Applying NOMA for Latency Reduction in Factory of the Future

Hung-Yun Hsieh

Graduate Institute of Communication Engineering &
Department of Electrical Engineering

National Taiwan University

June 19, 2019

Outline

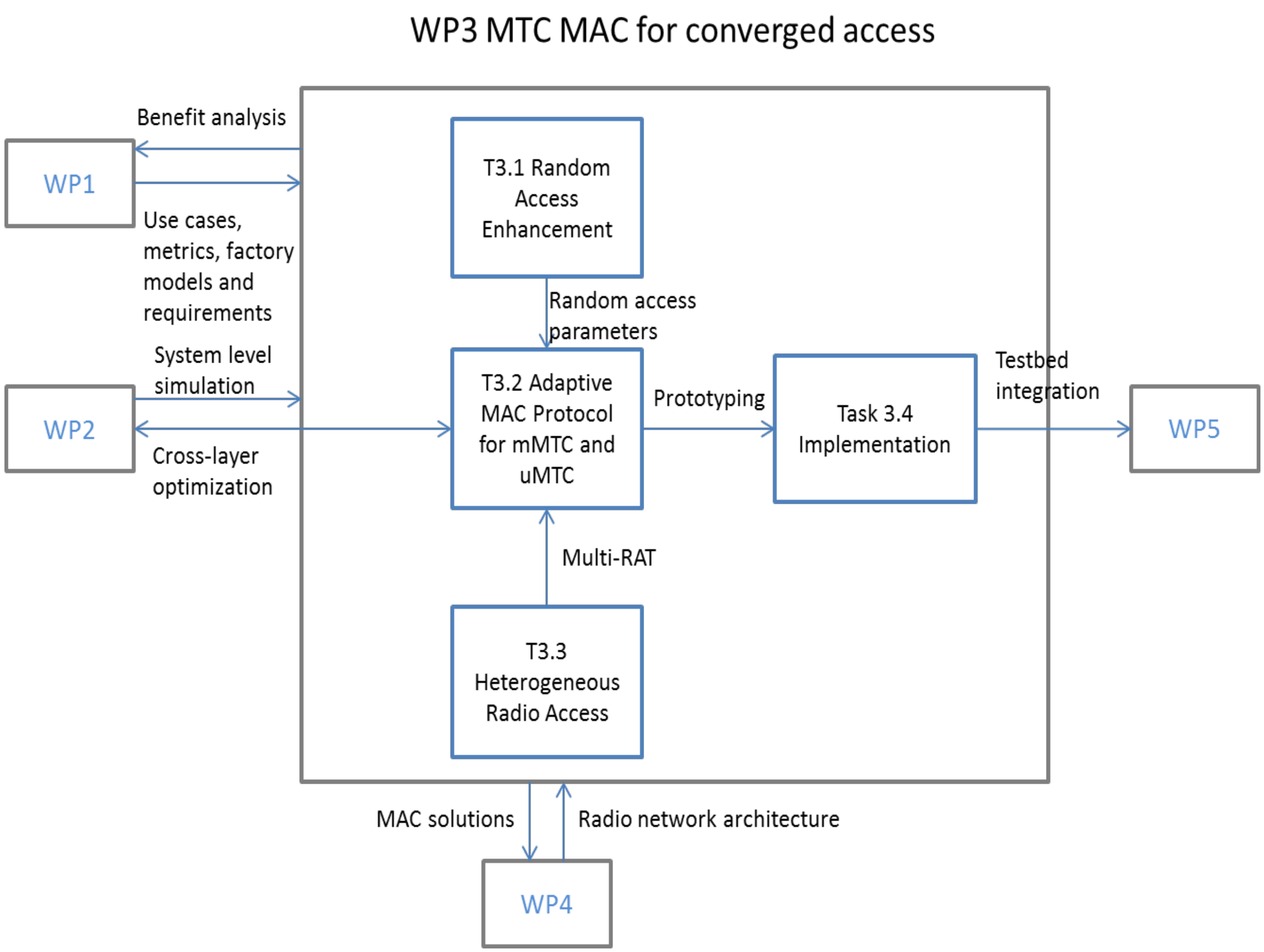
- Introduction
- III Contributions
- NTU Contributions
- Summary

Clear5G – WP3

- WP3 provides MAC-layer air interface enhancements for the MTC in the FoF use cases, including both the control plane and the user plane.
- WP3 investigates cross-layer (PHY and MAC) IoT traffic management and the impact of heterogeneous radio networks.

