6G Network AI Architecture for Everyone-Centric Customized Services

Yang Yang, Mulei Ma, Hequan Wu, Quan Yu, Ping Zhang, Xiaohu You, Jianjun Wu, Chenghui Peng, Tak-Shing Peter Yum, Sherman Shen, Hamid Aghvami, Geoffrey Y Li, Jiangzhou Wang, Tao Zhang, Guangyi Liu, Qi Bi, Xiongyan Tang, Chang Cao, John Thompson, Kit Wong, Shanzhi Chen, Zhiqin Wang, Merouane Debbah, Schahram Dustdar, Frank Eliassen, Yonggang Wen, Rui Tan, Tao Chen, Xiangyang Duan, Shaohui Sun, Xiaofeng Tao, Qinyu Zhang, Jianwei Huang, Shuguang Cui, Wenjun Zhang, Jie Li, Yue Gao, Honggang Zhang, Xu Chen, Xiaohu Ge, Yong Xiao, Chengxiang Wang, Zaichen Zhang, Zhisheng Niu, Song Ci, Guoqiang Mao, Changle Li, Ziyu Shao, Yong Zhou, Junrui Liang, Kai Li, Liantao Wu, Kunlun Wang, Zening Liu, Kun Yang, Jun Wang, Hongfeng Shu

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

Abstract
As a tradition of the telecom industry, researchers have developed previous mobile communication standards mainly for enhancing transmission and network Key Performance Indicators (KPI) by improving spectrum efficiency and utilizing more radio resources. However, how to use a wide range of technological solutions to address a diverse set of user requirements effectively remains an open problem. The future Sixth-Generation (6G) systems have the potential to solve this problem by taking advantage of pervasive intelligence and computing capabilities to support everyone-

centric customized services anywhere and anytime. In this paper, from a user's perspective, we define a novel concept of Service Requirement Zone (SRZ) to characterize the comprehensive service requirements of each task for any user. Then, from a system's perspective, we further propose the User Satisfaction Ratio (USR) as a fair quantifiable measure to evaluate the overall service capability of Artificial Intelligence (AI) architectures in supporting a variety of tasks with different SRZs. Through extensive simulations, we investigate and compare three AI service architectures with centralized or distributed computing resources. Our results show that the proposed network AI architecture can consistently offer higher USR performance than the cloud AI and the edge AI architectures, considering different task scheduling algorithms, network conditions, and operation scenarios.

I. Introduction

Over the last decade, the global development and application of Internet of Things (IoT) have accelerated the digitalization of the physical world and human society. To fully exploit the commercial values of massive data from IoT devices, we can use Artificial Intelligence (AI) algorithms to integrate user requirements, domain knowledge, operation procedures, and business models for 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 personalized preferences and requirements [1]. For manufacturing applications, data from industrial automated control devices in assembly lines can be analyzed by AI algorithms to improve efficiency, productivity, 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 a collaborative, adaptive, and intelligent network architecture [2].

The Sixth Generation (6G) systems will be different from the Fifth Generation (5G) systems in three important aspects. First, in terms of goals, 5G targets at radical improvements of several Key Performance Indicators (KPIs), such as 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 different "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 the same for urban, sub-urban, and rural users. This is the traditional "user-centric" service model which could only provide a statistically satisfactory performance for typical users. However, the goal of 6G is to guarantee Quality of Experience (QoE) in all application scenarios and network conditions to meet specific requirements of various tasks from many users. Built upon the digital world, 6G will seek to provide "everyone-centric customized" services according to each task's individual, integrated, and dynamic requirements [3]. Advanced IoT and AI technologies will further accelerate the evolution towards this ambitious goal of 6G, thus achieving the finest service granularity for satisfying every user with a personalized QoE.

