A Parallel Processing Method of Large-scale System Level Simulator for Advanced 5G System

Megumi Shibuya, Akira Yamaguchi, Takahide Murakami, and Hiroyuki Shinbo
KDDI Research, Inc.
Saitama, Japan
e-mail: {xmu-shibuya, ai-yamaguchi, tk-murakami, hi-shinbo}@kddi.com

Abstract—The advanced fifth generation mobile communication system (5G system) around 2025 is expected to introduce new technologies, such as virtualized Radio Access Network (vRAN) that can place base station functions on general servers, and base station function placement based on vRAN to fit the quality requirements of communication services. In order to evaluate the quality of end-to-end communication in mobile communication system, a System Level Simulator (SLS) is widely used. However, more simulation time for SLS with the advanced 5G system is required than with the 5G system. Because new technologies are added in SLS and it executed long-term and large-scale simulations are required. A reduction of simulation time for SLS is required for an effective evaluation. In this paper, we propose a software design of SLS for the advanced 5G system with RU-basis parallel processing by multiple computation nodes. Through the SLS executed on the supercomputer Fugaku, we confirmed that our proposed method can reduce SLS processing time.

Keywords - System level simulator; advanced 5G system; virtual RAN; parallel processing; MPI

I. Introduction

The fifth generation mobile communication system (5G system) has already become widespread in many countries. Around 2025, the 5G system will further advanced as the "*advanced 5G system*" accompanied by the introduction of new technologies. In order to introduce new technologies, it is necessary to evaluate the end-to-end communication quality of the entire advanced 5G system with the technologies due to clear how affect use-level packet. For example, the new technologies are wireless communication systems, such as massive Multiple Input Multiple Output (MIMO) [1], grant-free non-orthogonal multiple access for Internet of Things (IoT) [2], and virtualized Radio Access Network (vRAN). To evaluate the end-to-end communication quality with these technologies in a mobile communication system, a System Level Simulator (SLS) on computers is widely used.

The existing SLSs for the 5G system [3][4][5] can simulate the User Equipment (UE) layout and movement, Radio Unit (RU) layout, generation of application traffic, and the wireless communication technologies of the 5G system. For evaluations of the advanced 5G system, additional technologies will be simulated, such as a new wireless communication system and vRAN. In addition, the management methods for vRAN to maintain communication quality are also evaluated by SLS, such as base station function placement based on computation resources and transport resources [6], and radio resource assignment for each virtualized base station function [7]. The vRAN controls by the management methods are judged by the frequently changed status such as the status of the radio links between each of the RUs and UEs, and UE traffic generation. In addition, for example, base station function placement in vRAN is controlled by aggregating RAN-wide information, such as time-varying radio quality information and the generated traffic in each UE. Therefore, when base station function placement changes, the communication quality of user-level packets is affected by the placed base station function. Because of this interaction, the RAN portion and the radio portion must be simulated simultaneously. Although abstracted simulations are proposed [5], since it is not yet clear how the new technologies in the advanced 5G system will affect user-level packets, detailed simulations of individual technologies should be performed for the initial stage evaluation.

Figure 1 shows the simulation configuration of the advanced 5G system. There are approximately 1,000 RUs and 50,000 UEs inside approximately one square kilometer. The RUs create many areas of cells with various frequency bands. UEs (e.g., cars, IoT terminals, and smartphones) move in the areas. In addition, to judge the usefulness of the management methods, it is required an SLS simulation time of around 10 minutes to 1 hour period. The simulation will become large-scale about the number of RUs and UEs, and long-term about simulation time.

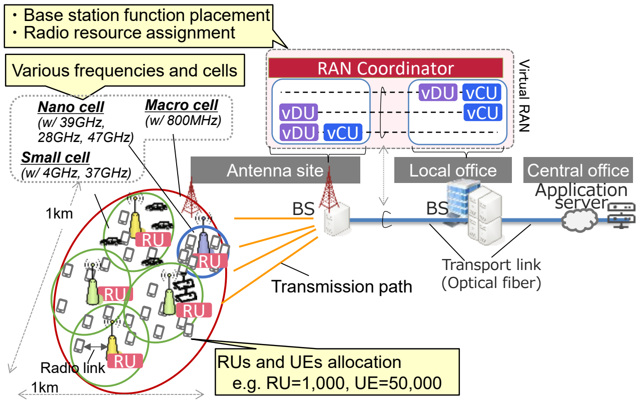
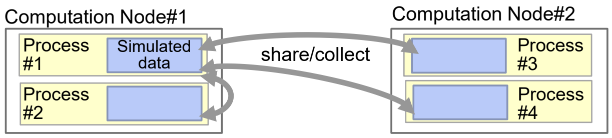


Figure 1. Simulation configuration of advanced 5G system.

Simulations with both new technologies and long-term/large-scale simulation takes a long time to execute the SLS for the advanced 5G system. The long simulation time

ISBN: 978-1-68558-052-0

affects effective evaluations by the SLS because many simulations will be executed to evaluate new wireless communication technologies and management methods. Therefore, we need to reduce the simulation time for the SLS. One method of reducing the simulation time of the SLS is a parallel processing method by one computation node with multi-cores [8]. However, since total computation power is limited, one computation node cannot effectively improve the simulation time. To obtain more computation power, there is a parallel processing method using multiple computation nodes with multi-cores [9]. To apply the SLS to this environment, the simulation result of each radio frame transmission (e.g., 1 msec) needs to be shared between the computation nodes because the next radio frame transmission is simulated based on previous radio frame transmissions. Since such data sharing and collecting between computation nodes is over a network, a "*memory access time*" is required. As shown in Figure 2, applying multiple computation nodes has pros and cons in that it can reduce SLS processing time in each computation node, but the memory access time between the computation nodes is increased. In order to reduce the total simulation time of the SLS, it is important to have balanced software design between the memory access time and the reduction of the SLS processing time.



Important balance :
- Pros. : Use multiple computation nodes → **reduce processing time**
- Cons. : Transfer large size data → **increase memory access time**

Figure 2. Pros and cons of data sharing in multiple computation nodes.