- Task 3.1: Random Access Enhancement
- Task 3.2: Adaptive MAC Protocol for mMTC and uMTC
- Task 3.3: Heterogeneous Radio Access
- Task 3.4: Implementation

Clear5G – WP3


WP3 MTC MAC for converged access
Benefit analysis
WP1
Use cases, metrics, factory models and requirements
System level simulation
WP2
Cross-layer optimization
T3.1 Random Access Enhancement
Random access parameters
T3.2 Adaptive MAC Protocol for mMTC and uMTC
Prototyping
Task 3.4 Implementation
Testbed integration
WP5
Multi-RAT
T3.3 Heterogeneous Radio Access
MAC solutions
Radio network architecture
WP4

Clear5G – Task 3.2

- Task 3.2: Adaptive MAC protocol for mMTC and uMTC

- Task 3.2 proposes an adaptive MAC protocol, which works in a contention-like manner in low-load traffic conditions to reduce the access latency. Moreover, the proposed MAC protocol will work in a scheduling-like manner in high-load traffic conditions to increase the system throughput and reliability.

Clear5G – Roles of III and NTU

- III (task leader) contributed to the design of MAC scheduler for massive connection, in addition to exploring designing flexible MAC concepts for reconfiguration of the amount of resources for RACH, system signaling, scheduling of low-latency users, etc.

- NTU proposed designs of low latency random access for massive MTC, with the coexistence of multiple radio access technologies.

III Contributions (D3.2)

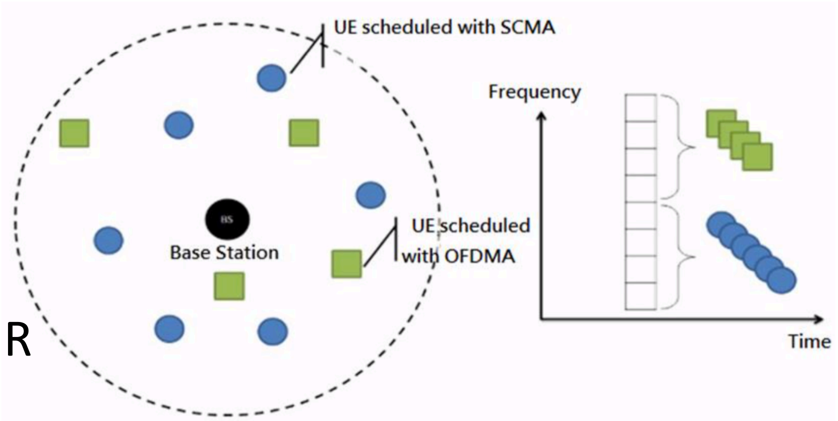
- III has designed a MAC-layer mechanism that performs resource allocation to users adaptively either in OMA or NOMA.
  - Code-domain NOMA (SCMA or LDS) and OMA (OFDMA or SC-FDMA) uplink multiple access schemes.
  - F candidate UEs to be allocated either in F resource blocks (OMA) or V>F resource blocks (NOMA)