Second, in terms of approaches, 5G improves a set of 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 the new and evolving applications, as KPIs are hard to satisfy without understanding the dynamic user requirements and traffic flows. As delay-sensitive broadband applications such as Virtual Reality/Augmented Reality (VR/AR) interactive games and autonomous driving grow explosively, 5G is unable to deliver massive data on time over a limited network bandwidth and therefore cloud computing cannot guarantee satisfactory QoE. In contrast, 6G will adopt an intelligent and sustainable "service-oriented" approach, which exploits ubiquitous sensing, communication, computing, storage, and control resources, as well as pervasive AI algorithms from cloud, to network, and to edge [4-10]. The service capabilities will continue all the way to user devices and things, and agilely address sudden changes due to reasons such as user behaviors, application scenarios and operation conditions. Heterogenous resources and AI algorithms will be fully shared and orchestrated to customize 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 direction of the intelligent transformation of future services, applications, and businesses across multiple domains and sectors. This vision will be realized not only by improving transmission KPIs for different application scenarios, but more importantly, by ubiquitous sensing, communication, computing, storage, control, and AI resources from the cloud to the edge. In addition, 6G will create novel cross-domain innovation ecosystems by enabling effective integration and collaboration of different types of data from different business domains, industrial sectors, application scenarios, and geographic locations. As the main battlefields of intelligent transformation, these ecosystems will simultaneously consider different requirements from multiple perspectives, develop feasible solutions with various objectives, and produce huge amounts of social and economic benefits. Novel digital infrastructures, application cases, collaboration paradigms, and business models will be deployed as the cornerstone for establishing our intelligent society [13, 14].

This paper proposes a new network AI architecture to fulfill the vision of 6G. Our key contributions are:

(i) To better specify the integrated service requirements of an arbitrary task from any user, we develop the concept of Service Requirement Zone (SRZ) in the multi-dimensional service space with multiple KPIs;

(ii) To better characterize the QoE of all served users in a 6G system according to their SRZs, we define User Satisfaction Ratio (USR) as a measure to evaluate the overall service capability of the system;

(iii) To better provide pervasive intelligence in 6G, we propose the network AI architecture, which integrates basic service functions, such as sensing, communication, computing, storage, control, and intelligent algorithms, to provide the native AI capability for serving the neighborhood;

(iv) We verify the performance of the proposed network AI architecture through extensive simulation studies, together with the cloud AI and edge AI architectures. Results show that the network AI architecture consistently achieve the highest USR under different network conditions.

The rest of this paper is organized as follows. Section II introduces the concept of SRZ for every task from any user. Then, Section III defines the performance metric of USR for 6G systems with pervasive intelligence. The network AI architecture is proposed in Section IV. Section V shows extensive simulation results under dynamic task and system parameters, together with our detailed analysis. Finally, Section VI concludes this paper and identifies future research directions.

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 represent the multi-dimensional Service Requirement Zone (SRZ) of every task to capture the user's service preferences and requirements. From a typical user's perspective, some system KPIs are irrelevant to his/her individual service experience, e.g., device density, peak data rate, and network capacity. However, many service KPIs are critical for his/her QoE since they jointly determine the personalized SRZ.

As an illustrative example, Fig. 1 shows eight service KPIs that define an eight-dimensional SRZ on an octagonal radar chart, i.e., the brown zone. Assume a user is playing a highly interactive VR/AR online game with a group of virtual friends. The SRZ of this task requests a low end-to-end 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. In order to guarantee QoE, 6G systems should satisfy this personalized SRZ under complex application scenarios and dynamic operation conditions. In other words, 6G will extend network slicing technology to the finest granularity, i.e., from a specific application to every task of any user, thus requiring intelligent algorithms to orchestrate necessary network resources for providing everyone-centric customized services anywhere and anytime.

