Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Service decoupling for open and intelligent service-based RAN*

Chunjing YUAN‡[1], Tong LEI[2], Ze XUE[1], Lin TIAN[1,3], Shuyuan ZHANG‡[4], Na LI[4], Zhou TONG[4]

[1]*Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China*

[2]*Hubei intelligent Cloud Network Operation Center,China Telecom, Wuhan 430024,China*

[3]*Nanjing Institute of InforSuperBahn, Nanjing 210008, China*

[4]*China Mobile Research Institute, Beijing 100053, China*

E-mail: yuanchunjing@ict.ac.cn; lt@gs.zzu.edu.cn; xueze22@mails.ucas.ac.cn; zhangshuyuan@chinamobile.com; tianlindd@ict.ac.cn; linawx@chinamobile.com; tongzhou@chinamobile.com

Received Apr. 1, 2024; Revision accepted Nov. 3, 2024; Crosschecked

Abstract: Task diversity is one of the biggest challenges for future 6G networks. Taking the task as the center and driving the dynamic 6G RAN (radio access network) with AI (artificial intelligence) are necessary to accurately meet the personalized demands of users. However, AI can only configure the parameters of a monolithic RAN and cannot schedule the functions. The development trend of 6G RANs is to enhance dynamic capability and scheduling ease. In this paper, we propose a service-based RAN architecture that can deploy decoupled RAN functions and customize according to tasks. Protocol analysis shows that the interactive relationship between RAN CP (control plane) functions is complex and needs to be decoupled according to the principle of high cohesion and low coupling. Based on graph theory rather than expert experience, we designed a RAN decoupling scheme. The functional connection and interaction of the CP are represented by constructing an undirected weighted graph, followed by achieving decoupling of the CP through a minimum spanning tree. Then an integration decoupling scheme of a RAN-CN is introduced considering the duplicate and redundant functions of the RAN and CN. The granularity of decoupling in a service-based RAN was determined by analyzing the flexibility of decoupling, complexity of signaling, and processing delay. We found that it is more appropriate to decouple the RAN CP into four services. The integration decoupling of the RAN-CN resolves the technical bottleneck of low serial efficiency in the Ng interface, supporting AI-based global service scheduling.

Key words: Service decoupling; Open and intelligent; Service-based RAN; Graph theory; Full service 6G

https://doi.org/10.1631/FITEE.2400248 **CLC number:**

unedited

1 Introduction

In the future, 6G AI (artificial intelligence) will be applied to a wide range of tasks, and there will be a diverse demand for RANs (radio access networks) (Uusitalo et al., 2021). For example, multimedia AI tasks require a relatively large transmission bandwidth (Yang et al., 2023). Machine control AI tasks require high timeliness of transmission (Xu et al., 2021), while those for autonomous driving have extremely high requirements for bandwidth, latency, and reliability (He et al., 2021). To support diverse AI tasks, a 6G RAN needs to have dynamic characteristics (Zhang and Zhu, 2023). The RAN needs to support the orchestration of functions based on the non-real-time requirements of AI services, as

‡ Corresponding author
ORCID: Chunjing YUAN, https://orcid.org/0000-0001-5490-971X;
© Zhejiang University Press 2024

well as the scheduling and optimization of functions based on real-time requirements. Dynamic customization of RANs for AI tasks can not only meet the task requirements but also improve the efficiency of RANs.

Due to the benefits of AI enabling RAN, the concept of an open and intelligent RAN, called O-RAN, has been proposed (Puligheddu et al., 2023). The transition to software-based, programmable, and virtualized environments will make cellular networks more flexible and dynamic systems, enabling rapid deployment and network programming (Polese et al., 2024). O-RAN is based on the idea that 6G networks can be decomposed into finer-grained microservices, which can operate relatively independently (O-RAN, 2023). 6GANA proposes that the control object of RAN systems with native-AI will change from "sessions" to "tasks" (6GANA, 2023). New bearers will be constructed for the task-centric transmission of signaling and data. RAN and CN (core network) functions will need to be pushed to the network edge and deployed in a distributed manner (6GANA, 2023). Logical separation of RAN and CN functions will be redefined for dynamic deployment at the same physical site. The AI-RAN Alliance is committed to promoting RAN performance and capability through AI (Khan and Schmid, 2024). It uses open, software-based and AI technologies to accelerate the development of new services and use cases (Upadhyaya et al.,2023). For future RANs, the above studies have proposed softwarization and decoupled architecture based on cloud-native software.

Native-AI needs an open decoupled service-based RAN (Zhang et al., 2022). Therefore, RANs need to undergo a disruptive evolution based on SBA (service-based architecture) characterized by software and decoupling (Li et al., 2022). That is, monolithic RANs are evolving into distributed service-based RANs. SDN (software defined networking) and NFV (network function virtualization) are the fundamental technologies for 6G (Cao et al., 2022). Softwarization liberates RANs from dedicated hardware. Compared to dedicated hardware, software is the key technology that supports a flexible, scalable, and sustainable evolutionary architecture. A 6G service-based RAN can not only support existing UP (user plane) and CP (control plane) functions but also extend its capabilities to include data plane, computing plane, or security plane functions (Yan et al., 2023). The modular functions of a service-based RAN can be deployed independently and scheduled flexibly. Different functions or services exhibit high interoperability (Polese et al., 2023). By using open interfaces of service-level, the network can acquire operational status and intelligently schedule network capabilities, thereby efficiently achieving the objective of an AI task.

Referring to the concept of microservices, the control plane is decoupled into a set of independent functions, including the Status Monitoring Function (SMF), Quality of Service Analysis Function (QAF), Policy Configuration Function (PCF), and Resource Control Function (RCF). By leveraging Kubernetes' support for instantiation schemes, differentiated services are provided (Ding et al., 2023). In accordance with the principles of high cohesion and low coupling, the functionalities of existing protocols have been decoupled and integrated into new services, such as RRC Message Convergence Service (RRC-MCS) and Connection Mobility Management Service (CMMS) (Du et al., 2023). Li et al. (2022) proposed a series of control plane services, including Radio Bearer Management Service (RBS), Connection Mobility Management Service (CMS), Local Location Service (LLS), Multicast Broadcast Service (MBS), Data Collection Service (DCS), Signaling Transport Service (STS), Paging Service (PS), RAN Exposure Service (RES), and Random Access Service (RAS). The control plane was also proposed to be split into the following functions: Radio Connection Control Function (RCCF), Radio Bearer Management Function (RBMF), Mobility Management Function (MMF), Radio Resource Management and Scheduling Function (RRMSF), and Radio Connection Security Function (RCSF) (Zong et al., 2023). Wang et al. (2022) proposed redefining the N2 interface to support service-based protocols, which enables direct interaction between RAN and AMF (access and mobility management) through service invocations. Khaturia et al. (2024) proposed an end-to-end service-based network architecture that includes transforming the interfaces between the RAN and CN, and enabling direct communication between the DU and CU-CP through SBI (service-based interfaces), thereby reducing

unedited

signaling latency.

