Network Control and Management Automation: Architecture Standardization Perspective

Ved P. Kafle, Takahiro Hirayama, Takaya Miyazawa, Masahiro Jibiki, and Hiroaki Harai

Abstract

Artificial intelligence (AI) and machine learning (ML) techniques are becoming essential to automate the control and management of complex systems, such as 5G and future networks and services, through the collection and timely processing of a large amount of control data from the system. The standardization activities of AI/ML integrated network architectures and frameworks have recently been progressing in various standards developing organizations (SDOs). This article provides a detailed survey of the AI/ML network standardization activities of the International Telecommunication Union. An overview of the related activities of other SDOs, such as the European Telecommunications Standards Institute and the 3rd Generation Partnership Project, is also provided. Finally, a set of related open issues that deserve further research and standardization are discussed.

Introduction

Fifth generation (5G) networks and beyond will be built on the convergence of virtualized networking and cloud/edge computing infrastructures. These networks are configured through software using the recently developed technologies of software-defined networking (SDN) and network function virtualization (NFV) [1]. A virtualized computing infrastructure can be segmented into multiple virtual machines or containers, each of which can be used to deploy and operate one or more cloud-native network functions or virtual network functions (VNFs) [2]. SDN enables the configuration of virtual network slices by orchestrating VNFs remotely through control functions implemented in the software. Moreover, the computational (e.g., CPU and memory), storage, and networking (e.g., buffer size and bandwidth) resources allocated to each VNF can be adjusted dynamically such that its performance can always meet the quality-of-service requirements of the delivered communication service [3].

In addition, 5G networks, which are also referred to as IMT-2020 networks in the International Telecommunication Union (ITU) standards, have been designed and developed to offer three types of communication services: enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low-latency communication (URLLC) [4]. High-definition video (e.g., 4K video) and augmented/virtual reality (AR/VR) applications that require very high bandwidth are supported by eMBB services, and Internet of Things (IoT) devices and smart utility meters connected to the network in large numbers are supported by an mMTC service, whereas automated vehicles that require ultra-high reliability and low latency are supported by a URLLC service. These three types of communication services can be offered from the same sharable virtualized cloud computing and networking infrastructure by configuring network slices containing the required VNFs [3].

Unlike traditional telecommunication infrastructures, in which the network functions are mostly implemented in hardware, the control and management of virtualized, programmable, software-driven, and distributed cloud/edge computing infrastructures are challenging because of the proliferation of vendor-specific platforms and management technologies. Moreover, the emerging demand for more complex services (e.g., context-aware and personalized services) for a large number of user subscriptions has made network control and management extremely challenging.

Starting from the network service design, resource provisioning, and deployment for performance monitoring, resource adjustment, failure detection, and recovery, network control and management operations require the analysis of a huge volume of control data collected from various components of the network. Under this complex situation, the capacity of the human operator does not meet the requirements of quickly processing a large volume of data and providing a timely response. To address this issue, advanced techniques of artificial intelligence (AI) and machine learning (ML) have recently been applied in network control and management functions [5, 6].

The effectiveness of ML applications for workload and resource utilization prediction, and accordingly adjusting the computational resources and reconfiguring the network to maintain the required quality of service at all times, was demonstrated in a paper presented at the ITU Kaleidoscope Academic Conference 2020 [7]. Extending the standardization aspects identified in that paper, this article provides a survey of AI/ML integrated network standardization activities of various standards developing organizations (SDOs), mainly focusing on the activities of the International Telecommunication Union (ITU). ITU's Telecommunication Standardization Sector (ITU-T) has recently produced several standard documents (i.e., ITU-T Recommendations) on the high-lev-

Digital Object Identifier:
10.1109/MCOMSTD.121.2100043

The authors are with the National Institute of Information and Communications Technology, Tokyo.

2471-2825/21/$25.00 © 2021 IEEE

el architectures and frameworks of ML-controlled networks. In this article, the latest developments in ITU-T AI/ML high-level architectures of data handling, network intelligence evaluation, network service provisioning, resource and fault management, and quality of service assurance are reviewed. Similarly, the reference architectures of network and service management developed by the European Telecommunications Standards Institute's (ETSI's) two Industry Specification Groups (ISGs), Experiential Networked Intelligence (ENI) and Zero Touch Network and Service Management (ZSM), as well as Network Data Analytics Function (NWDAF) of the 3rd Generation Partnership Project (3GPP) are reviewed. It also discusses a set of open issues that deserve further research and standardization for the realization of intelligent and autonomous network control and management.

AI/ML Network Standardization in ITU

ITU started a study and standardization of AI/ML integration in telecommunication networks in 2017 with the establishment of the Focus Group on Machine Learning for Future Networks, including 5G (FG ML5G) [8]. This section introduces the ITU-T Recommendations developed on the basis of FG ML5G deliverables, ITU-T Recommendations developed from contributions by ITU-T Study Group 13 (SG13), and an overview of the activities of the recently established Focus Group on Autonomous Networks.

ITU-T Recommendations from FG ML5G Deliverables

FG ML5G studied the use cases, requirements, data formats, interfaces, intelligence-level evaluation, and ML-enabling network architecture among three working groups. In its two-and-a-half-year lifetime (2018–2020), FG ML5G produced 10 deliverables (the list is available on the FG ML5G website [8]) and submitted them to ITU-T SG13 for the consideration as ITU-T Recommendations.

ITU-T SG13 further improved the content of the FG ML5G deliverables and approved the following documents as one ITU-T Supplement and four ITU-T Recommendations:

- ITU-T Supplement 55: Machine learning in future networks including IMT-2020: use cases
- ITU-T Recommendation Y.3172: Architectural framework for machine learning in future networks, including IMT-2020
- ITU-T Recommendation Y.3173: Framework for evaluating intelligence levels of future networks, including IMT-2020
- ITU-T Recommendation Y.3174: Framework for data handling to enable machine learning in future networks, including IMT-2020
- ITU-T Recommendation Y.3176: Machine learning marketplace integration in future networks including IMT-2020
- ITU-T Recommendation Y.3179: Architectural framework for ML model serving in future networks, including IMT-2020

ITU-T Supplement 55 describes 30 use cases and their requirements. The use cases are classified into the following five groups: network slice and service, user plane, applications, signaling and management, and security. Similarly, the requirements are mapped to three groups of functions: data collection, data storage and processing, and ML models. The requirements are labeled as critical, expected, and added values.

