IEEE Hong Kong 6G Wireless Summit

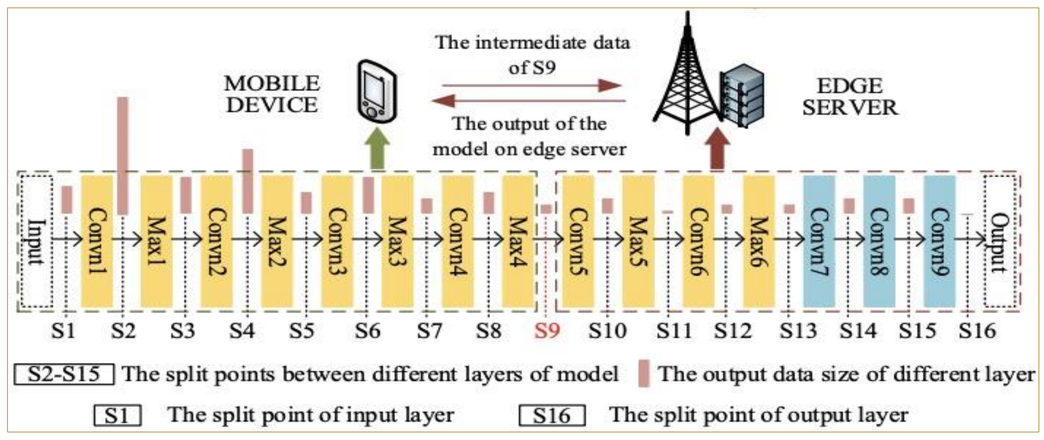
11-12 September 2024
Hong Kong, China

IEEE ComSoc IEEE Communications Society | IEEE

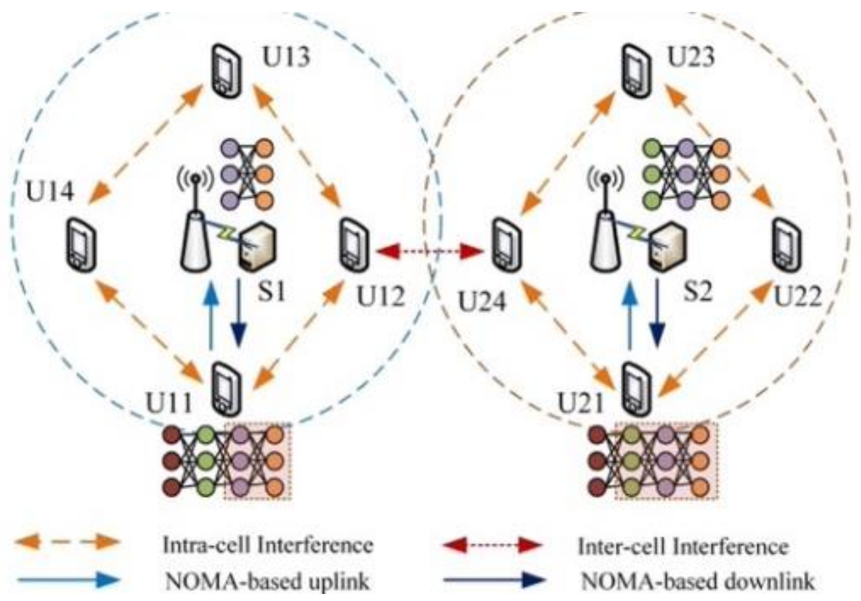
High Efficiency Inference Accelerating Algorithm for NOMA-based Edge Intelligence

The main contributions of this paper can be summarized as follows.

1. In this study, we integrate NOMA into the model split inference in EI. To the best of our knowledge, this is the first work to investigate the possibility and approach of using NOMA technology to improve the performance of model split inference in EI under a multi-user scenario. Moreover, the challenges and issues caused by integration are also discussed in this paper.
2. Since NOMA has a significant effect on model split point selection and energy consumption, we take both the energy consumption and inference delay into account to find the optimal model split strategy and resource allocation strategy (computing resource, channel resource, transmission power) for NOMA-based split inference in EI. Moreover, because the minimum energy consumption and minimum inference latency cannot be achieved simultaneously, the GD-based algorithm is adopted in this study to effectively achieve an optimal tradeoff between them.
3. Moreover, considering the complexity of this issue caused by uneven and discrete intermediate data size, we propose a Li-GD algorithm to improve the efficiency of the GD procedure. The key idea of the Li-GD algorithm is that: the initial value of the ith layer's GD procedure is selected from the optimal results of the former $(i-1)$ layers' GD procedure whose intermediate data size is the closest to the ith layer.
4. The properties of the proposed Li-GD algorithm are investigated. The Li-GD algorithm is convergent, and the convergence time is $K=\frac{\|x^0-x^*\|_2^2}{2\eta\epsilon}$,the complexity of the Li-GD is $O(XK\mathcal{F}Mx^3\ln^2(x))$, the approximate error is smaller than $\frac{\varepsilon}{\rho_{min}(1-B_{max})\log_2\left(1+\frac{P_{min}}{\Delta^*+\frac{\alpha P_{max}}{2}}\right)}$. Additionally, it can reduce the complexity and convergence time compared with the traditional GD approach.



