Towards Joint Optimization Problem for Computing and Resource Allocation in In-network Computing for Metaverse

Ibrahim Aliyu[1], Ibrahim Mohammed Abdullahi[2], Sang-joon Lee[3], Tai-Won Um[4], and Jinsul Kim*

[1,*]*Dept. of ICT Convergence System Engineering, Chonnam National University, Gwangju, South Korea*
[2]*Dept. of Computer Engineering, Federal University of Technology, Minna, Nigeria*
[3]*Graduate School of Data Science, Chonnam National University, Gwangju, South Korea*
[4]*School of Business Administration, Chonnam National University, Gwangju, South Korea*

aliyu@ieee.org[1], amibrahim@futminna.edu.ng[2], s-lee@jnu.ac.kr[3], *stwum@jnu.ac.kr[4], jsworld@jnu.ac.kr**

Abstract

Metaverse poses a huge challenge regarding networking and computing requirements for task processing, especially on mobile devices. In-network computing (COIN) is a promising solution for enabling computing in the network device for faster completion of computationally intensive tasks and efficient use of resources. This paper examines three modes of metaverse task processing in a mobile network to formulate the joint optimization problem to improve task processing efficiency through task offloading and resource allocation. The mode includes mobile/local in-network computing (LIN), fog in-network computing (FIN) and edge in-network computing. The corresponding task completion and energy consumption models for each mode are examined. Subsequently, the task offload problem is transformed into a joint optimization problem of task completion time and energy consumption.

Keywords: *In-network computing, Joint optimization, Metaverse, Resource allocation.*

1. Introduction

The proliferation towards massive adoption of the metaverse is causing some great concern in the research community and industry regarding the challenge posed by the metaverse's new massive network requirements, such as huge computing resources. The metaverse will only add to this predicament and the massive deployment of the metaverse is beyond the capacity of cloud computing and traditional network [1].

COIN suggest utilising unused computing, storage and network resources to perform some of the tasks within the network [2]. This will make it possible to handle the metaverse immersive computation-intensive tasks by partitioning them into atomic tasks processed by other network resources. To fully leverage the potential of COIN, particularly in a mobile network, attention is needed to understand metaverse tasks and the

computation-intensive tasks offloading to other nodes with the corresponding resources needed for the task.

Although adding more communication and computing resources reduces task execution delay, it also increases the power consumption of the mobile node [4]. To solve this competing issues, the joint optimization problem of time delay and energy consumption of metaverse tasks is formulated.

In this paper, we seek to define the joint optimization problem of metaverse task offloading and resource allocation. We investigated three task offloading modes, including LIN, FIN, and EIN, while modeling the delay and energy consumption. LIN deal with task execution with the mobile node, FIN describes task computation in nearby resources such as PC, while EIN focuses on a more distance, rich communication and computing resource in the operator's access network. Subsequently, we formulate the problem based on the delay model and the energy consumption, following closely the approaches of [4] and [5], which are based on edge-computing concepts. These approaches are suitable for COIN as they all have an element of offloading tasks to an external resource in the fog and edge.

2. System Model and Joint Optimization Problem modeling

Assume that computational tasks arrive at the mobile terminal *i* at time slot t is denoted by a variable $\Re_i(t)$ where $\Re_i(t) \epsilon \left[\Re_{i,min}, \Re_{i,max}\right]$. Each of these tasks can be split into subtasks that can be executed in parallel, series or combined. Various parts of the framework are modelled as follows:

2.1. Task queue model

Local CPU at mobile node (LIN) can perform limited computational tasks due to limited battery capacity and computing power [4]. However, due to the high resource requirement of the metaverse, the local CPU may not be able to perform all the tasks locally. Thus, some tasks need to be offloaded to external computing resources.

In this study, we consider two external resource scenarios- (1) a fog in-network computing node (FIN), which includes nearby resources such as nearby smart TV (2) edge in-network computing node (EIN) which includes more sufficient computing resources such as programmable network devices (PND) in the operator's access network. Let $\mathfrak{T}_{i,L}(t)$, $\mathfrak{T}_{i,F}(t)$ and $\mathfrak{T}_{i,E}(t)$ represent the number of tasks that can be performed by LIN, FIN and EIN, respectively.