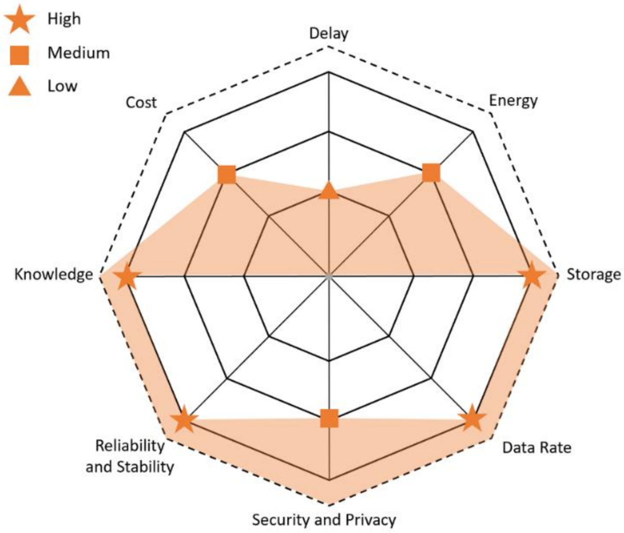


Fig. 1 Service Requirement Zone.

III. User Satisfaction Ratio

The SRZs of various tasks can be used as the QoE targets for service provisioning and performance optimization in 6G. From the perspective of network operator and service provider, we propose the User Satisfaction Ratio (USR) as an effective measure to evaluate the overall capability of a 6G system in serving a large number of tasks with different SRZs simultaneously.

Referring to the SRZ in Fig. 1, if the achieved system performance results in multiple dimensions are all located within the brown zone, the corresponding user will feel very satisfied. Otherwise, this service has failed. As its name implies, the USR is calculated as the ratio between the number of satisfied tasks and the total number of served tasks. It is an effective, fair, and general performance metric for evaluating a system's overall service capability in guaranteeing QoE for a variety of tasks at the same time, not regarding any specific user locations, application scenarios, or network operation conditions.

Consider different systems with a similar amount of network resources. The higher the USR is, the more intelligent a system is in utilizing limited resources for efficiently serving many tasks with various 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 efficiencies. However, 6G seeks to satisfy every user's personal and subjective requirements denoted by SRZs at the demand side. In 6G, network resources in multiple domains are effectively integrated to jointly enhance everyone's QoE and the system's USR.

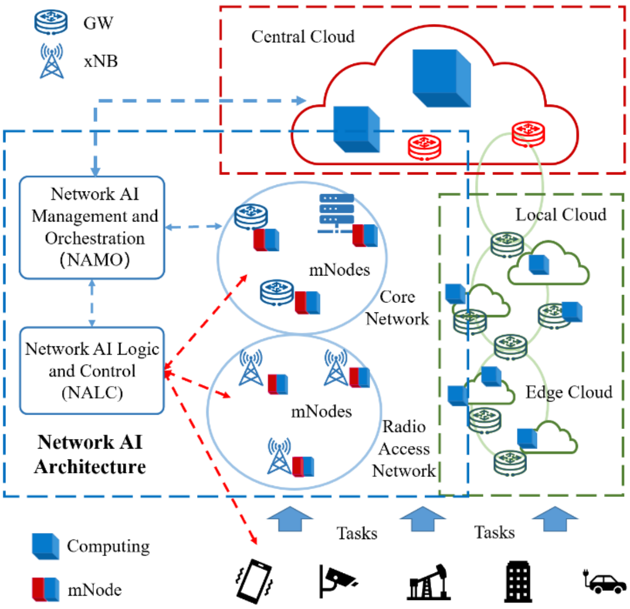
The calculation of USR is based on the binary, hard decision according to every task's SRZ, i.e., whether or not the system has satisfied the specified KPIs simultaneously. To loosen the restriction, two approaches could be applied to extend the definitions of SRZ and USR from the user side and the system 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 the varying degrees of importance of different KPIs. Second, we can introduce the soft-decision method to keep the decimals when the achieved system performance results are compared with a prespecified SRZ. Hence, the stepped USR is derived by taking into account the actual levels of satisfaction on different KPIs.

