This article has been accepted for inclusion in a future issue of this magazine.

6G Network AI Architecture for Everyone-Centric Customized Services

Yang Yang[1,2,3,†], Mulei Ma[1], Hequan Wu[4], Quan Yu[3,4], Ping Zhang[5,3,4], Xiaohu You[6,7,3], Jianjun Wu[8], Chenghui Peng[8], Tak-Shing Peter Yum[6], Sherman Shen[9], A. Hamid Aghvami[10], Geoffrey Y Li[11], Jiangzhou Wang[12], Guangyi Liu[13], Peng Gao[13], Xiongyan Tang[14], Chang Cao[14], John Thompson[15], Kat-Kit Wong[16], Shanzhi Chen[17], Merouane Debbah[18], Schahram Dustdar[19], Frank Eliassen[20], Tao Chen[21], Xiangyang Duan[22], Shaohui Sun[17], Xiaofeng Tao[5,3], Qinyu Zhang[23,3], Jianwei Huang[24], Shuguang Cui[24,3], Wenjun Zhang[25], Jie Li[25], Yue Gao[26,3], Honggang Zhang[27], Xu Chen[28], Xiaohu Ge[29,3], Yong Xiao[29,3], Cheng-Xiang Wang[6,7], Zaichen Zhang[6,7], Song Ci[30], Guoqiang Mao[31], Changle Li[31], Ziyu Shao[1], Yong Zhou[1], Junrui Liang[1], Kai Li[1], Liantao Wu[1], Fanglei Sun[1], Kunlun Wang[32], Zening Liu[7], Kun Yang[33], Jun Wang[16], Teng Gao[34], Hongfeng Shu[35]

1. ShanghaiTech University, China
2. Terminus Group, China
3. Peng Cheng Laboratory, China
4. Chinese Academy of Engineering, China
5. Beijing University of Posts and Telecommunications, China
6. Southeast University, China
7. Purple Mountain Laboratories, China
8. Huawei Technologies, China
9. University of Waterloo, Canada
10. King's College London, UK
11. Imperial College London, UK
12. University of Kent, UK
13. China Mobile, China
14. China Unicom, China
15. The University of Edinburgh, UK
16. University College London, UK
17. China Academy of Telecommunication Technology, China
18. Technology Innovation Institute, UAE
19. TU Wien, Austria
20. University of Oslo, Norway
21. VTT Technical Research Centre of Finland, Finland
22. ZTE Corporation, China
23. Harbin Institute of Technology (Shenzhen), China
24. The Chinese University of Hong Kong (Shenzhen), China
25. Shanghai Jiao Tong University, China
26. Fudan University, China
27. Zhejiang Lab, China
28. Sun Yat-sen University, China
29. Huazhong University of Science and Technology, China
30. Tsinghua University, China
31. Xidian University, China
32. East China Normal University, China
33. University of Electronic Science and Technology of China, China
34. Fuzhou Internet of Things Open Lab, China
35. Shenzhen Smart City Technology Development Group, China

Abstract—**Mobile communication standards were developed for enhancing transmission and network performance by using more radio resources and improving spectrum and energy efficiency. How to effectively address diverse user requirements and guarantee everyone's Quality of Experience (QoE) remains an open problem. The Sixth Generation (6G) mobile systems will solve this problem by utilizing heterogenous network resources and pervasive intelligence to support everyone-centric customized services anywhere and anytime. In this article, we first coin the concept of Service Requirement Zone (SRZ) on the user side to characterize and visualize the integrated service requirements and preferences of specific tasks of individual users. On the system side, we further introduce the concept of User Satisfaction Ratio (USR) to evaluate the system's overall service ability of satisfying a variety of tasks with different SRZs. Then, we propose a network Artificial Intelligence (AI) architecture with integrated network resources and pervasive AI capabilities for supporting customized services with guaranteed QoEs. Finally, extensive simulations show that the proposed network AI architecture can consistently offer a higher USR performance than the cloud AI and edge AI architectures with respect to different task scheduling algorithms, random service requirements, and dynamic network conditions.**

I. INTRODUCTION

Recently, the global development and application of Internet of Things (IoTs) have accelerated the digitalization of the physical world and human society. To exploit the commercial values of massive IoT data, we use Artificial Intelligence (AI) algorithms to integrate user requirements, domain knowledge, operation procedures, and business models in different application scenarios. To improve user satisfaction in public services, data from user devices and public facilities can be utilized by self-learning algorithms to meet each user's personal requirements and preferences [1]. For manufacturing applications, data from industrial automated control devices in assembly lines can be analyzed by AI algorithms to improve efficiency, productive force, and safety, and to reduce cost, energy consumption, and carbon emissions. Eventually, a digital world will emerge, where all kinds of distributed IoT devices/things will contribute to and benefit from an intelligent, adaptive, and collaborative network architecture [2].

[†] Yang Yang is the corresponding author. Email: dr.yangyang@terminusgroup.com.

The Sixth Generation (6G) mobile communication systems will be different from the Fifth Generation (5G) systems in three important aspects. First, in terms of goals, 5G aims at radical improvements of network Key Performance Indicators (KPIs), such as peak data rate, spectrum efficiency, energy efficiency, service coverage, device density, and air-interface delay, by at least ten times comparing to the Fourth Generation (4G) systems. 5G continues to provide predefined "standard" services, such as enhanced Mobile BroadBand (eMBB), Ultra Reliable Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC), for different groups of users, just like 4G did for urban, sub-urban, and rural users. This traditional "user-centric" service model could only provide good average performance for a group of typical users in similar locations or application scenarios. However, the goal of 6G is to provide "everyone-centric customized" services according to the integrated, dynamic, and multi-dimensional service requirements of different user tasks [3]. In order to guarantee everyone's Quality of Experience (QoE) in customized services, adaptive End-to-End (E2E) system formulation and service provisioning algorithms are needed for different application scenarios and network conditions [16]. Building upon the digital world, advanced IoT and AI technologies will accelerate the evolution towards this ambitious goal of 6G, thus achieving the finest service granularity at the task level for guaranteeing every user's personalized QoE.