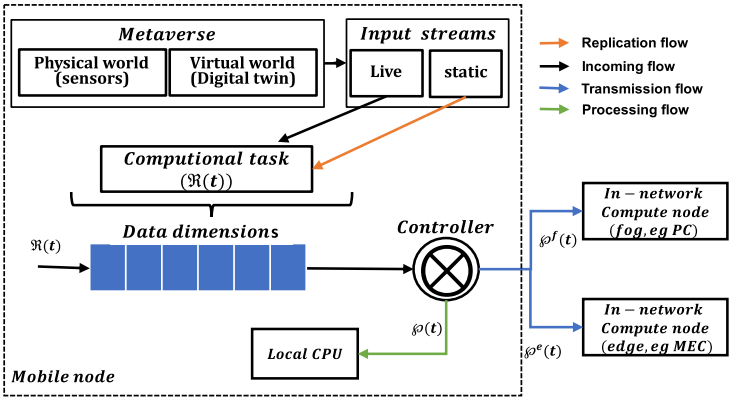

Figure 1. Metaverse in-network computing task offloading framework.

Tasks can generally be characterized by the size of the input data size K_i and complexity L_i which is the number of required CPU cycles. The task complexity L_i in relation to the input data size K_i can be expressed:

$$L_i = K_i \mathcal{U}_i \tag{1}$$

where $\mathcal{U}_i$ is the number of CPU cycle per data bit.

The capacity of a computing node/link *i* can be characterized in terms of uncontrollable system state (e.g. channel state) $w(t)$ and controllable resource allocation $\alpha(t)$ and represented as:

$$\theta_i(t) = \theta\big(w(t), \alpha(t)\big) \tag{2}$$

In other words, the node's capacity can be described as the number of tasks the node can handle. For our scenario, lets $\theta_{i,L}(t)$, $\theta_{i,F}(t)$ and $\theta_{i,E}(t)$ denotes the capacity of LIN, FIN and EIN, respectively. The constraint on the resource is then expressed as

$$\theta_{i,L}(t) \epsilon \left[{\theta_{iL}}^{min}, {\theta_{iL}}^{max}\right],\ \theta_{i,F}(t) = \left[{\theta_{i,F}}^{min}, {\theta_{i,F}}^{max}\right],\ \theta_{i,E}(t) = \left[{\theta_{i,E}}^{min}, {\theta_{i,E}}^{max}\right] \tag{3}$$

Generally, tasks arriving are placed in a buffer queue before execution. The queue length at time *t* in a mobile node *i* can be described as

$$Q_i(t+1) = \left[Q_i(t) - \mathfrak{T}_{i,\sum(t)},\ Q_i(0)\right] + \mathfrak{R}_i(t) \tag{4}$$

where $\mathfrak{T}_{i,\sum(t)} = \mathfrak{T}_{i,L}(t) + \mathfrak{T}_{i,F}(t) + \mathfrak{T}_{i,E}(t)$ describes the number of tasks exiting the queue [4]. The flexible queuing system is essential for network state evolution resulting from resource allocation and flow scheduling actions. In essence, the network system state $s(t)$ is characterize in terms of the uncontrollable state system $w(t)$ and the queue system ($Q(t)$) [5]:

$$s(t) = \big(w(t), Q(t)\big) \tag{5}$$

2.2. Task time delay and energy consumption model

The time delay in execution time is synonymous with task completion time. Each of the computing nodes, ie. LIN, FIN and EIN require a time delay model in order to comprehensively model the cost of operating the nodes.

LIN time delay model. The time taken to execute a given task (K_i, L_i) in LIN only includes the processing time on the local node at time *t* and is defined as:

$$T_i^L(t) = \frac{L_i}{F_i^L(t)} = \frac{K_i \mathcal{U}_i(t)}{F_i^L(t)},\ s.t\ F_i^L(t) \leq\ F_{i,max} \tag{6}$$

where F_i^L is the actual frequency at which a local node can execute tasks.