In this paper, in order to reduce the simulation time, we propose a software design of SLS for the advanced 5G system (A5G-SLS) with multiple computation nodes. The A5G-SLS introduces the RU-basis for parallel processing to reduce processing time. To evaluate the effectiveness of our proposed method, we use different types of parallel processing models, such as ALL MPIs and hybrid models, on the supercomputer Fugaku [10]. The rest of this paper is organized as follows. Section II presents the simulation targets and the problem of processing time reduction in the A5G-SLS. Section III proposes a software design for the A5G-SLS with parallel processing by multiple computation nodes. We present the evaluation of the proposed method in Section IV. Finally, we provide our conclusion in Section V.

II. System Level Simulator for Advanced 5G System and Parallel Processing Problem

A. *Overview of advanced 5G RAN*

In the advanced 5G system, the functions of the base stations are divided into a Central Unit (CU), a Distributed Unit (DU) and an RU. In the vRAN environment, the CU and DU are virtualized as the vCU and vDU, and they work on commodity servers. The servers are placed at various locations in the RAN, i.e., an antenna site, a local office, and a central office. An optical fiber connects the local office and the antenna site as a physical link in the physical configuration as shown in Figure 3 (a). The application server is located in the central office (or local office), and various types of service traffic are generated, such as high-definition video streaming, connected vehicles, drone control, and the IoT.

The logical networks are constructed by the vCUs and vDUs, they are works on the commodity servers and the RUs in an area shown in Figure 3 (b). In order to maintain communication quality, a RAN slice is created for each service [6]. For example, one RAN slice is created for high-capacity traffic, and another RAN slice is created for low-latency, and massive connections in the area. The vRAN management method controls base station function placement based on computer resources and transport resources, and radio resource assignment based on the radio link quality between each of the RUs and UEs, and UEs traffic generation. For example, in Figure 3 (b), one RAN slice is reallocated from high-capacity traffic to massive connections, and the assignment of radio resource is controlled. To obtain user packet level results for throughput, delays, and errors in end-to-end communications, both the RAN and radio links are simulated at the same time. In addition, radio resource assignment is performed in relatively short-term control (100 msec to 1 sec), and base station function placement is performed in long-term control (10 minutes to 1 hour). On the other hand, the radio resource scheduling in the vDU for the radio link is performed in ultra-short control (approximately 1 msec).

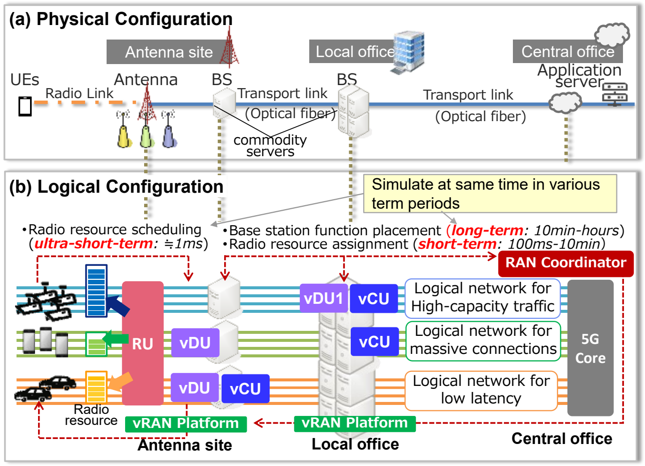

(a) Physical Configuration
Antenna site
Local office
Central office
UEs
Radio Link
Antenna
BS
Transport link
(Optical fiber)
BS
Transport link
(Optical fiber)
Application server
commodity servers
(b) Logical Configuration
Simulate at same time in various term periods
•Radio resource scheduling (ultra-short-term: ≒1ms)
•Base station function placement (long-term: 10min-hours)
•Radio resource assignment (short-term: 100ms-10min)
RAN Coordinator
RU
vDU
vDU
vCU
vDU1
vCU
vCU
Logical network for High-capacity traffic
Logical network for massive connections
Logical network for low latency
5G Core
Radio resource
vRAN Platform
vRAN Platform
Antenna site
Local office
Central office

Figure 3. Overview of 5G advanced RAN

B. *A5G-SLS Implementation and Points for Simulation Targets*

The A5G-SLS consists of three parts; initialization, preprocessing, and main simulation. As the initialization part, the physical and logical configurations, such as base stations (vCUs, vDUs, RUs), UEs, slices, and traffic servers, are set in the initial files as parameters. In addition, the services and simulation time are set in the same files. The traffic of each service for each UE and the propagation are created using the initial files in the preprocessing part. The main simulation conducts the simulation process, such as propagation at a transport link, the radio packet process includes radio resource scheduling at the vDU, and the radio link process includes error and retransmission in the radio links and the UE. The main simulation part conducts the processing for each time step until the simulation time is finished. We created the A5G-SLS for advanced 5G systems based on the SLS for 5G systems used in the evaluation of previous research [11].

From the above configuration, the A5G-SLS can simulate the following. A large number of RUs and UEs are placed, the RUs with various frequency bands, and UEs, which move in the area, are simulated. The propagations are simulated using a radio propagation model in 1 msec order, and the radio frames are sent/received using propagation data that include error and retransmission. In addition, the networks and transport links in a RAN are simulated, and the data packets are transferred through them. The data packets are transferred as radio signals or user data depending on changes in the base station functions. Furthermore, the UE can generate various services data, such as high-definition video streaming and connected vehicles. From these simulations, it is possible to clear how the generated traffic is affected by radio frames and transmission paths as communication, and the end-to-end communication quality can be simulated.

The simulation targets of the A5G-SLS are to evaluate the quality of end-to-end communication for each user in the large-scale environment of the advanced 5G system. There are two points for simulating the targets;

Point 1 is new technologies. One of the new technologies is the new wireless communication systems, such as MIMO. The other is vRAN and its management method, such as base station function placement and radio resource assignment. As described in section II A, the vRAN controls are judged by frequently changed statuses (e.g., radio links and UE traffic generation). In addition, the communication quality of user-level packets is affected by the changing placement of the base stations. Because of this interaction, the RAN portion and the radio portion must be simulated simultaneously.

