Enhancing Uplink Scheduling in 5G Enabled Vehicular Networks: A Cross-Layer Approach with Predictive Buffer Status Reporting

Veerendra Kumar Gautam, Venkatarami Reddy Chintapalli,
Bheemarjuna Reddy Tamma and C. Siva Ram Murthy

Networked **W**ireless **S**ystems **(NeWS)** Lab
Department of Computer Science & Engineering
Indian Institute of Technology Hyderabad

Contents

1. Introduction
2. System Model and Offloading Application
   - System Model
   - HD Map Task Offloading Application with OCTANE and RETALIN
3. Proposed Solution
   - Cross-layer framework and BSR prediction.
   - Grant-based UL procedure with BSR prediction.
   - Utility Metric
   - Scheduler running at MEC Server
   - Scheduler running at gNodeB
4. Performance Results
   - Performance Metrics
   - Offloading Success Rate Results
5. Conclusions

Motivation: High Definition (HD) Map over 5G NR V2X

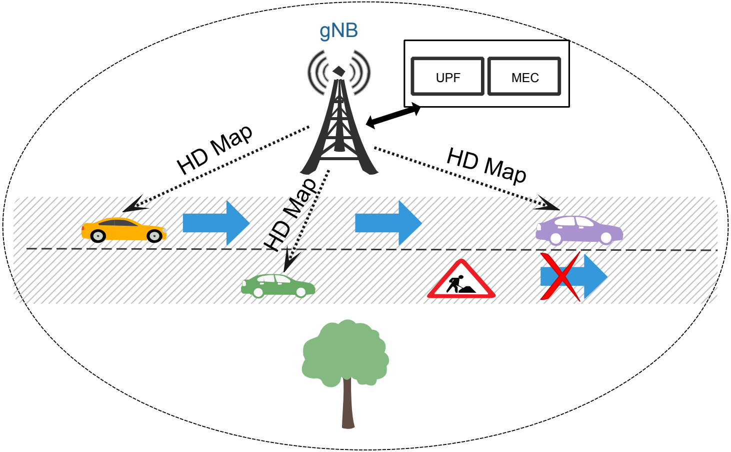


FIGURE 1: HD Map exchange between vehicles and MEC Server over 5G network

Challenges in supporting HD Map App over 5G NR V2X

- Vehicles must complete computational tasks within strict deadlines, even under varying channel conditions.
- The use of higher numerologies can lead to increased signaling overhead, particularly in terms of SRs in the UL, especially for applications with a heavy UL load.

Mobility-Aware Multi-User Offloading Optimization for Edge Computing[1]

Objective of the paper

- Computational offloading decision and computational resource allocation.
- They considered the trade-off between task latency and energy consumption.
- Vehicle mobility in the process of task offloading is considered in the optimization.
- In the paper, the authors designed constraints to account for resource limitations, user mobility, and task latency requirements. This optimization problem is formulated as a mixed-integer nonlinear programming (MINLP) problem.

[1] Reference: W. Zhan, C. Luo, G. Min, C. Wang, Q. Zhu and H. Duan, "Mobility-Aware Multi-User Offloading Optimization for Mobile Edge Computing," in IEEE Transactions on Vehicular Technology, vol. 69, no. 3, pp. 3341-3356.

Mobility-Aware Multi-User Offloading Optimization for Edge Computing[1]

Objective of the paper

- Computational offloading decision and computational resource allocation.
- They considered the trade-off between task latency and energy consumption.
- Vehicle mobility in the process of task offloading is considered in the optimization.
- In the paper, the authors designed constraints to account for resource limitations, user mobility, and task latency requirements. This optimization problem is formulated as a mixed-integer nonlinear programming (MINLP) problem.

Limitation

Jointly considering the radio and computational resources for offloading decision.

[1] Reference: W. Zhan, C. Luo, G. Min, C. Wang, Q. Zhu and H. Duan, "Mobility-Aware Multi-User Offloading Optimization for Mobile Edge Computing," in IEEE Transactions on Vehicular Technology, vol. 69, no. 3, pp. 3341-3356.