$$S_{OMA} = \sum_{i=1}^{F} (O_i / O_{max}) * (1 - BLER_i)$$

$$S_{NOMA} = \mathrm{V} * \mathrm{O} * (1 - BLER_{worst}) / O_{max}$$

O_{max}: max modulation order

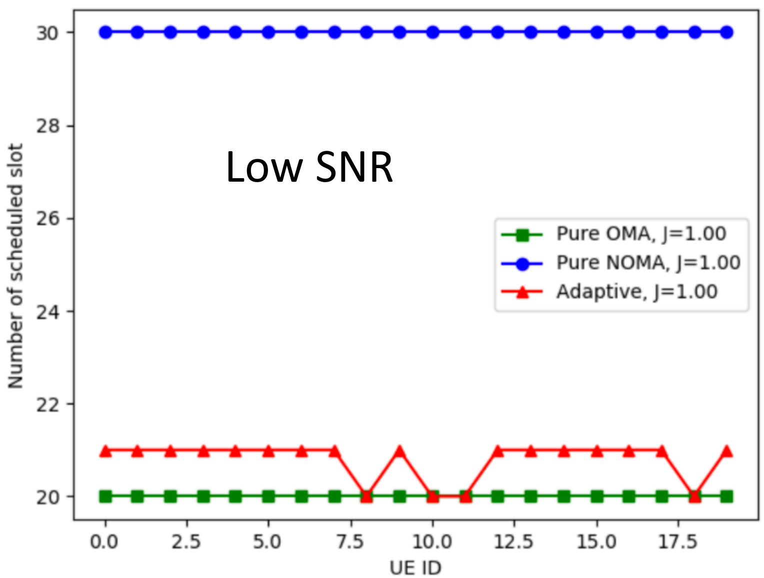
NOMA BLER based on the worst SNR

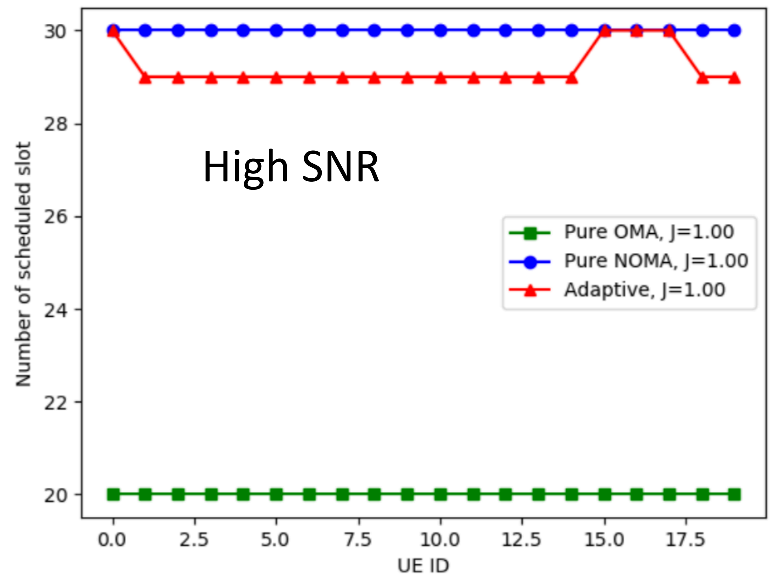


  - Choose the MA with the best throughput (S_{OMA} or S_{NOMA})

III Contributions (D3.2)

- Evaluation results with low and high SNRs
  - Low SNR: UE power = -20dBm
  - High SNR: UE power = -10dBm

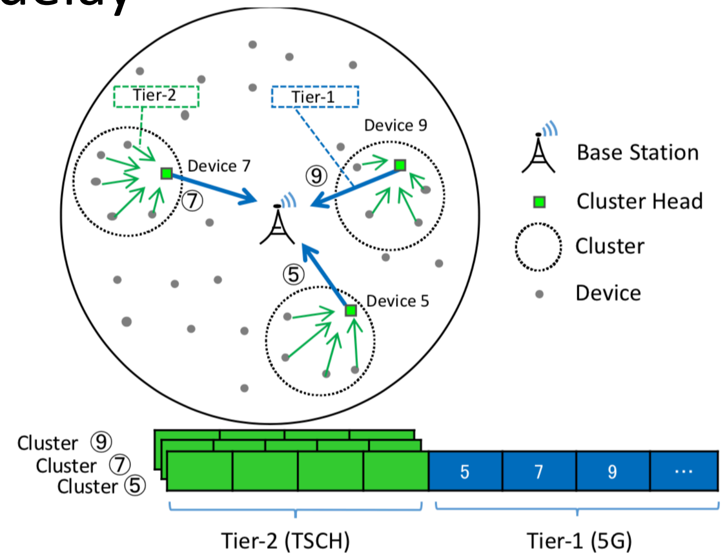




- Latency of the proposed mechanism can always achieve the lowest irrespective of the conditions

NTU Contributions (D3.2)

- Two-tier architecture with heterogeneous radios
  - Tier-1 (to/from base station): 5G
  - Tier-2 (between IIoT devices): short-range radio (6TiSCH)
- Latency reduction
  - Pure 5G -> two-tier: reduce bottleneck at the base station
  - Pure 6TiSCH -> reduce multi-hop delay
- Use NOMA at tier-1 can further reduce latency
- Some interesting results for IIoT devices



Data Characteristics of IIoT Devices

- IIoT devices are deployed to collectively gather (or report) data required by the target application
  - Collected data by individual devices is often *related*
  - Conventionally traffic from individual UEs are considered independent
- Contention of radio resource
  - Conventionally, it matters to provide QoS to each UE
  - For IIoT devices, it is possible that they contend resource to transmit same (similar) data
  - Does it make sense to provide QoS to each IIoT devices without considering the data they carry?
- Consider the problem of pairing NOMA UEs