Point 2 is large-scale simulation. One large-scale simulation is the environment. As described in Section I, there are approximately 1,000 RUs and 50,000 UEs in approximately one square kilometer. The RUs create many areas with various frequency bands, such as small-cell, nano-cell, and macro-cell. UEs are assumed to be the cars, IoT terminals, and smartphones, and they move in the areas. The other large-scale simulation is long-term simulation. To judge the usefulness of management methods, it is required a simulation time of SLS around 10 minutes to 1 hour period.

The above two points require that the simulations with both new technologies and long-term/large-scale simulation take a long time (e.g., a few days to a week) to execute on the SLS for the advanced 5G system. In addition, considering the new technologies, since the quantitative effects on user-level packets and end-to-end communication quality in the advanced 5G system are not yet clear, detailed simulations of how to work the radio frame transmission and user-level packets should be performed in 1 msec order. One reason is that a detailed simulation of the radio environment, such as each radio link between RUs and UE, is needed to evaluate new wireless communication systems. Since the effectiveness of new wireless communication system for user-level communication, such as communication quality, is unknown, the wireless communication system needs to evaluate the radio environment in detail on the order of milliseconds. The other reason is the evaluation of vRAN management methods and user-level communication quality based on the status of user traffic, usage of radio/transport/computation resources in RAN, and the status of radio links between the RUs and UEs. To evaluate the management methods and the communication quality, all items such as vRAN, user traffic generation, and radio links are required to simulate at the same time. In addition, long-term simulation is required because the management method controls vRAN per 1 or 10 minutes.

C. Related SLS (Existing System Level Simulators)

As the SLS for the 5G system, some simulation tools and open platforms are proposed in [3][4][12][13], and [14]. OMNet++ [3], NS-3 [4], and Veins [12] exist as popular event-based network simulators. In these simulators, most of the protocol stacks are modeled. However, in order to address computational complexity, it is difficult to simulate large-scale networks. 5G K-SimSys [13] has been developed to provide an open platform for evaluating SLS performance of the 5G standard. It is designed to be flexible, open, and has a modular form to make it easy to customize. To evaluate performance, a more complex testbed is required. OpenAirInterface [14] is implemented in part of the 3GPP LTE and provides an interface between the hardware platform and works as an emulator. When the complexity increases, it is difficult to conduct a large-scale simulation due to the number of nodes, which is limited.

The approaches of reducing the simulation time for the SLS have been conducted [15][16][8], and [5]. 5G-Lena [15], which is NS-3 based, introduces a method of reducing simulation time by abstracting the physical layer. D. S. Buse et al. [16] introduced an approach in which some part of the SLS process of the wireless signal attenuation model computation is asynchronously pushed to the background and offloaded. This approach is implemented in the Veins of VANET simulator. However, when the simulation scale becomes large and complex, the simulation takes a long time to compute due to the single-thread simulation run. Therefore, it has the limitation of reducing the simulation time for the SLS. As other approaches, Vienna5G [8] has been proposed. This simulator is based on MATLAB, and it can perform in a large-scale, multi-tier network with numerous types of network nodes. This approach uses multiple threads on a computation node. However, it is insufficient to reduce simulation time by only multi-threads processing in one computation node to increasing amount of calculation. Simu5G [5] is provided as an open-source simulator with an emulator function based on OMNeT++. In order to provide real-time emulation, it is introduced lightweight models of UEs and gNBs with abstracted limited functionalities for creating resource contention and interference.

Comparing the existing 5G SLS and the A5G-SLS, the existing 5G SLS in [16] presents the simulation result using a scenario, such as 30 RUs, 7,000 UEs, and 100 msec of simulation time. On the other hand, in the A5G-SLS, the simulation is performed using a scenario, such as approximately 1,000 RUs, 50,000 UEs, and one hour of simulation time. Hence, large-scale and long-term simulations are required. Moreover, in terms of the implementation of functions, the existing 5G SLS consists of the following functions: 1) services and the transmission paths process, 2) DU process, 3) RU process, and 5) other processes as shown in Table I. On the other hand, the A5G-SLS includes the function 4) the vRAN process in addition to the functions of the 5G SLS. These processes from 1) to 5) are performed repeatedly in the order of each time step (e.g., every 1 msec) until the end of the simulation. Namely, the A5G-SLS increases the simulation functions. Hence, to be adaptive of increasing calculation, it requires the parallel processing method using multiple computation nodes.

TABLE I. IMPLEMENTATION OF THE FUNCTIONS FOR EXISTING 5G SLS AND A5G-SLS

Function	Process	Existing 5G SLS	A5G-SLS
1) Services, transmission paths	• Traffic generation • Packet process (application server to vCU) , transmission paths	✓	✓
2) DU	• Wireless scheduler • Packet process (vDU to RU)	✓	✓
3) RU	• Packet process (RU to UE) • Radio communication quality measurement (calculate SINR) • Calculate receive power	✓	✓
4) vRAN	• Radio communication quality (calculate the RU changing indicator) • RAN coordinator (RUs and resources allocation)	-	✓
5) Others	• Radio communication (hand-over)	✓	✓

D. *Problem with Parallel Processing*

To realize parallel processing using multiple computation nodes, there are the following problems.

Problem-a) Transmission time of memory data: To use multiple computation nodes, in order to share the data between all processes, memory data are collected and shared between multiple computation nodes. In general, as shown in Figure 4, in existing scientific simulations, such as fluid dynamics [17] and weather forecasting [18], each process is an independent event, only the data used in each process is transferred, and the calculated result data is collected to approximately 320 Kbytes [18]. On the other hand, in the A5G-SLS, the radio allocation data of all UEs and RUs are required for calculating SINR. Therefore, the A5G-SLS transfers all simulated data in memory to all processes (shared data), and transfers all RUs and UEs information where changes, such as SINR, have occurred (collected data) (e.g., when the number of RUs is 598 and the number of UEs is 35 per RU, transmission data size, which includes shared data and collected data, is approximately 5 MBytes). Hence, the data transfer time becomes longer.