System Model

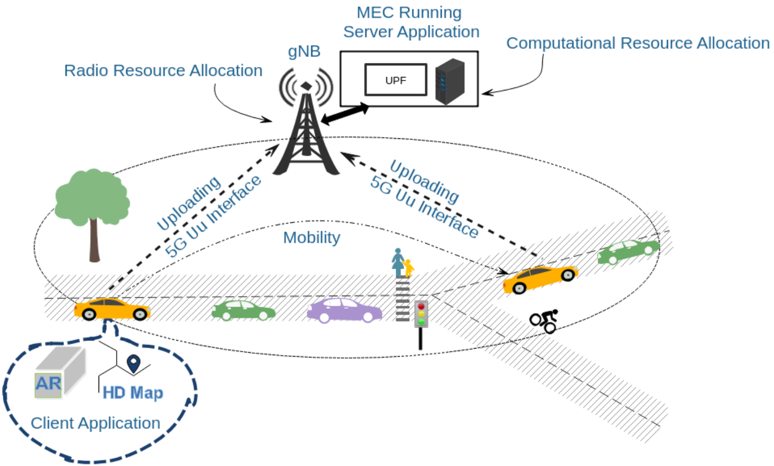


FIGURE 2: System model.

HD Map Task Offloading Application with OCTANE[2] and RETALIN[3]

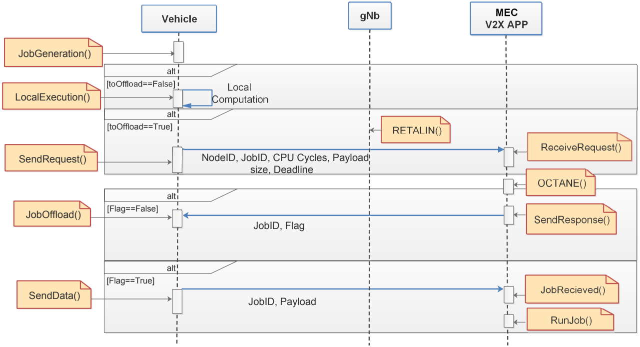


FIGURE 3: Sequence diagram of MapOffloading for HD MAP Use Case.

---

[2] Reference: "V. K. Gautam, C. Tompe, B. R. Tamma and A. F. A, "OCTANE: A Joint Computation Offloading and Resource Allocation Scheme for MEC Assisted 5G NR Vehicular Networks," 2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Hyderabad, India, 2021, pp. 60-65."

[3] Reference: V. K. Gautam and B. R. Tamma, "RETALIN: A Queue Aware Uplink Scheduling Scheme for Reducing Scheduling Signaling Overhead in 5G NR," in IEEE Access, vol. 12, pp. 16632-16651, 2024.

OCTANE: A Joint Computation Offloading and Resource Allocation Scheme for MEC Assisted 5G NR Vehicular Networks[2]

- We studied the joint job offloading decision, radio resource allocation, and computational resource allocation problem for latency-sensitive vehicular applications.
- We proposed OCTANE, which selects jobs for offloading by jointly considering deadlines, computational and communication delays of the jobs.
- We have used an HD Map application as a use case to study the effectiveness of OCTANE.

[2] Reference: "V. K. Gautam, C. Tompe, B. R. Tamma and A. F. A, "OCTANE: A Joint Computation Offloading and Resource Allocation Scheme for MEC Assisted 5G NR Vehicular Networks," 2021 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), Hyderabad, India, 2021, pp. 60-65."

RETALIN: A Queue Aware Uplink Scheduling Scheme for Reducing Scheduling Signaling Overhead in 5G NR[3]

- We proposed RETALIN, a two-phase UL grant-based scheme for radio resource allocation in 5G NR.
- RETALIN is capable of subsisting generation of Scheduling Request (SRs) in the network, thereby increasing PDR and reducing E2E delay of the UEs.
- RETALIN is capable of achieving a better trade-off between SR and BSR for higher numerologies with different packet sizes and traffic patterns.

[3] Reference: V. K. Gautam and B. R. Tamma, "RETALIN: A Queue Aware Uplink Scheduling Scheme for Reducing Scheduling Signaling Overhead in 5G NR," in IEEE Access, vol. 12, pp. 16632-16651, 2024.

Cross-layer framework and BSR prediction.