Sum-Rate Maximization Scheduling

- Conventional approach for NOMA user pairing is to maximize the sum data rate of the pair

$$\mathcal{R} = \sum_{k=1}^{K} R_k$$

  machines are indexed in decreasing channel gain

- Achievable rate of each machine

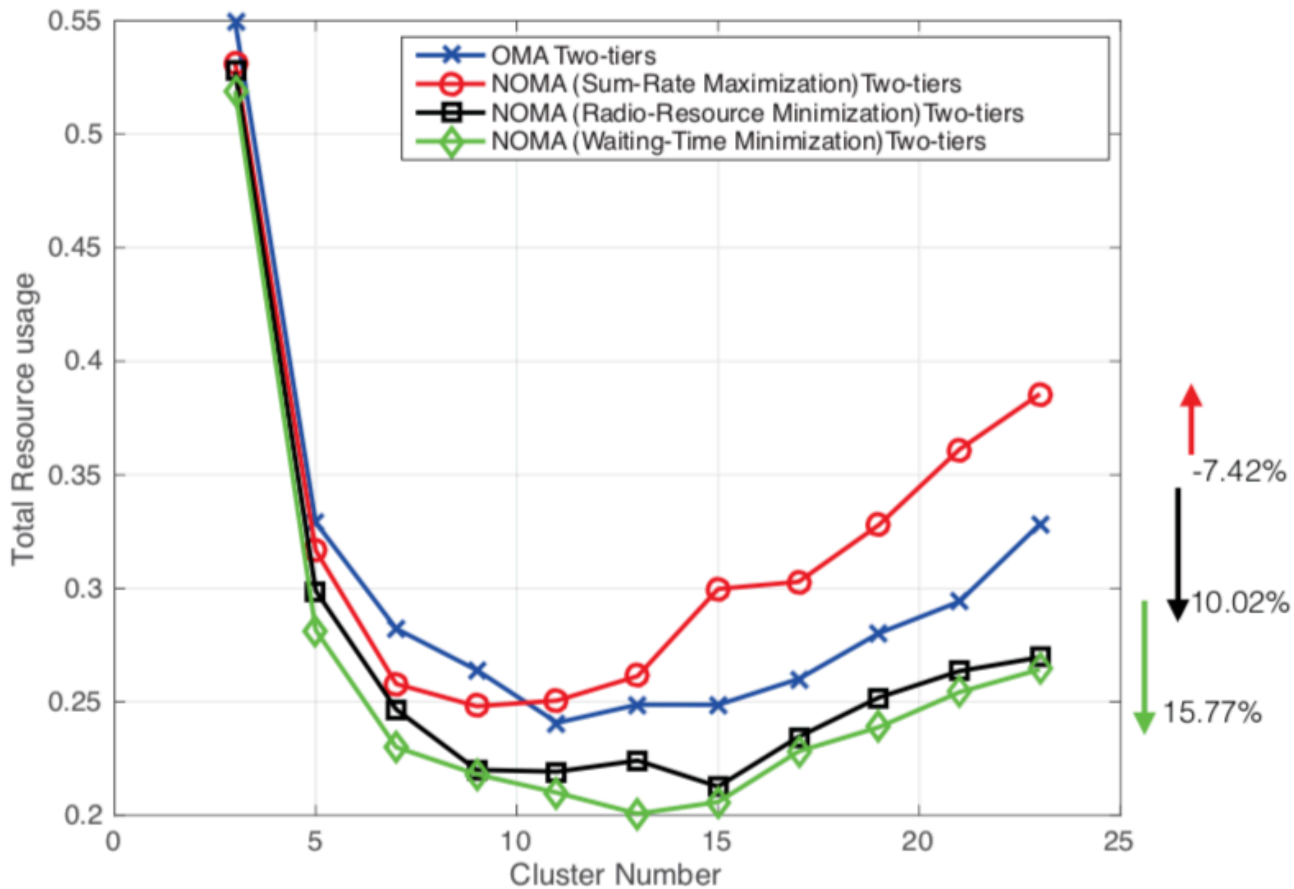
$$R_k = W \log_2 \left(1 + \frac{P G_{k0}}{\Gamma \left(W N_0 + \sum_{i=k+1}^{K} P G_{i0} \right)} \right)$$