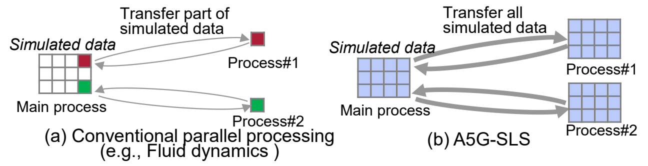

Transfer part of simulated data
Simulated data
Process#1
Main process
Process#2
(a) Conventional parallel processing (e.g., Fluid dynamics)
Transfer all simulated data
Simulated data
Process#1
Main process
Process#2
(b) A5G-SLS

Figure 4. Transmission data size for conventional parallel processing and A5G-SLS.

Problem-b) Process waiting time: Conventionally, the parallel processing method has the problem that it cannot proceed to the next process until all processes are finished because of the synchronization of all processes and the sharing of all simulation results between the computation nodes. In other words, the processes should wait until all processes reach this barrier of synchronization as shown in Figure 5. In the A5G-SLS, the simulation result of each radio frame transmission needs to be shared between computation nodes. However, increasing the waiting time increases the processing time for the simulation. In order to solve this problem, the A5G-SLS is required to reduce the waiting time. For this reason, it is necessary to design all parallel processes to have similar processing times as much as possible.

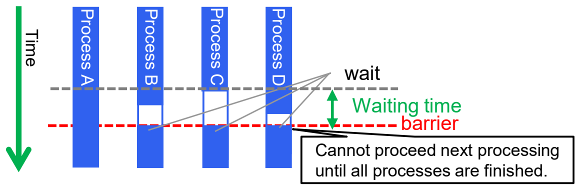
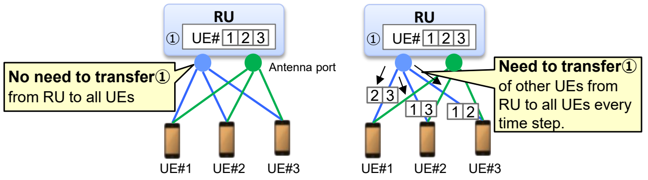

Time
Process A
Process B
Process C
Process D
wait
Waiting time
barrier
Cannot proceed next processing until all processes are finished.

Figure 5. Process waiting time.

Problem-c) Dividing basis for parallel processing: As the dividing basis, two types exist: RU unit (hereinafter "RU-basis") and the UE unit (hereinafter "UE-basis") as shown in Figure 6. The UE-basis processing needs to transfer the radio resource assignment with the channel status information (CSI) data of other UEs from the RU to all UEs for every time step. Therefore, the transference data size and times are increased. On the other hand, the RU-basis does not need to transfer the radio resource assignment with CSI data from the RU to all UEs because the RU already has them. Hence, if the dividing unit is inadequate, it takes a long processing time due to Problem-a) and Problem-b).


RU
① UE# 1 2 3
Antenna port
No need to transfer① from RU to all UEs
UE#1
UE#2
UE#3
RU
① UE# 1 2 3
2 3
1 3
1 2
Need to transfer① of other UEs from RU to all UEs every time step.
UE#1
UE#2
UE#3

・① : Radio resource assignment with Channel Status Information (CSI) data per UE

(a) RU-basis processing (b) UE-basis processing

Figure 6. Distribution of the radio resource assignment with CSI for each divided method.

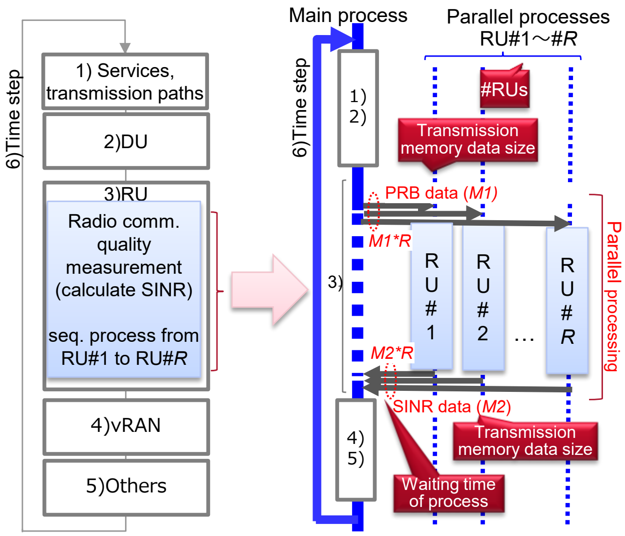
To address these problems, it requires reducing the simulation time by the parallel processing method using multiple computation nodes.

III. Proposal for A5G-SLS Design Method

In this section, we explain the parallel processing method for reducing simulation time by using multiple computation nodes to resolve Problem-a), Problem-b) and Problem-c) as described in the previous section.

For the design of the parallel processing target, the processing with a high computation load in the A5G-SLS is selected due to the effect of the processing time reduction. We analyzed the program execution time in the A5G-SLS from the following two viewpoints; the availability of parallel processing, and the process of the high computation load within A5G-SLS. As a result, the computation load imposed by the radio communication quality measurement (calculation of the Signal to Interference and Noise Ratio (SINR)) in RU processing is high (approximately 30% of the total) so this part is parallelized in multiple processing.

In addition, as a selection basis method for parallel processing, the UE-basis and RU-basis exist as shown in Figure 6. In order to calculate the SINR, the radio resource assignment data with the CSI of all UEs are required. In the case of the UE-basis, the RU transfers the radio resource assignment data with the CSI to all UEs at every time step in order to calculate the SINR at each UE. On the other hand, in the case of the RU-basis, it is not necessary to transfer the radio resource assignment data with the channel status information from each UE because the RU already has the data. By using these selection methods, the transmission volume is reduced, and the transmission time is reduced (Problem-a can be resolved). In addition, since there are no major differences in the process load of each RU, the processing waiting time may be reduced. Hence, we propose a method of reducing simulation time using the RU-basis (Problem-b and Problem-c can be resolved).