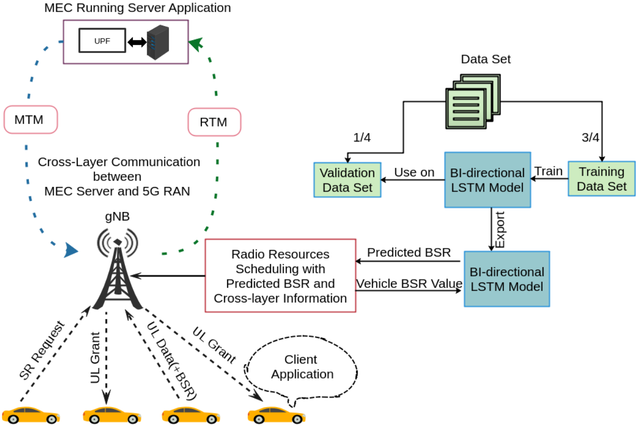


FIGURE 4: Cross-layer framework and BSR prediction.

Grant-based UL procedure with BSR prediction.

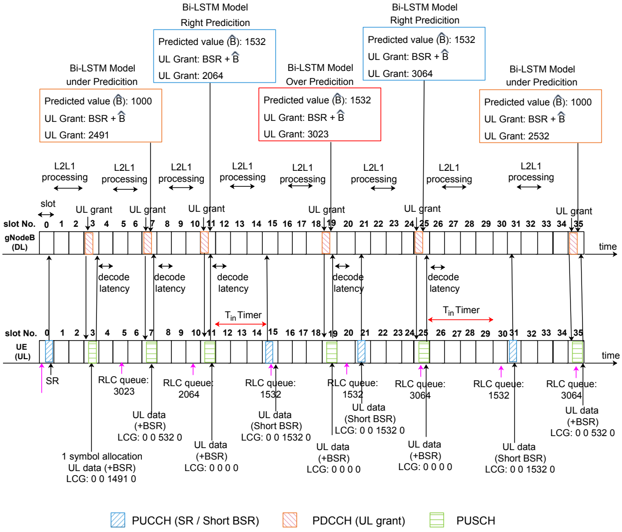


FIGURE 5: Grant-based UL procedure with BSR prediction in case of numerology 1 for packets of size 1500 bytes (with 32 bytes of header) with $IPAT = 5$ slots and $T_{in} = 5$ slots.

Utility Metric

Utility Metric $U_v(t)$

- We modified the utility metric $U_v(t)$ (defined earlier in *RETALIN*[3]), which is a PF variant. This modification involves incorporating the ranks of UEs, denoted as R_v, which are sent by the MEC scheduler using MTM messages and utilized for UL scheduling.

$$U_v(t) = \left[\frac{\alpha_v \times P_{v,NSR}(t) \times PF_v(t)}{R_v} \right] \tag{1}$$

- The probability of not generating an SR in a TTI is denoted by $P_{v,NSR}(t)$.
- The PF metric for UE v for a RB r in a TTI is represented as $PF_{v,r}(t)$.
- R_v refers to the ranking of vehicles.
- The normalized backlog ratio of UE v, denoted as (α_v), indicates the drift in the UL buffer length of UE v as compared to the aggregated UL buffer length of all UEs in a given TTI.

[3] Reference: V. K. Gautam and B. R. Tamma, "RETALIN: A Queue Aware Uplink Scheduling Scheme for Reducing Scheduling Signaling Overhead in 5G NR," in IEEE Access, vol. 12, pp. 16632-16651, 2024.

Scheduler running at MEC Server

ALGORITHM 1: **Scheduler running at MEC Server**

```
inputs : V_allJobs = {J_1, J_2, ..., J_V}, J_v = {j_1, j_2, ..., j_n}, (v ∈ 𝒱), V_SelectedJobs = {}, R_V = {}
if receive the message RTM then
    MCS_V ← extractMCSofVehicles(RTM)
    calculateRanking(V_allJobs, MCS_V)
else
    if new job request is received then
    else
        calculateRanking(V_allJobs, MCS_V)
    end
    GOTO line 1
end
calculateRanking V_allJobs, MCS_V
V_SelectedJobs ← OCTANE(V_allJobs, MCS_V)
JobSizeTotal ← 0, JobSize_V = {}
forall i ∈ V_SelectedJobs do
    JobSize_v ← sumOfJobs(i)
    JobSizeTotal ← JobSizeTotal + JobSize_v
    JobSize_V ← JobSize_V ∪ JobSize_v
end
forall v ∈ 𝒱 do
    R_v ← ⌈(JobSizeTotal − JobSize_v) / JobSizeTotal⌉
    R_V ← R_V ∪ R_v
end
Send MTM message with R_V to the RAN scheduler
```