Service-based RANs have been widely studied, with many researchers proposing their own decoupling solutions. However, the current state of arts is based mainly on the expertise of senior engineers. The protocol stack is decoupled based on the engineer's prior development experience with protocols and software, rather than using a theoretical approach. Therefore, is no widely accepted solution for RAN decoupling. The UP and CP have significant differences in their functional organizational architecture, requiring the design of separate decoupling schemes based on organizational characteristics. In general, the functions of the UP are sequential and can be decoupled sequentially. The functions of the CP, however, have more complex connection characteristics. The signaling procedure involves multiple network functions, each capable of appearing in multiple signaling procedures. Therefore, the focus should be on decoupling the research and design of the CP. The contributions of this paper are as follows:

1 In this paper, we examine the protocol stack characteristics and the interoperability relationship between functions. A decoupling approach, based on a theoretical framework for data analysis rather than expert experience, is proposed.

2 The functional connection and interaction of the CP are represented by constructing an undirected weighted graph, followed by achieving decoupling of the CP through a minimum spanning tree. The signaling delay performance of different decoupling granularities is evaluated.

3 The service-based RAN and CN are both SBA architectures, which can be connected and interacted through SBI. Considering the duplicate and redundant functions of the two, we studied the integration decoupling scheme of RAN and CN. A fully service-based 6G will provide an end-to-end virtualization environment for AI task programming and optimization.

2 RAN system based on service decoupling

A service-based RAN can be deployed based on general computing infrastructure. It is a distributed set of functions that can be organized by AI tasks to provide network-level services. AI tasks include AI-based decision scheduling network functions or AI-based application-based network functions. The service-based RAN is divided into two main parts: the Operation and Maintenance Domain and the Functional Domain (Fig. 1). The Operation and Maintenance Domain includes a data storage, AI decision and service library, which intelligently schedules network services by collecting and analyzing network data. The Functional Domain is the main body of the network, supporting the instantiation deployment of network services based on general computing infrastructure.

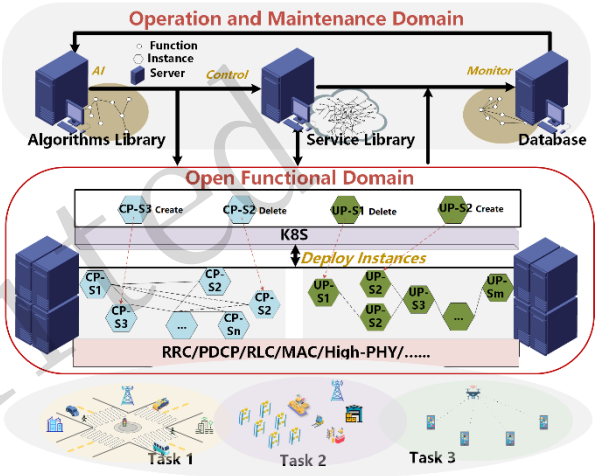


Fig. 1 Dynamic service-based RAN

2.1 Operation and maintenance domain

The distinguishing feature of a service-based RAN is that it dynamic and flexible, but the dynamic change of the network is complex. Accurate service scheduling can be achieved through AI decisions, which include network-level or function-level scheduling, optimization, and deployment. Network-level scheduling can realize simultaneous scheduling of multiple services. Function-level optimization supports scheduling of single services. The Operation and Maintenance Domain is responsible for monitoring and planning network capacity and realizes real-time and non-real-time operation and maintenance of the system according to tasks. The main functions include data management, an AI algorithm library, and a service library. Runtime data are collected at the function level, and heterogeneous data are stored in heterogeneous databases. Data

unedited

management is conducted through the use of metadata. AI algorithms generate metadata based on task demands and facilitate data management for data utilization and retrieval. The data management mechanism adopts a scheme that separates data collection and consumption. AI algorithms make network scheduling decisions based on data. The decision results are sent to scheduling execution modules, such as K8S, which pull service images from the service library, generate service instances, and deploy them in a general computing infrastructure.

2.2 Functional domain

A service-based RAN with native AI can effectively address the demanding requirements of dynamic tasks on network capacity while consistently maintaining streamlined network function and resource utilization. This requires flexible scheduling of the RAN at the functional level, allowing seamless addition, deletion, and reconstruction of services. It enables fine-grained scheduling and optimization of RAN capability through AI-based approaches or dynamic allocation of functions and resources for AI tasks. Consequently, the RAN UP and CP are all decoupled in this architecture. In the UP, service interactions follow a sequential connection mode in which uplink or downlink data undergo sequential processing. On the other hand, different signaling procedures in the CP traverse the same service, resulting in non-sequential connections between services. Leveraging general computing power infrastructure allows multiple instances of any RAN service to be deployed as needed. Services offer open interfaces for external access and are registered within the system to facilitate easy discovery and establishment of connections with other internal or external services. K8S (kubernetes) or other management methods are used to achieve service instantiation, and service instances are scheduled for deployment on the general computing infrastructure. Service instances can be dynamically created or deleted to enable function-level additions, deletions, and reconstructions as required by tasks. The functional domain exhibits scalability by connecting data functions, computing functions, or security functions with other services through registration and discovery mechanisms.

3 RAN decoupling scheme

3.1 Model of RAN CP

The RAN issues new signaling information after multiple processing when the CP receives a message from the CN or terminal. In the CP, multiple components are responsible for completing the signaling procedure (Fig. 2). Specifically, the component is accountable for processing the signaling. Input data D, after the operation O Output result A.

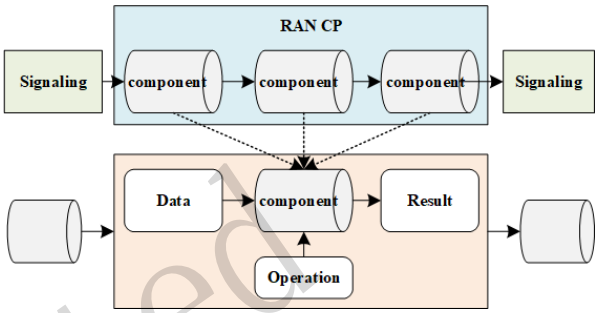


Fig. 2 CP handle signaling procedure

There are two types of data, local context and signaling-carried context, which can be represented as,

$$D = (D_C, D_S), \tag{1}$$

where D_C represents the context stored in the CP function and D_S represents the context carried in the signaling. 'Operation' generally refers to the processing of data. 'Output' generally refers to the results obtained after data processing. The results can be expressed as

$$A = (A_C, A_S), \tag{2}$$

where A_C represents the context saved by the CP function in the result, and A_S represents the context of signaling transmission in the result. There may be an overlap between the two. For example, when an RRC (radio resource control) message (Data) is received, ASN.1 (abstract syntax notation one) decoding (Operation) is called to get the specific content of the new RRC message (Output). A component is

unedited

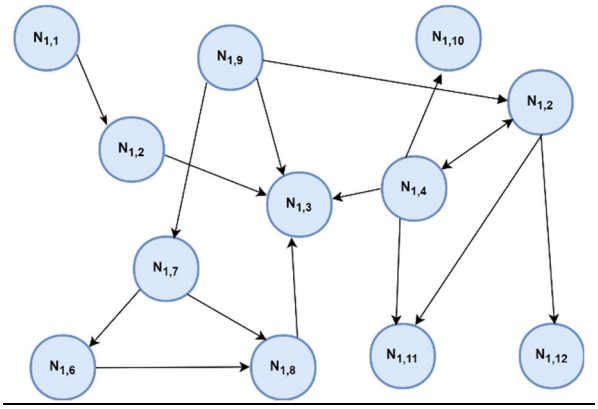
defined as a set of data, operations, and outputs.

$$N = (D, O, A). \tag{3}$$

To enhance network flexibility, the first step is to analyze and extract all components of each signaling procedure on the RAN CP. The processing of a signaling procedure needs to be forwarded and processed among different components. The signaling needs to be packaged and unpacked in each transmission, resulting in a significant network overhead and a decrease in the efficiency of signaling transmission. Simultaneously, the transmission of multiple messages will result in an increase in the total amount of signaling within the network, potentially surpassing the processing capacity of network signaling resources and leading to signaling storms. Decoupling the network by components can improve the flexibility of RAN, but it also significantly affects the network's performance. Therefore, to balance the flexibility and performance of the network, it is necessary to reasonably aggregate CP components into independently deployable services. Compared to making all components independent, this approach can significantly reduce the total amount of signaling in the network while maintaining flexibility.