6)Time step
1) Services, transmission paths
2)DU
3)RU
Radio comm. quality measurement (calculate SINR)
seq. process from RU#1 to RU#R
4)vRAN
5)Others
Main process
Parallel processes RU#1〜#R
6)Time step
1)
2)
#RUs
Transmission memory data size
PRB data (M1)
M1*R
3)
RU#1
RU#2
…
RU#R
Parallel processing
M2*R
SINR data (M2)
Transmission memory data size
4)
5)
Waiting time of process

(a) Non-parallel processing (b) Parallel processing of RU-basis

Figure 7. Sequence of parallel processing for A5G-SLS.

Figure 7 shows the specific processing sequence for the A5G-SLS. Non-parallel processing is shown in Figure 7 (a), and the parallel sequence of the RU-basis is shown in Figure 7 (b). We explain the sequence using the simulation scenario when the number of RUs is denoted as *R*. In the case of non-parallel processing, the SINR is calculated sequentially from RU#1 to RU#*R* in the measurement process of the radio communication quality in RU processing. On the other hand, in the proposed parallel processing method for the RU-basis, SINR calculation processing is divided into the *R* of each RU, then the *R* processes are conducted in parallel. Since introducing parallel processing, data transfer between the main process and each RU process is necessary due to sharing and collecting the data; when each process starts, the Physical Resource Blocks (PRBs) of data *M1* [Byte], which are allocated to all UEs by all RUs, are transferred from the main process to each RU process, and when each RU process ends, the SINR data of each UE *M2* [Byte] are transferred from each RU process to the main process. Hence, the total transmission memory data sizes from and to the main process of PRB and SINR are $M1 \times R$ [Byte], $M2 \times R$ [Byte], respectively.

IV. Evaluation

In this section, we verify our proposed method for the A5G-SLS to reduce the processing time on the RU-basis using parallel processing by multiple computation nodes.

A. Evaluation Viewpoints

We evaluate our proposed method from two viewpoints.

Viewpoint 1 is the reduction in processing time. The processing time is compared with some parallel models and a non-parallel model (the details are provided in Section IV.B.) varying the number of RUs and UEs.

Viewpoint 2 is the effect of Problem-a) and Problem-b). For Problem-a), it shows the balance between the memory access time and processing time. For Problem-b), it shows the relationship between the process waiting time and the increase in the number of RUs. From this relationship, we explain that the waiting time is reduced, and the calculation time is improved as a result.

B. Evaluation Method

As mentioned in subsection A, in order to evaluate the proposed method, four types of evaluation models, including three types of parallel model (Model-2, Model-3 and Model-4) and one non-parallel model (Model-1), which is used for comparing with the parallel processing models, are defined (shown in Figure 8). In the parallel models, considering the different configurations of execution types, such as process and/or thread, and interface types of transference memory data, such as Message Processing Interface (*MPI*) [19] and/or Open Multi-Processing (*Open*MP) [20], we set the following models:

- **Model-1) ALL threads :** Using multiple threads in one computation node, as "*non-parallel processing model*".

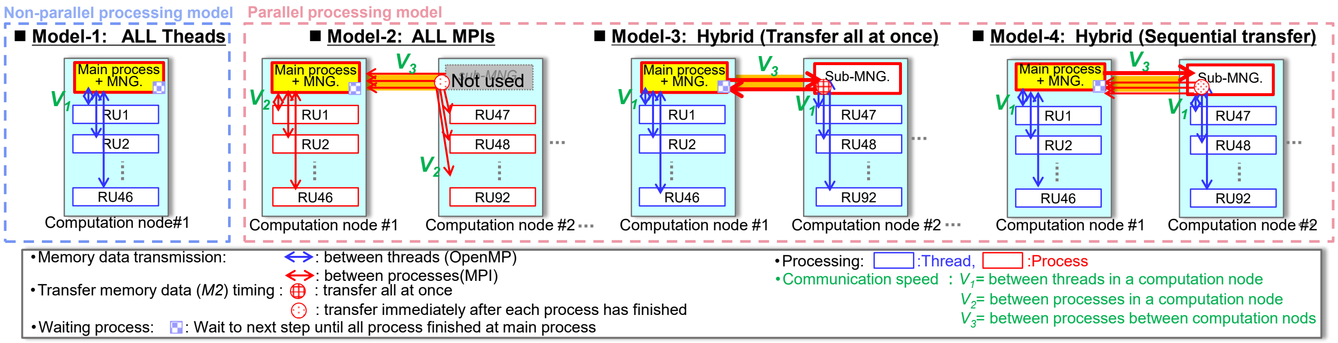

Non-parallel processing model
Parallel processing model
■ Model-1: ALL Theads
■ Model-2: ALL MPIs
■ Model-3: Hybrid (Transfer all at once)
■ Model-4: Hybrid (Sequential transfer)
Main process + MNG.
Sub-MNG.
Not used
RU1
RU2
RU46
RU47
RU48
RU92
V_1
V_2
V_3
Computation node#1
Computation node #1
Computation node #2
•Memory data transmission: ↔: between threads (OpenMP) ↔: between processes(MPI)
•Transfer memory data (M2) timing : ⊕ : transfer all at once ⊙ : transfer immediately after each process has finished
•Waiting process: ▨: Wait to next step until all process finished at main process
•Processing: □:Thread, □:Process
•Communication speed : V_1= between threads in a computation node
V_2= between processes in a computation node
V_3= between processes between computation nods

Figure 8. Parallel Processing Models (Model-1: Single computation node, Model-2, Model-3, Model-4: Multiple computation nodes).

- **Model-2) ALL-MPI :** All processes in the multiple computation nodes are connected to the management process in node #1 by MPI. Data *M2* are transferred separately to the main process when the SINR calculation has finished at each thread.

- **Model-3) Hybrid (Transfer all at once) :** Multiple threads are connected to the management process within the computation node, and each management process is connected to the management process of computation node #1 by MPI. Data *M1* and *M2* are transferred through the sub-management process together with all data of each node by one MPI. Data *M2* are transferred to the main process through the sub-management process all at once when the SINR calculation of all threads in a computation node has finished.