IV. Three AI Architectures and the System Model

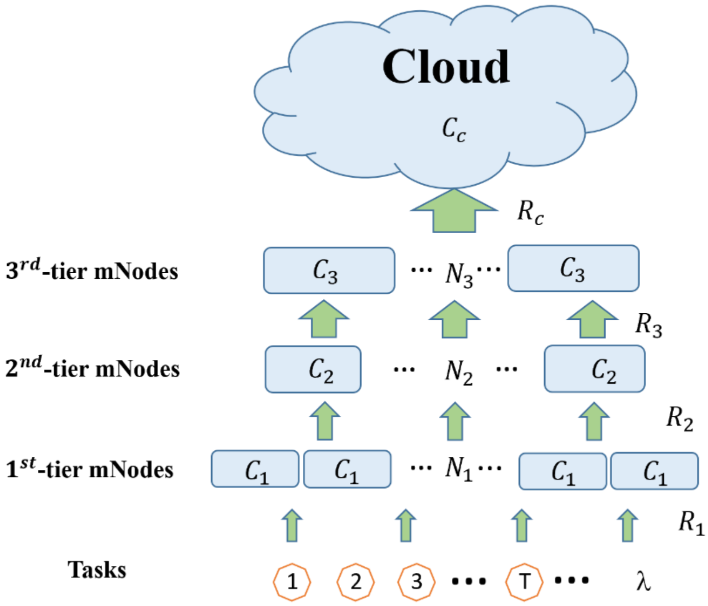
1. The Cloud AI and Edge AI Architectures

In the era of 5G, the cloud AI architecture has been widely adopted to provide centralized computing services, such as data analysis, 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).

In order to solve the problem of low speed, large 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.



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



(b) System Model.

Fig.2 Three AI Architectures and the System Model.

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. Strictly speaking, these computing resources are deployed as Over-The-Top (OTT) services and not part of the mobile networks. They are considered as affiliated AI resources for enhancing the AI capabilities at different network locations. Cross-domain resource coordination and service orchestration require in-depth domain knowledge and rich experiences, and hence are very complicated and time-consuming. This would 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 end-to-end QoE for sophisticated AI services in dynamic application environments and mobile network conditions.

2. The Network AI Architecture with Multi-tier mNodes

To address those challenging issues, two-level digital twins and edge-cloud cybertwins are proposed in the cyber space [8] and the service network [9], respectively, to fulfill the vision of 6G. In this paper, we propose the network AI architecture with multi-tier, multi-function Nodes (mNodes) to shift the classic design paradigm that assumes communication networks as the pipe only for data transmission.

As the key 6G network element, the mNode will integrate basic service functions, such as sensing, communication, computing, storage, control, and AI algorithms, to provide the native AI capability inside the mobile network for QoE-guaranteed, everyone-centric customized services. They will be deployed in multiple network tiers and locations for enhancing, and gradually replacing, the current 4G/5G Node Base-stations (xNBs) and Gateways (GWs) in the Radio Access Network (RAN) and the Core Network (CN), respectively. Depending on specific application scenarios, some tasks may have stringent SRZs due to bandwidth, delay, security, privacy, or energy constraints and should be served as locally as possible by nearby mNodes.

In Fig. 2 (a), the proposed network AI architecture consists of three key units and constructs a comprehensive, distributed, and pervasive AI environment for 6G. First, the AI infrastructure is composed of mobile networks and multi-tier mNodes with heterogeneous resources. Second, the Network AI Logic and Control (NALC) unit controls all the integrated resources and functions inside the AI infrastructure, thus providing realtime or semi-realtime (from milliseconds to tens of milliseconds) task scheduling and resource allocation for QoE guarantee. It also conducts the monitoring and management of customized service procedure and lifetime satisfaction for every task of any user in dynamic mobile environments and wireless channel conditions. Third, the Network AI Management and Orchestration (NAMO) unit contains an elastic and extensible AI as a Service (AIaaS) platform, which supports AI service orchestration and automatic management for diverse 6G applications. For the cases that other IT vendors are willing to contribute additional cloud and edge computing resources, the NAMO can help to coordinate and utilize all the resources to provide complex services across different AI architectures. In summary, the 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 the 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 by either end users enjoying mobile internet services or various devices/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 sensing, communication, computing, storage, and control functions separated (as in previous generations of mobile networks), would generate 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 1^{st}-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 2^{nd}-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 3^{rd}-tier mNode will be called upon to perform multi-domain resource orchestration 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], then multiple 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 over 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 to allocate multi-tier, pervasive intelligence in 6G systems.