Second, in terms of approaches, 5G has improved a set of network KPIs by committing more resources, such as frequency spectrum, transmission power, antenna arrays, denser cells, cloud computing, and complex algorithms. This "technology-driven" approach cannot suit new and evolving applications, as KPIs are hard to satisfy without understanding dynamic user requirements and traffic flows. As delay-sensitive broadband applications such as autonomous driving and interactive Virtual Reality/Augmented Reality (VR/AR) games grow explosively, 5G is unable to deliver massive data on time over a limited network bandwidth and, consequently, cloud computing cannot guarantee satisfactory QoEs. In contrast, 6G will adopt a sustainable "service-oriented" approach, which integrates and exploits ubiquitous system resources of Sensing, Storage, Communication, Computing, Control, and AI (S^2C^3A) from cloud, to network, and to edge for supporting different types of AI methods and customized services with multi-dimensional personal requirements [4-10]. This capability will continue all the way to user devices/things and can agilely address sudden changes due to the unexpected reasons such as user behaviors, application scenarios, and network conditions. Heterogenous network resources and pervasive AI algorithms will be shared and orchestrated to customize E2E service provisioning, optimize network operation, and achieve customer well-being at different locations and time scales [11] [12].

Third, in terms of impacts, 5G is playing the key role in the digital transformation, while 6G is envisioned to lead the intelligent transformation of services, applications, businesses, and societies for the future. This vision will be realized not only by improving network KPIs in different application scenarios, but more importantly, by utilizing heterogenous network resources and ubiquitous AI algorithms from the cloud to the edge. 6G will create novel cross-domain innovation ecosystems by enabling effective integration, analysis, and collaboration of disparate data from different business domains, industrial sectors, application scenarios, geographic locations, and digital societies. During the process of intelligent transformation, these ecosystems will jointly consider diverse requirements from multiple perspectives, develop holistic solutions with various objectives, and produce huge amounts of benefits for social progress and economic growth. Novel digital infrastructures, application cases, collaboration paradigms, and business models will be invented and deployed as the cornerstones for establishing our intelligent society [13] [14].

This article proposes a network AI architecture to facilitate the developments and applications of pervasive AI methods and intelligent customized services in future 6G mobile networks. Our main contributions are summarized as follows.

(i) To visualize the complex and dynamic requirements from each user task, we coin the concept of Service Requirement Zone (SRZ) that characterizes its multi-dimensional service requirements by using a set of E2E performance bounds, which jointly determine the user's overall QoE.
(ii) To measure a 6G system's service ability of guaranteeing everyone's QoE, we introduce the concept of User Satisfaction Ratio (USR) that calculates the percentage of satisfied tasks among all served tasks over a period of time by comparing their individual SRZs one-by-one against achieved performance results.
(iii) To pursue high QoE and USR in 6G systems, we propose the network AI architecture with multi-tier, multi-function Nodes (mNodes) as its basic elements that integrate local system resources of S^2C^3A to provide a native AI service platform for serving diverse tasks with customized SRZs.
(iv) To evaluate the performance of the proposed network AI architecture, we conduct extensive computer simulations, and the results show that it can achieve the highest USR under dynamic service requirements and network conditions, in comparison with the existing cloud AI and edge AI architectures.

The rest of this article is organized as follows. Section II introduces the concept of SRZ for every task from each user. Next, Section III defines the performance metric of USR for evaluating the overall service ability of a system. The network AI architecture is then proposed and discussed in Section IV. Section V shows and analyzes the extensive simulation results for three AI architectures under dynamic service requirements and network conditions. Several key research challenges are then identified and elaborated as the future work in Section VI. Finally, Section VII concludes this article.

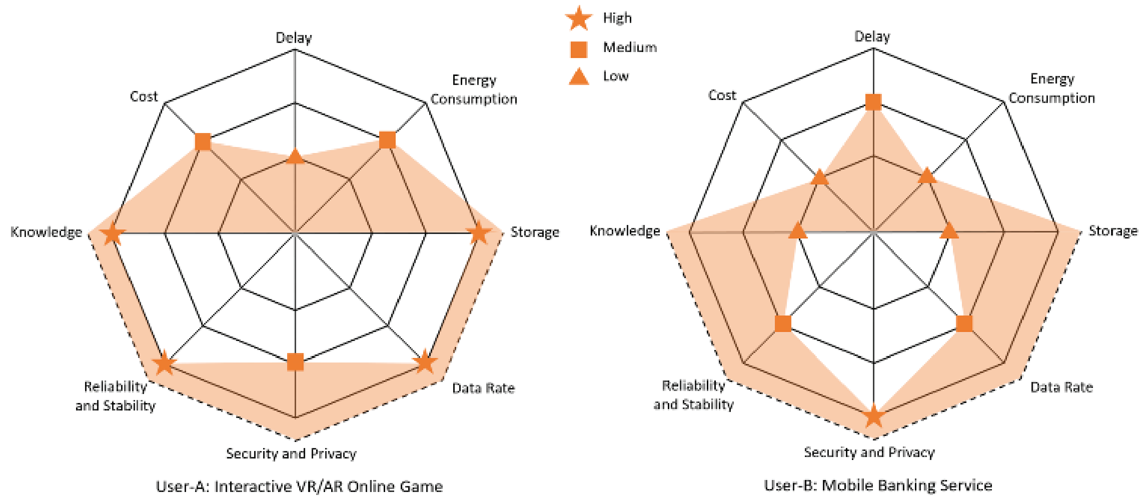


Fig. 1 Service Requirement Zone.

II. Service Requirement Zone

Radar charts with multiple KPIs have been widely used to indicate the technology advancements and capability enhancements from an aggregated system's perspective [4] [14]. Unlike this traditional approach, we apply radar charts to visualize the SRZ of every task for characterizing the user's integrated, multi-dimensional service requirements and preferences. Some network KPIs are not directly relevant to a user's own service experience, e.g., device density, peak data rate, and network capacity. However, many service KPIs are critical for his/her QoE because they jointly determine the personalized SRZ.