FIN time delay model. Offloading a task of input size (K_i, L_i) to external compute node such as FIN, consist of three delay parts. The first part deals with the time needed to transmit the input data K_i through an access point a and is expressed as:

$$T_{i,a}^t(t) = \frac{K_i}{\omega_{i,a}},\ s.t\ T_i^F(t) \leq \theta_{i,F}(t) \tag{7}$$

where $\omega_{i,a}$ is the uplink rate of the node $i \epsilon \mathcal{N}_a$ which depends on the specific set of connected node $\mathcal{N}_a$ The second part deals with the time taken to execute the task in the external resource FIN and is expressed as:

$$T_{i,F}^{exe}(t) = \frac{L_i}{F_i^F(t)} \tag{8}$$

where F_i^F is the actual frequency at which the FIN node can execute a task.

The last part of the delay is the time taken to return the computation results from the FIN node to the mobile/local node. However, this delay is observed to be negligible compared to the input data K_i [6]. Thus the total delay for offloading task to the FIN over an access point a can be defind as:

$$T_{i,a}^F(t) = T_{i,a}^t(t) + T_{i,F}^{exe}(t) \quad (9)$$

EIN time delay model. For simplicity, the EIN node is assumed to have sufficient resources to compute the offloaded task K_i that is transmitted via the cellular network. Thus, the processing time at the EIN can be considered negligible. At time t, the number of tasks that EIN received from the mobile node can express as:

$$T_{i,a}^E(t) = K_i r_i(t) \quad (10)$$

where $r_i(t)$ is transmission rate to the EIN component per bit.

Meanwhile, the total energy consumption of a mobile terminal is characterized by the node's execution energy consumption and transmission energy consumption. The transmission energy entails the energy required to offload the task to FIN and EIN. Each component's energy consumption is considered as follows:

LIN energy consumption model. The energy consumption of executing a given task (K_i, L_i) using LIN at frequency F_i^L is linearly proportional to the square of the $F_i^L(t)$ and is given as:

$$\mathcal{E}_i^L = \tau\,(F_i^L(t))^2\,L_i \quad (11)$$

where $\tau \sim 10^{-11}$ [6].

FIN energy consumption model. For offloading task a given task (K_i, L_i) to FIN through access point a, the energy consumption is characterized by the energy used in uploading the task input size a given task K_i, considering the energy for connection scanning is negligible. The energy consumption can be expressed as:

$$\mathcal{E}_{i,a}^F(t) = \frac{K_i \wp_{i,a}^F}{\omega_{i,a}}, \quad (12)$$

WHERE $\wp_{i,a}^F$ is the transmission power of the device through access point a.

EIN energy consumption model. The energy consumption at EIN is similar to FIN. However, the execution energy is negligible. Therefore, the EIN energy consumption can be modelled as:

$$\mathcal{E}_{i,a}^E(t) = K_i \wp_{i,a}^E \quad (13)$$

where $\wp_{i,a}^E$ is the transmission power the EIN through the access point a.

Total energy consumption. The total energy consumption of a mobile terminal *i* at a given time slot *t* is given as:

$$\mathcal{E}_{i,\sum(t)} = \mathcal{E}_i^L(t) + \mathcal{E}_{i,a}^F(t) + \mathcal{E}_{i,a}^E(t) \quad (14)$$

For mobile equipment, energy consumption in a particular time slot is expressed as:

$$\mathcal{E}_i(t) = \sum_{i=1}^N \mathcal{E}_{i,\sum(t)}(t) \quad (15)$$

3. The Joint Optimization Problem Formulation

In this study, we considered a mobile COIN system that consists of set of mobile computing node $|N| = N$, communication resource $|\mathcal{A}| = A$ and computing resource $|\mathrm{C}| = C$, following closely the formulation of [6]. A given task can be allowed access to communication and computing resources using task placement matrices $\mathcal{X} \epsilon \{0,1\}^{N \times A}$ and $\mathcal{Y} \epsilon \{0,1\}^{N \times C}$, respectively. The management policies on the allocation of the

communication and computation resource can express as $\mathcal{P}_{\mathcal{X}}: \rightarrow \mathbb{R}_{[0,1]}^{N\times A}$ and $\mathcal{P}_{\mathcal{Y}}: \rightarrow \mathbb{R}_{[0,1]}^{N\times C}$, respectively. Therefore, the system cost is given as $\mathcal{C}(\mathcal{X}, \mathcal{Y}, \mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}})$. By relying on the definitions, the joint optimization problem for the metaverse task computation and resource allocation can be formulated as follows:

