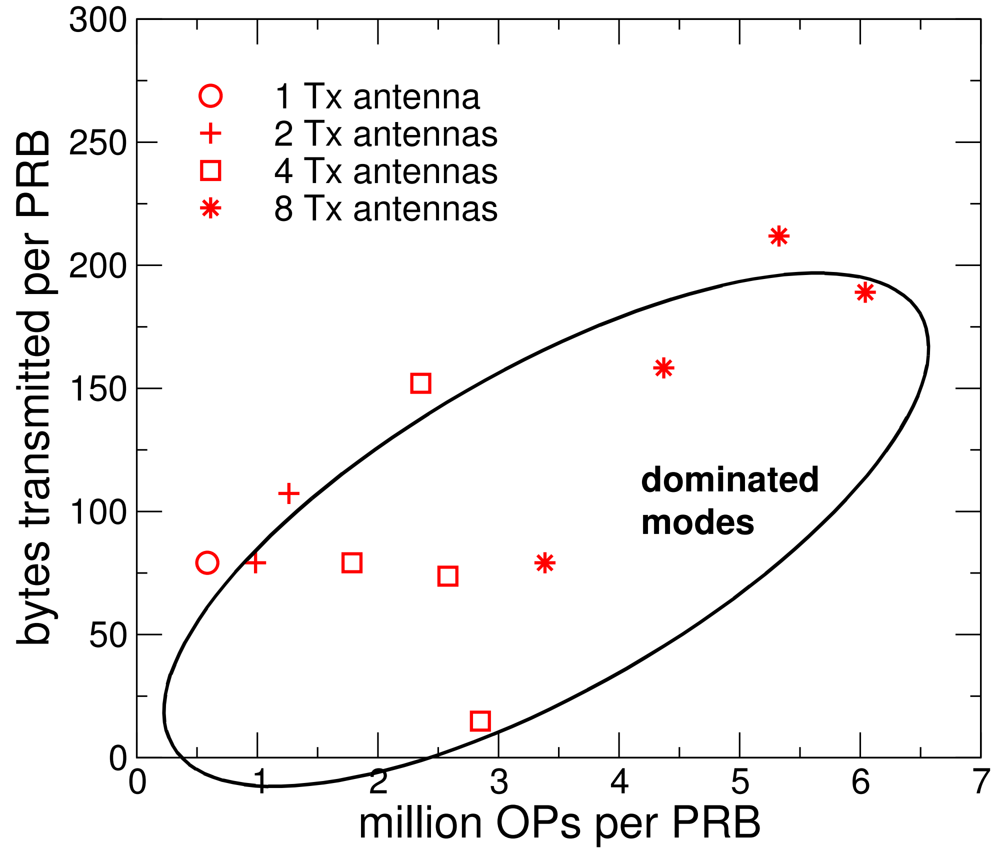


Proposed System

MIMO Mode Selection

- Mode selection can be seen as multiple-choice knapsack problem (see e.g. Kellerer et al., 2004)
- Some modes are dominated, i.e. there are others which provide more throughput with less effort