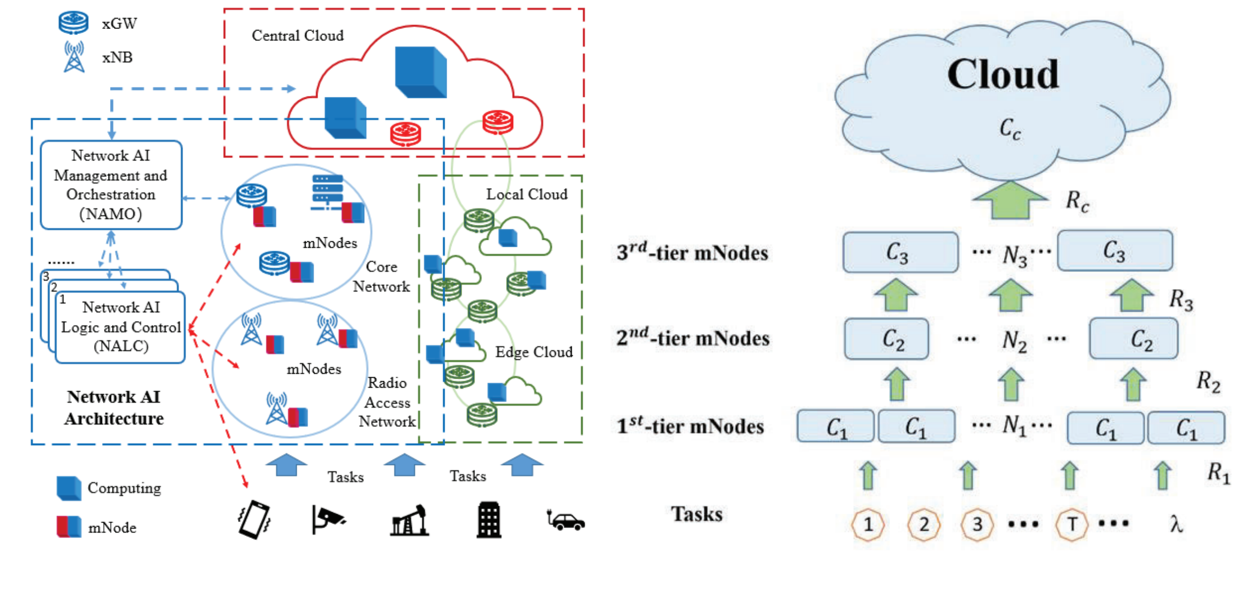
As an example, Fig. 1 shows eight service KPIs that define the eight-dimensional SRZ on an octagonal radar chart, i.e., the brown zone. Note that, for a particular task, the user requirements on storage, data rate, security, reliability, and knowledge are actually the performance lower bounds, while the requirements on cost, delay, and energy consumption are the upper bounds. Since the system can certainly achieve much better performance than these KPI bounds of a single user task, the radar chart is colored in from the origin (i.e., the minimal values for the three upper bounds) to the dashed lines outside the chart, which represent the maximal system performance values for the five lower bounds. The dimension and shape of each SRZ could be determined by different types of users and application requirements, such as by professional users in premium services and by application developers for general users in standard services. In general, a larger SRZ with wider area indicates lower service requirements, and vice versa.

Referring to Fig. 1, User-A on the left-hand side is playing an interactive VR/AR game with a group of virtual friends in the Metaverse. The SRZ of this task requests a low E2E service delay, a standard energy consumption, instant storage and caching of a large amount of user data, a high transmission data rate, normal security and privacy protection, an ultra-reliable and stable experience during the service process, rich domain-specific knowledge and capability for 3D graphic rendering, as well as a reasonable cost. On the right-hand side, User-B is using a mobile banking service for money transfer. The corresponding SRZ consists of a medium service delay, a low energy consumption, small data storage and caching, a normal transmission data rate, strong security and privacy protection, a standard service reliability, no additional domain-specific knowledge, and a low cost. To satisfy diverse SRZs, adaptive E2E service provisioning algorithms are crucial in supporting integrated, multi-dimensional service requirements from different tasks.

In order to guarantee each user's QoE, future 6G systems should integrate and orchestrate heterogenous network resources across multiple domains for providing everyone-centric customized services anywhere and anytime, thus shifting network slicing technology to the finest granularity at the task level. Such task-specific SRZs might look like a huge burden for the corresponding users. However, in practice, each type of tasks has the similar SRZ, i.e., the de facto service model for a group of users. The typical SRZs for interactive VR/AR online games and mobile banking services are given in Fig. 1. Despite dynamic user behaviors and service environments, these SRZs are quite stable because most users usually do not compromise their service requirements and QoEs, unless service continuity and high quality cannot be satisfied at the same time. In this case, some users may accept an expanded SRZ with lower requirements and degraded quality for maintaining service continuity, say in a high-speed train. 6G systems with pervasive intelligence should be able to efficiently identify, allocate, and manage heterogenous network resources for a variety of tasks in different user environments, application scenarios, and network conditions.

III. User Satisfaction Ratio

The dynamic SRZs of various tasks are used as the QoE targets for customized service provisioning and performance optimization in 6G. Referring to the SRZs in Fig. 1, if the achieved system



(a) Deployments of Cloud, Edge, and Network AI Architectures.

(b) System Model.

Fig.2 Three AI Architectures and the System Model

performance results in multiple dimensions are all located within the brown zone, the corresponding user will feel satisfied. Then, the counters for served tasks N_T and satisfied tasks N_S are both increased by one. Otherwise, this service has failed and only N_T is increased by one. For a given period of time, the USR is calculated as the ratio between the number of satisfied tasks N_S and the total number of served tasks N_T, i.e.,

$$USR = \frac{N_S}{N_T}. \quad (1)$$

Individually, every user could have these two counters and calculate the USR to indicate his/her personal QoE with the network operator or service provider. Collectively, the USR can be applied to evaluate a 6G system's overall service ability in satisfying individual SRZs of a variety of tasks, not regarding any specific user locations, application scenarios, network conditions, or operation environments. In the rest of this article, the USR is used as an effective, fair, and general performance metric of the whole system.

Consider different systems with a similar amount of network resources. The higher the USR is, the more intelligent a system is in utilizing limited network resources for efficiently serving diverse tasks with individual SRZs. 5G today is mainly focused on improving separate and objective KPIs at the supply side, such as signal strength, service coverage, device density, and spectrum and energy efficiency. However, 6G seeks to satisfy every user's personal and subjective requirements denoted by SRZs at the demand side. Heterogenous 6G network resources in multiple domains should be effectively integrated and exploited to jointly enhance everyone's QoE and the system's USR.