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 is based on real world experiences and best practices in typical CT and IT networks. Table 1 lists all the parameters about tasks, three AI architectures, and two task scheduling algorithms, together with reasonable values 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 this limited paper, only delay and energy consumption are chosen as the illustrative KPIs for constructing a two-dimensional SRZ for every task. If task T is served by an mNode in the h^{th} tier, the overall service delay D_T consists of communication delay and computation delay, and can be expressed as

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

where the effective computing power C_h includes the combined effects of queueing, execution, and storage delays at the service mNode. Similarly, the total energy consumption E_T consists of transmission energy consumption and computation energy consumption, and can be expressed as

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

where α_i and γ denote the average transmission power over the i^{th} hop and the constant related to the service mNode's computing hardware structure, respectively. For analysis, we set $\alpha_i = 0.1$ Watts and $\gamma = 1 \times 10^{-33}$ in this study. 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, i.e., 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.

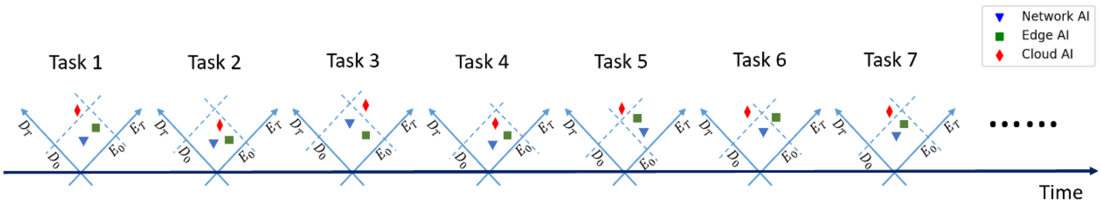


Fig. 3 Service Results of Representative Tasks with Different SRZs.

On the supply side, the cloud AI, edge AI, and network AI architectures are evaluated with the same total computing power of 14M 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 computing overhead for task scheduling and system resource management, so the effective computing power is $C = C_c = 14M$ teraFLOPS.