  P is the transmission power of each machine

  Γ accounts for the spectral gap to Shannon capacity

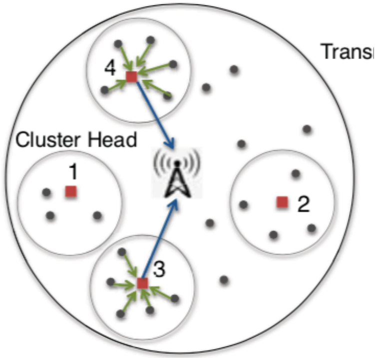
Overall Comparison

- Comparing different NOMA scheduling methods



Resource Minimization Scheduling

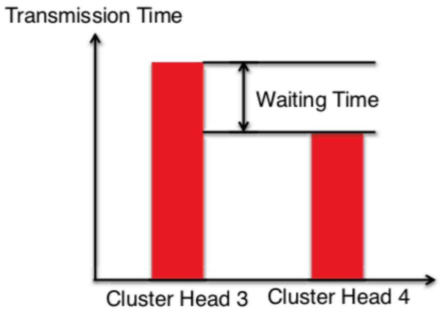
- Sum rate maximization is typically good for backlogged traffic sources (e.g. FTP), yet is it the case to M2M communications used for data gathering?
- Resource (time) needed for machine k



$$t_k = \frac{H(\mathcal{C}_k)}{R_k}$$

$H(\mathcal{C}_k)$ is the size of data

- Required resource for a given pair



$$\mathcal{T} = \max_{1 \le k \le K} t_k$$

- "Faster" machine needs to wait for the slower

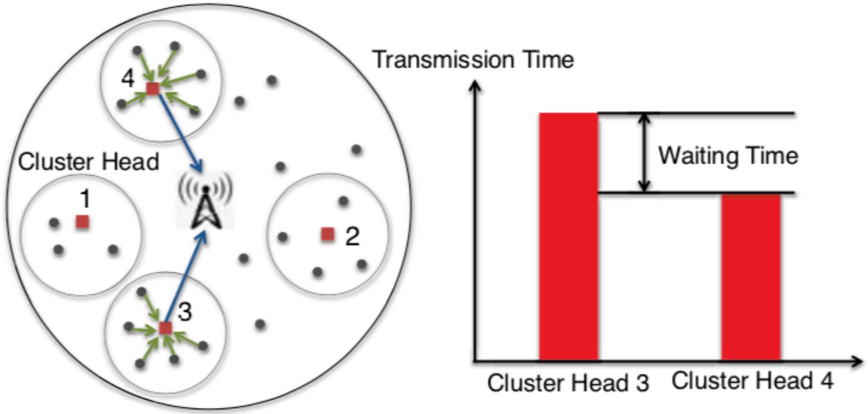
Waiting-Time Minimization

- Waiting time inside a time slot indicates a waste of radio resource

$$\mathcal{G} = \max_{1 \leq k,j \leq K} t_k - t_j$$

- Maximizing resource utilization is beneficial to resource minimization
- Scheduling metric