3.2 Signaling procedure of RA

Analyze the RAN CP protocol stack and extract all functional components $N = \{N_1, N_2, \dots, N_m\}$ (Fig. 3). Analyze the signaling procedures $P = \{P_1, P_2, \dots, P_n\}$. Each signaling procedure P_n is composed of multiple components. The same component may appear multiple times in a single signaling procedure or may be involved in different signaling procedures.



Component	ID	Component	ID
Update measurement config	$N_{1,1}$	Update security key	$N_{1,7}$
Handover decision	$N_{1,2}$	Generation security key	$N_{1,8}$
Update PDU context	$N_{1,3}$	Check information	$N_{1,9}$
Update DRB context	$N_{1,4}$	Construct F1 signaling	$N_{1,10}$
Update SRB ccontext	$N_{1,5}$	Construct RRC signaling	$N_{1,11}$
Establish UE context (CP)	$N_{1,6}$	Construct E1 signaling	$N_{1,12}$

PDU: protocol data unit; DRB: data radio bearer; SRB: signaling radio bearer; UE:user equipment.

Fig. 3 Signaling procedure connection diagram

The undirected graph $G(E, V)$ is established with CP components serving as vertices. The correlation between vertices is expressed by the number of consecutive occurrences of CP components in different signaling procedures, i.e. the weight w_{process} between two vertices in the undirected graph G.

The frequencies of different procedures are not the same, and their influence on the decoupling of the CP also varies. Therefore, the weight between vertices in the undirected graph G is expressed as:

$$w_{ij} = \sum_{1}^{n_p} f_{P_n} * w_{process_ij}, \tag{4}$$

where, w_{ij} is the weight between functional component i and functional component j, n_p represents the number of protocol procedures, f_{P_n} is the frequency of P_n process, and $w_{\text{process_}ij}$ is the number of continuous occurrences between functional component i and functional component j in the procedure.

unedited

3.3 RAN decoupling scheme

The high coupling between components can be considered as forming a service, which makes the capabilities of the service more focused and independent. The connecting lines between components or services represent the path of message transmission in the signaling procedure, so reducing coupling between services can decrease the number of signaling transmissions and enhance network performance. Based on the expectation of low coupling between services, it is necessary to minimize the weight ω_{nf} of the connecting lines between services.

The connecting lines between vertices in the undirected graph G are sorted in descending order based on their weight. Then the connecting line with the lowest weight is removed. The number of subgraphs in the graph after removing connecting lines can be calculated by the minimum spanning tree algorithm. The components corresponding to the vertices in the subgraph are the components included in the expected services. The removed connecting lines actually serve as connections between services, ensuring low coupling between them. The retained lines are highly weighted, which guarantees high cohesion of the service.

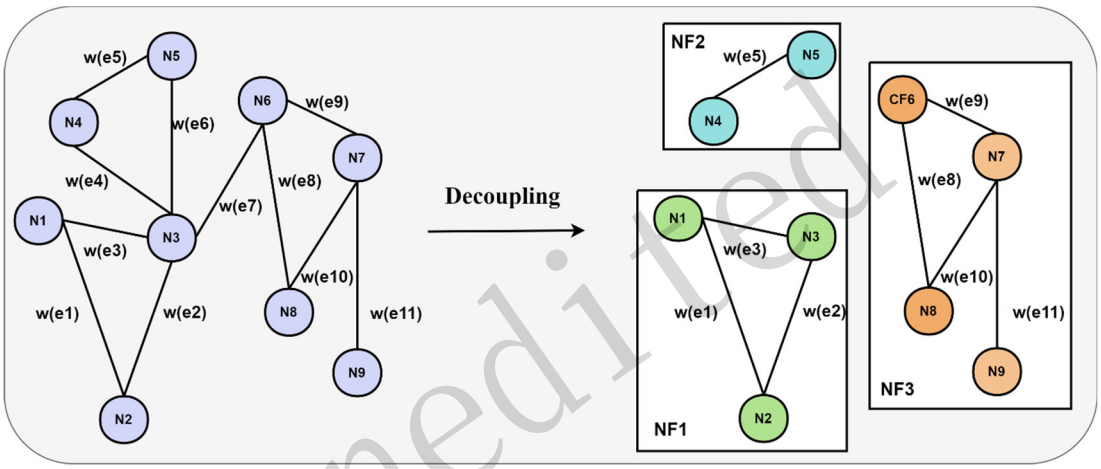


Fig. 4 Decoupling model

The algorithm itself cannot automatically determine when to stop removing connecting lines in the graph, so the number of subgraphs to be formed, namely the number of services, needs to be taken as an input parameter. Some CP components occur a few times in the procedure, or the procedure containing the component appears too infrequently. Therefore, the weight of the corresponding vertices connecting line is too low. Removing the connecting lines directly from these vertices will result in the formation of numerous "microservices" that consist of only one component or function. These "microservices" will lead to a surge of signaling in some procedures. If there are "microservices", the previously removed connecting lines are restored and the next cycle is executed. Finally, the decoupling scheme of the service-based RAN CP is determined (Fig. 4).

Algorithm 1 RAN decoupling scheme

Input: Expected number of services N; CP procedure P;
CP component O;
Output: Expected services
Calculate the weight ω between vertices, the number of lines M according to P and O;
Generate an undirected graph $G(E,V)$ from O and ω;
Sort the weight of the connection line;
While (number of subgraph=N) do;
For i=1 to M do;
Remove the connection;
If the number of subgraphs increases;
If “microservice”:
Restore line
Generate expected services based on subgraph

4 Integration decoupling scheme of RAN-CN

4.1 Interaction between the RAN and CN

The introduction of SBA in 5G CN enhances a network's flexibility. The functions of the CN communicate through the HTTP2 protocol.