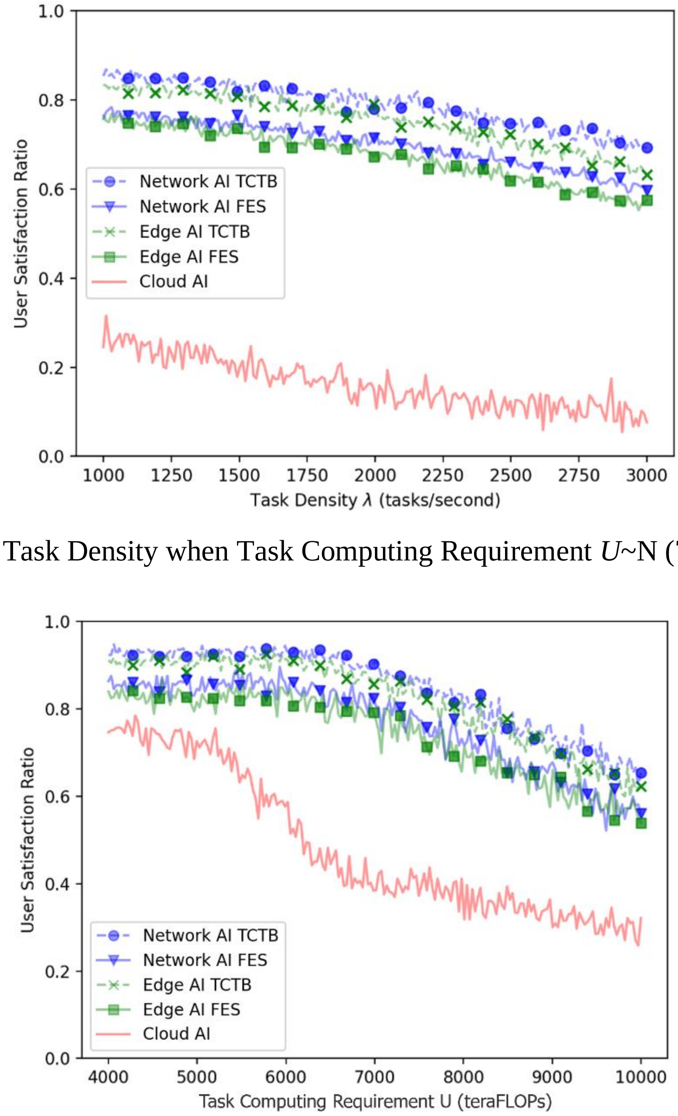
The edge AI architecture allocates a small amount of computing power among 1000 1^{st}-tier mNodes at the edge and the rest of computing power in the cloud. Assuming a 20% computing overhead for task scheduling and system resource management, the resulting effective computing power is equal to $C = N_1 \times C_1 + C_c = 11.12M$ teraFLOPS. In Table 1, two task scheduling algorithms are considered in performance evaluation. Under Fair Equal Scheduling (FES), all tasks are split in a random manner with half going to the edge and half to the cloud for services. While 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 dynamic operation scenarios and complex network conditions.

Table 1. Simulation Parameters.

<table>
<tr><th></th><th colspan="2">Parameter</th><th colspan="3">Value</th></tr>
<tr><td rowspan="5">User Task: Demand Side</td><td colspan="2">Task Density/Arrival Rate λ</td><td colspan="3">[1000, 3000] (tasks per second)</td></tr>
<tr><td colspan="2">Delay Bound D_0</td><td colspan="3">E[D_0]=1600 (seconds), Var(D_0)=50</td></tr>
<tr><td colspan="2">Energy Bound E_0</td><td colspan="3">E[E_0]=1.85(kW·h), Var(E_0)=0.05</td></tr>
<tr><td colspan="2">Task Size Z</td><td colspan="3">E[Z]$\in[4.8\times10^8, 7.2\times10^8]$(bytes)
Var(Z)=1×10^6</td></tr>
<tr><td colspan="2">Computing Requirement U</td><td colspan="3">E[U]$\in[0.4\times10^4, 1.0\times10^4]$ (teraFLOPS)
Var(U)=1×10^2</td></tr>
<tr><td rowspan="14">6G System: Supply Side</td><td colspan="2"></td><td>Cloud AI</td><td>Edge AI</td><td>Network AI</td></tr>
<tr><td colspan="2">Computing Overhead</td><td>0</td><td>2880000 (teraFLOPS)</td><td>3640000 (teraFLOPS)</td></tr>
<tr><td colspan="2">Effective Computing Power</td><td>14000000 (teraFLOPS)</td><td>11120000 (teraFLOPS)</td><td>10360000 (teraFLOPS)</td></tr>
<tr><td rowspan="2">Cloud</td><td>Computing Power C_c</td><td>14000000 (teraFLOPS)</td><td>10000000 (teraFLOPS)</td><td>7000000 (teraFLOPS)</td></tr>
<tr><td>Data Rate R_c</td><td colspan="3">2500 (Mbps)</td></tr>
<tr><td rowspan="3">3^{rd}-tier mNode</td><td>Number N_3</td><td>0</td><td>0</td><td>10</td></tr>
<tr><td>Computing Power C_3</td><td>-</td><td>-</td><td>112000 (teraFLOPS)</td></tr>
<tr><td>Data rate R_3</td><td colspan="3">E[R_3] $\in[1600, 2500]$ (Mbps), Var(R_3)=100</td></tr>
<tr><td rowspan="3">2^{nd}-tier mNode</td><td>Number N_2</td><td>0</td><td>0</td><td>100</td></tr>
<tr><td>Computing Power C_2</td><td>-</td><td>-</td><td>11200 (teraFLOPS)</td></tr>
<tr><td>Data Rate R_2</td><td colspan="3">E[R_2] $\in[400, 625]$ (Mbps), Var(R_2)=25</td></tr>
<tr><td rowspan="3">1^{st}-tier mNode</td><td>Number N_1</td><td>0</td><td>1000</td><td>1000</td></tr>
<tr><td>Computing Power C_1</td><td>-</td><td>1120 (teraFLOPS)</td><td>1120 (teraFLOPS)</td></tr>
<tr><td>Data Rate R_1</td><td colspan="3">E[R_1] $\in[56, 87.5]$ (Mbps), Var(R_1)=7</td></tr>
<tr><td rowspan="2">Algorithms: Supply Side</td><td colspan="2">Fair Equal Scheduling (FES)</td><td>100%</td><td>50% : 50%</td><td>25:25:25:25 %</td></tr>
<tr><td colspan="2">The Closer The Better (TCTB)</td><td>100%</td><td>80% : 20%</td><td>80: 10: 5: 5 %</td></tr>
</table>

