Prioritized Job Offloading in Cloudlet System for Smart City

Shibaura Institute of Technology
Yuki Yokota, Sumiko Miyata

Background and Objective

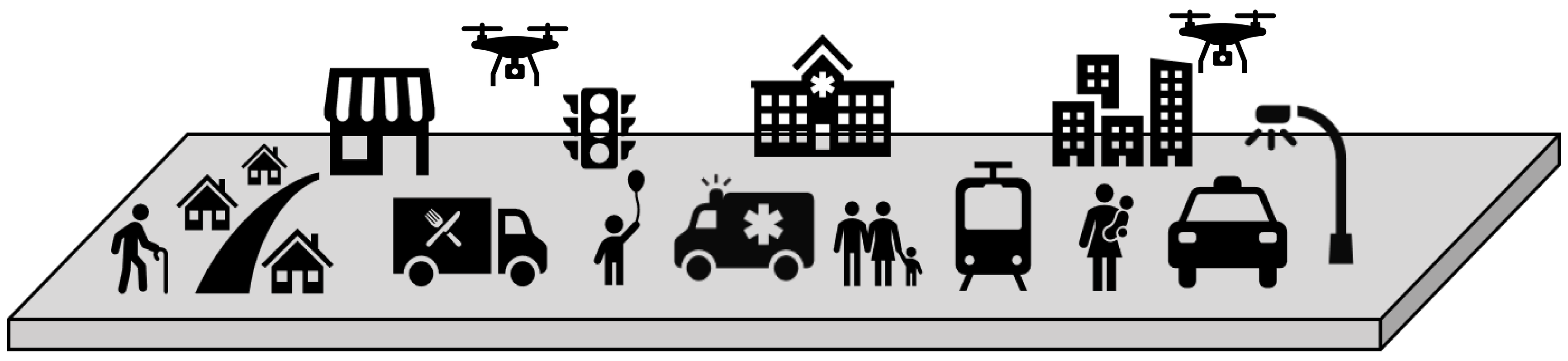
Background

・A smart city comprises processes with both strict and tolerant latency requirements[1].

・The use of cloudlets as one of the distributed processing for various computational processes is being researched[2].

☆ Edge server in a smaller scale with lower latency

➡ **Enables effective processing for latency sensitive applications**