ITU-T Recommendation Y.3072 specifies the requirement and high-level architecture for integrating ML in future networks, including IMT-2020. It also provides example guidelines for mapping this ML-based architecture framework with a standardized 5G network architecture. As shown in Fig. 1, the architecture framework consists of four subsystems: the ML pipeline subsystem, ML sandbox subsystem, management subsystem, and ML underlay networks. The ML pipeline subsystem contains a set of ML models and execution abstraction. It consists of several functional components, such as the input data collector (C), preprocessor (PP), ML models (M), policy (P), and output result distributor (D). The input data are collected through a source function (SRC), and the ML output results are distributed to the network controllers through a sink function (SINK). The ML sandbox subsystem also contains the same functional components as the ML pipeline subsystem. The ML sandbox subsystem uses simulated underlay networks to train and test ML models. Trained ML models are then deployed in the ML pipeline subsystem to monitor and control live ML underlay networks. The management subsystem consists of an ML function orchestration (MLFO) and other management functions. The MLFO function obtains the input of the ML intents, and manages and orchestrates the functional components of the ML pipeline and ML sandbox. A detailed description of these functions and interfaces can be found in [9].

ITU-T Recommendation Y.3173 starts with a description of the progress in network intelligence levels and then specifies methods and an architecture for the network intelligence level evaluation. The standard methods for evaluating network intelligence levels can provide a decision mechanism to operators, equipment vendors, and other network industry participants for planning network technology features and product roadmaps. The following five dimensions are provided for an intelligence level evaluation: demand mapping, data collection, analysis, decision, and action implementation. Demand mapping corresponds to the process of converting the network configuration or requirements given by a human operator into specific instructions that network components can understand and execute. Data collection and analysis dimensions are related to the process of collecting network monitoring and control data and conducting an analysis.

The decision and action implementation dimensions relate to the process of making network or service configuration decisions and, accordingly, command executions.

To evaluate the overall network intelligence level of an entire network, the intelligence levels of individual workflows and network subsystems need to be evaluated with respect to the five dimensions. The workflows constitute the tasks of planning, deployment, maintenance, optimization, and provisioning. Similarly, the network subsystem includes components such as network elements, management subsystems, and application platforms. To determine the overall network intelligence level, a basic method is provided, which is based on the network intelligence capability level

ITU-T Supplement 55 describes 30 use cases and their requirements. The use cases are classified into the following five groups: network slice and service, user plane, applications, signaling and management, and security. Similarly, the requirements are mapped to three groups of functions: data collection, data storage and processing, and ML models. The requirements are labeled as critical, expected, and added values.

The 3GPP has defined the network data analytics function (NWDAF) in the 5G service-based architecture (SBA) as the enabler of intelligence and autonomous network operation and management. The NWDAF collects data from various modules of the 5G system and provides analysis results.

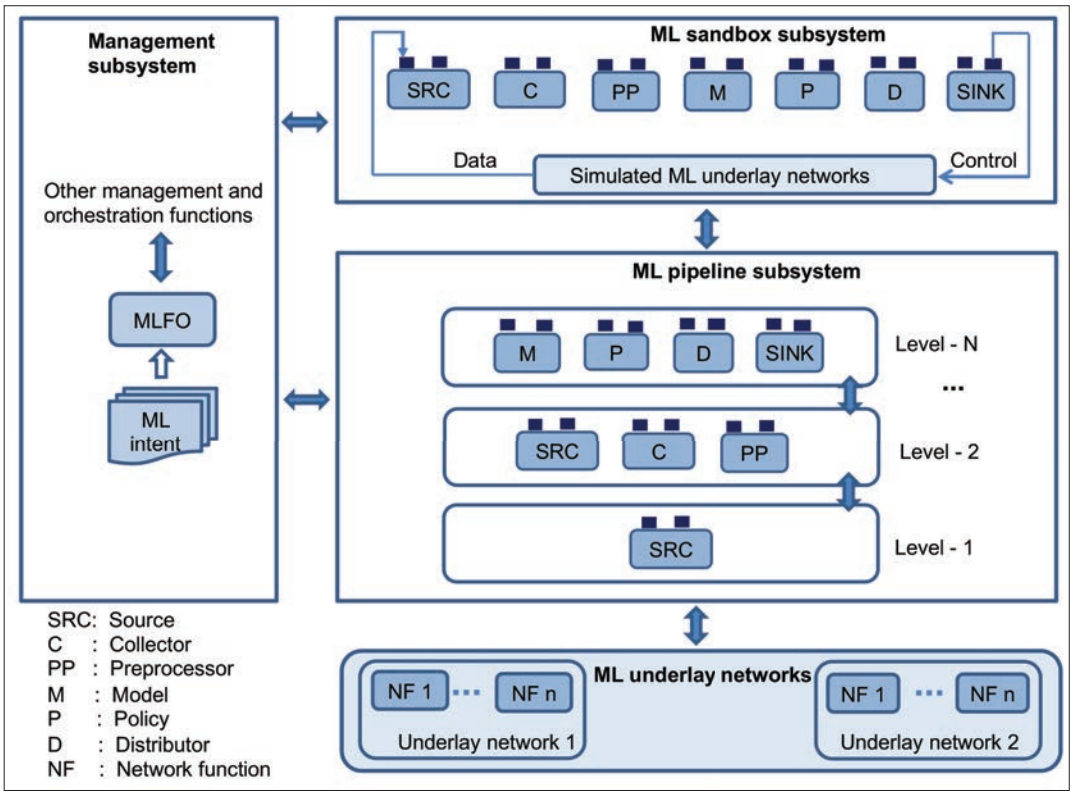

Figure 1. High-level architecture for integrating machine learning in future networks including IMT-2020.

of each of the five dimensions. Network intelligence capability levels are determined based on whether the function of a dimension is executed by a human only, a human and a system, or a system only. Based on the combinations of the intelligence capability levels of individual dimensions, six intelligence levels are defined, as shown in Table 1. In the lowest network intelligence level of L0, which corresponds to a manual network operation, all dimensions are executed by humans. Similarly, in the highest network intelligence level of L5, which corresponds to full intelligence, all dimensions are autonomously executed by the system only. In between, there are intermediate intelligence levels in which the tasks of some dimensions require human involvement.