$$\mathcal{J}_p : \min_{\mathcal{X},\mathcal{Y},\mathcal{P}_{\mathcal{X}},\mathcal{P}_{\mathcal{Y}}} \mathcal{C}(\mathcal{X}, \mathcal{Y}, \mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}}) \tag{16}$$

$$\begin{aligned} s.t. \quad & d1: \mathcal{B}_i(\mathcal{X}, \mathcal{Y}, \mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}}) \leq \delta_i, \forall i \in \mathcal{N}, \\ & d2: \rho(\mathcal{X}, \mathcal{P}_{\mathcal{X}}) \leq \delta_a, \forall a \in \mathcal{A}, \\ & d3: q(\mathcal{Y}, \mathcal{P}_{\mathcal{Y}}) \leq \delta_c, \forall c \in \mathbb{C}, \\ & d4: \sum_{a\epsilon\mathcal{A}} \sum_{c\epsilon\mathbb{C}\setminus\{i\}} \mathcal{X}_{i,a}\mathcal{Y}_{i,c} + \mathcal{Y}_{i,i} = 1, \forall i \in \mathcal{N}, \\ & d5: \mathcal{X}\epsilon\{0,1\}^{N\times A}, \quad \mathcal{Y}\epsilon\{0,1\}^{N\times C} \\ & d6: \mathcal{P}_{\mathcal{X}}: \rightarrow \mathbb{R}_{[0,1]}^{N\times A}, \quad \mathcal{P}_{\mathcal{Y}}: \rightarrow \mathbb{R}_{[0,1]}^{N\times C} \end{aligned}$$

The constraint $d1$ enforces the optimization to consider either completion time or energy consumption of the devices $i \in \mathcal{N}$ is within the threshold δ_i. Constraint $d2$ and $d3$ only limit operation to only communications and computing resources, respectively. The decision to perform computation locally or offload the task to an external resource is enforced by $d4$. The constraints $d5$ enforce placement decisions to be integers, while $d6$ describes resource allocation policies. Solving the entire problem is impractical in practice. But these models can be easily be adapted for different allocation problems in-network computing

4. Conclusion

This paper described a joint optimization problem formulation for metaverse in in-network computing. We consider three communication and computing task allocation modes: the LIN, FIN and EIN. We considered the performance of these modes in terms of task completion time and energy consumption. As the addition of COIN resources in the network adversely affects the overall network performance, formulating the optimization problem is necessary to strike a balance between competing communication and computing resources for optimal resource allocation. Future studies would focus on solving different problems relating to the joint optimization problem.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(NRF-2021R1I1A3060565) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-02068, Artificial Intelligence Innovation Hub).

References

[1] P. Liu and L. Geng. "Requirement of Computing in network. Working Internet Draft, draft-liu-coinrg-requirement-03." Internet Engineering Task Force (IETF). https://datatracker.ietf.org/doc/draft-liu-coinrg-requirement/03/ (accessed 04/12, 2022).

[2] S. Huang *et al.*, "Intelligent Eco Networking (IEN) III: A Shared In-network Computing Infrastructure towards Future Internet," in *2020 3rd International Conference on Hot Information-Centric Networking (HotICN)*, 2020: IEEE, pp. 47-52.

[3] I. Kunze *et al.* "Use Cases for In-Network Computing. Working Internet Draft, draft-irtf-coinrg-use-cases-02." Internet Engineering Task Force (IETF). https://datatracker.ietf.org/doc/draft-irtf-coinrg-use-cases/02/ (accessed 03/28, 2022).

[4] S. Yang, "A joint optimization scheme for task offloading and resource allocation based on edge computing in 5G communication networks," *Computer Communications,* vol. 160, pp. 759-768, 2020.

[5] Y. Cai, J. Llorca, A. M. Tulino, and A. F. Molisch, "Compute-and Data-Intensive Networks: The Key to the Metaverse," *arXiv preprint arXiv:2204.02001,* 2022.

[6] S. Josilo, "Task placement and resource allocation in edge computing systems," KTH Royal Institute of Technology, 2020.