Conventional researches

- Deadline-Constrained RSU-to-Vehicle Task Offloading Scheme for Vehicular Fog Networks [3]

➡ **Does not consider the variation of jobs in smart city.**

- Latency Focused Load Balancing Method between Cloudlets using Game theory [4]

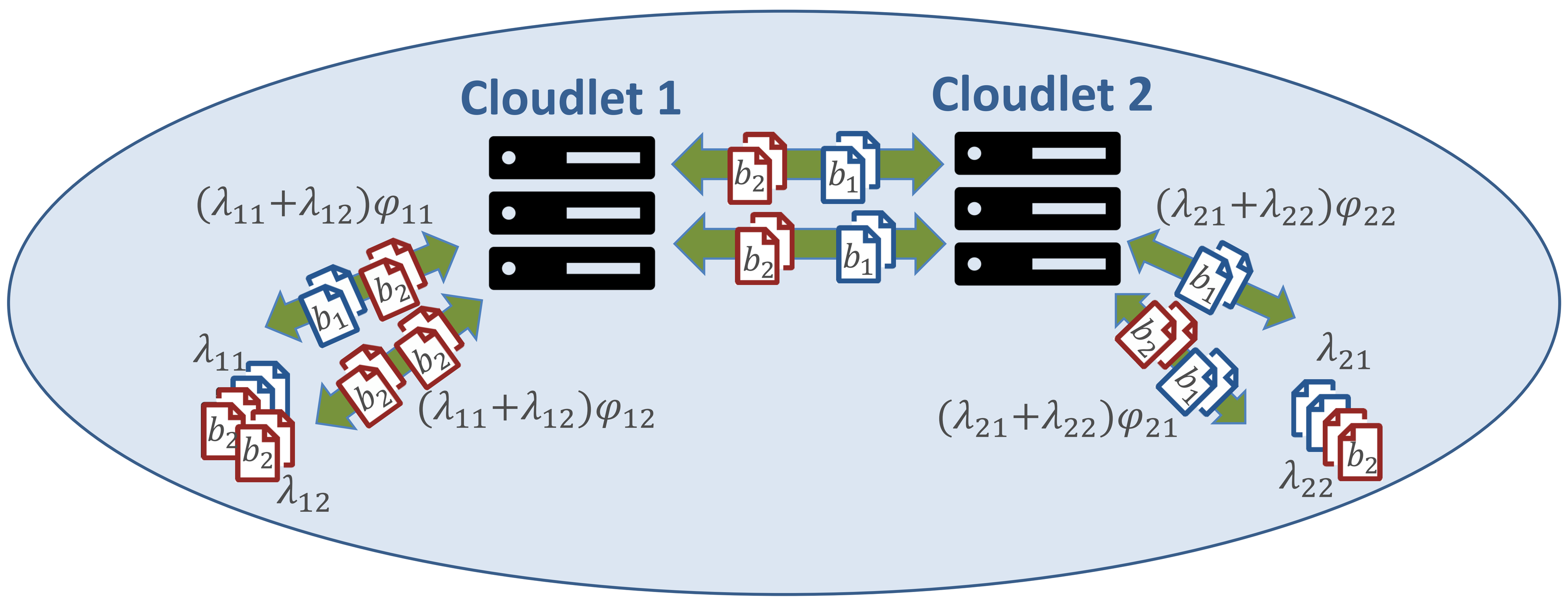
➡ **The main goal of the conventional research is to reduce latency for users and to balance the load between servers, and it does not focus on the latency minimization.**