unedited

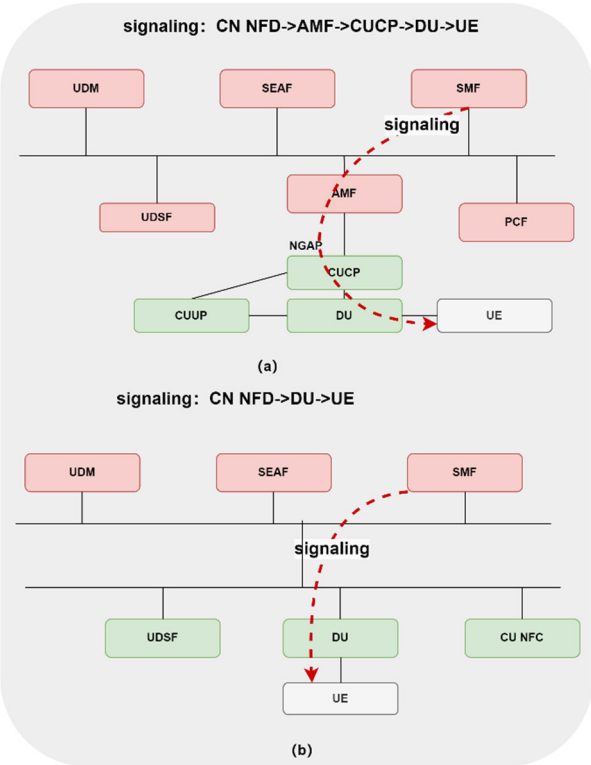


Fig. 5 Signaling processing flow

As depicted in Fig. 5a, any function within the CN that intends to establish or update RAN configuration must indirectly accomplish the task through the RAN CP. The RRC of the RAN CP exchanges signaling with the CN AMF through the NGAP (NG application protocol) interface. Considering that the RAN CP can be virtualized and deployed in the cloud with the CN after decoupling, the NGAP interface leads to inefficient network control and becomes a bottleneck.

As shown in Fig. 5b, to break the limitation of NGAP on the CP, we analyzed the possibility of the interaction of a service-based RAN and a light weight CN. RAN CP functions can communicate with CN functions through SBI. By breaking the limitation that control information of the CN must pass through AMF, 6G is promoted to become a fully decoupled and fully service-based network. For example, the process of PDU (protocol data unit) session modification needs to be operated only in the converged SMF (session management function), which controls session-related parameters in the service-based RAN and applies the operation results to UP entities. Based on the service-based RAN architecture, control signaling between the CP and SMF interacts through SBI without passing through AMF. This means that delay and signaling overhead can be expected to be reduced in the control procedure.

In private network scenarios such as intelligent manufacturing, there is demand for deploying a RAN and light weight CN in the same site. However, the current 5G architecture is designed mainly for separate deployment of the RAN and CN, which also leads to redundancy and duplication in functions. Although deployment in the same site reduces the problem of propagation delay, it still causes resource loss and complex signaling. The functions of the service-based RAN and CN are connected and interact through SBI within the framework of the fully serviced network. This provides an opportunity to merge the redundant functions of the RAN and CN. Therefore, the integration CP of the RAN and CN is proposed.

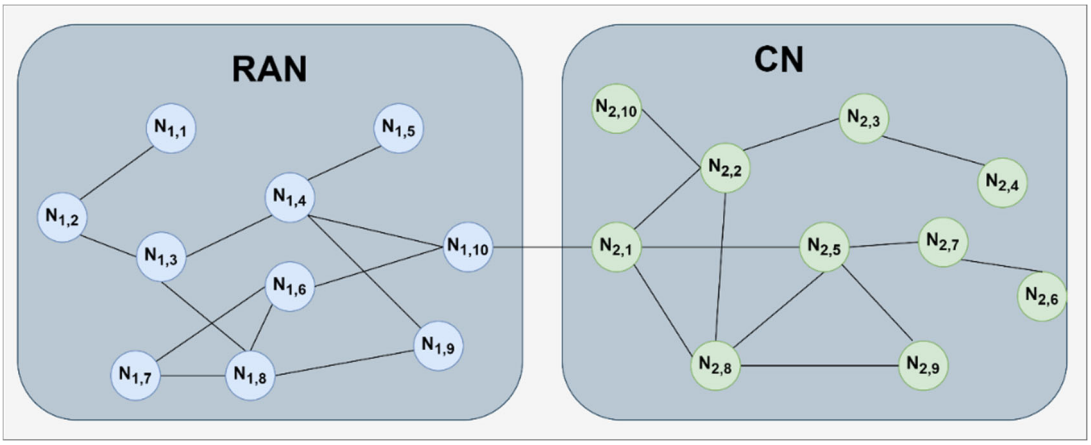
4.2 RAN and CN signaling procedure

The procedure of the RAN and CN is shown in Fig. 6. Extract RAN and CN CP components from 3GPP (the 3rd generation partnership project) protocols:

$$O = \{N_{11}, \dots, N_{1i}; N_{21}, \dots, N_{2j}\}, \tag{5}$$

where N_{1i} is a RAN CP component and N_{2j} is a CN CP component. Therefore, a signaling procedure can be defined as:

unedited



Component	Identifier	Component	Identifier	Component	Identifier
Parse NG singaling	$N_{2,1}$	Establish/Update SM context	$N_{2,5}$	PDU handover supervise	$N_{2,9}$
Establish/Update UE context	$N_{2,2}$	Provide access subscription data	$N_{2,6}$	Construct NG signaling	$N_{2,10}$
Obtain access police	$N_{2,3}$	Obtain access subscription data	$N_{2,7}$		
Provide access police	$N_{2,4}$	Configure UPF	$N_{2,8}$		

SM:Session management; UPF: user plane function

Fig. 6 Procedure undirected graph

$$P_s = \{N_{1\alpha}, \dots, N_{1\beta}; N_{2\varepsilon}, \dots, N_{2\epsilon}\}, \quad (6)$$

where, $N_{1\alpha}, \dots, N_{1\beta}$ are the component chains on the RAN side, and $N_{2\varepsilon}, \dots, N_{2\epsilon}$ are the component chains on the CN side. If the RAN sends commands to the CN, the signaling procedure is as shown in Eq. (6), if the CN sends commands to the RAN, Eq. (6) should be reversed. The undirected graph $G(E,V)$ is established with the CP components as vertices. The lines and weights on the edges represent the correlation between the CP components. The signaling interaction is used to establish the correlation.

In the signaling procedure P_s, ω_{s_ij} represents the weight extracted from V_i and V_j. The weight between V_i and V_j can be expressed as:

$$\omega_{p_ij} = \frac{\sum_1^{n_s} \omega_{s_ij}}{n_s}, \quad (7)$$

where n_s is the number of signaling interactions contained in procedure P_s. The weight of V_i and V_j in the graph can be expressed as:

$$\omega_{ij} = \frac{\sum_1^{n_p} \omega_{p_ij}}{n_p} = \frac{\sum_1^{n_p} \left(\frac{\sum_1^{n_s} \omega_{s_ij}}{n_s} \right)}{n_p}. \quad (8)$$

In this study, we selected six signaling procedures, including “RRC Establishment”, “PDU Establishment”, “RRC Inactive State Transition”, “RRC Recovery”, “PDU Modification”, and “AMF Handover”.

4.3 Integration decoupling scheme

The traditional monolithic RAN and CN are two distinct systems with different logics. For the same event, RAN and CN maintain different contexts. For example, for user information, CN uses PDU description and RAN uses DRB description. For the integration of RAN and CN CP, the connection between the contexts of RAN and CN must first be established. The interaction between components is graphed using signaling procedures, similar to Section 3, and the functions of the CN and RAN are aggregated together

unedited

through graph clustering. The aggregated components form a CP service.

Some components appear more frequently and others less frequently. The components that appear more frequently must be able to reflect their association with other components in the procedure. For components that appear less frequently, the overall perspective is that they are less relevant than other components. Therefore, when establishing the correlation between the components of the RAN and CN, the components that appear fewer times (represented as outliers in the Fig. 7) should be excluded first.