ITU-T Recommendation Y.3074 specifies the ML data handling framework to deal with the diversity of the control data produced by various components in the network. It introduces data models, brokers, and application program interfaces (APIs) in both the user and control planes. It also provides the requirements for input data collection, processing, and output data.

ITU-T Recommendation Y.3176 describes the challenges, motivations, requirements, and architecture for ML marketplace integration in networks. It defines the ML marketplace as a repository of interoperable trained AI/ML models. It also specifies a method that uses the ML intent and MLFO to select appropriate ML models from the ML marketplace, and interfaces to connect the ML marketplace with the ML sandbox and the ML pipeline.

Similarly, ITU-T Recommendation Y.3179 specifies an architectural framework for ML model serving, that is, the preparation and deployment of ML models in different deployment environments. The ML model service takes place in the following three fundamental stages: inference optimization, model deployment, and model inference. In the inference optimization stage, trained ML models are modified to improve the performance when executing an inference in a certain deployment environment according to the requirements of the use case and the current state of the network. In the model deployment stage, the ready-to-run ML model is deployed in a specific deployment environment. Finally, in the inference stage, the model inference output result (e.g., prediction or classification) is applied to ML pipelines.

In addition to the above ITU-T Recommendations produced by enriching FG ML5G deliverables, ITU-T SG13 has also produced Recommendations from contributions submitted by its delegates, which are described next.

ITU-T Recommendations from SG13 Contributions

From the contributions of its delegates, ITU-T SG13 has produced three Recommendations, Y.3175, Y.3177, and Y.3178, which are briefly described below.

ITU-T Recommendation Y.3175 specifies the functional architecture of ML-based quality of service (QoS) assurance, including the reference points. It also describes the procedures for ML-based QoS assurance in the IMT-2020 network.

By extending the basic ML architecture specified in ITU-T Y.3172, ITU-T Recommendation Y.3177 specifies a high-level architecture of AI/ML-based network automation for resource and fault management. Figure 2 shows the architecture, which is composed of four subsystems. The management subsystem contains the following three management functions: resource management, fault management, and other manage-

Network intelligence level		Dimensions				
		Action implementation	Data collection	Analysis	Decision	Demand mapping
L0	Manual network operation	Human	Human	Human	Human	Human
L1	Assisted network operation	Human and system	Human and system	Human	Human	Human
L2	Preliminary intelligence	System	Human and system	Human and system	Human	Human
L3	Intermediate intelligence	System	System	Human and system	Human and System	Human
L4	Advanced intelligence	System	System	System	System	Human and System
L5	Full intelligence	System	System	System	System	System

TABLE 1. Network intelligence levels.

ment functions. It also contains the ML functional orchestrator (MLFO), which takes ML intents as input. The AI/ML pipeline consists of six functional groups: data collection, fault detection, fault recovery, resource prediction, resource adaptation, and controller. The AI/ML sandbox is composed of an AI/ML pipeline for the purpose of training AI/ML models by using the data obtained from the simulated AI/ML underlay network. The reader is referred to [10] for a detailed description of these functional groups and interfaces.

Similarly, ITU-T Recommendation Y.3178 specifies a functional framework for AI-based network service provisioning. It starts with a description of the business-role-based model for AI-based network service provisioning, then provides a list of high-level requirements for the roles and their interactions from an AI-based operational perspective, as well as the functional framework showing the components and their interactions for AI-based operations for network service provisioning.

Focus Group on Autonomous Networks

ITU-T SG13 established the Focus Group on Autonomous Networks (FG AN) [11] in December 2020 to provide an open platform for conducting pre-standardization activities on autonomous networks. The terms of reference of FG AN define autonomous networks as those that possess the ability to monitor, operate, recover, heal, protect, optimize, and reconfigure themselves. FG AN studies the autonomy of various processes or network aspects, such as planning, security, audits, inventory, optimization, orchestration, and quality of experience. The group studies the various approaches of exploratory evolution, emergent behavior, and real-time responsive experimentations that can provide a new layer of abstraction for introducing evolution mechanisms leading to the realization of autonomous networks. The group also tries to address the questions associated with accountability for non-human decisions that affect customers and explore the approaches of exploratory evolution, emergent behavior, and real-time responsive experimentation to enable an autonomous network.

Over the course of a one-year period, FG AN has planned to produce deliverables defining the characteristics, use cases, requirements, proofs of concepts, high-level architecture, standardization gap analysis, specification languages, and representations of autonomous networks.