Objective

Minimizing the average latency by efficient job offloading between cloudlets considering the priority of jobs in smart city.

Modeling by Queueing Theory

- Each cloudlet can be modeled by $M/G/1$



- Using the parameters above the latency of both prioritized and unprioritized jobs can be shown as below.

$$L_{\text{high priority}} = \frac{(\lambda_{11}+\lambda_{12})\varphi_{12}b_1^2 + (\lambda_{21}+\lambda_{22})\varphi_{22}b_2^2}{2(1-(\lambda_{11}+\lambda_{12})\varphi_{12}b_1)}$$

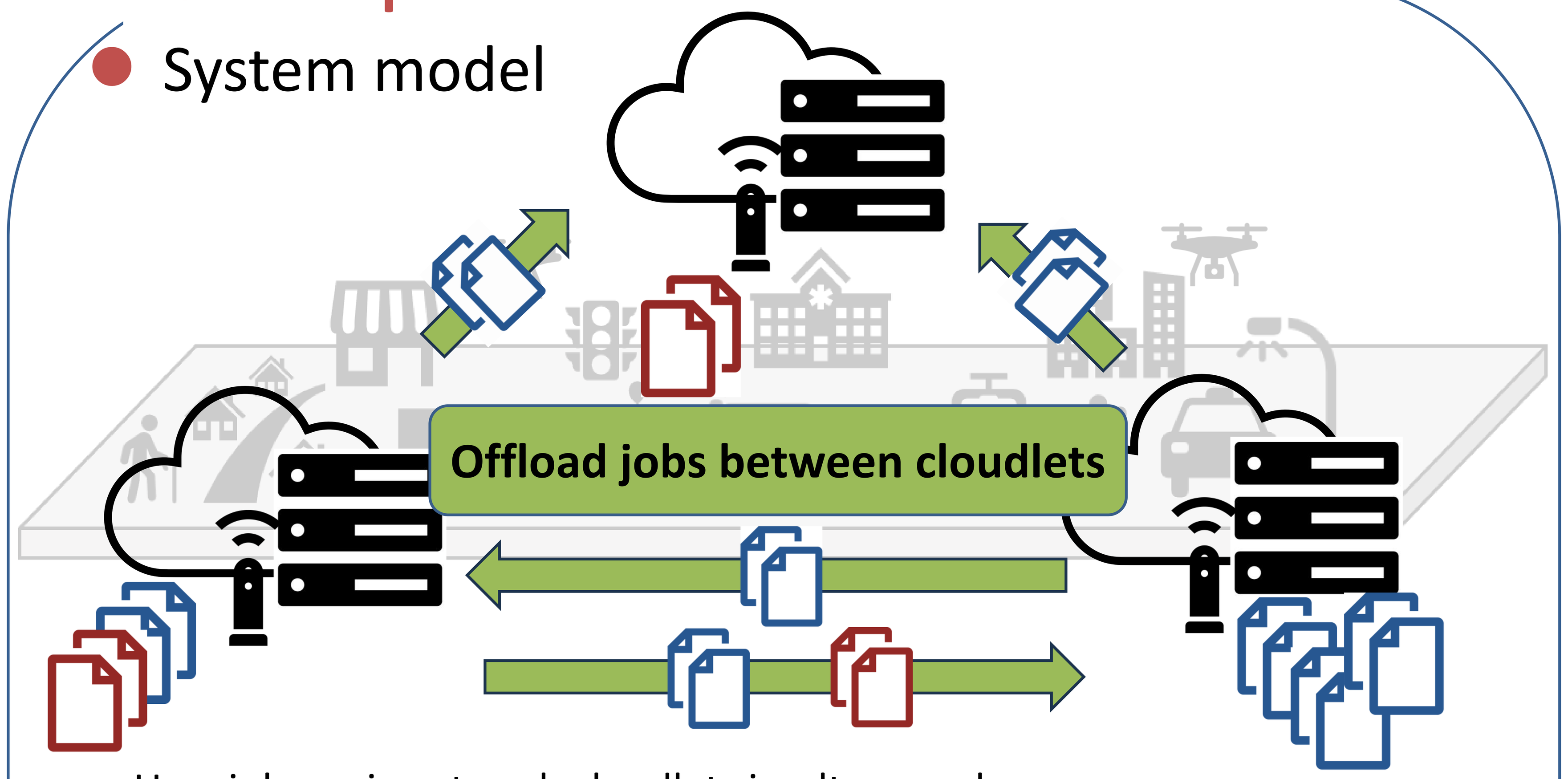
$$L_{\text{low priority}} = \frac{(\lambda_{11}+\lambda_{12})\varphi_{12}b_1^2 + (\lambda_{21}+\lambda_{22})\varphi_{22}b_2^2}{2(1-(\lambda_{11}+\lambda_{12})\varphi_{12}b_1)(1-(\lambda_{11}+\lambda_{12})\varphi_{12}b_1-(\lambda_{21}+\lambda_{22})\varphi_{22}b_2)}$$

$\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}$: Arrival rate of each jobs at each cloudlet [jobs/s]
b_1, b_2: Average processing time of each jobs [s/job]
$\varphi_{11}, \varphi_{12}, \varphi_{21}, \varphi_{22}$: Offload fraction

Compute the optimal $\varphi_{11}, \varphi_{12}, \varphi_{21}, \varphi_{22}$ that minimizes the average latency