The calculation of USR is based on the binary, hard decision according to every task's SRZ, i.e., whether or not the system can satisfy the task-specific KPIs simultaneously. Besides this binary classification method, the definitions of SRZ and USR can be extended to multiple scales from the user side and the system side, respectively. First, we can assign different coefficients to prioritize the KPIs that are more important to particular tasks or users. Hence, the weighed SRZ is obtained by considering different degrees of importance for selected KPIs. Second, we can introduce a multi-step, soft-decision method to produce an acceptable performance zone on top of a task's SRZ by loosening its requirements on some KPIs. For example, everyone likes watching high-definition videos at home, but most of us would accept low-quality (low data rate) videos in a high-speed train. Hence, the stepped USR is derived by considering different levels of satisfaction on selected KPIs.

IV. Three AI Architectures and the System Model

1. The Cloud AI and Edge AI Architectures

In the era of 5G, the cloud AI architecture has been widely adopted to provide centralized computing services, such as big data analysis and AI training and inference. The conventional "cloud-pipe-terminal" structure decouples the data sensing functions at user terminals, the communication functions in mobile networks (a.k.a. the pipe), and the computing functions or the AI-enabled analytical services on the cloud [12]. This is simply a combination of the existing infrastructures of Data Technology (DT), Communication Technology (CT), and Information Technology (IT). It is very challenging to coordinate these separate functions in multiple facilities for effectively providing an agile, smooth, and stable service with guaranteed QoE.

In order to solve the problem of low speed, long delay, poor privacy, and high carbon emissions in centralized AI applications on the cloud, the edge AI architecture extends the computing capability from the cloud to the locations physically closer to end users. Although the costs for deploying edge clouds (also called cloudlets) widely in the neighborhood are very high, this "cloud-edge-terminal" structure is getting popular in various application scenarios with high added values. This is because it is much more effective in supporting computing-intensive, delay-constrained, security-assured, and privacy-sensitive applications, such as interactive VR/AR games, autonomous driving, and intelligent manufacturing.

As shown in Fig. 2 (a), central, local, and edge clouds are connected by high-speed, expensive bearer networks, which are just the traffic pipes with huge bandwidth. They are considered as affiliated computing resources for enhancing the AI service capabilities in different application scenarios and network locations. Strictly speaking, local and edge clouds are deployed as affiliated Over-The-Top (OTT) services to support computing-intensive applications. They are usually co-located with the existing network elements, but not embedded in mobile networks. Thus, cross-domain resource coordination and service orchestration between these local/edge clouds and end users require round-trip data transmissions through the mobile network. The actual service procedure is very complicated, time-consuming, and expensive, and may generate a series of management and technical problems such as redundant deployment costs, circuitous data paths, and frequent desynchronized cooperation. It is very difficult for the cloud AI and edge AI architectures to guarantee E2E QoE for sophisticated cross-domain services in dynamic application scenarios and mobile environments.

2. The Network AI Architecture with Multi-tier mNodes

To address those challenging problems, two-level digital twins and edge-cloud cybertwins are proposed in the cyber space [8] and the service network [9], respectively. In this article, we propose the network AI architecture with multi-tier mNodes to integrate and coordinate cross-domain S^2C^3A resources for processing local/regional user data, executing distributed AI algorithms, and providing customized services for everyone as closely as possible. This architecture shifts the classic design paradigm that assumes mobile networks only as the pipe for data transmissions. Based on the hierarchy of mNodes, heterogeneous network resources and separate functions are effectively integrated to support cross-domain, wide-area, and delay-sensitive applications, e.g., autonomous driving. Compared with the edge AI, the proposed network AI architecture can achieve a better balance between E2E service performance, management overhead, and deployment and maintenance costs.

As the key 6G network element, an mNode will not only coordinate local resources as a Service Provider does for E2E service auction [16], but also integrate the basic S^2C^3A resources and multiple functions to support QoE-guaranteed, everyone-centric customized services. Different from traditional rigid hardware deployments with dedicated duties and separate functions in either Radio Access Network (RAN) or Core Network (CN), the mNodes will adopt advanced Network Function Virtualization (NFV) technologies and play different roles as needed inside 6G mobile networks, such as the e/g Node Base-station (xNB), the P/S-Gateway (xGW), the Access and Mobility Management Function (AMF), and edge/fog service nodes. Besides general-purpose computing units, it is envisaged that more and more AI processors will be widely integrated and shared by the mNodes to provide the 6G native AI service platform. Based on this, most tasks with smaller SRZs, i.e., stringent KPIs on data rate, delay, security, privacy, and energy consumption, will be automatically assigned to the nearby mNodes, thus satisfying everyone's QoE with personal service requirements in dynamic user environments, application scenarios, and network conditions.

In Fig. 2 (a), the proposed network AI architecture consists of three key units and constructs a comprehensive, distributed, and scalable AI as a Service (AIaaS) platform in 6G. First, the network infrastructure is composed of dispersive mNodes in multi-tier mobile networks. Second, each Network AI Logic and Control (NALC) unit is task-oriented and manages the multi-tier mNodes in a specific local/regional area through effective signaling schemes. In 6G mobile networks, a NALC coordinates the integrated S^2C^3A resources and functions for serving every task in realtime and near-realtime applications, i.e., E2E delay ranges from milliseconds to tens of milliseconds. The customized service procedure and personal QoE of every task are constantly monitored and optimized by a corresponding NALC. Third, a Network AI Management and Orchestration (NAMO) unit manages the AIaaS platform with multiple NALCs to support wide-area applications by cross-domain resource coordination, service orchestration, and E2E QoE guaranteeing protocols. In 6G systems, NALC and NAMO should work close together to effectively balance the service requirements on short E2E delay and wide service coverage in different application scenarios. For the cases that other IT vendors are willing to contribute additional cloud and edge computing resources, NAMO will coordinate multi-vendor resources to support complex applications across different AI architectures. Therefore, the proposed network AI architecture can either serve various tasks independently, or complement with the cloud AI and edge AI architectures to satisfy sophisticated user requirements with challenging SRZ targets.

3. System Model

To study a typical 6G system with dispersive computing resources and pervasive intelligence, Fig. 2 (b) shows a general system model for different AI architectures. Let us consider a series of tasks, each having a customized SRZ, arriving at the system with rate λ tasks per second. These tasks are generated randomly either by end users enjoying mobile internet services or by various devices and things embedded in industrial IoT applications. As discussed, simply deploying more computing resources as the affiliated AI capabilities in access networks and bearer networks, while keeping different service functions separated (as in previous generations of mobile networks), would generate

Table 1. Simulation Parameters.