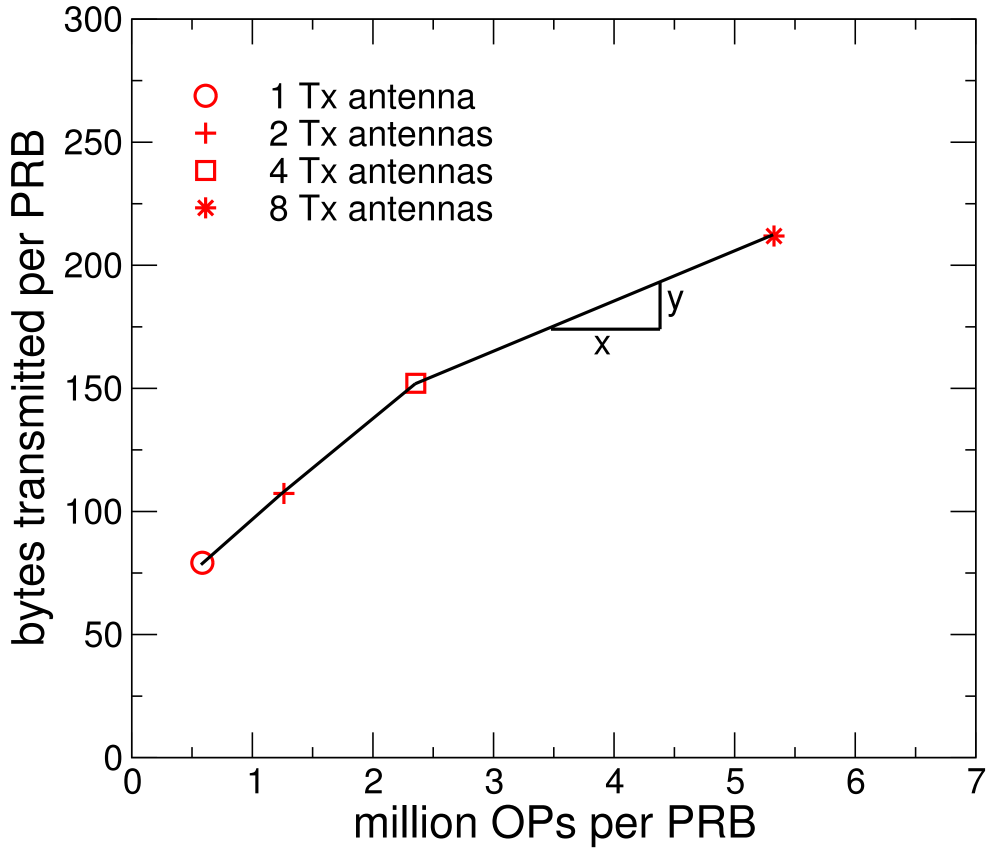
  → not considered further



Proposed System

MIMO Mode Selection

- Mode selection can be seen as multiple-choice knapsack problem (see e.g. Kellerer et al., 2004)
- Some modes are dominated, i.e. there are others which provide more throughput with less effort
  $\rightarrow$ not considered further
- Incremental efficiency defined as additional throughput divided by additional compute effort



Approach

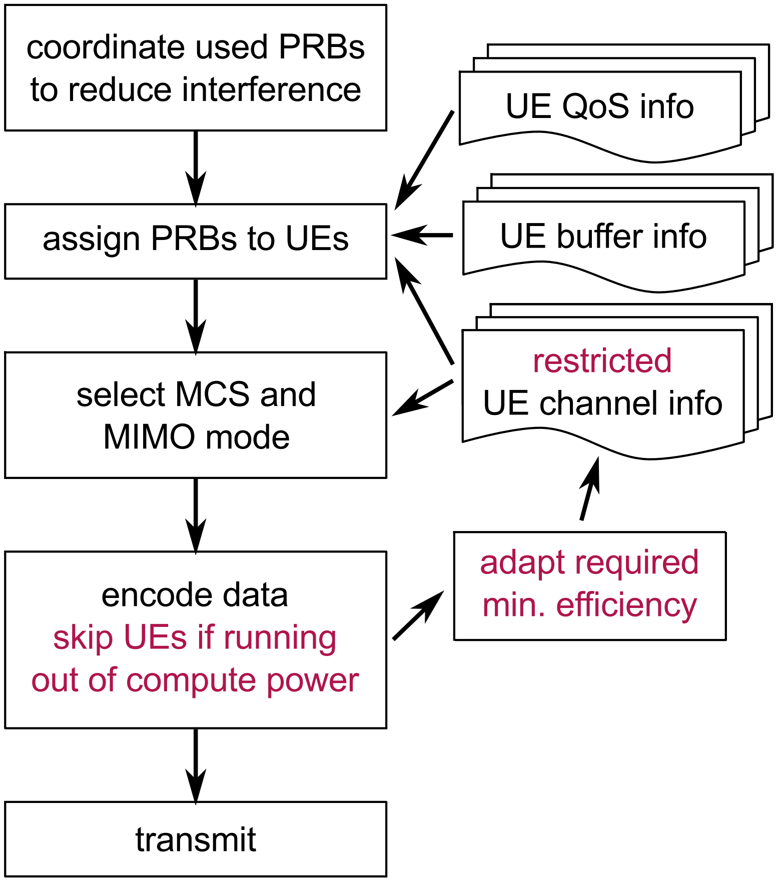
- Select mode upgrades with an efficiency higher than a threshold
- Use a simple control loop to set this threshold such that all resources are occupied

Proposed System

Integration with Scheduler

- Per-UE channel information is filtered based on MIMO mode efficiency

  → only efficient modes are visible
- Scheduler operates as before, but sees "worse channel conditions"
- Phy layer calculations just stop when the available compute time is over, remaining (random) UEs are skipped
- Feedback loop adjusts efficiency threshold, so that
  - skipping UEs happens infrequently
  - compute resources are used to capacity

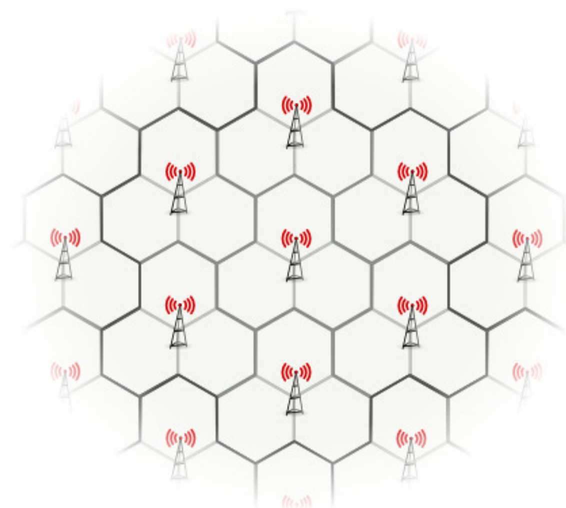


→ **Can this heuristic realize similar efficiency as the optimizer?**

Evaluation Model

Scenario

- 7 tri-sectorized sites; wrap-around
- BSs equipped with 8 antennas, UEs with 4 antennas