Proposed method

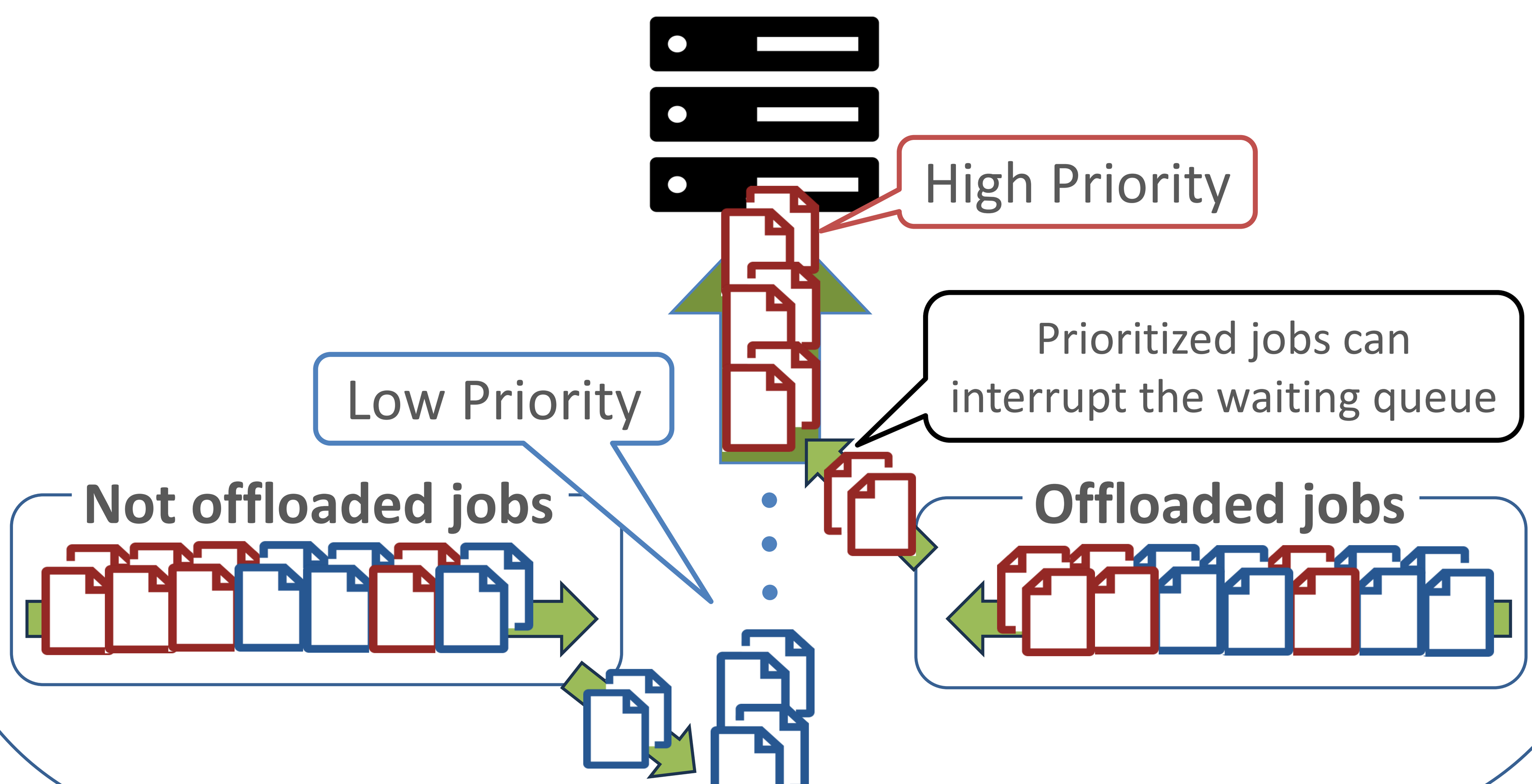
System model



- User jobs arrive at each cloudlet simultaneously.
- Jobs that are decided to be offloaded to other cloudlets will be processed at their destination and their results are sent back to users.