	Parameter		**Value**		
User Task: Demand Side	Task Density/Arrival Rate λ		[1000, 3000] (tasks per second)		
	Delay Bound D_0		E[D_0]=1600 (seconds), Var(D_0)=50		
	Energy Bound E_0		E[E_0]=1.85(kW·h), Var(E_0)=0.05		
	Task Size Z		E[Z]$\in[4.8\ \times 10^8, 7.2 \times 10^8]$(bytes) Var($Z$)=$1 \times 10^6$		
	Computing Requirement U		E[U]$\in[0.4 \times 10^2, 1.0 \times 10^2]$ (teraFLOPS) Var(U)=1.0		
6G System: Supply Side			**Cloud AI**	**Edge AI**	**Network AI**
	Computing Overhead		0	28800 (teraFLOPS)	36400 (teraFLOPS)
	Effective Computing Power		140000 (teraFLOPS)	111200 (teraFLOPS)	103600 (teraFLOPS)
	Cloud	Computing Power C_c	140000 (teraFLOPS)	100000 (teraFLOPS)	70000 (teraFLOPS)
		Data Rate R_c	2500 (Mbps)		
	3^{rd}-tier mNode	Number N_3	0	0	10
		Computing Power C_3	-	-	1120 (teraFLOPS)
		Data rate R_3	E[R_3] $\in[1600, 2500]$ (Mbps), Var(R_3)=100		
	2 nd-tier mNode	Number N_2	0	0	100
		Computing Power C_2	-	-	112 (teraFLOPS)
		Data Rate R_2	E[R_2] $\in[400, 625]$ (Mbps), Var(R_2)=25		
	1 st-tier mNode	Number N_1	0	1000	1000
		Computing Power C_1	-	11.2(teraFLOPS)	11.2 (teraFLOPS)
		Data Rate R_1	E[R_1] $\in[56, 87.5]$ (Mbps), Var(R_1)=7		
Algorithms: Supply Side	Fair Equal Scheduling (FES)		100%	50% : 50%	25:25:25:25 %
	The Closer The Better (TCTB)		100%	80% : 20%	80: 10: 5: 5 %

significant management and technical problems. Therefore, without loss of generality, we consider a three-tier network AI architecture with three types of mNodes, which are represented by blue rectangular boxes. The number of mNodes, the computing power (FLOPS: floating-point operations per second), and the network data rate (bytes per second) in the i^{th}-tier are denoted by N_i, C_i, and R_i, respectively. Above them sits a cloud, which has the highest data rate R_c and the strongest computing power C_c . This system model can be easily simplified to represent the cloud AI and edge AI architectures by setting $N_i = 0$ for $i \geq 1$ and $i \geq 2$, respectively.

For an arbitrary task T, the corresponding service provisioning procedure is determined by the specific task scheduling algorithm. Upon the arrival of task T, its SRZ is first checked by a nearby *1st*-tier mNode at the edge, which analyzes the possibility of satisfying that SRZ with the network resources available in the vicinity. If local resources are sufficient, task T will be immediately served by this mNode. If not, a more powerful *2nd*-tier mNode will be initiated to lead the effort of identifying feasible network resources in a bigger neighborhood. If regional resources are still not sufficient, an even stronger *3rd*-tier mNode will be called upon to perform multi-domain resource coordination over a much wider area. In some cases, task T is so complex that a large amount of network resources will be used to collect and process not only local and regional data, but also global data. If task T can be split into multiple subtasks [15], the same number of mNodes in the horizontal or vertical directions can share their resources and capabilities to collectively serve task T. Otherwise, task T cannot be split and has to be uploaded to the cloud through the multi-tier network, thus increasing the end-to-end transmission delay, energy consumption, and total cost. Traditional cloud AI architecture relies on remote super-powerful computing resources, while recent edge AI architecture takes advantage of local light-weight computing resources. As the next stage, the network AI architecture incorporates both cloud and edge AI resources to allocate multi-tier, pervasive intelligence in 6G systems.

This article has been accepted for inclusion in a future issue of this magazine.

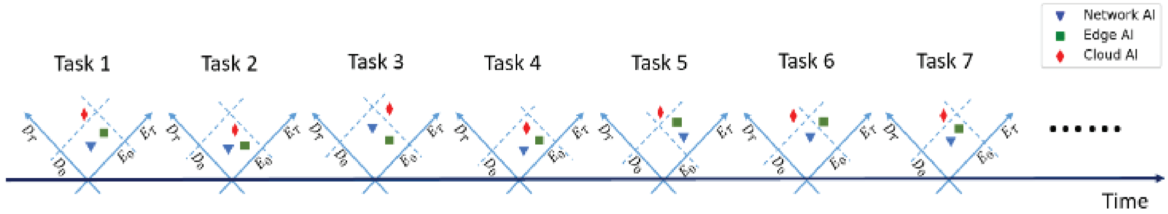


Fig. 3 Service Results of Representative Tasks with Different SRZs.

V. System Parameters and Simulation Results

Different from the DeFog benchmarks built on representative applications (https://github.com/qub-blesson/DeFog), the simulation study of different AI architectures in this article is based on real world experiences and best practices in typical CT and IT networks. Table 1 lists all the parameters and their assumed values about tasks, three AI architectures, and two task scheduling algorithms for extensive computer simulations. On the demand side, different users continuously generate λ tasks per second. Assume a non-splittable task T have a size of Z bytes and a computing requirement of U teraFLOPS. To demonstrate the key results within limited space, only delay and energy consumption are chosen as the illustrative service KPIs for constructing a two-dimensional SRZ for every task. If task T is served by an mNode in the h^{th} tier, the E2E service delay D_T consists of (i) the h-hop transmission delay which is determined by task size and random data rate at each hop, and (ii) the computation delay at the serving mNode, which is affected by task computing requirement, shared computing power at the mNode, dynamic queueing delay due to multiple competing tasks, and limited I/O speed for data storage. These negative effects at the mNode prolong the computation delay of every task. After considering their combined impact, the effective computing power C_h seen by the tasks is proportionally reduced. Therefore, the overall service delay D_T can be expressed as