Time and Frequency Resources

- Grid of 50 PRBs x 100 subframes
- Multiple repetitions to avoid transient effects in simulation

Channel

- Channel correlation according to SCM (3km/h)
- No IfCo, i.e. all BSs transmit on all resources
- Ideal channel knowledge for all MIMO modes

Processing Effort

Downlink coding, modulation, and MIMO precoding

Traffic & Scheduler

Full buffer, proportional fair

$$P = \frac{R}{10} \cdot \left(3A + AL + \frac{LMC}{3}\right)$$

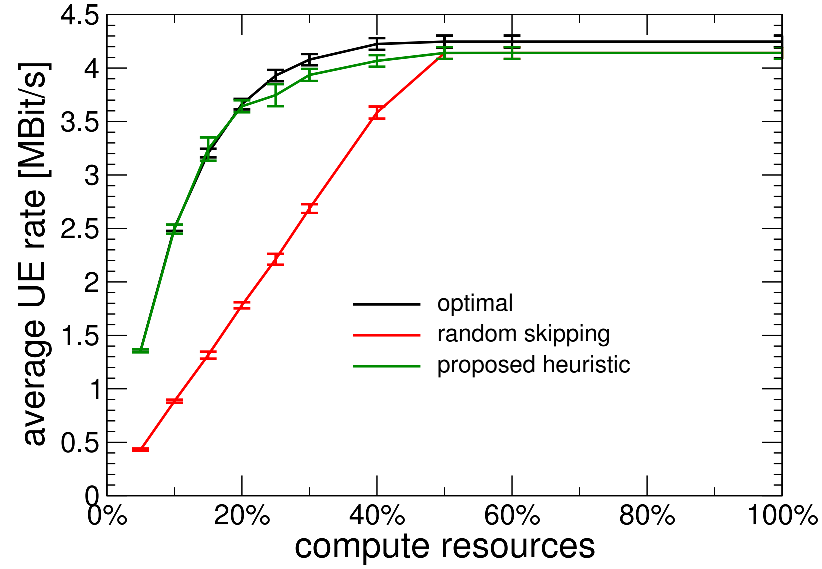
P compute effort in GOPS
R number of PRBs
A number of BS antennas
L number of MIMO layers
M modulation (bits per sym.)
C FEC Code Rate

[C. Desset et al, 2012: Flexible Power Model for LTE Base Stations]

Evaluation of the Heuristic

Evaluation

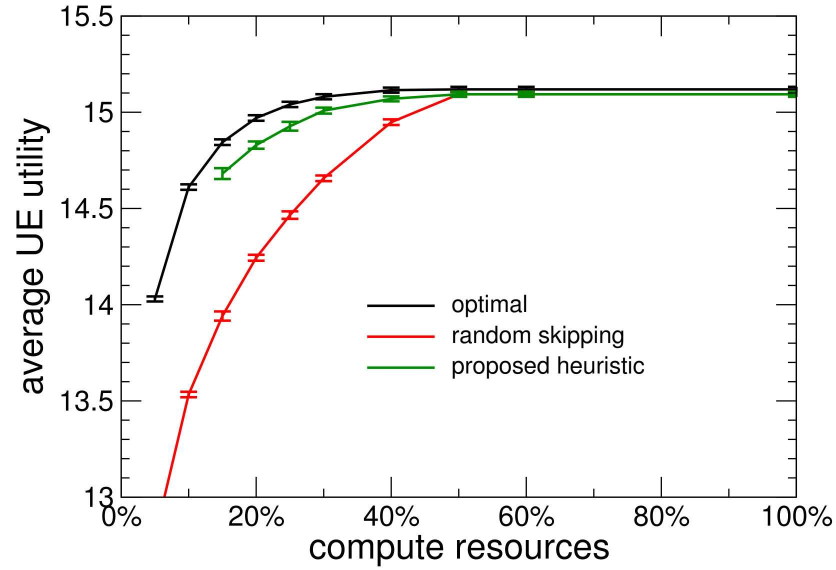
- Compare heuristic with
  - adapted optimizer, simultaneously allocating resources for all base stations and 100 TTIs
  - trivial heuristic:
    skip UEs if running out of compute resources
- Metrics: UE rate and prop. fair utility (log rate)



Simulation Results

- With unrestricted compute resources, scheduler achieves nearly optimal performance
- Random skipping results in linear decrease
- Proposed heuristic shows only slight deviation from optimal solution

→ **Heuristic can maintain acceptable performance**



© 2015 Universität Stuttgart • IKR

Conclusion

Summary

- Elastic scaling of compute requirements beneficial for C-RAN implementation
- Optimizer has shown that system performance degrades gradually
- Heuristic adapts MIMO mode selection
- Performance close to optimum

Conclusion

Elastic scaling of Phy layer compute effort is possible and efficient

Next Steps

- Evaluation in dynamic traffic scenario
- Publications & Thesis