Priority setting

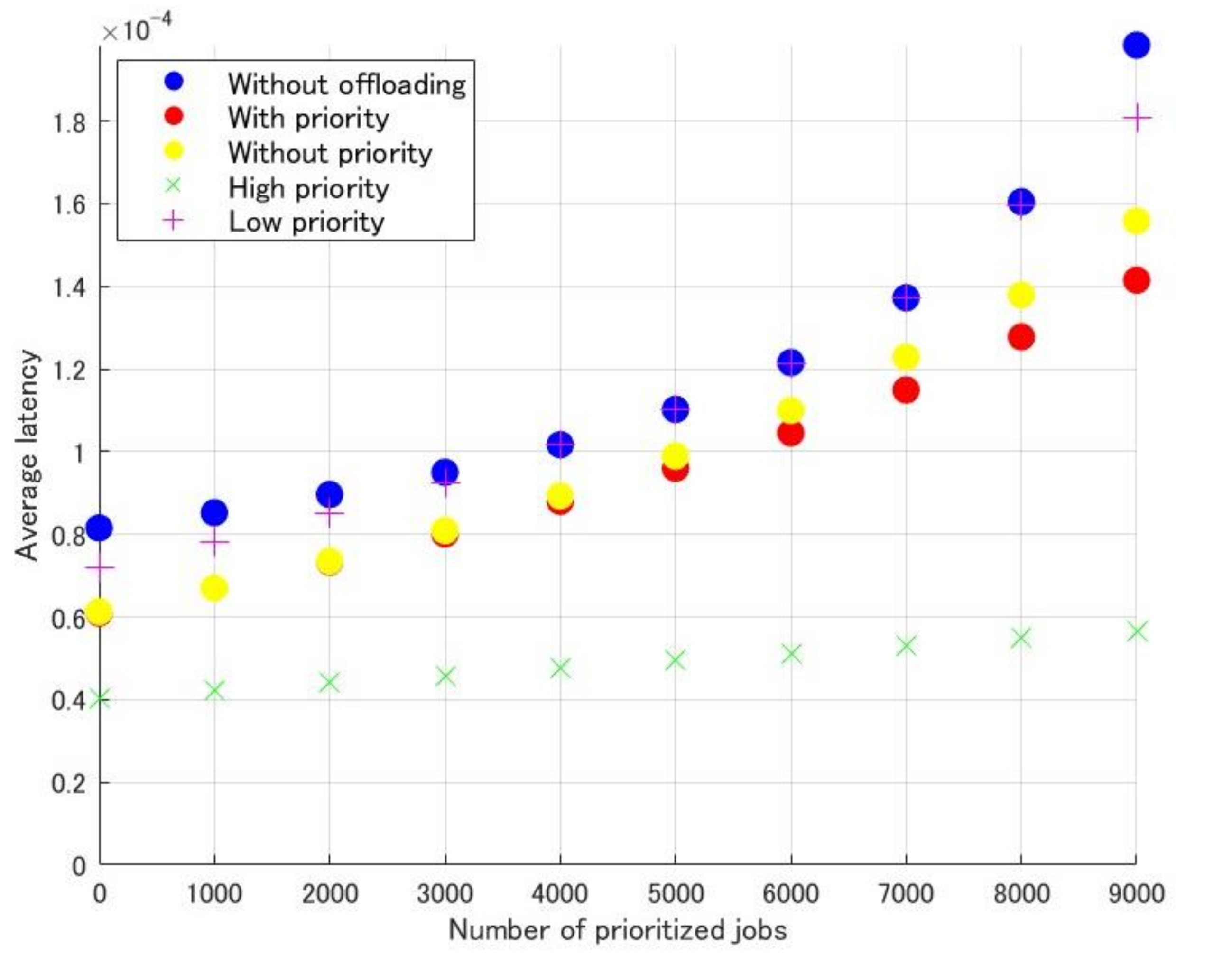
- Priority of jobs are determined according to their average latency based on processing time and arrival rate.



Result

Parameters

- λ_{11}: 2000
- λ_{12}: 0-9000
- λ_{21}: 0
- λ_{22}: 2000
- b_1: 0.6*10^-4
- b_2: 1.2*10^-4



- Setting the priority has decreased the average latency of all jobs as the number of computationally heavy jobs increases.

Summary and Future Goals

- The proposed priority setting has achieved a lower average latency compared to scenarios without priority and without offloading + priority.
- We will investigate the implications of changing the number of cloudlets and job variations.
- Additionally, we will examine the effect of priority settings on different job arrival environments.

Acknowledgement

These research results were obtained from the commissioned research(No.JPJ012368C05601) by National Institute of Information and Communications Technology (NICT) , Japan. This work was supported by JSPS KAKENHI Grant Numbers 19K11947, 22K12015, 22H00247.

References

[1] C. Goumopoulos, "Smart City Middleware: A Survey and a Conceptual Framework," in IEEE Access, vol. 12, pp. 4015-4047, Jan. 2024.
[2] M. Babar, M. S. Khan, F. Ali, M. Imran and M. Shoaib, "Cloudlet Computing: Recent Advances, Taxonomy, and Challenges," in IEEE Access, vol. 9, pp. 29609-29622, Feb. 2021.
[3] M. Khabbaz, "Deadline-Constrained RSU-to-Vehicle Task Offloading Scheme for Vehicular Fog Networks," in IEEE Transactions on Vehicular Technology, vol. 72, no. 11, pp. 14955-14961, Nov. 2023.
[4] Y. Yokota, and S. Miyata, "Latency Focused Load Balancing Method between Cloudlets using Game theory," in Nonlinear Theory and Its Applications, IEICE, vol. E-15N, no. 2, Apr. 2024 (to appear).