This is an example of YOLOv2



Network model

OPTIMAL MODEL SPLIT AND RESOURCE ALLOCATION ALGORITHM

A. Loop iteration GD algorithm

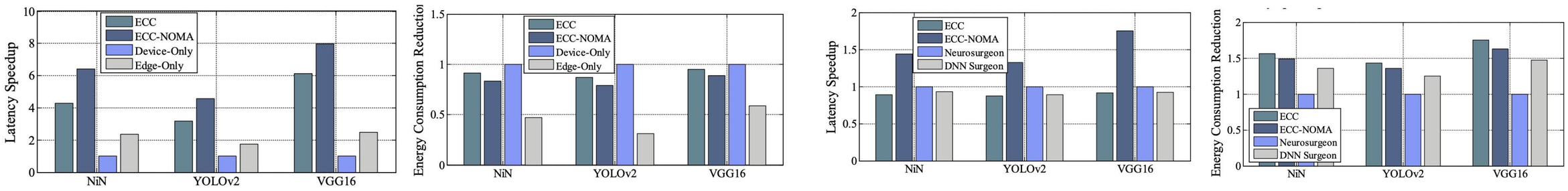
Since the optimization objectives shown in P0 are opposite, we introduce the weight-based approach to construct the utility function for each mobile user that contains both these objectives, which can be expressed as: $U_i=\omega_T T_i+\omega_E E_i$ where ω^T and ω^E are the weights of inference delay and energy consumption, respectively, and $\omega_T+\omega_E=1$. The weight represents the importance of each optimal objective to users.

B. The properties of Li-GD algorithm

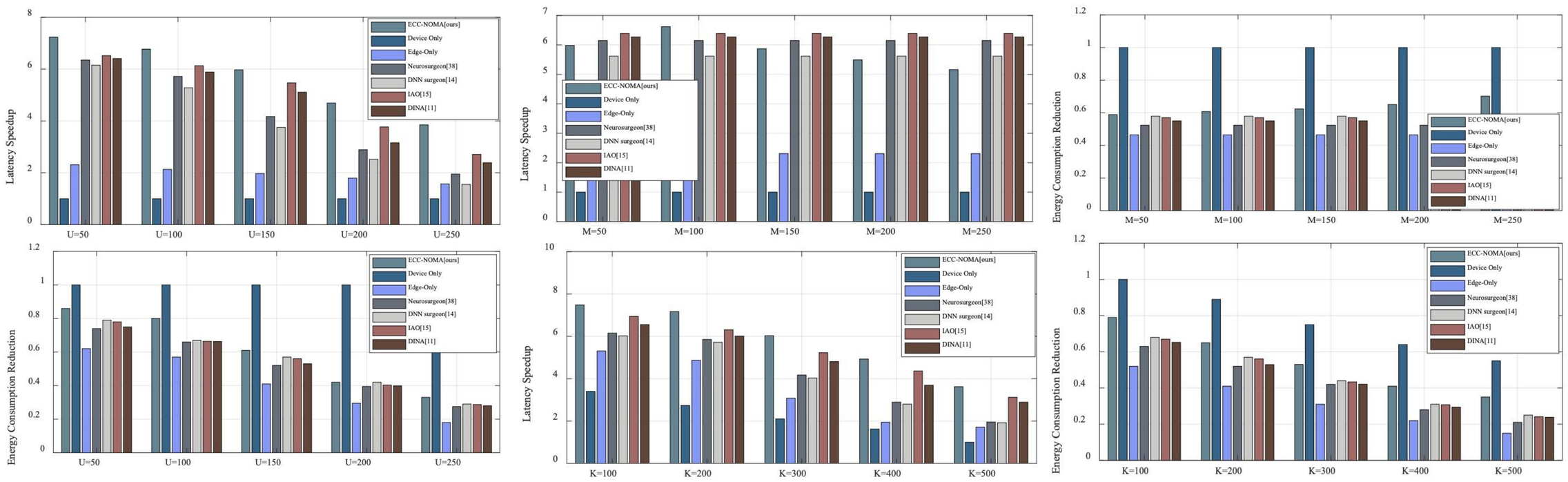
Corollary 1. When the values of f_l^i, f_e^i, and w_{s_i} are know in advance, and we loose the constraints of $\beta_{n,i}^m \in \{0,1\}$ and $\beta_{j,i}^k \in \{0,1\}$ to $\beta_{n,i}^m \in [0\ 1]$ and $\beta_j, i^k \in [0\ 1]$, the utility function shown in before is differentiable.

Corollary 2. The Li-GD algorithm is convergent, and the convergence time $K=\frac{\|x^0-x^*\|_2^2}{2\eta\epsilon}$, where η is the step size and $\eta \le \frac{1}{L}$, ϵ is the threshold of accuracy.

PERFORMANCE EVALUATION



PERFORMANCE UNDER DIFFERENT NETWORK CONDITIONS



ACKNOWLEDGMENT

This work was supported in part by the Research Grants Council of Hong Kong under the Areas of Excellence scheme grant AoE/E-601/22-R.