Scheduler running at gNodeB

Algorithm 2: Scheduler running at gNodeB

inputs : $V_{\text{allBSR}} = \{BSR_1, BSR_2, \ldots, BSR_v, \ldots, BSR_V\}$,
$BSR_v = \left\{bsr_1, bsr_2, \ldots, bsr_{Bi-LSTM_w}\right\}, (v \in \mathcal{V})$
$R_V = \{\}, QueueSize_v^{pred} = \{\}, (\text{v} \in \mathcal{V})$

if *receive the message MTM* **then**
 $R_V \leftarrow \texttt{extractRanksOfVehicles}(MTM)$
 $MCS_V \leftarrow \texttt{callRETALIN}(V_{\text{allBSR}})$
 Send *RTM* message with MCS_V to the MEC scheduler
else
end
$\texttt{callRETALIN}(V_{\text{allBSR}})$
callRETALINV_{allBSR}
`// Predict BSR values using Bi-LSTM model`
forall $v \in V_{allBSR}$ **do**
 $\widehat{B} \leftarrow \texttt{Bi-LSTMpredict}(BSR_v)$
 $QueueSize_v^{pred} \leftarrow (bsr_{Bi-LSTM_w} + \widehat{B})$
end
$MCS_V \leftarrow \texttt{RETALIN}(R_v, QueueSize_v^{pred})$
Return MCS_V

System analysis of QoS schedulers for XR traffic in 5G NR (Compared Work)[4]

Objective

- QoS scheduler introduces a delay budget factor (D) that represents the weight sensitive to delay, considering HOL delay and Packet Delay Budget (PDB). The calculation for this factor is given as D = PDB / (PDB - HOL).
- QoS scheduler utilizes multiple factors including the default priority level of the flow, the PF metric and D to allocate the radio resources.

[4] Reference: K. Koutlia, B. Bojovic, S. Lagen, X. Zhang, P. Wang, and J. Liu, "System analysis of QoS schedulers for XR traffic in 5G NR," Simulation Modelling Practice and Theory, vol. 125, p. 102745, 2023.

NS-3 Simulation Parameters

TABLE 1: Simulation Parameters

Parameter	**Value**
Scenario	Urban Macro Cell
Number of Vehicles $\lvert\mathcal{V}\rvert$	30
Mobility Model	Krauss
Average Vehicle Velocity ($\mathcal{V}_{vel}$)	60 kmph
5G NR Base Station/Vehicle TX power	46/23 dBm
5G NR Base Station Antenna Pattern	Canadian dataset
5G NR Base Station Antenna Tilt	15°
5G NR Base Station/Vehicle Antenna Height	25 m / 1.5 m
Carrier Frequency	6 GHz
Channel Model	3GPP, Line-Of-Sight
Channel Bandwidth	30 MHz
5G NR Numerology μ	0, 1, 2
Channel model	UMa_LoS
MEC Task scheduler	*OCTANE*
5G NR MAC Scheduler	*RETALIN*, PF, QoS
5G QoS Identifier (5QI)	75, GBR_V2X
$Bi-LSTM_w$	50
Packet size (L)	1000 Bytes
Job generation per vehicle	0.1 sec

Performance Metrics

Performance Metrics

- Offloading Success Rate (OSR) is calculated at MEC server for different schemes in a vehicular scenario. Success is accomplished when job is offloaded to the MEC server and completed within its deadline. If a vehicle decides to run a job locally, it is not counted as failure or success.

$$OSR = \left[\frac{NumberOfJobsDone}{TotalNumberOfRequestsReceivedByMECServer} \right] \times 100 \quad (2)$$

- SR_p: It the number of SR request generated over the total number of packets.