$$D_T = \sum_{i=1}^{h} \frac{Z}{R_i} + \frac{U}{C_h}, \tag{2}$$

Similarly, the total energy consumption E_T consists of the h-hop transmission energy consumption and computation energy consumptions of the task, i.e.,

$$E_T = \sum_{i=1}^{h} \alpha_i \frac{Z}{R_i} + \gamma {C_h}^2 U, \tag{3}$$

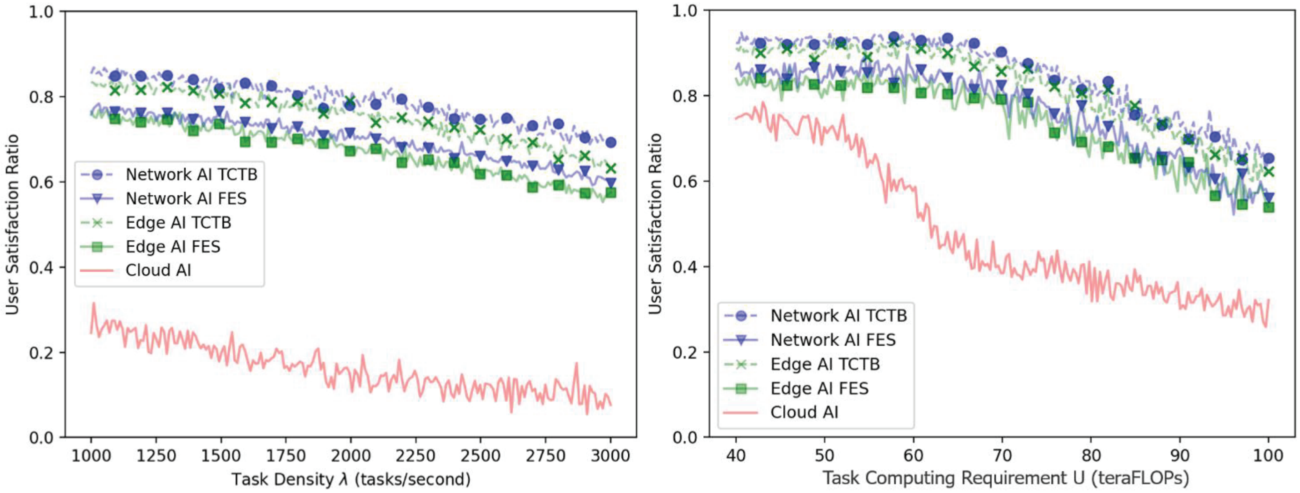
where α_i denotes the average transmission power over the i^{th} hop, which is set to be 0.1 Watts for typical network elements. The coefficient γ represents the effective switched capacitance, which is related to the chip architecture at the serving mNode. According to the previous study [17], it is an extremely small constant and can be set as $\gamma = 1 \times 10^{-27}$. The condition for user satisfaction is therefore $D_T \leq D_0$ and $E_T \leq E_0$, where D_0 and E_0 are the upper bounds of service delay and energy consumption, as specified by the SRZ of task T. Without loss of generality, the values of Z, U, D_0, and E_0 are randomly generated according to different Gaussian distributions.

For a sequence of tasks, Fig. 3 shows their customized SRZs as rectangular zones bounded by the actual values of D_0 and E_0, represented by two dashed lines. The service results of the delay and energy consumption performance are denoted by three markers for different AI architectures. Taking Task 1 as an example, both the network AI and edge AI architectures can achieve satisfied QoEs since their markers are located inside the SRZ. On the contrary, the cloud AI architecture fails to provide acceptable delay performance.

On the supply side, the cloud AI, edge AI, and network AI architectures are evaluated with the same total computing power of 140K teraFLOPS. For a fair comparison, they are composed of a cloud and a three-tier network for serving tasks with different SRZs. For the cloud AI architecture, all tasks are transmitted over the network and served in the cloud. There is no additional computing overhead for task scheduling and resource management outside the cloud, so the effective computing power is $C = C_c = 140\text{K}$ teraFLOPS.

The edge AI architecture allocates a small amount of computing power among 1000 1st-tier mNodes at the edge and the rest of computing power in the cloud. Assuming a 20% computing overhead for task scheduling and resource management at the edge, the resulting effective computing power is equal to $C = N_1 \times C_1 + C_c = 111.2\text{K}$ teraFLOPS. In Table 1, two task scheduling algorithms are considered in performance evaluation. The Fair Equal Scheduling (FES) algorithm assigns all the tasks in a random manner, with half going to the edge and half to the cloud for services. The-Closer-The-Better (TCTB) algorithm follows the Pareto principle, or the 80/20 rule, so that 80% and 20% of all the tasks go to the edge and the cloud, respectively. The use of FES and TCTB algorithms will demonstrate the fundamental differences among the three AI architectures and provide standard benchmarks for developing more sophisticated algorithms for complex application scenarios and dynamic network conditions.

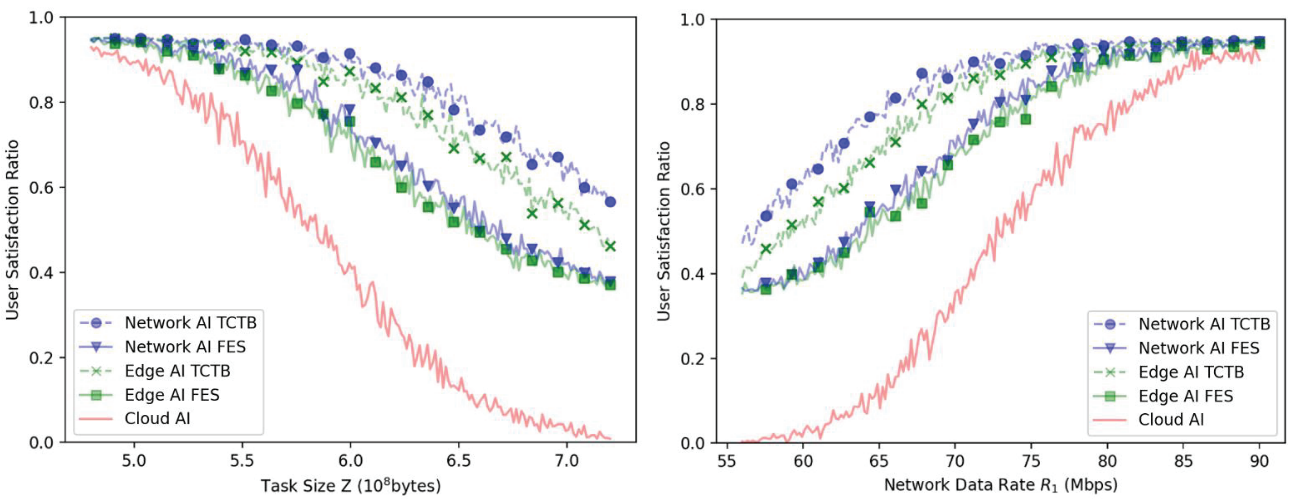
The network AI architecture is comprised of more mNodes with different capabilities in three network tiers, thus the additional computing overhead due to system and algorithm complexities is higher and assumed to be 36.4K teraFLOPS. The total effective computing power is then derived as $C = N_1 \times C_1 + N_2 \times C_2 + N_3 \times C_3 + C_c = 103.6\text{K}$ teraFLOPS. Usually, an upper-tier mNode covers a larger geographical or logical area in the network and therefore is more capable of serving more tasks. Specifically, as network tier increases, we assume that the