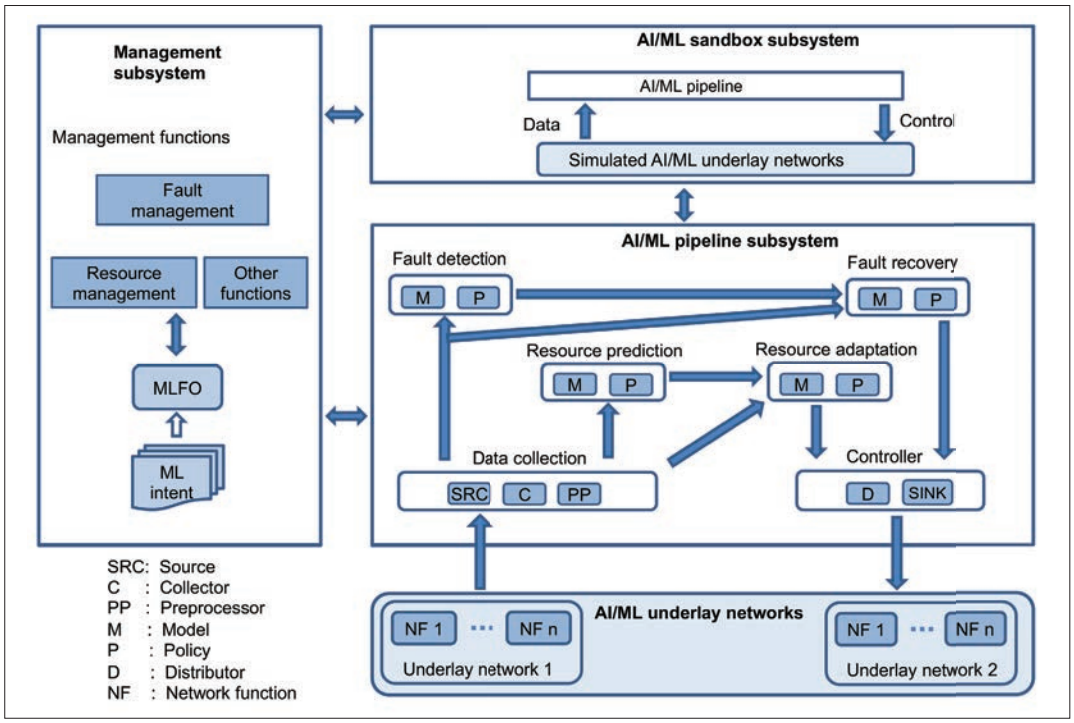

FIGURE 2. High-level architecture of AI/ML-based network automation for resource and fault management.

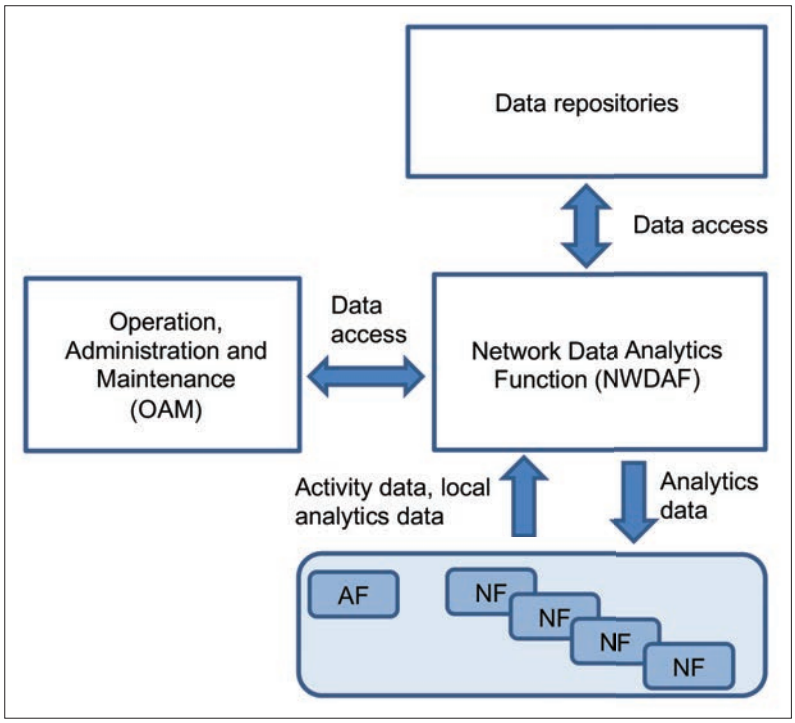

Figure 3. General framework of 3GPP 5G network automation using network data analytics function.

AI/ML Network Standardization in Other SDOs

This section presents a review of network-automation-related activities of 3GPP and ETSI.

3GPP Network Automation for 5G

3GPP has defined the network data analytics function (NWDAF) in the 5G service-based architecture (SBA) as the enabler of intelligence and autonomous network operation and management. The NWDAF collects data from various modules of the 5G system and provides analysis results.

3GPP TR 23.791 specifies a general framework for 5G network automation [12]. As shown in Fig. 3, the NWDAF collects data from the operation, administration, and maintenance (OAM) module, application functions (AFs), 5G core network functions (NFs), and data repositories. In addition, 3GPP TS 23.288 specifies the reference architecture and detailed procedures for data analytics [13]. It also provides a description of NWDAF discovery and selection by NFs/AFs, network performance analytics, and user data congestion analytics.

NWDAF analyzes the data by leveraging AI/ML models. The analytic results are then delivered to the NFs/AFs that have requested the NWDAF. The NFs/AFs utilize analytical output data to make appropriate decisions for network operation and management actions. NWDAF utilizes existing 5G service-based interfaces to collect data from NFs/AFs and OAM, as well as deliver analytical output data to them.

Because the input data of the NWDAF may come from multiple sources, such as mobility management, session management, QoS management, application layer, security management, and NF life cycle management, the resulting actions that an NF or AF takes according to the analytical feedback may be concerned with many of these functional domains. TR 23.791 provides guidelines for the application of NWDAF to various use cases belonging to one or several domains. In the description of each use case, the following six aspects are included: domain (e.g., performance, QoS, resilience, and security) and timescale, nature of input data (e.g., systems logs, KPI, and events), nature of the output data, types of NFs consuming NWDAF analytics output data, examples of actions taken, and benefits (e.g., revenue, resource saving, quality of experience assurance, and reputation enhancement).

ETSI Experiential Networked Intelligence

ETSI's Experiential Networked Intelligence Industry Specification Group (ENI ISG) is targeting the development of efficient and extensible standards-based mechanisms to provide context-aware services. It has specified an experiential architecture that uses AI/ML and other mechanisms to improve its understanding of the network environment, and thus the operator experience, over time. It can adapt its functionality based on contextual changes in user requirements, network conditions, and business goals [14].