- **Model-4) Hybrid (Transfer immediately) :** The configuration is the same as Model-2. Data *M1* are transferred through the sub-management process together with all data of each node by one MPI. Data *M2* are transferred separately to the main process through the sub-management process when the SINR calculation has finished at each thread.

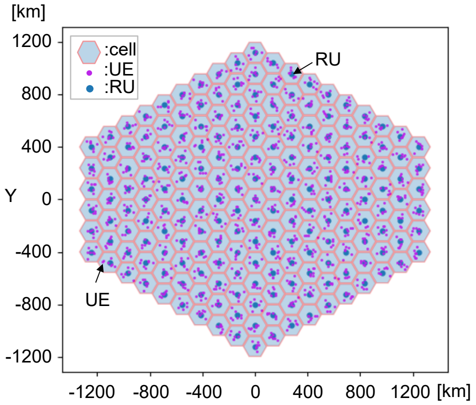


Figure 9. Example of RUs and UEs allocation in the scenario.

The evaluation scenario is that multiple RUs exist and a huge number of UEs are moving in many directions in a nano-area in the scenario (see Figure 9). We measure the processing time and waiting time when the number of RUs *R* varies from 46 to 1,012, and the number of UEs *U* varies from 690 to 55,660. We set the simulation time to 60 [min]. The other simulation parameters are shown in Table II. To obtain evaluation values, the A5G-SLS executes five times on Fugaku as shown in Table III.

TABLE II. SERVICE OPERATING AND MANAGEMENT DATA

Parameter	Value
# of RUs (*R*)	46 - 1,012 (46 RU intervals)
# of UEs (*U*)	15, 35, 55 (per RU) (Total UEs: 690 – 55,660)
# of cores	46 cores / node
#of PRBs	273
Simulation time	10 [min]
# of Time steps (*T*)	60,000
Communication speed (*Vi*)	V_1=8,192.0, V_2=159.6, V_3=50.1 [Gbps]

TABLE III. NODE SPECIFICATION ON FUGAKU@RIKEN [21]

Hardware	
Parameter	***Value***
CPU, # of Core	A64FX, 48 Cores/Node
Available # of nodes	Max 384
Node IF specification	Tofu Interconnected D (28 Gbps x 2 lane x 10 port)
Software	
Parameter	***Value***
OS	Red Hat Enterprise Linux 8
Compiler	C++ 17
MPI	FUJITSU MPI Library 4.0 (based on Open MPI)
OpenMP	OpenMP 4.5

The memory access time of the whole simulation *C* [sec] is calculated by equation (1):

$$C[sec] = MT\sum_{i=1}^{3} W_i/V_i \tag{1}$$

where, *i* = (*1,2,3*) denotes the transference points; *i* = 1 is between the management process and each thread in a computation node (Model-1, Model-3 and Model-4), *i* = 2 is between the management process and process in a computation node (Model-2), and i = 3 is between the management process and the sub-management process among two computation nodes (Model-2, Model-3 and Model-4), respectively. W_i is the number of transfer times per time step of i. V_i is the communication speed of point i [bps], each communication speed is V_1 and uses 8,192 [Gbps] referred from[17], V_2 and V_3 use the measurement values 159.6 [Gbps], and 50.1 [Gbps], respectively. Furthermore, T is the number of time steps, M is the transmission memory data size per RU [Byte] calculated by $M = M1 + M2$. From our simulation, M, $M1$, and $M2$ are the following values in Figure 10. $M1$ is the fixed size for every simulation step because it is calculated by the number of RUs and UEs. $M2$ is a different size for each simulation step, however, $M2$ size is in proportion to the number of RUs and EUs. The maximum transmission memory data size of M is approximately 13.9 [MBytes] per RU, when R is 1,012 and U is 55.

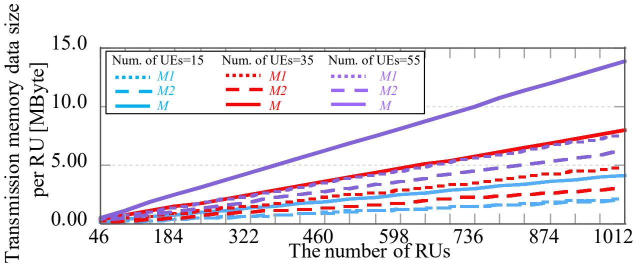


Figure 10. Transmission data size *M1*, *M2* and, *M* of each RU.

C. Simulation Results

1) Improvement of Processing Time

First, we verify that our proposed method improves the processing time compared to ALL Threads (Model-1). Figure 11 shows the processing time varying R with a comparison between the parallel processing using multiple computation nodes and the non-parallel model using one computation node. The graph is normalized to the maximum processing time. From Figure 11, the parallel processing model can reduce the processing time more than the non-parallel processing model. Although the parallel processing model and the non-parallel processing model are almost same when R is small, the difference in the processing time is larger when R increases. Specifically, in the case of R=1021 and U=35 and 15, the processing time of Model-2 can be reduced by 1.8% and 12.5% compared to Model-1, respectively. The greatest reduction at U=35 is 11.3% when R=522. In addition, the processing time of Model-3 and Model-4 is reduced compared to Model-1. When U=15, it is obtained the greatest reduction, Model-3 and Model-4 reduce the processing time by 16.16% and 17.51%, respectively. Furthermore, Model-4 achieves more reduction compared to Model-3.

However, in cases where U=55 per RU (total 45,540 UEs), when R exceeded 828, Model-2 has a longer processing time than Model-1. This reason is estimated that the data size is larger (over 10MBytes per RU process), and transmission time is increased.

From this result, we confirm that the proposed method can reduce the processing time, unless in the case of Model-2 and the transmission data size becomes extremely large (e.g., over 10 MB per RU).

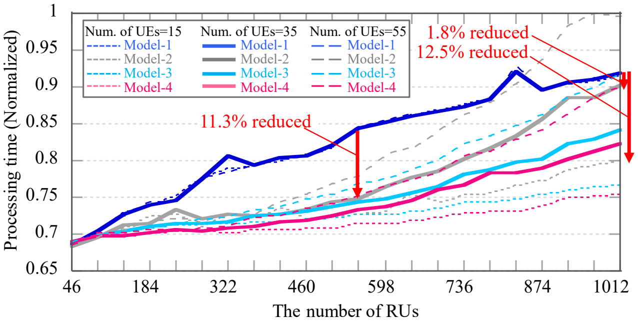


Figure 11. Improvement of SLS processing time.