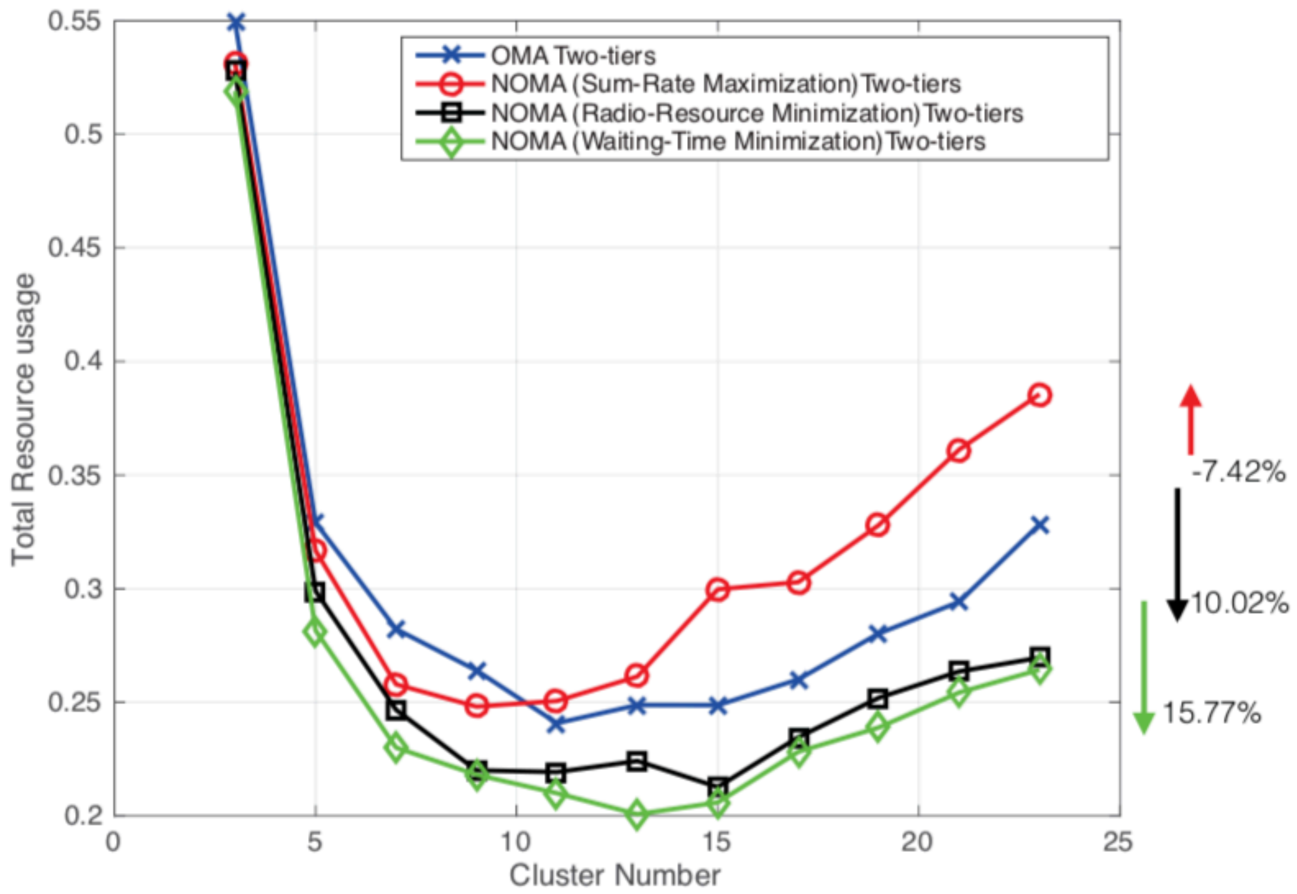
$$\mathcal{W} = \max_{1 \leq k \leq K} t_k + \alpha \mathcal{G}$$



- Scheduling pairs in decreasing metric value

Overall Comparison

- Comparing different NOMA scheduling methods



Two-User NOMA Model

- Consider the case with two NOMA users

Symbols	Definition
α	The ratio of data of node 2 to node 1, $\frac{H_2}{H_1}$
β	The ratio of signal of node 2 to node 1, $\frac{P_2G_2}{P_1G_1}$
P	The SNR of node 1, $\frac{P_1G_1}{WN_0}$