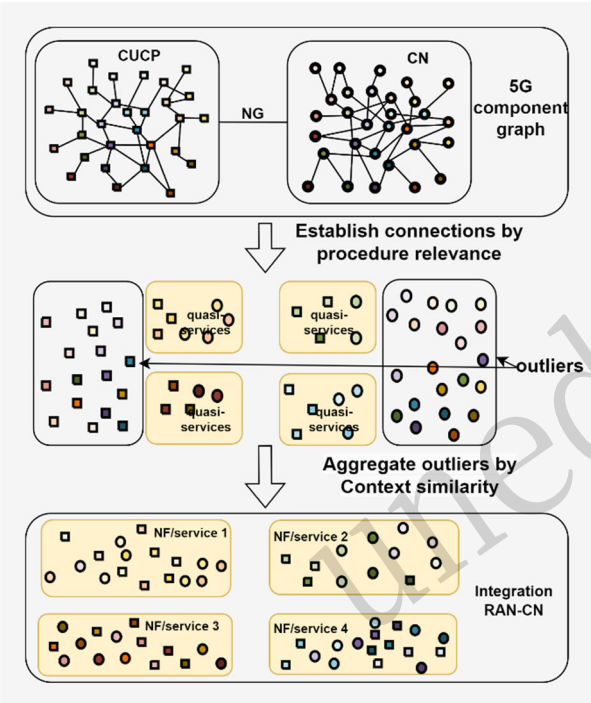


Fig. 7 Integration decoupling steps

Fig. 7 shows the steps of integration decoupling of the RAN-CN. The colored square represents the RAN component, and the colored sphere represents the CN component. The CN component is extracted mainly from various services of the CN but does not correspond to all services of the CN.

After the graph is established, the weight of the edge is normalized. The normalized weight of the edge is:

$$\omega' = \frac{\omega - min\ (\omega)}{max(\omega) - min\ (\omega)}, \tag{9}$$

where ω is the edge weight and ω' is the normalized edge weight. ω_0 is the threshold value for determining outliers. All edges with $\omega' < \omega_0$ are removed, and the points not connected to the graph are considered outliers. The removal of outliers ensures that the graph remains connected. The edges in the connected graph are sorted in order of weight from high to low. The number of subgraphs in the graph is calculated by determining the minimum spanning tree and removing its edges. When the number of corresponding subgraphs is equal to the expected number of services, quasi-services are generated. The algorithm of Integration Decoupling is as follows:

Algorithm 2 Integration decoupling

Input: Expected number of services N; CP procedure P_s; CP component O;
Output: Quasi-services
Calculate the weight ω between vertices, the number of lines M according to P_s and O;
Generate an undirected graph $G(E,V)$ from O and ω_{ij};
Exclusion of outliers;
Sort the weight of the connection line;
While (number of subgraph=N) do;
For i=1 to M do;
Removal the connection;
Generate quasi-services based on subgraph

Because removing outliers may result in the removal of some functional components, the quasi-service obtained at this stage is incomplete. It is necessary to put outliers back into the quasi-service according to the context similarity between outlier components and the quasi-service. Algorithm 3 describes this process. First, the relevance between components and quasi-service is calculated. The Jaccard coefficient is used to measure the similarity between components and quasi-services.

$$S = \frac{(D_o, A_o) \cap (D_{\mathrm{NF}}, A_{\mathrm{NF}})}{(D_o, A_o) \cup (D_F, A_F)}, \tag{10}$$

(D_o, A_o) represents the context in the component, while (D_F, A_F) represents the context in the quasi-service. The intersection of these two contexts, denoted by $(D_o, A_o) \cap (D_F, A_F)$, signifies the degree of overlap in terms of words or elements, indicating the extent to which the component and quasi-service share common functionality or data dependencies. The union of the two contexts, represented by

unedited

$(D_o, A_o) \cup (D_F, A_F)$, encompasses all the elements from both contexts. The outliers that have a similarity of more than 50% are added to the quasi-service. The exclusion of more outliers during the outlier exclusion stage could result in the failure of outlier matching. After each round of matching, the context similarity between the outlier and the quasi-service is recalculated and rematched until all outliers are placed in the quasi-service.

Algorithm 3 Outlier matching

Input: Quasi-services; outlier V
Output: Expected services
1 Calculate context relevance
2 While (Outliers exist) do;
3 for (int i=0; i<k; i++)
4 if (similarity S>50%)
5 Outlier points are placed in quasi-service
6 Recalculate context relevance
7 Generate expected services

5 Analysis and result

The ICT/CAS (Institute of Computing Technology, Chinese Academy of Sciences) has developed 5G RAN protocol stack software that conforms to the 3GPP standard. A part of the protocol stack has been published as open source in OpenXG. Based on the analysis of protocol stack software and standard documents, we selected some common procedures as objects for analysis and modeling. The selected procedures are the most important ones in a real wireless network environment. The labels of procedure are shown in Table 1.

Table 1 Labels of procedure

Procedure	Label
RRC establish	$P1$
RRC inactivate	$P2$
RRC resume	$P3$
PDU establish	$P4$
PDU modification	$P5$
Cell/AMF handover	$P6$

5.1 Flexibility analysis

The monolithic RAN is decoupled into a set of services, and the resulting decoupling is evaluated. Different services can be generated for different decoupling granularity. In most commercial systems, the ratio between each procedure is 1:1:1:1:2:72 (Choi et al., 2022). The decoupling granularity was 3, 4 and 5. The decoupling results are shown in Table 2.

unedited

Table 2 Decoupling results of different granularity

Decoupling granularity	Network service	Component			
3 ($E3$)	NF 3-1	Establish/Update UE context	Construct NG signaling	Select AMF	Generation security key
		Update security key			
	NF 3-2	Construct RRC signaling	Update DRB context	Construct $F1$ signaling	Update DRB context
		Check information	Construct $E1$ signaling	Update PDU context	
	NF 3-3	Handover decision	Update measurement config		
4 ($E4$)	NF 4-1	Establish/Update UE context	Construct NG signaling	Select AMF	
	NF 4-2	Generation security key	Update security key		
	NF 4-3	Construct RRC signaling	Update SRB context	Construct $F1$ signaling	Update DRB context
		Check information	Construct $E1$ signaling	Update PDU context	
	NF 4-4	Handover decision	Update measurement config		

To be continued

Table 2 (continued)

5 ($E5$)	NF 5-1	Establish/Update UE context	Construct NG signaling	Select AMF	Update PDU context
	NF 5-2	Generation security key	Update security key		
	NF 5-3	Construct RRC signaling	Update SRB context	Construct $F1$ signaling	
	NF 5-4	Update DRB context	Check information	Construct $E1$ signaling	
	NF 5-5	Handover decision	Update measurement config		

When the decoupling granularity is set to 3, the service NF 3-3 comprises "Handover Decision" and "Update Measurement Config", both of which are associated with handover. In scenarios such as smart factories or smart homes, in which the terminals are stationary or have limited mobility, the functionality related to handover is not required. Therefore, the deployment of NF 3-3 can be omitted, which makes the network lighter and simpler.

When the decoupling granularity is set to 4, it can be observed that the network services NF 4-1 and NF 4-2 are essentially derived from the splitting of the network service NF 3-1 at a decoupling granularity of 3. NF 4-1 consists mainly of "Establish/Update UE Context" and "NAS message construction". The possible reason for this is that in certain procedures, the "Update UE Context" results in changes to the UE's selection of the AMF.