2) Effectiveness of Parallel Processing

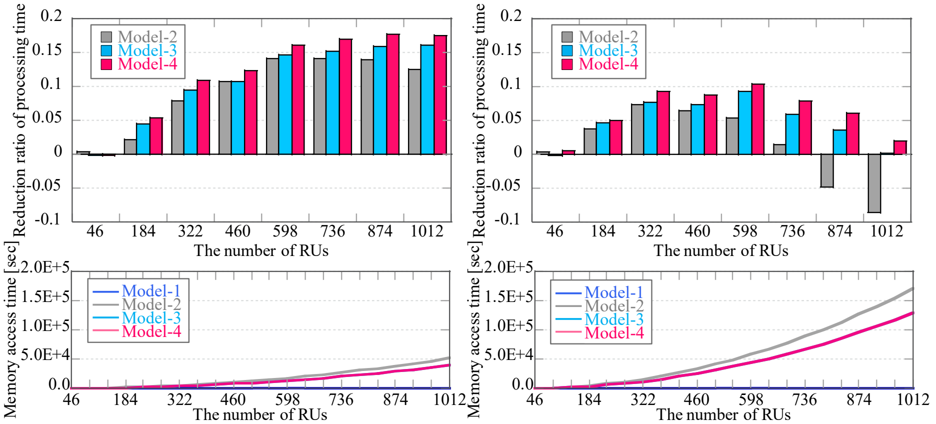
a) Memory Data Sharing

Next, the effect of processing time reduction and the balance between the memory access time and the processing time is confirmed. The upper graph in Figure 12 shows the reduction ratio of the processing time compared to Model-1, and Figure 12 below shows the memory access time. Both graphs vary R in which U=15 and 55, respectively. The reduction ratio is defined as the ratio of the difference between Model-1 and each model at each R. In the case of U=15, the reductions of processing time are obtained from Model-1. The maximum reduction ratio of the processing time is for Model-2, Model-3, and Model-4 when R=1,012, and they are 0.13, 0.16 and 0.18, respectively. On the other hand, in the case of U=55, the reduction ratio is improved compared to Model-1, and the ratio is less than when U=15. The maximum reduction ratio of the processing time is for Model-2, Model-3, and Model-4 when R=598, and they are 0.05, 0.09 and 0.10, respectively. However, these ratios are reduced after R=598, R exceeds 874 in Model-2, and the reduction ratio is less then Model-1.

The memory access time when U=15 and 55 is increased as R is increased, especially when U=55, and it is increased rapidly. When R=1,012 and U=55, the memory access time takes 3.3 times longer than when U=15. Hence, the reduction ratio is decreased due to the memory access time being increased. From this result, it can be seen that it is important to balance the processing time and memory access time.

b) Process Waiting Time

We evaluate the process waiting time of each model. Figure 13 shows the variance in waiting time varying R in the four models. The variance of waiting time for Model-2, Model-3, and Model-4 is smaller than that for Model-1. However, the variances of these models are almost constant despite the increase in the number of RUs when RU is less than or equal to 460. This result indicates that because the RU-basis divided method keeps the waiting time constant, the processing time can be improved. From another viewpoint,



(a) The number of UEs (U)=15

(b) The number of UEs (U)=55

Figure 12. Improvement of reduction ratio of processing time comparing the non-parallel processing model (Upper figure) and memory access times (Lower figure) of each UE.

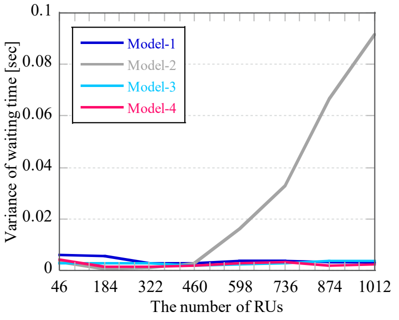


Figure 13. Process waiting time (The number of UEs (U) = 55).

Model-2 of R exceeds 460, and the variance of the waiting time is increased. We assume the reason for this is that the transmission data size when R is over 460 is large, such as 6MBytes, and the communication speed of Model-2 is slower than the other models.

From Figure 12 and Figure 13, it confirms that when the waiting time is long, the reduction of the processing time cannot be increased.

These results indicate that the hybrid Model-4 (transfer immediately) is the best method due to the balance among the reduction in processing time, memory access time and waiting time.

D. Discussion

In the above evaluation, we have verified the reduction of the processing time of A5G-SLS using three parallel processing models, Model-2, Model-3 and Model-4. Next, we evaluate quantitatively the best parallel model to execute A5G-SLS.

In order to conduct a quantitative evaluation, an evaluation score is introduced. The evaluation score E is calculated by (2) using three indications of processing time, memory access time and waiting time.

$$E = \alpha + \beta + \gamma \qquad (2)$$

where, α is the processing time, β is the memory access time and γ is the waiting time, and all indications are normalized by the maximum value of each indication. Therefore, E is from 0.00 to 3.00 due to using the three indications. The smallest E indicates the best model. The evaluation score E is calculated for each model and each RU.

Figure 14 shows the evaluation score E of each model varying the number of RUs R. From this graph, when R is small, all models obtain almost the same score for E, which is less than 1. However, when R is large, the score for E is increased. Specifically, when R=1,012 and U=55, the evaluation scores E of Model-2, Model-3 and Model-4 are 3.00, 1.72 and 1.68, respectively. In comparing two hybrid models, E of Model-4 is 0.04 smaller than that of Model-3.

From the above results, in order to reduce the processing time, especially for large-scale simulation, it is desirable to conduct the simulation using hybrid Model-4 (transfer immediately).