The network AI system comprises of more mNodes with different capabilities in three network tiers, thus the computing overhead due to system and algorithm complexity is assumed higher, let us say $3.64M$ teraFLOPS. The total effective computing power is derived as $C = N_1\times C_1 + N_2\times C_2 + N_3\times C_3 + C_c = 10.36M$ teraFLOPS. Usually, a higher-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 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 are 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.



(a) Impact of Task Density when Task Computing Requirement U~N $(7 \times 10^3,\ 10^2)$.

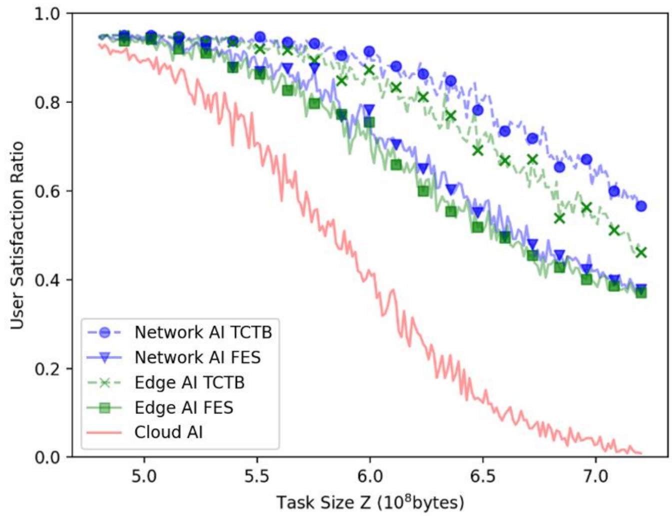
(b) Impact of Task Computing Requirement when Task Density λ=1000.

Fig. 4 USR versus Task Density and Computing Requirement.