Consequently, CP functionality for "Construct NG Signaling" follows immediately. The NF 4-2 is associated mainly with network security, demonstrating two aspects of flexibility:

1 In extreme cases, such as emergency calls, bypassing network functions related to security can improve network access speed.

2 The independent deployment of security-related network functions allows for the flexible addition, deletion, and upgrading of security and authentication algorithms without interfering with other network functions. This design can enhance network security.

When the decoupling granularity is set to 5, the actions of "Update DRB Context" and "Update SRB Context" should be grouped within the same network service. Similarly, "Construct $E1$ Messages", "Construct $F1$ Messages" and "Construct Air Interface Message" should be grouped within the same network service. However, from a procedural perspective, message construction and transmission are interrelated, and various functions for constructing messages may not occur consecutively. As a result, their association may not be immediately apparent, resulting in their allocation to different expected network services.

5.2 Decoupling performance

Six procedures were selected for decoupling and numerical evaluation. According to 3GPP, (2024), the processing delay of signaling within a network function during simulation is 3 ms, while the propagation delay of signaling between network functions is 1 ms. The processing delay for a single procedure is:

$$t = \sum_{1}^{n} t_p + t_s + t_q, \tag{11}$$

where n is the total number of signaling messages, t_p is the processing delay of the signaling, t_s is the transmission delay of the signaling, and t_q is the queuing delay at the next network function after receiving the signaling. The queuing delay is directly proportional to the number of signaling messages in the message queue.

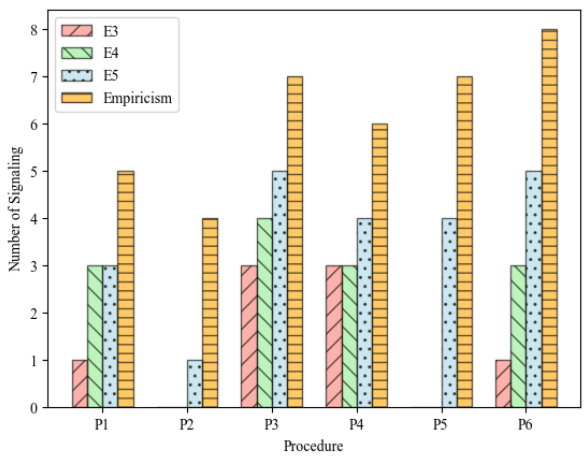


Fig. 8 Number of serviced RAN signaling messages

unedited

The number of CP signaling messages for different decoupling granularity under a light load is shown in Fig. 8. Compared to $E3$, $E4$, and $E5$, the use of expert experience in decoupling the RAN results in the highest number of signaling messages. When decoupling is done with expert experience, components are divided into more services, leading to an increase in the number of extra signaling messages. In $E3$, there are 10 CP signaling messages. In $E4$, that number increases by 30% to 13. However, in $E5$, the number of CP signaling messages rises sharply to 23, marking a 130% increase. In both $E3$ and $E4$, there are no CP signaling transmissions in the "RRC Inactive Transition" and "PDU Modification" procedures. This is because the CP signaling needs to be processed by only one network function before being routed to the CN or UE.

Fig. 9 shows the processing time of CP procedures at different decoupling granularity under a light load. The processing time is defined by Eq. (11), where t_q equals 0. Compared to $E3$, $E4$, and $E5$, the processing latency is highest when expert experience is used to decouple the RAN. The expert experience decoupling involves more services, resulting in a significant increase in additional transmission time. In both $E3$ and $E4$, the "RRC Inactive Transition" and "PDU Modification" procedures involve fewer signaling messages, resulting in lower CP processing times. For these procedures, $E3$ and $E4$ show performance equivalent to a monolithic RAN. For more frequent events such as "Cell Handover" and "PDU Modification", the processing times of $E3$ and $E4$ are identical. Similarly, in scenarios with less frequent events such as "RRC Establishment" and "PDU Establishment", the processing times of $E4$ and $E5$ are identical. We conclude that under a light load, the performance of $E3$ and $E4$ is comparable, while the performance of $E5$ is inferior.

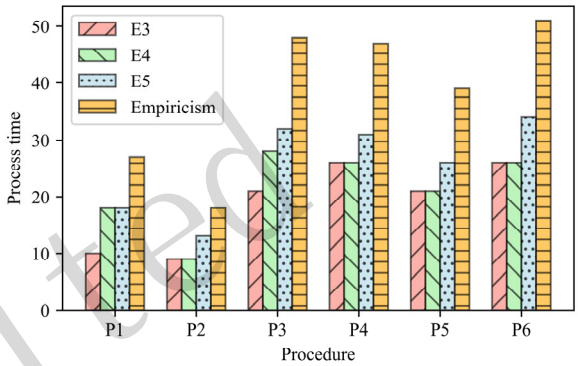


Fig. 9 Service-based RAN CP processing time

unedited

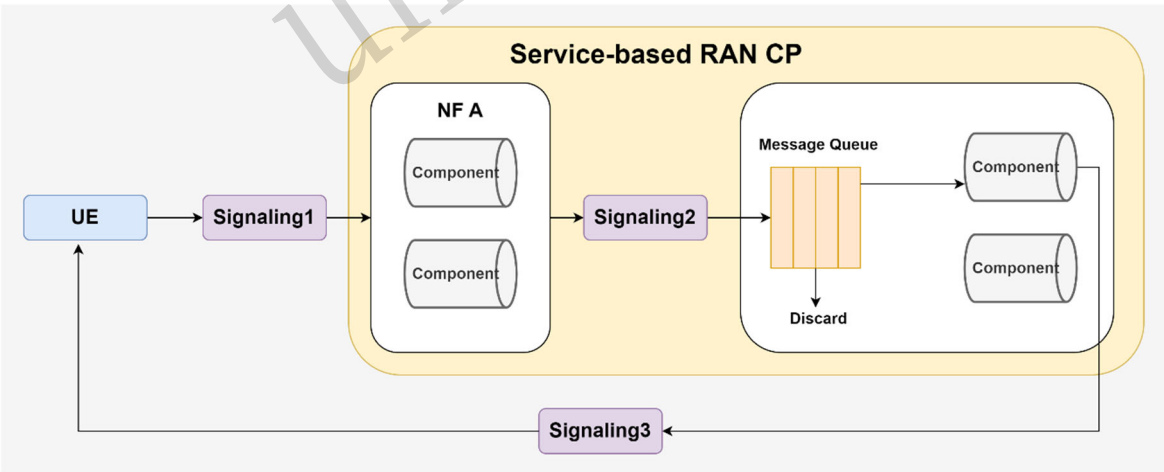


Fig.10 Service-based RAN signaling procedure

Fig. 10 illustrates the signaling procedure of a service-based RAN. The UE sends signaling message 1 to service A, which then generates signaling message 2 and forwards it to service B. Upon receiving signaling message 2, service B places it into a message queue. After signaling message 2 has been queued and processed by internal components, signaling message 3 is generated and sent back to the UE. The process from sending signaling message 1 to the UE receiving signaling message 3 can be

considered as one cycle.