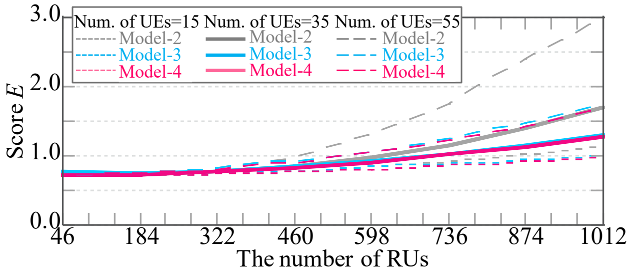


Figure 14. Score of each model.

V. Conclusion

Considering the advanced 5G system of approximately 2025, in this paper to reduce the simulation time, we propose a software design of SLS for the A5G-SLS with multiple computation nodes. Specifically, in the A5G-SLS, since the computation load of the radio communication quality measurement is high, in which the SINR calculation in the RU processing, the processing is divided into the RU-basis, and parallel processing is performed by multiple computation nodes. From our simulation results, we confirm that our proposed parallel processing method can improve the processing time compared to non-parallel processing models. However, when the number of RUs is large, the reduction ratio of the ALL_MPI model is less than that of the non-parallel processing model. In addition, as a result of the quantitative evaluation of four processing models, we verified that the hybrid Model-4 (Transfer immediately) is the best for large-scale simulations.

As a future work, in order to further reduction of the processing time, we study the reduction method of the transmission data size and the memory usage in the A5G-SLS. Furthermore, as another approach, a method is needed to reduce the processing time by abstracting the process while not affecting evaluations of the new technologies. The A5G-SLS is a closed software due to introducing our invention. Initially, it will be used for the evaluation of the advanced 5G technologies, and then we will publish our obtained research results.

ACKNOWLEDGMENT

This research is supported by the Ministry of Internal Affairs and Communications in Japan (JPJ000254). This work used computational resources of supercomputer Fugaku provided by the RIKEN Center for Computational Science through the HPCI System Research Project (Project ID: hp220145).

REFERENCES

[1] T. Murakami et al., “Research Project to Realize Various High-reliability Communications in Advanced 5G Network,” 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020, pp. 1-8.

[2] T. Hara, H. Iimori, and K. Ishibashi, “Grant-Free NOMA Using Time-Delay Domain for Low-Latency Massive Access over MIMO-OFDM,” IEEE International Conference on Communications (ICC 2022), May 2022.

[3] OMNeT++, “OMNeT++,” Simulation Models and Tools. [online] Available from: https://omnetpp.org/download/models-and-tools 2023.01.30.

[4] Network Simulator 3 (ns-3). [online] Available from: https://www.nsnam.org/ 2023.01.30.

[5] G. Nardini, G. Stea, and A. Virdis, “Scalable Real-time Emulation of 5G Networks with Simu5G”, IEEE Access, Vol. 9, pp. 148504-148520, 2021.

[6] Y. Tsukamoto, H. Hirayama, S. I. Moon, and H. Shinbo, “Adaptive Function Placement with Distributed Deep Reinforcement Learning in RAN Slicing,” 2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring), June 2022.

[7] H. Hirayama, Y. Tsukamoto, and H. Shinbo, “Feedback Control for QoS-Aware Radio Resource Allocation in Adaptive RAN,” IEEE Access, Vol. 10, pp. 21563-21573, Feb. 2022.

[8] M. K. Muller et al., “Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator,” EURASIP Journal on Wireless Communications and Networking, 227, pp.1-17, Sep. 2018.

[9] W. Tan, P. Lin, B. Liang, and H. Deng, “Influence of Network Bandwidth on Parallel Computing Performance with Intra-node and Inter-node Communication,” 2009 Second International Conference on Intelligent Networks and Intelligent Systems, pp. 534-537, Dec. 2009.

[10] Fugaku. [online] Available from: https://www.r-ccs.riken.jp/fugaku/system/ 2023.01.30.

[11] T. Ohseki, and Y. Suegara, “Fast outer-loop link adaptation scheme realizing low-latency transmission in LTE-Advanced and future wireless networks,” 2016 IEEE Radio and Wireless Symposium (RWS), pp. 1-3, Jan. 2016.

[12] M. Gutlein, R. German, and A. Djanatliev, “Performance Gains in V2X Experiments Using Distributed Simulation in the Veins Framework,” 2019 IEEE/ACM International Symposium on Distributed Simulation and Real Time Application (DS-RT), Oct. 2019.

[13] J. Lee, M. Han, M. Rim, and C. G. Kang, “5G K-SimSys for Open/Modular/Flexible System-Level Simulation: Overview and its Application to Evaluation of 5G Massive MIMO,” IEEE Access Vol. 9, pp. 94017-04032, Jun. 2021.

[14] N. Nikaein et al., “OpenAirInterface: A flexible platform for 5G research,” ACM SIGCOMM Comput. Commun. Rev. 44(5), pp.33-38, 2014.

[15] S. Lagen et al., “New Radio Physical Layer Abstraction for System-Level Simulations of 5G Networks,” 2020 IEEE International Conference on Communications (ICC), Jun. 2020.

[16] D. S. Buse, G. Echterling, and F. Dressler, “Accelerating the Simulation of Wireless Communication Protocols using Asynchronous,” Proc. MSWiM ’21, ACM, pp. 55-67, Nov. 2018.

[17] X. Guo et al., “Improving performance for simulating complex fluids on massively parallel computers by component loop-unrolling and communication hiding,” 2020 IEEE 22nd International Conference on High Performance Computing and Communications, pp.130-137, Dec. 2020.

[18] T. Saito et al., “Consideration of Data Transfer between Jobs,” IPSJ SIG Technical Report Vol. 2014-HPC-143 No.2, pp.1-6, Mar. 2014, (in Japanese).

[19] MPI Forum. [online] Available from: https://www.mpi-forum.org/ 2023.01.30.

[20] OpenMP, “Openmp application program interface version 4.6,” The OpenMP Forum Tech. Rep, 2008.

[21] Y. Nakamura, “Basic Performance of Fujitsu MPI on Fugaku,” The 7th meeting for application code tuning on A64FX computer systems, Jan. 2022.

ISBN: 978-1-68558-052-0