(a) Impact of Task Density when Task Computing Requirement U~N (70,1).

(b) Impact of Task Computing Requirement when Task Density λ=1000.

Fig. 4 USR versus Task Density and Computing Requirement.



(a) Impact of Task Size when Network Data Rate R_1~N (70, 7).

(b) Impact of Network Data Rate when Task Size Z~N $(6 \times 10^8, 10^6)$.

Fig. 5 USR versus Task Size and Network Data Rate.

number of mNodes decreases exponentially while the computing power of each mNode increases exponentially. The FES algorithm randomly assigns each task to a network tier or the cloud, thus a portion of 25% tasks is served in each network tier and the cloud. The TCTB algorithm gives much higher priorities to lower network tiers, so the proportions of task assignments to the *1st*-tier, *2nd*-tier, *3rd*-tier, and cloud are reasonably set as 80%, 10%, 5%, and 5%, respectively.

As defined, the overall USR can be calculated by comparing the number of satisfied tasks against the total number of served tasks. When the Gaussian distributions of task size and network data rates are fixed, i.e., Z~N $(6 \times 10^8, 10^6)$, R_1~N (70, 7), R_2~N (500, 25), and R_3~N (2000, 100), Fig. 4 illustrates the USR performance of the three AI architectures under dynamic task densities and computing requirements. In Fig. 4 (a), the task density has a linear impact on the decline of the USR curves under different AI architectures. For TCTB, when λ is equal to 1500, 2000, and 2500 tasks per second, respectively, the network AI architecture can achieve 3.8%, 5.3%, and 7.4% higher USR than the edge AI architecture, while 315.0%, 393.8%, and 461.5% higher USR than the cloud AI architecture, respectively.

In Fig. 4 (b), the USR curve of the cloud AI architecture has two knee points at about U=48 teraFLOPS and U=66 teraFLOPS. The transition region between them has a steep slope, which implies that the energy consumptions for executing all the tasks in the cloud increase very rapidly when the average computing requirement increases. Under both TCTB and FES algorithms, the green and blue curves of the edge AI and network AI architectures are much less sensitive to this change, which is due to the efficient services by mNodes in the neighborhood. The turning points for TCTB and FES curves are around U=68 teraFLOPS and U=71 teraFLOPS respectively, where the gradients climb roughly from 0 to 0.36.

In Fig. 5 (a), for fixed task density λ=1000 and task computing requirement U~N(70,1), when task size increases, the USR curve of the cloud AI architecture degrades dramatically because long-distance transmissions of bigger tasks become more time-consuming and energy-intensive, thus adversely impacting the USR. On the contrary, the USR curves of the edge AI and network AI architectures are much less sensitive to task size changes, thanks to the computing resources deployed at the edge and in the network. Compared with FES, TCTB is more effective in satisfying different SRZs simultaneously by transmitting most tasks to local and regional mNodes. The turning points of TCTB curves are around Z=6 $\times 10^8$ bytes where the gradients are doubled from 0.17 to 0.38.

Fig. 5 (b) demonstrates the influence of network data rates on the USR performance. Specifically, we assume that R_1, R_2 and R_3 are Gaussian random variables with different mean values, but at a fixed ratio of E[R_1]:E[R_2]:E[R_3]=7:50:200. So, only E[R_1] is shown as the X-axis in the figure. Very interestingly, these curves are like the mirror flips of those in Fig. 5 (a), because higher network data rates and smaller task sizes both imply lower transmission delays. Therefore, increasing network data rates and reducing task size have almost equivalent impact on the USR performance. When network data rate is high, e.g., E[R_1] > 85 Mbps, the USR curve of the cloud AI architecture gets very close to the curves of the edge AI and network AI architectures, just like the case when the average task size E[Z] < 4.95×10^8 bytes in Fig. 5 (a).

VI. Research Challenges

We believe the following research challenges and technical problems require further discussions and investigations.

(1) **Statistical Models of Diverse SRZs:** integrated service requirements of different types of realistic tasks should be studied in complex application scenarios and dynamic network conditions. New KPIs on pervasive intelligence, QoE, and social benefits will be investigated. Priorities should be given to mission-critical tasks and elderly users.

(2) **Service Capacity of 6G Systems:** practical mechanisms should be developed to map customized SRZs onto heterogenous system resources and AI capabilities across multiple tiers and domains. Theoretical analysis of system service capacity is crucial for improving service efficiency, resource utilization, and everyone-centric QoE.

(3) **Cross-domain Service Provisioning:** the design of mNodes, NALC, and NAMO should be promoted to support a series of effective interfaces, protocols, and algorithms for cross-domain resource allocation, E2E service provisioning, customized task scheduling, multi-node collaborations, mobility management, user behavior monitoring, and QoE performance optimization.

(4) **E2E Security and Privacy Protection:** considering randomly distributed users with a variety of access devices, a zero-trust architecture should be developed together with the network AI architecture. Context-aware security and privacy protection methods should support everyone-centric customized services under different user locations, mobile terminals, wireless environments, application scenarios, and network conditions.