If the message queue is full, service B will immediately discard signaling message 2 without generating signaling message 3. After a certain period of time without receiving signaling message 3, the UE will request to send signaling message 1 again, thus re-starting the procedure. Multiple requests from the UE will cause the message queue to be constantly full, resulting in many signaling messages being discarded and network congestion.

Occasional queuing delays may cause timers in the UE to expire, resulting in signaling retransmissions. The retransmissions will reactivate the signaling procedure. Therefore, the overall time delay of the process is:

$$t_t = \sum_{1}^{m-1} t_o + t, \tag{12}$$

where m represents the total number of retransmissions, t denotes the processing time of the last retransmission, and t_o stands for the timeout period, which refers to the time for terminal timers to expire and resend signaling messages. In this experiment, the network service's queue length was set to 10. Congestion occurs when the number of waiting signaling messages exceeds 10. In cases of network function congestion, any new signaling message received will be discarded, resulting in signaling retransmissions.

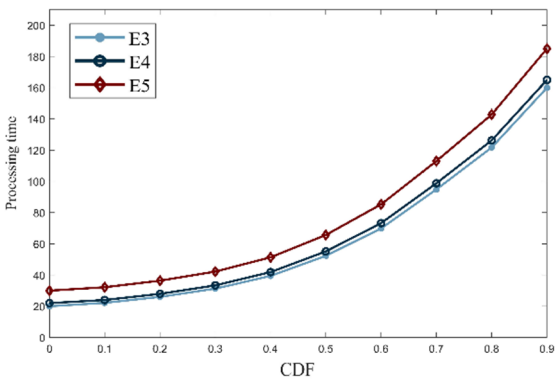


Fig.11 Call frequency and processing time in congestion

Fig.11 illustrates the processing delay at various decoupling granularities during congestion. In this scenario, 100 UEs request services at different frequencies. The total processing delay of the CP is defined by Eq. (12). The CDF (cumulative distribution function) represents the probability of each UE issuing a signaling request per second. It is evident that the processing time sharply increases when the CDF exceeds 0.3, indicating congestion in some network services.

The processing times of $E3$ and $E4$ are similar, which can be attributed to two reasons. Firstly, under a light load, the processing times of $E3$ and $E4$ are similar. Secondly, in $E3$, each network service needs to process more signaling messages compared to $E4$. In situations with high load or high signaling density, $E3$ is more susceptible to congestion and subsequent retransmissions. Each network function in E4 must process a higher volume of signaling messages, which may result in congestion and longer processing times.

Fig. 12 compares the average processing delay between a service-based RAN and a monolithic RAN. The average processing delay represents the time it takes for different procedures initiated by the UE to be completed. The average processing delay:

$$t_a = \frac{\sum_1^n t_i * \omega_i}{\sum_1^n \omega_i}, \tag{13}$$

where t_i represents the processing time of a certain procedure under a light load, and ω_i represents the occurrence frequency of the procedure. The performance of $E3$ and $E4$ is similar to that of a monolithic RAN. Due to the excessively fine granularity of decoupling in $E5$, its performance is inferior. The evaluation results indicate that $E4$ is the optimal decoupling result. Its performance is comparable to that of the monolithic RAN, offering greater flexibility in terms of handover and security.

unedited

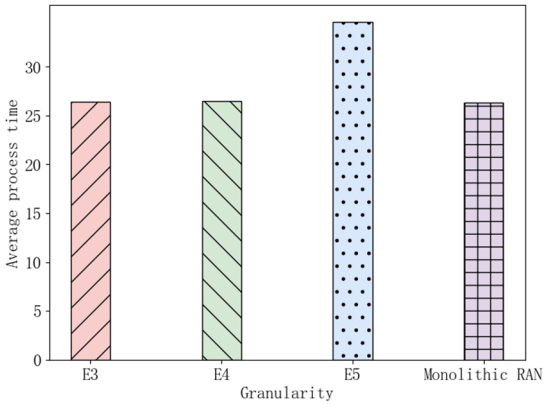


Fig. 12 Average processing delay

5.3 Integration decoupling performance

During the phase of generating quasi-services, when the granularity of the service is 4 or higher, some of the generated quasi-services consist of only one component, indicating that these quasi-services are actually outliers. The outliers will be integrated with quasi-services, and the result is the same as with a granularity of 3. Therefore, in the performance evaluation part, the granularity was 3. The integration RAN-CN CP services are shown in Table 3.

Table 3 Integration RAN-CN CP services

NF	Component		
NF A	Establish/Update UE context (CP)	Obtain access police	AMF select
	Establish/Update UE context (CN)	Provide access police	Update measurement config
	Provide access subscription data	Obtain access subscription data	
NF B	Update DRB context	Update SRB context	Update PDU context
	Configure UPF	Provide SM subscription data	Obtain QOS subscription data
	Provide QOS polices	Select UPF	PDU handover supervise
NF C	Establish/Update SM context	Obtain SM subscription data	
	Generation security key	Update security key	Calculate authentication vector
	Select SSC	Activate integrity protection	

QOS: quality of service; SSC: session and service continuity

unedited

The request ratio for each procedure was set to 1:1:1:1:2:72 (Choi et al., 2022). Due to significant structural changes in the network, messages are no longer being forwarded through the AMF. Table 4 presents the processing and propagation delays among different network functions (3GPP, 2024).

Table 4 Processing delay and propagation delay

Processing delay (ms)		
Entity	5G	Integration CP
UE	RRC Up Link: 2, RRC Down Link: 5	
DU	1	1
CU	2	N/A
NF	3	3
Propagation delay (ms)		
Interface	5G	Integration CP
UE-DU	0.5	
DU-CU	10	10
CU-NF	1	
NF-NF	1	

The numbers of signaling messages for the monolithic RAN, service-based RAN, and integration RAN-CN CP are shown in Fig. 13. In the monolithic RAN, there is no signaling transmission in the CP. The signaling messages interact between the RAN and the CN, as well as internal CN services. The

three procedures of “RRC Establishment”, “RRC Inactive State Transition”, and “RRC Recovery” are related to the CP and have no connection with CN, so the total number of signaling interactions is small. In the “PDU Establishment”, “PDU Change” and “PDU Handover Processes”, which benefit from the simplified network structure, the integration CP reduces the signaling events by 34.5% and 36.8% compared to the monolithic and service-based RANs, respectively.

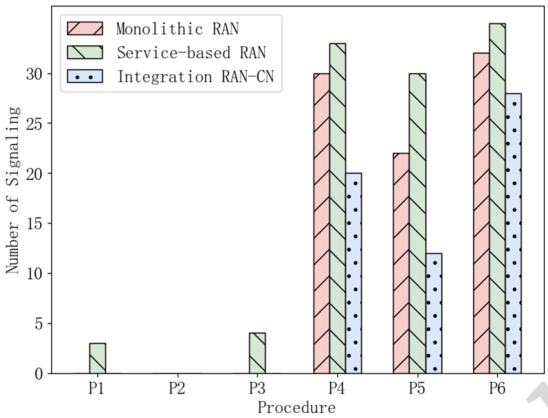


Fig. 13 Number of CP Signaling messages

The total processing delay for the three entities is shown in Fig. 14, where the processing delay is calculated according to Eq. (11). Compared to the monolithic RAN and the service-based RAN, the integration RAN-CN reduces the processing times by 17.9% and 23.0% respectively.