Figure 4 shows the ETSI ENI system architecture, which consists of three modules: the input processing module, analysis module, and output generation module. The figure also shows the API broker; however, the ENI system can function with or without the API broker. The API broker serves as a gateway between different systems. It possesses the translation mechanisms to translate data communicated from the external system into a normalized form that all ENI functional blocks can understand, as well as translate recommendations and commands from the normalized form of the ENI system to a form that the external system can understand. Thus, it enables heterogeneous types of external systems such as infrastructure, applications, and users to interoperate with the ENI system.

The input processing and normalization module possesses data ingestion and normalization functional blocks to process the input data such that the other functional blocks in an ENI system can interpret and understand the data in a unified and consistent manner.

The analysis module includes context-aware, knowledge management, cognitive processing, situation-aware, model-driven, and policy management functional blocks. The context-aware functional block describes the state and environment in which a set of entities in the controlled or assisted system (i.e., the system being assisted and/or controlled by the ENI system) exist. The context consists of measured and inferred knowledge that may change over time. The knowledge management functional block includes the mechanisms for knowledge representation, inference, and reasoning to represent information about both the ENI system and the controlled external system.

The cognition processing functional block includes a mechanism to understand normalized ingested data and information, as well as the context that defines how those data were produced. Based on data interpretation results, it determines whether any action needs to be taken to ensure that the goals and objectives (e.g., improving or

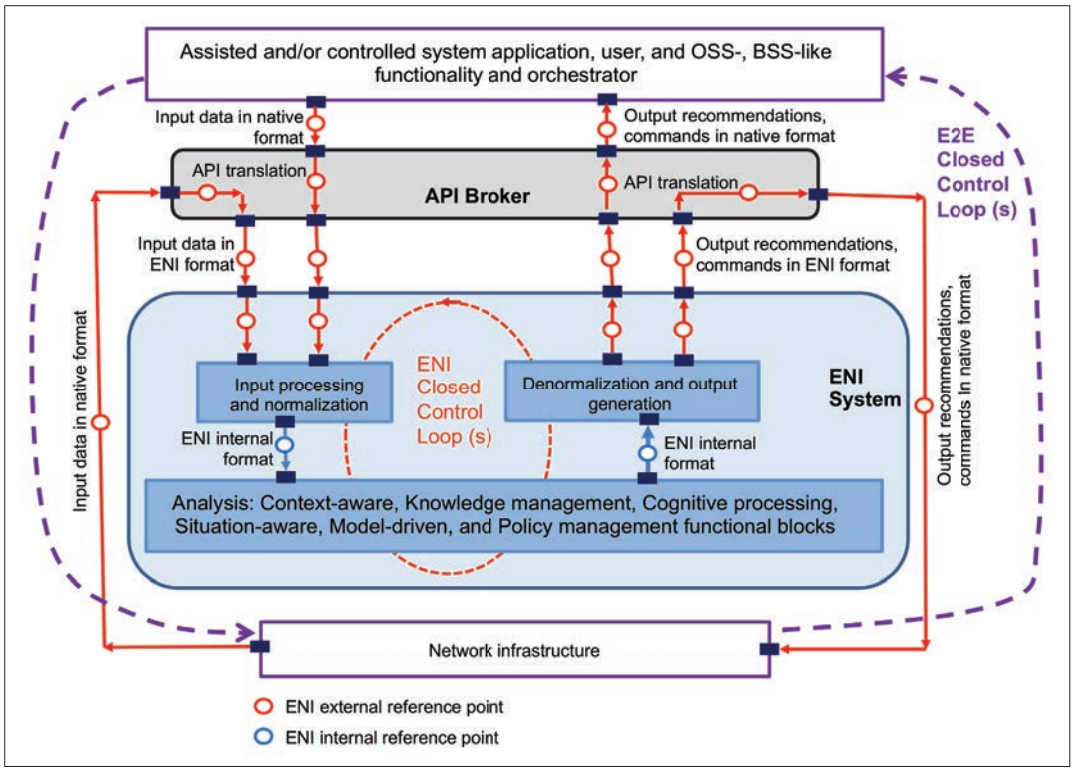

Figure 4. High-level ENI system architecture with API broker.

optimizing the performance, reliability, and/or availability) of the ENI system have been met. The situation awareness functional block includes a mechanism to enable the ENI system to be aware of events and behaviors that are relevant to the controlled system. It has the capacity to understand how information, events, and recommendation commands given by the ENI system impact the management and operational goals and behavior in the short and long terms.

The model-driven engineering functional block contains a set of models that collectively abstract all important concepts for managing the control system governed by the ENI system. The policy management functional block provides a set of rules to manage the system in such a way that the system goals and objectives are met.

Similarly, the denormalization functional block of the denormalization and output generation module includes mechanisms to process and translate policies, recommendations, and data received from other functional blocks of the ENI system into an intermediate form that can be subsequently translated or transcoded by the output generation functional block into a form that the controlled systems are able to understand and use. If an API broker exists, the output is sent to the API broker; otherwise, it is sent directly to the controlled system.

The ENI system architecture was designed based on the key assumption that the ENI system functionality evolves over time to meet emerging functional requirements such as network and service planning requirements, service provisioning, deployment, optimization, data collection, modeling, analysis, policy specifications, interoperability with other systems, non-functional requirements of system performance (latency, accuracy, and efficiency), and scalability, among other factors. The ENI architecture can be applied to various aspects of network management, such as infrastructure management, network operation, service orchestration, and the management of numerous use cases.

ETSI Zero Touch Network and Service Management

ETSI ISG Zero Touch Network and Service Management (ZSM) specifies the architecture, functional, and operational requirements for E2E networks, and service automation based on the closed-loop control and integration of AI/ML techniques. The ZSM architecture aims to address the challenges of technological and managerial heterogeneity in E2E cross-domain network management by defining a holistic management framework that can reuse the management capabilities available in various standard technologies. It follows the principles of modularity, extensibility, scalability, model-driven open interfaces, closed-loop management automation, support of stateless management functions, resilience, intent-driven interfaces, and simplicity, among others [15].