OMA

$$R_1 = W\log_2(1+P),$$

$$R_2 = W\log_2(1+\beta P),$$

$$T = \frac{H_1}{W\log_2(1+P)} + \frac{\alpha H_1}{W\log_2(1+\beta P)},$$

NOMA

$$R_1' = W\log_2(1+\frac{P}{1+\beta P}),$$

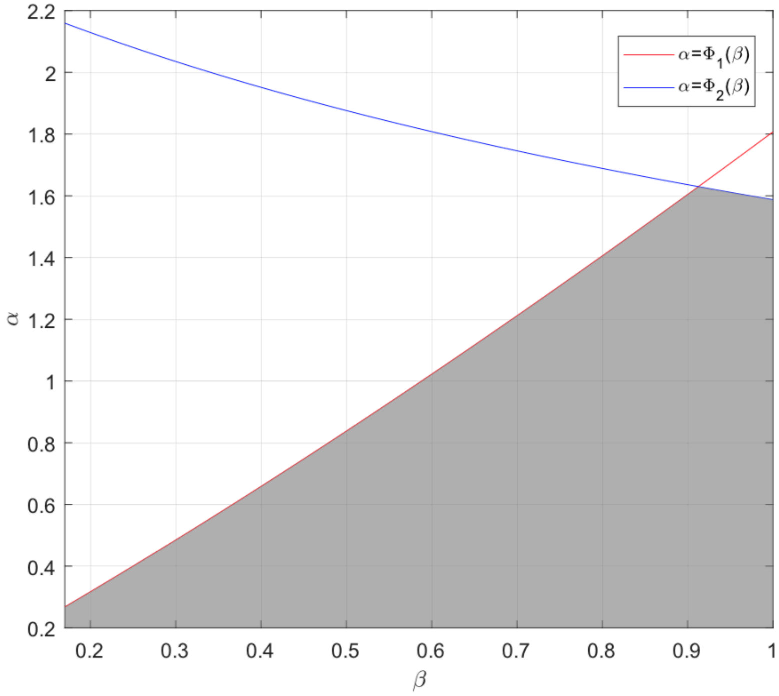
$$R_2' = W\log_2(1+\beta P),$$

$$T' = \max\left(\frac{H_1}{W\log_2(1+\frac{P}{1+\beta P})}, \frac{\alpha H_1}{W\log_2(1+\beta P)}\right)$$

- When will NOMA perform worse than OMA (T’>T)?

Two-User NOMA Model

- The region where T’>T (NOMA is worse)



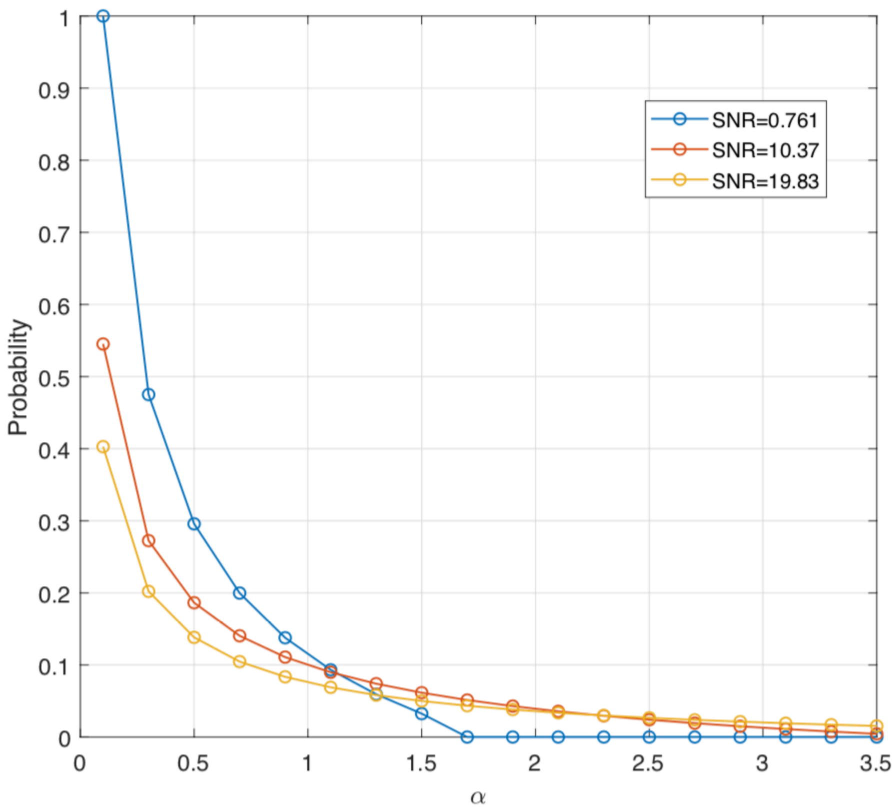
$$\alpha < \min\left(\Phi_1(\beta), \Phi_2(\beta)\right),$$

$$\Phi_1(\beta) = \frac{\log_2(1+\beta P)}{\log_2(1+\frac{P}{1+\beta P})},$$

$$\Phi_2(\beta) = \frac{\log_2(1+\beta P)}{1/\log_2(1+\frac{P}{1+\beta P}) - 1/\log_2(1+P)}.$$

Two-User NOMA Model

- Node 2 is randomly scattered while Node 1 is fixed
- Probability that NOMA is worse than OMA



SNR of Node 1

α is the data size ratio (H_2/H_1)

Summary

- Reducing latency in FoF using NOMA
- Adaptive MAC to choose the best multiple access scheme depending on the achievable throughput
- Two-tier architecture provides more flexibility with further reduction in end-to-end latency
- For IIoT devices with limited amount of data to send, throughput maximization does not necessarily lead to latency minimization
- Better user NOMA pairing and scheduling methods can be designed optimize latency performance in FoF