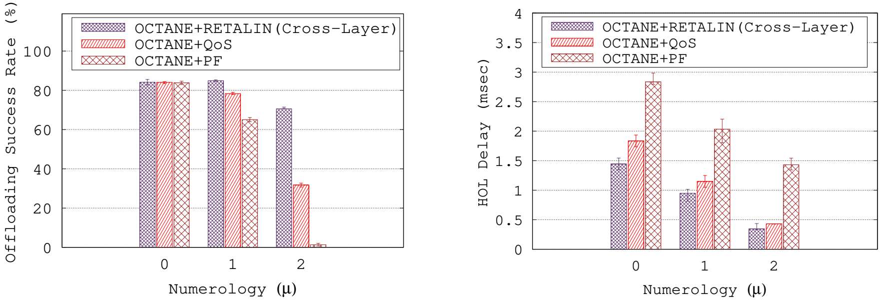
$$PercentageofSR = \left[\frac{NumberofSRrequest}{TotalnumberofPackets} \right] \times 100 \quad (3)$$

- BSR_{avg}: BSR_{avg} is the average number of BSR messages used to transmit a packet in a network.

$$PercentageofBSR = \left[\frac{NumberofBSRrequest}{TotalnumberofPackets} \right] \times 100 \quad (4)$$

- *Head-Of-Line (HOL) Delay*: It represents the delay experienced by a packet at the head of a queue waiting to be transmitted. HOL delay is a measure of the time a packet spends in a UE queue before it is transmitted.

Performance Results



((a)) OSR. ((b)) HOL Delay.

FIGURE 6: Result observed for HD Map application by varying numerology for $V = 30$ with $V_{speed} = 60\ kmph$ where $L = 1000$ bytes and $Bi - LSTM_w = 50$.

Observations

- As we increase the Numerology, Offloading Success Rate (OSR) decreases.
- OCTANE+RETALIN (Cross-Layer) is able to achieve 19% higher OSR as compared to OCTANE+QoS.
- OCTANE+RETALIN (Cross-Layer) is able to achieve 25% reduction in Head-Of-Line(HOL) delay as compared to OCTANE+QoS.

Performance Results

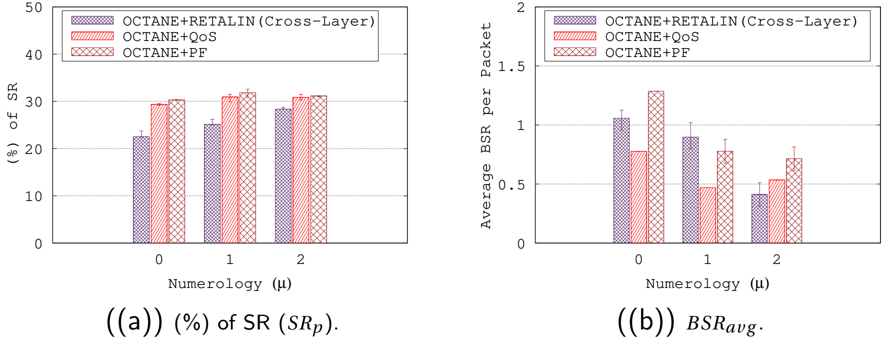


FIGURE 7: Result observed for HD Map application by varying numerology for $V = 30$ with $V_{speed} = 60\ kmph$ where $L = 1000$ bytes and $Bi - LSTM_w = 50$.

Observations

- SR_p increases for Numerology 2 for OCTANE+QoS. But OCTANE+RETALIN (Cross-Layer) is able to reduce SR_p as compared to OCTANE+QoS.
- BSR_{avg} for OCTANE+RETALIN (Cross-Layer) increases but BSR_{avg} doesn't effect E2E delay.

Conclusions

- Proposed cross-layer framework that facilitated information exchange between RAN and MEC schedulers, leveraging ranking and channel condition data for efficient task offloading in case of V2X applications.
- Bi-directional LSTM, trained on the Berlin V2X dataset, enhanced the model's capability to learn traffic inter-arrival patterns and predict future UL grants, assisting the RAN scheduler.

Thank You!

Email: tbr@cse.iith.ac.in

Visit NeWS Lab @ IIT Hyderabad for more details: https://newslab.iith.ac.in/