Figure 5 shows the ETSI ZSM reference architecture, which is composed of a two-layer hierarchical structure of management domains. In the lower layer, there are multiple management domains (MDs), each of which is responsible for managing a domain-managed infrastructure; in the upper layer, this is an end-to-end (E2E) service management domain, which orchestrates the management services provided by individual management domains to realize E2E cross-domain management. Both the individual and E2E management domains expose a set of management services that they provide.

Management domain services can be categorized into the following groups. Domain data

The ENI system architecture was designed based on the key assumption that the ENI system functionality evolves over time to meet emerging functional requirements such as network and service planning requirements, service provisioning, deployment, optimization, data collection, modeling, analysis, policy specifications, interoperability with other systems, non-functional requirements of system performance , and scalability, among other factors.

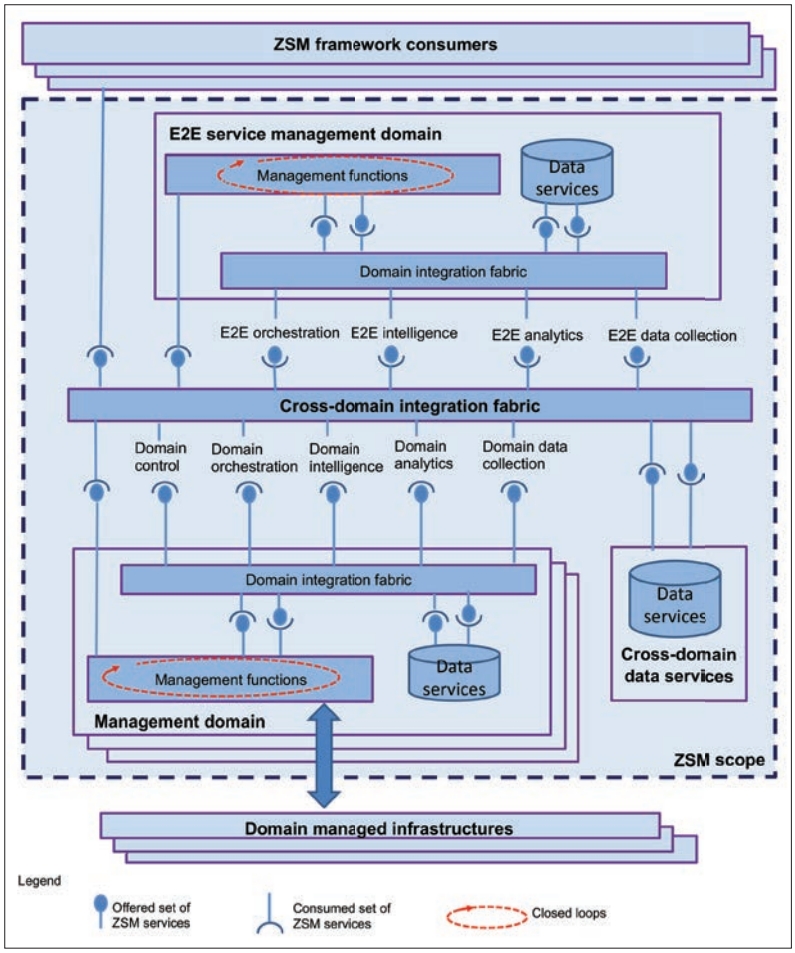

Figure 5. ETSI ZSM framework reference architecture.

collection services monitor the domain-managed infrastructure and collect data as demanded by other management functions. Domain analytics services apply various data analytics to the collected data to generate domain-specific knowledge and predict domain-specific behavior. Domain intelligence services execute closed-loop automation functions of decision support, decision making, and action planning to enable autonomous management of the domain-managed infrastructure. Domain orchestration services automate workflows and processes inside a management domain to handle the life cycle management of the domain-managed infrastructure. Domain control services facilitate the control of each management entity.

Similarly, the E2E service management domain contains the following groups of management services: E2E service orchestration services are responsible for the coordination of provisioning, configuration, and life cycle management of various services across management domains. E2E service intelligence services are responsible for driving closed-loop automation in E2E networks and service management. E2E service analytics services are responsible for deriving E2E service insight for managing E2E network service performance. E2E service data collection services are responsible for collecting E2E service-related data. As in the individual management domains, each E2E management service is offered through one or more interfaces.

Data services provide the means of data sharing with authorized consumers within and across management domains while taking care of information security and data privacy regulations. Data services may store and provide performance monitoring data, trace data, configuration data, topology data, and inventory data, among other data types.

An integrated fabric exists inside each management domain. There is also a cross-domain integration fabric between each management domain and the E2E service management domain. The integration fabric provides a set of ZSM services to facilitate the interoperation of management functions with regard to the offered and consumed management services. Services offered by the integration fabric include the registration/de-registration of management services, discovery of registered management services and the means to access them, the means to support the invocation of management services, and the means to support synchronous and asynchronous communication. Similarly, the cross-domain integration fabric offers the services of registration, discovery, communication, invocation, and capability exposure configuration of management services.

Key Architecture Features

As common features of intelligent and autonomous network management framework architectures being developed in ITU-T, 3GPP, ETSI ENI, and ZSM, all are based on data collection (from various network functions, application functions, and other repositories) and data processing and analysis using AI/ML models. The key features of these architectures are listed in the second column of Table 2.

The last column mentions the target networks that fall in the scope of each of these architectures. Both ITU-T and 3GPP have considered 5G networks; however, the ITU-T ML architecture is focused mainly on fixed wireline networks, whereas 3GPP SBA 5G architecture also includes radio access networks in the scope. ETSI does not assume any specific network type but considers a generalized network architecture. In contrast to the ITU-T, 3GPP, and ETSI ENI framework architectures, which mainly focus on the management of a single administration domain, the ETSI ZSM architecture specifies a holistic management framework to enable the interworking of various domain-specific network management standards to support E2E service management across different technology and administrative domains.

Network Automation Research Issues

Although several framework architectures have been developed in various SDOs, the detailed specifications of the related protocols, mechanisms, and technologies remain to be developed. This section discusses a set of related open research and standardization issues for intelligent and autonomous networks and service management.