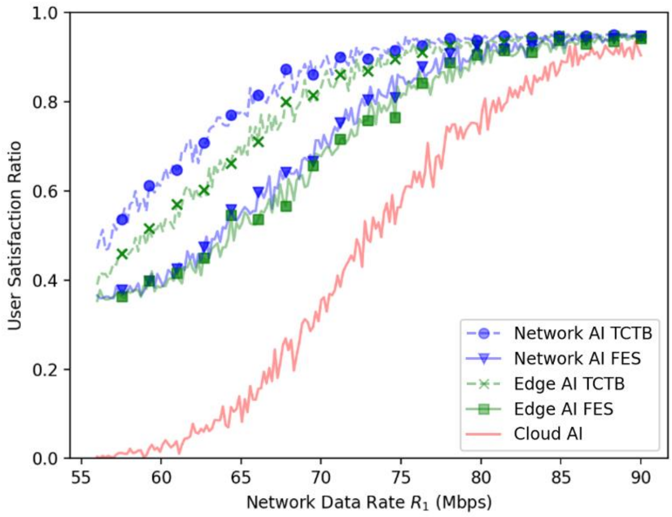
As defined, the overall USR can be derived by comparing the number of satisfied tasks against the total number of served tasks. When the Gaussian distributions of task size and network data rates are fixed, i.e., Z~N $(6 \times 10^8,\ 10^6)$, R_1~N (70, 7), R_2~N (500, 25), and R_3~N (2000, 100), Fig. 4 illustrate 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 5.1%, 8.1%, and 13.0% higher USR than the edge AI architecture, while 315.0%, 406.2%, and 457.1% higher USR than the cloud

AI architecture, respectively.

In Fig. 4 (b), the USR curve of cloud AI has two knee points at about *U*=4800 and *U*=6600 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 edge AI and network AI are much less sensitive to this change, which is due to the multi-tier mNodes deployed in the neighborhood. The turning points for TCTB and FES curves are around *U*=6800 teraFLOPS and *U*=7200 teraFLOPS respectively, where the gradients climb roughly from 0.31 to 0.43.



(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 (6×10^8, 10^6).

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

In Fig. 5 (a), for fixed task density λ=1000 and task computing requirement *U*~N (7×10^3, 10^2), when task size increases, the USR curve of cloud AI degrades dramatically because long-distance transmissions of bigger tasks become more time- and energy-consuming, thus adversely impacting the USR. On the contrary, the USR curves of edge AI and network AI 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 more 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 0.35.

Fig. 5 (b) demonstrates the influence of network data rates on the USR performance. Specifically, we assume R_1, R_2 and R_3 have different Gaussian distributions with dynamic 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 cloud AI gets very close to the curves of edge AI and network AI, just like the case when the average task size E[Z]<4.95×10^8 bytes in Fig. 5 (a).

VI. 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 individual requirements. From a user's perspective, this paper first proposed the concept of SRZ to characterize each task's combined performance requirements. Next, from a system's perspective, the concept of USR was introduced to evaluate the system's overall capability of satisfying personalized 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 can achieve the highest USR in all network conditions and operation scenarios. 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 games, intelligent manufacturing, and autonomous driving.

As to future work, the following open problems require further discussions and investigations from the community:

(1) **Statistic Models of SRZs:** we should study the combined service requirements of different types of realistic tasks in various application scenarios and operation conditions. We will research the new KPIs related to QoE, pervasive intelligence, and social benefits. Priorities should be considered for mission-critical tasks and elderly users.

(2) **Service Capacity of 6G Systems:** we should develop the practical mechanism of mapping 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 Provision:** we should complete the design of NALC and NAMO by developing a series of effective interfaces, protocols and algorithms for multi-tier,

cross-domain resource allocation, multi-node collaborations, task scheduling, behavior monitoring, security guarantee, mobility management, and performance optimization.

(4) **Implementation of 6G Network AI Architecture:** we should conduct a series of real experiments with multi-tier mNodes deployed in the 6G system for supporting different SRZs. Specifically, we will investigate some important practical issues such as training data splitting, AI model and algorithm dependency, and system complexity.

Acknowledgement

Dr. Yang Yang would like to thank Prof. Lajos Hanzo at University of Southampton and Prof. Raymond Yeung at the Chinese University of Hong Kong for their valuable comments on a draft version of this paper. This work was partially supported by the National Key Research and Development Program of China (2020YFB2104300), the Major Key Project of the Peng Cheng Laboratory (PCL2021A15), and the Joint Funds of the National Natural Science Foundation of China (U21B2002).

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.