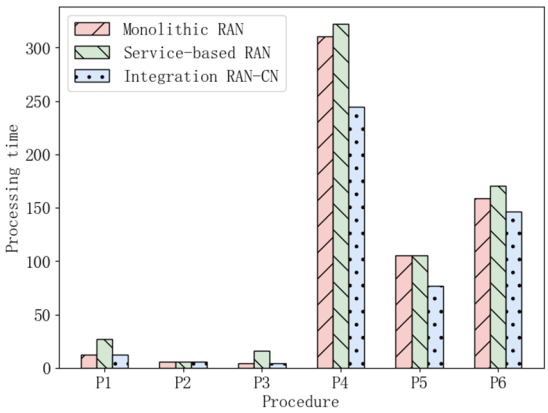


Fig. 14 Processing delay

6 Conclusions

In this paper, we propose a virtualization and decoupling service-based RAN architecture to meet dynamic task requirements. Services can be independently deployed, achieving high flexibility and schedulability at the function level based on AI. It offers an open RAN environment for AI optimization and a customized network. The decoupling of the RAN and the integration decoupling of RAN-CN are clarified by analyzing the protocol stack. A decoupling scheme based on a minimum spanning tree is proposed. The performance of the service-based RAN is comparable to that of the monolithic RAN, and offers greater flexibility in terms of handover and security. The integration decoupling breaks through the serial signaling bottleneck for both, overcoming the restrictions of the Ng interface. Under the six selected common functions, considering the number of signaling messages and delay, we recommend decoupling RAN CP into four services, and RAN-CN CP integration decoupling into three services.

unedited

Contributors

Chunjing YUAN and Tong LEI designed the research. Tong LEI and Ze XUE processed the data. Chunjing YUAN drafted the manuscript. Lin TIAN helped organize the manuscript. Lin TIAN revised and finalized the paper. Shuyuan ZHANG, Na LI, Zhou TONG helped revise the paper.

Conflict of interest

All authors declare that they have no conflict of interest.

References

3GPP, 2024. Study on NR positioning enhancements. TR 38.857. Generation Partnership Project (3GPP).

6GANA, 2023a. User-Centric Friendly Network (UCFN)-Concept and Requirements.

6GANA, 2023b. White paper on Task-oriented Intelligent-native RAN Architecture.

Cao HT, Du JB, Zhao HT, et al., 2022. Toward tailored resource allocation of slices in 6G networks with softwarization and virtualization. *IEEE Internet Things J*, 9(9):6623-6637.

https://doi.org/10.1109/JIOT.2021.3111644

Choi J, Sharma N, Gantha SS, et al., 2022. RAN-CN converged control-plane for 6G cellular networks. IEEE Global Communications Conf, p.1253-1258. https://doi.org/10.1109/GLOBECOM48099.2022.10001281

Ding HY, Wang YF, Zheng XY, et al., 2023. Design and implementation of a service-based radio access network. IEEE 97th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTC2023-Spring57618.2023.10199509

Du KL, Wang LH, Zhu ZS, et al., 2023. Converged service-based architecture for next-generation mobile communication networks. IEEE Wireless Communications and Networking Conf, p.1-6. https://doi.org/10.1109/WCNC55385.2023.10118793

He JH, Yang K, Chen HH, 2021. 6G cellular networks and connected autonomous vehicles. *IEEE Netw*, 35(4):255-261. https://doi.org/10.1109/MNET.011.2000541

Khan NA, Schmid S, 2024. AI-RAN in 6G networks: state-of-the-art and challenges. *IEEE Open J Commun Soc*, 5:294-311. https://doi.org/10.1109/OJCOMS.2023.3343069

Khaturia M, Sharma N, Choi J, et al., 2024. Service-based architecture evolution: towards enhanced signaling in beyond 5G/6G networks. IEEE Wireless Communications and Networking Conf, p.1-6. https://doi.org/10.1109/WCNC57260.2024.10571233

Li N, Liu GY, Zhang HM, et al., 2022. Service-based RAN: the next phase of cloud RAN. IEEE Globecom Workshops, p.1206-1211. https://doi.org/10.1109/GCWkshps56602.2022.10008666

O-RAN, 2023. O-RAN next Generation Research Group (nGRG) Research Report: O-RAN Towards 6G.

Polese M, Bonati L, D'Oro S, et al., 2023. Understanding O-RAN: architecture, interfaces, algorithms, security, and research challenges. *IEEE Commun Surv Tut*, 25(2):1376-1411. https://doi.org/10.1109/COMST.2023.3239220

Polese M, Dohler M, Dressler F, et al., 2024. Empowering the 6G cellular architecture with open RAN. *IEEE J Sel Areas Commun*, 42(2):245-262. https://doi.org/10.1109/JSAC.2023.3334610

Puligheddu C, Ashdown J, Chiasserini CF, et al., 2023. SEM-O-RAN: semantic and flexible O-RAN slicing for NextG edge-assisted mobile systems. IEEE Conf on Computer Communications, p.1-10. https://doi.org/10.1109/INFOCOM53939.2023.10228870

Upadhyaya PS, Tripathi N, Gaeddert J, et al., 2023. Open AI Cellular (OAIC): An open source 5G O-RAN testbed for design and testing of AI-based ran management algorithms. *IEEE Netw*, 37(5):7-15. https://doi.org/10.1109/MNET.2023.3320933

Uusitalo MA, Rugeland P, Boldi MR, et al., 2021. 6G vision, value, use cases and technologies from European 6G flagship project Hexa-X. *IEEE Access*, 9:160004-160020. https://doi.org/10.1109/ACCESS.2021.3130030

Wang XY, Sun T, Duan XD, et al., 2022. Holistic service-based architecture for space-air-ground integrated network for 5G-advanced and beyond. *China Commun*, 19(1):14-28. https://doi.org/10.23919/JCC.2022.01.002

Xu HS, Wu J, Li JH, et al., 2021. Deep-reinforcement-learning-based cybertwin architecture for 6G IIoT: An integrated design of control, communication, and computing. *IEEE Internet Things J*, 8(22):16337-16348. https://doi.org/10.1109/JIOT.2021.3098441

Yan XQ, An XL, Ye WX, et al., 2023. User-centric network architecture design for 6G mobile communication systems. Joint European Conf on Networks and Communications & 6G Summit, p.305-310. https://doi.org/10.1109/EuCNC/6GSummit58263.2023.10188283

Yang ZM, Hu DL, Guo Q, et al., 2023. Visual E^2C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. *IEEE Wirel Commun*, 30(3):204-210. https://doi.org/10.1109/MWC.019.2200518

Zhang HM, Liu GY, Li N, et al., 2022. Performance analysis of service-based RAN via multi-state Markov chain. IEEE 8th Int Conf on Computer and Communications, p.1561-1565. https://doi.org/10.1109/ICCC56324.2022.10066043

Zhang X, Zhu QX, 2023. AI-enabled network-functions virtualization and software-defined architectures for customized statistical QoS over 6G massive MIMO mobile wireless networks. *IEEE Netw*, 37(2):30-37. https://doi.org/10.1109/MNET.005.2200408

Zong JY, Huang X, Liu HT, et al., 2023. Service-oriented wireless network architecture and edge network convergence design. IEEE Int Symp on Broadband Multimedia Systems and Broadcasting, p. 1-5. https://doi.org/10.1109/BMSB58369.2023.10211242

unedited