Heterogeneity in Multiple Domains: The first issue is the challenges in the E2E management of heterogeneous technology and administration domains. As stated earlier, the E2E management of network services in cloud-native future net-

Architectures	Key features	Target networks
ITU-T ML architecture	• Pipelining of ML models • Defining network intelligence levels • Integrating ML model marketplace • Integrating ML sandbox for model training	5G and beyond, mainly fixed wireline network aspects
3GPP 5G SBA architecture	• NWDAF as key component • Using 5G service-based interfaces to access data from OAM module, AF, NF, and repositories, and deliver analysis results • Applicability of NWDAF in various use cases	5G architecture, including radio access network aspects
ETSI ENI architecture	• Generalized adaptive architecture capable of learning from the context and evolving the system functionality over time • Focused on data processing and analysis, together with knowledge management • Applicable to various aspects of network management, e.g., infrastructure, operation, service orchestration	Not specific to 5G, assuming a generalized architecture
ETSI ZSM architecture	• Holistic management framework for reusing management capabilities available in various standard technologies • E2E service management through bridging individual network management domains	Not specific to 5G, focused on interworking of multiple technological and administrative domains

TABLE 2. Key architecture features.

Hierarchical management layers have been proposed for E2E network and service management (e.g., the ZSM architecture contains open interfaces, model-driven services, and resource abstraction). However, they lack detailed specifications of interfaces that can be followed to implement mechanisms for the exchange of management capabilities and related data among multiple management domains.

works formed by the convergence of networking and cloud/edge computing infrastructures is challenging due to the involvement of different technologies and administrative domains, such as the domains of mobile operators, edge-computing service providers, core network operators, and cloud-computing service providers. Each of these domains may employ different types of network resource management mechanisms. In a multi-vendor E2E communication service delivery environment, each vendor network may have its own operation and business support system (OSS/BSS) with different control interfaces, resource virtualization and SDN technologies, AI/ML techniques, and models. To address this issue, the 5G service-based architecture of 3GPP applies SBI between the NWDAF and other network functions; in addition, ZSM has an E2E service management domain, ENI has an API broker, and the ML network architecture of ITU-T has an MLFO. However, they lack detailed specifications for the design and implementation of an easy-to-deploy multi-domain management mechanism that autonomously re-composes and reconfigures itself to provide near-optimal performance for different types of network services. For example, the standard representations of ML models, managed resources, and management methods to be used in each management domain have yet to be defined. Only after having standardized representations can these entities be orchestrated from the overlay E2E service management domain.

Abstraction of Hierarchical Management Layers: Hierarchical management layers have been proposed for E2E network and service management (e.g., the ZSM architecture contains open interfaces, model-driven services, and resource abstraction). However, they lack detailed specifications of interfaces that can be followed to implement mechanisms for the exchange of management capabilities and related data among multiple management domains. Moreover, because the volume of management data produced in each management domain is extremely large, an appropriate level of abstraction in each management domain is required to expose only the relevant information in the standard form and size.

Scalability of Telemetry Data Collection: For intelligent and autonomous E2E network and service management, the existence of standard technology for agile monitoring and control of all involved network functions is essential. The monitoring and control functions include a process for network telemetry data collection, which needs to be efficient to avoid incurring a high overhead hampering network performance and delaying the execution of control commands. The intelligent and autonomous telemetry data collection may require attaching an AI/ML process to each network function to carry out optimal decisions regarding the right amount and time of telemetry data collection to ensure the efficiency and scalability of the overall system.

Data Models: Standard data models are essential for sharing cross-domain control data for the realization of autonomous E2E networks and service management. Cross-domain data services in the ZSM reference architecture and data handling framework in ITU-T Y.3174 have been defined and specified at the conceptual level. The ZSM framework architecture also provides definitions and requirements for data collection, data storage, data persistence, and data processing services. However, they lack detailed specifications of data models that can be followed to implement scalable mechanisms for efficiently exchanging relevant data across multiple management domains.

AI/ML Pipelining: To realize scalable E2E network management, there must be a mechanism for collaboration between the AI/ML models used within the same local management domain as well as across different management domains. The ML pipelining concept has been mentioned in the ITU-T ML architecture [9]. Similarly, the ZSM framework's hierarchical management structure also assumes the existence of domain-specific ML or data analytics models [15]. However, the detailed design of management architecture that can fully leverage AI/ML capabilities for cross-do-

Although framework architectures are being developed in various SDOs, detailed specifications of related protocols, mechanisms, and technologies are still missing. Therefore, future research and standardization are required to develop detailed specifications that can be followed to implement the framework architectures.

main management is still missing, which deserves further research and standardization.

Conclusion

This article surveys the status of the standardization of AI/ML-based network and service management architectures in ITU-T, 3GPP, and two ISGs of ETSI: ENI and ZSM. The ITU-T architectures are based on AI/ML model pipelining for the autonomous operations of control information handling, network intelligence evaluation, network service provisioning, resource and fault management, and quality of service assurance. Similarly, the 3GPP 5G service-based architecture has the NWDAF function that uses 5G service-based interfaces to access data from OAM modules, application functions, network functions, and repositories, and delivers analysis results to them. The ETSI ENI architecture is capable of learning from context and evolving in terms of system functionality over time. The ETSI ZSM architecture presents a holistic E2E service management framework for reusing management capabilities available in various standard technologies used in individual network management domains.

Although framework architectures are being developed in various SDOs, detailed specifications of related protocols, mechanisms, and technologies are still missing. Therefore, future research and standardization are required to develop detailed specifications that can be followed to implement the framework architectures. For future study, the issues of interworking of heterogeneous multiple domains, abstraction of management layers, scalable telemetry, data models, and AI/ML pipelining are also discussed. To realize the concept of AI/ML pipelining, standardization activities are required to specify the design, implementation, and operational cycles that allow AI/ML models from different domains to be combined in the same way that software libraries and modules are combined in general-purpose computing platforms.