(5) **Implementation of Native AI Capability:** to enable the native AI capability in the network AI architecture, a joint design methodology should be studied to support effective development and evaluation of collective AI methods using distributed, heterogenous network resources. Such localized but federated AI algorithms could greatly reduce the training time and the size of action space. Some implementation issues from physical layer to application layers should be studied for real-world applications, such as user requirement and mobility models, wireless channel characteristics, task arrival statistics, network traffic dynamics, system and algorithm complexities, training data splitting, distributed AI collaborations, AI service coverage and handoff, and stable QoE performance.

VII. Conclusions

Unlike existing 4G/5G systems that offer standard mobile services for different application scenarios, 6G systems should be able to tailor customized services to meet everyone's personal requirements. From a user's perspective, we first coined the concept of SRZ to characterize each task's integrated performance requirements. Next, from a system's perspective, we introduced the concept of USR to evaluate the system's overall service ability of satisfying individual SRZs of different tasks. Then, the cloud, edge, and network AI architectures were studied and compared under dynamic task densities, task sizes, computing requirements, network data rates, and two task scheduling algorithms. By deploying multi-tier mNodes, the proposed network AI architecture with integrated S^2C^3A resources can effectively support customized services for a variety of user tasks with different SRZs, thus achieving the highest USR under random service requirements and dynamic network conditions. In contrast, the centralized cloud AI architecture has difficulties in meeting stringent delay and energy consumption bounds, thus not suitable for delay-sensitive broadband applications such as interactive VR/AR games, autonomous driving, and intelligent manufacturing.

ACKNOWLEDGMENT

Yang Yang would like to thank the Associate Editor Professor Dusit Niyato and four anonymous reviewers for their constructive comments. He is also very grateful to Prof. Lajos Hanzo from University of Southampton, UK, Dr. Qi Bi from China Telecom, China, Prof. Zhisheng Niu from Tsinghua University, China, Dr. Tao Zhang from the National Institute of Standards and Technology, USA, Prof. Raymond Wei-Ho Yeung from the Chinese University of Hong Kong, China, and Prof. Rui Tan from Nanyang Technological University, Singapore, for their valuable comments on a draft version of this article. This work was supported in part by the National Key Research and Development Program of China (2020YFB2104300), the Major Key Project of the Peng Cheng Laboratory (PCL2021A15), and the National Natural Science Foundation of China (U21B2002 and 61932014).

REFERENCES

[1] Y. Xiao, G. Shi, Y. Li, W. Saad, and H. V. Poor, “Toward Self-learning Edge Intelligence in 6G,” IEEE Communications Magazine, vol. 58, no. 12, pp. 34-40, Dec. 2020.

[2] Y. Yang, “Multi-tier Computing Networks for Intelligent IoT,” Nature Electronics, vol. 2, pp. 4-5, Jan. 2019.

[3] Y. Yang, X. Chen, R. Tan, and Y. Xiao, “Intelligent IoT for the Digital World,” ISBN: 9781119593546, Wiley, 2021.

[4] X. H. You, C. X. Wang, et al., “Towards 6G Wireless Communication Networks: Vision, Enabling Technologies, and New Paradigm Shifts,” Science China Information Sciences, Vol. 64, No. 1, Jan. 2021.

[5] Z. Feng, Z. Wei, X. Chen, H. Yang, Q. Zhang, and P. Zhang, “Joint Communication, Sensing, and Computation Enabled 6G Intelligent Machine System,” IEEE Network, pp. 34-42, Nov./Dec. 2021.

[6] W. Saad, M. Bennis, and M. Chen, “A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems,” IEEE Network, pp.134-142, May/Jun. 2020.

[7] S. Chen, Y. C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, “Vision, Requirements, and Technology Trend of 6G: How to Tackle the Challenges of System Coverage, Capacity, User Data-Rate and Movement Speed,” IEEE Wireless Communications, vol. 27, no. 2, pp. 218-228, Apr. 2020.

[8] X. Shen, J. Gao, W. Wu, M. Li, C. Zhou, and W. Zhuang, “Holistic Network Virtualization and Pervasive Network Intelligence for 6G,” IEEE Communications Surveys & Tutorials, Vol. 24, No. 1, pp. 1-30, First Quarter, 2022.

[9] Q. Yu, J. Ren, Y. Fu, Y. Li, and W. Zhang, “Cybertwin: An Origin of Next Generation Network Architecture,” IEEE Wireless Communications, pp. 111-117, Dec. 2019.

[10] 6G Alliance of Network AI (6GANA), “From Cloud AI to Network AI, a View from 6GANA,” May 2021. Available at http://www.6g-ana.com/upload/file/20210619/6375969458505193666851527.pdf

[11] NTT DOCOMO, “White Paper: 5G Evolution and 6G (Version 4.0),” Jan. 2022. Available at: https://www.docomo.ne.jp/english/corporate/technology/whitepaper_6g/

[12] N. Chen, Y. Yang, T. Zhang, M. T. Zhou, X. L. Luo, and J. Zao, “Fog as a Service Technology,” IEEE Communications Magazine, Vol. 56, No. 11, pp. 95-101, Nov. 2018.

[13] Next Generation Mobile Networks (NGMN) Alliance, “6G Drivers and Vision,” Apr. 2021. Available at https://www.ngmn.org/wp-content/uploads/NGMN-6GDrivers-and-Vision-V1.0_final.pdf

[14] The 5G Infrastructure Association (5GIA), “European Vision for the 6G Network Ecosystem,” Jun. 2021. Available at https://5g-ppp.eu/wp-content/uploads/2021/06/WhitePaper-6G-Europe.pdf

[15] Z. Liu, Y. Yang, K. Wang, Z. Shao, and J. Zhang, “POST: Parallel Offloading of Splittable Tasks in Heterogeneous Fog Networks,” IEEE Internet of Things Journal, Vol. 7, No. 4, pp. 3170-3183, Apr. 2020.

[16] X. Chen, Y. Deng, G. Zhu, D. Wang, and Y. Fang, “From Resource Auction to Service Auction: An Auction Paradigm Shift in Wireless Networks,” IEEE Wireless Communications, early access, May 2022.

[17] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, “Stochastic Joint Radio and Computational Resource Management for Multi-User Mobile-Edge Computing Systems,” IEEE Transactions on Wireless Communications, Vol. 16, No. 9, pp. 5994-6009, Sept. 2017.