Acknowledgment

This study was conducted partially as a part of the project "Research and development for innovative AI network integrated infrastructure technologies (JPMI00316)," supported by the Ministry of Internal Affairs and Communications, Japan.

References

[1] ITU-T Rec. Y.3104, "Architecture of the IMT-2020 Network," Dec. 2018.
[2] S. Sharma, R. Miller, and A. Francini, "A Cloud-Native Approach to 5G Network Slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, Aug. 2017, pp. 120127.
[3] V.P. Kafle *et al.*, "Adaptive Virtual Network Slices for Diverse IoT Services," *IEEE Commun. Stds. Mag.*, vol. 2, no. 4, Dec. 2018, pp. 33–41.
[4] ITU-R Rec. M.2083-0, "IMT Vision "Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond," Sept. 2015.
[5] V. P. Kafle, P. Martinez-Julia, and T. Miyazawa, "Automation of 5G Network Slice Control Functions With Machine Learning," *IEEE Commun. Stds. Mag.*, vol. 3, no. 3, Sept. 2019, pp. 54–62.
[6] D. Rafique and L. Velasco, "Machine Learning for Network Automation: Overview, Architecture and Applications," *IEEE/OSA J. Opt. Commun. Net.*, vol. 10, no. 10, Oct. 2018,
[7] V. P. Kafle and A.H.A. Muktadir, "Automation of Computational Resource Control of Cyber-Physical Systems With Machine Learning," *Proc. ITU Kaleidoscope Academic Conf.*, Dec. 2020
[8] ITU-T Focus Group on Machine Learning for Future Networks Including 5G; https://www.itu.int/en/ITU-T/focusgroups/ ml5g/Pages/default.aspx, accessed 15 Apr. 2021.
[9] ITU-T Rec. Y.3172, "Architectural Framework for Machine Learning in Future Networks Including IMT-2020," June 2019.
[10] ITU-T Rec. Y.3177, "Architectural Framework of Artificial Intelligence-Based Network Automation for Resource and Fault Management in Future Networks Including IMT-2020," Feb. 2021.
[11] ITU-T Focus Group on Autonomous Networks, https://www.itu.int/en/ITU-T/focusgroups/an/Pages/ default.aspx), accessed 15 Apr. 2021.
[12] 3GPP TR 23.791, "Study of Enablers for Network Automation for 5G," v. 16.2.0, 2019.
[13] 3GPP TS 23.288, "Architecture Enhancements for 5G System (5GS) to Support Network Data Analytics Services," v. 16.6.0, Release 16, 2021.
[14] ETSI GS ENI 005, "Experiential Networked Intelligence (ENI) System Architecture, v. 1.1.1, 2019.
[15] ETSI GS ZSM 002, "Zero Touch Network and Service Management (ZSM) Reference Architecture," v. 1.1.1, 2019.

Biographies

Ved P. Kafle [S'04, M'07, SM'14] received his B.E. degree in electronics and communications engineering from Punjab Engineering College, Chandigarh, India, his M.S. degree in computer science and engineering from Seoul National University, South Korea, and his Ph.D. degree in informatics from the Graduate University for Advanced Studies, Japan. He joined the National Institute of Information and Communications Technology (NICT), Tokyo, Japan, as a researcher in 2006. He is currently working as a research manager at NICT, concurrently holding a visiting associate professor position at the University of Electro-Communications, Tokyo, Japan. His recent research interests include 5G and beyond network architectures, IoT, network resource management, network operation automation by AI/ML, and standardization. He has been serving as a Co-Rapporteur of ITU-T Study Group 13 since 2014. He received the ITU Association of Japan's Encouragement Award and Accomplishment Award in 2009 and 2017, respectively. He has also received four Best Paper Awards from the ITU Kaleidoscope Academic Conferences in 2009, 2014, 2018, and 2020. He is a Fellow of ITU-T Study Group 13 and a member of IEICE.

Takahiro Hirayama [M'18] received his M.S. and Ph.D. degrees in information science from Osaka University, Japan, in 2010 and 2013, respectively. He joined NICT as a researcher in 2013, where he is currently working as a senior researcher. His research interests are in optical networks, software defined networking, network function virtualization, and machine learning for network management. He received the Best Paper Award from the IEICE Communications Society in 2019. He is a member of IEICE.

Takaya Miyazawa [S'03, M'06] received his M.E. and Ph.D. degrees in information and computer science from Keio University, Yokohama, Japan, in 2004 and 2006, respectively. From April 2006 to March 2007, he was a visiting researcher at the University of California, Davis. He joined NICT as a researcher in 2007, where he is currently a research manager. From April 2019 to July 2020, he worked at the Ministry of Internal Affairs and Communications, Japan, as a deputy director, and returned to NICT in August 2020. His research interests include the automation of network control, management, and virtualization technologies. He served as Secretary of the IEEE Communications Society Asia-Pacific Board from 2016 to 2017. He received the Hiroshi Ando Memorial Young Engineer Award in 2007, the Funai Young Researcher Award in 2010, and the Best Paper Award from the ITU Kaleidoscope Academic Conference in 2018. He is a member of IEICE.

Masahiro Jibiki [M'21] received his Ph.D. degree in systems management from the University of Tsukuba, Japan, in 2003. He joined NEC Corporation in 1992 and worked as a researcher in the Central Research Laboratories until 2011. From 2006 to 2009, he was also a visiting professor at the University of Wakayama, Japan. Currently, he is a senior researcher at NICT. His research interests include networking, software science, and mathematical models. He received two Best Paper Awards from the IEICE Communications Society in 2010 and 2019. He is a member of IEICE.

Hiroaki Harai [M'97] received his M.E. and Ph.D. degrees in information and computer science from Osaka University, Japan, in 1995 and 1998, respectively. He is currently a director general of the Network Research Institute at NICT. His research interests include new network architectures and optical networks. He received the Outstanding Young Researcher Award from the IEEE Communications Society Asia-Pacific Region in 2007 and the Young Researcher Award from the Ministry of Education, Culture, Sports, Science and Technology, Japan, in 2009. He is a member of IEICE.